LABEL-EFFICIENT RISK CERTIFICATION FOR MEDICAL QUESTION ANSWERING AND TRIAGE VIA PREDICTIVE CALIBRATION ORDERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Approving medical AI policies for repeated use can require a high-confidence population-error certificate, even when expectation-based methods offer higher yield. With scarce calibration labels, however, these strict certificates often force low release rates. We introduce Predictive Calibration Ordering (PCO), which forecasts dependent certification outcomes from development data and uses exact dynamic programming to optimize a first-failure testing sequence, preserving the finite-sample selected-risk guarantee of Learn then Test. We evaluate PCO across medical question answering (HealthAdvice and 1,000 MedQA questions) and current-action triage (CARE-Bench). Across 80 target-averaged settings, Exact PCO achieves higher useful yield than Holm-certified adaptations of SCoRE, LEC and SCRC. At 20 labels and 90% certification confidence, MedQA useful yield reaches 51.52% and 43.12% for two models, versus 18.68% and 12.20% for SCoRE+Holm. Matched controls isolate joint planning, with small-budget gains across all three benchmarks of up to 2.4 percentage points over marginal ordering and up to 7.8 over Learn then Test. By increasing certified yield from scarce labels, PCO could help clinical AI systems release more answers to patient questions and triage recommendations without individual review.

1 INTRODUCTION

Clinical AI deployment (Singhal et al., 2023) requires both useful decisions and a rule for when to release them. With a fixed labeling budget, tighter error targets usually reduce the fraction of decisions that can be certified. We call the attained coverage and useful yield across these targets the *clinical autonomy–risk frontier*. Improving that frontier means releasing more useful decisions at the same risk target, calibration confidence and label budget.

We study the operational decision to approve one calibrated release policy for repeated use. The decision maker specifies an acceptable population error rate α and a maximum probability δ that calibration approves a rule exceeding it. A bound on expected error can coexist with above-target policies in some calibration outcomes. Repeated use makes this distinction consequential: one unfavorable approval can affect many future cases. The objective is therefore to maximize useful release subject to this approval requirement.

Patient-portal communication supplies a concrete opportunity. AI-generated drafts already reach physician inboxes for review (Tai-Seale et al., 2024). Consider a prospective service answering routine questions about existing, clinic-approved follow-up instructions. It would release a certified subset and route remaining messages to clinicians. Approval would concern the subset's rate of clinically meaningful errors before repeated use. Section 6 describes how PCO could support this decision.

Our benchmarks evaluate the two decisions such a service would make: answering patient-posed medical questions (HealthAdvice) and recommending a triage action (CARE-Bench), with 1,000 MedQA questions as a larger transfer.

Fixed-sequence Learn then Test (LTT) permits development-learned testing orders while controlling certification error (Bauer, 1991; Angelopoulos et al., 2025). Its first failed test ends the sequence,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

potentially blocking useful later rules. Marginal predictive ordering evaluates candidates individually. PCO instead forecasts their joint passage patterns and optimizes the useful mass released by the complete stopped sequence. Its exact dynamic program uses dependence to improve certification efficiency while retaining LTT's risk guarantee.

PCO contributes a joint-sequence optimizer, an implementation-aware utility certificate and three-benchmark evidence. We compare certified autonomy–risk curves, isolate planning with policy, forecast and labels fixed, and report native SCoRE/LEC/SCRC trade-offs. At 20 labels and 90% certification, Exact achieves 51.52% useful yield on MedQA/GPT-4o, versus 18.68% for SCoRE+Holm. Across all three benchmarks, small-budget planning gains are positive over both marginal ordering and matched LTT.

2 CERTIFYING A RELEASE POLICY

Let $Z = (X, Y) \sim P$. A development-fitted policy supplies a decision $f(X)$, risk score $s(X) \in [0, 1]$, and binary loss $L(f(X), Y)$. The loss marks a problematic answer on HealthAdvice, an incorrect current action on CARE, or an incorrect MedQA answer. A development-fixed acquisition envelope E restricts release to

$$\begin{aligned} A_t(X) &= \mathbf{1}\{X \in E,\ s(X) \le t\}, & C(t) &= \mathbb{E}[A_t(X)], \\ R(t) &= \mathbb{E}[L \mid A_t(X) = 1], & U(t) &= \mathbb{E}[A_t(X)(1 - L)] = C(t)(1 - R(t)). \end{aligned} \tag{1}$$

Coverage C measures decision autonomy, the fraction released without per-case review; useful yield U counts released decisions without a benchmark error over the full population. Abstention gives zero coverage and yield. CARE counts a correct clarification as a useful current action; episode completion is evaluated separately.

For risk target α and certification confidence $1 - \delta$, probability over the independent calibration sample must satisfy

$$\Pr\{\widehat{t} \neq \varnothing,\ R(\widehat{t}) > \alpha\} \le \delta. \tag{2}$$

For a concrete distinction, suppose a rule has the same positive coverage in every calibration outcome, but selected population risk 0 in 80% of outcomes and 5α in 20%, with $0 < \alpha \le 0.2$. Its expected risk equals α, yet its probability of exceeding α is 0.2, violating Equation 2 whenever $\delta < 0.2$. Averaging over other calibration outcomes does not correct the risk of the particular rule approved for deployment.

For development-fixed thresholds $t_1 < \cdots < t_m$, selected calibration counts N_j and error counts K_j define

$$p_j = F_{\mathrm{Bin}(N_j,\alpha)}(K_j), \qquad B_j = \mathbf{1}\{N_j \ge 10,\ p_j < \delta\}. \tag{3}$$

Here $F_{\mathrm{Bin}(N,\alpha)}$ is the binomial CDF. Even with zero errors, strict rejection requires $(1 - \alpha)^N < \delta$. The minimum feasible support is therefore $n_{\min} = \max\{10, \lfloor \log \delta / \log(1 - \alpha) \rfloor + 1\}$; at $\alpha = \delta = 0.10$, it is 22. A cap below this forces abstention regardless of ordering.

The procedure follows a frozen sequence until its first failed test. Among rejected hypotheses it selects the cutoff with greatest calibration support, breaking ties toward the smaller cutoff. With no rejection, it abstains everywhere.

Focused acquisition preserves the risk target. Calibration scans an iid pool in a label-blind order and acquires up to n labels inside E. Conditional on the accepted count, observations follow $Q_E = P(\cdot \mid E)$. For every release set $A \subseteq E$,

$$Q_E(L = 1 \mid A) = P(L = 1 \mid A). \tag{4}$$

Focusing therefore preserves selected risk. Population utility still includes the factor $P(E)$, and finite-pool shortages can reduce actual labels below the cap. The primary comparison uses one shared, development-selected envelope for every planner (Appendix C.2).

Proposition 1 (Inherited fixed-sequence validity)**.** *Conditional on development data and planning randomness, suppose the policy, envelope, thresholds and sequence are fixed independently of an iid calibration pool. Valid selected-risk p-values and first-failure stopping imply Equation 2, including under focused acquisition. Assume $P(E) > 0$; otherwise the controller abstains. Risk is relevant only for positive-mass release sets.*

Rejecting any true null requires rejecting the first true null in the sequence, with probability at most δ. The support restriction only removes rejections. The proof, including focused acquisition, is given in Appendix A.

3 PREDICTIVE CALIBRATION ORDERING

3.1 FORECAST THE JOINT CERTIFICATION OUTCOME

The controller forecasts how much useful mass each sequence would release at the available label budget. Development scores and losses define a predictive distribution inside E, augmented by local loss-quantile pseudo-observations. Each predictive world samples conditional case weights, the envelope's population mass, the eligible finite-pool count, and a calibration sample. It produces passage indicators $B_j^{(b)}$ and population useful masses $u_j^{(b)}$ for all cutoffs simultaneously. Sharing a world across nested thresholds preserves their dependence.

The Exact experiments use 2,048 training worlds and 2,048 independent check worlds. Candidates come from 21 development-score quantiles within the envelope plus endpoints, with at least five development cases except for the mandatory terminal cutoff. The terminal cutoff is included inside the optimizer. The local predictive distribution permits fractional simulated losses; calibration tests use integer binary errors. All planners receive identical forecasts and candidate arrays. Appendix C specifies the conditional simulator and every constant.

3.2 OPTIMIZE USEFUL RELEASE AFTER THE FIRST FAILURE

For an increasing sequence $S = (j_1, \ldots, j_k)$, set $j_0 = 0$ and $u_0^{(b)} = 0$. The last-passing-threshold surrogate is

$$\widehat{F}(S) = \frac{1}{B} \sum_{b=1}^{B} \sum_{\ell=1}^{k} \left(\prod_{h=1}^{\ell} B_{j_h}^{(b)} \right) \left(u_{j_\ell}^{(b)} - u_{j_{\ell-1}}^{(b)} \right). \tag{5}$$

An increment contributes only in worlds that pass the entire preceding sequence. Marginal predictive ordering ranks candidates by their individual expected useful release,

$$q_j = \frac{1}{B} \sum_{b=1}^{B} u_j^{(b)} B_j^{(b)}. \tag{6}$$

Development-p-value LTT supplies another ordering control. Marginal and LTT retain ascending record thresholds and complete the remaining higher tail, sharing Exact's candidates and terminal cutoff. This completion is fixed before calibration; every resulting sequence still stops at its first failed test (Appendix A).

Why marginal power is insufficient. Consider two thresholds with utilities $(0.2, 0.5)$ and pass probabilities $(0.8, 0.6)$. Testing only the second yields 0.30. Testing both yields $0.16 + 0.30 \Pr(B_1 = B_2 = 1)$: 0.28 when joint passage is 0.4, but 0.34 when it is 0.6. Identical marginal probabilities can favor different sequences. This mechanism motivates joint forecasting; its empirical value is tested with the forecast held fixed.

Compress worlds by joint pass pattern. A state (i, A) records the last threshold i and surviving worlds A. With $A_j = \{b \in A : B_j^{(b)} = 1\}$, the recurrence is

$$V(i, A) = \max \left\{ 0, \max_{j > i} \left[\frac{1}{B} \sum_{b \in A_j} (u_j^{(b)} - u_i^{(b)}) + V(j, A_j) \right] \right\}. \tag{7}$$

The entry state is $V(0, \{1, \ldots, B\})$, with $u_0^{(b)} = 0$; $V(m, A) = V(i, \varnothing) = 0$. Empty continuation maxima contribute zero. Reconstruction resolves equal objectives by fewer tests, then lexicographically larger indices; the implementation uses tolerance 10^{-13}. The terminal candidate belongs to the optimization family, and completing a tied optimum to it does not reduce the surrogate.

Theorem 1 (Finite-simulation optimality)**.** *In exact arithmetic, for fixed pass indicators and nonnegative, nondecreasing utilities, Equation 7 globally maximizes Equation 5 over increasing subsequences of the eligible simulated thresholds.*

Backward induction over the next cutoff proves the result; memoization reuses surviving-world sets (Appendix A). The last-passing surrogate telescopes with state (i, A). The implemented selector can retain a narrower cutoff at support ties, motivating the correction below. Direct optimization of that selector would augment the state with each surviving world's retained support-tie choice.

3.3 Connect planning to the implemented controller

Let $\mathcal{S}$ contain the ascending sequences ending at the terminal candidate. In each predictive world, write $F_b(S)$ for last-passed utility and $G_b(S)$ for the implemented support-selected utility, with hats denoting world averages. Nested nonnegative useful mass gives $G_b(S) \le F_b(S)$. For the returned plan $\widehat{S}$, define tie loss $\widehat{\tau} = \widehat{F}(\widehat{S}) - \widehat{G}(\widehat{S})$ and let ζ bound its surrogate optimization residual. The operational certificate is

$$\max_{S \in \mathcal{S}} \widehat{G}(S) - \widehat{G}(\widehat{S}) \le \widehat{\tau} + \zeta. \tag{8}$$

An exact surrogate optimum with zero tie loss is therefore also an exact operational optimum on the predictive bank. Appendix A.4 proves this certificate and extends it to new forecast worlds and population utility under an explicit forecast-discrepancy assumption. Appendix A.5 gives the variance-sensitive bound for a fixed plan pair. These utility results quantify planning quality; Proposition 1 supplies the separate calibration risk guarantee. Held-out experiments measure the realized useful-yield gain. Tie-loss diagnostics appear in Appendix E.1.

A shared architecture for policy selection. The matched ordering contrast fixes the clinical policy. An Expanded arm plans within each development-defined policy template, then uses independent check worlds to choose one complete template for calibration. Every planner receives the same bank and check budget. Selecting one template before calibration adds no calibration multiplicity. Appendix C.2 gives a finite-bank selection bound.

4 Experimental design

Three benchmarks and frozen predictions. HealthAdvice contains 222 patient-posed medical questions with four physician-evaluated model responses each (Draelos et al., 2026). A development-trained warning selector chooses the response and score. Two configurations each use five grouped views with 97–98 development, 80 calibration and 44–45 evaluation questions. CARE's pinned snapshot contains 439 patients and 925 prefixes (Hua et al., 2026). Two triage action-model families train on the same 100 development patients; five allocations rotate the other 339 between 271–272 calibration and 67–68 evaluation patients, using one fixed certification prefix per patient. Repeated views are grouped by question or patient.

The larger transfer uses 1,000 unique MedQA-USMLE questions (Jin et al., 2021) with cached four-option answers from GPT-4o and GPT-4o-mini, whose unfiltered accuracies are 87.3% and 76.1%. Each model has two configurations of five folds, with 400 development, 400 calibration and 200 evaluation questions per view. We reuse the original development-fitted scalar error scores and calibration permutations. The transfer applies the manuscript's Exact implementation verbatim; the cached Fractional Joint PCO remains a separate reference. No additional language-model calls or scorer fits are used for these comparisons. Appendix B.1 records prediction provenance; Appendix G specifies the transfer and adaptations.

Matched resources and operating grid. Label caps are $\{20, 40, 80\}$ for HealthAdvice, $\{20, 40, 80, 160, \text{full}\}$ for CARE, and $\{20, 40, 80, 160, 320, 400\}$ for MedQA, additional to development labels. All experiments retain targets $\alpha \in \{0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}$ and

$\delta \in \{0.10, 0.05\}$. Reduced caps use 200 shared label-blind calibration permutations; full pools use one. Actual labels and screened cases are retained when focusing exhausts the eligible pool.

The HealthAdvice/CARE factorial crosses Current/Expanded policies, LTT/Marginal/Exact planning and uniform/shared-focused acquisition. The ordering comparison fixes Current policy. The native-method comparison uses the same Expanded Exact-selected policy, score, envelope and labels for every method; it does not independently optimize each baseline's policy bank. Uniform acquisition is primary on MedQA, preserving its cached design; shared focusing is a separate transfer. MedQA has no Expanded answer-policy bank. Both acquisition rules and all budgets remain in the complete results.

Native and certified comparators. SCoRE-SDR controls expected batch selected loss; LEC controls expected selected errors relative to expected selections (Bai & Jin, 2026; Wang et al., 2026). Neither criterion automatically supplies Equation 2. We report SCoRE's designated homogeneous boosted version in the main tables, retaining general and binary variants. SCRC-I provides a calibration-only PAC guarantee for prediction sets on selected cases (Xu et al., 2025). Our caches supply scalar correctness scores, so its explicit binary adaptation forms sets over incorrect/correct and releases only admitted singleton $\{\text{correct}\}$ answers. Its selected-set theorem does not automatically certify this extra singleton filter's answer-error risk; we label it SCRC† throughout.

