Contextual Restless Multi-Armed Bandit (CRMAB) for Dynamic Treatment Plan Optimization in Chronic Disease Management

Chelluri Ashish Dheer
Guided by: Dr. Vidhya Kamakshi

October 7, 2025

Abstract

Chronic diseases such as diabetes and cardiovascular disorders evolve dynamically, requiring adaptive and personalized treatment strategies. Static protocols are not suitable for handling nonlinear progression and variability between patients. We propose a Contextual Restless Multi-Armed Bandit (CR-MAB) framework for dynamic treatment regimes (DTRs), using patient-specific covariates from MIMIC-III to guide intervention under resource constraints. Building on classic RMAB and sequential decision-making principles, we integrate online learning, non-stationary adaptation, and fairness-aware objectives. A prototype simulation with synthetic transitions demonstrates feasibility; future phases will incorporate real clinical rewards, advanced reward engineering (e.g., IRL, LLM-driven elicitation), and deployment considerations such as interpretability and equity. This work bridges RL theory and clinical practice, providing a roadmap to scalable, ethical, and effective chronic care management.

1 Overview

This progress report consolidates the full pipeline design and implementation for the Contextual Restless Multi-Armed Bandit (CRMAB) project. The goal is to convert raw ICU event data into structured, decision-ready sequences that can train and evaluate RMAB-based policies.

The report documents:

- The theoretical background motivating CRMAB for chronic disease management.
- The heart-failure (HF) cohort extraction and preprocessing methodology.
- The trajectory construction process and resulting dataset statistics.
- The integration of these trajectories in RMAB simulation and learning.
- Current progress and produced artifacts.

2 Background and Motivation

Chronic disease management involves long-term, sequential decisions about treatment intensity and timing. Traditional clinical guidelines follow fixed protocols, but such static approaches cannot adapt to patient-specific dynamics or evolving physiological responses.

Restless Multi-Armed Bandits (RMABs) generalize the multi-armed bandit problem by allowing arms to transition state even when not selected; this property makes RMABs suitable for modeling patients whose conditions change continuously over time [3, 5]. Contextual RMABs additionally integrate patient covariates (demographics, vitals, labs) to personalize interventions.

The importance of fairness and equitable allocation has been emphasized in recent RMAB work; without fairness-aware objectives, naive index policies can result in inequitable treatment allocations [2, 4]. Previous applied works demonstrate that RMAB frameworks can inform public-health interventions and preventive healthcare decision-making [1, 3].

3 Implementation

The implementation in `crmab.py` forms an end-to-end data engineering and preprocessing pipeline.

Cohort Selection and HF Extraction: Patients with ICD-9 codes beginning with '428' were identified as part of the heart-failure cohort. Admissions were merged with demographic and ICU stay tables; ages were top-coded at 90 years for privacy.

ITEMID Detection and Measurement Extraction: All ITEMIDs associated with heart rate, systolic blood pressure, and creatinine were detected using dictionary keyword search over `D_ITEMS` and `D_LABITEMS`. Data are read in chunks for memory efficiency, and the latest non-null measurement per (subject, hadm, itemid) is retained.

Unit Normalization and Quality Assurance: Unit inconsistencies (e.g., creatinine in µmol/L vs mg/dL) are detected and standardized to a canonical unit using conversion heuristics. Measurements outside physiological plausibility bounds are masked and logged for QC.

Aggregation and Imputation: Multiple ITEMID-derived values are collapsed using the median per admission. Missing values are filled with a cascading fallback: admission-level → closest-in-time (within configurable window) → patient-latest → cohort median. The source of each imputed value is recorded for traceability.

Action Mapping and Deduplication: We map raw events from `PRESCRIPTIONS`, `INPUTEVENTS_MV`, `INPUTEVENTS_CV`, and `PROCEDUREEVENTS` to a concise action set: `no_action`, `vasopressor`, `fluid_bolus`, `diuretic`, `antibiotic`, `insulin`, and `other`. Text-based regex rules, route and unit heuristics, and time-range explosion are used; conflicting labels per timestep are resolved using a deterministic priority mapping.

