Reinforcement Learning for Antibiotic Stewardship: Optimizing Prescribing Policies Under Antimicrobial Resistance Dynamics

Joyce Lee Seth Blumberg

2026-03-12

Contents

1 Abstract **2**

2 Introduction **3**

3 Materials and Methods **4**

3.1 Reinforcement Learning Framework 4

3.2 Agent-environment interaction loop 4

3.3 Environment subcomponents 6

3.3.1 Patient Population and Observability 6

3.3.2 Antibiotic Resistance Dynamics 6

3.3.3 Treatment Decisions 6

3.3.4 Reward Function 6

3.4 Reinforcement learning agents 7

3.5 Experimental Pipeline 7

3.6 Fixed Prescribing Rules 8

3.7 Experiments: varying partial observability of patient attributes and AMR levels 9

4 Results **11**

4.1 Value Iteration Optimal Policy (benchmark for Experiment Sets #1 - #3) 11

4.2 Experiment Set #1: Perfect Observability 12

4.3 Experiment Set #2: Delayed, Noisy, Biased AMR 16

4.4 Experiment Set #3: Heterogeneous Patient Attributes, Varied Bias 18

4.5 Experiment Set #4: Combined Observational Noise, Patient Heterogeneity 20

5 Discussion **22**
5.1 Individual Experiment Set Findings . . . 22
5.2 Cross-Experiment Set Findings . . . 23
5.3 Finite-Horizon Effects and Exploitation by Hierarchical Agents . . . 24
5.4 Limitations . . . 24
5.4.1 Abstraction of pathogen identity . . . 24
5.4.2 Stationarity assumptions . . . 25
5.4.3 Single centralized prescriber . . . 25
5.4.4 Reinforcement learning algorithm selection . . . 25
5.4.5 AMR resistance dynamic assumptions . . . 25
5.4.6 Finite-horizon artifact in hierarchical agent training . . . 26
5.4.7 Scope of findings and generalizability . . . 26
5.5 Future Work . . . 26
5.5.1 Removing timestep awareness from hierarchical agents . . . 26
5.5.2 Multi-agent, multi-locale environments . . . 26
5.5.3 Non-stationary dynamics . . . 27
5.5.4 Upper bounds on personalized antibiogram prediction . . . 27
5.6 Conclusions . . . 27

References **28**

1 Abstract

Antimicrobial resistance (AMR) threatens antibiotic effectiveness, but quantitatively evaluating stewardship strategies under partial observability and delayed feedback remains difficult in real-world data. We developed `abx_amr_simulator`, a Gymnasium-compatible simulation framework, and used it to benchmark reinforcement learning (RL) prescribing policies against value-iteration (VI) benchmarks and fixed prescribing rules across four experiment sets of increasing complexity, with varying levels of information degradation.

Across scenarios, temporal abstraction was consistently important: flat PPO was competitive only in simpler settings, whereas hierarchical PPO was generally needed when prescribing decisions had delayed, coupled effects on future resistance. We found that adding recurrent memory did not uniformly improve performance; its value was context-dependent. In some degraded-information settings, memoryless policies performed better by adopting conservative update-responsive behavior, while in more complex, multi-signal partially observable settings, recurrent memory provided modest advantages.

Patient heterogeneity and risk-stratification signals were major determinants of policy quality. When agents could differentiate higher- from lower-risk patients, they more reliably learned selective treatment behavior, stabilized AMR, and improved clinical outcomes. Exaggerated risk stratification modestly outperformed accurate stratification, while compressed stratification produced moderate degradation. In more realistic settings combining noisy patient observations, delayed AMR surveillance, and multi-patient decisions, hierarchical agents outperformed fixed prescribing rules across both stewardship and clinical metrics, converging to conservative low-AMR equilibria with reduced cross-seed variance.

Across experiments, results support the utility of hierarchical RL as a best-case policy-analysis tool for stewardship under uncertainty, while also highlighting that performance estimates are sensitive to observation

structure and training horizon design. The framework provides a controlled environment for hypothesis generation and for stress-testing prescribing strategies before translation to policy-relevant settings.

2 Introduction

Rising antimicrobial resistance (AMR) is widely recognized as a major threat to global public health, adversely affecting the clinical care and outcomes of millions of patients worldwide by reducing the effectiveness of currently available antibiotics (WHO 2015). Recent estimates suggest that 4.95 million deaths globally in 2019 were associated with antibiotic-resistant bacterial infections, with a disproportionate burden borne by those living in resource-limited settings (Murray et al. 2022). In response, a range of interventions have been proposed and implemented to curb the emergence and spread of AMR (America (IDSA) 2011). In the United States, one of the primary strategies has been the establishment of antibiotic stewardship programs (ASPs) (K. F. Sutton and Ashley 2024), which are multidisciplinary efforts encompassing clinician education, dissemination of prescribing guidelines, compilation of antibiograms from microbiological data, and prospective review and optimization of antibiotic regimens, particularly in inpatient settings (Barlam et al. 2016).

Despite the widespread adoption of such interventions, the quantitative evidence base evaluating their population-level impact remains limited. Many studies assessing the effectiveness of ASPs or related strategies have been constrained in geographic scope, duration, or both (Bertollo, Lutkemeyer, and Levin 2018). A fundamental challenge to performing quantitative impact evaluation of these programs is the pervasive partial observability of key system components (Laxminarayan et al. 2013). Even in resource-rich settings, where antibiotic prescription records are often available, true selection pressure on pathogen populations is incompletely observed due to unmeasured sources of antibiotic exposure, including agricultural use and environmental contamination (Van Boeckel et al. 2015). Measurement of AMR itself presents an even greater challenge: while initiatives such as the WHO's Global Antimicrobial Resistance and Use Surveillance System (GLASS) have sought to standardize surveillance, participation remains voluntary and coverage incomplete, and no universally adopted scalar metric of resistance at the community level yet exists (Organization 2022; Leth and Schultsz 2023). One commonly used proxy, the antibiogram, is subject to substantial biases, as it usually reflects only cultured clinical isolates from patients seeking care and is typically updated infrequently, often on an annual basis (Truong et al. 2021). Consequently, antibiograms provide an incomplete and temporally lagged view of underlying resistance dynamics.

Given these challenges, it is difficult to directly quantify the long-term effects of antibiotic stewardship program interventions using observational or interventional studies alone (Bertollo, Lutkemeyer, and Levin 2018; Schweitzer et al. 2019). Simulation-based approaches offer a complementary framework in which key mechanisms can be explicitly specified and systematically varied, enabling controlled investigation of trade-offs that are difficult or impossible to observe in real-world settings.

Prior simulation-based work in this domain has largely operated at the pathogen level, modeling evolutionary dynamics of resistance under drug pressure (King et al. 2025; Weaver et al. 2024). While these models have yielded important insights into resistance evolution, they do not capture the patient-level clinical dynamics — including treatment decisions, outcomes, and feedback between prescribing behavior and population-level resistance — that are central to evaluating antibiotic stewardship interventions. Separately, there has been growing interest in applying machine learning to AMR-related problems, including resistance prediction and empirical antibiotic selection (Haredasht et al. 2025; Harandi et al. 2025). However, these approaches predominantly utilize supervised learning to focus on prediction rather than optimization, and typically operate on static snapshots of patient data rather than modeling the dynamic, longitudinal consequences of prescribing decisions at the population level. To our knowledge, the only prior application of reinforcement learning to an AMR-related problem is Weaver et al. (Weaver et al. 2024), which uses RL to identify optimal treatment cycling strategies at the pathogen level. Our work differs fundamentally in that we apply RL to optimize clinical prescribing policy within a patient-level simulator, with the explicit goal of evaluating trade-offs between individual clinical outcomes and population-level resistance.

In this study, we conduct a series of reinforcement learning experiments using a simulation framework to investigate antibiotic prescribing strategies under varying levels of observability and uncertainty. The experiments were performed using `abx_amr_simulator`, a Python-based simulation environment designed to model the relationship between prescribing decisions and the evolution of AMR (Lee 2026). Here we focus on its application to experimental settings.

Using the simulator, we explored the trade-off between immediate clinical benefit and long-term resistance management. In simplified, fully observable settings, the exact optimal prescribing policy can be derived using dynamic programming techniques. However, as uncertainty and delayed effects are introduced, deriving optimal solutions becomes intractable, suggesting the need for more flexible approaches such as reinforcement learning (RL). We hypothesize that RL algorithms are capable of identifying near-optimal prescribing policies under conditions of partial observability, even as the degree of observability varies. By training different types of RL agents across a range of simulated scenarios, we perform upper-bound policy analyses to characterize the magnitude and structure of potential gains achievable through improved prescribing strategies.

In the following sections, we describe our experimental design, including how we simulate synthetic patient populations, define the reward function, and vary the quality and timeliness of information available to the RL prescribing agent.

3 Materials and Methods

3.1 Reinforcement Learning Framework

Optimizing antibiotic prescribing is a sequential decision-making problem under uncertainty: prescribing decisions affect both immediate patient outcomes and longer-term resistance dynamics, and the true underlying state of the system is rarely fully observable. We address this using reinforcement learning (RL), a class of machine learning methods designed to maximize cumulative rewards in such settings. The experiments were conducted using `abx_amr_simulator`, a Python-based simulation framework designed to model antibiotic prescribing decisions and antimicrobial resistance (AMR) dynamics, built on the standard Gymnasium API. A detailed description of the simulator's architecture, functionality, and configuration options is available in the `abx_amr_simulator` GitHub documentation (Lee 2026). This section provides a brief overview of key components and experimental settings relevant to the study.

3.2 Agent-environment interaction loop

In reinforcement learning, the way that agents learn to optimize sequential decision making is by interacting with the environment. In this work, the environment is created by `abx_amr_simulator` according to the user's specifications in order to construct a specific simulated scenario. The environment consists of three main subcomponents: a `PatientGenerator`, an array of `AMR_LeakyBalloon` models that keep track of each antibiotic's resistance level, and a `RewardCalculator` which computes the reward for the agent's action at each timestep based on patient outcomes. (For more detailed information on the `abx_amr_simulator` package and the subcomponents of the environment class, see S2 in Supplementary Materials, also documentation in Github repository.)

For the reinforcement learning agents used in our experiments, we used PPO (proximal policy optimization) agent implementations from the Python package `stable-baselines3`. Each of the subcomponents of the environment is discussed further below, as are the various types of RL agents used in our experiments.

A schematic overview of the simulator components and the agent–environment interaction loop is shown in Figure 1.

Note: this figure is reproduced from the `abx_amr_simulator` package documentation (Lee 2026).