For a common population-risk requirement, a separate panel applies strict exact-binomial Holm testing to the same development-fixed threshold family. Each native proposal selects an entire simultaneously certified grid policy whose batch release set is contained in its proposal, maximizing calibration support with the common tie rule. Selection from this family satisfies Equation 2, even if the proposal uses calibration or unlabeled batch scores. Appendix G.2 proves the guarantee for the full selected grid policy and specifies the selection rule. These are comparisons of certification wrappers. Holm's initial cutoff is δ/m, followed by step-down thresholds; the development-fixed PCO sequence tests at δ and stops on failure. The yield gap therefore includes the different multiplicity/power trade-offs. Holm grid without a proposal cap isolates the certifier's contribution. Native decision rules are evaluated separately in Section 5.3; Exact–Marginal and Exact–LTT isolate ordering under the same fixed-sequence requirement.

Outcomes and uncertainty. Useful yield is the fraction of all evaluation cases released without a benchmark error; abstention is one minus coverage. Pooled error divides error mass by release mass. Exceedance frequency is the fraction of recorded held-out batches whose released-error fraction exceeds their own α; empty batches count as no exceedance. Severity is the mean excess above α conditional on an exceedance, and is undefined if none occurs. These dependent finite-batch diagnostics do not estimate the population false-certification probability in Equation 2 and are not compared directly with δ.

Target-averaged summaries weight all six targets equally. Paired 95% intervals use 5,000 question/patient-cluster bootstrap draws, conditional on cached fits, plans and calibration realizations. The primary factorial retains its 756-contrast simultaneous family. The clinical comparator analysis retains its complete 2,412-contrast audit family, including archived comparisons beyond those displayed here. MedQA uses one 456-contrast family across both models, acquisition rules, budgets and confidence levels, with shared question resampling. These conditional performance intervals are distinct from risk-certification confidence; Appendix B.3 gives the family definitions and critical values.

5 Results

5.1 The certified autonomy–risk frontier

Figure 1 makes the frontier explicit: at a fixed cap of 20 labels and 90% certification, it plots coverage and useful yield over every tested risk target. The comparison uses the same release policy, scores and calibration labels within each panel. PCO attains higher coverage and useful yield across the nonzero portions of the curves than the three tested Holm adaptations. The horizontal axis is the requested risk tolerance; observed errors are reported separately in Table 2 and Appendix D.2.

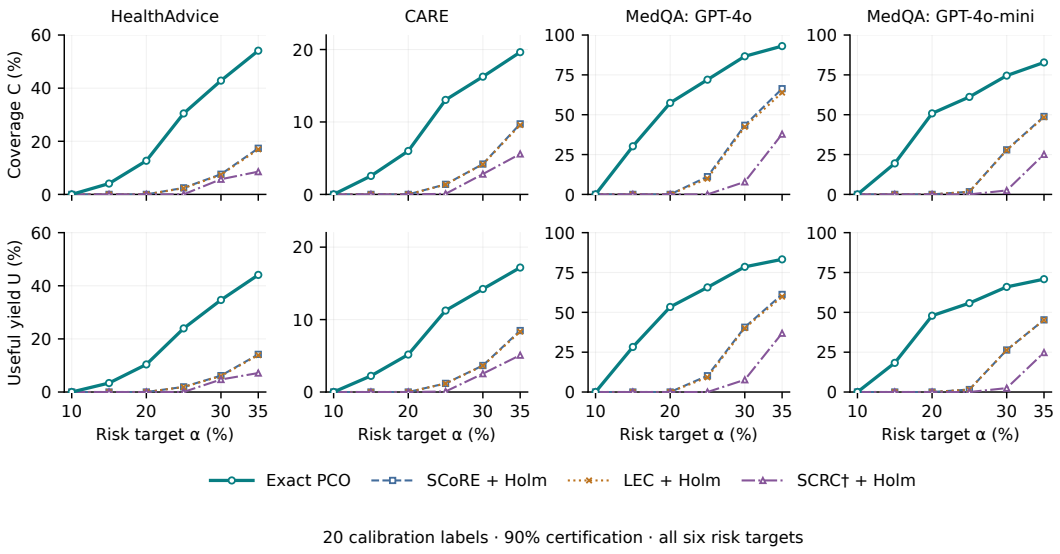

Figure 1: **The autonomy–risk frontier at a fixed certification budget.** Coverage C (top) and useful yield U (bottom) versus risk target α, using 20 calibration labels and 90% certification. Points are observed means at all six targets; connecting lines guide the eye. HealthAdvice/CARE use Expanded policy and shared focusing; MedQA uses fixed cached answers and uniform acquisition. Every curve satisfies the same population-risk requirement using its stated certification procedure. SCRC† is the correctness-set adaptation. Paired and simultaneous uncertainty is reported for the target-averaged comparisons.

Table 1: **Useful yield (%) under the same 90% population-risk requirement.** HealthAdvice/CARE use Expanded policy and shared focusing; MedQA uses uniform acquisition. All six risk targets are averaged. +H denotes the stated Holm adaptation; Holm is the uncapped grid certifier. SCRC† denotes binary correctness sets with singleton release.

Benchmark / model	Cap	Exact	Holm	SCoRE+H	LEC+H	SCRC†+H
HealthAdvice	20	19.41	3.71	3.71	3.63	1.98
HealthAdvice	40	31.03	12.95	12.95	12.69	8.79
CARE	20	8.34	2.22	2.22	2.19	1.28
CARE	40	12.19	7.71	7.71	7.58	5.21
MedQA / GPT-4o	20	51.52	18.68	18.68	18.20	7.43
MedQA / GPT-4o	40	73.56	50.36	50.35	49.80	35.32
MedQA / GPT-4o-mini	20	43.12	12.20	12.20	12.17	4.54
MedQA / GPT-4o-mini	40	66.18	42.08	41.77	42.06	32.71

Table 1 compares the specified certification wrappers under Equation 2; Table 2 evaluates the native release rules. At 20 labels and 90% certification, Exact yields 19.41% and 8.34% useful releases on HealthAdvice and CARE, versus 3.71% and 2.22% for SCoRE+Holm. On MedQA, the corresponding yields are 51.52% versus 18.68% for GPT-4o and 43.12% versus 12.20% for GPT-4o-mini. These are differences between specified certification procedures; the smaller matched Exact–Marginal gains isolate joint planning itself. Section 5.2 compares Exact with matched fixed-sequence LTT under the same requirement.

Across both acquisition rules, all budgets and both confidence levels, Exact exceeds each of the three Holm-certified adaptations in all 32 Expanded HealthAdvice/CARE cells and all 48 MedQA cells. Every corresponding interval is positive in its study's simultaneous family: 96 clinical and 144 MedQA contrasts. The yield gap often narrows with more labels: at 400 MedQA labels and 90%, Exact versus SCoRE+Holm is 86.50% versus 86.08% for GPT-4o and 75.08% versus 74.68% for GPT-4o-mini. The largest automation gains occur where calibration labels are scarce.

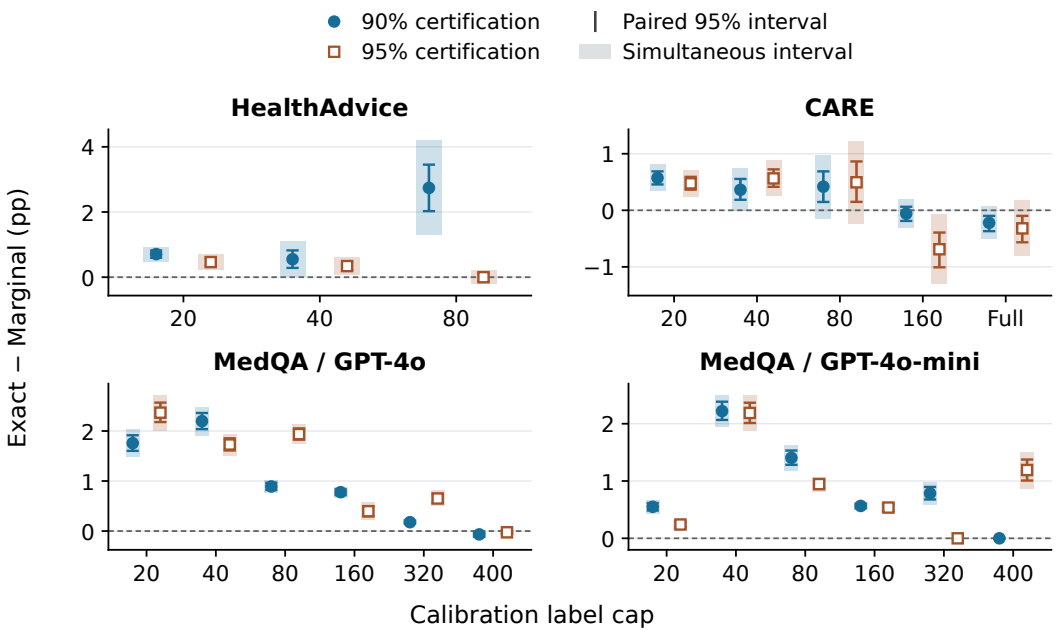

Figure 2: **Matched planning gains across every label cap.** Exact minus Marginal useful yield, averaged over all six targets. HealthAdvice/CARE fix Current policy and shared focusing; MedQA fixes cached answers and uniform acquisition. Capped bars show paired 95% intervals; shaded bands show simultaneous intervals in the 756-contrast and MedQA 456-contrast families. CARE Full is 271–272 available labels. Focusing can use fewer than the nominal cap.

5.2 Joint planning improves small-budget certification

Under shared focused acquisition, Current Exact exceeds matched Marginal at caps 20 and 40 on HealthAdvice and caps 20, 40 and 80 on CARE, at both certification levels: ten gains of 0.34–0.71 pp, all with positive paired intervals, seven with positive simultaneous intervals. At 20 labels and 90% certification, useful yield increases from 17.65% to 18.36% on HealthAdvice and from 7.77% to 8.34% on CARE. These comparisons hold clinical policy, forecast, thresholds, acquired labels and screening fixed; the added resource is offline planning.

On MedQA under uniform acquisition, the matched ordering gains at 20 and 40 labels range from 0.24 to 2.37 pp. All eight model/budget/confidence comparisons have positive paired and simultaneous intervals. At 20 labels and 95% certification, GPT-4o gains 2.37 pp, with simultaneous interval $[2.02, 2.72]$; at 40 labels and 90%, GPT-4o-mini gains 2.22 pp, with interval $[1.94, 2.50]$. Thus the joint-planning effect transfers to a larger independent question set and is visible beyond 20 labels.

Matched LTT. With the same policy, forecast, candidates and labels, Exact also exceeds development-p-value LTT at 20 labels on all three benchmarks. At 90% certification, useful yield rises from 16.32% to 18.36% on HealthAdvice and from 0.53% to 8.34% on CARE, with positive simultaneous intervals at both certification levels, and CARE remains positive at 40 labels. On MedQA, all eight 20–40-label model and confidence comparisons favor Exact (Table 19).

Budget dependence. Figure 2 retains the complete cap grid. Current Exact's full-cap CARE gains are -0.22 and -0.32 pp at 90% and 95%, with negative paired intervals; uniform HealthAdvice/CARE findings are mixed. Across all 24 uniform MedQA model/budget/confidence cells, Exact exceeds matched Marginal in 20, falls below it in three and ties once; 20 simultaneous intervals are positive and none negative. Full grids appear in Appendices D and G.3.

Choosing a planner at the intended budget. The demonstrated small-budget regime is 20–40 labels; nominal $n \geq 100$ alone is not a switching criterion because selected support, score quality

Table 2: **Native yield–error trade-offs at 20 labels.** Same panels and target averaging as Table 1; Exact uses 90% certification. Error is pooled error among released cases. Exceed is held-out batch frequency above that batch's target; Excess is conditional overshoot severity. A dash means no exceedance. The diagnostic definitions and their conditioning are given in Section 4. SCRC† is the correctness-set/singleton adaptation.

Method	Yield (%)	Abstain (%)	Error (%)	Exceed (%)	Excess (pp)
HealthAdvice					
Exact	19.41	75.97	19.22	3.03	3.24
SCoRE	51.26	37.15	18.43	16.18	5.99
LEC	47.56	41.99	18.02	13.91	6.11
SCRC†	51.47	38.57	16.21	28.08	7.99
CARE					
Exact	8.34	90.42	12.92	1.83	2.17
SCoRE	17.12	80.35	12.88	11.51	4.43
LEC	16.23	81.56	11.98	9.37	4.07
SCRC†	18.58	78.74	12.63	20.71	9.91
MedQA / GPT-4o					
Exact	51.52	43.44	8.90	0.00	–
SCoRE	82.26	6.70	11.84	6.60	2.67
LEC	80.55	11.51	8.97	2.01	1.78
SCRC†	67.43	30.32	3.23	0.12	0.45
MedQA / GPT-4o-mini					
Exact	43.12	51.86	10.42	0.55	2.94
SCoRE	70.32	13.81	18.40	17.86	5.28
LEC	71.00	17.73	13.69	5.65	3.56
SCRC†	63.53	34.83	2.51	0.08	0.10

and target also determine power. At larger budgets, compare Exact, Marginal and development-order LTT on a separate development validation split at the intended (n, α, δ), then freeze one sequence before final calibration. Reserve that validation split from policy fitting and planning, or obtain it separately. This is deployment guidance, not a newly evaluated switching algorithm. A failed calibration test ends the chosen sequence; switching or combining tests afterward requires an error allocation specified in advance. Holm is an alternative when simultaneous certification of the whole grid is desired.

LTT and Pareto comparisons. The supplementary uniform nine-method comparison provides additional evidence. At 20 labels and 90% certification, Joint fractional PCO yields 16.93% on HealthAdvice versus 15.45% for LTT and 15.53% for Pareto Testing; the CARE means are 2.07%, 0.00% and 0.00%. Appendix J retains the nine-method comparison, including all budgets, both confidence levels, other baselines and PCO variants under their recorded protocols.

5.3 Yield and error under native release rules

Table 2 reports the native release rules before adding a population-risk certificate. At 20 labels, native SCoRE and LEC typically release substantially more correct answers and abstain less than Exact. On HealthAdvice, Exact's held-out exceedance frequency is 3.03%, compared with 16.18% for SCoRE and 13.91% for LEC; conditional excess is 3.24 pp versus 5.99 and 6.11 pp. CARE shows the same direction, with frequencies 1.83%, 11.51% and 9.37%. Table 2 reports yield and abstention alongside these frequencies.

On uniform MedQA, native SCoRE adds 27–31 correct releases per 100 questions relative to Exact, with about 6–11 additional wrong releases; LEC adds 28–29 correct and 3–6 wrong releases. SCRC† is especially competitive: it adds 16–20 correct releases while releasing about three fewer wrong answers, with exceedance frequencies of 0.12% and 0.08%, compared with Exact's 0% and 0.55%. The native results quantify the automation available under each method's contract; the certified comparison in Section 5.1 addresses the reusable-policy approval requirement. SCRC's set

diagnostics, other SCoRE variants, both acquisition rules and target-specific results are retained in Appendix G.3 and the source.

5.4 What accounts for the gains?

At 20 labels and 90%, the Expanded pipeline exceeds the highest observed simple-method mean by 1.10 pp on HealthAdvice and 0.57 pp on CARE, or about 11 and 6 useful releases per 1,000 decisions. HealthAdvice includes a 1.05 pp policy-expansion contribution; CARE's corresponding contribution is 0.0004 pp. The reference is an observed envelope across recorded simple methods, recomputed within each bootstrap, not one selected deployment policy. Appendix D separates policy, ordering and reference effects and retains all larger-budget gaps.

