Hierarchical & Shrinkage Thompson Sampling for Stable Off-Policy Evaluation On MIMIC-IV

Qianruo Tan

Department of Statistics, University of Connecticut, Storrs CT 06268, USA

qianruo.tan@uconn.edu

Abstract. Trial-and-error is costly in intensive care. Clinicians face high-stakes, noisy decisions where experimentation is risky. We study how to make Thompson Sampling (TS) reliably evaluable offline before deployment. We introduce two variants—Hierarchical Empirical-Bayes TS and Hierarchical Shrinkage TS—that share information by using log across arms to stabilize exploration, especially in early stage. To enable fair and high-power improvement, we adopt an ε-mix shared-log design that improves action overlap while keeping a fixed log for all target policies. Off-policy evaluation uses Inverse Propensity Scoring with weight diagnostics, with effective sample size, upper quantiles, max, and paired bootstrap over seeds for uncertainty. On synthetic data and MIMIC-IV with 6-hour grid setting. Our pipeline yields small but consistent gains over standard TS together with healthy weight distributions and large effective sample sizes. Without ε-mix, early exploration is imbalanced and IPS becomes unstable. Overall, coverage-aware logging plus hierarchical and shrinkage priors provides a reproducible pathway to assess TS policies safely and credibly for ICU decision support.

Keywords: Thompson Sampling, Hierarchical Bayes, Shrinkage prior, ε-mix shared log, Off-Policy Evaluation.

1. Introduction

Clinical decision-making in the intensive care unit (ICU) is high-risk and time sensitive. Trial-and-error approaches pose significant risks to patients. Algorithmic decision support offers a principled method to reduce human error by mitigating biases and distractions in judgment through established expertise. This proves invaluable when clinicians must act amid uncertainty and information overload. In such scenarios, decision learning frameworks can serve as tools for offline evaluation of treatment allocation strategies prior to deployment. This eliminates the need for high-risk experiments [1]. This perspective aligns with recent research findings on judgment errors and confounding factors in intensive care units (ICUs). It underscores the complementary value of structured, data-driven support systems to clinical expertise [2].

Among multi-armed bandit (MAB) methods, Thompson Sampling (TS) provides a natural approach to sequential decision making by balancing exploration with exploitation. TS has strong empirical performance across recommendation, adaptive experimentation, and clinical treatment selection tasks [3][4]. However, a well-known drawback emerges during the early learning phase: exploration may become imbalanced across different policy branches. When certain actions are sampled too infrequently, the coverage of recorded data becomes insufficient. This leads to variance inflation in policy evaluation metrics (OPE), particularly the inverse propensity score (IPS), thereby undermining the reliability and reproducibility of policy value estimates [5][6].

Therefore, enhancing the reliability of open-ended exploration (OPE) is a prerequisite for applying TS to real-world scenarios. Two approaches can address this requirement: First, hierarchical Bayesian modeling combined with shrinkage priors facilitates information sharing across groups, curbing TS's initial overconfidence. This enables a smoother exploration process and more stable posterior inference [1]. Secondly, in terms of logging, injecting a small amount of random behavior through the ε-mix logging strategy can improve action overlap while maintaining overall behavior that is clinically plausible. This facilitates more uniform exploration [6]. These design choices collectively

aim to control the weight tails that cause IPS instability, thereby achieving a larger effective sample size (ESS), a tighter uncertainty range, and more reliable conclusions [5][6].

In this study, we develop and explore two variants of TS for early-stage stability. One is the Hierarchical Empirical Bayesian TS (Hier-EB), and the other is the Hierarchical Shrinkage TS (Hier-Shrink). Both models share statistical strength across strategies. The Hier-Shrink model further introduces a stronger neutral prior shrinkage mechanism to reduce variance and guard against early overfitting. Evaluation employs an ε-mix shared log protocol, where all target policies are assessed on identical fixed logs. Fair and highly efficient comparisons are achieved through multi random seed paired statistics. Out-of-policy (OOP) performance is evaluated using only IPS, supplemented by weighted diagnostics and cross-seed pairing guidance to derive confidence intervals and approximate p-values [5][6]. This protocol emphasizes reproducibility and coverage-aware evaluation rather than reliance on single-run scores.

