

Multi Armed Bandit Based Adaptive Model Selection for Clinical AI Governance An Infrastructure Approach to Safer AI Use in U.S. Healthcare

Julian Borges

fxmedbrasil@gmail.com

Boston University <https://orcid.org/0009-0001-9929-3135>

Research Article

Keywords: clinical artificial intelligence, AI governance, multi armed bandits, best arm identification, delayed feedback, model deployment, intensive care, MIMIC-IV, reproducibility

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8651972/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Multi armed bandit based adaptive model selection for clinical AI governance: a fixed confidence simulation study on MIMIC-IV with delayed feedback

Julian Borges, MD, MS

Julian Borges, MD, MS

Department of Computer Science, Boston University Metropolitan College, Boston, MA, USA

Corresponding author: Julian Borges, Department of Computer Science, Boston University Metropolitan College, 1010 Commonwealth Avenue, Boston, MA 02215, USA. Email: jyborges@bu.edu. ORCID 0009-0001-9929-3135.

Word count (main text, excluding abstract, tables, figure legends and references): 9,815

Abstract

Health systems increasingly hold several validated models for one prediction task, yet deployment fixes one on retrospective validation. We evaluate a governance layer that treats model selection as a sequential decision problem: three frozen ICU mortality models with different operational footprints are decision arms, a composite reward combines predictive quality, a declared cost and a penalty for missed deaths with explicit governance weights, and fixed confidence action elimination selects among arms under delayed feedback with a bounded probability of retiring the best. It was compared with Thompson sampling, UCB1, random allocation, static deployment, ensemble averaging and heuristic escalation on 7,383 MIMIC-IV v3.1 ICU stays from 2020 to 2022, with models developed on 20,000 earlier stays (100 replicates). Under the primary weights the cheapest arm was preferred, with gaps of 0.015 and 0.040 made entirely by the cost term. Elimination reached no decision within the horizon with either radius: the stream supplies about 2,461 outcomes per arm against 7,508 (6,814 to 8,209) needed to retire the most expensive arm, as a 30,000 decision resample of the same stays confirmed. Thompson sampling attained the lowest regret but uncontrolled exposure. When the validation chosen model was degraded on the stream, regret minimising policies beat its static deployment within the stream, while confident retirement of even a broken model needed about 3,060 outcomes per arm. The magnitudes are specific to this portfolio, database and cost model,

but the method quantifies the evidence horizon confident retirement requires. Every number comes from one public pipeline run.

Keywords: clinical artificial intelligence; AI governance; multi armed bandits; best arm identification; delayed feedback; model deployment; intensive care; MIMIC-IV; reproducibility

Highlights

- Deployment model selection is formalised as fixed confidence best arm identification
- A composite reward makes cost, safety and alert weights explicit governance settings
- No arm retired on 7,383 ICU stays; the gaps imply 23,000 to 118,000 decisions
- A wrong validation choice is abandoned by regret minimisers within the stream
- Thompson sampling lowers regret but leaves exposure to inferior models uncontrolled

Statement of significance

Problem	Institutions that hold several validated clinical prediction models deploy one of them statically and have no principled, auditable mechanism to decide when evidence from routine use justifies switching or retiring a model, under delayed outcomes and explicit cost and safety priorities.
What is already known	Multi armed bandit and best arm identification algorithms offer confidence controlled sequential decisions, and adaptive trial designs have used them; external validations show that deployed models drift and that error patterns differ across subgroups. Their combination as a deployment governance layer, with delayed feedback, cost and safety terms, and quantified sample complexity, has not been evaluated on real ICU data.
What this paper adds	A fully specified, reproducible governance framework with a composite reward whose weights and cost model are institutional parameters; a simulation on a calendar ordered MIMIC-IV stream showing that fixed confidence elimination cannot retire any of three realistic arms within the stream, together with the evidence horizon that would be needed, its uncertainty, its dependence on the cost model and its confirmation on an extended horizon; a demonstration that the layer retires a model whose validation choice is wrong on the stream; and a comparison with regret minimising policies that quantifies the exposure they leave uncontrolled.
Who would benefit	Clinical AI governance committees planning model portfolios and retirement decisions, informatics teams implementing monitoring of deployed models, and methodologists designing adaptive deployment studies.

1. Introduction

Clinical prediction models are usually deployed one at a time. A health system validates a candidate on retrospective data, selects the one with the best aggregate discrimination, and applies it to every patient until a scheduled review. This static pattern treats model selection as a configuration decision made once, on average performance, rather than as a decision that recurs

with every patient and that an institution remains accountable for over time. External validations of widely implemented models have shown that aggregate performance can differ materially from the performance observed in a new setting or period [1,2], that error patterns cluster in patient subgroups [3,4], and that models can rely on shortcut features that fail under contextual change [5,6]. Audits identify such failures but do not control model exposure during routine use.

A growing number of institutions now hold several validated models for the same prediction task, differing in the inputs they require, their computational footprint and their calibration. Choosing among them for each patient is a sequential decision under uncertainty: which model to run, what it costs to run it, which errors are most harmful, and when enough evidence has accumulated to retire an alternative. Multi armed bandit algorithms formalise exactly this problem [7]. In the fixed confidence best arm identification setting, an algorithm eliminates arms whose confidence intervals fall below a competitor and stops with a stated error probability [8-10], which maps directly onto the question a governance committee asks: with what confidence can we retire this model. Deployment adds constraints that classical formulations ignore: outcomes arrive hours to days after a prediction [11,12], running a model has an operational cost, and some errors (a missed death) matter more than others.

This study specifies and evaluates such a governance layer. Three frozen in hospital mortality models with different operational footprints are treated as decision arms. A composite reward combines predictive quality, operational cost and a safety penalty for missed deaths, with weights that are governance parameters rather than tuned quantities. A fixed confidence action elimination policy selects among the arms under delayed feedback, and is compared with Thompson sampling, UCB1, uniform random allocation, static deployment, ensemble averaging and a heuristic escalation rule in a simulated deployment on a temporally external stream of 7,383 ICU admissions from MIMIC-IV v3.1 [13,14].

The contribution is methodological and governance oriented. The framework is specified so that every quantity an oversight body would need is explicit and reproducible: the weights that define the preferred model, the confidence level that bounds the probability of a wrong retirement, the delay structure, and the number of observed outcomes a confident decision requires. The results quantify what fixed confidence elimination can and cannot decide within a realistic deployment

horizon, how that depends on the governance weights, and how regret minimising alternatives trade decision certainty for lower exposure to inferior models.

This article replaces an earlier preprint of the same project [15], which carried a different title and reported a positive result on 3,000 admissions. An audit of the archived code found four defects: the three models had been trained and evaluated on the same stays, the elimination policy described in the text had been replaced in the code by an epsilon greedy placeholder without delayed feedback or a safety term, the comparator strategies had been chosen on the evaluation data, and several reported quantities had never been computed by the code. The analysis was rebuilt in full, the direction of the primary comparison reversed, and every number reported here is produced by a single run of the public pipeline (Section 2.12 and the code availability statement).

2. Methods

2.1. Study design

We conducted a retrospective simulation of sequential model deployment. The unit of decision was an index ICU stay. At each decision a governance policy observed the patient context available 24 hours after ICU admission, selected one of three frozen mortality models, deployed its calibrated prediction, and received the in hospital mortality outcome only after a stochastic delay. Models were trained, calibrated and thresholded on stays that occurred before 2020 and deployed on stays of patients anchored in 2020 to 2022, so that the evaluation includes the temporal drift a governance layer must manage. The study is a phase zero evaluation: it characterises deployment behaviour in simulation before any prospective use, and it makes no clinical recommendation.

2.2. Data source and ethics

Data came from the Medical Information Mart for Intensive Care IV, version 3.1 (MIMIC-IV v3.1), a de identified electronic health record database of hospital and ICU admissions at the Beth Israel Deaconess Medical Center [13,14], accessed through PhysioNet [16] under the credentialed data use agreement. Seven tables were read: patients, admissions and ICU stays (cohort), chart events with the item dictionary (vital signs), and laboratory events with the laboratory item dictionary. No other table was used. The collection of MIMIC-IV was approved by the institutional review boards of the Beth Israel Deaconess Medical Center and the

Massachusetts Institute of Technology with a waiver of informed consent, and the database is distributed de identified under that approval [13,14]; the present secondary analysis of de identified data therefore did not constitute human subjects research and required no further institutional review board approval.

2.3. Cohort

Eligible stays were index ICU stays (the first ICU stay of a hospital admission) of adults aged 18 years or older at admission, with complete admission, discharge and ICU timestamps in valid order, recorded sex, and an ICU length of stay between 24 hours and 60 days. The lower bound ensures that the 24 hour feature window is fully observed and that no outcome occurs inside it; the upper bound removes implausible records. The outcome was in hospital mortality.

Development and deployment sets were disjoint at the patient level and ordered in calendar time. MIMIC-IV assigns each patient an anchor year group, the real three year window in which the patient's anchor year fell, and shifts every date of the patient by one patient specific offset, so a stay admitted k years after the anchor year occurred in that window shifted by k years. Patients anchored in 2020 to 2022 formed the deployment pool; every stay of such a patient occurred in or after 2020. Patients anchored in 2008 to 2019 formed the development pool, from which a stay was retained only if its shifted window ended before 2020, so that every development stay is guaranteed to precede every deployment stay; stays that could not be placed before 2020 (those admitted long enough after the anchor year for the shifted window to reach 2020) were excluded, and the pipeline asserts the ordering. One index stay per patient, drawn at random among the patient's eligible stays, was retained in each pool to limit within patient correlation. From the development pool, 20,000 stays were sampled and split by patient into a training set (75 percent) and a validation set (25 percent); the validation set was further halved by patient into a half used for early stopping and tuning and a half used for calibration, thresholds and the choice of the static comparator. All 7,383 patients of the deployment pool formed the deployment stream. Because the date shift differs between patients, stays cannot be ordered chronologically across patients within a pool; the stream was therefore ordered by a seeded random permutation in each simulation replicate. Race as recorded on the admission was grouped into five categories (White, Black, Hispanic or Latino, Asian, other or unknown) for descriptive subgroup analyses.

2.4. Features

For each stay, vital signs and laboratory tests charted in the first 24 hours after ICU admission were summarised as mean, minimum, maximum and last value. Vital signs comprised heart rate, respiratory rate, temperature, oxygen saturation, systolic, diastolic and mean arterial pressure, and the Glasgow Coma Scale total computed from components charted at the same time. Laboratory tests comprised white blood cell count, hemoglobin, platelets, sodium, potassium, chloride, bicarbonate, urea nitrogen, creatinine and glucose (the core panel) and lactate, total bilirubin and INR (the extended panel). Item identifiers were fixed in advance and verified against the MIMIC-IV dictionaries at run time (Supplementary Table S7); values outside prespecified physiologic ranges were discarded. Extraction used DuckDB [17] with predicate pushdown on the PhysioNet files. Each laboratory concept carried a missing indicator; remaining gaps were filled with training set medians. Age and sex completed the context vector.

2.5. Decision arms

Three models with distinct operational footprints were trained on the training set and then frozen (Table 1). Arm A was a ridge penalised logistic regression [18] on age, sex and the vital sign features (normalised cost 0.25). Arm B was a gradient boosted tree model [19] adding the core laboratory panel and its missing indicators (cost 0.60). Arm C was a random forest [20,21] adding the extended laboratory panel (cost 1.00). The normalised cost represents the operational footprint of running a model with its full input set (data integration, acquisition and maintenance burden of the inputs it depends on) as an institution would assign it; it is a declared constant, not a measured quantity, and in this retrospective stream no test was ordered by any arm, since every laboratory value had been obtained for clinical reasons before the decision point. Because the arm defined by the extended panel is charged its full cost even when the panel is absent from the record and its inputs are imputed, an acquisition conditional costing scenario in which arm C is charged at arm B's cost whenever the extended panel is absent is reported as a sensitivity analysis, together with alternative cost vectors (Section 2.11). All three arms were tuned by the same protocol on the early stopping half of the validation set: the penalty of arm A by tenfold cross validation within the training set, the number of boosting rounds of arm B by early stopping, after which the model was refitted with that number of rounds, and the number of candidate variables per split and the minimum node size of arm C by log loss over a grid of nine settings with 500 trees. Each arm was calibrated by Platt scaling [22] on the calibration half of

the validation set and its operating threshold was set there to achieve 80 percent sensitivity. Discrimination was summarised by the area under the receiver operating characteristic curve (AUROC) with DeLong confidence intervals [23], the area under the precision recall curve, the Brier score [24] and the logistic calibration intercept and slope [25]. No stream observation was used for training, tuning, calibration or threshold selection; the pipeline asserts the disjointness of patient identifiers.

2.6. Problem formulation

Let $\mathcal{A} = \{1, \dots, K\}$ denote the arms ($K = 3$) and T the number of decisions in the stream. At decision t the policy observes the context x_t , selects an arm $a_t \in \mathcal{A}_t \subseteq \mathcal{A}$ from the currently active set, and later observes the reward $r_{a_t,t} \in [0,1]$ defined in Section 2.7. Because every frozen model can be scored on every patient after the fact, the reward of each unchosen arm is also available for evaluation, although the policy never sees it. Two regret definitions are used. The primary one compares the policy with the best fixed arm in hindsight,

$$R_T = \sum_{t=1}^T (r_{a^\circ,t} - r_{a_t,t}), \quad a^\circ = \operatorname{argmax}_{a \in \mathcal{A}} \sum_{t=1}^T r_{a,t},$$

that is, the cumulative reward forgone relative to deploying, for the whole stream, the single arm that turns out best over that stream. The secondary one compares the policy with a per decision oracle that always picks the arm with the highest realised reward,

$$R_T^{\text{oracle}} = \sum_{t=1}^T \left(\max_{a \in \mathcal{A}} r_{a,t} - r_{a_t,t} \right),$$

which no implementable policy can attain and which bounds the value of context dependent selection.

2.7. Composite reward

For arm a with calibrated probability $p_{a,t}$ and label y_t , the raw reward combined predictive quality, cost and safety,

$$r_{a,t}^{\text{raw}} = \left(1 - (p_{a,t} - y_t)^2 \right) - \lambda_c c_a - \lambda_s s_{a,t}, \quad s_{a,t} = \mathbb{1}\{y_t = 1 \text{ and } p_{a,t} < \tau_a\},$$

where $1 - (p - y)^2$ is one minus the Brier loss of the deployed prediction (a proper scoring rule bounded in $[0,1]$), $c_a \in [0,1]$ is the normalised operational cost of the arm, $s_{a,t}$ indicates a

missed death at the arm's operating threshold τ_a , and λ_c and λ_s are governance weights. The reward was rescaled to the unit interval,

$$r_{a,t} = \frac{r_{a,t}^{\text{raw}} + \lambda_c + \lambda_s}{1 + \lambda_c + \lambda_s} \in [0,1],$$

which is required by the concentration inequalities of Section 2.8 and holds because $c_a \leq 1$. The primary weights were $\lambda_c = 0.10$ and $\lambda_s = 0.50$; they encode a preference for the cheapest adequate model and a strong penalty on missed deaths, and were varied in sensitivity analyses. The primary reward carries no term for a false alert, because at thresholds fixed to a common sensitivity the alert burden of an arm is a property of its specificity that the Brier utility already rewards in part, and because the weight an institution places on an unnecessary alert is less settled than the weight on a missed death; the omission is consequential, however, since the arms differ in alert rate, so a false positive term $-\lambda_f \mathbb{1}\{y_t = 0 \text{ and } p_{a,t} \geq \tau_a\}$ with the rescaling extended to $1 + \lambda_c + \lambda_s + \lambda_f$ was added in sensitivity analyses, and the break even weight at which the preferred arm changes is reported. Because the Brier utility rewards calibration as well as discrimination, a further sensitivity analysis replaced $1 - (p_{a,t} - y_t)^2$ by binary correctness at the operating threshold, $\mathbb{1}\{(p_{a,t} \geq \tau_a) = (y_t = 1)\}$, which is also bounded in $[0,1]$ and leaves the rescaling unchanged. Composite strategies that run more than one model (Section 2.9) can incur a cost above 1; their rewards can then fall below 0 and are reported on the same scale without clipping, outside the range to which the concentration inequalities of Section 2.8 apply, which is immaterial because these strategies learn nothing from the rewards.

2.8. Fixed confidence action elimination

Let n_a be the number of outcomes observed for arm a and $\hat{\mu}_a$ the mean of its observed rewards. The policy maintains an active set, cycles through it in round robin order so that evidence accrues evenly, updates n_a and $\hat{\mu}_a$ only when delayed feedback arrives, and removes arm i from the active set when

$$\hat{\mu}_i + \beta_i(n_i) < \max_{j \neq i, j \in \mathcal{A}_t} (\hat{\mu}_j - \beta_j(n_j)),$$

that is, when the upper confidence bound of i lies below the highest lower confidence bound of another active arm [8,9]. Two confidence radii were used. The Hoeffding radius [26],

217

$$\beta^H(n) = \sqrt{\frac{\log(4Kn^2/\delta)}{2n}},$$

218

depends only on the sample size. The empirical Bernstein radius [27],

219

$$\beta^B(n) = \sqrt{\frac{2 \hat{V}_n L_n}{n}} + \frac{7 L_n}{3(n-1)}, \quad L_n = \log\left(\frac{16Kn^2}{\delta}\right),$$

220

uses the sample variance $\hat{V}_n$ of the observed rewards and is tighter when rewards concentrate, as composite rewards do. In both cases the factor Kn^2 inside the logarithm pays for a union bound over the K arms and over all sample sizes (using $\sum_{n \geq 1} n^{-2} < 2$), so that the probability that any arm's estimate ever leaves its interval, and therefore the probability that the best arm is ever eliminated, is at most δ . The primary confidence level was $\delta = 0.05$. Until an arm is eliminated the policy cycles round robin over all K arms, so its allocation, cost and reward up to that point are those of uniform random allocation up to the randomness of the order; identical results of the two radii and of the three confidence levels within a horizon without elimination are therefore identities of the design, not findings, and are reported as such. A stratified variant ran independent eliminations within four prespecified context strata, defined by whether all three extended laboratory tests were observed in the feature window and by age 65 years, each at level $\delta/4$, to allow the preferred arm to differ by context; the stratum sizes are reported with the results, and because each stratum needs $K n^*$ decisions of its own at a tighter level, the variant can decide within the horizon only where the whole stream could.