Saved stopping traces support the proposed mechanism. At 20 labels/90%, the frequency of missing an individually passing candidate falls from 9.10% to 7.91% on HealthAdvice and from 16.45% to 14.17% on CARE. This reduction occurs in all ten highlighted small-budget cells. Appendix E gives the full stopping decomposition, budget crossover and independent forecast checks. Across 960 recorded focused Current conditions, median planning time is 46.6 ms for Exact and 13.9 ms for Marginal; the median paired increment is 26.8 ms. These timings include sequence construction and forecast-utility evaluation, excluding the shared forecast generation. The exact optimizer remains exponential in the worst case. Appendix E.3 supplies a joint-increment greedy alternative with $O(Bm^2)$ cost and a cache-based timing demonstration through 100 requested quantiles.

6 Related work and deployment rationale

LTT provides risk certification from valid tests, including fixed sequences (Angelopoulos et al., 2025); RCPS provides high-probability population-loss bounds (Bates et al., 2021). Conformal risk control targets expectation guarantees for suitable losses (Angelopoulos et al., 2024). SCoRE, LEC and SCRC address selective risk through different functionals and output objects, motivating our separate native and adapted comparisons. Pareto and guided testing optimize configuration search (Laufer-Goldshtein et al., 2023; 2024); adaptive LTT develops sequential evidence (Zecchin et al., 2025); reliability-graph testing uses configuration structure with an FDR formulation (Farzaneh & Simeone, 2025). Our FWER adaptations are identified explicitly. PCO's distinct contribution is optimization of dependent passage events for a development-fixed first-failure sequence.

When to choose PCO. PCO fits deployments requiring Equation 2 with development-trained policies and scarce additional certification labels. Higher native yield alone does not satisfy that approval requirement. If expectation-based control meets the objective, the larger native SCoRE/LEC yields favor those methods. LTT and other valid certifiers remain alternatives under Equation 2; PCO contributes joint planning efficiency. Across all three benchmarks, its most consistent matched gains occur at 20–40 labels; Section 5.2 gives development-only planner selection for other budgets. The Holm conversions compare particular certification procedures; LTT/Pareto provide closer controls.

A prospective portal deployment. For the proposed portal service, clinicians would define eligible requests, a loss for clinically meaningful inaccuracies and omissions, and institution-specific (α, δ). Development data would fix the model, score, envelope and testing order; independent, clinician-labeled calibration cases would certify one cutoff. Section 2's support bound determines the required labels. The frozen rule would release eligible replies below that cutoff and route all others for clinician review. The opportunity is useful replies without individual pre-release review, under a bound on approving an above-target rule. Scarce local labels make planning efficiency relevant even when many future messages use the approved rule. The HealthAdvice and CARE results show this gain on answers to patient questions and triage recommendations; a prospective study would measure it on live messages.

The certificate concerns the selected population under the calibration distribution, without protection against shift or bounds for each patient or batch. Model or population changes require renewed validation. CARE's current-action endpoint is distinct from episode completion, below 0.4% (Appendix H). A prospective portal study would measure clinical error, handoff burden and review time at prespecified clinical tolerances; these deployment outcomes are not measured by the benchmarks.

AI USE STATEMENT

Generative AI tools were used to assist with proofreading, wording revisions, and figure generation.

ETHICS STATEMENT

This is a retrospective analysis of research benchmark material, without a new patient intervention or newly recruited clinical reviewers. HealthAdvice's problematic-answer labels, CARE's action labels and MedQA's answer keys are proxies with different clinical meanings. A correct label does not certify the safety of associated prose or successful human handoff. No tested risk level is presented as a clinically endorsed tolerance. A clinical implementation would require independent validation, appropriate oversight, and explicit treatment of escalation and abstention. Reuse and redistribution of benchmark material remain subject to the source terms; this manuscript does not claim new ownership of those data.

REPRODUCIBILITY STATEMENT

An anonymized package containing all code, stored model outputs, frozen protocols and complete results will be provided to reviewers during the review period and released publicly upon publication. HealthAdvice (Draelos et al., 2026), CARE-Bench (Hua et al., 2026) and MedQA-USMLE (Jin et al., 2021) are public benchmarks; all language-model outputs are stored, so reproduction requires no API calls, and Appendix B.1 records model identifiers, prompts and input hashes. Self-contained notebooks reproduce each of the three studies on a CPU in a few hours, and build scripts regenerate every table and figure from saved outputs. The procedure, constants, seeds and uncertainty methods are in Appendices B–C. Replays establish computational consistency, not independent clinical replication.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2208.02814.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://www.nowpublishers.com/article/Details/MAL-101.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. *arXiv preprint arXiv:2603.24704*, 2026. URL https://arxiv.org/abs/2603.24704.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *arXiv preprint arXiv:2101.02703*, 2021. URL https://arxiv.org/abs/2101.02703.

P. Bauer. Multiple testing in clinical trials. *Statistics in Medicine*, 10(6):871–890, 1991. doi: 10.1002/sim.4780100609. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780100609.

Rachel L. Draelos, Samina Afreen, Barbara Blasko, Tiffany L. Brazile, Natasha Chase, Dimple Patel Desai, Jessica Evert, Heather L. Gardner, Lauren Herrmann, Aswathy Vaikom House, Stephanie Kass, Marianne Kavan, Kirshma Khemani, Amanda Koire, Lauren M. McDonald, Zahraa Rabeeah, and Amy Shah. Large language models provide unsafe answers to patient-posed medical questions. *npj Digital Medicine*, 9:241, 2026. doi: 10.1038/s41746-026-02428-5.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Amirmohammad Farzaneh and Osvaldo Simeone. Multi-objective hyperparameter selection via hypothesis testing on reliability graphs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1324. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/38853020527b0f84187870114ff7a686-Abstract-Conference.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Yining Hua, Hongbin Na, and Cyrus Ayubcha. CARE-Bench: Benchmarking patient-facing LLM triage. *arXiv preprint arXiv:2608.03731*, 2026. URL https://arxiv.org/abs/2608.03731.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. URL https://www.mdpi.com/2076-3417/11/14/6421.

Ying Jin and Emmanuel J. Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023. URL https://jmlr.org/papers/v24/22-1176.html.

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi Jaakkola. Efficiently controlling multiple risks with Pareto Testing. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cyg2YXn_BqF.

Bracha Laufer-Goldshtein, Adam Fisch, Regina Barzilay, and Tommi Jaakkola. Risk-controlling model selection via guided bayesian optimization. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2312.01692.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6147–6157, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://www.cs.mcgill.ca/~colt2009/papers/012.pdf.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pzUhfQ74c5.

Aaditya Ramdas, Jianbo Chen, Martin J. Wainwright, and Michael I. Jordan. A sequential algorithm for false discovery rate control on directed acyclic graphs. *Biometrika*, 106(1):69–86, 2019. doi: 10.1093/biomet/asy066. URL https://academic.oup.com/biomet/article/106/1/69/5303997.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023. doi: 10.1038/s41586-023-06291-2. URL https://www.nature.com/articles/s41586-023-06291-2.

Ming Tai-Seale, Sally L. Baxter, Florin Vaida, Amanda Walker, Amy M. Sitapati, Chad Osborne, Joseph Diaz, Nimit Desai, Sophie Webb, Gregory Polston, Teresa Helsten, Erin Gross, Jessica Thackaberry, Ammar Mandvi, Dustin Lillie, Steve Li, Geneen Gin, Suraj Achar, Heather Hofflich, Christopher Sharp, Marlene Millen, and Christopher A. Longhurst. AI-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Network Open*, 7(4):e246565, 2024. doi: 10.1001/jamanetworkopen.2024.6565. URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2817615.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, first edition, 2005. doi: 10.1007/b106715. URL https://link.springer.com/book/10.1007/b106715.

Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 2022. doi: 10.1111/rssb.12489. URL https://academic.oup.com/jrsssb/article/84/3/822/7056146.

Zhiyuan Wang, Aniri, Tianlong Chen, Yue Zhang, Heng Tao Shen, Xiaoshuang Shi, and Kaidi Xu. LEC: Linear expectation constraints for selection-conditioned risk control in selective prediction and routing systems. *arXiv preprint arXiv:2512.01556v3*, 2026. URL https://arxiv.org/abs/2512.01556v3.

Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control. *arXiv preprint arXiv:2512.12844*, 2025. URL https://arxiv.org/abs/2512.12844.

Matteo Zecchin, Sangwoo Park, and Osvaldo Simeone. Adaptive Learn-Then-Test: Statistically valid and efficient hyperparameter selection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74018–74036, 2025. URL https://proceedings.mlr.press/v267/zecchin25a.html.

A PROOFS AND SCOPE OF THE OPTIMIZATION CLAIMS

A.1 RISK CERTIFICATION

Condition on the development sample, all fitted components, and planning randomness. The candidate family and order are now fixed. Under the iid assumption, the released losses for a fixed threshold are independent Bernoulli observations with its selective risk, conditional on their count. Their lower-tail binomial p-value is super-uniform under $R(t) \geq \alpha$. Let j^* be the first true null in the tested order. If any true null is rejected, first-failure stopping requires $p_{j^*} < \delta$. That event has probability at most δ. Selecting any rejected hypothesis, including using calibration support to select among rejections, inherits this simultaneous guarantee for the rejected family. If no true null occurs or no hypothesis is rejected, there is no erroneous certification. Averaging over independent development randomness preserves the bound. A support filter adds possible stopping events and cannot invalidate it. The predictive forecast may be misspecified.

This result requires an iid calibration population independent of the complete planned procedure. Exchangeability alone without the conditional binomial model is insufficient for this particular test. Case grouping prevents treating correlated answer candidates or conversation prefixes as independent trials. A different α, a different confidence setting, or a different fitted view is a separate procedure. Reporting an entire grid does not supply a joint guarantee for deployment selection over that grid.

A.2 EXACT OPTIMIZATION OF THE JOINT SURROGATE

For simulation b, let r_b be the largest passed prefix position in S. The sum in Equation 5 telescopes to $u^{(b)}_{j_{r_b}}$, with zero for $r_b = 0$. At a state (i, A), simulations outside A have already stopped and their achieved utilities cannot change. Choosing a next threshold $j > i$ increases utility only for A_j, by $u_j - u_i$. Those simulations enter the subproblem (j, A_j). All possible increasing continuations are enumerated by this choice, or by stopping with no further gain. Induction over the number of remaining thresholds proves the recurrence and global optimality over the finite candidate family.

Grouping simulations by pass pattern does not change the objective: the implementation sums their utility vectors with mass $1/B$ and retains each pattern's common passage mask. Utilities may be correlated with passage, and no independence factorization is required. Nonnegative utility increments permit appending the largest simulated threshold without lowering the surrogate. In the primary conditional experiment, the terminal cutoff is included in the optimization family. Joint fractional PCO instead appends the $t = 1$ fallback after optimization; that separate endpoint is outside its simulation-optimality statement. These surrogate-optimality statements alone do not imply optimality under the operational smaller-threshold tie rule or the true data distribution; Appendix A.4 supplies the operational certificate.

The memoized state is a threshold index and an active set of pass patterns. The number of reachable active sets may be exponential in the threshold count. Our families are small development-quantile grids; the method is not presented as a general polynomial-time optimizer for arbitrary hypothesis families.

A.3 SEQUENCE COMPLETION

For an ordered list of threshold indices, retain each index that exceeds every preceding retained index, then append indices higher than its last retained index. An empty order remains empty. A discarded non-record threshold cannot broaden the release set. If such a test fails in the original order, removing it permits continued testing; if it passes, the original selected set is no broader than one already available. A retained record that fails is encountered only after the original procedure either reached it or stopped earlier. Appending the higher tail cannot erase a previous rejection. Consequently completion cannot lower released coverage or useful release for a fixed labeled population under nested release sets and nonnegative per-case usefulness. It can affect different orderings by different amounts, as the HealthAdvice full-budget results show. Its validity still follows from a development-fixed sequence and first-failure stopping.

A.4 Operational utility certificate and forecast regret

Let $\mathcal{S}$ contain all ascending subsequences of the m eligible candidates ending at terminal m, so $M = 2^{m-1}$. Let $F_b(S)$ denote last-passed utility and $G_b(S)$ support-selected utility in world b, and use hats for their B-world averages. Ascending order and nested nonnegative useful mass imply $G_b(S) \le F_b(S)$. Define $G_Q(S) = \mathbb{E}_{W\sim Q} G_W(S)$ and $G_P(S) = \mathbb{E}_{\mathcal{D}_{\text{cal}}\sim P} G_{\mathcal{D}_{\text{cal}}}(S)$, with development fixed and population utility evaluated under the respective process. For the returned plan $\widehat{S}$, let $\zeta \ge \max_S \widehat{F}(S) - \widehat{F}(\widehat{S})$ bound its optimization residual and define empirical tie loss $\widehat{\tau} = \widehat{F}(\widehat{S}) - \widehat{G}(\widehat{S}) \ge 0$.

Theorem 2 (Implementation-aware utility regret)**.** *Suppose utility and support are nondecreasing, utilities lie in* $[0,1]$*, and* $\mathcal{S}$ *is the development-fixed ascending sequence family just defined. For* B *iid predictive worlds from* Q*, with probability at least* $1-\eta$*,*

$$\max_{S\in\mathcal{S}} G_Q(S) - G_Q(\widehat{S}) \le \widehat{\tau} + \zeta + 2\sqrt{\frac{\log(2M/\eta)}{2B}}. \tag{9}$$

If $\sup_S |G_P(S) - G_Q(S)| \le \epsilon_{\text{forecast}}$*, the population regret adds at most* $2\epsilon_{\text{forecast}}$*.*

The terminal candidate participates inside the primary optimizer. The following proof also establishes the deterministic operational certificate used in the main text.

Proof of Theorem 2 and Equation 8. For every $S \in \mathcal{S}$,

$$\widehat{G}(S) \le \widehat{F}(S) \le \widehat{F}(\widehat{S}) + \zeta \tag{10}$$

$$= \widehat{G}(\widehat{S}) + \widehat{\tau} + \zeta. \tag{11}$$

Maximizing the left side proves the deterministic certificate. In particular, $\widehat{\tau} = \zeta = 0$ implies that $\widehat{S}$ maximizes empirical operational utility, even though the optimizer is written for F. For a fixed S, the B independent values $G_b(S)$ lie in $[0,1]$. Hoeffding's inequality (Hoeffding, 1963) and a union bound give

$$\Pr\left\{\sup_{S\in\mathcal{S}} |\widehat{G}(S) - G_Q(S)| > \sqrt{\frac{\log(2M/\eta)}{2B}}\right\} \le \eta. \tag{12}$$

On the complementary event, whose probability is at least $1-\eta$, applying the deviation bound to an expected-utility maximizer and to $\widehat{S}$ adds twice the radius to the deterministic certificate. A uniform forecast-to-population discrepancy adds at most $2\epsilon_{\text{forecast}}$. The candidate family is fixed by development data before the predictive draws; neither cross-sequence independence nor a correct forecast is needed for the finite-bank certificate. Population transfer does require the stated discrepancy assumption. □

The exact-arithmetic recurrence gives $\zeta = 0$. Floating-point comparisons with tolerance 10^{-13} require a residual that covers the complete optimization error, not merely one comparison. The archived exhaustive checks examined 10,200 sequences over 320 arrays with zero measured surrogate residual and verified the operational regret inequality. These checks support the implementation and do not replace the proof.

