A Cohort-Level Evaluation of Thompson Sampling for Reducing Asthma Risk

Yuxiao Chen

School of Art and Science, University of Washington, Seattle WA 98105, USA

Harrychen2004@gmail.com

Abstract. Public-health programs are continually deciding which outreach action will most reduce risk in the populations they serve, often with limited data and changing conditions. In this study, I apply Thompson Sampling (TS) to cohort-level decision making for asthma-related interventions using the Behavioral Risk Factor Surveillance System (BRFSS) 2015 data, I defined seven intervention "arms" that plausibly affect asthma outcomes including baseline, smoking cessation, vaccinations, inhaler adherence education, preventive check-ups, weight control, and air-quality awareness. I also compared two algorithms: (i) fixed Thompson Sampling (TS) with a Beta(1,1) prior per arm and (ii) Empirical-Bayes Thompson Sampling (EB-TS) that fits Beta priors from offline estimates. Over 10 replications of 30,000 rounds, fixed TS achieves mean expected reward 0.2660 and mean regret 0.002065 per round, concentrating ~99% of pulls on the best arm. EB-TS increases the overall reward to 0.2681 and lower the mean regret to 0.000046 by stabilizing early decisions with data-informed priors. Results suggest fixed TS is a strong, hyperparameter-free baseline for cohort-level outreach, while EB-TS helps when offline reward estimates are reliable and improved reward with dense value feedback.

Keywords: Multi-armed bandits, Thompson Sampling, Empirical Bayes, mHealth, BRFSS.

1. Introduction

Exploration vs. exploitation is a key challenge in decision-making under uncertainty. In healthcare and public health, resources are limited. We want to know which outreach actions reduce risk. At the same time, we aim to avoid wasting money on ineffective options. Multi-armed bandits (MABs) offer a smart way to allocate resources sequentially. Among various MAB algorithms, Thompson Sampling (TS) is notable. TS stands out for its simplicity and Bayesian foundation. It also comes with strong theoretical guarantees. These features make TS a powerful approach. Its effectiveness has been shown in many applications. TS is well-supported in literature [1, 2, 3]. In this paper, I will investigate whether TS can guide cohort-level outreach. I use data from the BRFSS 2015. My research question is straightforward. If I need to choose an intervention for the next person, can TS learn this choice quickly? And can it do so reliably? I also want to know if the Empirical-Bayes (EB) version of TS performs better [4]. My contributions are threefold. First, I mapped the BRFSS data into seven intervention arms. These arms are relevant to asthma risk. Second, I evaluated the performance of fixed TS versus EB-TS. I compared their rewards and regrets [5]. Finally, I summarize the findings and present practical takeaways. These takeaways could support public-health operations.

2. Source and preprocess

My study uses the 2015 BRFSS dataset [6]. This dataset is a large survey in the U.S. It collects data on health behaviors and outcomes. I focused on a specific file from this dataset. Each record in the file is linked to seven predicted intervention effects. These effects are measured by reductions in adverse asthma outcomes. The outcomes include symptom severity, ER visits, and hospitalizations. The predicted values serve as inputs for bandit evaluation. The interventions in my study represent realistic, deployable public health measures. There are seven candidate interventions in total. The first is general education, which serves as the baseline. The second intervention is smoking-cessation programs. The third is vaccination. The fourth focuses on inhaler adherence education. The fifth is preventive checkups. The sixth is weight management and physical activity encouragement. The

seventh intervention promotes air-quality awareness. It also includes rescue-medication readiness. Each of these interventions is treated as an "arm" in the multi-armed bandit framework.

A single round of the simulated decision process proceeds by drawing one BRFSS record with replacement. For that record, each arm is associated with its predicted value. When the algorithm chooses an arm, the reward is the corresponding predicted reduction; the regret for that round is the gap between the best available value for the record and the value of the chosen arm. This setup follows standard definitions in the bandit literature and provides a principled way to evaluate sequential public-health decision-making [2, 3].

3. Methods

I used Thompson Sampling (TS) as the decision policy. TS maintains a belief distribution over each arm's mean reward, samples a value from this belief, selects the arm with the highest draw, and updates the belief based on feedback. This sample–then–act rule, with Bayesian updating, follows the modern tutorial by Russo et al. (2018) and traces back to Thompson's original 1933 formulation. For broader background on bandit models, regret notions, and algorithmic variants, I also follow Lattimore and Szepesvári [7]. Furthermore, I adopt the Beta–Bernoulli framework, where each arm's belief is a Beta distribution updated incrementally. Because rewards here are continuous predictions on [0,1], values closer to one are treated as positive feedback and values near zero as negative, preserving the canonical Beta bookkeeping without extra complexity [2, 5].