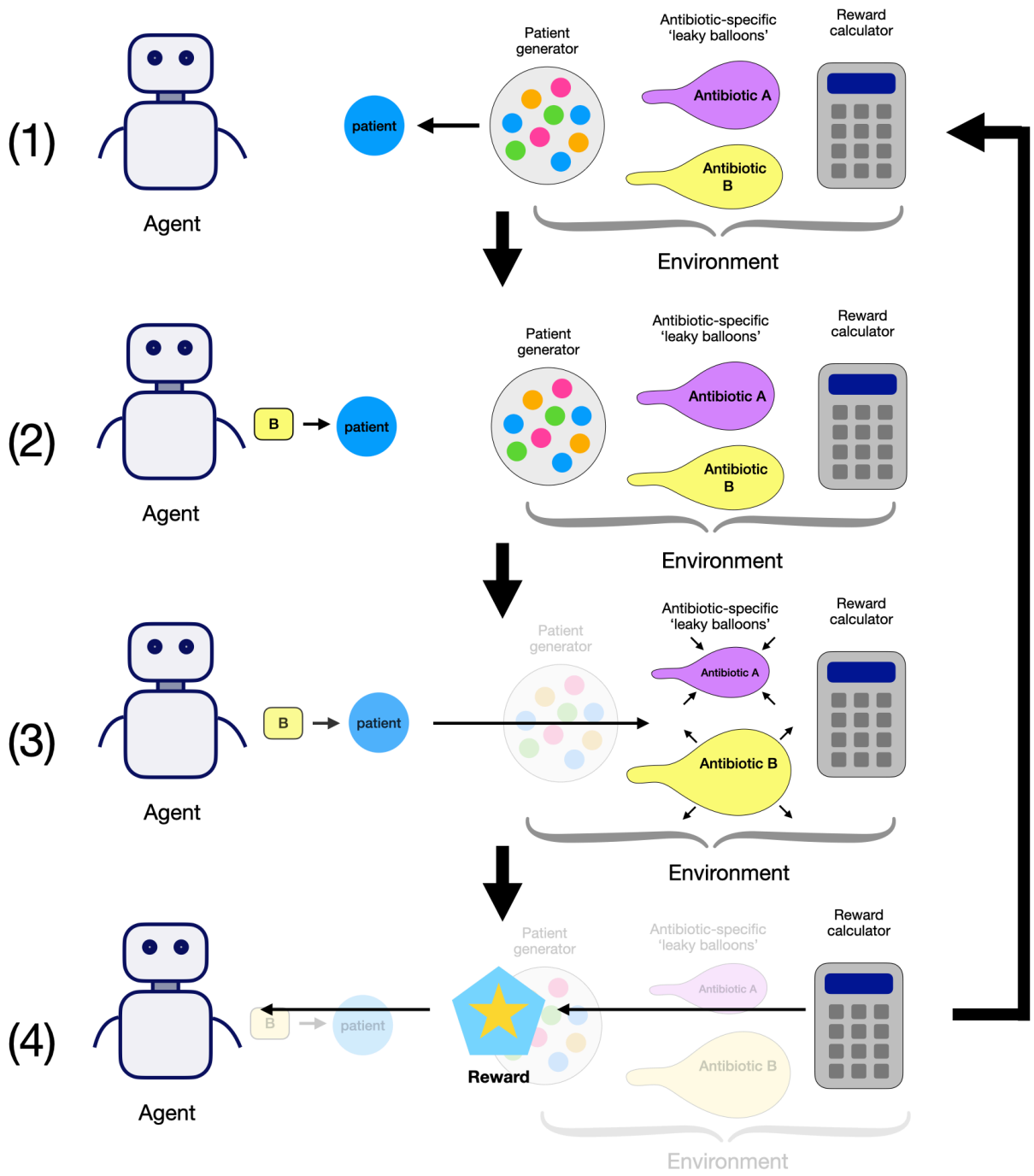


Figure 1: Schematic overview of the environment and its subcomponents (patient generator, antibiotic leaky balloons, and reward calculator), and the interaction loop between the agent and environment. 1) The environment uses the PatientGenerator to create a patient and present that patient to the agent. 2) The agent decides to prescribe Antibiotic B for the patient. 3) The agent informs the environment that it decided to prescribe Antibiotic B for the current patient. 4) The environment uses the RewardCalculator to compute the reward for the current action, updates each antibiotic's AMR LeakyBalloon, and advances to the next timestep and next patient.

3.3 Environment subcomponents

3.3.1 Patient Population and Observability

At each timestep, the environment uses the `PatientGenerator` subcomponent to generate a cohort of synthetic patients, which are characterized by attributes that determine their clinical profile and treatment outcomes. These attributes include infection probability, clinical benefit and failure multipliers, and spontaneous recovery probability. Attribute values can be set to constant values (creating homogenous patient populations), or drawn from a probability distribution (creating heterogeneous populations). Observations available to the RL agent can be manipulated by introducing noise or bias, or by concealing certain attributes altogether, therefore enabling exploration of prescribing strategies under varying levels of partial observability.

3.3.2 Antibiotic Resistance Dynamics

Antimicrobial resistance dynamics are modeled using the `AMR_LeakyBalloon` class, which tracks resistance to each antibiotic as a function of cumulative prescribing pressure. The model operates as a soft-bounded accumulator: prescribing a given antibiotic increases the internal latent 'pressure', while in the absence of prescribing, this internal pressure decays over time as selection pressure is relieved — analogous to a balloon that inflates as air is pumped in, but slowly deflates when left alone. The observable resistance level is derived from this latent pressure via mapping through a sigmoid function, and represents the probability that a new infection at the current time is resistant to the given antibiotic. The simulator also supports cross-resistance, allowing users to specify how prescribing one antibiotic influences resistance to others.

As with patient attributes, the fidelity of resistance level observations can be manipulated by introducing noise, bias, or delay, facilitating systematic investigation of prescribing strategies under imperfect surveillance.

3.3.3 Treatment Decisions

At each timestep, the agent makes an independent treatment decision for each patient: prescribe antibiotic A, prescribe antibiotic B, ..., or withhold treatment. The agent cannot prescribe multiple antibiotics to the same patient simultaneously, reflecting standard clinical practice of monotherapy (single-drug treatment) for straightforward infections. This constraint simplifies the decision space and aligns with real-world prescribing norms.

3.3.4 Reward Function

The reward function as calculated by the `RewardCalculator` balances immediate clinical outcomes against long-term antimicrobial resistance considerations. It consists of two components: the individual reward, which reflects the clinical benefit or failure of treating patients, and the community reward, which penalizes high levels of antimicrobial resistance (AMR) across all antibiotics. These components are combined into a single scalar reward using a tunable weight parameter $\lambda \in [0, 1]$, allowing the user to control the relative importance of individual versus community objectives. In all experiments reported in this manuscript, we fixed $\lambda = 0$, so agents were optimized using only individual clinical reward. This design allows us to test whether AMR-preserving behavior can emerge from long-horizon environment dynamics and policy architecture without explicit AMR reward shaping.

The individual reward is calculated based on patient-specific attributes, antibiotic-specific properties, and discrete clinical outcomes (e.g., treatment success, failure, or adverse effects). The community reward, on the other hand, is based on the observed AMR levels of all antibiotics, ensuring consistency with realistic surveillance data rather than ground truth resistance levels. This design encourages agents to learn prescribing strategies that work within realistic informational constraints.

For mathematical definitions of the reward function, including equations, parameter descriptions, and details on stochastic modeling of clinical outcomes, see Supplementary Materials S2.

3.4 Reinforcement learning agents

To evaluate how policy architecture affects antibiotic prescribing under uncertainty, we implemented four PPO-based agent classes using the `stable-baselines3` Python library: flat memoryless PPO, flat recurrent PPO, hierarchical memoryless PPO, and hierarchical recurrent PPO.

The memoryless agents make each decision from the current observation only. By contrast, recurrent agents carry a short internal memory across timesteps, so recent observation history can influence the current decision. This distinction is important when AMR information is delayed or intermittently updated: agents that retain recent history may better approximate the latent current resistance state.

We also compared flat and hierarchical decision structures. Flat agents choose a treatment action directly at each timestep. Hierarchical agents instead choose among higher-level options (also called workers), where each option encodes a clinically interpretable prescribing strategy over one or more steps. In our experiments, we used two different option types: deterministic-sequence options, which followed fixed prescribing patterns, and heuristic-worker options, which used fixed risk-based rules to choose treatment for each patient. Our use of hierarchical agents was motivated by the long-horizon nature of AMR dynamics, in which current prescribing can affect resistance and treatment utility many timesteps later. (For more information, please see Supplementary Materials S3: Options and Option Libraries, which discusses the option types and instances that were used in these experiments.)

We additionally performed preliminary implementation and exploration of other long-horizon approaches (including temporal feature augmentation and an MBPO-style approach). In early testing, hierarchical PPO showed the most promising and robust behavior for our simulator settings, and was therefore prioritized for the experiments reported in this manuscript.

3.5 Experimental Pipeline

Effective reinforcement learning requires careful hyperparameter tuning. The `abx_amr_simulator` package uses the `Optuna` library (a hyperparameter optimization framework) for automated hyperparameter tuning (Akiba et al. 2019). This allows us to build modular end-to-end pipelines with three phases:

Tuning Phase: Optuna explores the hyperparameter search space, evaluating candidate configurations on preliminary environments. For each trial, the agent undergoes truncated training across 5 random seeds, with the study ultimately selecting the hyperparameter set that maximized a stability-adjusted objective (mean reward minus a modest variance penalty across seeds).

Training Phase: Multiple agent copies (20 per experiment) are initialized with optimized hyperparameters and trained on target environments, each using a distinct random seed. This approach accounts for the empirical sensitivity of learned policies to initialization: we observed for certain scenarios and agent architectures, environmental stochasticity (particularly patient infection and resistance status across seeds) substantially influences policy acquisition, even with identical hyperparameters.

Evaluation Phase: Following training, each seed-deployed agent executes 3 evaluative rollouts with logging enabled. During evaluation, we record three categories of cumulative metrics:

- *Clinical Outcomes*: patients experiencing benefit, failure, or adverse effects
- *Patient Status*: counts by true infection status and treatment received
- *Antibiotic Response*: prescriptions per antibiotic, and sensitive vs. resistant infections treated

These metrics are agnostic to the decision algorithm—applicable to RL agents, fixed prescribing rules, or any strategy—enabling quantitative policy comparison. When the reward function is held constant while

evaluating different policies, raw reward provides a valid shorthand for relative performance. However, examining cumulative count metrics remain essential for interpretation, since they show which outcome components (for example, clinical successes, failures, and adverse events) are driving those reward differences. All experiments in this study used the same reward function with fixed `RewardCalculator` hyperparameter values.

Figure 2 shows a schematic of the full experimental pipeline.

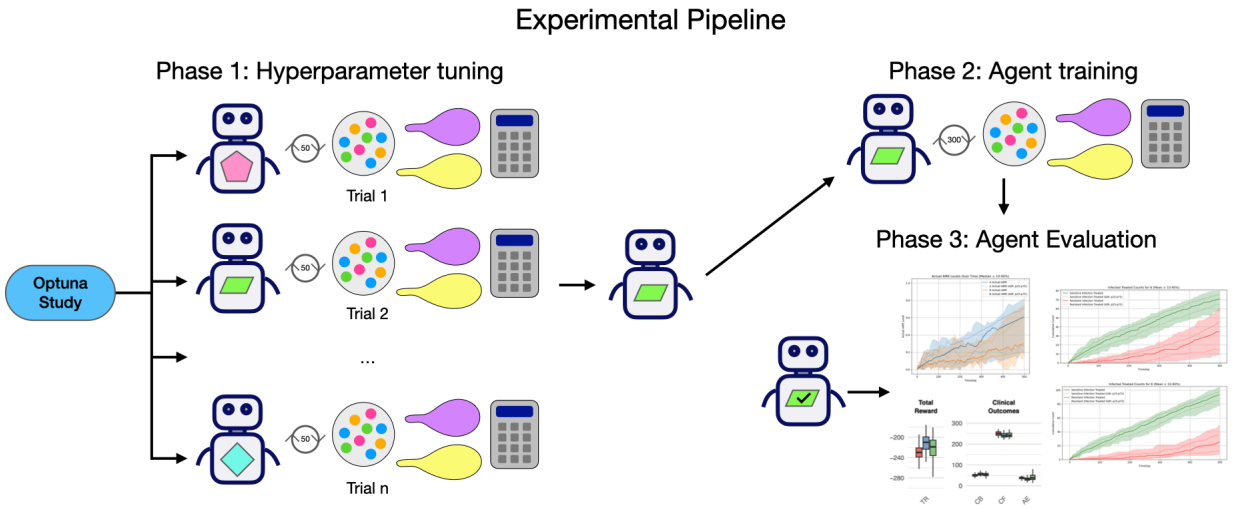


Figure 2: The experimental pipeline consists of three phases. Phase 1 (Tuning): An Optuna hyperparameter search evaluates candidate configurations via truncated training across 5 random seeds, selecting the configuration that maximizes a stability-adjusted objective (mean reward minus a modest variance penalty across seeds). Phase 2 (Training): Twenty agent copies are initialized with the tuned hyperparameters and undergo prolonged training, each using a distinct random seed. Phase 3 (Evaluation): Each trained agent executes three evaluative rollouts in the environment across unseen random seeds; clinical outcome, patient status, and antibiotic response metrics are logged and aggregated for analysis.