Support ties are the specific reason that optimizing F need not optimize G. The appended-endpoint variant appends a full-release endpoint after optimization; its proof requires an additional nonnegative endpoint term. That variant, its counterexamples and numerical diagnostics remain in the source archive. The primary terminal-included procedure uses the certificate above. For the primary $B = 2048$ forecast and $\eta = 0.05$, the generic concentration term ranges from 9.50 to 12.83 pp across HealthAdvice plans and 12.29 to 13.09 pp across CARE plans; case-weighted means are 12.02 and 12.53 pp. The smaller terms use a different 4,096-world protocol. These are concentration terms, not estimated clinical losses or superiority certificates. The fixed-pair check below avoids enumerating the sequence family but still refers to the forecast distribution.

A.5 AN INDEPENDENT PAIRED BOUND FOR OPERATIONAL PLANNING GAIN

Theorem 2 bounds regret over a finite family. To assess the practical comparison with a specified marginal plan, condition on development data and all simulation draws used to choose exact and marginal plans S_E, S_M. On $B \geq 2$ new iid worlds from their forecast Q, independent of plan construction, evaluate the operational utilities G_b^+ including the common support-tie rule (and the appended endpoint when applied to the appended-endpoint variant). Define

$$Z_b = G_b^+(S_E) - G_b^+(S_M) \in [-1, 1], \quad D_Q = \mathbb{E}_Q Z_b, \quad \overline{Z} = \frac{1}{B}\sum_b Z_b. \tag{13}$$

Let $s_Z^2 = (B-1)^{-1}\sum_{b=1}^{B}(Z_b - \overline{Z})^2$. A two-sided application of Maurer & Pontil (2009, Theorem 4) gives the following corollary, stated as a proposition to match the manuscript's numbering.

Proposition 2 (Paired forecast gain)**.** *Under the preceding independence and boundedness assumptions, for $\eta \in (0,1)$, with probability at least $1-\eta$,*

$$|D_Q - \overline{Z}| \leq r_B(\eta) := \sqrt{\frac{2s_Z^2 \log(4/\eta)}{B}} + \frac{14\log(4/\eta)}{3(B-1)}. \tag{14}$$

Proof. Apply the one-sided empirical Bernstein inequality to $X_b = (Z_b+1)/2 \in [0,1]$ and $1-X_b$, each with failure probability $\eta/2$. Their sample variance is $s_Z^2/4$. Multiplying each bound by two and taking a union bound yields Equation 14. This is a direct corollary of the cited inequality, not a new concentration result. □

The check stream must be independent of every draw used to choose the final plans, including any policy-template selection. Because the two plans are fixed before that stream, the bound does not enumerate the entire sequence family. Direct operational evaluation includes the support/endpoint rules and needs no surrogate correction. It answers a narrower question than optimal-family regret. Identical order lists imply zero utility difference for all worlds and are recorded as exact ties; observing zero sample variance alone is not treated as a proof of equality.

Forecast-to-population transfer. Let D_P be the expected gain under the actual calibration/deployment process. If an explicit assumption supplies $|D_P - D_Q| \leq \epsilon_{\text{pair}}$, then

$$D_P \geq \overline{Z} - r_B(\eta) - \epsilon_{\text{pair}}. \tag{15}$$

Bounds of size $\epsilon_{\text{forecast}}$ on each plan's forecast error imply $\epsilon_{\text{pair}} \leq 2\epsilon_{\text{forecast}}$. Thus a positive population gain needs the simulated lower bound to exceed forecast discrepancy. No distribution-free numerical discrepancy bound is estimated here. Clinical paired bootstrap intervals, calibration pass-frequency diagnostics, and the simulation interval have distinct interpretations. None is substituted for this assumption. Proposition 1 continues to provide the separate calibration risk guarantee.

B DATA, COMPARATORS AND UNCERTAINTY

B.1 DATA AND PREDICTION PROVENANCE

HealthAdvice. The source is the released HealthAdvice dataset with evaluations in the LLMPatientSafety repository. It contains 888 answer records nested within 222 questions, with released model labels Claude, GPT-4o, Gemini, and Llama. We preserve these labels without inventing provider snapshots. The response text and physician annotations are cached. The warning selector uses the maximum of four weighted cached channels and chooses the answer with the lowest resulting score. Its 65 candidate weight vectors are ranked on development data by a predictive useful-release objective with an unsafe-release penalty, then by development discrimination and tie rules. This selection is upstream of every calibration comparator and is repeated within each HealthAdvice view.

The stronger judge identifier recorded in the cache is `gpt-5.6-sol`; its configuration uses medium reasoning and a 4096-token completion cap. The judge receives question and answer text,

not physician evaluation annotations. The analysis uses cached stronger structured and overall-unsafe channels. Original prompts, response identifiers, and score records accompany the experiment. Initial and refit views use distinct recorded partitions and deterministic selector tie conventions. They reuse the same 222-question cohort and are not independent replications.

CARE. The cache uses a pinned dataset snapshot, identified in the source README. The original 439-case generation cache comprised 100 development, 277 calibration, and 62 evaluation cases. The analysis combines the latter 339 cases into five calibration/evaluation allocations, preserving the 100-case training pool. All prefixes of a case share their source role, and only the preselected certification prefix enters the binary calibration trial. Thus 925 prefix records do not supply 925 independent calibration cases.

The four action classes are information needed, self-care/monitoring, nonurgent care, and urgent care. Input features to the numerical action models are cached overall risk, information risk, one-hot original action, and current round index divided by three. Prefix training weights make total weight equal per case. The direct model is histogram gradient boosting (Pedregosa et al., 2011). The information-first model averages logistic models with $C = 0.1$ and $C = 1$, extra trees, and histogram gradient boosting; it first estimates whether clarification is needed and then predicts among the other three actions. Five grouped development folds produce out-of-fold development predictions; models fitted to all 100 development cases predict the remaining cases. Direct scores are $1 - p_{\max}$; information-first scores are $(1 - p_{\max} + p_{\text{runnerup}})/2$.

The CARE cache records `gpt-5.6-sol` for generation and mapping. Generation used Chat Completions, no reasoning, and a 300-token cap; mapping used the official version-3 prompt, low reasoning, and a 1024-token cap. Overall/information assessment used medium reasoning and a 4096-token cap. The published benchmark runner differs in endpoint, generation reasoning, mapper model, and evaluated population. One unavailable information-only score retained the frozen fallback of 1. These differences are held fixed across the nine calibration methods and preclude interpreting our numerical results as direct reproduction of the original model leaderboard.

MedQA. The fixed cache contains 1,000 unique MedQA-USMLE four-option questions assembled in the recorded MedHELM v4 run. The answer models are `gpt-4o-2024-05-13` and `gpt-4o-mini-2024-07-18`. Binary loss compares the extracted answer with the dataset key. Original development-fitted logistic error scores use the cache's agreement, answer-length and choice features; no model or scorer is refitted in this transfer. Each model has ten fitted views (two configurations, five folds), each with 400 development, 400 calibration and 200 evaluation questions. Questions stay grouped, and paired inference shares question resamples across models. The original archive's 278-entry SHA-256 manifest was verified before use. The source README maps the archive and kernel provenance records.

Meaning of development and validation. All numerical fits and threshold plans exclude their own view's calibration and evaluation labels. CARE's internal grouped folds are used to form development predictions. HealthAdvice selector choice is made on its view's development data. Evaluation outcomes are used to measure performance and are not used to tune the methods. The archived fold audit verifies within-view separation. Nested PCO refers to independently seeded predictive simulations; it does not refer to nested cross-validation. Repeated evaluation contributions retain their original case identifiers for the paired analysis.

B.2 Primary and supplementary comparator definitions

Primary LTT, unpenalized Marginal and Exact share the 2,048-world forecast in Section 3; primary Holm tests the same candidate family. The supplementary nine-method comparison uses a 4,096-world forecast and the following recorded implementations; full constants and kernel checks are indexed in the source README.

Fixed-sequence LTT. Order by development exact-binomial p-value, breaking ties by descending development support and ascending threshold; apply record-high completion and Equation 3 with first-failure stopping.

Pareto Testing. This binary selective-risk adaptation retains candidates nondominated in development selective risk and negative useful count (weak componentwise dominance with one strict improvement), orders them by development binomial p-value with index tie-breaking, and applies completion. Its dominance/tie details, completion and selected-support objective specify the task adaptation.

Bonferroni. Test all m candidates at exact-binomial $p < \delta/m$, require support at least ten, and select maximum calibration support with the common tie rule.

aLTT-FWER. Use online-Newton-step bets $1 - \gamma_t(L_t - \alpha)$ on selected cases, with $\gamma_1 = 1$, past-loss updates clipped to $[0, 0.75/(1-\alpha)]$, a 10^{-10} stabilizer, and a shared frozen label order. Inverse running-maximum wealth supplies anytime p-values tested at δ/m with support at least ten: a fixed-budget FWER adaptation, distinct from adaptive acquisition.

RG-PT, FWER-matched. On the development Pareto frontier, assign unselected losses α, form Hoeffding–Bentkus reliability scores, cluster into at most five levels, and link successive levels by positive Lasso with penalty 0.1. Root budgets sum to δ in proportion to recursive leaf mass; all parents must pass before a node is tested, with rejected-node budgets split by fixed development weights—an explicit graph-gated FWER adaptation of the published DAGGER-FDR pipeline (Ramdas et al., 2019).

Marginal simulated-power ordering. Reuse Regular PCO's draws, rank by mean minus 0.25 standard deviations of $u_j B_j$, break ties by development support and threshold, and apply completion; joint-sequence passage is not scored.

Holm. Use support-filtered p-values in ascending order and thresholds $\delta/(m - k + 1)$ with the common selector. Its rejection family contains Bonferroni's; support ties prevent interpreting this as a guaranteed evaluation-yield increase.

PCO variants. Regular PCO and marginal simulation use 4,096 worlds with pseudo-concentration 20. Nested PCO uses 128×32 worlds, $\lambda = 0.25$, three insertion/deletion rounds and an independent 128×32 check; its development-only starting sequences include regular, marginal, development-order, Pareto and rollout/prefix alternatives. Joint fractional PCO uses 4,096 local worlds, simulated support at least ten and exact ascending-sequence optimization; candidates require five development cases and include the highest simulated threshold as terminal. All settings are fixed across the comparison; shared calibration-permutation prefixes, seed formulas and complete implementations are in the source package.

B.3 Case aggregation and simultaneous uncertainty

Let a_{icrq} be the release indicator for case i, configuration c, calibration repeat r, and risk target q, with loss ℓ_{ic}. Within a given benchmark, budget, and certification level, the reported useful yield is

$$\widehat{U} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{CQ} \sum_{c=1}^{C} \sum_{q=1}^{Q} \frac{1}{R_c} \sum_{r=1}^{R_c} a_{icrq}(1 - \ell_{ic}) \right], \tag{16}$$

where $C = 2$, $Q = 6$, and $R_c = 200$ at reduced budgets and 1 at full budget. Each case's assigned evaluation view determines the relevant calibration pool. Coverage substitutes a for $a(1 - \ell)$; error mass substitutes $a\ell$. Frontier points fix q instead of averaging it. Percentages are obtained by multiplying by 100. Equal weighting is over cases, not over folds of slightly different size.

The 5,000 recorded paired bootstrap replicates draw N case identifiers with replacement, using the fixed benchmark-specific seeds recorded in the source README. The same counts multiply both methods' casewise contributions. Percentile intervals use the 2.5th and 97.5th percentiles. All repeated appearances of an identifier remain together. The bootstrap does not create additional independent cases from the 200 subsamples, targets, or configurations. In particular, a narrow interval must not be interpreted as observing thousands of independent patients.

Primary comparison family. The primary study uses 756 target-averaged contrasts: 18 contrasts in each of 42 cells. There are 32 current-action cells (two acquisitions, two certification levels and $3 + 5$ benchmark-specific budgets) and ten episode cells. The 18 contrasts are three policy effects,

two Exact–Marginal effects, one policy-by-planner interaction, six two-chain effects and six Holm effects. If s_j is the bootstrap standard deviation of contrast j, the simultaneous sensitivity band is $\widehat{\Delta}_j \pm c_{0.95} s_j$, where

$$c_{0.95} = \text{quantile}_{0.95}\left[\max_{j:s_j>0}\left|\frac{\Delta_j^* - \widehat{\Delta}_j}{s_j}\right|\right] = 3.971. \tag{17}$$

The same family is used for positive and negative differences. Tables reporting Exact minus Holm reverse the archived Holm-minus-Exact estimate and both interval endpoints; they use the same comparisons and resampling. The supplementary uniform nine-method analysis has a separate 288-contrast family (critical value 3.3496); the targeted binary forecast comparison has 80 contrasts, and the episode-score comparison has 620. These families and their complete outputs remain separate.

Clinical Exact–LTT intervals add 84 Exact–LTT contrasts to this family (840 contrasts, critical value 3.993).

These are approximate conditional performance intervals, distinct from the calibration FWER guarantee. They keep model outputs, numerical fits, planned sequences and recorded calibration draws fixed. Constant case contrasts have zero-width sensitivity bands under this resampling procedure, not necessarily zero uncertainty for a new population. All patients/questions remain the resampling units.

Comparator families. The clinical native comparison has a 660-contrast family; the larger certification audit has 2,412 contrasts and critical value 3.727650. We retain the latter family for all reported clinical adapted-method intervals, including comparisons outside the manuscript's displayed subset. The family is not reduced when excluding PCO-certificate hybrids from the manuscript. Both analyses preserve the original case-cluster resampling and conditional estimand.

MedQA uses 5,000 question-cluster draws, with shared sampled question identifiers for both answer models. Its 456 primary contrasts are $48 \times 9 + 24$: 48 cells (two models, two acquisition rules, six caps, two confidence levels) compare Exact with boosted SCoRE, LEC, SCRC correctness sets, their three Holm adaptations, matched Marginal, matched LTT and Holm grid; 24 uniform cells additionally compare cached Fractional Joint PCO. The maximum-standardized critical value is 3.461142. General/binary SCoRE and individual-target intervals are secondary pointwise results. All full grids and family memberships are supplied. No independence is attributed to repeated held-out batches.

C PRIMARY CONTROLLER AND IMPLEMENTATION

This section records the implementation constants accompanying the main-text controller. It restates the frozen procedure.

Shared conditional forecast. The forecast receives development scores/losses, the calibration cap, screening-pool size and (α, δ). The shared envelope is the development-fixed marginal-selected envelope on the original score. It is not reselected for Exact. Candidate thresholds are distinct values from 21 score quantiles within the envelope, plus endpoints, retaining at least five development cases except for the mandatory terminal cutoff.

The conditional forecast augments development observations inside E with 55 local pseudo-observations: five neighboring loss quantiles at each of 11 score quantiles, using up to 60 neighbors. Real observations have unit Dirichlet concentration; pseudo-observations share total mass 20. Each world draws conditional weights $w^{(b)}$, envelope mass $q_E^{(b)}$ from its development Jeffreys beta posterior, an eligible count in the finite screening pool, and up to n pseudo-calibration cases. These draws jointly determine selected counts, errors and passage at every cutoff, and population useful mass

$$u_j^{(b)} = q_E^{(b)} \sum_i w_i^{(b)} \mathbf{1}\{s_i \le t_j\}(1 - y_i). \tag{18}$$

The same draw affects all nested thresholds. Fractional local pseudo-losses and their binomial-CDF transform are planning heuristics; calibration itself uses actual binary losses and integer counts. The

primary study uses 2,048 training worlds and 2,048 independent check worlds per template. Its terminal cutoff is included inside the optimizer. Global-forecast and appended-endpoint settings form separate supplementary comparisons; absolute changes across these experiments are not one-factor ablations.