234

The sample complexity of elimination follows from the same radii. Arm i with reward gap $\Delta_i = \mu_{a^*} - \mu_i$ is eliminated once $\beta_i(n) + \beta_{a^*}(n) < \Delta_i$, so the smallest such n , denoted n_i^* , is the number of outcomes per arm a confident retirement needs, and under round robin allocation about $K n_i^*$ decisions plus the feedback delay. We report n_i^* for each arm from the gaps and variances observed on the stream, found exactly by bisection since the sum of the radii decreases in n , together with the number of outcomes per arm the stream actually supplies under round robin allocation, $\lfloor T/K \rfloor$, and the confidence level δ_i^* at which that supply would just suffice, obtained by solving $\beta_i(\lfloor T/K \rfloor) + \beta_{a^*}(\lfloor T/K \rfloor) = \Delta_i$ for δ ; a value of δ_i^* above 1 means that no admissible confidence level permits elimination within the horizon.

242

2.9. Comparator policies

Thompson sampling [28,29] kept a Beta(1,1) posterior per arm, updated with a Bernoulli draw of the observed reward, and selected the arm with the largest posterior sample; the Bernoulli reduction preserves the mean but discards the concentration of the composite reward, so a second variant kept a Gaussian posterior on each arm mean with the running sample variance of its rewards, the same moments the empirical Bernstein radius uses, so that both paradigms see the reward signal at the same fidelity. UCB1 [30] selected the arm maximising $\hat{\mu}_a + \sqrt{2\log m/n_a}$ with m the number of outcomes observed so far; arms without any observed outcome were selected first, as were the arms of the Gaussian variant. Uniform random allocation selected each arm with equal probability. Static deployment applied one arm throughout; the static comparator was the arm with the highest mean composite reward on the calibration half of the validation set under the scenario's own weights, never on the stream. Each single arm was also run statically for reference. Ensemble averaging deployed the mean of the three calibrated probabilities with its own validation threshold, at cost 1.00 because it requires every input. Heuristic escalation ran the cheapest arm first and escalated to the most complete arm when the cheap arm's probability reached its threshold, using the complete arm's prediction and threshold for escalated decisions, with additive cost; it represents the rule based switching that institutions often improvise. The additive convention counts the inputs shared by the two arms twice and is therefore conservative against the heuristic; charging escalated decisions at the cost of the complete arm alone would lower its mean cost by 0.25 times the escalation share.

2.10. Delayed feedback

Decision times followed a Poisson process with rate estimated as the number of eligible stays divided by the 15 years spanned by the database, the institutional accrual rate of eligible stays (0.52 per hour). The stream itself, divided by its three year anchor window, gives a lower rate (0.28 per hour) because patients anchored before 2020 who were admitted during 2020 to 2022 belong to earlier anchor groups by construction and cannot enter the stream; calendar horizons are therefore reported at both rates, the database rate as the estimate of institutional accrual and the stream own rate as a conservative bound. Each outcome became available after a delay drawn uniformly between 6 and 72 hours, independently of the outcome and of the arm. Delay acts on a policy only through the outcomes it has not yet seen when it must decide; a policy whose allocation does not depend on observed outcomes, as round robin allocation does not until an

arm is eliminated, is unaffected by it within such a horizon, so delay matters for the regret minimising policies throughout, for elimination only once a retirement is within reach, and for every policy in the drift scenarios of Section 2.11. A policy could use the outcome of decision s when making decision t only if the availability time of s was no later than the decision time of t . The simulator counts violations of this condition and the pipeline stops if any occurs. Delay scenarios of no delay and of 24 to 168 hours were run as sensitivity analyses.

2.11. Evaluation, replication and sensitivity analyses

The primary metric was the cumulative regret against the best fixed arm in hindsight, the exposure measure that every policy shares. The fixed confidence policies are designed to bound the error of a retirement decision rather than to minimise regret, so for them the quantities of interest are additionally whether and when each arm was eliminated, whether the best arm was ever eliminated, and the implied sample complexity of Section 2.8; the regret comparison between the two families measures what the error guarantee costs in exposure, not which family is the better learner. Secondary metrics were the oracle regret, the mean composite reward, the AUROC and Brier score of the deployed predictions, the false negative rate at the operating thresholds, the mean cost per decision, the share of decisions allocated to each arm, an exploration proxy (the share of decisions in each decile of the stream allocated to arms other than the arm that dominated the final decile), and elimination times. Subgroup consistency was described by three quantities: the signed difference in sensitivity of the deployed predictions between age groups (under 65 minus 65 and over) and between sexes (female minus male), the disparity a patient actually experiences under a mixing policy (the share weighted average of the absolute per arm sensitivity gaps, since each patient is scored by one model and experiences that model's disparity), and the range of sensitivity across race groups with at least 30 deaths. Each arm's sensitivity by age group, sex and race group on the stream is reported with Wilson intervals. Every policy was run in 100 replicates with seeded randomness governing the stream order, decision times, delays and policy draws; results are reported as means with 2.5th and 97.5th percentiles across replicates, and differences against static deployment as paired means with 2,000 resample percentile bootstrap intervals. Sensitivity analyses with 30 replicates each, with the static comparator re chosen on validation under each scenario's own reward, varied the weights ($\lambda_c \in \{0, 0.10, 0.20\}$ by $\lambda_s \in \{0, 0.50, 1.00\}$), the confidence level ($\delta \in \{0.01, 0.05, 0.10\}$), the delay, the utility term of the reward (binary correctness in place of the Brier utility), the false

positive weight ($\lambda_f \in \{0.10, 0.20, 0.40\}$), the cost vector itself (compressed to 0.45, 0.60 and 0.75; expanded to 0.10, 0.55 and 1.00; and equal at 0.60) and the acquisition conditional costing of arm C. Because the validation ordering and the stream ordering of the arms coincide in the primary scenario, so that static deployment of the validation chosen arm carries no regret by construction and no policy can show a benefit over it, four drift scenarios made the validation choice wrong on the stream: Gaussian noise was added to the logit of arm A's probabilities from the first decision or from one third of the stream onward, with standard deviation 1.5 (a moderate loss of discrimination) or 4.0 (a broken model whose discrimination is near chance), while validation data and thresholds were left untouched; each scenario was run under no delay, the primary delay and the long delay. A development window sensitivity analysis repeated the whole pipeline with the development pool restricted to patients anchored in 2017 to 2019, the most recent window before deployment. Replicate percentiles describe the variability induced by the stream order, the feedback timing and the policy randomness for one fixed set of patients; the sampling uncertainty of the reward gaps, arm means and implied sample sizes was quantified by 1,000 patient level bootstrap resamples of the deployment stream (2.5th and 97.5th percentiles). To test whether elimination occurs at the horizon implied by the observed gaps, an extended horizon of 30,000 decisions was simulated by resampling stream stays with replacement (30 replicates), under the primary scenario and under the severe degradation of arm A; this tests the planning formula, not the behaviour of new patients.

2.12. Reproducibility

All analyses are implemented in R (version 4.5.2) as a scripted pipeline with every parameter in a version controlled configuration file. The pipeline writes a run manifest recording the commit hash of the analysis code, package versions, the configuration hash, the elapsed time of each step and, for every MIMIC-IV file read, the SHA256 checksum published by PhysioNet and whether the local file matched it (Supplementary Table S8; the checksums themselves are in the manifest file of the repository). A synthetic data mode with the same schema allows the entire pipeline and its unit tests to run without credentialed data, which the continuous integration workflow does on every commit. The tests cover the reward bounds, the coverage of both confidence radii, the behaviour of the elimination rule on arms with known gaps, the delayed feedback condition and the metric identities. Patient level derived data never leave the local environment.

3. Results

3.1. Cohort and deployment stream

Of 94,458 ICU stays joined to admissions, 95 were excluded for incomplete or inconsistent timestamps or missing sex, none for age (all were adults), 19,598 for an ICU length of stay below 24 hours or above 60 days and 6,262 because they were not the index stay of the admission, leaving 68,503 eligible index stays (Supplementary Table S1). Patients anchored in 2008 to 2019 contributed 60,496 eligible stays, of which 7,680 (12.7 percent) were admitted long enough after the anchor year for their calendar window to reach 2020 and were excluded from the development pool; 72.9 percent of the retained training stays fell within the anchor year itself (Supplementary Table S17). The development set comprised 15,000 training and 5,000 validation stays from 20,000 patients; the deployment stream comprised all 7,383 patients anchored in 2020 to 2022, every one of whose stays occurred in or after 2020, with 982 in hospital deaths (13.3 percent) against 11.7 percent in the training set (Table 2). Stream stays were from slightly younger patients (mean age 64.0 versus 65.0 years), had longer ICU stays (median 68.4 versus 55.3 hours) and a smaller share of patients recorded as White (56.0 versus 67.6 percent) with a larger share recorded as other or unknown (30.4 versus 17.1 percent). The availability of the extended laboratory panel changed between development and deployment: bilirubin was missing for 59.0 percent of training stays and 45.6 percent of stream stays, INR for 16.0 and 11.8 percent, and lactate for 40.6 and 39.7 percent, while core laboratory tests and vital signs changed by less than two percentage points (Supplementary Table S2); all three extended tests were present for 34.9 percent of stream stays. The institutional accrual rate of eligible stays was 0.52 per hour over the 15 years of the database; the stream itself, over its three year anchor window, corresponds to 0.28 per hour, and the 10,906 eligible stays whose calendar window lies entirely in 2020 to 2022 to 0.41 per hour.

3.2. Frozen arms on validation and on the deployment stream

Table 3 reports the three arms on the calibration half of the validation set and on the deployment stream. On the stream, the vital sign only arm A reached an AUROC of 0.855 (0.843 to 0.867), the core laboratory arm B 0.897 (0.888 to 0.906) and the extended laboratory arm C 0.893 (0.883 to 0.902); the equal weight ensemble reached 0.897 (0.888 to 0.906). Brier scores were 0.0878, 0.0777 and 0.0779. Calibration slopes on the stream were 0.98, 1.03 and 1.06 with intercepts of 0.06, 0.01 and 0.03. At the operating thresholds fixed on validation to 80 percent sensitivity (τ_A

= 0.086, $\tau_B = 0.117$, $\tau_C = 0.133$), stream sensitivity was 0.838, 0.855 and 0.864 with false negative rates of 0.162, 0.145 and 0.136. The arms differed in alert burden: arm A alerted on 0.361 of stays (2,665 alerts), arm B on 0.313 (2,308) and arm C on 0.318 (2,346), so the cheapest arm issued 357 more alerts than arm B while detecting 17 fewer deaths. The arms therefore differed modestly in discrimination and substantially in cost and alert rate, which is the regime in which the governance weights decide the preferred model.

3.3. Composite reward and the best fixed arm

Under the primary weights ($\lambda_c = 0.10$, $\lambda_s = 0.50$, no false positive term) the mean composite reward on the calibration half of the validation set was 0.9265 for arm A, 0.9097 for arm B and 0.8843 for arm C, so arm A was the static comparator. On the stream the full information mean rewards were 0.9228, 0.9079 and 0.8831, giving reward gaps of 0.0148 (0.0130 to 0.0168) for arm B and 0.0396 (0.0377 to 0.0417) for arm C relative to arm A (patient level bootstrap intervals), equal to 0.10 and 0.29 standard deviations of the per decision reward (Table 6); arm A had the highest mean reward in every resample. The gaps were made entirely by the cost term: on the rescaled reward, the Brier utility favoured arm B by 0.0063 and arm C by 0.0062 over arm A, the safety term favoured them by 0.0007 and 0.0011, and the cost term counted 0.0219 against arm B and 0.0469 against arm C (Supplementary Table S11, Figure 10), so the preferred arm under these weights is the cheapest arm despite being the least accurate. Arm A also had the highest mean reward in each of the four context strata (complete extended panel by age group, 2,164, 2,645, 1,232 and 1,342 stays; Supplementary Table S4), so under these weights no context dependent preference existed for a stratified policy to exploit. The best fixed arm changed with the weights: with $\lambda_c = 0$ the best fixed arm on the stream was arm B when $\lambda_s = 0$ and arm C when λ_s was 0.50 or 1.00, while any positive cost weight made arm A the best fixed arm (Table 7).

3.4. Primary simulation

Table 4 summarises the 100 replicates and Figure 2 shows the cumulative regret trajectories. Fixed confidence action elimination did not eliminate any arm within the 7,383 decision horizon with either radius (Table 6). Until an arm is eliminated the policy cycles round robin over the three arms, so its allocation is that of uniform random allocation up to the order of the cycle; the cumulative regret of the elimination policies against the best fixed arm, 134.2 (126.3 to 141.0) reward units with either radius, is therefore the same quantity as that of uniform random

allocation (133.4 (124.7 to 141.5)) by construction and not an empirical agreement between two bounds, and the stratified variant behaved the same (133.2 (124.7 to 141.2)). Because the validation ordering and the stream ordering of the arms coincided, static deployment of the validation chosen arm carried no regret by construction, and every adaptive policy paid an exploration cost: Thompson sampling with the Bernoulli reduction attained 33.5 (7.4 to 89.3) and concentrated 76 percent of decisions on arm A, 19 percent on arm B and 4 percent on arm C; the Gaussian posterior variant, which sees the reward at the same fidelity as the elimination radii, attained 13.4 (3.0 to 28.0) with 90 percent of decisions on arm A; UCB1 was intermediate (93.2 (83.5 to 105.3)). Static deployment of the most complete arm C incurred a regret of 292.5 and ensemble averaging 295.9, both driven by cost; heuristic escalation incurred 150.2. Against the per decision oracle, static arm A left 140.3 units of regret, the value that perfect context dependent selection could add at most.

The policies traded cost, discrimination and safety differently (Table 4, Figure 5). Mean cost per decision was 0.250 under static arm A, 0.290 (0.263 to 0.332) under Gaussian Thompson sampling, 0.350 (0.267 to 0.516) under Beta Thompson sampling, 0.512 (0.490 to 0.530) under UCB1 and 0.617 under elimination and random allocation. Deployed prediction AUROC was 0.855 for static arm A, 0.865 (0.855 to 0.885) for Beta Thompson sampling and 0.881 (0.877 to 0.886) for elimination. False negative rates at the operating thresholds were 0.162 for static arm A, 0.148 (0.137 to 0.160) for elimination and 0.157 (0.140 to 0.167) for Beta Thompson sampling; the paired difference in false negative rate between elimination and static deployment was -0.0141 (-0.0153 to -0.0130) (Table 5), reflecting the higher sensitivity of the laboratory informed arms on the stream. The heuristic escalation rule, which lets the cheap arm's negatives stand, raised the false negative rate to 0.230.

Subgroup consistency depended on which arms were deployed rather than on the policy mechanism, and the arms disagreed in direction. The signed sensitivity gap between age groups (under 65 minus 65 and over) was -0.060 for arm A, -0.030 for arm B and 0.049 for arm C on the stream (Supplementary Table S12, Figure 12): arms A and B detected deaths better in older patients and arm C better in younger patients, so the absolute gap of any policy that mixes the arms partly cancels rather than improves. Under static arm A the deployed signed gap was -0.060 , under elimination -0.011 (-0.039 to 0.015) and under Beta Thompson sampling -0.050 (-0.067 to -0.019); the disparity a patient actually experiences, the share weighted absolute gap

of the arm that scored the patient, was 0.060 under static arm A, 0.046 (0.046 to 0.046) under elimination and 0.054 (0.039 to 0.060) under Beta Thompson sampling (Table 4). Signed gaps between sexes (female minus male) were 0.007, -0.001 and 0.002 for the three arms. Across race groups with at least 30 deaths (White, Other or unknown, Black, Hispanic or Latino), the range of sensitivity was 0.153 for arm A, 0.057 for arm B and 0.099 for arm C, and 0.109 (0.060 to 0.183) under elimination against 0.153 under static arm A; these quantities describe the arms and are not evidence about the policies.

Exposure dynamics are shown in Figure 4 and Supplementary Table S3: Beta Thompson sampling allocated 48 percent of decisions to non dominant arms in the first decile of the stream and 9 percent in the last, the Gaussian variant 37 and 2 percent, UCB1 62 and 46 percent, and elimination 67 percent throughout. The confidence bound trajectories of replicate 1 (Figure 3) show the intervals of arms A and B overlapping for the entire stream with both radii.