3.6 Fixed Prescribing Rules

To evaluate whether learned policies provide value beyond static decision rules, we compared trained RL agents against two fixed prescribing baselines that perform no learning or adaptation.

Expected-reward (lowest AMR): For each patient, compute the expected net clinical benefit (expected reward) of treating with each available antibiotic based on observed patient attributes and known scenario parameters. If no antibiotic has positive expected reward, withhold treatment. Otherwise, prescribe the antibiotic with the lowest currently visible AMR level.

Expected-reward (greedy): For each patient, compute the same expected reward as described in the prior rule. If no antibiotic has positive expected reward, withhold treatment. Otherwise, prescribe the antibiotic with the highest expected reward.

These rules are intended to represent clinically interpretable policies, akin to static practice guidelines: they use available patient information and visible AMR data, but do not update behavior through experience. This makes them suitable non-learning comparators for isolating the incremental value of RL-based adaptation.

For fair comparison, fixed-rule policies are evaluated under the same rollout framework used for trained agents (multiple seeds and repeated evaluation episodes per seed), so differences reflect policy behavior rather than evaluation protocol artifacts.

3.7 Experiments: varying partial observability of patient attributes and AMR levels

Using the `abx_amr_simulator` package, we evaluated policies across three different antibiotic scenarios of increasing complexity: (1) a single-antibiotic scenario, (2) a two-antibiotic scenario without cross-resistance, and (3) a two-antibiotic scenario with moderate asymmetric cross-resistance.

For each of these antibiotic scenarios, we varied patient-population structure and information quality across four sets of experiments (Table 1). Populations were either homogeneous (fixed patient attribute values) or heterogeneous (attribute values sampled from Gaussian distributions), including mixed high-risk/low-risk cohorts in the heterogeneous settings. Importantly, the reward function was held constant across all experiment sets and antibiotic scenarios (with fixed `RewardCalculator` hyperparameter values), so performance differences reflect policy behavior under different observability and population conditions rather than changes in reward design.

We now discuss the benchmark scenario construction and value-iteration (VI) reference policies, then describe the four experiment sets that progressively degrade observability and increase heterogeneity.

Table 1: Overview of experimental conditions across four experiment sets. Observability, agent architectures, and comparators were systematically varied. All sets evaluated three antibiotic scenarios: single-antibiotic, two-antibiotic without cross-resistance, and two-antibiotic with cross-resistance.

	Experiment Set	**Patient Observability**	**AMR Observability**	**Patients / Timestep**	**Agent Architectures Evaluated**	**Hierarchical Agent Worker Type**	**Number of RL Experiments**	**Comparators**
1	Perfect Observability	Perfect	Perfect	1	Flat PPO; Hierarchical PPO	Deterministic Sequences	6	Value iteration-derived policy benchmark
2	Delayed, Noisy, Biased AMR	Perfect	Delayed + noisy + biased	1	Flat PPO; Flat Recurrent PPO; Hierarchical PPO; Hierarchical Recurrent PPO	Deterministic Sequences	12	Value iteration-derived policy benchmark
3	Heterogeneous Patient Attributes, Varied Bias	Biased observations	Perfect	1	Hierarchical PPO	Heuristic Workers	9	Value iteration-derived policy benchmark
4	Combined Observational Noise, Patient Heterogeneity	Noisy + biased + mixed visibility	Delayed + noisy	10	Hierarchical PPO; Hierarchical Recurrent PPO	Heuristic Workers	6	Fixed prescribing rules (Greedy, Lowest AMR)

Benchmark scenario design and value iteration-derived reference policies: In designing the three different antibiotic scenarios, we selected environment parameter values in order to produce nontrivial stewardship dynamics (i.e., situations that were required policies with meaningful trade-offs, rather than trivial always-treat or never-treat behavior). In particular, for the homogeneous benchmark population we fixed baseline infection probability at $p = 0.7$, so treatment withholding is not trivially optimal and agents must balance near-term clinical benefit against longer-term resistance consequences.

In order to compute the optimal policy for each of our three antibiotic scenarios, we turned to a dynamic programming. Our benchmark scenarios, where the state is fully observable and tractable (homogeneous patient population, small number of antibiotics), can be formulated as Markov Decision Processes (MDP), which means it is possible to compute optimal policies based on expected reward using value iteration (hereafter referred to as VI), which is a dynamic programming algorithm (R. S. Sutton and Barto 2018). We discretized AMR state variables, precomputed transition and reward terms from the leaky-balloon dynamics, and iteratively solved the Bellman updates to obtain optimal policy tables. Full implementation details, including state-space construction for multi-antibiotic scenarios, are provided in Supplementary Materials S1.

VI-derived policies serve as reference benchmarks for Experiment Sets 1–3. To preserve comparability with those VI benchmarks, Experiment Sets 1–3 use single-patient-per-timestep configurations. Experiment Set 4 increases patient volume to 10 patients per timestep and adds further complexity with differential partial observability of patients, and is therefore excluded from VI comparison due to state-action intractability; fixed prescribing rules are used as comparators instead (Section 3.6).

Experiment Set 1: Perfect Observability:

We first conducted a set of experiments in which the agent had full observability of the environment: all patient attributes and antibiotic-specific AMR levels were presented to the agent without noise, bias, or delay. We used the same homogeneous baseline patient population defined in the Value Iteration Benchmark section. These experiments serve as a best-case baseline, establishing whether reinforcement learning agents can recover near-optimal prescribing policies when the true state of the system is fully known. We tested both flat and hierarchical PPO architectures; the rationale for the hierarchical architecture and the option designs used are described in Section 3.4.

Experiment Set 2: Delayed, Noisy, Biased AMR:

In the second set of experiments, we introduced partial observability in the form of noise, bias, and temporal delay in the AMR levels visible to the agent, while maintaining perfect observability of patient attributes. We again used the same homogeneous baseline patient population defined in the Value Iteration Benchmark section (and used in Experiment Set 1). This type of degraded information regarding AMR levels does resemble real-world clinical practice: as discussed in the Introduction, antibiograms provide an incomplete and typically delayed proxy for community resistance levels, requiring clinicians to make prescribing decisions using outdated and potentially biased information.

Introducing delayed updates to AMR observations renders the environment partially observable, as the agent now only has periodic access to updated resistance levels; otherwise, during inter-update periods, this information is stale. For this reason, in addition to memoryless PPO agents, we evaluated recurrent flat and hierarchical PPO agents with LSTM-based policies, which are commonly used to handle temporal partial observability by allowing agents to integrate information over time. Intuitively, such agents may learn to maintain an internal estimate of latent resistance dynamics based on prior observations and prescribing outcomes, analogous to a clinician forming informal beliefs from accumulated experience.

Experiment Set 3: Heterogeneous Patient Attributes, Varied Bias:

In Experiment Set 3, we introduced heterogeneous patient populations with true differences in infection risk and treatment response while maintaining population-level averages matching the homogeneous baseline (Experiment Set 1). Patients were drawn from two subpopulations (high-risk and low-risk) with equal probability, where infection probability followed Gaussian distributions ($\mu = 0.84$ vs. $\mu = 0.56$, $\sigma = 0.1$) rather than constant values. This creates ground-truth risk stratification that optimal policies should exploit.

We then manipulated observational bias in how agents perceive these patient attributes, testing three conditions for each antibiotic scenario (see Table 2): (1) Accurate: true risk levels correctly observed, (2) Exaggerated: high-risk patients appear even higher risk, low-risk appear even lower, and (3) Compressed: risk differences appear smaller than they truly are. This isolates the cost of misperceiving patient heterogeneity when AMR information remains perfect and up-to-date.

All experiments in this set employed HRL PPO agents with heuristic workers. While Experiment Sets 1 and 2 used deterministic workers (precomputed multi-step regimens with fixed antibiotic prescribing sequences), these are unsuitable for heterogeneous populations where no single fixed sequence is optimal across all individuals. Heuristic workers use the observable patient attributes to compute an expected 'reward' for each available action, then uses rule-based strategies to select which action to take (e.g., "treat high-risk patients with antibiotic A," "withhold from low-risk patients"). The manager then learns to select among these context-appropriate options based on patient characteristics and AMR state.

Unlike the fixed prescribing rules described in Section 3.6, heuristic workers serve as options within a learning architecture—the manager learns to dynamically select which worker to deploy based on AMR state and population context. Heuristic workers themselves encode deterministic clinical logic (similar to the fixed rules), but the key difference is that the hierarchical agent learns when to switch between them, enabling adaptive cycling strategies that fixed rules cannot achieve.

Experiment Set 4: Combined Observational Noise, Patient Heterogeneity:

In the fourth and final set of experiments, we combine all sources of uncertainty from Experiment Sets 2 and 3: agents face noisy/biased patient observations alongside noisy/biased/delayed AMR surveillance.

Table 2: Experimental conditions for perceived vs. true patient heterogeneity in Experiment Set 3. Each of these three risk stratification conditions was tested across all three antibiotic scenarios (single-antibiotic, two-antibiotic without cross-resistance, two-antibiotic with cross-resistance). All experiments use mixed low-risk and high-risk heterogeneous populations as ground truth.

Perceived heterogeneity	**Clinical interpretation**	**Agent Perception of Patient Risk**
Accurate	Accurate risk stratification	True risk levels accurately observed
Exaggerated	Over-stratification	High-risk appear higher, low-risk appear lower
Compressed	Under-stratification	High-risk appear less high, low-risk appear less low

We also substantially expand patient heterogeneity beyond Experiment Set 3. Whereas Experiment Set 3 introduced Gaussian variation only in infection probability (with other treatment response parameters held constant within subpopulations), Experiment Set 4 samples all available patient attributes from Gaussian distributions. This creates both between-subpopulation differences (high-risk vs. low-risk mean parameter values) and within-subpopulation variation ($\sigma = 0.1$ for all attributes), more closely approximating real-world patient diversity.

Two additional design changes increase realism. First, we increase patient volume to 10 patients per timestep. Second, we implement differential observability: low-risk patients present with minimal workup (only 2 observable attributes: infection probability and spontaneous recovery probability), while high-risk patients undergo comprehensive assessment (all patient attributes observable). This mirrors real-world clinical triage, where different patients have differential observability based on risk assessment.