Policy templates. HealthAdvice's expanded bank includes the original selector, the development-error-minimizing selector (ties by unsafe AURC and index), the development-AURC-minimizing selector and equal warning weights. CARE's bank includes original argmax and fixed probability multipliers for information needed (1.35 or 0.75), self-care (1.50), and urgent care (1.25). For each template, training worlds plan the sequence. Independent check worlds evaluate its actual support-selected utility, and one complete template is chosen before calibration. Every planner receives this same bank and check budget. The reproducible definitions and selection bounds are in Appendices C.2 and D.

C.1 End-to-end procedure

1. Fit the policy, score, acquisition envelope and threshold candidates using development data only.
2. Generate one shared bank of 2,048 predictive worlds per template; compute all candidate pass indicators and useful masses jointly.
3. Construct the LTT, Marginal or Exact sequence. Exact uses Equation 7; all three share the same eligible candidates and terminal cutoff.
4. For the expanded policy arm, evaluate each complete plan on 2,048 independent check worlds and choose one template. The Current arm keeps its fixed template.
5. Acquire calibration labels through the frozen envelope, apply the specified tests, stop the single chain at its first failure, and select among rejections by maximum support with the smaller-cutoff tie rule.
6. Evaluate the frozen controller on the grouped evaluation cases. Pair methods on identical calibration permutations and evaluation identifiers.

The simulation randomness and all choices in steps 1–4 precede calibration. Evaluation labels are used only in step 6. Holm instead tests the same family with its order-invariant step-down procedure. Candidate thresholds, policies and acquired labels remain common within the Current-policy comparison.

C.2 Focused labels and complete clinical policies

This design separates three sources of useful yield: the clinical policy supplying an answer/action, the calibration cases acquired, and the sequence of certification tests. Within each frozen experiment, numerical policy choices use development data. The uniform-sampling experiments remain separate controls.

C.2.1 Focused calibration and the sparse-support mechanism

Fix an envelope $E \subseteq \mathcal{X}$ using development data and restrict every deployable release set to E. Scan a fixed iid calibration pool until n eligible cases are found or the pool is exhausted, without looking at labels. Conditional on the number accepted, these cases are iid from $Q = P(\cdot \mid E)$. For every $A \subseteq E$,

$$Q(L = 1 \mid A) = P(L = 1 \mid A). \tag{19}$$

Consequently the original selected-risk test remains valid. Forecasting and reporting must use population useful mass $P(E)\mathbb{E}_Q[\mathbf{1}_A(1-L)]$, rather than replacing population yield with yield conditional on the acquisition region. Actual labels and screened cases are different costs.

The focused-acquisition experiment crosses five controllers with uniform, individually selected, shared marginal, and shared joint envelopes. Candidate envelopes correspond to 25%, 50%, 75%,

and 90% development coverage, plus the full region. A separate predictive check selects each controller's own envelope. The complete grid has 19,200 selected arms and 1,133,280 case contributions. All five uniform controls reproduce their recorded results. All 192 primary target-averaged contrasts are reported.

On CARE with 20 labels and 90% certification, Joint fractional useful yield rises from 2.07% under uniform acquisition to 7.70% under its selected envelope, a gain of 5.63 pp $[4.15, 7.14]$. The simultaneous band is $[2.91, 8.35]$. Local/Marginal rises from 2.08% to 7.66%. Against the strongest simple acquisition arm, the Joint fractional difference is 0.04 pp $[-0.02, 0.07]$. At the same budget on HealthAdvice, focused Local/Marginal yields 17.78% and focused Joint fractional yields 17.20%; the latter's own acquisition gain is 0.27 pp $[-0.27, 0.78]$. These results identify an acquisition benefit shared with simple predictive planning.

The CARE fraction of 20-label calibrations with no individually certifiable cutoff falls from 91.52% to 47.36% under the joint-selected envelope. For a 10% useful-yield goal at $\alpha = 0.25$, $\delta = 0.1$, the first tested label cap is 160 under uniform acquisition and 40 under focused acquisition for LTT, the marginal controllers, and Joint fractional PCO. Focused Local/Marginal and Joint fractional both reach 18.35% $[14.61, 22.16]$ with 40 labels and 166.22 screened cases. This is an observed grid-based 75% reduction in labels with about 4% more screening, not a guaranteed saving or a gain uniquely attributable to exact planning. At larger caps, finite eligible pools may prevent filling the nominal cap; actual counts are retained in every table.

C.2.2 Selecting one complete policy before testing

Let a development-defined bank contain K complete policies π_k, each supplying an action f_k and score s_k. Candidate sets within a template are

$$A_{kj}(X) = \mathbf{1}\{X \in E,\ s_k(X) \le t_{kj}\}. \tag{20}$$

For each intended budget and risk setting, the procedure trains one sequence per template and chooses one template using independent check simulations. Only that chosen template is tested on calibration cases. This avoids applying a nested-threshold optimizer to nonnested action-switching candidates, and avoids silently charging one answer label for several tested answers.

Proposition 3 (Complete-policy validity)**.** *Conditional on development data and its simulation randomness, suppose the chosen policy, envelope, thresholds, and testing procedure are fixed independently of calibration. Fixed-sequence testing at level δ controls the probability of certifying any threshold with selected risk exceeding α by δ, regardless of how the development stage chose among the K templates.*

Proof. After conditioning, the selected template is fixed. Its candidate p-values are valid under their respective selected-risk nulls, including under focused acquisition by the preceding conditional-risk identity. Rejecting a true null in the chosen sequence entails rejecting its first true null, with probability at most δ. Integrating over development randomness gives the result. No calibration multiplicity penalty for K is needed because the other templates are not selected using calibration outcomes. □

Proposition 4 (Finite-bank predictive selection)**.** *Condition on K training-generated plans. Let $G_k \in [0, 1]$ be plan k's expected operational utility under its forecast and let $\widehat{G}_k$ average B_c independent check worlds. For $\widehat{k} \in \arg\max_k \widehat{G}_k$, with probability at least $1 - \eta$,*

$$\max_k G_k - G_{\widehat{k}} \le 2\sqrt{\frac{\log(2K/\eta)}{2B_c}}. \tag{21}$$

If every forecast operational expectation differs from its actual population/calibration expectation by at most ϵ, add 2ϵ to the bound.

Proof. Hoeffding's inequality and a union bound give a simultaneous bound $|G_k - \widehat{G}_k| \le b$, where b is the square-root term. If k^* maximizes G_k, then $G_{k^*} - G_{\widehat{k}} = (G_{k^*} - \widehat{G}_{k^*}) + (\widehat{G}_{k^*} - \widehat{G}_{\widehat{k}}) + (\widehat{G}_{\widehat{k}} - G_{\widehat{k}}) \le b + 0 + b$, since $\widehat{k}$ maximizes $\widehat{G}_k$. The population version adds one discrepancy for each compared plan. Cross-template independence is unnecessary for the union bound; check worlds must be independent of the corresponding training-generated plan. □

This is a standard concentration argument applied to the implemented outer choice. It is not a new distribution-free utility guarantee. With $B_c = 2048$, $K = 5$, and $\eta = 0.05$, the generic bound is about 7.19 pp and can be substantially larger than observed effects. The inner dynamic program remains exact for its supplied last-passing simulation objective; check selection uses actual support ties. In this design the terminal candidate is optimized within the family, so there is no separately appended endpoint term.

C.2.3 A VALID FALLBACK COMPARISON

The primary chain and a complete ascending fallback chain receive $\delta/2$ each in advance. Each stops at its first failure. Their rejected candidates are combined, and the same maximum-support selector is applied. Each chain has FWER at most $\delta/2$, so their union has FWER at most δ, without independence and even if candidates overlap. A failed test never frees its allocated significance. Holm step-down over the identical family is an additional comparison. These are established multiple-testing constructions available in the LTT framework (Angelopoulos et al., 2025; Holm, 1979); the empirical question is whether recovered intermediate cutoffs compensate for reduced per-chain power.

The power cost follows directly from the discrete binomial boundary. At $\alpha = 0.25$ and zero errors, ten selected cases give $p = 0.75^{10} = 0.0563$: a cutoff passes at $\delta = 0.10$ but fails at $\delta/2 = 0.05$. Eleven cases give $0.75^{11} = 0.0422$ and pass both. More generally, each chain changes the zero-error support requirement to $\max\{10, \lfloor \log(\delta/2)/\log(1-\alpha) \rfloor + 1\}$. This explains a concrete source of power loss; the saved comparison measures its net effect alongside any recovered cutoffs.

D COMPLETE CONTROLLED RESULTS AND RESOURCE ACCOUNTING

The primary experiment contains 2,520 task/view/budget/risk/acquisition conditions, 15,120 selected policy/controller plans and 2,772,144 case contributions. The tables report every target-averaged Current-policy ordering comparison under both acquisition rules, including null and negative settings. The complete matched Holm table appears in Appendix J; the adapted-baseline grid appears in Appendix G.3. Exact–Marginal holds the testing procedure fixed and isolates the planner. Exact–Holm also changes certification from a fixed sequence to order-invariant step-down testing; it measures the combined consequence of that choice, not the joint-planning increment alone. Holm outputs are identical across Current-policy planner labels because its rejection rule ignores their planned orders. The 756-contrast simultaneous family covers both comparisons.

Table 3: Current Exact minus Marginal under shared focused acquisition. Useful-yield gains and interval endpoints are percentage points; all six targets and both configurations are averaged. Full means 80 available labels for HealthAdvice and 271–272 for CARE.

Benchmark	Cap	Cert.	Gain (pp)	Paired 95%	Simultaneous
HealthAdvice	20	90%	+0.71	$[+0.60, +0.82]$	$[+0.48, +0.94]$
HealthAdvice	20	95%	+0.46	$[+0.35, +0.59]$	$[+0.23, +0.70]$
HealthAdvice	40	90%	+0.55	$[+0.28, +0.82]$	$[+0.001, +1.10]$
HealthAdvice	40	95%	+0.34	$[+0.21, +0.47]$	$[+0.08, +0.60]$
HealthAdvice	Full	90%	+2.74	$[+2.03, +3.45]$	$[+1.30, +4.18]$
HealthAdvice	Full	95%	+0.00	$[-0.11, +0.11]$	$[-0.21, +0.21]$
CARE	20	90%	+0.57	$[+0.46, +0.69]$	$[+0.34, +0.81]$
CARE	20	95%	+0.48	$[+0.36, +0.59]$	$[+0.25, +0.71]$
CARE	40	90%	+0.36	$[+0.19, +0.55]$	$[-0.02, +0.74]$
CARE	40	95%	+0.56	$[+0.42, +0.72]$	$[+0.25, +0.88]$
CARE	80	90%	+0.42	$[+0.15, +0.69]$	$[-0.14, +0.98]$
CARE	80	95%	+0.50	$[+0.15, +0.86]$	$[-0.24, +1.23]$
CARE	160	90%	-0.06	$[-0.19, +0.06]$	$[-0.31, +0.20]$
CARE	160	95%	-0.69	$[-1.01, -0.39]$	$[-1.30, -0.07]$
CARE	Full	90%	-0.22	$[-0.37, -0.10]$	$[-0.51, +0.07]$
CARE	Full	95%	-0.32	$[-0.57, -0.10]$	$[-0.81, +0.17]$

Among the ten small-cap gains highlighted in Section 5.2, the three simultaneous intervals containing zero are CARE at 40/90%, 80/90% and 80/95%. A single calibration error changes an error

fraction by 2.5–5 pp at support 20–40; this discrete step is a different quantity from a useful-yield gain averaged across questions, targets and calibration draws.

Table 4: Current Exact minus Marginal under uniform acquisition. Useful-yield gains and interval endpoints are percentage points; all six targets and both configurations are averaged. Full means 80 available labels for HealthAdvice and 271–272 for CARE.

Benchmark	Cap	Cert.	Gain (pp)	Paired 95%	Simultaneous
HealthAdvice	20	90%	+0.74	$[+0.39, +1.12]$	$[-0.01, +1.48]$
HealthAdvice	20	95%	-0.01	$[-0.12, +0.12]$	$[-0.26, +0.24]$
HealthAdvice	40	90%	+0.08	$[-0.40, +0.58]$	$[-0.93, +1.08]$
HealthAdvice	40	95%	+0.08	$[-0.23, +0.40]$	$[-0.57, +0.72]$
HealthAdvice	Full	90%	+0.41	$[-0.15, +0.98]$	$[-0.73, +1.56]$
HealthAdvice	Full	95%	+0.04	$[-0.11, +0.23]$	$[-0.30, +0.37]$
CARE	20	90%	-0.02	$[-0.03, -0.01]$	$[-0.03, -0.01]$
CARE	20	95%	-0.02	$[-0.03, -0.01]$	$[-0.03, -0.01]$
CARE	40	90%	-0.15	$[-0.20, -0.11]$	$[-0.24, -0.07]$
CARE	40	95%	-0.06	$[-0.09, -0.02]$	$[-0.12, +0.01]$
CARE	80	90%	+0.15	$[+0.05, +0.24]$	$[-0.04, +0.33]$
CARE	80	95%	+0.18	$[+0.10, +0.27]$	$[+0.003, +0.36]$
CARE	160	90%	-0.49	$[-0.70, -0.30]$	$[-0.91, -0.07]$
CARE	160	95%	-0.19	$[-0.30, -0.08]$	$[-0.41, +0.04]$
CARE	Full	90%	-0.10	$[-0.27, +0.05]$	$[-0.41, +0.21]$
CARE	Full	95%	-0.17	$[-0.39, +0.02]$	$[-0.60, +0.25]$

D.1 Policy expansion and the broad simple reference

For Exact at 90% certification, policy expansion adds 1.05 pp on HealthAdvice, with paired interval $[0.66, 1.40]$ and simultaneous band $[0.30, 1.80]$. The policy-family by planner interaction is 0.40 pp, with simultaneous band $[0.07, 0.73]$. CARE's corresponding policy-expansion gain is only 0.0004 pp. Consequently the CARE result at 90% is an ordering gain under the specified conditional experiment, rather than an improvement attributable to changing the selected action policy.

The broad simple reference includes all Current/Expanded LTT and Marginal procedures under Single, Two-chain, and Holm testing, all twelve simple focused-acquisition arms, and the six non-PCO uniform comparators. Every method shares the same maximum label and screening caps; different envelopes can consume different actual labels at larger caps. The maximum reference mean is recomputed in each bootstrap and is not a deployable evaluation-selected rule. The primary Exact–Marginal contrast compares the two specified methods directly. The paired maximum intervals are descriptive and can be nonregular at ties. The reference is interpreted as a descriptive envelope across the recorded methods. Table 5 reports the four low-label comparisons. At 90%, the expanded pipeline adds about 11.0 and 5.7 correct releases per 1,000 decisions over this reference on HealthAdvice and CARE, respectively.