3.5. Elimination and implied sample complexity

Under round robin allocation the 7,383 decision stream supplies about 2,461 observed outcomes per arm. With the observed gaps and reward variances, the smallest per arm sample at which arm C would be eliminated is 7,508 (6,814 to 8,209) outcomes with the empirical Bernstein radius and 33,541 (30,051 to 37,283) with the Hoeffding radius (patient level bootstrap intervals), 3.05 and 13.6 times what the stream supplies, corresponding to about 22,524 and 100,623 round robin decisions before feedback delay (Table 6, Figure 8A). For arm B, whose gap is 0.0148, the corresponding numbers are 39,194 (31,249 to 50,550) and 277,052 (212,598 to 368,939) outcomes per arm, 15.9 and 112.6 times the supply, or 117,582 and 831,156 decisions. No admissible confidence level would have permitted elimination within the stream: the level at which 2,461 outcomes per arm would just suffice is about 88,000 for arm C and about 50 million for arm B with the empirical Bernstein radius, values far above 1. In the bootstrap, the round robin decisions needed to retire arm C fitted within the stream in none of the 1,000 resamples, so the conclusion does not depend on the particular patients sampled.

In calendar time, at the institutional accrual rate of 0.52 eligible stays per hour, retiring arm C with the variance adaptive bound would take about 4.9 years of admissions and retiring arm B about 25.7 years, while the stream corresponds to 1.6 years; at the stream own rate of 0.28 per hour the same decisions would take 9.2 and 47.8 years and the stream itself spans its three year

anchor window. These horizons are specific to this model portfolio, this database and the declared cost vector; the arithmetic that produces them from the running means and variances of the arm rewards is what transfers.

3.6. Sensitivity analyses

Governance weights. Table 7 and Figure 6 report the reward weight grid. Raising the cost weight widened the gap of arm C and allowed elimination within the horizon: with $\lambda_c = 0.20$ and $\lambda_s = 0$, the empirical Bernstein policy eliminated arm C in 100 percent of replicates and reduced its regret to 325.5 against 410.0 for the Hoeffding policy and 410.3 for random allocation; with $\lambda_c = 0.20$ and $\lambda_s = 0.50$ elimination of arm C occurred in 67 percent of replicates. Arm B was never eliminated in any cell. With $\lambda_c = 0$ the ordering of the arms reversed: arm B ($\lambda_s = 0$) or arm C (λ_s at or above 0.50) became the best fixed arm and ensemble averaging, whose cost no longer counted against it, outperformed every single arm. The Gaussian Thompson sampling variant kept the lowest regret among adaptive policies in all nine cells (7.2 to 15.6).

Cost vector. Because the gaps are made by the cost term, the cost assignment itself was varied (Supplementary Table S14). Compressing the vector to 0.45, 0.60 and 0.75 kept arm A as the best fixed arm but reduced the gaps to 0.0023 (arm B) and 0.0115 (arm C), raising the implied per arm sample for arm C to 60,236 outcomes; expanding it to 0.10, 0.55 and 1.00 widened the gaps to 0.0211 and 0.0490 and lowered the implied sample for arm C to 5,402, with the empirical Bernstein policy eliminating arm C in none of the 30 replicates within the stream; with equal costs of 0.60 the best fixed arm became arm C and the gaps of the other arms were 0.0002 and 0.0073. In no cost vector, and in no other sensitivity scenario, did the stream supply the outcomes a confident retirement of the closest competitor needs (Figure 8B).

Alert burden. The primary reward carries no false positive term. Adding one changes the preferred arm at a weight of $\lambda_f = 0.469$ for the comparison between arms A and B and 1.360 for arms A and C, the weights at which the extra alerts of arm A (0.249 false positive rate against 0.199 and 0.203) offset its cost advantage; an institution that weighted an unnecessary alert at 94 percent of the weight it places on a missed death would prefer arm B. At λ_f of 0.10 and 0.20 arm A remained the best fixed arm on the stream and the validation choice; at 0.40 the validation choice became arm B while arm A remained best on the stream, so static deployment of the validation choice carried a regret of 12.9 against 10.1 for Gaussian Thompson sampling

(Supplementary Table S13, Figure 11). No arm was eliminated within the stream at any false positive weight.

Cost semantics. Charging arm C at arm B's cost whenever the extended panel was absent from the record, which was the case for 65.1 percent of stream stays, lowered its mean cost from 1.00 to 0.739 and its gap to arm A from 0.0396 to 0.0233, raising the implied per arm sample to 17,858 outcomes with the empirical Bernstein radius; arm A remained the best fixed arm and no arm was eliminated (Supplementary Table S15).

Confidence level and delay. The confidence level had no effect within the horizon because no elimination occurred at δ of 0.01, 0.05 or 0.10, which follows from the policy reducing to round robin allocation absent an elimination (Supplementary Table S5); the gaps place the confidence level that would permit elimination within the stream beyond any admissible value (Section 3.5). For the same reason the elimination policies were unaffected by delay. For the regret minimising policies the effect of delay was small relative to the spread across replicates: Beta Thompson sampling's mean regret was 32.6 with no delay, 31.7 with 6 to 72 hour delays and 35.5 with 24 to 168 hour delays (2.5th to 97.5th percentiles 5.6 to 99.7 at the primary delay), and the Gaussian variant's 11.7, 12.2 and 12.4 (Supplementary Table S6); in a small gap regime the arms are nearly interchangeable, so the outcomes a policy has not yet seen cost it little, whereas the drift scenarios below show delay acting in the expected direction once the preferred arm changes.

Utility term. Replacing the Brier utility by binary correctness at the operating threshold changed which arm was preferred but not the elimination result (Supplementary Table S9). Under the accuracy utility the full information mean rewards on the stream were 0.8083, 0.8202 and 0.7937 for arms A, B and C, so arm B was the best fixed arm on the stream and also the static arm chosen on validation under that utility; the gaps of arms A and C were 0.0119 and 0.0265, with implied per arm samples of 161,172 and 32,390 outcomes with the empirical Bernstein radius. No arm was eliminated within the stream (arm C eliminated in none of the 30 replicates), and the heuristic escalation rule, which alerts only when arms A and C both alert, became the best performing strategy under this utility (regret -143.1, below zero) at the price of the highest false negative rate (0.230).

Development window. Restricting the development pool to patients anchored in 2017 to 2019, the most recent window before deployment, left 8,887 development stays (6,665 training and

2,222 validation) against 20,000 in the primary analysis. On the same stream the arms reached AUROCs of 0.853, 0.890 and 0.889 against 0.855, 0.897 and 0.893, arm A remained the best fixed arm with gaps of 0.0165 (arm B) and 0.0383 (arm C), the implied per arm samples were 8,405 outcomes for arm C and 36,952 for arm B, and no arm was eliminated within the stream in any of the 100 replicates (Supplementary Table S16). The twelve year development window therefore did not create the small gaps.

3.7. Drift scenarios: when the validation choice is wrong

Table 8 and Figure 9 report the scenarios in which arm A, the validation chosen arm, degrades on the stream while the validation data and thresholds stay as they were. A moderate degradation (Gaussian noise with standard deviation 1.5 on the logit of its probabilities) reduced its stream AUROC from 0.855 to 0.752, raised its Brier score from 0.0878 to 0.1148 and lowered its sensitivity at the fixed threshold from 0.838 to 0.751; a severe degradation (standard deviation 4.0, a broken model) reduced the AUROC to 0.626 and raised the Brier score to 0.2328. Because the cost advantage of arm A is large on the reward scale, the moderate degradation made arm B the best fixed arm by a gap of only 0.0056, while the severe degradation produced a gap of 0.0846, 2.1 times the primary gap of arm C.

Under the moderate degradation from the first decision, static deployment of the validation choice carried a regret of 41.4 reward units under the primary delay, the cost of deploying the wrong model for the whole stream; Gaussian Thompson sampling carried 16.7, Beta Thompson sampling 32.8 and UCB1 57.8, each below static deployment, while the elimination policies carried 73.8 and retired arm A in none of the 30 replicates, because a gap of 0.0056 is below the resolving power of the stream. Under the severe degradation from the first decision, static deployment carried 624.6, Gaussian Thompson sampling 12.9, Beta Thompson sampling 25.9, UCB1 126.2 and random allocation 270.4; the elimination policies carried 269.4 and still retired arm A in none of the 30 replicates within the stream, because the degraded arm's rewards are also more variable (standard deviation 0.261) and the anytime radii at about 2,461 outcomes per arm still sum to more than the gap: the implied sample is 3,060 outcomes per arm, about 9,180 round robin decisions, with the empirical Bernstein radius. On the extended horizon of 30,000 resampled decisions under the same degradation the empirical Bernstein policy retired arm A in 100 percent of replicates at a median decision of 9,570, in agreement with the implied decisions, while the Hoeffding policy retired it in 100 percent (Table 9). The crossover between the cost of

exploration and the cost of deploying the wrong model therefore lies inside the stream for the regret minimising policies at either magnitude of degradation, and inside about 2.0 years of admissions for a confident retirement of a broken model at $\delta = 0.05$, with the error guarantee intact. Delay acted in the expected direction in this regime: the regret of Gaussian Thompson sampling under the severe degradation was 11.3 with no delay, 12.9 with the primary delay and 14.4 with the long delay, and that of Beta Thompson sampling 21.5, 25.9 and 29.9.

When the degradation began one third of the way through the stream, the best fixed arm over the whole stream was arm A under the moderate and arm B under the severe degradation, and static deployment of arm A carried 0.1 and 381.8; under the severe degradation the elimination policies carried 189.5, Gaussian Thompson sampling 33.8, Beta Thompson sampling 61.4 and random allocation 186.3. A policy that had already concentrated on arm A before the drift had to unlearn that preference, which the posterior sampling policies did within the stream while the elimination policy, whose running mean of arm A still carried its pre drift rewards, could not; a non stationary variant that discounts old evidence would shorten that lag (Section 4.6).

3.8. Extended horizon

When the stream was resampled with replacement to 30,000 decisions from the same stays, the empirical Bernstein policy eliminated arm C in 100 percent of 30 replicates at a median decision index of 22,233, in agreement with the 22,524 decisions implied by the observed gap, while the Hoeffding policy eliminated nothing, as its implied horizon of 100,623 decisions predicts (Table 9, Figure 13). Neither arm A nor arm B was eliminated in any replicate with either radius, so no erroneous retirement occurred, consistent with the guarantee that the best arm is eliminated with probability at most 0.05. Regret at the extended horizon was 464.4 for the empirical Bernstein policy against 542.0 for the Hoeffding policy, 542.8 for random allocation, 277.3 for UCB1, 39.9 (9.8 to 133.7) for Beta Thompson sampling and 16.7 (6.6 to 36.1) for the Gaussian variant. Under the severe degradation of arm A on the same extended horizon, static deployment of arm A carried 2516.4, the empirical Bernstein policy 594.0 with arm A retired in 100 percent of replicates, and Gaussian Thompson sampling 15.7.

4. Discussion

4.1. Principal findings

Treating model selection as a sequential decision problem made four things explicit that static deployment leaves implicit. First, the preferred model is a function of the governance weights and of the cost model, not a property of the models: with any positive cost weight the vital sign only arm was best, without a cost weight the laboratory informed arms were, and a false positive weight of 0.47 on the same scale as the missed death penalty, or equal costs, moved the preference to a laboratory informed arm; under the primary weights the cost term accounts for the whole of the separation between the arms and the predictive and safety terms run the other way. Second, the evidence a confident retirement needs is large relative to a realistic deployment horizon. On 7,383 stays, equivalent to 1.6 years of eligible admissions at the institution's accrual rate, fixed confidence action elimination could not retire any arm at $\delta = 0.05$, because the reward gaps between validated models were 0.015 and 0.040 on the unit scale: the stream supplies about 2,461 outcomes per arm under round robin allocation against the 7,508 the most expensive arm needs with a variance adaptive bound and the 39,194 the intermediate arm needs, shortfalls of 3.1 and 16 fold that no admissible confidence level closes; the implied horizon of about 22,524 decisions for the most expensive arm was confirmed when the stream was resampled to 30,000 decisions. Third, regret minimising policies reduce exposure to the inferior arms sooner but without a stated error guarantee: Thompson sampling with a Gaussian posterior attained the lowest regret of the adaptive policies and converged on the preferred arm in most replicates, yet its regret varied widely across replicates and it kept allocating a share of decisions to the most expensive arm, which a governance body cannot bound in advance. Fourth, when the validation choice was wrong on the stream, the case a governance layer exists for, the regret minimising policies beat static deployment of the wrong model inside the stream at either magnitude of degradation (12.9 against 624.6 for Gaussian Thompson sampling against static deployment of a broken model), whereas a confident retirement at $\delta = 0.05$ needed about 3,060 outcomes per arm even for a broken model, because its rewards are more variable and the anytime radii at a few thousand outcomes are still of the order of the gap; the elimination policy retired it on the extended horizon in 100 percent of replicates, after about 2.0 years of admissions at the institution's accrual rate. The crossover between the cost of exploration and the cost of deploying

the wrong model therefore lies inside the stream for regret minimisation and beyond it for a fixed confidence decision.

4.2. Interpretation for governance design

The results do not argue against fixed confidence governance; they quantify what it costs, and they supply the method for quantifying it elsewhere. A committee that wants a bounded probability of wrongly retiring a model must either accept a long evidence horizon, pool outcomes across sites or services so that the horizon is reached sooner, or set weights that create larger reward gaps. The sensitivity analyses showed the third lever directly: doubling the cost weight or expanding the cost vector produced elimination of the most expensive arm within the horizon in most or all replicates, compressing the cost vector lengthened the horizon, and the Hoeffding radius never decided where the empirical Bernstein radius did. The implied sample complexity is therefore a planning instrument whose arithmetic transfers while its inputs do not: it needs only the running means and variances of the arm rewards, which are available from the first weeks of a deployment, and it returns the number of outcomes a retirement decision will take under the chosen δ and weights before any decision is made. The magnitudes reported here belong to this model portfolio, this database, this cost model and this accrual rate; the calendar horizons double at the stream own accrual rate, and an institution with twice the eligible admissions would halve them.

Three further design points follow. The composite reward compressed the differences between validated models because the Brier loss of calibrated models of similar quality differs by hundredths; the cost term, which is an institutional choice, accounted for the whole of the gaps and more, since the utility term favoured the laboratory informed arms and the safety term was negligible (Supplementary Table S11). Governance weights and the cost model are thus not a detail of the reward but the main determinant of what adaptive deployment can learn, and the alert burden that the primary reward omits would reverse the preference at a weight below the one placed on a missed death. The utility term matters in the same way: scoring the deployed prediction by binary correctness at the operating threshold instead of the Brier loss moved the preferred arm from the cheapest to the intermediate model and made a two stage escalation rule the best performing strategy despite its higher false negative rate. A committee should therefore treat the choice of utility, the cost model and the alert burden weight as governance decisions with the same standing as the weights, and the sensitivity analyses show that none of these

choices shortened the evidence horizon in the primary regime. And the operating thresholds, fixed on validation at a common sensitivity, left the laboratory informed arms more sensitive than the cheapest arm on the stream, so the safety penalty discriminated little between arms here and the cheapest arm's advantage was bought with more alerts and more missed deaths; where thresholds are set differently or models drift to different degrees, the penalty would carry more of the gap.

4.3. Relation to prior work

Best arm identification with action elimination and its sample complexity are classical [8-10,31,32], as are Thompson sampling and UCB1 [28-30]; bandit designs have a long history in clinical trials [33], where the arms are treatments and the ethical constraints differ from those of model selection. Delayed feedback has been analysed for online learning in general [11] and for conversion type outcomes [12]. Contextual bandits that adapt the choice to covariates [34] are the natural extension when the preferred arm differs by context, which it did not under the primary weights here. The present contribution is the specification and evaluation of these components as a deployment governance layer on real ICU data with frozen, externally deployed models, together with the finding that realistic gaps make confident decisions slow and the demonstration of the regime in which the layer pays for itself. This complements work on continual monitoring and updating of clinical AI [35], on reporting standards for prediction models and early stage decision support [36,37], on real world integration of deployed models [38], and on audits that identify shortcut learning and misclassification but do not by themselves control exposure [5,6,39,40]. The framework is the control counterpart of such audits.

4.4. Strengths

The evaluation uses frozen models deployed on a stream whose every stay follows every development stay in calendar time, so that drift between development and deployment is part of the evaluated condition rather than averaged away; every policy sees only delayed feedback and the simulator asserts that no outcome is used early; the static comparator is chosen on validation data under each scenario's own reward, never in hindsight on the stream; the planning quantities are reported with sampling intervals and with the confidence level and sample the stream actually supplies; the dependence of the result on the cost model, the alert burden weight, the utility term and the development window is shown rather than asserted; and every number reported is produced by one run of a public pipeline whose manifest records the analysis commit,

the environment and the checksums of the data files. The sample complexity formula and its confirmation on an extended horizon give the main result a mechanism rather than a description, and the drift scenarios show the mechanism working in the other direction.