We validated this framework using synthetic data and the MIMIC-IV dataset represented in a 6-hour grid. Following the cohort construction and preprocessing methods established in prior studies [7][8][9][10]. Specifically, we compared two complementary experimental settings. The first is a stress-test baseline without ε-mixing and using a large batch size (B=100) to characterize instability during the early exploration phase. The second is a practical setting featuring ε-mixing logging (ε=0.05), shared logs, B=50, and n=20 seeds. Here, we quantify improvements over standard TS using seed-level pairing guidance. Under these configurations, we observe modest gains in directional consistency for both Hier-EB and Hier-Shrink, alongside healthier weight diagnostics and larger effective sample sizes. This indicates that hierarchical/shrink prior and coverage-aware logging enable TS to perform more reliably on real ICU data.

In this paper, the contributions are listed as follows:

1. Two early-stabilized TS variants-Hier-EB and Hier-Shrink-that share information across arms via hierarchical Empirical-Bayes and shrinkage priors.
2. A repeatable open-experiment (OPE) protocol integrating ε-mix shared log evaluation, IPS, weight diagnostics, and seed-level pairing guidance, adhering to best practices for trusted comparisons. [5][6].
3. An end-to-end ICU case study on MIMIC-IV (6-hour grid) showing small, consistent gains over standard TS and markedly improved evaluation stability; this supports safer, offline-first assessment of decision policies prior to clinical deployment [4][5][6][11]. We propose two early-stabilized TS variants—Hier-EB and Hier-Shrink—that share information across arms via hierarchical Empirical-Bayes and shrinkage priors.

2. Data & Preprocessing

We use MIMIC-IV v3.1, a large, de-identified critical-care database with rich longitudinal signals from ICU stays [8]. Our pipeline turns raw EHR tables into a bandit-ready cohort and event log following established practice in MIMIC-Extract and related clinical studies [7][9][10].

Guided by the inclusion/exclusion logic in Timing of vasopressin initiation and mortality in septic shock (applied to MIMIC-III/IV) [9], we restrict the study to adult patients (≥18 years) and retain only the first ICU admission per subject to avoid within-patient dependence. We remove records with missing or clearly invalid demographics (e.g., age, sex, admission timestamps). The resulting cohort.csv stores subject identifiers and ICU stay metadata (demographics, admission/discharge times). This pragmatic, safety-oriented screen mirrors prior sepsis analyses while keeping the population broadly representative [9].

To align heterogeneous measurements, we aggregate raw time-stamped records into fixed 6-hour windows—from events.csv.gz to events_6h.csv.gz. The grid synchronizes vitals, labs, interventions, and outcomes across patients, reduces sparsity, and preserves clinically meaningful dynamics. Following recommendations on the temporal-resolution vs. missingness trade-off [10], the 6-hour window captures early physiological changes without the noise amplification typical of finer bins.

Missing values are handled with carry-forward imputation for quasi-continuous streams and median imputation otherwise, as in MIMIC-Extract [7]. Feature definitions, value domains, and aggregation rules are documented in feature_schema.json and feature_schema_6h.json.

We organize configurations—TS variant (Standard, Hier-EB, Hier-Shrink), batch size (B), ε-mix setting, and global/random seeds—into dedicated directories. Each configuration binds a fixed cohort, grid events, and schema, enabling deterministic reruns and systematic comparison under controlled conditions. This grid-cohort framework provides a stable basis for downstream IPS-based OPE and seed-level paired bootstrap [7][10].

3. Methods

3.1. Three TS Variants (Keywords definition well specified)

We implemented two setups in this study for comparison:

1. Setup A — No-ε, independent logs (B=100).

Each policy is executed separately and produces its own log (no ε-mix). This regime stresses early-phase instability (uneven exploration, poor overlap) and is used as a diagnostic baseline for OPE variance.

2. Setup B — ε-mix, shared log (ε=0.05, B=50, n=20 seeds).

A single shared exploration log is generated with ε-mix logging and reused to evaluate all target policies. This enables paired comparisons across seeds, improves action overlap, and increases statistical power. This is our primary evaluation regime.