For a fixed Expanded Marginal comparator M, the observed reference gap decomposes as

$$\widehat{U}_E - \max_j \widehat{U}_j = (\widehat{U}_E - \widehat{U}_M) - (\max_j \widehat{U}_j - \widehat{U}_M). \tag{22}$$

At full budget and 90% certification, Expanded Exact exceeds the matched Expanded Marginal pipeline by 1.88 pp on HealthAdvice and 0.02 pp on CARE; the broad reference exceeds that Marginal pipeline by 2.82 and 0.71 pp, yielding the reported -0.94 and -0.69 pp gaps. At 95%, the corresponding pairs are $(2.21, 2.55)$ and $(-0.02, 0.98)$ pp. This exact accounting separates the observed reference premium from the matched pipeline difference; it does not estimate the pure sampling effect of maximization. Full accounting for all 16 settings is in the source. The primary fixed-policy ordering contrast remains separate.

Table 6 retains the larger-budget comparisons. Positive low-label increments do not establish superiority throughout the budget grid. In particular, the expanded focused Exact pipeline remains below the strongest simple reference at the maximum cap on both benchmarks. Focused maximum-cap conditions may label fewer cases than the pool size and are distinct from the completion-off full-budget comparison in the source package.

Table 5: Expanded Exact and the strongest tested simple reference at 20 labels. Means are useful-yield percentages; gain intervals are paired percentage points.

Benchmark	Cert.	Expanded Exact (%)	Best simple (%)	Gain (pp) [paired 95% CI]
HealthAdvice	90%	19.41	18.31	$+1.10$ $[+0.91, +1.30]$
HealthAdvice	95%	13.50	13.30	$+0.20$ $[+0.11, +0.32]$
CARE	90%	8.34	7.77	$+0.57$ $[+0.45, +0.69]$
CARE	95%	5.70	5.23	$+0.47$ $[+0.33, +0.60]$

Table 6: Expanded focused Exact minus the broad simple reference over every calibration cap. Entries are gain [paired 95% interval] in percentage points. Full denotes the available-pool label cap, not necessarily that many acquired labels.

Benchmark	Label cap	90% certification	95% certification
HealthAdvice	20	$+1.10$ $[+0.91, +1.30]$	$+0.20$ $[+0.11, +0.32]$
HealthAdvice	40	$+0.46$ $[+0.24, +0.59]$	-0.16 $[-0.55, +0.05]$
HealthAdvice	full	-0.94 $[-1.54, -0.38]$	-0.34 $[-0.60, -0.08]$
CARE	20	$+0.57$ $[+0.45, +0.69]$	$+0.47$ $[+0.33, +0.60]$
CARE	40	-1.08 $[-1.39, -0.79]$	-0.86 $[-1.20, -0.55]$
CARE	80	-1.38 $[-1.87, -0.91]$	-1.14 $[-1.53, -0.78]$
CARE	160	-2.80 $[-3.30, -2.31]$	-1.29 $[-1.80, -0.81]$
CARE	full	-0.69 $[-1.08, -0.42]$	-1.01 $[-1.38, -0.69]$

The envelope is fixed before calibration for each intended budget and risk setting, but can differ between those settings. It is not one common envelope held across the entire budget grid; a 20-label eligible fraction therefore does not determine the full-cap label count. Consequently, the roughly 65 labels acquired on CARE at cap 80 must not be extrapolated to caps 160 or full: their means are roughly 104 and 206–211. Finite-pool exhaustion and the envelope's release ceiling still distinguish this comparison from unrestricted acquisition. Equal nominal caps do not imply equal realized labels across the broad-reference arms.

D.2 Observed errors and evaluation denominators

The next tables show the primary 20-label comparison at each target. Risk is pooled error mass divided by pooled release mass; no risk is assigned to an empty release set. Across two configurations and 200 recorded calibrations there are 88,800 repeated evaluation opportunities from 222 HealthAdvice questions and 135,600 from 339 CARE patients. Coverage is releases divided by the corresponding opportunity count. Count columns use comma thousands separators; decimal points denote fractional rates. These repeated events are not additional independent patients; the bootstrap groups them by their original identifier.

Finite evaluation ratios need not fall below every nominal target. They do not estimate how frequently the calibration procedure selects a rule whose true population risk exceeds the target. The theorem concerns that latter event under its stated assumptions.

Table 7: Actual labels acquired by the focused Exact pipeline, averaged over the same targets and cases as the performance summaries. Current and Expanded share these counts; all matched planners do too. Full caps are 80 for HealthAdvice and 271–272 for CARE.

Benchmark	Cap	Labels (90%)	Screened (90%)	Labels (95%)	Screened (95%)
HealthAdvice	20	20.00	22.55	20.00	22.58
HealthAdvice	40	39.97	46.24	39.97	47.87
HealthAdvice	full	77.93	80.00	78.36	80.00
CARE	20	20.00	73.15	20.00	73.15
CARE	40	40.00	166.22	40.00	166.22
CARE	80	65.20	271.20	65.77	266.34
CARE	160	103.99	269.42	103.66	267.73
CARE	full	205.93	271.20	210.68	271.20

Table 8: Observed primary comparison at 90% certification. Targets, coverage C and pooled risks are percentages. M/E denote Marginal/Exact; rel./err. are repeated release/error counts.

Benchmark	Target	M C	M rel.	M err.	M risk	E C	E rel.	E err.	E risk
HealthAdvice	10	0.00	0	0	–	0.00	0	0	–
HealthAdvice	15	4.09	3,636	629	17.30	4.06	3,602	620	17.21
HealthAdvice	20	10.81	9,601	1,737	18.09	12.20	10,835	1,978	18.26
HealthAdvice	25	26.16	23,231	4,879	21.00	28.02	24,884	5,215	20.96
HealthAdvice	30	37.99	33,739	6,546	19.40	39.65	35,205	6,809	19.34
HealthAdvice	35	51.84	46,036	8,392	18.23	52.19	46,343	8,413	18.15
CARE	10	0.00	0	0	–	0.00	0	0	–
CARE	15	2.56	3,478	458	13.17	2.55	3,458	435	12.58
CARE	20	5.95	8,062	1,063	13.19	5.99	8,127	1,114	13.71
CARE	25	9.35	12,680	1,866	14.72	13.05	17,691	2,440	13.79
CARE	30	16.20	21,963	2,749	12.52	16.25	22,038	2,749	12.47
CARE	35	19.60	26,576	3,405	12.81	19.63	26,622	3,327	12.50

Table 9: Observed primary comparison at 95% certification. Targets, coverage C and pooled risks are percentages. M/E denote Marginal/Exact; rel./err. are repeated release/error counts.

Benchmark	Target	M C	M rel.	M err.	M risk	E C	E rel.	E err.	E risk
HealthAdvice	10	0.00	0	0	–	0.00	0	0	–
HealthAdvice	15	1.83	1,628	304	18.67	1.84	1,637	313	19.12
HealthAdvice	20	4.22	3,747	690	18.41	4.72	4,189	755	18.02
HealthAdvice	25	10.81	9,601	1,737	18.09	12.20	10,835	1,978	18.26
HealthAdvice	30	27.66	24,565	4,810	19.58	27.70	24,601	4,863	19.77
HealthAdvice	35	48.93	43,451	8,828	20.32	50.35	44,711	8,970	20.06
CARE	10	0.00	0	0	–	0.00	0	0	–
CARE	15	0.61	827	136	16.44	1.00	1,360	222	16.32
CARE	20	2.67	3,618	479	13.24	2.65	3,597	457	12.71
CARE	25	5.95	8,062	1,063	13.19	5.99	8,127	1,114	13.71
CARE	30	10.14	13,748	1,997	14.53	11.55	15,659	2,141	13.67
CARE	35	16.09	21,821	3,058	14.01	17.42	23,615	3,202	13.56

E STOPPING DIAGNOSTICS AND COMPUTATION

For each fixed clinical policy and realized calibration sample, let $\mathcal{C} = \{j : N_j \geq 10,\ p_j < \delta\}$. A single-chain controller can reject only candidates in $\mathcal{C}$. Its probability of producing no rejection decomposes exactly as

$$\Pr(\text{no rejection}) = \Pr(\mathcal{C} = \varnothing) + \Pr(\text{no rejection},\ \mathcal{C} \neq \varnothing). \tag{23}$$

The first term is common to the matched planners and cannot be repaired by changing their order. The second records a passing candidate missed by first-failure stopping. It can change when only the planned sequence changes. "No rejection" concerns the calibration decision, not whether a particular evaluation patient falls inside the resulting release set.

The table uses the saved calibration diagnostics, averaged over calibration draws and then over targets and fitted views with their evaluation-case weights, exactly as in the archived method summaries. Percentages describe dependent calibration realizations; no patient-level confidence interval is claimed. The diagnostic partition was checked in all 3,840 Current-policy Exact/Marginal rows across both acquisition rules. The complete uniform and focused diagnostics accompany the source.

Table 10: Calibration stopping diagnostic under shared focused acquisition. None pass is common to both methods. Missed means no rejection despite at least one individually passing candidate. All entries after certification are percentages or percentage-point differences.

Benchmark	Cap	Cert.	None pass	Missed: M	Missed: E	M − E
HealthAdvice	20	90%	64.36	9.10	7.91	+1.19
HealthAdvice	20	95%	74.94	7.00	6.25	+0.75
HealthAdvice	40	90%	47.90	10.77	9.39	+1.38
HealthAdvice	40	95%	57.82	10.23	9.51	+0.71
HealthAdvice	Full	90%	38.44	8.37	3.30	+5.07
HealthAdvice	Full	95%	48.39	8.33	8.33	+0.00
CARE	20	90%	43.74	16.45	14.17	+2.28
CARE	20	95%	56.55	17.26	15.06	+2.20
CARE	40	90%	23.90	15.33	12.21	+3.12
CARE	40	95%	32.96	17.33	13.57	+3.76
CARE	80	90%	16.69	11.70	10.03	+1.67
CARE	80	95%	23.35	10.00	8.33	+1.67
CARE	160	90%	16.69	10.03	10.03	+0.00
CARE	160	95%	23.35	8.33	11.68	-3.34
CARE	Full	90%	16.69	5.01	6.69	-1.67
CARE	Full	95%	23.35	6.64	8.31	-1.67

The missed-candidate frequency is lower for Exact in all ten highlighted small-cap conditions. At full CARE budgets it is higher by about 1.67 pp at both confidence levels, matching the direction of the reported useful-yield crossover. At full HealthAdvice/90%, it falls from 8.37% to 3.30%; at 95% both are 8.33%. The diagnostic identifies one observable mechanism; it is not an additive decomposition of yield, since two nonempty rejected sets can have different population useful mass.

Independent forecast checks at larger budgets. The saved Current-policy plans also record support-selected utility on their 2,048 training worlds and independent 2,048 check worlds. At full CARE budgets, Exact-minus-Marginal forecast gains are 0.38/0.37 pp on training/check worlds at 90% certification, and 0.30/0.27 pp at 95%. The forecast preference persists out of the planning bank, whereas the observed clinical differences are -0.22 and -0.32 pp. Thus these records do not establish finite-world overfitting as the cause of the crossover. The matched stopping counts identify an observed contribution. The tie audit in Appendix E.1 measures the surrogate/selector discrepancy separately; the forecast-to-clinical difference still includes forecast mismatch and realized calibration variation. These are descriptive forecast checks, not clinical intervals. The source provides all 960 plans and the complete 16-setting aggregation.

E.1 Measured surrogate-to-selector tie loss

We reconstruct the recorded 2,048 training and 2,048 independent-check worlds for each selected plan and evaluate its frozen single-chain sequence twice: last-passing utility F_b and actual maximum-support utility G_b. Every reconstructed operational mean exactly matches its archived value. The check bank is independent of inner sequence construction and also serves the Expanded-template selection. The denominator is 15,120 selected plan records across 2,520 task/view/budget/risk/acquisition conditions, including Current/Expanded, all three planners and the episode study. Repeated or identical plans retain their recorded multiplicity; these counts are not independent experiments.

Training tie loss exceeds numerical tolerance 10^{-12} in 8,540/15,120 records (56.48%). Across all records its mean, median, 95th percentile and maximum are 0.03, 0.001, 0.17 and 1.17 pp. For the 5,040 Exact records the corresponding values are 0.03, 0.001, 0.17 and 1.17 pp; 2,979 have positive loss. The independent-check mean is 0.03 pp, with 2,915 positive Exact records. Thus a nonzero

plan-average discrepancy is common but usually small. The source includes every plan, world-level positive fractions, and median/90th/95th/99th percentiles by task, acquisition, policy, planner, cap and confidence.

Table 11: Surrogate-to-selector tie loss on recorded predictive banks. Positive denotes a plan-average loss greater than 10^{-12}; mean, p95 and maximum are percentage points of useful yield. Current refers to the current-action task and includes both policy families.

Plan group	Bank	Plans	Positive (%)	Mean	p95	Max
All plans	Check	15,120	56.38	0.03	0.17	1.19
Exact, all tasks	Check	5,040	57.84	0.03	0.16	1.19
HealthAdvice current Exact	Check	1,440	74.79	0.08	0.39	1.19
CARE current Exact	Check	2,400	71.21	0.01	0.02	0.12
CARE episode Exact	Check	1,200	10.75	0.0001	0.001	0.01
All plans	Train	15,120	56.48	0.03	0.17	1.17
Exact, all tasks	Train	5,040	59.11	0.03	0.17	1.17
HealthAdvice current Exact	Train	1,440	74.44	0.08	0.40	1.17
CARE current Exact	Train	2,400	72.83	0.01	0.02	0.12
CARE episode Exact	Train	1,200	13.25	0.0001	0.001	0.01

E.2 Recorded planning cost

The 960 shared-focused Current-policy timing conditions cover both benchmarks, all budgets, targets and confidence levels. Timers include full- and half-level sequence construction and training/check utility evaluation, excluding the shared forecasts. The median paired added time is 26.8 ms; the source package reports the complete timing table. Worst-case exact optimization is exponential.

E.3 A bounded-cost planner for denser grids

An optional greedy approximation uses the same predictive worlds and state (i, A) as Exact. Start at $i = 0$, with $u_0 = 0$ and all worlds active. At each step choose the largest positive immediate joint increment

$$j^* \in \arg\max_{j>i} \frac{1}{B} \sum_{b \in A : B_j^{(b)} = 1} (u_j^{(b)} - u_i^{(b)}), \tag{24}$$

breaking ties toward larger j. Append j^*, replace A by its passing worlds and set $i = j^*$. Stop if all increments are zero or no world survives, then complete to the terminal candidate if needed. There are at most m stages with at most m candidate evaluations of B worlds each: $O(Bm^2)$ time and $O(Bm)$ storage including the input arrays. All choices precede actual calibration, so Proposition 1 applies unchanged. This approximation does not inherit the exact DP objective guarantee; it is a distinct optional planner and is not substituted into the clinical results.

Table 12 measures this kernel on one cached development view per benchmark, at cap 40, $\alpha = 0.25$, $\delta = 0.10$, uniform acquisition and $B = 2048$. Each requested quantile grid is deduplicated and pruned by the original five-development-case rule, with the terminal retained. We report the resulting m, which is smaller than the requested density when scores repeat. Times are medians of nine post-warm-up calls on a single-threaded numerical runtime and exclude shared forecast generation. Exact is timed only on grids with at most 21 candidates; dashes indicate that it was not run. This is a computational demonstration, with clinical performance of the approximation left unmeasured.

Table 12: Sequence-construction time as the requested grid density increases. Actual m counts distinct eligible thresholds including the terminal. Greedy complexity is polynomial; no timing extrapolation for Exact is made.

