Contextual Bandits for Medication Personalization in ICU Patients

Juliane Mercoli - University of Washington, Spring 2025

Introduction

Today, early medication decisions in the ICU are still made under severe uncertainty. This project aims to personalize medication recommendations for ICU patients using contextual bandit algorithms. Specifically, the project recommends medication treatments using initial patient contexts such as vital signs, lab results and demographics collected early in the ICU stay, and optimizing for patient survival and shorter ICU durations (the rewards).
We model the first drug given to each admission as an action and ask: how well could simple contextual bandit policies have performed, had they been deployed instead of the clinicians' choices of first administered drug (logged policy)?

Data and Model Choices

- We use the eICU Collaborative Research Database Demo from PhysioNet (1), which contains 3,614 stays with context. We decided to focus on the 6 first hours after admission as this is when the first drug is usually administered.
- We build the context features vector as a vector of numeric values containing, across the six first hours, age, gender, mean and standard deviation of the temperature, mean and standard deviation of the heart rate, mean and standard deviation of the respiratory rate, mean and standard deviation of the lab results.
- We map the first drug administered as belonging to one of the: fluid (582 cases), analgesic (328 cases), antiemetic (266 cases) group. This first drug administered is the arm.
- We define two rewards model: Survival based with a binary reward model and ICU lenght of stay (LoS) based where the reward is defined as 1/ICU-LoS (days), which is then z-scored so that the reward is > 0 if LoS is inferior to the average duration, and < 0 otherwise.
- The baseline action is the clinician's choice of first administered drug, approximated by a logistic regression model.
- We define three estimators:
  - IPS, which is unbiased when propensities are correct but has high variance
  - SN-IPS, which removes the largest part of that variance by re-normalising the weights
  - DR, which further reduce variance. With limited data and binary (or z-scored) target, complex models would over-fit. We therefore use the constant baseline $\bar{r}$ as the simplest DR learner.

Methods

We approximate the baseline (logging) policy pi0 with a multinomial logistic regression that takes the 10-dimension context and outputs class-posterior probabilities (the propensities).
Logistic is known to be well-calibrated for sparse-multiclass problems, easily interpretable, and fast to fit under cross-validation, thus making it a strong choice to approximate the clinicians' choices of medicine.
Because the three arms are imbalanced (fluid 50%, analgesic 28%, antiemetic 22%), we set class

weight="balanced" so that the loss up-weights minority classes and avoids propensities collapsing towards zero, which is a key requirement for IPS consistency according to multiple studies. Even with balancing, the smallest fitted probability was 3.4×10^{-229} because the clinicians never give certain drugs to certain group of people. However, this would explode the IPS weights as it is defined as $\mathrm{R_{IPS}} = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{p_i}$. So, we clip the probabilities at a threshold ε. With $= 0.1$, the residual bias which is monitored by SN-IPS is below 1%.

Off-policy implementation

1. Estimate propensities for each patient context (the probabilities that the logging policy choose each arm)

2. Simulate the bandits policies on the same context, record the chosen action but update the policy with the logged reward to avoid rewriting factual outcomes

    - We run: UCB1, which is context-free ; LinUCB ($\alpha = 1$) which has a linear pay-off model in 10-dimensions context ; and Thompson Sampling, with v = 1
    - Simulation protocol: Patients are replayed in chronological order (using stream order) Every policy observes the context x_i, selects an action a_i, but updates only on the logged action. We then build a history table containing chosen action, logged action and the propensity (used to undo the clinicians' selection bias) for the estimators

3. Compute the off-policy estimators (IPS and DR)

Results

The overlap between bandits chosen and clinicians' logged actions is about 23% for UCB1, 26% for LinUCB and 34% for Thompson. This is sufficient for stable IPS. Also, the two estimators DR and IPS agree within 5% on both rewards, strengthening our conclusions.