We implemented three variants of Thompson Sampling (TS) for comparison:

1. Standard TS: Independent per-arm learning with a regularized linear model. No information sharing across arms. This serves as the canonical TS baseline and represents how TS behaves without hierarchical coupling.

2. Hierarchical Empirical Bayes TS (Hier-EBTS): Adds a lightweight hierarchical layer that borrows strength across arms. Concretely, each arm's parameters are softly pulled toward a data-driven global center estimated by Empirical Bayes and re-estimated every B steps. Intuition: in early, data-sparse phases, this reduces over-confidence on rarely sampled arms, yielding more balanced exploration and smoother weight distributions for OPE.

3. Hierarchical Shrinkage TS (Hier-ShrinkTS): Builds on Hier-EB but applies stronger shrinkage toward the global center. This is a more conservative variant designed to further damp early variance and prevent arm starvation. It trades a bit of short-term flexibility for greater stability and overlap, which is desirable in high-risk settings like ICU care.

3.2. Thompson Sampling

Thompson sampling selects actions by sampling model parameters from the posterior distribution and making greedy decisions under the sampled model. This rule naturally balances exploration and exploitation and has shown strong empirical performance in contextual bandits [3][4].

We implement contextual TS with a linear–Gaussian model. For each decision, we draw linear parameters, compute action scores on the current patient features, and select the action that scores highest. This is the standard LinearTS formulation for contextual bandits with linear payoffs, as in [12]. In code, we use a regularized linear model (ridge prior/noise assumption) and Monte-Carlo posterior sampling to obtain action probabilities $\mu(a \mid x)$; the same probabilities are then used consistently for both online simulations/logging and off-policy evaluation (OPE). Our per-arm linearization is equivalent to the shared-parameter formulation in [12] via a block-feature view, it preserves the same LinearTS mechanics while matching our feature layout.

LinearTS is fast, stable, and well-suited to our 6-hour grid features; it pairs cleanly with IPS-based OPE (propensity logging, weight diagnostics, ESS) and provides a transparent baseline against which we add structure. The two variants introduced next—Hierarchical Empirical-Bayes TS (Hier-EB) and

Hierarchical Shrinkage TS (Hier-Shrink)—keep the same LinearTS sampling mechanism but introduce cross-arm information sharing to stabilize early exploration and improve overlap [3][4][12].

3.3. ε-mix & Shared Exploration Log

The motivation of us to develop TS by ε-mix and shared logging has been triggered by [6]. Early-phase Thompson Sampling (TS) can under-sample some actions, producing poor overlap and heavy-tailed IPS weights, which inflates variance and makes OPE unstable [3][4]. In high-stakes ICU applications we cannot "explore online until it stabilizes"; we need offline, coverage-aware logs that keep propensities away from zero while remaining clinically plausible [1][2][11].

For ε-mix logging, we introduce minimal explicit randomization during data collection. This ensures that actions are selected uniformly at probability ε from the acceptable action set, while actions with probability $1-\varepsilon$ follow the baseline behavior policy. This establishes a non-zero lower bound on propensity, mitigates the action-greedy problem, and increases the effective sample size (ESS), thereby yielding more stable IPS weights. These are widely recognized as essential prerequisites for credible Operational Performance Evaluation (OPE). Through this design, we endow the model with a non-zero bias threshold that mitigates branch starvation while enhancing effective sample size (ESS), yielding more stable IPS weights. These are universally acknowledged as critical foundations for achieving reliable Operational Performance Evaluation (OPE) [3][4][5][6][11]. We set $\varepsilon=0.05$ to minimally perturb routine clinical behavior while still ensuring adequate early coverage.