Benchmark	Requested quantiles	Actual m	Greedy (ms)	Exact (ms)
HealthAdvice	10	10	0.19	7.13
HealthAdvice	20	18	0.37	71.76
HealthAdvice	50	36	0.74	–
HealthAdvice	100	60	1.24	–
CARE	10	10	0.45	3.78
CARE	20	20	0.69	7.96
CARE	50	49	1.62	–
CARE	100	90	3.51	–

F SUPPORTING CONTROLS AND SUPPLEMENTARY COMPARISONS

The complete Current-policy uniform-acquisition comparisons are in Appendix D. Supplementary uniform-acquisition experiments use different forecast and endpoint conventions and are retained as supporting controls. Their full target-specific tables, interval families and code are in the source package.

The controlled two-forecast by three-planner study separates forecast construction from ordering. Including matched unpenalized marginal/greedy references leaves no observed discrete label savings in 144 fixed-goal comparisons. Appendix J retains the nine-method comparison; the source README maps the single-ordering ablation, complete six-way comparison and full 288-contrast uncertainty analysis. Their outcomes are not substituted for the primary matched conditional experiment. For a predeclared goal g and tested budget grid $\mathcal{N}$, define $n_M(g) = \min\{n \in \mathcal{N} : \widehat{U}_M(n) \geq g\}$, with $+\infty$ if the goal is never reached; coverage goals use $\widehat{C}_M$ instead. A discrete label saving compares these first attained caps, not an interpolated sample-complexity guarantee. The 144 matched-reference comparisons contain zero savings, 64 ties, four slower attainments, one reference-only attainment and 75 goals unreached by both methods.

F.1 FRACTIONAL VERSUS BINARY PREDICTIVE LOSSES

A targeted forecast ablation replaces each fractional local pseudo-error contribution $C_a r_a$ with $K_a \mid C_a \sim \mathrm{Binomial}(C_a, r_a)$, sharing the sampled atom count across all containing thresholds. It holds policies, candidates, means, acquisition, calibration and evaluation fixed. Expected pseudo-error totals are preserved, but their variance and predicted passage probabilities can change. Actual calibration uses binary errors in both arms. Both forecasts receive matched Marginal and Exact planners.

The complete sixteen-cell forecast comparison is in the source package. Exact has two positive, nine negative and five unchanged point estimates; neither positive effect remains positive under the 80-contrast simultaneous sensitivity. All full-budget case contributions are identical across forecasts. Thus the particular fractional-error approximation does not provide a general explanation for ordering performance. The source README maps the full moment calculations, forecast checks and matched plans.

F.2 TWO-CHAIN FALLBACK

The primary and complete fallback chains each receive $\delta/2$ in advance, preserving the total error budget by a union bound. Synthetic traces show both recovery of a skipped cutoff and a loss of power from splitting the budget. Across all 42 Expanded Exact task/budget/confidence/acquisition cells, no two-chain-minus-single paired interval is positive and 38 are negative. Holm remains a separate order-invariant comparator. The source README maps the complete contrast grid.

G COMPARATOR METHODS AND MEDQA TRANSFER

G.1 SHARED INPUTS AND EXACT IMPLEMENTATIONS

All comparison arms receive the same answer/action, error score, development-fixed acquisition envelope, calibration labels and evaluation cases within each condition. Ineligible cases contribute zero coverage and yield to the population denominator. In the clinical Expanded panel, the common policy is the template selected by the original Exact development procedure. This tests controllers conditional on a shared policy; it does not claim independently optimal policy selection for each comparator. The policy-by-planner factorial compares planners with their own development-selected Expanded templates, while Current isolates ordering.

MedQA's primary uniform panel retains the cached 200 reduced-budget calibration permutations and complete 400-question pool at full budget. The secondary focused panel selects a common envelope by the original development-only Local/Marginal procedure: 4,096 training/check worlds, candidate development coverages near 25%, 50%, 75%, 90% and 100%, and widest-envelope ties. Every method uses the same resulting eligible calibration pool and release ceiling. Conditional Exact uses 2,048 training and 2,048 check worlds, the recorded seed formulas, 21 within-envelope score quantiles plus endpoints, development support five and the mandatory terminal inside the dynamic program. Matched LTT and Marginal use these same arrays. The cached Fractional Joint reference retains its distinct grid and frozen orders; it is not relabeled Exact. Uniform Exact has eight higher, fourteen lower and two tied means against that reference, with seven positive and eight negative simultaneous intervals. Both results remain available.

SCoRE. We implement SCoRE-SDR Algorithm 3 (Bai & Jin, 2026) with $\gamma = \alpha$ and e-BH (Wang & Ramdas, 2022), its homogeneous boosted version, and its binary specialization (Corollary 5.4). The main comparison designates boosted SCoRE and retains both alternatives. Source checks match the pinned author implementation, including tied scores and step-up boundaries; its commit is recorded in the source README. Homogeneous randomization uses fixed shared draws across matched settings. The native procedure uses the unlabeled evaluation batch, so its guarantee and resampling conditioning include that batch. Its expectation guarantee is not replaced by a per-calibration assertion in native tables.

LEC. Version 3, Equation (5), selects the largest observed calibration-score cutoff whose full tied prefix satisfies $K + 1 \leq \alpha N$. This native cutoff need not lie on PCO's development grid. The implementation preserves complete score ties and is checked against the stated criterion, including small-sample and all-correct/all-error cases. Its target is expected selected errors divided by expected selections. SCoRE and LEC do not take δ; uniform MedQA native results consequently repeat across confidence panels. Focused results can change when the shared development envelope changes with δ.

SCRC correctness sets. The scalar score s supplies two-class scores $[s, 1-s]$ for incorrect/correct. SCRC-I uses the original calibration-only construction, fixed selection target $\xi = 0.7$ and eleven selection cutoffs. We match the public rank transformation and feasibility convention at the pinned source commit recorded in the README. Set thresholds are solved at exact observed breakpoints rather than approximate bisection. Native admission, selected-set size and set miscoverage are retained separately from the additional singleton-$\{\text{correct}\}$ answer-release rule. Scores here are not full answer-option probabilities; this is not a native multiclass SCRC MedQA comparison. SCRC-I's PAC selected-set theorem is respected, but does not automatically transfer to the singleton-conditioned answer-error ratio.

G.2 A COMMON CERTIFICATE FOR ADAPTIVE PROPOSALS

Conditional on development, fix the finite threshold policies $A_j = \{X \in E, s(X) \leq t_j\}$ and risks R_j. Compute the exact-binomial p_j from independent calibration, assigning $p_j = 1$ when $N_j < 10$. For a true null $R_j > \alpha$, p_j is super-uniform. Strict Holm step-down testing produces a rejection set $\mathcal{C}_H$ satisfying

$$\Pr\{\exists j \in \mathcal{C}_H : R_j > \alpha\} \leq \delta. \tag{25}$$

For any possibly data-dependent choice $\widehat{j} \in \mathcal{C}_H$, the event $R_{\widehat{j}} > \alpha$ is contained in this familywise event. Thus Equation 2 holds even when choice within the certified family uses a native calibration-dependent proposal, randomization or the unlabeled batch. No new label split is needed.

For each native proposal, consider certified grid policies whose entire release sets on the current batch are contained in that proposal. Choose greatest calibration support with lower-threshold ties, or abstain if none qualifies. The returned object is the full grid policy $A_{\widehat{j}}$, whose population risk is certified. Batch containment is the selection criterion; it is not a claim of global containment in the native rule. An arbitrary subset or intersection could have a different conditional risk and would not inherit this proof. Holm grid uses the same certificate without the proposal restriction. The comparison measures these specified conservative conversions; more powerful valid conversions may exist. The certification kernel was checked over 53,130 finite count configurations and against independent scalar certifiers.

G.3 Risk diagnostics and complete reporting

For held-out batch b, let M_b and H_b be released cases and released errors. Define $r_b = H_b/M_b$ when $M_b > 0$, and zero otherwise. For that batch's target α_b, record $e_b = \mathbf{1}\{r_b > \alpha_b\}$ and $v_b = (r_b - \alpha_b)_+$. Equality is handled with tolerance 10^{-12}. Frequency is $\overline{e}$; severity is $\overline{v}/\overline{e}$ when $\overline{e} > 0$. Each draw is weighted within its unit, then units and targets equally. Empty-batch and nonempty-only exceedance rates, exceedance by more than five points, mean batch risk, actual labels and screening are also preserved. These repeated batches reuse cases and calibration pools; independent-binomial intervals would be inappropriate. The displayed frequency and severity are descriptive diagnostics. A true safe policy can exceed its target on a finite evaluation sample, and an expectation guarantee can coexist with individual exceedances.

Tables 13–14 retain the complete certified useful-yield grid for the primary acquisition panels. Table 15 retains every uniform MedQA Exact–Marginal contrast, including negative and zero settings. The separate focused MedQA grid is in the source package. Full native variants, target-specific results, SCRC set diagnostics and all clinical Current/Expanded acquisition panels are in the adjacent result CSVs and notebooks. They are not replaced by averages from a chosen target. In particular, the stronger native MedQA SCRC results in the main text remain part of the interpretation.

Table 13: Complete useful-yield percentages at 90% certification. HealthAdvice/CARE: Expanded, shared focused; MedQA: uniform. +H denotes the specified Holm certificate; SCRC† is the correctness-set adaptation. Full caps are 80 and 271–272, respectively. All six targets are averaged.

Benchmark / model	Cap	Exact	Holm	SCoRE+H	LEC+H	SCRC†+H
HealthAdvice	20	19.41	3.71	3.71	3.63	1.98
HealthAdvice	40	31.03	12.95	12.95	12.69	8.79
HealthAdvice	full	42.34	25.19	25.19	25.04	17.57
CARE	20	8.34	2.22	2.22	2.19	1.28
CARE	40	12.19	7.71	7.71	7.58	5.21
CARE	80	14.53	10.99	10.99	10.89	7.45
CARE	160	18.91	13.58	13.56	13.56	11.22
CARE	full	22.91	14.87	14.87	14.87	14.26
MedQA / GPT-4o	20	51.52	18.68	18.68	18.20	7.43
MedQA / GPT-4o	40	73.56	50.36	50.35	49.80	35.32
MedQA / GPT-4o	80	82.04	71.61	71.52	71.07	54.39
MedQA / GPT-4o	160	85.21	82.27	82.14	81.88	61.05
MedQA / GPT-4o	320	86.38	85.73	85.67	85.52	61.01
MedQA / GPT-4o	400	86.50	86.08	86.08	85.91	59.61
MedQA / GPT-4o-mini	20	43.12	12.20	12.20	12.17	4.54
MedQA / GPT-4o-mini	40	66.18	42.08	41.77	42.06	32.71
MedQA / GPT-4o-mini	80	72.51	63.43	62.78	63.40	52.61
MedQA / GPT-4o-mini	160	74.53	72.28	71.80	72.25	59.86
MedQA / GPT-4o-mini	320	74.98	74.50	74.22	74.49	60.64
MedQA / GPT-4o-mini	400	75.08	74.68	74.68	74.68	60.27

Table 14: The same complete useful-yield comparison at 95% certification. Policies, label budgets and target averaging match the preceding table.

Benchmark / model	Cap	Exact	Holm	SCoRE+H	LEC+H	SCRC†+H
HealthAdvice	20	13.50	2.59	2.59	2.34	1.09
HealthAdvice	40	22.54	9.68	9.68	9.57	6.50
HealthAdvice	full	33.45	21.51	21.51	21.43	15.47
CARE	20	5.71	1.51	1.51	1.44	0.71
CARE	40	10.37	6.27	6.27	6.17	4.22
CARE	80	13.46	10.36	10.36	10.30	7.41
CARE	160	16.85	12.13	12.13	12.11	10.30
CARE	full	19.76	12.95	12.66	12.95	12.44
MedQA / GPT-4o	20	41.07	11.79	11.79	11.16	4.07
MedQA / GPT-4o	40	65.02	43.88	43.88	43.44	32.85
MedQA / GPT-4o	80	79.24	67.79	67.73	67.30	51.88
MedQA / GPT-4o	160	84.21	80.86	80.75	80.49	60.40
MedQA / GPT-4o	320	86.09	85.42	85.37	85.22	60.06
MedQA / GPT-4o	400	86.15	85.88	85.88	85.71	59.86
MedQA / GPT-4o-mini	20	33.06	6.01	6.01	5.99	1.28
MedQA / GPT-4o-mini	40	57.67	36.29	36.08	36.28	29.32
MedQA / GPT-4o-mini	80	70.83	58.73	58.19	58.70	48.98
MedQA / GPT-4o-mini	160	74.05	71.16	70.70	71.13	58.49
MedQA / GPT-4o-mini	320	74.83	74.30	74.03	74.29	60.04
MedQA / GPT-4o-mini	400	74.89	74.49	74.49	74.49	60.27

Table 15: Every uniform MedQA Exact–Marginal useful-yield contrast. Gains and interval endpoints are percentage points; simultaneous intervals retain the full 456-contrast family.

Model	Cap	Cert.	Gain (pp)	Paired 95%	Simultaneous
GPT-4o	20	90%	+1.76	$[+1.60, +1.92]$	$[+1.48, +2.03]$
GPT-4o	20	95%	+2.37	$[+2.18, +2.57]$	$[+2.02, +2.72]$
GPT-4o	40	90%	+2.20	$[+2.04, +2.36]$	$[+1.91, +2.48]$
GPT-4o	40	95%	+1.73	$[+1.61, +1.85]$	$[+1.52, +1.94]$
GPT-4o	80	90%	+0.89	$[+0.82, +0.97]$	$[+0.76, +1.02]$
GPT-4o	80	95%	+1.94	$[+1.83, +2.05]$	$[+1.74, +2.14]$
GPT-4o	160	90%	+0.78	$[+0.72, +0.83]$	$[+0.67, +0.88]$
GPT-4o	160	95%	+0.40	$[+0.30, +0.49]$	$[+0.22, +0.57]$
GPT-4o	320	90%	+0.18	$[+0.15, +0.21]$	$[+0.12, +0.23]$
GPT-4o	320	95%	+0.65	$[+0.56, +0.74]$	$[+0.50, +0.80]$
GPT-4o	400	90%	-0.07	$[-0.12, -0.02]$	$[-0.15, +0.01]$
GPT-4o	400	95%	-0.02	$[-0.06, +0.00]$	$[-0.08, +0.03]$
GPT-4o-mini	20	90%	+0.55	$[+0.48, +0.62]$	$[+0.43, +0.67]$
GPT-4o-mini	20	95%	+0.24	$[+0.19, +0.29]$	$[+0.16, +0.32]$
GPT-4o-mini	40	90%	+2.22	$[+2.07, +2.38]$	$[+1.94, +2.50]$
GPT-4o-mini	40	95%	+2.19	$[+2.01, +2.37]$	$[+1.88, +2.50]$
GPT-4o-mini	80	90%	+1.40	$[+1.28, +1.53]$	$[+1.18, +1.63]$
GPT-4o-mini	80	95%	+0.95	$[+0.88, +1.02]$	$[+0.82, +1.07]$
GPT-4o-mini	160	90%	+0.56	$[+0.53, +0.60]$	$[+0.50, +0.63]$
GPT-4o-mini	160	95%	+0.54	$[+0.48, +0.59]$	$[+0.44, +0.63]$
GPT-4o-mini	320	90%	+0.78	$[+0.68, +0.90]$	$[+0.59, +0.98]$
GPT-4o-mini	320	95%	0.00	$[-0.001, +0.00]$	$[-0.001, +0.00]$
GPT-4o-mini	400	90%	+0.00	$[+0.00, +0.00]$	$[+0.00, +0.00]$
GPT-4o-mini	400	95%	+1.19	$[+1.01, +1.37]$	$[+0.87, +1.51]$

H EPISODE CONTRACT AND SUPPORTING RESULTS

H.1 EPISODE RISK AND COMPLETION ARE DISTINCT

A CARE episode begins at its first cached prefix. An INFO action advances to the next recorded prefix; a care action terminates; exhaustion or a fixed later confidence gate causes handoff. Let L_π indicate any incorrect autonomous action along the trajectory, including an error before subsequent handoff. Let V_π indicate a correct terminal care action with no preceding action error. Then

$$0 \le V_\pi \le 1 - L_\pi, \tag{26}$$

and inequality can be strict for a safe but unfinished episode. Thus episode completion is not coverage minus error.