I also compared two policies. The first is a fixed, non-tunable TS, which assumes a uniform prior over each arm's mean reward. TS maintains for each arm a a belief about its mean reward θ_a using a Beta distribution. At round t it samples a plausible value for each arm and acts greedily on that sample:

$$\theta_a^{(t)} \sim Beta\big(\alpha_{a,t-1}, \beta_{a,t-1}\big), \tag{1}$$

$$A_t = \arg\max_a \ \theta_a^{(t)}. \tag{2}$$

After observing Y_t from each chosen arm A_t, TS performs the standard incremental Beta update (treating $Y_t \in [0,1]$ as fractional evidence):

$$\alpha_{A_t,t} \leftarrow \alpha_{A_t,t-1} + Y_t, \tag{3}$$

$$\beta_{A_t,t} \leftarrow \beta_{A_t,t-1} + (1 - Y_t), \tag{4}$$

and leaves the other arms' parameters unchanged [2,5]. The posterior mean for arm a after t rounds is:

$$E[\,\theta_a \mid \mathcal{D}_t\,] = \frac{\alpha_{a,t}}{\alpha_{a,t} + \beta_{a,t}}. \tag{5}$$

Fixed, non-tunable TS. I initialize each arm with a uniform prior

$$\alpha_{a,0} = 1, \tag{6}$$

$$\beta_{a,0} = 1 \quad (\forall a). \tag{7}$$

Such a design is simple to implement, and it also requires no hyperparameter calibration [2, 5]. The second is an Empirical-Bayes Thompson Sampling (EB-TS) design (upon practical MHealth work) (Tomkins et al., 2021). Eb-TS, I would start with setting a belief to each arm. I apply informed priority to this. The value of the priority is centered according to the average value of the arm in the analytic file. I pick the strength of the prior by trying to maximize a marginal-likelihood objective. This is being performed in a Beta-Binomial structure. After initialization, EB-TS operates similarly to the fixed version. Empirically, Thompson Sampling variants perform competitively across large-scale applications [8].

The key focus here is not whether TS learns. Its learning behavior is well-established [2, 3]. Instead, the focus is on whether starting with a well-informed priority provides a measurable advantage. This is particularly relevant in cohort-level public-health settings. In these settings, the reward signal is a dense prediction, not a sparse binary outcome.

4. Experimental Design

To approximate continuous, population-level decision making, I evaluate each policy over 30000 horizons per replication, this setup ensures that any initial transients are not over-interpreted and that stable allocation patterns can emerge. I also perform ten replications with independently random seed to summarize variability due to stochastic choice. In each decision round, the simulator draws one record uniformly with replacement from the analytic file and exposes the seven predicted values for that record. The policy selects an arm, receives the selected value as reward, registers the per-round regret as the difference between the best available value and the chosen value, and updates its internal beliefs accordingly.

Performance is summarized using two conventional metrics from the bandit literature [2]. The first is overall average reward, i.e., the average predicted reduction achieved by the chosen arms across the horizon. The second is average regret per round, i.e., the average missed value relative to the best arm for each record.

This design purposefully avoids personalization: records are sampled without conditioning on features, and the policy treats each decision as another draw from the same cohort. In operational terms, the question is, "If I had to choose a single action at scale, which one should it be, and how quickly can I identify it?" The non-personalized framing sidesteps the added modeling choices that accompany contextual bandits while still providing a realistic lens on resource allocation. As a natural extension, contextual bandits—such as linear Thompson Sampling—would enable personalization while following the same evaluation conventions [9].

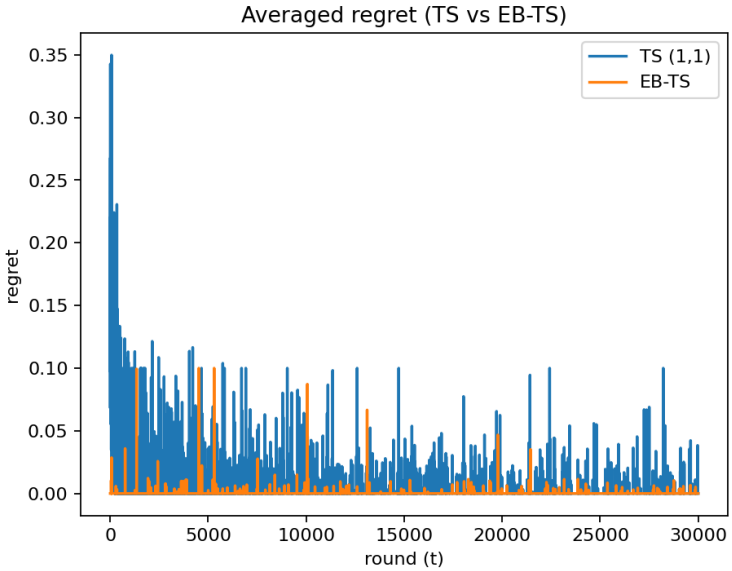

Fig. 1. Comparison of average regret of standard TS method and EB-TS methods

5. Results

The fixed, non-tunable TS policy exhibits rapid and decisive learning. Across ten replications, the overall average reward is 0.2660, and the average regret per round is 0.002065. These numbers indicate that, on average, only about two-thousandths of potential value is left uncollected per decision, a small gap in practical terms given the seven-way choice. As shown in Fig. 1, the allocation pattern explains why: out of 30,000 decisions, the policy allocates roughly 29,700 to the same arm—

air-quality awareness with rescue-medication readiness—while the remaining arms are sampled only sporadically to confirm their inferiority. This behavior is the hallmark of Thompson Sampling in stationary settings with a dominant option: early exploration gives way to long-run exploitation as evidence accumulates [2].