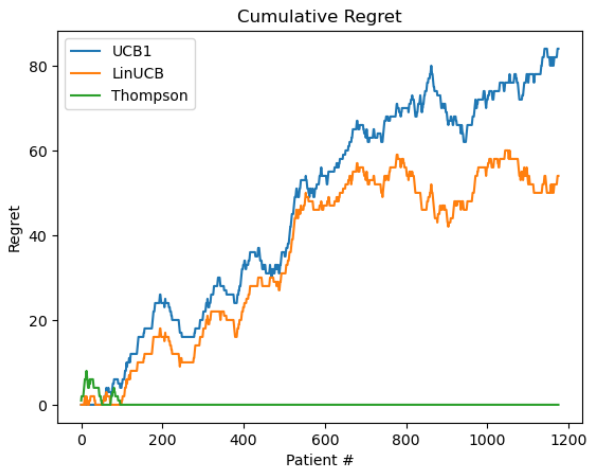

Figure 1: Cumulative regret – survival reward

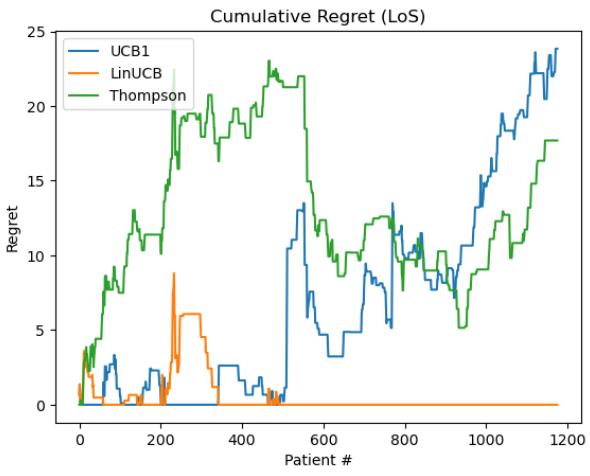

Figure 2: Cumulative regret – LoS reward

Survival reward			
	IPS	SN-IPS	DR
UCB1	2.81	1.00	0.94
LinUCB	5.19	1.00	0.81
Thompson	**6.28**	1.00	**0.99**
LoS reward			
	IPS	SN-IPS	DR
UCB1	0.79	1.00	0.79
LinUCB	**1.12**	1.00	**1.12**
Thompson	0.86	1.00	0.86

- **Survival Model** Thompson sampling leads to a potential 0.99 survival rate against 0.92 with the logged data.
- **Length-of-stay Model** LinUCB wins, cutting the z-scored reward by $+1.12\sigma$; corresponding to 1 hour shorter on average.

Discussion

Assumptions vs. data: The multiclass logistic propensity model ignores clinical subtleties. However, with only three coarse arms it gives plausible probabilities Also, linear reward models are probably misspecified, but they still exploit obvious correlations.

Class imbalance: Survival is highly skewed (92% alive). We therefore set class weight to balanced in the propensity model so that minority classes receive higher weight, mitigating variance in IPS estimates.

Propensity clipping: Without clipping, the minimum estimated probability was 3.4×10^{-229}, leading to absurd IPS ($> 10^{15}$). A clip at 0.1 trades bias for variance and produces more stable numbers while keeping the residual bias low.

Why Thompson dominates survival but not LoS: Survival is nearly deterministic given our arms. Thompson's posterior sampling encourages persistent exploration, yielding more logged-action matches. In LoS, rewards are continuous with substantial noise; LinUCB's explicit uncertainty bonus fits such settings better.

Limitations and Possible Imporvements

This is a limited dataset (demo) that omits a lot of the important signs that are taken into account when making the first drug choice by the clinicians. Patients' context is key for this task and a more complete dataset would be needed to create more representative context vectors. The very small fitted propensities (minimum $\approx 10^{-229}$) are a red flag: many patients receive drugs that our model regards as almost impossible.

Binary survival is coarse and suffers from class imbalance (91.8% survived). We partially mitigated this by reporting self-normalised IPS and DR. Some kind of data augmentation would be needed. Length-of-stay shrinks extreme values by z-scoring after a log transform, but still ignores post-ICU morbidity. A richer composite outcome (for example, days alive and later labs report within 28 days) would be preferable given more data. We could also use stronger propensity models and swap the logistic regression for gradient-boosted trees to enhance the model calibration as IPS / DR are only unbiased if propensities are correct. We could also try different reward models for DR and replace the constant baseline with multitask Random-Forest / XGBoost as a better $\hat{r}$ shrinks the IPS residual and yields DR estimates with lower variance.

References

[1] Pollard *et al.* (2018). *eICU Collaborative Research Database Demo v2.0.1*. PhysioNet. `https://physionet.org/content/eicu-crd-demo/2.0.1/`

[2] Varatharajah, Y. & Berry, B. (2022). A contextual-bandit approach for informed decision-making in clinical trials. *Life*, 12(1277). PMCID: PMC9410371. `https://pmc.ncbi.nlm.nih.gov/articles/PMC9410371/`