4.5. Limitations

The study is a simulation. Rewards were computed from recorded outcomes, so the governance layer could not influence care, and the clinical consequences of deploying one model over another are represented only through the reward terms. The costs of the arms are declared operational footprints, not measured resource use, and no test was ordered by any arm in this retrospective stream, so the cost charged for the extended panel is a tariff on data the health system had already obtained; the acquisition conditional scenario bounds the effect of that mismatch but does not measure cost. Feedback delays were drawn independently of the outcome and of the arm, whereas in practice the outcome becomes known at death or discharge, so the delay equals the remaining hospital stay and can differ systematically between survivors and non survivors. The development pool averages over twelve years of practice and is not what an institution deploying in 2020 would have trained on; the development window sensitivity analysis shows that restricting it to the most recent window leaves the gaps and the horizon unchanged. The anchor year group separates development and deployment at the patient level and, after the restriction applied here, at the stay level, but it does not order stays chronologically within a pool, so drift within 2020 to 2022 is not represented and the stream order is random. The availability of the extended laboratory panel rose between development and deployment, a measurement drift concentrated in the tests that define the most expensive arm, which the evaluation treats as part of the deployed condition rather than correcting for it. The three arms are of similar quality, which is realistic for validated models but limits the gaps; arms of very different quality separate quickly, as the drift scenarios and the unit tests on synthetic arms show. Thompson sampling with the Bernoulli reduction receives a noisier reward signal than the elimination radii; the Gaussian variant removes that handicap and is the fairer comparison. Only one data source and one outcome were studied, and no human factors, workflow or latency aspects were evaluated. The extended horizon resamples the same patients and tests the planning formula, not generalisation to new patients, and the percentile intervals across replicates describe simulation variability for one fixed set of patients, which the patient level bootstrap addresses for the planning quantities while still conditioning on the single site and period studied. Finally, the

subgroup results are descriptive and depend on which arms were deployed; they are not evidence of bias mitigation, the race field of MIMIC-IV is recorded administratively and was other or unknown for a larger share of the stream than of the development set, and the disparity a mixing policy produces is an average over patients who each experienced one model.

4.6. Future work

Shadow deployment, in which the governance layer runs in parallel with care without influencing it, is the next stage; it would measure latency, logging fidelity and the feasibility of pooling outcomes across units to shorten the evidence horizon. Contextual elimination with covariate dependent arm preferences, non stationary variants that discount old evidence so that a drift after a long stable period is detected sooner, and reward terms derived from measured costs, measured alert burden and clinician review of errors are natural extensions. Bayesian fixed confidence procedures such as top two Thompson sampling [41], which keep the sampling behaviour of Thompson sampling while targeting a posterior confidence stopping rule, are a natural competitor to action elimination and would test whether the gap between certainty and speed observed here can be narrowed. Multi site evaluation would test whether the sample complexity planning formula transfers.

5. Conclusion

A fixed confidence multi armed bandit layer can govern selection among validated clinical prediction models with an explicit error bound, delayed feedback and institutional cost and safety weights, but on a realistic deployment stream it reached no retirement decision because the reward gaps between validated models, which under the declared cost model are made by the cost term, are small. The framework turns that fact into a planning quantity: the number of outcomes a confident decision needs under the chosen weights, cost model and confidence level, reported with its uncertainty and against the outcomes a stream supplies. When the validation choice is wrong on the stream, regret minimising policies stop using the wrong model within the stream without a bound, and the fixed confidence layer retires it with its guarantee only after an evidence horizon that even a broken model does not shorten to less than about two years at this accrual rate. Governance bodies choosing between these options are choosing between certainty and speed, and the weights and the cost model they set determine both which model is preferred and how soon the data can say so.

Declarations

CRedit authorship contribution statement. Julian Borges: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing, original draft, Writing, review and editing, Visualization, Project administration.

Funding. This research received no specific grant from any funding agency in the public, commercial or not for profit sectors.

Declaration of competing interest. The author declares no competing financial or personal interests that could have influenced the work reported in this article.

Ethics. The study is a secondary analysis of de identified data from MIMIC-IV v3.1, accessed under the PhysioNet Credentialed Health Data Use Agreement after completion of the required human research training. The collection of MIMIC-IV was approved by the institutional review boards of the Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology with a waiver of informed consent [13,14]; the present analysis did not constitute human subjects research and required no further institutional review board approval or consent.

Data availability. MIMIC-IV v3.1 is available to credentialed researchers from PhysioNet (<https://doi.org/10.13026/kpb9-mt58>). The data use agreement prohibits redistribution of patient level data; no patient level derived data are shared. All aggregate results, tables and figures are included in the code repository.

Code availability. The complete pipeline (cohort construction, feature extraction, model training, governance simulation, tables and figures), its configuration, unit tests, a synthetic data mode and the run manifest of the reported analysis are maintained at <https://github.com/julian-borges-md/multiarm-bandit-clinical-ai-governance> under the Apache 2.0 license. The reported analysis is release 3.1.0 of that repository (analysis code commit 3f02ce510b97), deposited on Zenodo as version 3.1.0 under the concept DOI <https://doi.org/10.5281/zenodo.18248285>, which resolves to the latest deposited version; version 1.0.0 of the deposit corresponds to the superseded preprint and must not be used.

Preprint. An earlier version of this work was posted as a preprint under the title Multi armed bandit based adaptive model selection for clinical AI governance: an infrastructure approach to safer AI use in U.S. healthcare (Research Square, <https://doi.org/10.21203/rs.3.rs-8651972/v1>) [15]. An audit of the archived code found that the quantitative results of that version could not be

reproduced from it (Section 1); its abstract reports lower regret, lower cost and lower false negative rates for adaptive selection than for static deployment on 3,000 admissions, the opposite of the primary comparison reported here. The analysis was rebuilt in full and the present article supersedes the preprint. A version 2 of the preprint carrying the corrected analysis and a correction statement has been posted to the same record (<https://doi.org/10.21203/rs.3.rs-8651972/v2>); version 1 remains visible in the record's history as the platform requires.

References

1. Wong A, Otlés E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med.* 2021;181(8):1065-1070. <https://doi.org/10.1001/jamainternmed.2021.2626>
2. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med.* 2021;385(3):283-286. <https://doi.org/10.1056/NEJMc2104626>
3. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-453. <https://doi.org/10.1126/science.aax2342>
4. Celi LA, Cellini J, Charpignon ML, Dee EC, Dernoncourt F, Eber R, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities: a global review. *PLOS Digit Health.* 2022;1(3):e0000022. <https://doi.org/10.1371/journal.pdig.0000022>
5. Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. *Nat Mach Intell.* 2020;2(11):665-673. <https://doi.org/10.1038/s42256-020-00257-z>
6. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat Mach Intell.* 2021;3(7):610-619. <https://doi.org/10.1038/s42256-021-00338-7>
7. Lattimore T, Szepesvári C. *Bandit algorithms.* Cambridge: Cambridge University Press; 2020. <https://doi.org/10.1017/9781108571401>
8. Even-Dar E, Mannor S, Mansour Y. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J Mach Learn Res.* 2006;7:1079-1105.

787 9. Audibert JY, Bubeck S, Munos R. Best arm identification in multi-armed bandits. In:
788 Proceedings of the 23rd Annual Conference on Learning Theory (COLT 2010). 2010. p. 41-53.

789 10. Kaufmann E, Cappé O, Garivier A. On the complexity of best-arm identification in multi-
790 armed bandit models. J Mach Learn Res. 2016;17(1):1-42.

791 11. Joulani P, György A, Szepesvári C. Online learning under delayed feedback. In: Proceedings
792 of the 30th International Conference on Machine Learning (ICML 2013). PMLR 28(3); 2013. p.
793 1453-1461.

794 12. Vernade C, Cappé O, Perchet V. Stochastic bandit models for delayed conversions. In:
795 Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017). 2017.

796 13. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a
797 freely accessible electronic health record dataset. Sci Data. 2023;10(1):1.
798 <https://doi.org/10.1038/s41597-022-01899-x>

799 14. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, et al. MIMIC-IV (version
800 3.1). PhysioNet; 2024. <https://doi.org/10.13026/kpb9-mt58>

801 15. Borges J. Multi armed bandit based adaptive model selection for clinical AI governance: an
802 infrastructure approach to safer AI use in U.S. healthcare. Research Square. 2026; preprint,
803 version 1. <https://doi.org/10.21203/rs.3.rs-8651972/v1>

804 16. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al.
805 PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex
806 physiologic signals. Circulation. 2000;101(23):e215-e220.
807 <https://doi.org/10.1161/01.CIR.101.23.e215>

808 17. Raasveldt M, Mühleisen H. DuckDB: an embeddable analytical database. In: Proceedings of
809 the 2019 International Conference on Management of Data (SIGMOD 2019). 2019. p. 1981-
810 1984. <https://doi.org/10.1145/3299869.3320212>

811 18. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via
812 coordinate descent. J Stat Softw. 2010;33(1):1-22. <https://doi.org/10.18637/jss.v033.i01>

813 19. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd
814 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD
815 2016). 2016. p. 785-794. <https://doi.org/10.1145/2939672.2939785>

816 20. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5-32.
817 <https://doi.org/10.1023/A:1010933404324>

818 21. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional
819 data in C++ and R. *J Stat Softw.* 2017;77(1):1-17. <https://doi.org/10.18637/jss.v077.i01>

820 22. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized
821 likelihood methods. In: Smola AJ, Bartlett P, Schölkopf B, Schuurmans D, editors. *Advances in*
822 *large margin classifiers*. Cambridge, MA: MIT Press; 2000. p. 61-74.

823 23. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more
824 correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.*
825 1988;44(3):837-845. <https://doi.org/10.2307/2531595>

826 24. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev.*
827 1950;78(1):1-3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)

828 25. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing
829 the performance of prediction models: a framework for traditional and novel measures.
830 *Epidemiology.* 2010;21(1):128-138. <https://doi.org/10.1097/EDE.0b013e3181c30fb2>

831 26. Hoeffding W. Probability inequalities for sums of bounded random variables. *J Am Stat*
832 *Assoc.* 1963;58(301):13-30. <https://doi.org/10.1080/01621459.1963.10500830>

833 27. Maurer A, Pontil M. Empirical Bernstein bounds and sample variance penalization. In:
834 *Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2009)*. 2009.

835 28. Thompson WR. On the likelihood that one unknown probability exceeds another in view of
836 the evidence of two samples. *Biometrika.* 1933;25(3-4):285-294.
837 <https://doi.org/10.1093/biomet/25.3-4.285>

838 29. Agrawal S, Goyal N. Analysis of Thompson sampling for the multi-armed bandit problem.
839 In: *Proceedings of the 25th Annual Conference on Learning Theory (COLT 2012)*. PMLR 23;
840 2012. p. 39.1-39.26.

841 30. Auer P, Cesa-Bianchi N, Fischer P. Finite-time analysis of the multiarmed bandit problem.
842 Mach Learn. 2002;47(2-3):235-256. <https://doi.org/10.1023/A:1013689704352>

843 31. Jamieson K, Nowak R. Best-arm identification algorithms for multi-armed bandits in the
844 fixed confidence setting. In: 2014 48th Annual Conference on Information Sciences and Systems
845 (CISS). IEEE; 2014. p. 1-6. <https://doi.org/10.1109/CISS.2014.6814096>

846 32. Garivier A, Kaufmann E. Optimal best arm identification with fixed confidence. In:
847 Proceedings of the 29th Annual Conference on Learning Theory (COLT 2016). PMLR 49; 2016.
848 p. 998-1027.

849 33. Villar SS, Bowden J, Wason J. Multi-armed bandit models for the optimal design of clinical
850 trials: benefits and challenges. Stat Sci. 2015;30(2):199-215. <https://doi.org/10.1214/14-STS504>

851 34. Li L, Chu W, Langford J, Schapire RE. A contextual-bandit approach to personalized news
852 article recommendation. In: Proceedings of the 19th International Conference on World Wide
853 Web (WWW 2010). 2010. p. 661-670. <https://doi.org/10.1145/1772690.1772758>

854 35. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial
855 intelligence quality improvement: towards continual monitoring and updating of AI algorithms
856 in healthcare. NPJ Digit Med. 2022;5(1):66. <https://doi.org/10.1038/s41746-022-00611-y>

857 36. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al.
858 TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use
859 regression or machine learning methods. BMJ. 2024;385:e078378. [https://doi.org/10.1136/bmj-](https://doi.org/10.1136/bmj-2023-078378)
860 2023-078378

861 37. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting
862 guideline for the early-stage clinical evaluation of decision support systems driven by artificial
863 intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-933. [https://doi.org/10.1038/s41591-022-](https://doi.org/10.1038/s41591-022-01772-9)
864 01772-9

865 38. Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-world integration
866 of a sepsis deep learning technology into routine clinical care: implementation study. JMIR Med
867 Inform. 2020;8(7):e15182. <https://doi.org/10.2196/15182>

- 868 39. Borges JYV. Auditing shortcut learning and misclassification in artificial intelligence-based
869 breast cancer genomic subtyping. *JAMIA Open*. 2026;9(2):ooaf177.
870 <https://doi.org/10.1093/jamiaopen/ooaf177>
- 871 40. Borges J. Clinical artificial intelligence as a sociotechnical system: structural failure modes
872 and governance requirements. *Am J Biomed Sci Res*. 2026;30(1):1026-1032.
873 <https://doi.org/10.34297/AJBSR.2026.30.003888>
- 874 41. Russo D. Simple Bayesian algorithms for best-arm identification. *Oper Res*.
875 2020;68(6):1625-1647. <https://doi.org/10.1287/opre.2019.1911>

Tables

Table 1. Decision arms. Thresholds give 80 percent sensitivity on the calibration half of the validation set; the validation mean reward uses the primary weights ($\lambda_c = 0.10$, $\lambda_s = 0.50$) and determines the static comparator (arm A). The normalised cost is a declared operational footprint, not a measured quantity (Section 2.5).

Arm	Learner	Inputs	Features	Normalised cost	Operating threshold	Validation mean reward
A	Ridge logistic regression	Age, sex, vital signs	34	0.25	0.086	0.9265
B	Gradient boosted trees	A plus core laboratory panel and missing indicators	84	0.60	0.117	0.9097
C	Random forest	B plus lactate, bilirubin, INR and missing indicators	99	1.00	0.133	0.8843

Table 2. Development set and deployment stream. Every development stay has a calendar window ending before 2020 and every stream stay occurred in or after 2020 (Section 2.3). Race is the category recorded on the admission.

Characteristic	Training	Validation	Deployment stream
Stays (one per patient)	15,000	5,000	7,383
Anchor year groups of the patients	2008 to 2019	2008 to 2019	2020 to 2022
Stays within the anchor year, percent	72.9	73.0	93.9
In hospital mortality, percent	11.7	11.2	13.3
Age, mean (SD), years	65.0 (16.8)	65.2 (16.7)	64.0 (16.3)
Female, percent	43.4	42.4	40.2
Race recorded as White, percent	67.6	67.0	56.0
Black, percent	8.5	9.2	6.4
Hispanic or Latino, percent	3.9	3.7	3.9
Asian, percent	2.8	2.5	3.3
Other or unknown, percent	17.1	17.6	30.4
ICU length of stay, median (IQR), hours	55.3 (65.2)	55.9 (64.0)	68.4 (93.0)
Hospital length of stay, median (IQR), days	7.2 (7.7)	7.1 (7.7)	8.7 (10.7)

Table 3. Performance of the frozen arms and of the equal weight ensemble (Ens.) on the calibration half of the validation set (Validation) and on the deployment stream (Stream).

Classification metrics use the operating thresholds fixed on the validation set. Calibration intercept and slope are from the logistic recalibration of the calibrated probabilities; values of 0 and 1 indicate perfect calibration.

Set	Model	n	Deaths	AUROC (95% CI)	AUPRC	Brier	Cal. intercept / slope	Sens.	Spec.	PPV	Alert rate (alerts)
Validation	A	2,500	285	0.834 (0.810 to 0.858)	0.433	0.0812	0.00 / 1.00	0.800	0.690	0.249	0.366 (915)
Validation	B	2,500	285	0.880 (0.861 to 0.899)	0.532	0.0730	-0.00 / 1.00	0.800	0.797	0.336	0.271 (678)
Validation	C	2,500	285	0.872 (0.852 to 0.892)	0.526	0.0737	0.00 / 1.00	0.800	0.776	0.315	0.290 (724)
Validation	Ens.	2,500	285	0.882 (0.863 to 0.900)	0.532	0.0728	0.32 / 1.22	0.800	0.802	0.342	0.267 (667)
Stream	A	7,383	982	0.855 (0.843 to 0.867)	0.511	0.0878	0.06 / 0.98	0.838	0.712	0.309	0.361 (2,665)
Stream	B	7,383	982	0.897 (0.888 to 0.906)	0.598	0.0777	0.01 / 1.03	0.855	0.771	0.364	0.313 (2,308)
Stream	C	7,383	982	0.893 (0.883 to 0.902)	0.596	0.0779	0.03 / 1.06	0.864	0.766	0.361	0.318 (2,346)
Stream	Ens.	7,383	982	0.897 (0.888 to 0.906)	0.604	0.0775	0.27 / 1.21	0.846	0.786	0.378	0.298 (2,201)

Table 4. Primary simulation on the deployment stream (7,383 decisions, 100 replicates): mean and 2.5th to 97.5th percentiles across replicates. Primary weights $\lambda_c = 0.10$, $\lambda_s = 0.50$, $\delta = 0.05$, feedback delay 6 to 72 hours. FCAE, fixed confidence action elimination. Static arm A is the validation chosen comparator, whose regret is zero by construction because the validation and stream orderings coincide. The signed gap is the sensitivity of the deployed predictions in patients under 65 minus that in patients 65 and over; the experienced gap is the share weighted absolute gap of the arm that scored each patient (not defined for strategies that are not mixtures of the arms).