Because delayed AMR observations render the environment partially observable over time, we evaluated both hierarchical PPO and hierarchical recurrent PPO agents to assess the impact of adding memory to the agent. This experiment set allowed us to evaluate how increasing noise, bias in both information streams and delay in resistance information affected learned prescribing policies. We also then compared the performance of these learned policies to fixed prescribing rules that were operating under the same conditions of increased information degradation.

4 Results

4.1 Value Iteration Optimal Policy (benchmark for Experiment Sets #1 - #3)

We used value iteration (VI) to derive expected-reward optimal prescribing policies under perfect observability for the three antibiotic scenarios, treating the resulting distribution of outcomes as benchmark ceilings for all RL comparisons in Experiment Sets 1–3. To characterize realized performance under stochastic outcomes, each VI policy was rolled out across 100 seeds per scenario; AMR trajectories (top panels) and cumulative clinical count metrics (bottom panels) are summarized in Figure 3. The cumulative clinical response plots allow us to evaluate antibiotic efficacy throughout the episode by plotting counts of sensitive vs. resistant infections treated per antibiotic; sustained separation, with sensitive counts above resistant counts, indicates preserved clinical utility over the episode.

- Single-antibiotic scenario: VI induces threshold-like prescribing that stabilizes AMR near an equilibrium of approximately 0.556 (Figure 3, left column).
- Two-antibiotic scenario (no cross-resistance): VI converges to two distinct equilibria for each antibiotic: approximately 0.557 for antibiotic A and 0.357 for antibiotic B (Figure 3, middle column).
- Two-antibiotic scenario (with cross-resistance): The VI policy produces a slow alternation pattern rather than a fixed equilibrium, with periodic cycling of AMR levels for both drugs over the episode (Figure 3, right column).

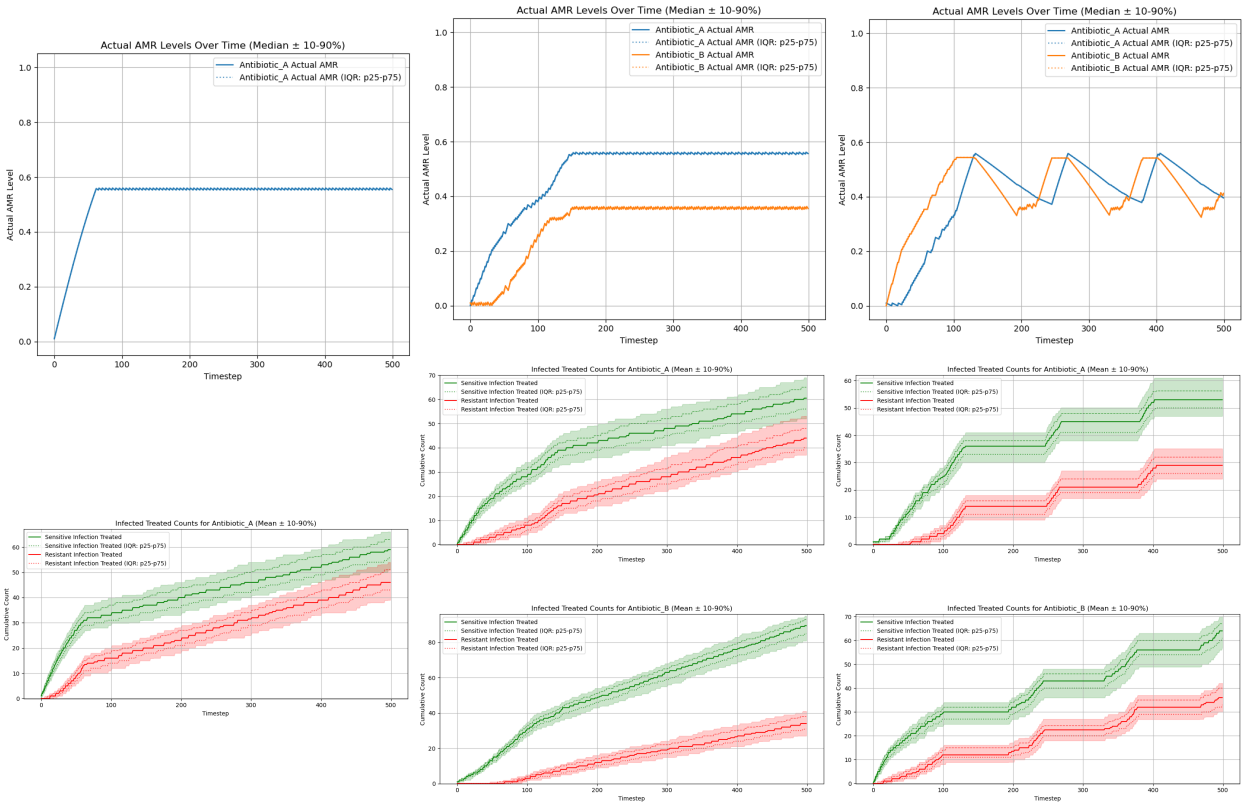


Figure 3: AMR levels evolution and cumulative count of sensitive vs. resistant infections treated per antibiotic, when value iteration-derived optimal policy is followed (where policy is computed using perfectly observed environments). Left: single-antibiotic scenario. Middle: two-antibiotic scenario without cross-resistance. Right: two-antibiotic scenario with cross-resistance. Each column shows evolution of true AMR levels (top) and cumulative infected/treated counts per antibiotic over time (bottom rows).

4.2 Experiment Set #1: Perfect Observability

Under perfect observability, we first asked a basic feasibility question: can reinforcement learning recover reasonable, near-optimal prescribing policies when all patient attributes and AMR levels are fully and accurately observed at every timestep? This best-case setting removes information degradation and therefore isolates the core learning challenge in these environments—long-horizon credit assignment.

Flat PPO showed only limited success in this setting. In the single-antibiotic scenario, it learned policies that stabilized AMR near a steady-state equilibrium, but the learned equilibrium varied across seeds and did not consistently match the VI-derived optimum (Figure 4, left column). In both two-antibiotic scenarios, flat PPO performed substantially worse: equilibrium AMR levels were highly variable across seeds, trajectories in the cross-resistance setting showed little evidence of the cycling seen in the VI benchmark, and cumulative sensitive-versus-resistant treated counts showed weaker separation, indicating poorer preservation of antibiotic efficacy over the episode (Figure 4, middle and right columns).

Hierarchical PPO produced qualitatively different behavior across all three scenarios. Its trajectories differed from both flat PPO and the VI benchmark, tending toward monotonic AMR escalation with high cross-seed variance (Figure 5). In the cross-resistance scenario, some moderate cycling was present, though less distinctly than in the VI-derived policy. Notably, despite these qualitatively different dynamics, hierarchical PPO achieved quantitative performance comparable to the VI benchmark in the single-antibiotic scenario and exceeded the VI benchmark in both two-antibiotic scenarios (Figure 6).

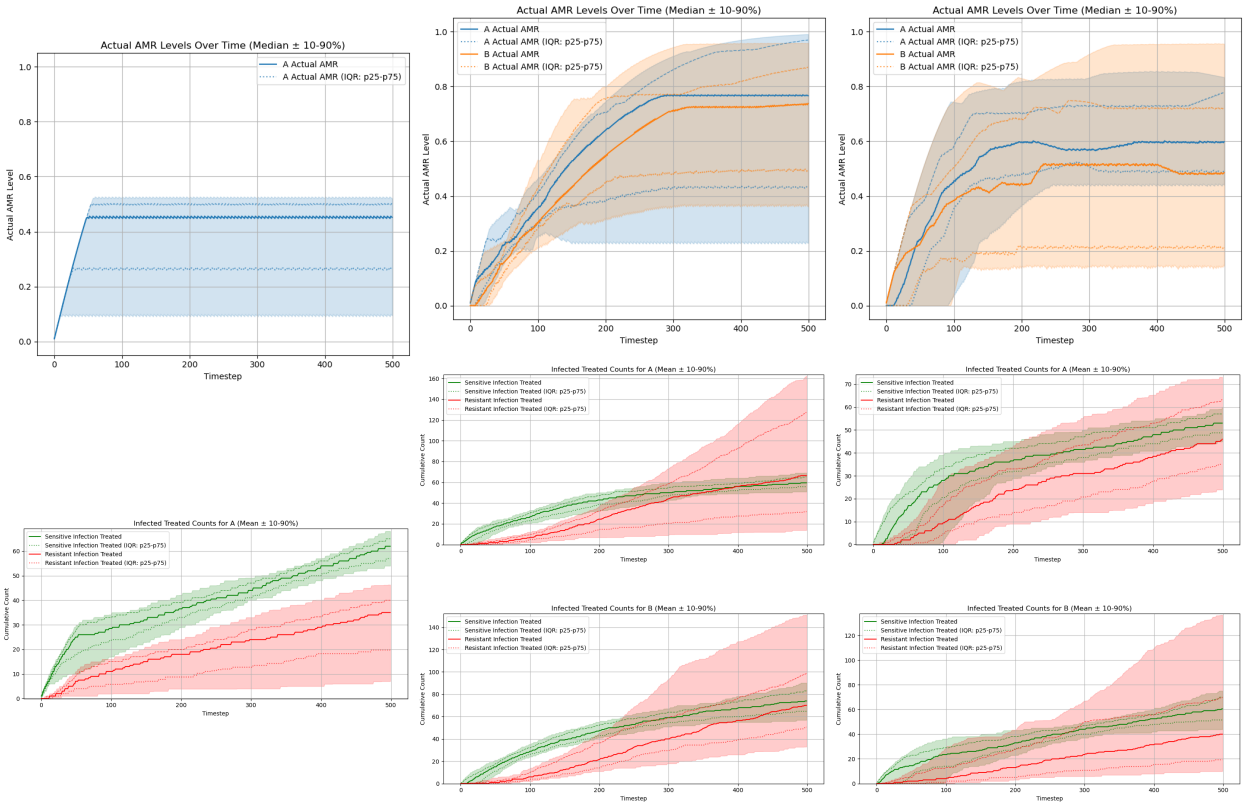


Figure 4: AMR levels evolution and cumulative count of sensitive vs. resistant infections treated per antibiotic, when tuned Flat PPO agent is trained in perfectly observed environment and then has evaluative rollouts performed. Left: single-antibiotic scenario. Middle: two-antibiotic scenario without cross-resistance. Right: two-antibiotic scenario with cross-resistance. Each column shows evolution of true AMR levels (top) and cumulative infected/treated counts per antibiotic over time (bottom rows).