The outer admission score depends only on the initial prefix. The admitted trajectory, including later gates, does not change when the outer threshold changes. Hence

$$U_\pi(t) = \mathbb{E}[\mathbf{1}\{s_\pi(X_1) \le t\} V_\pi] \tag{27}$$

is nonnegative and nondecreasing, and the inner optimizer applies. An episode contains at least one autonomous action if admitted. Applying the exact-binomial test to independent patient-level losses L_π controls the probability of certifying an episode policy whose any-action error probability among admitted episodes exceeds α. This derives a single episode-level contract; it does not compose per-turn assurances into a conversation guarantee or certify error conditional on terminal completion alone. Safe unfinished episodes contribute to admission coverage, but not completion utility.

For a fixed policy and single-chain test, let $\mathcal{C} = \{j : N_j \ge 10,\ p_j < \delta\}$ be the individually certifiable cutoffs. Any returned cutoff belongs to $\mathcal{C}$, so

$$U_\pi(\widehat{t}) \le \max_{j \in \mathcal{C}} U_\pi(t_j), \qquad U_\pi(\varnothing) = 0, \quad \max_{j \in \varnothing} U_\pi(t_j) := 0. \tag{28}$$

If $\mathcal{C}$ is empty, the controller abstains and completion utility is zero regardless of ordering.

The cache contains 439 patients and 925 recorded prefixes. All prefixes of a patient retain their assigned data role. The trajectory follows the prerecorded disclosures, with budgets counted in patients and prefix-label use recorded separately.

H.2 FIXED AND LEARNED TRAJECTORY CONTROLS

The controls compare a fixed five-template bank with six development-learned Trajectory policies, using nested patient folds within the same 100 development patients and matched calibration draws. Completion averages six targets and both action-model families across five grouped patient allocations. At full budget and 90% certification, fixed-bank Expanded Exact and Marginal tie at 0.15%; Trajectory/Selected/Exact reaches 0.32%, versus 0.39% for the strongest LTT/Marginal reference across recorded banks and scores. At 95%, all three Trajectory/Selected pipelines tie at 0.07%. No matched Exact-minus-Marginal paired interval has a strictly positive lower bound across the ten cells; the 80-episode/90% difference is -0.02 pp with simultaneous band $[-0.04, -0.002]$ across 620 contrasts. At full budget and 90%, 81.7% of Trajectory/Selected/Exact conditions have no individually certifiable cutoff and 5.0% miss a passing candidate, giving 86.7% no release. The source package retains all learning settings, completion grids, prefix-label costs, 244,080 fixed-bank and 1,464,480 learned-policy contributions, and 620 contrasts.

I RISK CONTRACTS AND NOTATION

The following table aligns the frameworks' output objects and statistical contracts (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). The implemented task adaptations are specified in Appendices B.2 and G.

Sources are LTT (Angelopoulos et al., 2025; Quach et al., 2024), RCPS (Bates et al., 2021), CRC (Angelopoulos et al., 2024), SCoRE (Bai & Jin, 2026), LEC (Wang et al., 2026) and SCRC (Xu et al., 2025). The Pareto, adaptive LTT and reliability-graph procedures are specified in Appendix B.2.

Table 16: Conceptual comparison of risk-control frameworks. Distinct risk functionals should be reconciled before empirical comparison. Empirical comparisons use the explicitly matched procedures described in Appendix G.

Framework	Output and guarantee	Relation to this experiment
LTT / PCO	A selected rule; high-probability control over calibration of its population selected risk	PCO plans the first-failure order; LTT supplies validity under the required valid tests
RCPS	Prediction sets satisfying a high-probability population-loss bound	Related PAC risk control; loss and output object must be matched
CRC	Prediction sets (Vovk et al., 2005; Angelopoulos & Bates, 2023) or nested rules with expectation control for suitable monotone losses	Its marginal expectation contract is distinct from Equation 2
SCoRE	Selective predictions (Jin & Candès, 2023); e-values control marginal or batch selective risk functionals for general losses	Expected selected/batch risk is not the same conditioning as a calibration-PAC selected-population guarantee
LEC	Selective prediction/routing (Madras et al., 2018; Mozannar & Sontag, 2020); controls expected selected errors relative to expected selections	A ratio of expectations differs from a high-probability bound on the selected rule's conditional risk
SCRC	Prediction sets on selected cases; transductive and calibration-only PAC variants	SCRC-I includes a PAC formulation; selection, prediction-set loss and output objects still differ

Symbol / quantity	Meaning and conditioning
α, δ	Selected-risk target and calibration error probability; $1-\delta$ is certification confidence
η	Failure probability for a predictive-simulation utility bound; separate from δ and empirical bootstrap intervals
E, n, N_j	Development-fixed acquisition envelope, nominal label cap, selected calibration support for cutoff j; screened cases and actual acquired labels differ
$B_j^{(b)}$, B	Pass indicator for cutoff j in world b, and number of iid predictive worlds; no claim that repeated clinical splits are iid new patients
$\widehat{F}$, $\widehat{G}$	Last-passing surrogate and implemented support-selected utility averaged over worlds; they can differ at support ties
$\widehat{\tau}$, ζ	Observed surrogate-to-selector tie loss for the returned plan and a bound on its optimization residual
$\epsilon_{\text{forecast}}$	Assumed uniform discrepancy between forecast and population utility; not numerically bounded by the reported experiments

I.1 Notation and boundaries of the claims

The iid calibration and positive-mass release assumptions apply to the stated binomial test. The zero-coverage decision is abstention; no conditional risk is assigned to an empty release set. A numerical threshold becoming certified says nothing about clinical acceptability of the chosen α. Observed evaluation error ratios, calibration validity, predictive regret and conditional bootstrap performance uncertainty remain four distinct objects.

J Additional comparison tables

This appendix retains all nine supplementary methods, the primary small-budget Exact–LTT contrasts and the complete primary Exact–Holm grid. Full forecasting, episode, completion-off and secondary MedQA tables remain in the source package, indexed by its README.

J.1 SUPPLEMENTARY NINE-METHOD COMPARISONS

These uniform-acquisition experiments use the recorded variants: Regular PCO, Nested PCO and Joint fractional PCO, with 4,096-world forecasts and their recorded completion conventions (Appendix B.2). They are distinct from the primary conditional Exact experiment. All budgets, both certification levels and all nine methods are retained. Bold identifies the largest observed mean in a column.

Table 17: HealthAdvice useful yield (%) for all nine methods, every budget and both certification levels. Means average all six risk targets and both configurations.

	90% certification			95% certification		
Method	20	40	80 (full)	20	40	80 (full)
Fixed-sequence LTT	15.45	29.28	42.38	9.47	21.07	**32.62**
aLTT-FWER	1.20	5.76	14.90	0.38	3.62	11.11
Pareto Testing	15.53	**29.34**	42.38	9.47	21.09	**32.62**
RG-PT (FWER)	9.45	20.70	33.71	5.73	15.24	28.23
Bonferroni	3.13	12.04	24.92	1.92	8.30	20.42
Marginal simulation	15.65	28.19	41.33	9.30	20.78	31.46
Regular PCO	15.71	28.20	41.89	9.33	20.77	31.38
Nested PCO	16.91	29.10	42.34	10.91	**21.63**	32.47
Joint fractional PCO	**16.93**	28.66	**42.42**	**11.46**	20.88	32.47

Table 18: CARE current-action useful yield (%) for all nine methods, every budget and both certification levels. Means average all six risk targets and both configurations.

	90% certification					95% certification				
Method	20	40	80	160	Full	20	40	80	160	Full
Fixed-sequence LTT	0.00	0.04	3.39	16.07	22.69	0.00	0.02	2.09	11.84	19.69
aLTT-FWER	0.01	0.22	1.13	4.25	10.52	0.00	0.11	0.58	2.91	9.00
Pareto Testing	0.00	0.04	3.36	15.92	22.66	0.00	0.02	2.08	11.75	19.52
RG-PT (FWER)	0.16	0.88	3.09	8.35	13.84	0.05	0.41	1.82	6.27	12.68
Bonferroni	0.09	0.71	2.74	7.82	13.35	0.03	0.36	1.68	5.92	11.80
Marginal simulation	0.01	1.95	8.57	16.05	19.67	0.01	0.02	4.96	12.58	16.91
Regular PCO	0.00	2.01	9.63	18.19	23.11	0.00	0.00	5.82	14.42	**20.11**
Nested PCO	1.74	**3.78**	9.89	18.19	23.11	0.58	1.95	6.05	14.42	**20.11**
Joint fractional PCO	**2.07**	3.74	**10.43**	**18.35**	**23.35**	**0.94**	**1.97**	**6.51**	**14.74**	20.06

J.2 SMALL-BUDGET MATCHED EXACT–LTT COMPARISONS

The clinical rows fix Current policy and shared focused acquisition; MedQA uses uniform acquisition with the cached answers. All contrasts hold the policy, forecast, candidate grid and calibration labels fixed. Clinical simultaneous intervals use the 840-contrast family in Appendix B.3; MedQA retains its 456-contrast family. The source package includes the complete contrast grids.

Table 19: Exact minus matched development-p-value LTT at 20–40 labels. Useful-yield gains, paired 95% intervals and simultaneous intervals are percentage points, averaged over all six risk targets.

Benchmark/model	Cap	Cert.	Gain (pp)	Paired 95%	Simultaneous
HealthAdvice	20	90%	+2.04	$[+1.80, +2.29]$	$[+1.54, +2.55]$
HealthAdvice	20	95%	+2.91	$[+2.63, +3.21]$	$[+2.33, +3.49]$
HealthAdvice	40	90%	+0.28	$[+0.10, +0.46]$	$[-0.08, +0.64]$
HealthAdvice	40	95%	-0.52	$[-0.79, -0.23]$	$[-1.09, +0.06]$
CARE	20	90%	+7.81	$[+6.23, +9.44]$	$[+4.55, +11.07]$
CARE	20	95%	+5.26	$[+4.18, +6.38]$	$[+3.03, +7.49]$
CARE	40	90%	+2.78	$[+2.10, +3.49]$	$[+1.36, +4.20]$
CARE	40	95%	+3.38	$[+2.64, +4.15]$	$[+1.84, +4.93]$
MedQA/GPT-4o	20	90%	+2.13	$[+2.04, +2.23]$	$[+1.96, +2.30]$
MedQA/GPT-4o	20	95%	+2.65	$[+2.56, +2.75]$	$[+2.49, +2.82]$
MedQA/GPT-4o	40	90%	+0.02	$[-0.01, +0.05]$	$[-0.04, +0.08]$
MedQA/GPT-4o	40	95%	+0.55	$[+0.48, +0.63]$	$[+0.42, +0.69]$
MedQA/GPT-4o-mini	20	90%	+1.04	$[+0.94, +1.15]$	$[+0.86, +1.23]$
MedQA/GPT-4o-mini	20	95%	+0.49	$[+0.39, +0.59]$	$[+0.31, +0.66]$
MedQA/GPT-4o-mini	40	90%	+0.34	$[+0.29, +0.39]$	$[+0.25, +0.43]$
MedQA/GPT-4o-mini	40	95%	+0.73	$[+0.69, +0.76]$	$[+0.66, +0.79]$

J.3 Complete primary Exact–Holm comparisons

Current-policy Exact–Marginal gains remain in Table 3; coverage, risk and repeated-event counts appear in Tables 8–9. Table 20 retains every focused and uniform Exact–Holm contrast under the unchanged 756-contrast simultaneous family.

Table 20: Current Exact minus matched Holm at every budget and certification level. Gain, paired 95% interval and simultaneous interval are percentage points.

Shared focused acquisition

Benchmark	Cap	Cert.	Gain (pp)	Paired 95%	Simultaneous
HealthAdvice	20	90%	+14.69	[+13.44, +15.90]	[+12.19, +17.19]
HealthAdvice	20	95%	+10.67	[+9.83, +11.47]	[+8.99, +12.34]
HealthAdvice	40	90%	+17.68	[+16.32, +18.98]	[+14.98, +20.38]
HealthAdvice	40	95%	+12.96	[+11.83, +14.06]	[+10.71, +15.20]
HealthAdvice	Full	90%	+17.23	[+15.88, +18.58]	[+14.49, +19.97]
HealthAdvice	Full	95%	+11.67	[+10.62, +12.69]	[+9.59, +13.76]
CARE	20	90%	+6.12	[+4.90, +7.37]	[+3.62, +8.62]
CARE	20	95%	+4.05	[+3.23, +4.89]	[+2.36, +5.74]
CARE	40	90%	+4.64	[+3.72, +5.60]	[+2.72, +6.55]
CARE	40	95%	+4.10	[+3.27, +4.93]	[+2.41, +5.78]
CARE	80	90%	+3.54	[+2.80, +4.30]	[+1.99, +5.09]
CARE	80	95%	+3.56	[+2.84, +4.30]	[+2.08, +5.05]
CARE	160	90%	+5.34	[+4.50, +6.20]	[+3.62, +7.06]
CARE	160	95%	+4.78	[+4.04, +5.55]	[+3.26, +6.31]
CARE	Full	90%	+8.38	[+7.40, +9.41]	[+6.33, +10.44]
CARE	Full	95%	+6.81	[+5.95, +7.67]	[+5.03, +8.59]

Uniform acquisition

Benchmark	Cap	Cert.	Gain (pp)	Paired 95%	Simultaneous
HealthAdvice	20	90%	+13.82	[+12.84, +14.77]	[+11.85, +15.79]
HealthAdvice	20	95%	+9.62	[+8.91, +10.27]	[+8.25, +10.98]
HealthAdvice	40	90%	+16.51	[+15.35, +17.64]	[+14.17, +18.86]
HealthAdvice	40	95%	+12.45	[+11.51, +13.38]	[+10.55, +14.35]
HealthAdvice	Full	90%	+16.82	[+15.47, +18.13]	[+14.08, +19.55]
HealthAdvice	Full	95%	+11.75	[+10.70, +12.76]	[+9.68, +13.82]
CARE	20	90%	+1.88	[+1.69, +2.05]	[+1.51, +2.25]
CARE	20	95%	+0.84	[+0.76, +0.92]	[+0.67, +1.01]
CARE	40	90%	+2.94	[+2.59, +3.29]	[+2.23, +3.65]
CARE	40	95%	+1.58	[+1.37, +1.78]	[+1.16, +2.00]
CARE	80	90%	+7.56	[+6.56, +8.56]	[+5.53, +9.60]
CARE	80	95%	+4.90	[+4.14, +5.67]	[+3.36, +6.45]
CARE	160	90%	+10.01	[+8.83, +11.19]	[+7.63, +12.40]
CARE	160	95%	+8.36	[+7.25, +9.45]	[+6.13, +10.58]
CARE	Full	90%	+9.44	[+8.36, +10.50]	[+7.26, +11.62]
CARE	Full	95%	+7.84	[+6.83, +8.82]	[+5.82, +9.86]