Policy	Regret		Mean		False negative rate	Deployed AUROC	Decisions on A / B / C, percent	Signed sensitivity gap by age	Experienced sensitivity gap by age
	vs best fixed arm	Regret vs oracle	reward	cost					

Policy	Regret vs best fixed arm	Regret vs oracle	Mean reward	Mean cost	False negative rate	Deployed AUROC	Decisions on A / B / C, percent	Signed sensitivity gap by age	Experienced sensitivity gap by age
FCAE, Hoeffding	134.2 (126.3 to 141.0)	274.5 (266.7 to 281.3)	0.9046 (0.9037 to 0.9057)	0.617 (0.617 to 0.617)	0.148 (0.137 to 0.160)	0.881 (0.877 to 0.886)	33.3 / 33.3 / 33.3	−0.011 (−0.039 to 0.015)	0.046 (0.046 to 0.046)
FCAE, empirical Bernstein	134.2 (126.3 to 141.0)	274.5 (266.7 to 281.3)	0.9046 (0.9037 to 0.9057)	0.617 (0.617 to 0.617)	0.148 (0.137 to 0.160)	0.881 (0.877 to 0.886)	33.3 / 33.3 / 33.3	−0.011 (−0.039 to 0.015)	0.046 (0.046 to 0.046)
FCAE, stratified	133.2 (124.7 to 141.2)	273.5 (265.1 to 281.5)	0.9047 (0.9036 to 0.9059)	0.616 (0.616 to 0.616)	0.146 (0.133 to 0.161)	0.881 (0.876 to 0.886)	33.4 / 33.3 / 33.3	−0.015 (−0.038 to 0.012)	0.046 (0.046 to 0.046)
Thompson sampling, Beta	33.5 (7.4 to 89.3)	173.9 (147.8 to 229.7)	0.9182 (0.9107 to 0.9218)	0.350 (0.267 to 0.516)	0.157 (0.140 to 0.167)	0.865 (0.855 to 0.885)	76.4 / 19.3 / 4.3	−0.050 (−0.067 to −0.019)	0.054 (0.039 to 0.060)
Thompson sampling, Gaussian	13.4 (3.0 to 28.0)	153.8 (143.3 to 168.4)	0.9209 (0.9190 to 0.9224)	0.290 (0.263 to 0.332)	0.160 (0.154 to 0.166)	0.859 (0.856 to 0.865)	90.1 / 8.5 / 1.4	−0.056 (−0.068 to −0.040)	0.057 (0.054 to 0.059)
UCB1	93.2 (83.5 to 105.3)	233.6 (223.8 to 245.6)	0.9101 (0.9085 to 0.9115)	0.512 (0.490 to 0.530)	0.151 (0.140 to 0.163)	0.876 (0.872 to 0.880)	47.8 / 32.4 / 19.8	−0.028 (−0.047 to −0.005)	0.048 (0.047 to 0.049)
Uniform random	133.4 (124.7 to 141.5)	273.7 (265.0 to 281.8)	0.9047 (0.9036 to 0.9059)	0.616 (0.608 to 0.624)	0.148 (0.134 to 0.162)	0.881 (0.877 to 0.885)	33.4 / 33.3 / 33.3	−0.016 (−0.044 to 0.012)	0.046 (0.046 to 0.047)
Static, validation best (arm A)	0.0 (0.0 to 0.0)	140.3 (140.3 to 140.3)	0.9228 (0.9228 to 0.9228)	0.250 (0.250 to 0.250)	0.162 (0.162 to 0.162)	0.855 (0.855 to 0.855)	100.0 / 0.0 / 0.0	−0.060 (−0.060 to −0.060)	0.060 (0.060 to 0.060)
Static, arm B	109.6 (109.6 to to)	250.0 (250.0 to to)	0.9079 (0.9079 to to)	0.600 (0.600 to to)	0.145 (0.145 to 0.145)	0.897 (0.897 to 0.897)	0.0 / 100.0 / 0.0	−0.030 (−0.030 to −0.030)	0.030 (0.030 to 0.030)

Policy	Regret vs best fixed arm	Regret vs oracle	Mean reward	Mean cost	False negative rate	Deployed AUROC	Decisions on A / B / C, percent	Signed sensitivity gap by age	Experienced sensitivity gap by age
	109.6)	250.0)	0.9079)	0.600)					
	292.5 (292.5 to 292.5)	432.8 (432.8 to 432.8)	0.8831 (0.8831 to 0.8831)	1.000 (1.000 to 1.000)	0.136 (0.136 to 0.136)	0.893 (0.893 to 0.893)	0.0 / 0.0 / 100.0	0.049 (0.049 to 0.049)	0.049 (0.049 to 0.049)
Static, arm C									
Ensemble average	295.9 (295.9 to 295.9)	436.2 (436.2 to 436.2)	0.8827 (0.8827 to 0.8827)	1.000 (1.000 to 1.000)	0.154 (0.154 to 0.154)	0.897 (0.897 to 0.897)	not applicable	0.002 (0.002 to 0.002)	not applicable
Heuristic escalation	150.2 (150.2 to 150.2)	290.5 (290.5 to 290.5)	0.9024 (0.9024 to 0.9024)	0.611 (0.611 to 0.611)	0.230 (0.230 to 0.230)	0.872 (0.872 to 0.872)	not applicable	-0.018 (-0.018 to -0.018)	not applicable

Table 5. Paired differences against static deployment of the validation chosen arm A across replicates (percentile bootstrap, 2,000 resamples). Intervals of width zero arise for deterministic strategies whose outcome does not depend on the stream order. The experienced sensitivity gap is the share weighted absolute per arm gap between age groups.

Policy	Metric	Policy mean	Static arm A mean	Paired difference (95% bootstrap interval)
FCAE, Hoeffding	Cumulative regret	134.2	0.0	134.2 (133.4 to 134.9)
FCAE, Hoeffding	Mean cost	0.617	0.250	0.367 (0.367 to 0.367)
FCAE, Hoeffding	False negative rate	0.1478	0.1619	-0.0141 (-0.0153 to -0.0130)
FCAE, Hoeffding	Deployed AUROC	0.8812	0.8550	0.0262 (0.0258 to 0.0266)
FCAE, Hoeffding	Experienced sensitivity gap, age	0.0465	0.0601	-0.0137 (-0.0137 to -0.0137)
FCAE, empirical Bernstein	Cumulative regret	134.2	0.0	134.2 (133.4 to 134.9)
FCAE, empirical Bernstein	Mean cost	0.617	0.250	0.367 (0.367 to 0.367)
FCAE, empirical Bernstein	False negative rate	0.1478	0.1619	-0.0141 (-0.0153 to -0.0130)
FCAE, empirical	Deployed AUROC	0.8812	0.8550	0.0262 (0.0258 to 0.0266)

Policy	Metric	Policy mean	Static arm A mean	Paired difference (95% bootstrap interval)
Bernstein				
FCAE, empirical Bernstein	Experienced sensitivity gap, age	0.0465	0.0601	-0.0137 (-0.0137 to -0.0137)
FCAE, stratified	Cumulative regret	133.2	0.0	133.2 (132.3 to 134.1)
FCAE, stratified	Mean cost	0.616	0.250	0.366 (0.366 to 0.366)
FCAE, stratified	False negative rate	0.1464	0.1619	-0.0156 (-0.0169 to -0.0142)
FCAE, stratified	Deployed AUROC	0.8813	0.8550	0.0263 (0.0259 to 0.0268)
FCAE, stratified	Experienced sensitivity gap, age	0.0465	0.0601	-0.0137 (-0.0137 to -0.0137)
Thompson sampling, Beta	Cumulative regret	33.5	0.0	33.5 (29.0 to 38.2)
Thompson sampling, Beta	Mean cost	0.350	0.250	0.100 (0.086 to 0.114)
Thompson sampling, Beta	False negative rate	0.1569	0.1619	-0.0050 (-0.0064 to -0.0037)
Thompson sampling, Beta	Deployed AUROC	0.8650	0.8550	0.0100 (0.0085 to 0.0116)
Thompson sampling, Beta	Experienced sensitivity gap, age	0.0539	0.0601	-0.0062 (-0.0073 to -0.0052)
Thompson sampling, Gaussian	Cumulative regret	13.4	0.0	13.4 (11.8 to 15.6)
Thompson sampling, Gaussian	Mean cost	0.290	0.250	0.040 (0.035 to 0.047)
Thompson sampling, Gaussian	False negative rate	0.1603	0.1619	-0.0016 (-0.0022 to -0.0010)
Thompson sampling, Gaussian	Deployed AUROC	0.8593	0.8550	0.0043 (0.0037 to 0.0051)
Thompson sampling, Gaussian	Experienced sensitivity gap, age	0.0574	0.0601	-0.0027 (-0.0032 to -0.0023)
UCB1	Cumulative regret	93.2	0.0	93.2 (92.1 to 94.4)
UCB1	Mean cost	0.512	0.250	0.262 (0.260 to 0.264)
UCB1	False negative rate	0.1515	0.1619	-0.0104 (-0.0115 to -0.0093)
UCB1	Deployed AUROC	0.8761	0.8550	0.0211 (0.0207 to 0.0215)
UCB1	Experienced sensitivity gap, age	0.0483	0.0601	-0.0119 (-0.0120 to -0.0118)
Uniform random	Cumulative regret	133.4	0.0	133.4 (132.5 to 134.3)
Uniform random	Mean cost	0.616	0.250	0.366 (0.365 to 0.367)

Policy	Metric	Policy mean	Static arm A mean	Paired difference (95% bootstrap interval)
Uniform random	False negative rate	0.1476	0.1619	−0.0143 (−0.0157 to −0.0129)
Uniform random	Deployed AUROC	0.8812	0.8550	0.0262 (0.0259 to 0.0266)
Uniform random	Experienced sensitivity gap, age	0.0465	0.0601	−0.0137 (−0.0137 to −0.0136)
Static, arm B	Cumulative regret	109.6	0.0	109.6 (109.6 to 109.6)
Static, arm B	Mean cost	0.600	0.250	0.350 (0.350 to 0.350)
Static, arm B	False negative rate	0.1446	0.1619	−0.0173 (−0.0173 to −0.0173)
Static, arm B	Deployed AUROC	0.8966	0.8550	0.0416 (0.0416 to 0.0416)
Static, arm B	Experienced sensitivity gap, age	0.0304	0.0601	−0.0297 (−0.0297 to −0.0297)
Static, arm C	Cumulative regret	292.5	0.0	292.5 (292.5 to 292.5)
Static, arm C	Mean cost	1.000	0.250	0.750 (0.750 to 0.750)
Static, arm C	False negative rate	0.1365	0.1619	−0.0255 (−0.0255 to −0.0255)
Static, arm C	Deployed AUROC	0.8926	0.8550	0.0376 (0.0376 to 0.0376)
Static, arm C	Experienced sensitivity gap, age	0.0488	0.0601	−0.0113 (−0.0113 to −0.0113)
Ensemble average	Cumulative regret	295.9	0.0	295.9 (295.9 to 295.9)
Ensemble average	Mean cost	1.000	0.250	0.750 (0.750 to 0.750)
Ensemble average	False negative rate	0.1538	0.1619	−0.0081 (−0.0081 to −0.0081)
Ensemble average	Deployed AUROC	0.8974	0.8550	0.0424 (0.0424 to 0.0424)
Heuristic escalation	Cumulative regret	150.2	0.0	150.2 (150.2 to 150.2)
Heuristic escalation	Mean cost	0.611	0.250	0.361 (0.361 to 0.361)
Heuristic escalation	False negative rate	0.2301	0.1619	0.0682 (0.0682 to 0.0682)
Heuristic escalation	Deployed AUROC	0.8718	0.8550	0.0168 (0.0168 to 0.0168)

Table 6. Observed reward gaps on the deployment stream under the primary weights and cost vector (0.25, 0.60, 1.00) and the implied sample complexity of a confident elimination at $\delta = 0.05$, found exactly by bisection. Intervals are 2.5th to 97.5th percentiles over 1,000 patient level bootstrap resamples of the stream (Supplementary Table S10). The stream provides 7,383 decisions, about 2,461 outcomes per arm under round robin allocation; the shortfall factor is the ratio of outcomes needed to outcomes available, and the required δ is the confidence level at which the available outcomes would just suffice. No arm was eliminated within the stream in any of the 100 replicates with either radius.

Arm	Mean reward (95% bootstrap interval)	SD of reward	Gap to best arm (95% bootstrap interval)	Gap in SD units	Outcomes per arm available	Outcomes per arm needed, Hoeffding (95% interval)	Outcomes per arm needed, Bernstein (95% interval)	Shortfall factor, Bernstein	Decisions, Bernstein	Required δ , Bernstein
A	0.9228 (0.9192 to 0.9263)	0.160	best arm	0.00	2,461	best arm	best arm	best arm	best arm	best arm
B	0.9079 (0.9045 to 0.9111)	0.144	0.0148 (0.0130 to 0.0168)	0.10	2,461	277,052 (212,598 to 368,939)	39,194 (31,249 to 50,550)	15.93	117,582	above 1 (none admissible)
C	0.8831 (0.8800 to 0.8864)	0.139	0.0396 (0.0377 to 0.0417)	0.29	2,461	33,541 (30,051 to 37,283)	7,508 (6,814 to 8,209)	3.05	22,524	above 1 (none admissible)

Table 7. Sensitivity of mean cumulative regret (against the best fixed arm on the stream) to the governance weights, 30 replicates per cell. Best (stream): best fixed arm on the stream; Static (val.): static arm chosen on validation under the same weights; FCAE Hoeff. and FCAE Bern.: fixed confidence action elimination with the Hoeffding and empirical Bernstein radii; TS Beta and TS Gauss.: Thompson sampling with the Beta and the Gaussian posterior; Arm C retired: percent of replicates in which FCAE Bern. eliminated arm C within the stream. Negative values occur when a composite strategy that is not itself an arm outperforms every fixed arm.

		FCA											Arm C
		Best	Stati	E	FCA	TS	TS						
		(stream	c	Hoeff	E	Bet	Gauss	UCB	Rando	Stati	Ensembl	Heuristi	retired
λ_c	λ_s	)	(val.)	.	Bern.	a	.	l	m	c	e	c	, %
0.0	0.0	B	B	25.0	25.0	12.4	9.2	20.7	24.9	0.0	-1.8	14.8	0
0	0												
0.0	0.5	C	B	18.9	18.9	9.7	8.7	16.7	18.8	1.8	3.6	39.7	0
0	0												
0.0	1.0	C	B	16.5	16.5	11.0	8.7	15.2	17.4	3.4	7.0	52.8	0
0	0												
0.1	0.0	A	A	201.1	201.1	28.2	12.4	116.9	200.3	0.0	434.0	188.0	0

λ_c	λ_s	FCA											Arm C retired, %
		Best (stream)	Static (val.)	E Hoeffding	FCA E Bern.	TS Beta	TS Gaussian	UCB1	Random	Static	Ensemble	Heuristic	
0	0												
0.1	0.5	A	A	133.4	133.4	31.7	12.2	92.9	133.1	0.0	295.9	150.2	0
0	0												
0.1	1.0	A	A	97.9	97.9	31.6	15.6	74.0	98.6	0.0	223.5	130.4	0
0	0												
0.2	0.0	A	A	410.0	325.5	20.7	7.2	142.2	410.3	0.0	859.3	394.4	100
0	0												
0.2	0.5	A	A	284.8	277.0	37.1	7.4	136.4	283.5	0.0	604.2	298.1	67
0	0												
0.2	1.0	A	A	216.5	216.5	19.0	8.1	123.8	216.5	0.0	465.1	245.6	0
0	0												

Table 8. Drift scenarios in which the validation chosen arm A becomes inferior on the stream (Gaussian noise on the logit of its probabilities with standard deviation 1.5, moderate, or 4.0, severe; validation data and thresholds untouched), 30 replicates per cell. Static deployment of arm A now carries the cost of deploying the wrong model; the adaptive policies carry the cost of discovering it.