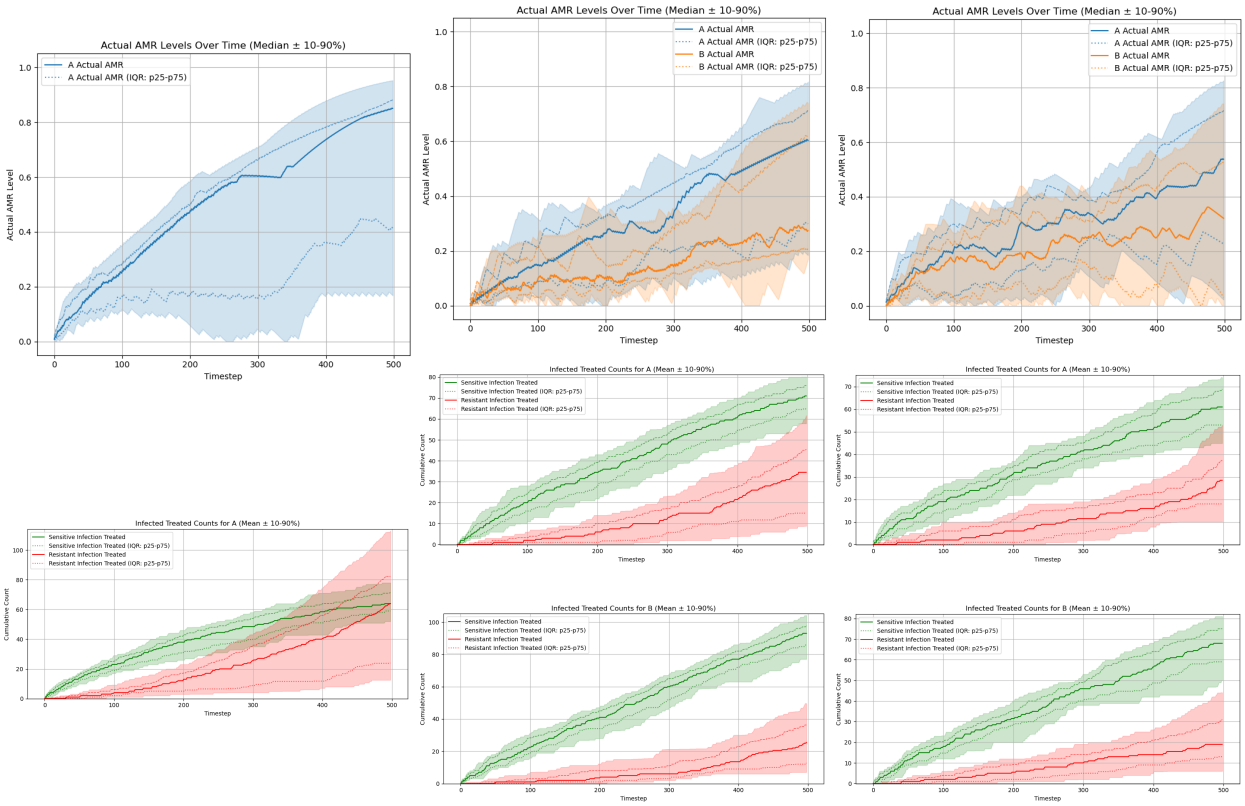
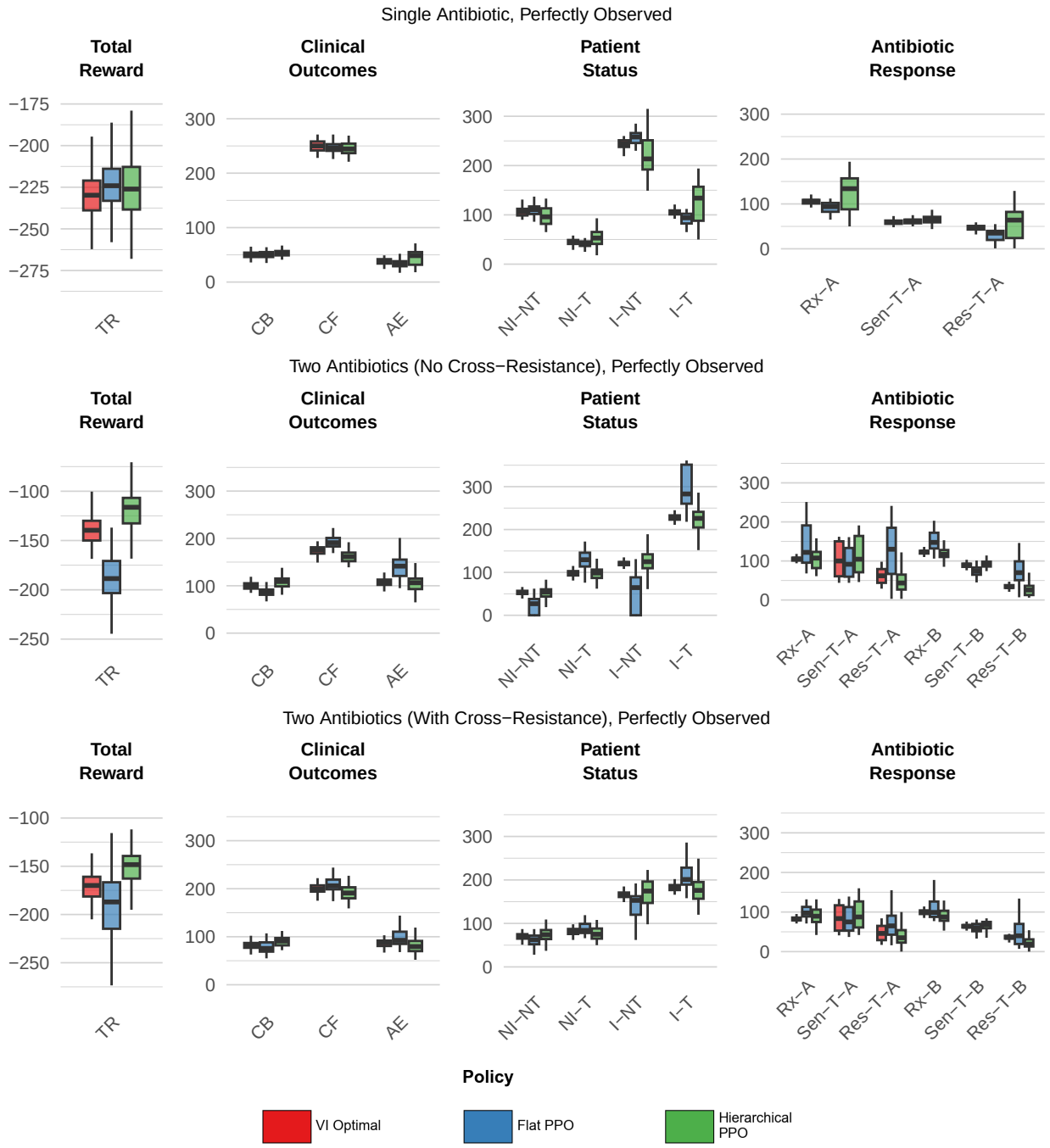


Figure 5: AMR levels evolution and cumulative count of sensitive vs. resistant infections treated per antibiotic, when tuned Hierarchical PPO agent is trained in perfectly observed environment and then has evaluative rollouts performed. Left: single-antibiotic scenario. Middle: two-antibiotic scenario without cross-resistance. Right: two-antibiotic scenario with cross-resistance. Each column shows evolution of true AMR levels (top) and cumulative infected/treated counts per antibiotic over time (bottom rows).



Abbrev	Description	Abbrev	Description	Abbrev	Description
TR	Total Reward	NI-T	Not Infected, Treated	Res-T-A	Resistant Treated: A
CB	Clinical Benefits	I-NT	Infected, No Treatment	Rx-B	Prescriptions: B
CF	Clinical Failures	I-T	Infected, Treated	Sen-T-B	Sensitive Treated: B
AE	Adverse Events	Rx-A	Prescriptions: A	Res-T-B	Resistant Treated: B
NI-NT	Not Infected, No Treatment	Sen-T-A	Sensitive Treated: A		

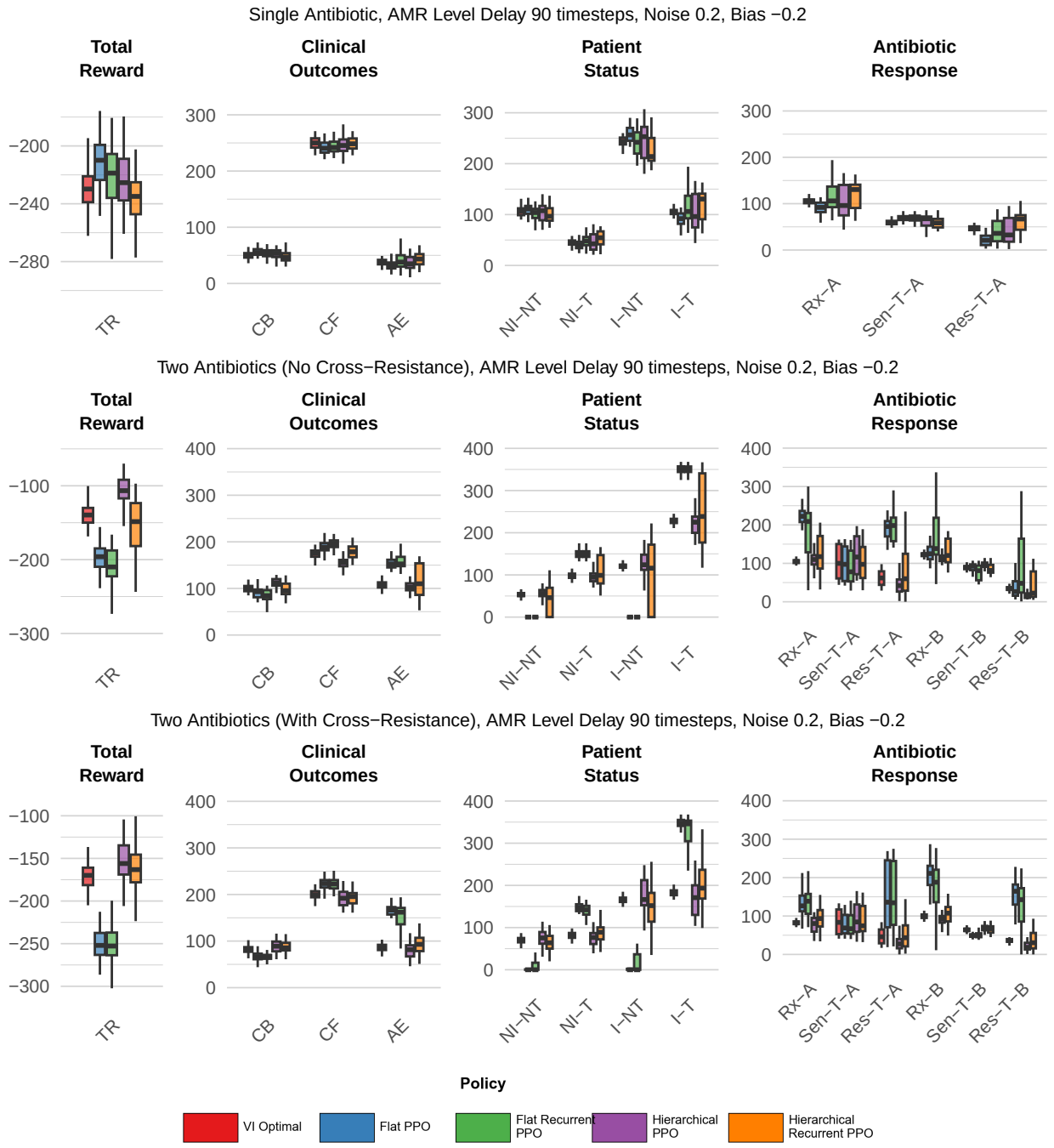
Figure 6: Comparison of cumulative performance metrics of Value Iteration-Derived Optimal Policy vs. Flat PPO vs. Hierarchical PPO in perfectly observed environment.