We opt to share exploration logs rather than generate separate logs for each candidate policy. By generating a single ε-mix exploration log and reusing it to evaluate all target policies, this approach offers three key advantages. First, it ensures fairness and comparability. All policies are evaluated under identical state-action distributions and overlap conditions, eliminating performance biases caused by the idiosyncrasies of distinct logs. Second, statistical power is significantly enhanced, as unified logs enable pairwise analysis. All strategies are evaluated under identical state-action distributions and overlapping conditions, preventing performance discrepancies from being obscured by the idiosyncrasies of independent logs. Furthermore, this approach achieves greater statistical power. Utilizing a unified log across strategies enables random seed pairing analysis, thereby reducing variance and providing more precise confidence intervals for strategy differences [5][6]. Finally, there is its practicality and safety. The single-log design minimizes the number of data collection sessions required in clinical settings and aligns with the offline evaluation principle adopted in previous multi-arm bandit/reinforcement learning studies in the medical field [11].

This paper presents two systems: Setup A and Setup B. We log each interaction, recording contextual features, selected actions, observed rewards, and behavioral propensities used to compute IPS. During logging, an ε-mixing policy is applied, and action probabilities are stored after normalization. The evaluation phase employs a conservative threshold (clip=20) to clip weights while monitoring weight summaries and ESS to ensure estimation stability [5][6]. Seeds are fixed and enumerated (20 values) to guarantee reproducibility across runs.

However, trade-offs still need to emphasis. A larger ε improves overlap but perturbs behavior more; a smaller ε better preserves usual practice but risks heavier tails. Our choice ($\varepsilon=0.05$) follows the OPE literature's guidance to use minimal but sufficient randomization for stability [5][6], and it fits the offline-first, risk-averse setting of ICU decision support [1][2][11].

3.4. Off-Policy Evaluation (OPE)

We use off-policy evaluation to estimate the performance of candidate policies from logged interaction. Without additional online experimentation. In healthcare, and specifically with MIMIC data. Importance-sampling–based OPE is a common, safety-preserving approach for counterfactual assessment [11]. In our study, OPE serves two purposes. The first is to provide an unbiased estimator under correct propensities. The second is to enable apples-to-apples comparisons across policy variants under identical coverage conditions.

Our primary estimator is Inverse Propensity Scoring (IPS) with conservative weight clipping. For each recorded interaction, we store its context, selected action, observed reward, and behavioral propensity. Action probabilities are normalized and persisted to support reproducible IPS computation [5][6][11]. During the evaluation process, we report standard-weighted diagnostic metrics, the mean (approximately 1 when the model is fully specified), upper quantiles, and maximum values. We also provide the effective sample size (ESS) to monitor stability [5][6].

To characterize the variability of estimators and compare different strategies, we employ seed-based paired bootstrapping. All strategies reuse the same random seed, enabling pairwise difference calculations. This method provides more precise confidence intervals and approximate p-values for strategy comparisons [5][6]. This protocol emphasizes reproducibility, overlapping awareness, and transparent uncertainty reporting consistent with OPE practices [11].

4. Experimental Design

We evaluate the three TS variants under two complementary logging regimes. The goal is to expose early-phase instability without extra randomization (diagnostic baseline) and test the variants fairly under identical coverage with higher statistical power.

4.1. Common setup (applies to both experiments)

Dataset: MIMIC-IV v3.1 [8]; cohort construction and 6-hour grid per Section 3 (following [7][9][10]).

Arms (actions): K=3 discrete treatment options defined at the 6-hour decision points.

Reward: Binary/continuous reward computed per grid step (definition as in Sec. 3; unchanged across runs).

Behavior model (for propensities): Regularized multinomial logistic regression fit on the logged data; predicted class probabilities are stored and used for IPS.

Target policies: Standard TS; Hier-EBTS; Hier-ShrinkTS (definitions in Sec. 3.1).

OPE estimator: IPS with weight clipping at clip = 20 and probability floor prob_floor = 1e-6.

Diagnostics: Report weight mean (≈1), p95/p99/max, and ESS.

Uncertainty: Paired bootstrap over seeds (B=2000 resamples) to form 95% CIs and approximate p-values for differences vs. Standard.

Other fixed knobs:

1. TS sampling draws per decision: ts_samples = 200
2. Hour cap for evaluation window: cap_hours = 48

Global seed grid for runs: Seeds = {1,3,7,11,17,23,29,31,35,40,43,51,55,56,63,67,68,73,78,91} (20 seeds)

Primary metric: Mean IPS; we also report lift vs. Standard and the diagnostic tuple (ESS, w_p95, w_p99, w_max).