Trajectory Construction: Each admission is discretized into 6-hour decision epochs (configurable). For each epoch we assemble the state vector (vitals, labs, demographics) and the corresponding mapped action (single label per epoch). The canonical output is `traj_with_mapped_actions.csv`.

4 HF Cohort: Extraction and Processing

The HF cohort was created by filtering `DIAGNOSES_ICD` for ICD-9 prefix '428' and merging with `PATIENTS` and `ADMISSIONS`. Measurement extraction for vitals and labs followed the ITEMID and unit-normalization pipeline.

Representative diagnostic counts from a sample pipeline run:

- Heart rate and SBP chart rows: 466
- Creatinine lab rows: 13,865
- Total HF admissions extracted: 125

These processed features feed directly into the per-epoch state vectors.

5 Trajectory Construction (TRAJ)

The trajectory dataset represents each admission as a sequence of (s_t, a_t) pairs.

Design choices:

- Epoch length: 6 hours (configurable).
- State features: heart rate, systolic BP, creatinine, plus static covariates.
- Action space: seven discrete categories as described above.

Each row in `traj_with_mapped_actions.csv` includes identifiers (`subject_id`, `hadm_id`), temporal indices (`timestep`, `time_since_admit_hours`), the state vector, and the mapped action.

Representative dataset statistics:

Metric	Value
Admissions processed	∼753
State-action pairs	∼6,000
Fluid bolus (%)	53%
No action (%)	34%
Insulin (%)	12%
Other / antibiotic (%)	<1%

Table 1: Representative trajectory dataset statistics (sample run).

6 Purpose of `traj_with_mapped_actions.csv` and Its Use

This trajectory file is the canonical input for downstream RMAB modeling and evaluation.

Contents (per row):

- Identifiers: `subject_id`, `hadm_id`, `timestep`, `time_since_admit_hours`.
- State features: `heart_rate`, `sys_bp`, `creatinine`, static covariates.
- Action fields: `mapped_action`, `action_code`.

Primary downstream uses:

1. Reward engineering (e.g., relative SBP improvement).
2. Transition model estimation $P(s_{t+1} \mid s_t, a_t)$.
3. Index/policy learning (Whittle-style indices; learned RL policies).
4. Offline policy evaluation (importance sampling, doubly-robust methods).

Example reward:

$$r_t = \frac{\mathrm{SBP}_{t+1} - \mathrm{SBP}_t}{\mathrm{SBP}_t}.$$

7 Action Mapping and Validation

Action mapping condenses heterogeneous event logs into interpretable treatment classes. Mappings are audited via the raw and deduplicated action files (`actions_raw.csv`, `actions_top.csv`) and sampled CV/MV joins are exported for clinician review.

8 Generated files

- `merged_tidy.csv`, `df_keep.csv` (cohort snapshots)
- `actions_raw.csv`, `actions_top.csv` (mapping outputs)
- `traj_with_mapped_actions.csv`, `traj_mv_with_actions.csv` (trajectories)
- Diagnostic logs and audit samples

References

[1] Arpita Biswas, Gaurav Aggarwal, Pradeep Varakantham, and Milind Tambe. Learn to intervene: An adaptive learning policy for restless bandits in application to preventive healthcare. *arXiv preprint arXiv:2105.07965*, 2021. Available at `https://arxiv.org/abs/2105.07965`.

[2] Jackson Killian, Manish Jain, Yugang Jia, and Jonathan Amar. Equitable restless multi-armed bandits: A general framework inspired by digital health. *arXiv preprint arXiv:2308.09726*, 2023. Available at `https://arxiv.org/abs/2308.09726`.

[3] Aditya Mate, Andrew Perrault, and Milind Tambe. Risk-aware interventions in public health: Planning with restless multi-armed bandits. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2021.

[4] A. Sood and et al. Fairness of exposure in online restless multi-armed bandits. In *Proceedings of ACM (conference)*, 2024.

[5] Kai Wang, Shresth Verma, Aditya Mate, Sanket Shah, and Aparna Taneja. Scalable decision-focused learning in restless multi-armed bandits with application to maternal and child health. *arXiv preprint arXiv:2202.00916*, 2022. Available at `https://arxiv.org/abs/2202.00916`.