Degradation of arm A	Delay	Policy	Best fixed arm	Cumulative regret, mean (2.5th to 97.5th)	Mean cost	False negative rate	Arm A retired, %	Median retirement decision
Moderate, from the first decision	6 to 72 h	FCAE, Hoeffding	B	73.8 (63.8 to 81.4)	0.617	0.171	0	none
Moderate, from the first decision	6 to 72 h	FCAE, empirical Bernstein	B	73.8 (63.8 to 81.4)	0.617	0.171	0	none
Moderate, from the first decision	6 to 72 h	Thompson sampling, Beta	B	32.8 (8.5 to 74.3)	0.526	0.177	not applicable	not applicable
Moderate, from the first decision	6 to 72 h	Thompson sampling, Gaussian	B	16.7 (6.2 to 33.9)	0.520	0.172	not applicable	not applicable
Moderate, from	6 to 72	UCB1	B	57.8 (44.1 to	0.568	0.179	not	not

Degradation of arm A	Delay	Policy	Best fixed arm	Cumulative regret, mean (2.5th to 97.5th)	Mean cost	False negative rate	Arm A retired, %	Median retirement decision
the first decision	h			68.8)			applicable	applicable
Moderate, from the first decision	6 to 72 h	Uniform random	B	74.9 (64.0 to 88.0)	0.616	0.177	not applicable	not applicable
Moderate, from the first decision	6 to 72 h	Static, validation best	B	41.4 (27.0 to 53.0)	0.250	0.245	not applicable	not applicable
Moderate, from the first decision	6 to 72 h	Ensemble average	B	196.9 (192.3 to 201.0)	1.000	0.141	not applicable	not applicable
Moderate, from the first decision	6 to 72 h	Heuristic escalation	B	118.7 (106.2 to 130.7)	0.686	0.309	not applicable	not applicable
Moderate, from one third	6 to 72 h	FCAE, Hoeffding	A	67.1 (57.8 to 76.2)	0.617	0.166	0	none
Moderate, from one third	6 to 72 h	FCAE, empirical Bernstein	A	67.1 (57.8 to 76.2)	0.617	0.166	0	none
Moderate, from one third	6 to 72 h	Thompson sampling, Beta	A	14.1 (−0.9 to 32.8)	0.387	0.194	not applicable	not applicable
Moderate, from one third	6 to 72 h	Thompson sampling, Gaussian	A	7.2 (−4.0 to 21.8)	0.335	0.204	not applicable	not applicable
Moderate, from one third	6 to 72 h	UCB1	A	45.0 (29.4 to 63.6)	0.550	0.171	not applicable	not applicable
Moderate, from one third	6 to 72 h	Uniform random	A	64.8 (54.7 to 79.2)	0.616	0.165	not applicable	not applicable
Moderate, from one third	6 to 72 h	Static, validation best	A	0.1 (0.0 to 0.9)	0.250	0.219	not applicable	not applicable
Moderate, from one third	6 to 72 h	Ensemble average	A	201.9 (193.7 to 211.1)	1.000	0.145	not applicable	not applicable
Moderate, from one third	6 to 72 h	Heuristic escalation	A	101.7 (91.0 to 112.2)	0.662	0.284	not applicable	not applicable

Degradation of arm A	Delay	Policy	Best fixed arm	Cumulative regret, mean (2.5th to 97.5th)	Mean cost	False negative rate	Arm A retired, %	Median retirement decision
Severe, from the first decision	none	FCAE, Hoeffding	B	269.4 (249.9 to 281.9)	0.617	0.212	0	none
Severe, from the first decision	none	FCAE, empirical Bernstein	B	269.4 (249.9 to 281.9)	0.617	0.212	0	none
Severe, from the first decision	none	Thompson sampling, Beta	B	21.5 (7.0 to 39.6)	0.624	0.147	not applicable	not applicable
Severe, from the first decision	none	Thompson sampling, Gaussian	B	11.3 (6.2 to 16.2)	0.607	0.146	not applicable	not applicable
Severe, from the first decision	none	UCB1	B	125.3 (112.6 to 140.8)	0.684	0.165	not applicable	not applicable
Severe, from the first decision	none	Uniform random	B	271.8 (249.5 to 299.8)	0.615	0.217	not applicable	not applicable
Severe, from the first decision	none	Static, validation best	B	624.6 (593.3 to 657.2)	0.250	0.365	not applicable	not applicable
Severe, from the first decision	none	Ensemble average	B	254.5 (244.9 to 261.8)	1.000	0.125	not applicable	not applicable
Severe, from the first decision	none	Heuristic escalation	B	217.8 (201.8 to 228.5)	0.729	0.432	not applicable	not applicable
Severe, from the first decision	6 to 72 h	FCAE, Hoeffding	B	269.4 (249.9 to 281.9)	0.617	0.212	0	none
Severe, from the first decision	6 to 72 h	FCAE, empirical Bernstein	B	269.4 (249.9 to 281.9)	0.617	0.212	0	none
Severe, from the first decision	6 to 72 h	Thompson sampling,	B	25.9 (10.5 to 56.1)	0.632	0.147	not applicable	not applicable

Degradation of arm A	Delay	Policy	Best fixed arm	Cumulative regret, mean (2.5th to 97.5th)	Mean cost	False negative rate	Arm A retired, %	Median retirement decision
decision		Beta						
Severe, from the first decision	6 to 72 h	Thompson sampling, Gaussian	B	12.9 (8.4 to 18.0)	0.609	0.146	not applicable	not applicable
Severe, from the first decision	6 to 72 h	UCB1	B	126.2 (116.6 to 136.6)	0.683	0.168	not applicable	not applicable
Severe, from the first decision	6 to 72 h	Uniform random	B	270.4 (249.2 to 286.0)	0.616	0.218	not applicable	not applicable
Severe, from the first decision	6 to 72 h	Static, validation best	B	624.6 (593.3 to 657.2)	0.250	0.365	not applicable	not applicable
Severe, from the first decision	6 to 72 h	Ensemble average	B	254.5 (244.9 to 261.8)	1.000	0.125	not applicable	not applicable
Severe, from the first decision	6 to 72 h	Heuristic escalation	B	217.8 (201.8 to 228.5)	0.729	0.432	not applicable	not applicable
Severe, from the first decision	24 to 168 h	FCAE, Hoeffding	B	269.4 (249.9 to 281.9)	0.617	0.212	0	none
Severe, from the first decision	24 to 168 h	FCAE, empirical Bernstein	B	269.4 (249.9 to 281.9)	0.617	0.212	0	none
Severe, from the first decision	24 to 168 h	Thompson sampling, Beta	B	29.9 (9.7 to 91.1)	0.639	0.148	not applicable	not applicable
Severe, from the first decision	24 to 168 h	Thompson sampling, Gaussian	B	14.4 (9.2 to 22.8)	0.610	0.147	not applicable	not applicable
Severe, from the first decision	24 to 168 h	UCB1	B	124.6 (113.7 to 134.6)	0.682	0.167	not applicable	not applicable
Severe, from	24 to	Uniform	B	270.5 (251.0 to	0.616	0.216	not	not

Degradation of arm A	Delay	Policy	Best fixed arm	Cumulative regret, mean (2.5th to 97.5th)	Mean cost	False negative rate	Arm A retired, %	Median retirement decision
the first decision	168 h	random		285.1)			applicable	applicable
Severe, from the first decision	24 to 168 h	Static, validation best	B	624.6 (593.3 to 657.2)	0.250	0.365	not applicable	not applicable
Severe, from the first decision	24 to 168 h	Ensemble average	B	254.5 (244.9 to 261.8)	1.000	0.125	not applicable	not applicable
Severe, from the first decision	24 to 168 h	Heuristic escalation	B	217.8 (201.8 to 228.5)	0.729	0.432	not applicable	not applicable
Severe, from one third	6 to 72 h	FCAE, Hoeffding	B	189.5 (179.0 to 204.9)	0.617	0.190	0	none
Severe, from one third	6 to 72 h	FCAE, empirical Bernstein	B	189.5 (179.0 to 204.9)	0.617	0.190	0	none
Severe, from one third	6 to 72 h	Thompson sampling, Beta	B	61.4 (22.9 to 125.5)	0.517	0.172	not applicable	not applicable
Severe, from one third	6 to 72 h	Thompson sampling, Gaussian	B	33.8 (16.5 to 64.3)	0.488	0.170	not applicable	not applicable
Severe, from one third	6 to 72 h	UCB1	B	99.9 (90.7 to 107.7)	0.619	0.167	not applicable	not applicable
Severe, from one third	6 to 72 h	Uniform random	B	186.3 (165.5 to 203.5)	0.616	0.190	not applicable	not applicable
Severe, from one third	6 to 72 h	Static, validation best	B	381.8 (362.8 to 401.2)	0.250	0.295	not applicable	not applicable
Severe, from one third	6 to 72 h	Ensemble average	B	232.0 (226.8 to 238.5)	1.000	0.135	not applicable	not applicable
Severe, from one third	6 to 72 h	Heuristic escalation	B	156.2 (148.6 to 169.6)	0.690	0.361	not applicable	not applicable

Table 9. Extended horizon of 30,000 decisions obtained by resampling stream stays with replacement (30 replicates), a check of the implied sample complexity rather than an evaluation on new patients, under the primary scenario (elimination of arm C) and under the severe

degradation of arm A from the first decision (elimination of arm A). In the primary scenario arms A and B were never eliminated.

Scenario	Policy	Cumulative regret, mean (2.5th to 97.5th percentile)	Mean reward	Mean cost	False negative rate	Arm eliminated, percent of replicates	Median elimination decision
Primary (no drift)	FCAE, Hoeffding	542.0 (518.1 to 570.3)	0.9043	0.617	0.147	C: 0	none
Primary (no drift)	FCAE, empirical Bernstein	464.4 (419.4 to 519.2)	0.9069	0.571	0.148	C: 100	22,233
Primary (no drift)	Thompson sampling, Beta	39.9 (9.8 to 133.7)	0.9211	0.279	0.161	C: 0	none
Primary (no drift)	Thompson sampling, Gaussian	16.7 (6.6 to 36.1)	0.9218	0.262	0.162	C: 0	none
Primary (no drift)	UCB1	277.3 (244.8 to 307.0)	0.9132	0.446	0.153	C: 0	none
Primary (no drift)	Uniform random	542.8 (514.4 to 576.9)	0.9043	0.617	0.148	C: 0	none
Primary (no drift)	Static, validation best (arm A)	0.0 (0.0 to 0.0)	0.9224	0.250	0.162	C: 0	none
Severe degradation of arm A	FCAE, Hoeffding	838.1 (754.4 to 936.3)	0.8797	0.681	0.189	A: 100	19,784
Severe degradation of arm A	FCAE, empirical Bernstein	594.0 (546.0 to 636.4)	0.8879	0.742	0.163	A: 100	9,569
Severe degradation of arm A	Thompson sampling, Beta	33.9 (13.0 to 76.7)	0.9065	0.609	0.146	A: 0	none
Severe degradation of arm A	Thompson sampling, Gaussian	15.7 (11.3 to 21.3)	0.9071	0.603	0.145	A: 0	none
Severe degradation of arm A	UCB1	294.6 (273.1 to 319.7)	0.8978	0.671	0.154	A: 0	none

Scenario	Policy	Cumulative regret, mean (2.5th to 97.5th percentile)	Mean reward	Mean cost	False negative rate	Arm eliminated, percent of replicates	Median elimination decision
Severe degradation of arm A	Uniform random	1087.7 (1042.4 to 1129.4)	0.8714	0.616	0.216	A: 0	none
Severe degradation of arm A	Static, validation best (arm A)	2516.4 (2459.0 to 2568.0)	0.8238	0.250	0.365	A: 0	none

Figure legends

Figure 1. Cohort flow: eligibility steps from 94,458 ICU stays to 68,503 eligible index stays (grey), and the analysis pools and sets (blue): the 60,496 stays of patients anchored 2008 to 2019, the 52,816 of them placed before 2020, the 20,000 stay development set, the 8,007 stays of patients anchored 2020 to 2022 and the 7,383 stay deployment stream.

Figure 2. Cumulative regret against the best fixed arm over the deployment stream for each policy under the primary weights, including static deployment of arms B and C: mean across 100 replicates with 2.5th to 97.5th percentile bands.

Figure 3. Estimated composite reward with confidence bounds for each arm in replicate 1 (shown because it is the first seeded replicate, not selected) under fixed confidence action elimination with the Hoeffding radius (top) and the empirical Bernstein radius (bottom). The intervals of arms A and B overlap throughout; no elimination occurs.

Figure 4. Exploration proxy by decile of the stream: share of decisions allocated to arms other than the arm dominating the final decile (mean and 2.5th to 97.5th percentiles across replicates).

Figure 5. Running mean normalised cost per decision under each policy (mean across replicates).

Figure 6. Difference in cumulative regret between fixed confidence action elimination with the empirical Bernstein radius and static deployment of the validation chosen arm, across the grid of cost and safety weights (30 replicates per cell).

Figure 7. Mean composite reward of each arm on the deployment stream by context stratum (complete extended laboratory panel by age group) under the primary weights.

Figure 8. The evidence horizon. (A) Outcomes per arm needed to retire arm B and arm C at $\delta = 0.05$ with the Hoeffding and the empirical Bernstein radii, with patient level bootstrap intervals (1,000 resamples), against the 2,461 outcomes per arm the stream supplies under round robin allocation (dashed line) and the 10,000 per arm of the extended horizon (dotted line). (B) The same quantity with the empirical Bernstein radius under each sensitivity scenario (utility term, cost vectors, acquisition conditional costing and alert burden weights); the arm shown is the arm that would be retired, which changes with the scenario. Logarithmic scale.

Figure 9. Drift scenarios in which the validation chosen arm A degrades on the stream (Gaussian noise on the logit of its probabilities, standard deviation 1.5 or 4.0, from the start or from one third of the stream) under three delay settings: cumulative regret against the best fixed arm on the stream for static deployment of arm A and for the adaptive policies (mean of 30 replicates with 2.5th to 97.5th percentiles; the static value has no replicate variation).

Figure 10. Composition of the reward gap of arms B and C relative to arm A on the deployment stream under the primary weights: the utility term (Brier loss), the safety term (missed deaths) and the cost term (declared cost), each on the rescaled unit reward. Diamonds mark the total gap; positive values favour arm A.

Figure 11. Reward gap of arms B and C relative to arm A as a function of an alert burden weight λ_f added to the composite reward, computed from the stream error rates (lines), with the simulated sensitivity grid (points). The gap crosses zero at $\lambda_f = 0.47$ for arm B and 1.36 for arm C, the weights at which the preferred arm changes.

Figure 12. Sensitivity of each arm at its validation threshold on the deployment stream by age group, sex and administrative race group, with Wilson 95 percent intervals; the dashed line marks the 80 percent sensitivity at which each threshold was set on validation data.

Figure 13. Extended horizon of 30,000 decisions: empirical distribution across 30 replicates of the decision index at which the empirical Bernstein policy retired arm C under the primary weights and retired arm A under severe degradation of arm A, against the stream length (dashed line) and the decision indices implied by the sample complexity formula for each case (dotted lines).

Figures

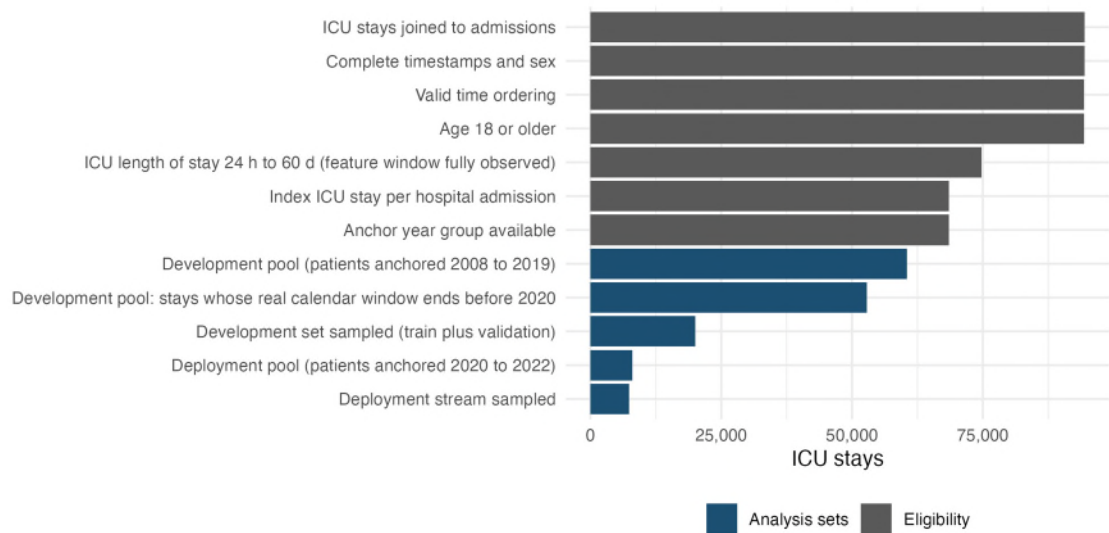


Figure 1

Figure 1.

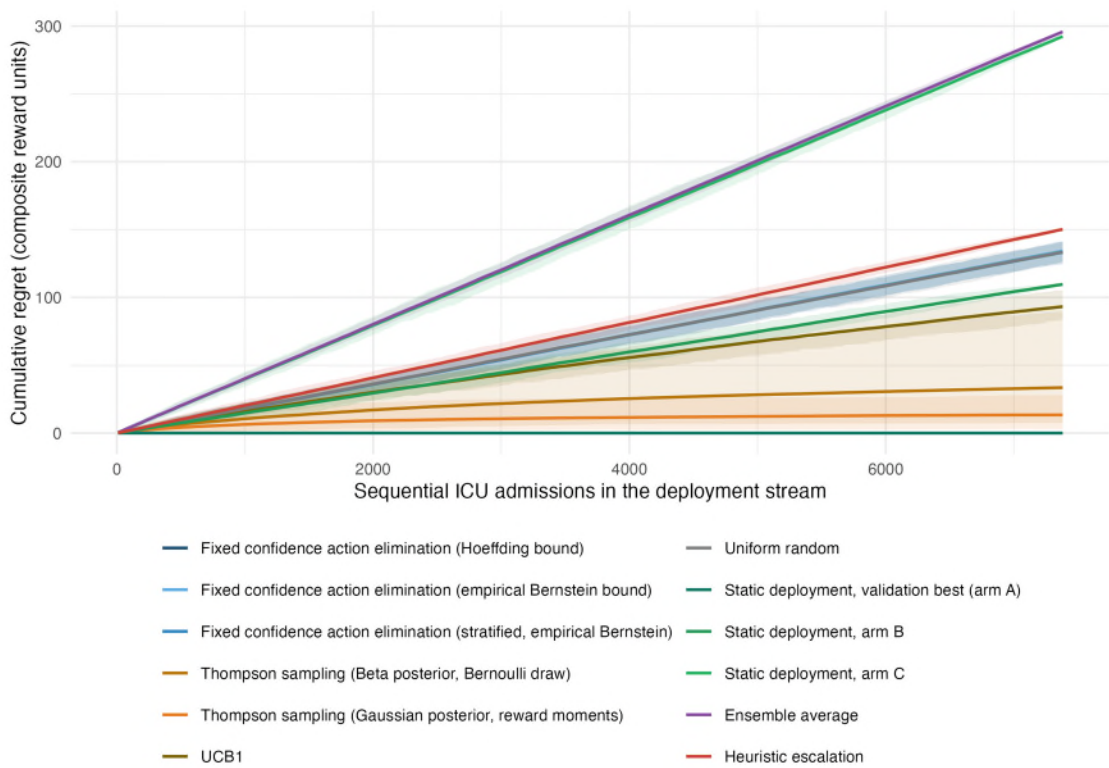


Figure 2

Figure 2.