As shown in Fig. 2, the EB-TS variant raises overall average reward to 0.2681 and lowered the mean regret per round to 0.000046. In the beginning, the learning trajectory is smoother. There are fewer uncertain attempts at clearly worse options. This steady start is due to a well-calibrated prior. When the analytic file already shows strong differences between options, it reduces early uncertainty. The policy can then focus on high-yield actions. After enough data, the EB and fixed policies behave similarly. This happens because posterior beliefs are shaped by observed feedback, not initial conditions [2, 4].

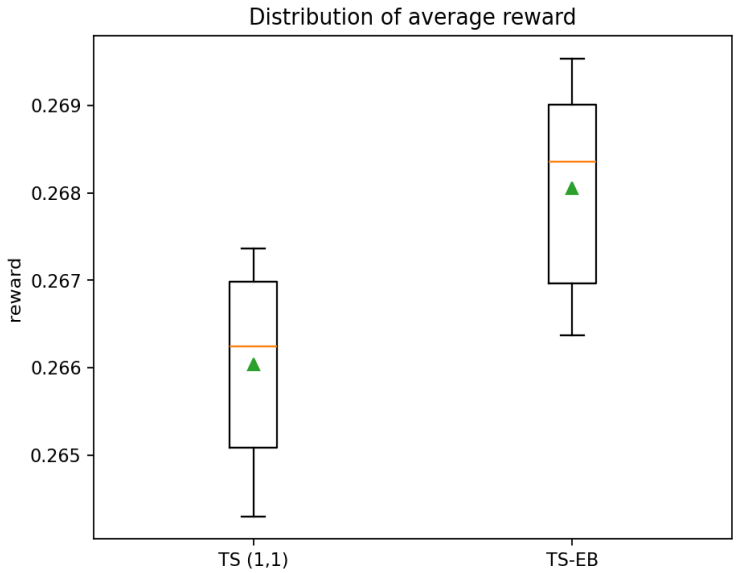

Fig. 2. Comparison of average reward of standard TS method and EB-TS methods

6. Conclusion

These results offer two key insights for public health. First, plain Thompson Sampling provides low regret and effectively targets the best intervention. It does this with a simple uniform prior. This makes it an ideal baseline for departments. They can use it with minimal setup and configuration. Second, when public health programs rely on their offline estimates, using an Empirical-Bayes prior can bring consistent gains. These gains are modest but meaningful. They help stabilize early decisions. While the improvements might seem small, they are valuable. They can lead to significant risk reduction on a larger scale. However, it's important to clarify one point. These findings should not be interpreted as claims about the causal effects of interventions. The preference for air-quality and rescue-readiness is a result of the specific values in the analytic file. Therefore, the bandit's guidance should be viewed as a policy prioritization. It is based on a particular reward structure, not on inherent intervention effectiveness.

References

[1] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 1933, 25(3–4): 285–294.

[2] Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen. A Tutorial on Thompson Sampling. Foundations and Trends in Machine Learning, 2018, 11(1): 1–96.

[3] Shipra Agrawal, Navin Goyal. Near-Optimal Regret Bounds for Thompson Sampling. Journal of the ACM, 2017, 64(5): 1–24.

[4] Steven Tomkins, Pei Liao, Predrag Klasnja, Susan A. Murphy. IntelligentPooling: Practical Thompson Sampling for mHealth. Machine Learning, 2021, 110(9): 2685–2727.

[5] Aleksandrs Slivkins. Introduction to Multi-Armed Bandits. Foundations and Trends® in Machine Learning, 2019, 12(1–2): 1–286.

[6] Centers for Disease Control and Prevention (CDC). Behavioral Risk Factor Surveillance System (BRFSS) 2015: Survey Data and Documentation. Technical report / public dataset, 2016. (Atlanta, GA: U.S. Department of Health and Human Services, CDC.)

[7] Tor Lattimore, Csaba Szepesvári. Bandit Algorithms. Cambridge University Press, 2020.

[8] Olivier Chapelle, Lihong Li. An Empirical Evaluation of Thompson Sampling. Advances in Neural Information Processing Systems, 2011, 24: 2249–2257.

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