4.2. Experiment-A — No ε, independent logs (B=100)

Purpose: Stress-test early-phase behavior without extra randomization; reveal overlap issues and heavy-tailed weights.

Protocol:

1. Each policy (Standard, Hier-EB, Hier-Shrink) is executed separately; each produces its own interaction log (no ε-mix).
2. Batch size: B=100 updates.
3. For every seed in the grid, run all three policies and evaluate each run with IPS (same clipping/diagnostics as above).

Expected observation: Larger w_p99 / w_max and smaller ESS; useful as a diagnostic baseline but not our primary comparison regime.

4.3. Experiment-B — ε-mix, shared log (ε=0.05, B=50, n=20 seeds) [Primary]

Purpose: Enforce minimal randomized coverage and evaluate **all** target policies on the **same** log for apples-to-apples comparison and higher power.
Protocol:

1. Generate a single shared exploration log per seed using ε-mix logging with ε = 0.05 and batch size B=50.
2. Store the behavior propensities for every interaction.
3. Evaluate all three target policies off-policy on the same log; compute paired differences vs. Standard per seed; aggregate with paired bootstrap.

Rationale: Shared logs align the state–action distribution across policies and enable paired analysis, which reduces variance of differences and yields tighter CIs under identical coverage.

5. Experiment result

We report off-policy performance for the three TS variants under two complementary regimes. **Exp-A** is a *diagnostic baseline* with **no ε-mix** and **independent logs (B=100)**; it highlights how early-phase exploration affects overlap and IPS stability. **Exp-B** is the *primary evaluation* with **ε-mix=0.05**, a **shared exploration log (B=50)**, and **n=20 seeds**, enabling paired comparisons. Across both, we present **IPS** as the main metric, plus **weight/ESS diagnostics.** In Exp-B we add **seed-level paired bootstrap 95% CIs and approximate p-values** to support inferential claims.

Table 1. Exp-A (B=100, ε=0, independent logs), Off-policy evaluation.

	Policy policy	ips	w_mean	w_max
0	hier-eb	0.551089	1.040072	18.968473
1	hier-shrink	0.550419	1.04256	27.367826
2	standard	0.536524	1.012042	26.762724

As shown in Table 1, the off-policy IPS estimates under Exp-A indicate small lifts of the hierarchical variants over the standard TS; these are descriptive because logs are independent (no paired seeds). Under Exp-A, hier-eb attains IPS=0.5511 and hier-shrink IPS≈0.5504, versus 0.5365 for standard. These point estimates suggest small lifts of estimate +0.0146 (≈+2.7%). For hier-eb and +0.0139 (≈+2.6%) for hier-shrink. Because this setting uses independent logs (no shared log, no paired seeds), these numbers should be considered descriptive rather than inferential; variations may depend on the specific overlaps inherent in each run [5][6].

Table 2. Exp-A (B=100, ε=0), Weight diagnostics & effective sample size.

	policyw_mean	w_p95	w_p99	w_max	ESS
standard	1.008972	1.011236	3.974096	20.000000	4828.585197
hier-eb	1.040072	2.07879	9.499058	18.968473	3308.965553
hier-shrink	1.034941	1.47671	9.589768	20.000000	2886.376240

As reported in Table 2, the weight diagnostics for Exp-A and the ESS reveal unstable overlap compared with Exp-B, motivating the ε-mix shared-log design. Diagnostic results indicate that hierarchical variants exhibit heavier tail distributions and smaller effective sample sizes. Hier-eb has w_p95≈2.08, w_p99≈9.50, ESS≈3309; hier-shrink w_p95≈1.48, w_p99≈9.59, ESS≈2886; whereas standard shows w_p95≈1.01, w_p99≈3.97, ESS≈4829. Although w_max is capped at 20 (stability control), the elevated tail quantiles and reduced ESS values reveal instability in independent logging and IPS sensitivity to overlap. This is precisely why we shifted to ε-mix shared logging and paired seed analysis in Experiment B. [5][6].