4.3 Experiment Set #2: Delayed, Noisy, Biased AMR

In Experiment Set 2, policy performance is evaluated when AMR surveillance is partially observable: AMR levels are updated every 90 timesteps, and the visible AMR signal is both noisy and biased (noise = 0.2, bias = -0.2). Figure 7 compares four PPO-based agents (flat memoryless, flat recurrent, hierarchical memoryless, hierarchical recurrent) against the VI-derived benchmark across all three antibiotic scenarios.

Among the flat agents, adding recurrent memory did not improve performance. The memoryless flat PPO agent outperformed the recurrent flat PPO agent in the single-antibiotic and two-antibiotic no cross-resistance scenarios, with similar performance in the cross-resistance setting. These differences were accompanied by distinct prescribing patterns. In the single-antibiotic scenario, the memoryless flat agent concentrated prescribing shortly after each AMR observation update and then largely ceased prescribing until the next update, whereas the recurrent flat agent reduced prescribing during stale-information periods but did not stop as completely. In the two-antibiotic scenario without cross-resistance, this same pattern appeared as relatively clean on/off alternation between antibiotics A and B for the memoryless flat agent, whereas the recurrent flat agent reduced use of the disfavored antibiotic without fully discontinuing it. When cross-resistance was introduced, the distinction between flat memoryless and recurrent prescribing patterns became less pronounced. Detailed prescribing time series and AMR trajectories for these agents are provided in Supplementary Materials S4.

The hierarchical agents showed a similar ranking, with hierarchical memoryless PPO outperforming hierarchical recurrent PPO across scenarios. As in Experiment Set 1, both hierarchical variants exhibited increased late-episode prescribing and corresponding upward AMR drift in later timesteps. Their prescribing patterns also appeared responsive to the AMR update cycle, although the deterministic-sequence macro-actions made these transitions less visually distinct than in the flat-agent case. Notably, hierarchical memoryless PPO again matched or exceeded the VI-derived benchmark in the two-antibiotic scenarios, despite these qualitatively different trajectories compared to the VI-derived benchmark policies (Figure 7).



Abbrev	Description	Abbrev	Description	Abbrev	Description
TR	Total Reward	NI-T	Not Infected, Treated	Res-T-A	Resistant Treated: A
CB	Clinical Benefits	I-NT	Infected, No Treatment	Rx-B	Prescriptions: B
CF	Clinical Failures	I-T	Infected, Treated	Sen-T-B	Sensitive Treated: B
AE	Adverse Events	Rx-A	Prescriptions: A	Res-T-B	Resistant Treated: B
NI-NT	Not Infected, No Treatment	Sen-T-A	Sensitive Treated: A		

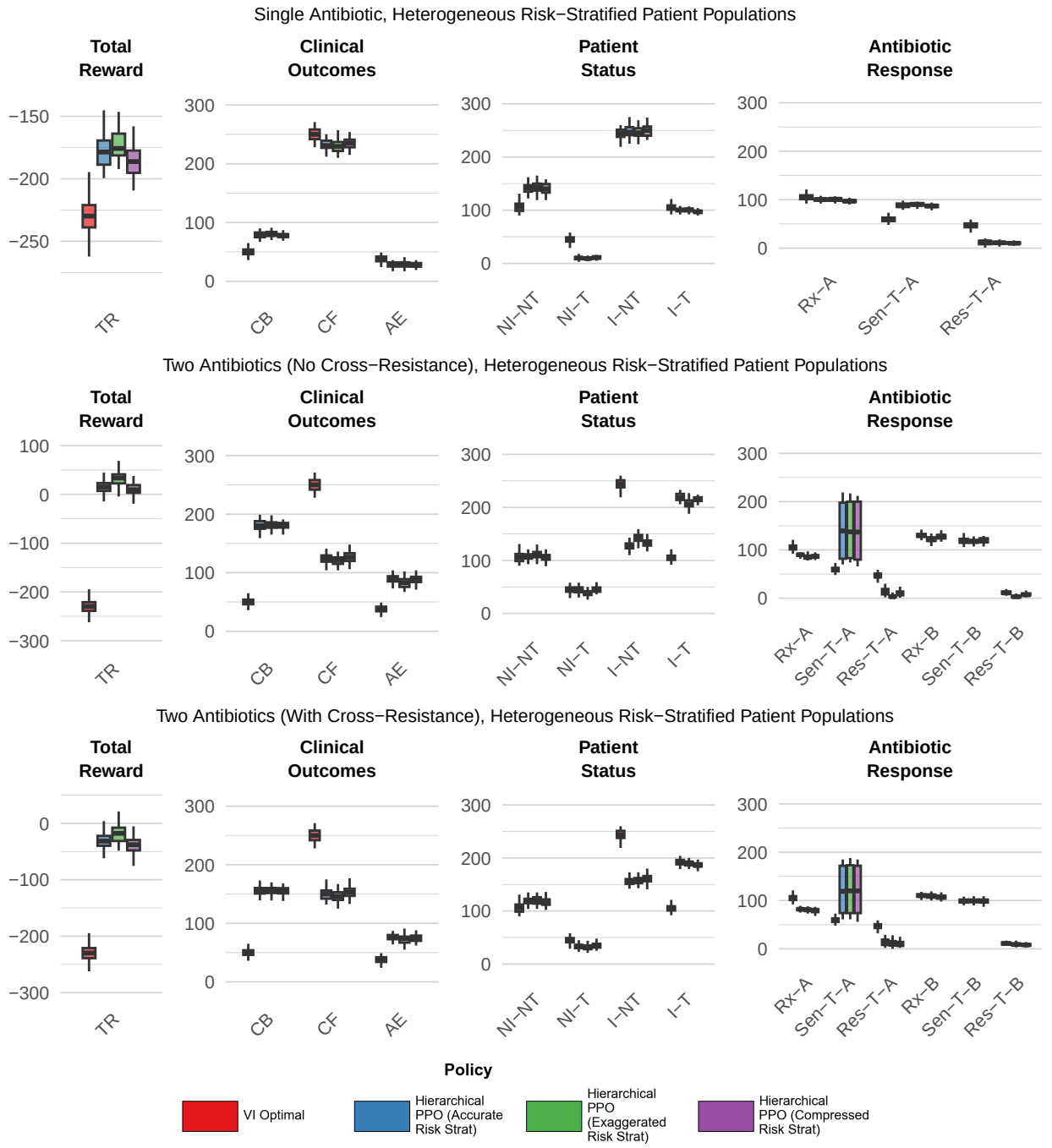
Figure 7: Comparison of cumulative performance metrics of Value Iteration-Derived Optimal Policy vs. Flat PPO vs. Flat Recurrent PPO vs. Hierarchical Recurrent PPO, in environment with degraded AMR level information. AMR levels affected by update delay = 90 timesteps, noise = 0.2, bias = -0.2.

4.4 Experiment Set #3: Heterogeneous Patient Attributes, Varied Bias

In Experiment Set 3, we isolate the effect of patient-level perception error by introducing bias in observed patient attributes while keeping AMR observations perfect and up to date. We modeled two subpopulations (high-risk and low-risk) sampled with equal probability, with subpopulation parameters chosen so that population-average attribute values match the homogeneous baseline used in Experiment Set 1, allowing direct comparison to the VI benchmark.

Figure 8 shows a consistent pattern across all three antibiotic scenarios: introducing true risk heterogeneity substantially improves both clinical outcomes and antibiotic stewardship relative to the homogeneous baseline. Under accurate risk stratification, AMR levels stabilized at low equilibria with tight cross-seed variance — a marked contrast to the monotonic escalation and high variance observed in Experiment Sets 1 and 2. Performance under exaggerated risk stratification was modestly better than accurate stratification across all three scenarios. Performance under compressed risk stratification was slightly worse than accurate stratification, though substantially better than the homogeneous population VI policy baseline.

The largest performance gains across all stratification conditions appeared in the two-antibiotic scenario without cross-resistance (Figure 8, middle panels). In the cross-resistance scenario, these gains were attenuated relative to the no cross-resistance case (Figure 8, right panels).



Abbrev	Description	Abbrev	Description	Abbrev	Description
TR	Total Reward	NI-T	Not Infected, Treated	Res-T-A	Resistant Treated: A
CB	Clinical Benefits	I-NT	Infected, No Treatment	Rx-B	Prescriptions: B
CF	Clinical Failures	I-T	Infected, Treated	Sen-T-B	Sensitive Treated: B
AE	Adverse Events	Rx-A	Prescriptions: A	Res-T-B	Resistant Treated: B
NI-NT	Not Infected, No Treatment	Sen-T-A	Sensitive Treated: A		

Figure 8: Comparison of cumulative performance metrics of Value Iteration-Derived Optimal Policy vs. Hierarchical PPO (Accurate Risk Stratification Environment) vs. Hierarchical PPO (Exaggerated Risk Stratification Environment) vs. Hierarchical PPO (Compressed Risk Stratification Environment), in environments with heterogeneous risk-stratified patient population with varying levels of bias.

4.5 Experiment Set #4: Combined Observational Noise, Patient Heterogeneity

In Experiment Set 4, we combined both sources of uncertainty from Experiment Sets 2 and 3 and also increase patient volume: noisy and biased patient observations, delayed and noisy AMR surveillance, and ten patients per timestep. We then compared the performance of hierarchical PPO and hierarchical recurrent PPO against the two fixed prescribing rules described in Section 3.6.

Both hierarchical architectures adopted markedly more conservative prescribing strategies than either fixed-rule baseline across all three antibiotic scenarios (Figure 9). Under the hierarchical learned policies, equilibrium AMR levels remained low and stable with tight cross-seed variance across all scenarios (Tables 3, 4, 5). In contrast, both fixed prescribing rules produced wide cross-seed variance in AMR trajectories, and also exhibited monotonically increasing AMR levels over the episode. Crucially, the hierarchical policies were not merely trading off short-term clinical outcomes for long-term stewardship: they achieved more clinical successes, fewer clinical failures, and significantly lower long-term AMR levels compared to either fixed rule.

We observe here that equilibrium AMR levels under the hierarchical learned policies for these experiments were substantially lower than those observed under the VI-derived policy in Experiment Sets 1–3. We also note here that within the hierarchical pair, hierarchical recurrent PPO modestly outperformed hierarchical memoryless PPO in both two-antibiotic scenarios, with the advantage more pronounced in the cross-resistance setting; this was achieved through slightly more aggressive prescribing and selection of slightly higher AMR equilibria with less variance, yielding better clinical outcomes with smaller variance compared to the memoryless variant.

Table 3: Experiment Set 4: Single-Antibiotic Scenario. Values are mean ± SD across seeds. HPPO = Hierarchical PPO, HRPPO = Hierarchical Recurrent PPO.

Metric	HPPO	HRPPO	Greedy	Lowest AMR
Final AMR level - Antibiotic A	0.084 +/- 0.058	0.093 +/- 0.054	1.000 +/- 0.000	0.984 +/- 0.086
Total Clinical Benefits	292.850 +/- 15.309	297.017 +/- 15.334	236.050 +/- 22.996	252.733 +/- 19.034
Total Clinical Failures	2785.333 +/- 38.212	2787.483 +/- 35.462	2859.017 +/- 39.068	2846.083 +/- 29.492
Total Adverse Events	26.600 +/- 4.570	25.967 +/- 4.995	230.533 +/- 81.331	109.150 +/- 19.479
Count Antibiotic A Prescriptions	94.167 +/- 6.201	95.683 +/- 5.771	775.633 +/- 281.717	361.383 +/- 58.758

Table 4: Experiment Set 4: Two-Antibiotic Scenario (No Cross-Resistance). Values are mean ± SD across seeds. HPPO = Hierarchical PPO, HRPPO = Hierarchical Recurrent PPO.