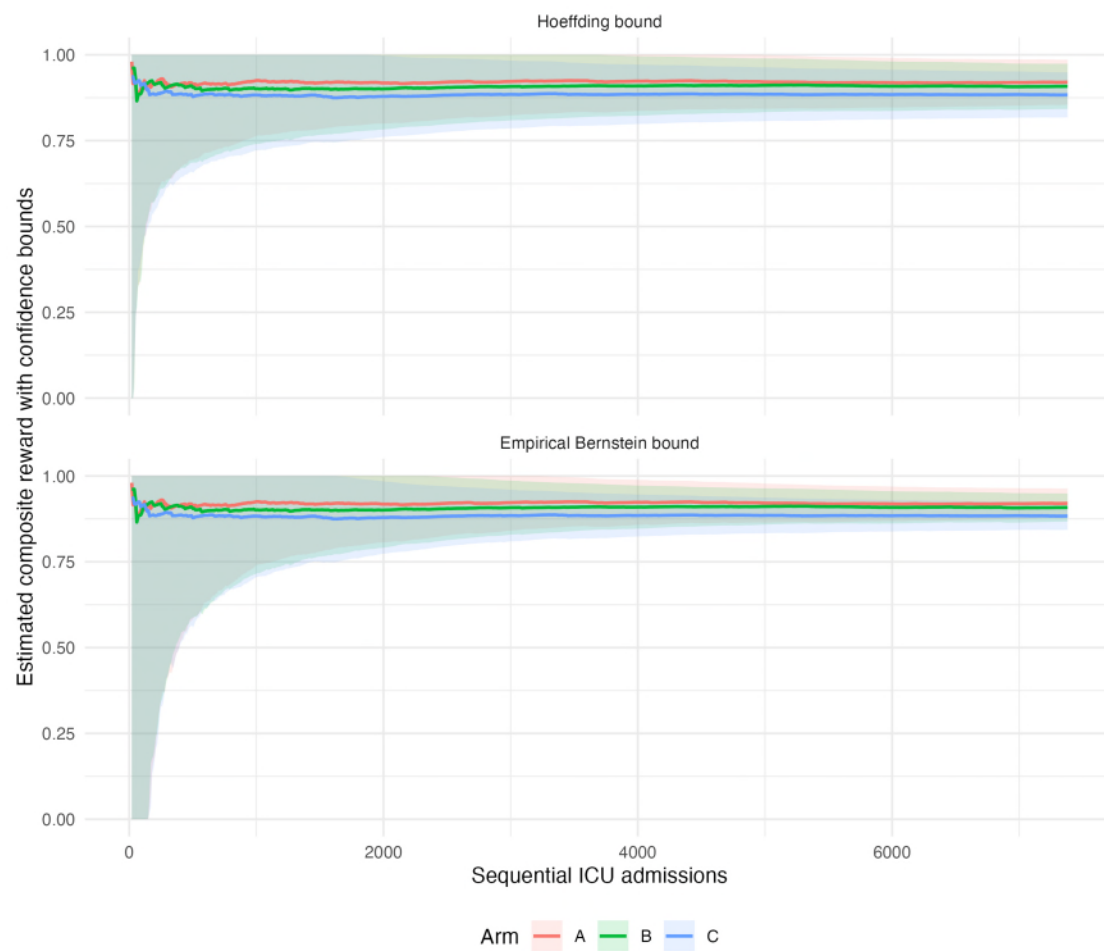


Figure 3

Figure 3.

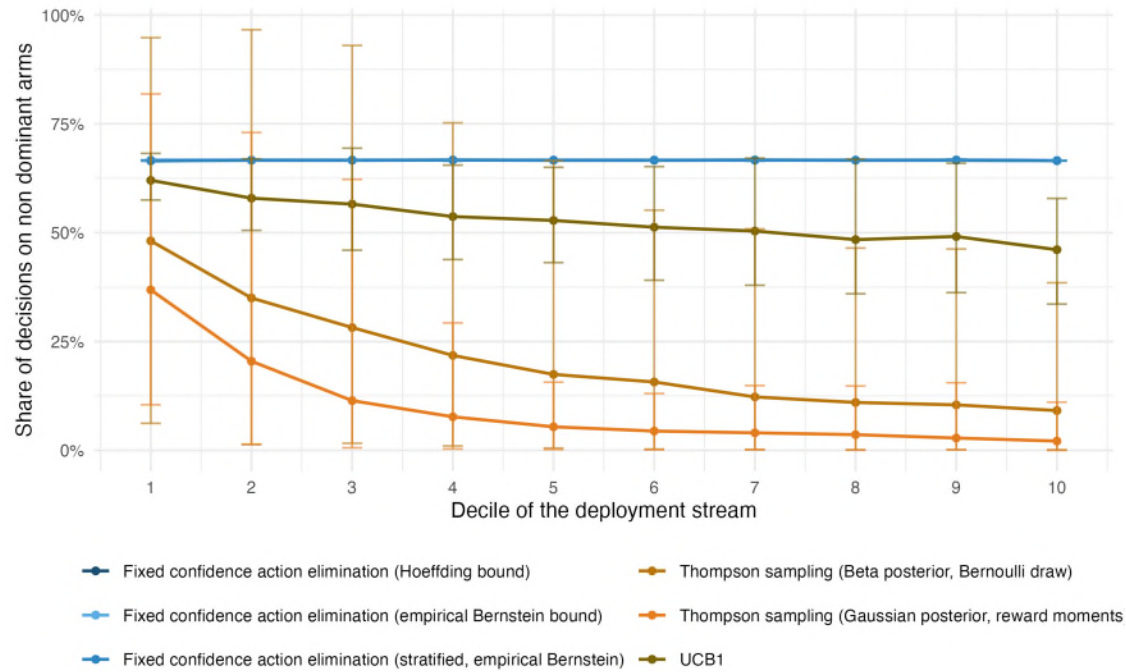


Figure 4

Figure 4.

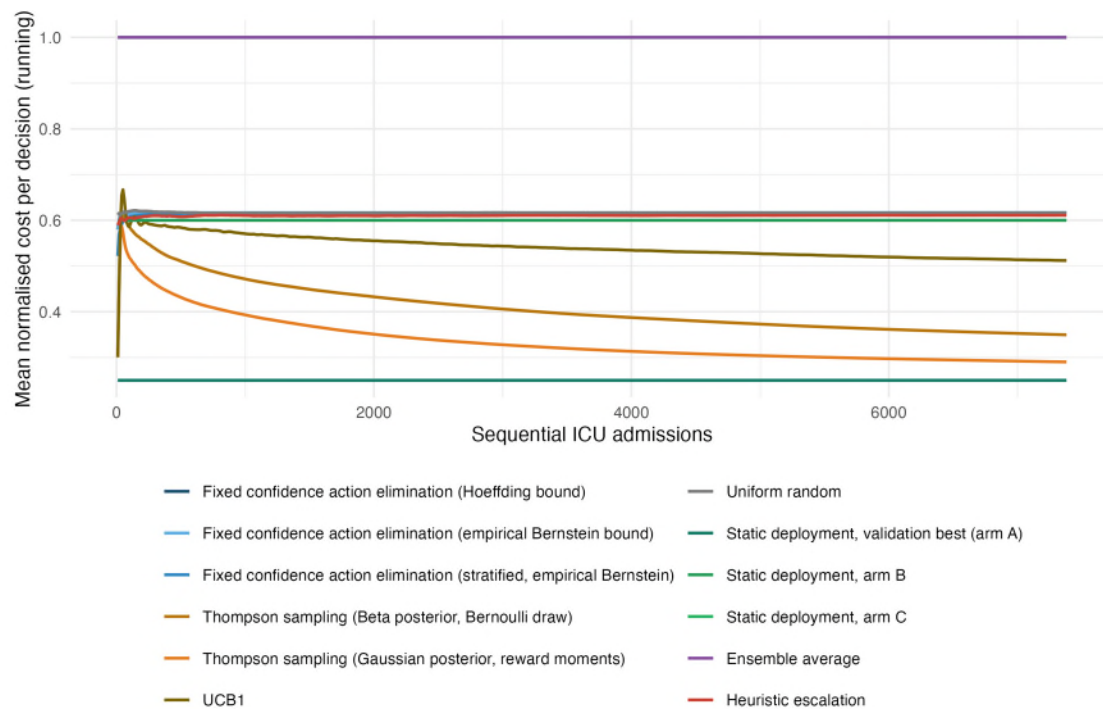


Figure 5

Figure 5.

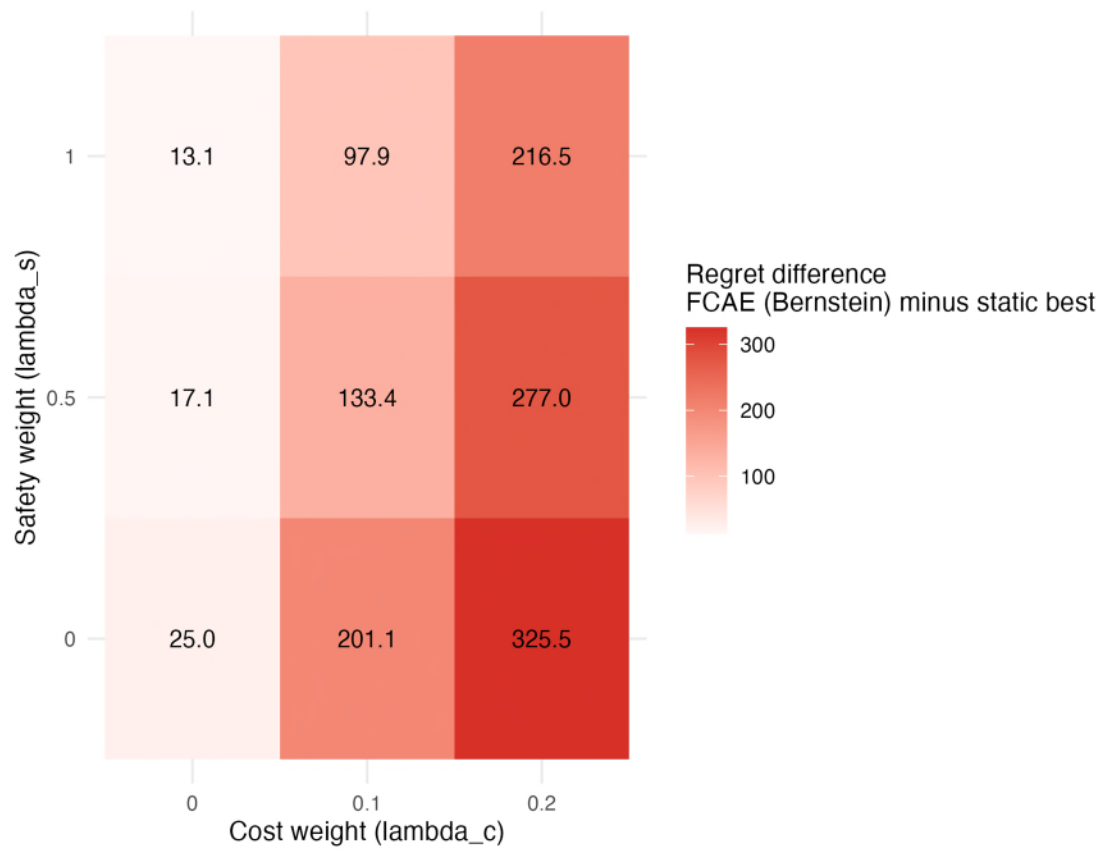


Figure 6

Figure 6.

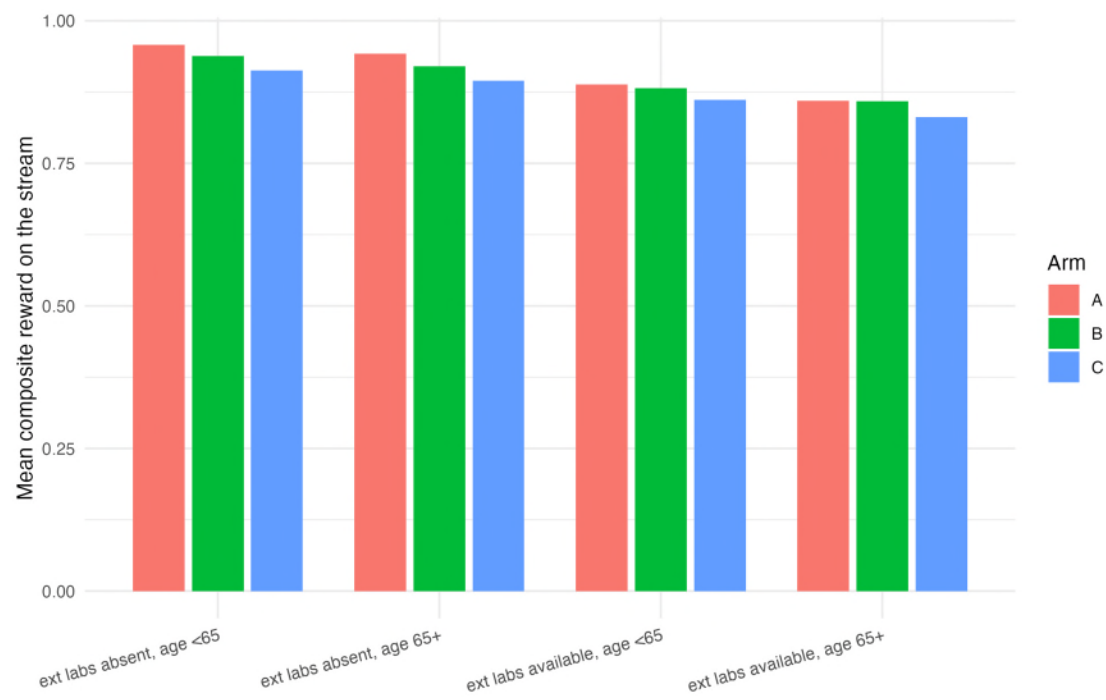


Figure 7

Figure 7.