Table 3. Exp-B (ε-mix=0.05, B=50, shared log, n=20 seeds) — Per-policy IPS with 95% bootstrap CI.

	poIPS IPS	IPS_CI_lo	IPS_CI_hi
policy			
standard	0.524773	0.522918	0.526718
hier-eb	0.547040	0.545818	0.548205
hier-shrink	0.542991	0.541468	0.544534

As shown in Table 3, under the ε-mix shared-log regime the per-policy IPS with 95% bootstrap CIs are tight and exhibit small, direction-consistent gains for Hier-EB/Hier-Shrink relative to Standard. On the ε-mix shared log (ε=0.05, B=50, n=20 seeds), IPS is 0.5248 [0.5229, 0.5267] for standard, 0.5470 [0.5458, 0.5482] for hier-eb, and 0.5430 [0.5415, 0.5445] for hier-shrink. The narrow confidence interval indicates low estimation uncertainty and good reproducibility under seed resampling, meeting optimal practice standards [5][6].

Table 4. Exp-B — Paired-seed bootstrap vs. Standard: ΔIPS with 95% CI and approximate p-value.

	Pol diff_IPS	diff_CI_lo	diff_CI_hi	p_approx
policy_vs_standard				
hier-eb	0.022266	0.020264	0.024217	0.0
hier-shrink	0.018218	0.017077	0.019408	0.0

As detailed in Table 4, seed-level paired comparisons against the Standard policy show positive diff_IPS with confidence intervals not crossing zero and near-zero p-values, indicating statistically significant improvements. To ensure fairness and control variance, a paired bootstrap method based on the same 20 seed values was adopted [6], hier-eb improves IPS by +0.0223 [+0.0203, +0.0242] (≈+4.2% relative), and hier-shrink by +0.0182 [+0.0171, +0.0194] (≈+3.5%). All 95% confidence intervals lie above zero, and p≈0.0, indicating that the increase is statistically significant.

Table 5. Exp-B — Weight diagnostics and effective sample size across seeds.

	Policy ESS_mean	ESS_median	ESS_min	ESS/N_mean
policy				
standard	3862.007959	3872.384967	3315.052964	0.386201
hier-eb	2582.273134	2575.246638	2165.483877	0.258227
hier-shrink	2398.283801	2406.225182	2080.674379	0.239828

As summarized in Table 5, Exp-B achieves large effective sample sizes (ESS) across policies, supporting stable IPS estimation under the shared-log setting. Aggregate diagnostics show ESS_mean (ESS/N_mean) of 3862 (0.386) for standard, 2582 (0.258) for hier-eb, and 2398 (0.240) for hier-shrink. Given that w_mean≈1 and w_max is capped at 20, these metrics indicate substantial overlap and stable operation of the IPS mechanism [5][6]. The ESS value of the stratified variant decreased slightly, reflecting a minor deviation in its distribution relative to the logging policy. This represents a reasonable trade-off compared to the sustained significant gains achieved in IPS performance.

Experiment A demonstrates that without ε mixing, the log-log model may exhibit heavier weight tails and lower effective sample sizes, rendering IPS differences merely descriptive rather than decisive. In contrast, Experiment B employs a shared log-log, paired-seed design to yield narrower confidence intervals, revealing statistically significant improvements for Hier-EB and Hier-Shrink over the baseline model, with Hier-EB typically performing optimally. Weight diagnostic metrics remain healthy (w_mean≈1, tails constrained). Although ESS slightly decreases in the hierarchical variant, it remains within reasonable bounds—a trade-off expected to achieve better coverage and stability. Thus, Experiment B constitutes the core conclusion, while Experiment A serves as an illustrative benchmark, clarifying the necessity of ε-mixing and shared logs for reliable OPE.

6. Conclusion

We investigated how to make Thompson sampling (TS) safer and easier to evaluate in high-risk environments like ICUs. Two elements proved most effective:

1. Structure in TS: Adding a light hierarchical layer (Hier-EB) and a stronger shrinkage variant (Hier-Shrink) promotes early, more balanced exploration across arms.
2. Evaluation that respects overlap: Using a small Œμ-mix during logging and a shared exploration log provides sufficient propensities for all actions, enabling stable IPS-based OPE with meaningful uncertainty.