Metric	HPPO	HRPPO	Greedy	Lowest AMR
Final AMR level - Antibiotic A	0.096 +/- 0.038	0.109 +/- 0.020	1.000 +/- 0.000	0.998 +/- 0.013
Final AMR level - Antibiotic B	0.100 +/- 0.027	0.108 +/- 0.016	1.000 +/- 0.000	0.992 +/- 0.052
Total Clinical Benefits	375.350 +/- 17.259	374.867 +/- 18.063	241.400 +/- 18.659	256.383 +/- 14.887
Total Clinical Failures	2675.400 +/- 35.337	2668.167 +/- 37.394	2829.367 +/- 27.614	2819.767 +/- 33.512
Total Adverse Events	87.283 +/- 10.291	85.483 +/- 9.530	449.433 +/- 96.485	284.167 +/- 27.407
Count Antibiotic A Prescriptions	89.517 +/- 17.056	92.283 +/- 13.237	539.183 +/- 180.049	371.300 +/- 32.095
Count Antibiotic B Prescriptions	137.683 +/- 7.084	138.167 +/- 12.038	609.417 +/- 183.578	364.283 +/- 43.499

Table 5: Experiment Set 4: Two-Antibiotic Scenario (With Cross-Resistance). Values are mean ± SD across seeds. HPPO = Hierarchical PPO, HRPPO = Hierarchical Recurrent PPO.

Metric	HPPO	HRPPO	Greedy	Lowest AMR
Final AMR level - Antibiotic A	0.087 +/- 0.051	0.108 +/- 0.031	1.000 +/- 0.000	0.961 +/- 0.136
Final AMR level - Antibiotic B	0.093 +/- 0.043	0.116 +/- 0.015	0.998 +/- 0.011	0.926 +/- 0.227
Total Clinical Benefits	351.800 +/- 20.786	358.483 +/- 18.728	230.533 +/- 19.948	257.250 +/- 15.987
Total Clinical Failures	2717.167 +/- 41.811	2701.883 +/- 29.036	2857.650 +/- 32.930	2820.183 +/- 34.720
Total Adverse Events	67.100 +/- 14.248	73.883 +/- 7.018	385.617 +/- 115.153	229.433 +/- 55.623
Count Antibiotic A Prescriptions	78.250 +/- 9.400	77.517 +/- 19.109	542.750 +/- 182.256	345.017 +/- 87.790
Count Antibiotic B Prescriptions	104.367 +/- 25.099	118.117 +/- 6.206	474.467 +/- 243.455	264.150 +/- 116.789

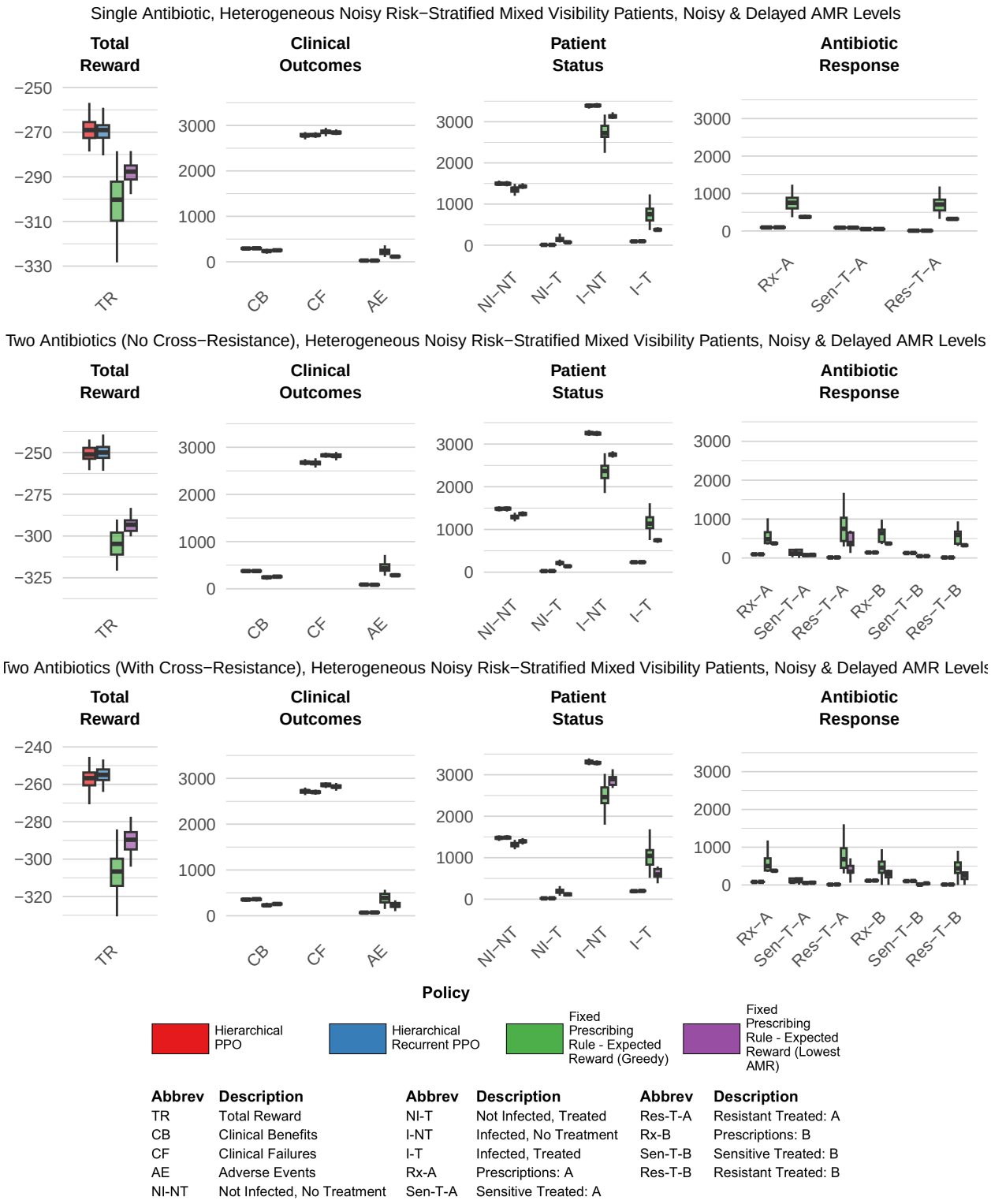


Figure 9: Comparison of cumulative performance metrics of Hierarchical PPO Recurrent PPO vs. Fixed Prescribing Rule Expected Reward (Greedy) vs. Fixed Prescribing Rule Expected Reward (Lowest AMR), in environments with heterogeneous noisy risk-stratified mixed visibility patient populations (noise=0.1), noisy and delayed AMR Levels (delay = 90 timesteps, noise=0.2)

5 Discussion

Across four experiment sets, we found that reinforcement learning agents can recover strong antibiotic prescribing policies, but performance and policy behavior are highly sensitive to the available information and environment structure.

5.1 Individual Experiment Set Findings

Experiment Set 1 (Perfect Observability). Under perfect observability, flat PPO was sufficient for the single-antibiotic scenario but failed to learn effective policies in both two-antibiotic settings, even with complete state information. This suggests the failure mode is not informational but architectural: flat agents struggle with the long-horizon credit assignment required when prescribing decisions have delayed, coupled effects on future resistance levels. Switching to a hierarchical architecture — which decomposes the long-horizon problem into sequences of macro-actions — proved sufficient to recover competitive performance across all three scenarios.

However, the quantitative advantage of hierarchical PPO over the VI benchmark in the two-antibiotic scenarios warrants caution in interpretation. The learned AMR trajectories — characterized by monotonic AMR escalation and high cross-seed variance — are qualitatively inconsistent with the stable equilibria and cycling behavior dictated by the VI-derived policy, suggesting these agents were not learning genuinely superior long-horizon strategies. We hypothesize that this performance advantage is instead an artifact of finite-horizon exploitation: because our hierarchical agents were given access to timestep information during the episode, they could learn policies conditioned on episode progress, increasing prescribing aggressively near the episode boundary when future AMR costs are discounted by limited remaining horizon. This issue is discussed further in Section 5.3.

Experiment Set 2 (Delayed, Noisy, Biased AMR). Under delayed and noisy AMR surveillance, recurrent memory consistently degraded policy performance across both flat and hierarchical architectures. We hypothesize that this counterintuitive result is driven by a behavioral difference between memoryless and recurrent policies: memoryless agents learn to treat AMR update events as discrete signals, adopting binary on/off prescribing aligned with update cycles, while recurrent agents maintain more sustained prescribing throughout — including during stale-information periods when conservative behavior would better preserve antibiotic efficacy. The longer inter-update recovery windows afforded by the memoryless on/off strategy appear to outweigh any benefit from integrating observation history. This advantage was most pronounced in single-antibiotic and two-antibiotic no cross-resistance scenarios; when prescribing one antibiotic also increases resistance to another, the coupling structure in AMR dynamics limits the gains available from inter-update conservation.

The finding that hierarchical memoryless PPO matched or exceeded VI benchmark performance in the two-antibiotic scenarios under degraded AMR surveillance compounds the concern raised in the discussion of the results from Experiment Set 1. There, we hypothesized that hierarchical overperformance of the VI benchmark reflected finite-horizon exploitation rather than genuinely superior policy learning. Here, that interpretation is strengthened: AMR information degradation should reduce performance relative to the perfect-observability benchmark, yet the overperformance persists. This suggests that finite-horizon exploitation is robust enough to overwhelm the degradative effects of delayed and noisy AMR surveillance, further underscoring the need for the architectural fix discussed in Section 5.3.

Experiment Set 3 (Heterogeneous Patient Attributes, Varied Bias). The performance gains from risk stratification in Experiment Set 3 appear to be driven by a single core mechanism: observable heterogeneous patient attributes give the agent reliable triage signals, enabling it to selectively treat high-risk patients while confidently withholding treatment from low-risk ones. This is a fundamental contrast to the homogeneous populations in Experiment Sets 1 and 2, where all patients shared identical observable attributes and the agent had no basis for differential treatment decisions. The availability of these triage signals also appears sufficient to stabilize long-horizon learning: AMR trajectories converged to low, stable equilibria with tight cross-seed variance, with no evidence of the finite-horizon exploitation seen in Experiment Sets 1 and

2. We interpret this as evidence that patient heterogeneity, by providing high-frequency decision-relevant information, reduces the agent's reliance on episode-boundary strategies to achieve competitive returns.

The finding that exaggerated risk stratification modestly outperformed accurate stratification is somewhat surprising, but we think the likely mechanism is that stronger perceived separation between risk groups increases the agent's confidence in withholding treatment from low-risk patients, further reducing unnecessary prescribing and preserving antibiotic efficacy. Compressed risk stratification produced only modest degradation in performance, suggesting that the gains from heterogeneity are fairly robust to miscalibration as long as relative risk ordering is preserved — though the direction of miscalibration matters, with over-stratification preferable to under-stratification in the scenarios tested here.

The largest gains across stratification conditions appeared in the two-antibiotic scenario with no cross-resistance. Here, risk-stratified treatment can be combined with antibiotic rotation: resting one antibiotic allows its resistance level to decay while the other is used, compounding the efficacy-preservation benefit of selective treatment. In the cross-resistance scenario these gains are attenuated, since prescribing one antibiotic also drives up resistance to the other, reducing the benefit of rotation.