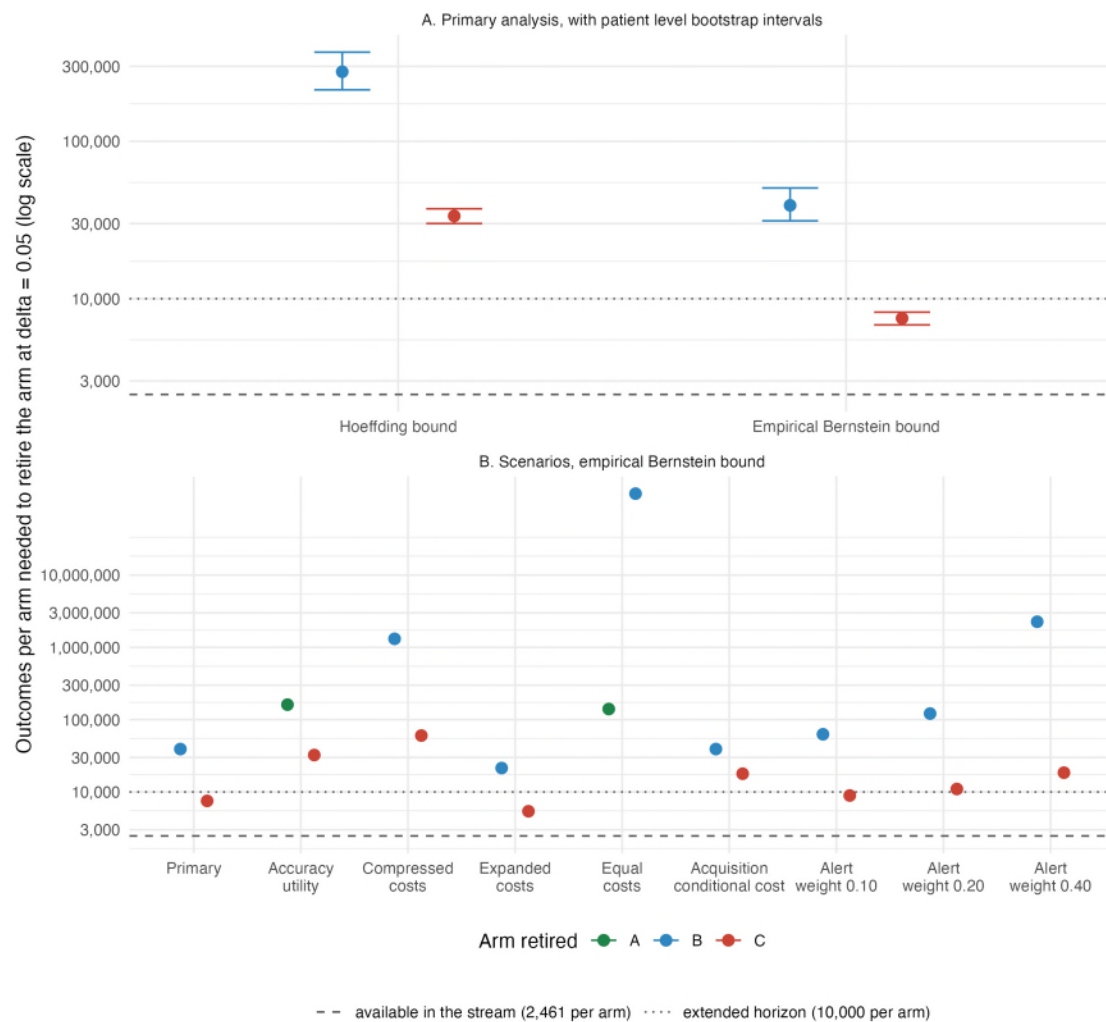
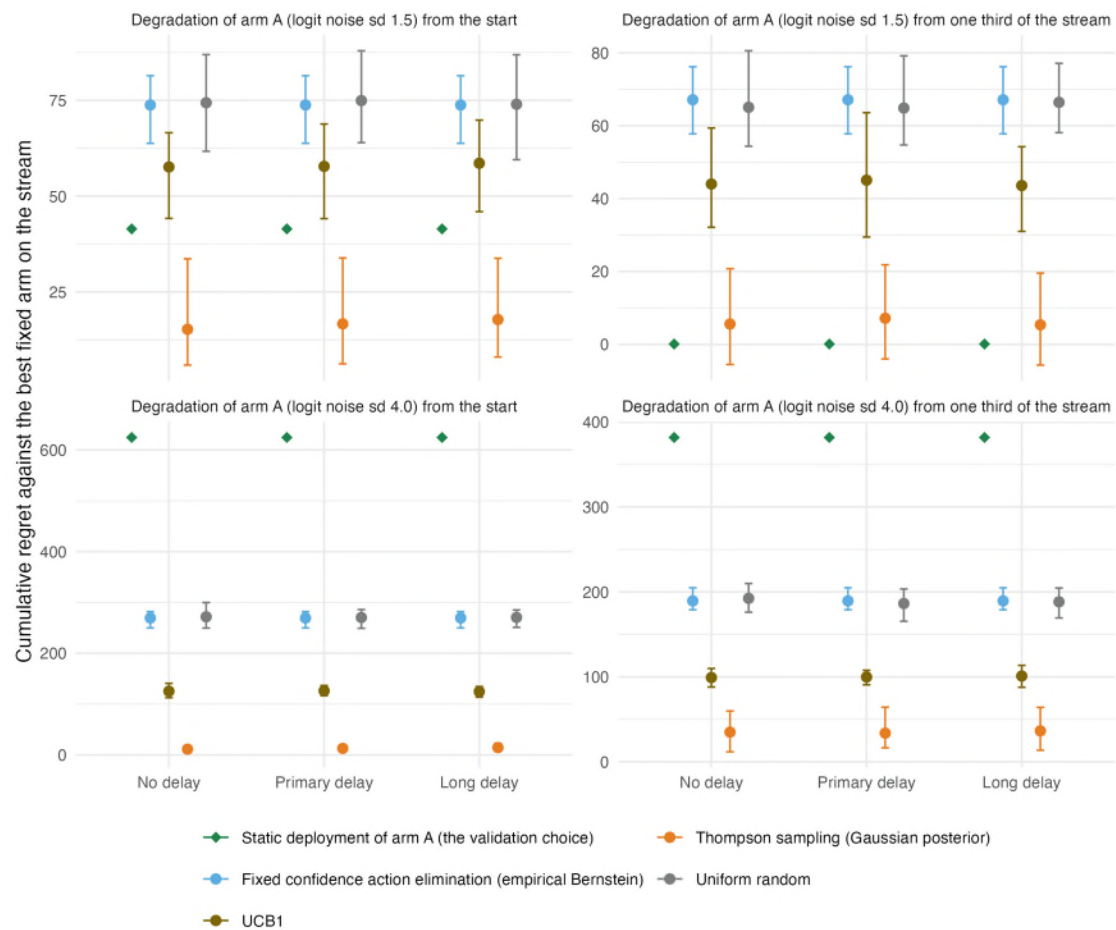


Figure 8

Figure 8.



Static deployment of arm A is shown without an interval; the regret of the other policies is the mean over 30 replicates with its 2.5 and 97.5 percentiles.

Figure 9

Figure 9.

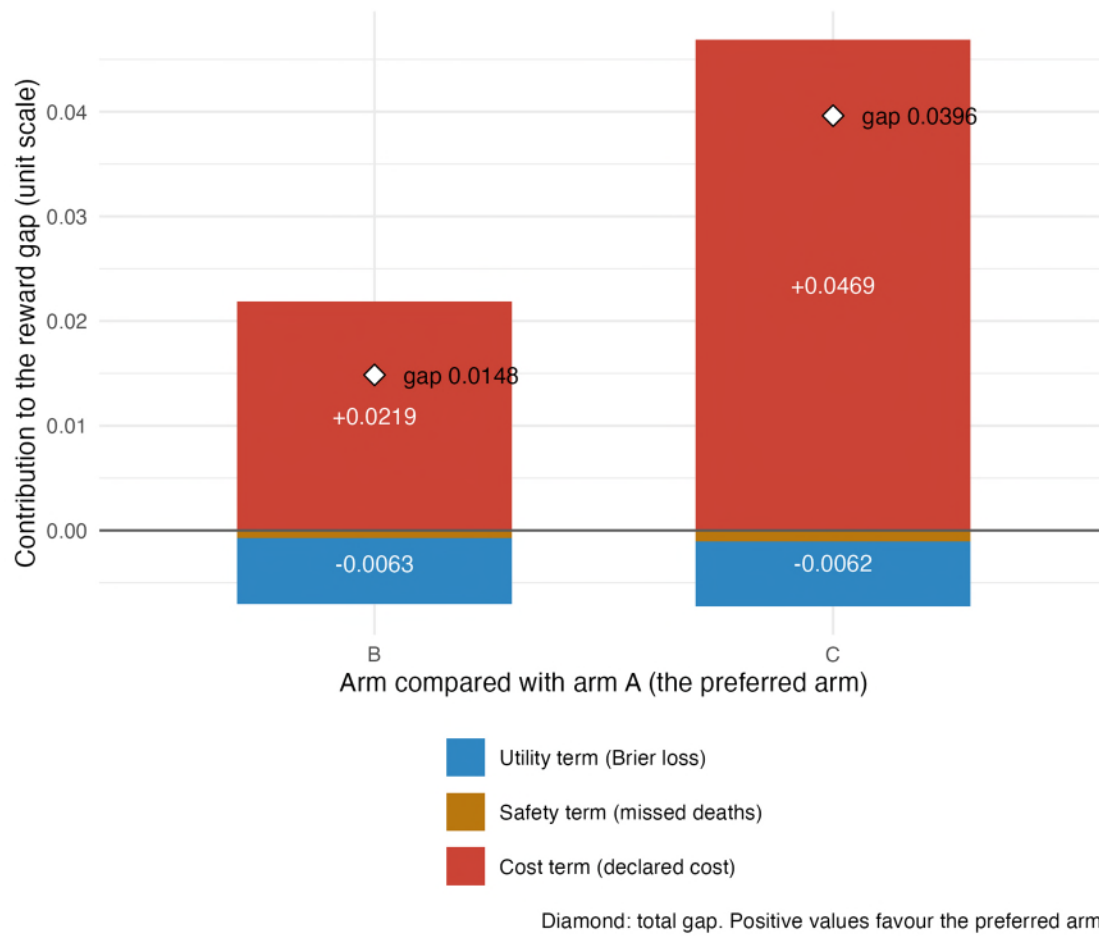


Figure 10

Figure 10.

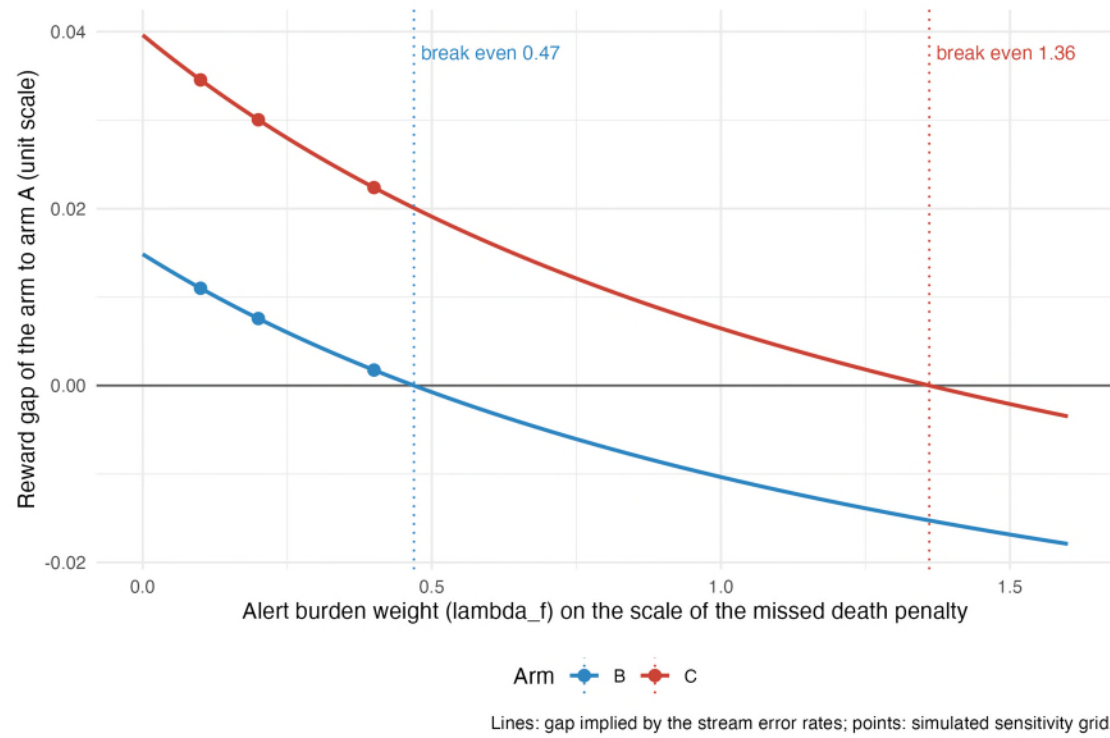


Figure 11

Figure 11.

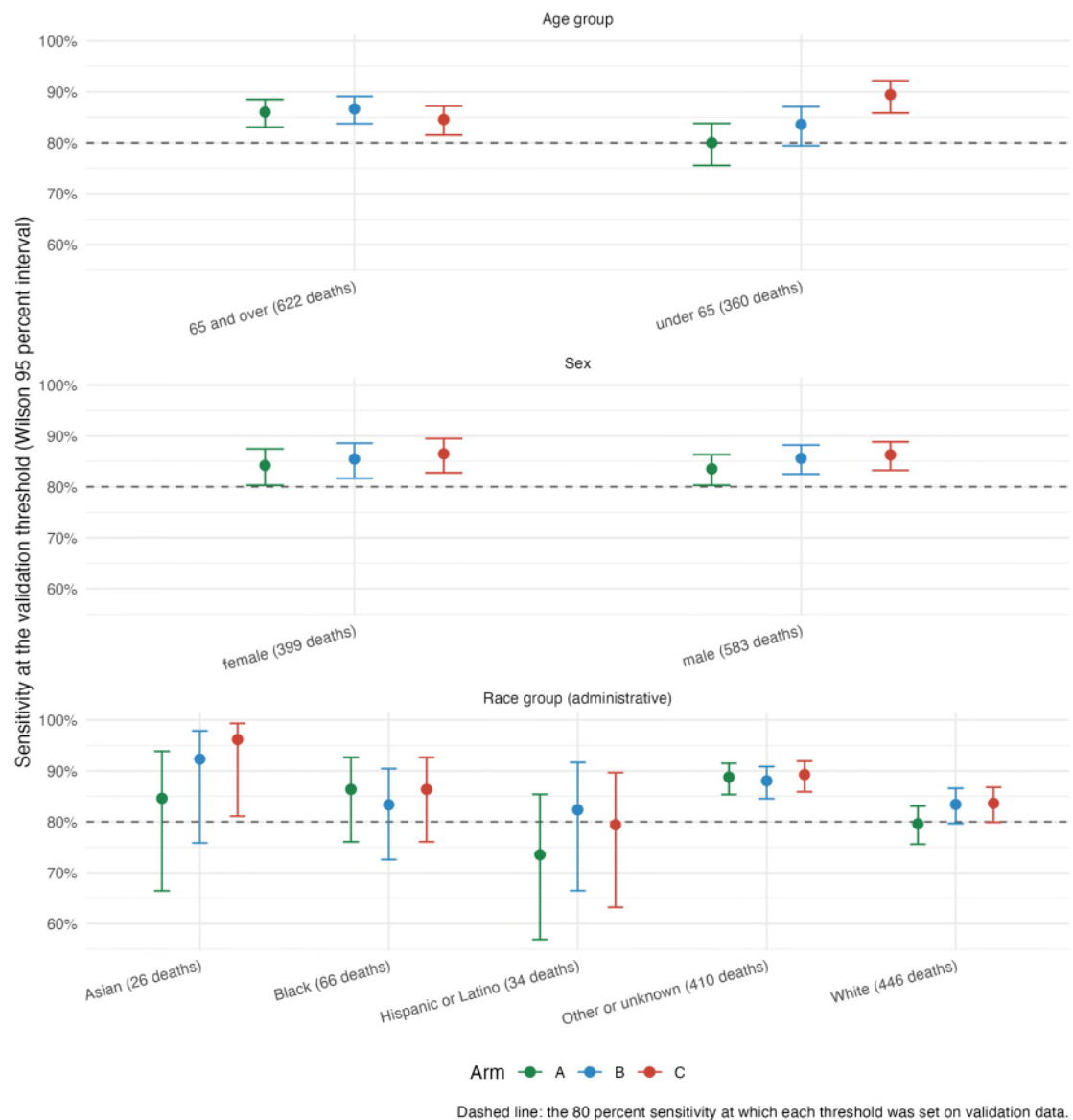


Figure 12

Figure 12.

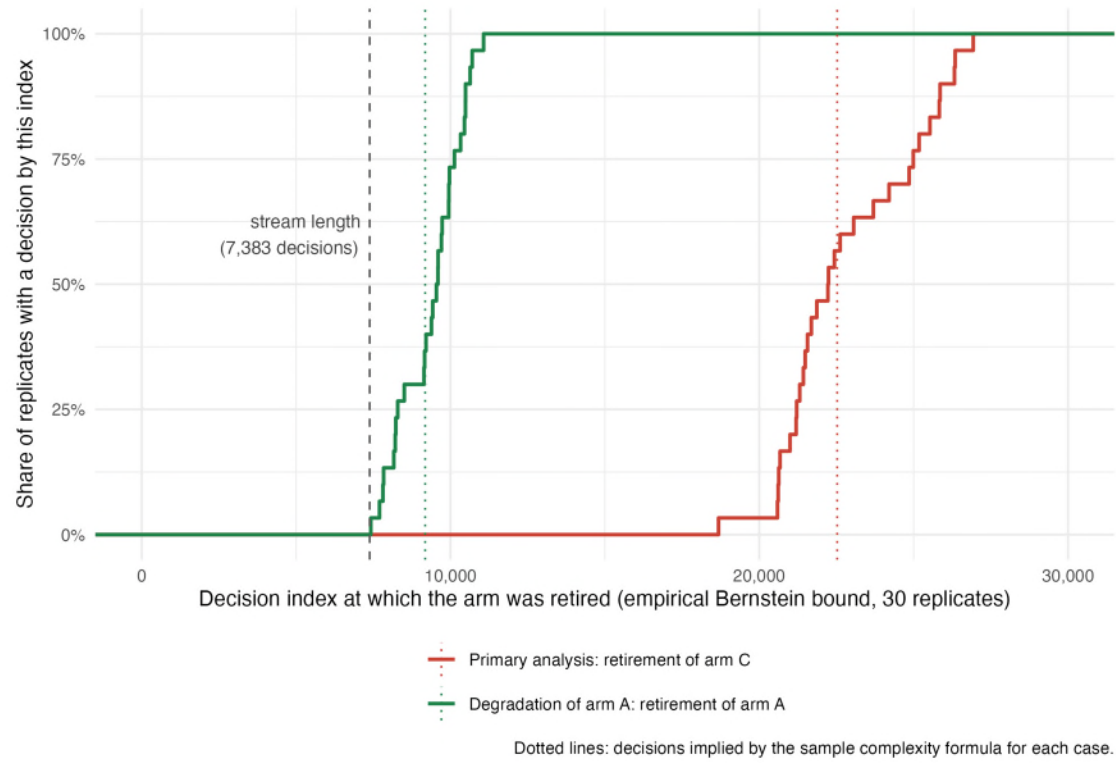


Figure 13

Figure 13.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Borgesrs8651972v2supplementarymaterial.docx](#)