In both experimental settings, the differences are evident. In Experiment A (no ε term, independent log-likelihood, B=100), we observe heavy-tailed importance weights and reduced effective sample size, with these symptoms of insufficient overlap leading to unstable importance weights. In Experiment B (ε=0.05, shared log, B=50, 20 seeds), weight diagnostics were healthy (mean≈1, constrained tails, effective sample size reaching thousands). Seeded pairing guidance showed Hier-EB and Hier-Shrink delivered small but stable gains over standard TS, with significant differences under the shared log setting. In summary, ε-mix resolves evaluation issues, and when evaluations are reliable, hierarchical/shrink TS yields modest yet dependable gains.

In practice, this work provides a reproducible analytical workflow for the MIMIC-IV hourly grid. Its core components include: standardized cohort partitioning, ε-blending logging, IPS processing with conservative truncation, weight/ESS diagnostics, and seed-level pairing-based guidance. This approach prioritizes robustness over superficial gains—a stance aligned with clinical decision support needs, as it emphasizes variance control and transparency.

Overall, stratified/shrinkage TS with ε-mix shared logging represents a practical and theoretically coherent approach that stabilizes early exploration, maintains clinical plausibility, and enables trustworthy off-policy evaluation—marking a crucial practical step toward reliable multi-armed bandit decision support in intensive care units.

In subsequent studies, we will exclusively employ IPS (based on design considerations to ensure transparency), enabling exploration of extensions to other operational probability estimators and more robust behavioral/target models without modifying the research workflow. The current reward design is relatively simplified; adopting richer clinical outcome measures may reveal more pronounced effects.

References

[1] Varatharajah, Y., Berry, B.A.: A Contextual-Bandit-Based Approach for Informed Decision-Making in Clinical Trials. Life 12, 1277 (2022).

[2] Peringa, I.P., Cox, E.G.M., Wiersema, R., van der Horst, I.C.C., Meijer, R.R., Koeze, J.: Human judgment error in the intensive care unit: a perspective on bias and noise. Critical Care 29, Article 86 (2025).

[3] Chapelle, O., Li, L.: An Empirical Evaluation of Thompson Sampling. Yahoo! Research Technical Report (2011).

[4] Agrawal, S., Goyal, N.: Near-Optimal Regret Bounds for Thompson Sampling. Journal of the ACM 64(5), Article 30 (2017). https://doi.org/10.1145/3088510

[5] Bayesian Off-Policy Evaluation and Learning for Large Action Spaces. arXiv preprint (2024).

[6] Rome, S., Chen, T., Kreisel, M., Zhou, D., et al.: Lessons on off-policy methods from a notification component of a chatbot. Machine Learning 110, 2577–2602 (2021).

[7] Howell, N., Neamatullah, I., Lin, K., Kuo, T.-T., et al.: MIMIC-Extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. In: Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI 2020). IEEE (2020).

[8] Johnson, A.E.W., Bulgarelli, L., Pollard, T.J., Gow, B., Moody, B., Horng, S., Celi, L.A., Mark, R.G.: MIMIC-IV (version 3.1). PhysioNet (2024). RRID:SCR_007345. https://doi.org/10.13026/kpb9-mt58

[9] Xu, J., Cai, H., Zheng, X.: Timing of vasopressin initiation and mortality in patients with septic shock: analysis of the MIMIC-III and MIMIC-IV databases. BMC Infectious Diseases 23, 199 (2023). https://doi.org/10.1186/s12879-023-08147-6

[10] Cao, J., et al.: Generalizability of an acute kidney injury prediction model across health systems. Nature Machine Intelligence (2022). (author manuscript)

[11] Komorowski, M., Celi, L.A., Badawi, O., Gordon, A.C., Faisal, A.A.: The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine 24(11), 1716–1720 (2018). https://doi.org/10.1038/s41591-018-0213-5

[12] Agrawal, S., Goyal, N.: Thompson Sampling for Contextual Bandits with Linear Payoffs. In: Proceedings of the 30th International Conference on Machine Learning (ICML 2013), pp. 127–135 (2013).