Experiment Set 4 (Combined Observational Noise, Patient Heterogeneity). In the most complex scenarios tested in this study, hierarchical PPO agents dominated fixed prescribing rules across both clinical and stewardship metrics — not by trading one off against the other, but by outperforming on both simultaneously. Fixed prescribing rules treated aggressively early, driving resistance upward and eroding future treatment capacity; the learned policies, by contrast, converged to conservative low-AMR equilibria that preserved efficacy throughout the episode. The substantially lower equilibrium AMR levels in Experiment Set 4 compared to Experiment Sets 1–3 are consistent with the higher patient volume applying more immediate and frequent pressure on the agent to preserve antibiotic efficacy: with ten patients per timestep, the consequences of resistance escalation accumulate faster, and the learned policy responds correspondingly.

The absence of finite-horizon exploitation in Experiment Set 4 — in contrast to Experiment Sets 1 and 2 — is consistent with this interpretation. When patient volume is high and triage signals are available, the agent faces a sufficiently rich and high-frequency learning signal that episode-boundary strategies offer no particular advantage. The result is stable, low-variance convergence to genuinely conservative policies, mirroring the pattern observed in Experiment Set 3.

The modest memory advantage of hierarchical recurrent PPO over hierarchical memoryless PPO in the two-antibiotic scenarios — reversing the pattern seen in Experiment Set 2 — is also worth noting. In Experiment Set 2, recurrent memory degraded performance by sustaining prescribing during stale-information periods. Here, in the most informationally degraded environment in the study, memory appears to confer a benefit instead: we hypothesize that when environmental signals are sufficiently noisy and delayed across multiple information streams simultaneously, the ability to maintain an internal belief state about the world outweighs the cost of imperfect memory-based inference. This suggests that the value of recurrent memory is context-dependent, and may be most beneficial precisely when information degradation is most severe.

5.2 Cross-Experiment Set Findings

Across all four experiment sets, the results confirm our central hypothesis: reinforcement learning agents can successfully learn antibiotic prescribing policies that perform comparably to static benchmark policies. However, the conditions under which this holds are architecture-dependent. Flat agents were insufficient for all but the simplest single-antibiotic scenario; hierarchical RL with temporal abstraction was necessary to address the long-horizon credit assignment structure inherent in the antibiotic prescribing problem. We make no claim that hierarchical RL is the only viable strategy for these environments, nor that it would scale without modification to substantially more complex simulations, i.e. more antibiotics, larger patient cohorts, or richer resistance dynamics. Nevertheless, even within these relatively controlled scenarios, the experiments surfaced several non-obvious insights: notably, that over-stratification of patient risk may yield better aggregate outcomes than accurate risk calibration, and that recurrent memory is not always beneficial, and may be most advantageous in settings where information degradation is more severe.

Agents achieved these outcomes despite being optimized using only individual clinical reward ($\lambda = 0$, with no explicit penalty for increasing community-level antimicrobial resistance). This indicates that when the environment itself creates long-term consequences for short-sighted actions — as antimicrobial resistance dynamics do — appropriately designed agent architectures can learn farsighted prescribing strategies without requiring explicit stewardship reward shaping. The consistent performance advantage of hierarchical over flat agents, particularly in cross-resistance scenarios, underscores the importance of policy architecture in settings where treatment decisions have delayed, coupled effects on future treatment efficacy.

5.3 Finite-Horizon Effects and Exploitation by Hierarchical Agents

In Experiment Sets 1 and 2, hierarchical agents learned policies characterized by monotonically increasing AMR levels over the course of each episode. This pattern is consistent with end-of-episode exploitation under a fixed training horizon: when the agent can observe the current timestep and therefore infer how much of the episode remains, it learns to increase prescribing aggressively near the episode boundary, collecting immediate clinical reward at the cost of resistance escalation that it will not be penalized for sustaining. This behavior was an unintended consequence of providing hierarchical agents with timestep awareness — a design choice made to support timed macro-action control, specifically to prevent long macro-actions from being truncated or clipped at episode boundaries. The results of Experiment Sets 1 and 2 reveal that this design has the second-order effect of enabling timestep-conditioned policies that exploit the fixed horizon structure, making it difficult to cleanly attribute quantitative performance differences to information degradation per se. In the two-antibiotic scenarios of Experiment Set 2, hierarchical agents outperformed the VI-derived benchmark despite operating under degraded AMR surveillance — a result we interpret as finite-horizon exploitation overwhelming the degradative effects of information corruption, rather than as evidence of genuinely superior policy learning.

The most straightforward remedy would be to remove timestep information from the manager's observation, and instead handle episode-boundary macro-action truncation through boundary-aware clipping in the wrapped environment. Additional mitigations include increasing episode length or randomizing horizon timing to reduce the predictability of episode boundaries. If the performance advantage of hierarchical agents over the VI benchmark disappears under these modifications, this would provide empirical confirmation that finite-horizon exploitation was the primary driver of the performance advantage of the hierarchical agents over VI-derived policies that we observed in Experiment Sets 1 and 2. Such tests would cleanly separate genuine policy quality from artifacts induced by fixed-horizon training structure.

Notably, finite-horizon exploitation was absent in Experiment Sets 3 and 4, where hierarchical agents instead converged to stable low-AMR policies with tight cross-seed variance and no evidence of boundary gaming. We attribute this contrast to the richer learning signal available in those experiment sets: heterogeneous patient populations provide high-frequency triage signals that give the agent a reliable basis for differential treatment decisions at every timestep, reducing its reliance on episode-boundary strategies to achieve competitive returns. In Experiment Sets 1 and 2, where all patients shared identical observable attributes, the agent had no such basis for discrimination, and boundary exploitation emerged as a fallback strategy for maximizing cumulative reward under an otherwise undifferentiated decision problem. We therefore interpret finite-horizon exploitation as a difficulty-triggered fallback under fixed-horizon training, rather than as the primary mechanism of hierarchical RL success — the behavior of agents in Experiment Sets 3 and 4 demonstrates that the hierarchical architecture is capable of robust, genuinely conservative policy learning when the environment provides sufficient decision-relevant structure.

5.4 Limitations

5.4.1 Abstraction of pathogen identity

Our simulation environment necessarily represents a simplified and idealized abstraction of real-world antibiotic prescribing and antimicrobial resistance dynamics. The most significant simplification involves our

treatment of pathogens: we deliberately abstract away individual pathogen identity, even though in clinical practice pathogens serve as the natural unit for measuring antimicrobial resistance. Real-world antibiograms stratify resistance by pathogen species, tallying sensitive and resistant isolates separately for each organism; our simulator omits this pathogen-level granularity entirely. We made this choice deliberately in order to distill clinical decision-making regarding antibiotic therapy to its most essential elements: (1) the probability that a given patient has a bacterial infection requiring antibiotic treatment, (2) the likelihood that a specific antibiotic will effectively treat that infection if present, and (3) the clinical trade-offs associated with prescribing. By reducing the decision to these core questions, we gain interpretability — both in how patients and AMR levels are represented, and in how the policies learned by RL agents can be understood. We also hypothesize that this simplification likely represents a theoretical upper bound on achievable clinical benefit; in a more stochastic, multi-pathogen environment, policy performance would likely degrade further.

5.4.2 Stationarity assumptions

In this study, we model both the patient population and the antimicrobial resistance response curve for each antibiotic as stationary entities — that is, patient attribute distributions and resistance dynamics remain constant throughout each episode and across episodes. In reality, neither is static: patient populations shift over time due to demographic change, seasonal variation in infection prevalence, and emergence of high-risk subpopulations, while resistance mechanisms can evolve in response to sustained selection pressure in ways that may be partially or fully irreversible. Policies learned under stationary assumptions may therefore be poorly calibrated to real-world settings where the underlying dynamics are drifting, and our results should be interpreted accordingly as characterizing performance under idealized stable conditions. Extensions to non-stationary dynamics are discussed in the Future Work section.

5.4.3 Single centralized prescriber

The simulator models a single, centralized prescriber with complete information about all patients and system-level AMR dynamics. This is unrealistic: real-world antibiotic prescribing involves multiple independent clinicians with limited visibility into their peers' prescribing patterns and heterogeneous access to surveillance data. We plan to address this limitation in future work by extending the framework to support multi-agent, multi-locale environments where individual prescribers optimize locally with incomplete information (see Section 5.5).

5.4.4 Reinforcement learning algorithm selection

PPO was selected primarily for pragmatic reasons: it natively supports multidiscrete action spaces, as required for our multi-patient, multi-antibiotic setup, and is computationally tractable for the experimental scope of this study. Other RL algorithms or agent architectures may yield different performance characteristics. The modular design of the `abx_amr_simulator` package — which conforms to the Gymnasium API — allows interested users to straightforwardly implement and evaluate alternative agent classes beyond those tested here.

5.4.5 AMR resistance dynamic assumptions

The `AMR_LeakyBalloon` model makes specific assumptions about the shape of the antibiotic resistance response curve: latent resistance pressure is mapped to an observable AMR level through a sigmoid function, yielding a smooth, bounded accumulation and decay dynamic. While we believe this is a reasonable abstraction of resistance emergence and relief, the true functional form of resistance dynamics in real bacterial populations is not well characterized at the community level, and alternative response curve shapes could meaningfully alter the prescribing strategies learned by RL agents.

Additionally, our current implementation models cross-resistance as strictly positive — prescribing antibiotic A increases resistance to antibiotic B — whereas certain antibiotic combinations in fact exhibit collateral sensitivity, where use of one agent decreases resistance to another (Wang et al. 2025). Incorporating negative cross-resistance could meaningfully alter optimal prescribing strategies, potentially favoring cycling or mixing regimens that exploit these collateral effects.

Both of these limitations are mitigated by the modular architecture of the simulator. AMR dynamics are implemented through an abstract interface (`AMRDynamicsBase`), with `AMR_LeakyBalloon` as the concrete implementation used in this study. Users can provide alternative implementations of `AMRDynamicsBase` to specify different response curve shapes or cross-resistance structures, enabling systematic investigation of how resistance dynamic assumptions affect learned prescribing policies without requiring modification to the broader simulation framework.

5.4.6 Finite-horizon artifact in hierarchical agent training

In Experiment Sets 1 and 2, hierarchical agent performance should be interpreted with caution due to finite-horizon exploitation: because hierarchical agents were provided with timestep information during training, they learned policies conditioned on episode progress, increasing prescribing aggressively near episode boundaries in a manner that inflated cumulative reward metrics without reflecting genuine long-horizon policy quality. This artifact and its implications are discussed in detail in Section 5.3; a planned architectural fix is described in the Future Work section.

5.4.7 Scope of findings and generalizability

Because the simulator is intentionally abstract, findings should not be interpreted as direct clinical policy recommendations. The framework is most useful for directional analysis under controlled conditions — quantifying the effect of delayed, noisy, or biased antibiogram updates, biased patient-risk perception, and varying levels of patient-level information on prescribing decisions and the clinical-stewardship trade-off. Findings are hypothesis-generating and intended to inform the design of stewardship interventions and surveillance investments, not to prescribe specific clinical protocols.

5.5 Future Work

Several natural extensions of the `abx_amr_simulator` framework could substantially broaden its applicability and scientific scope.

5.5.1 Removing timestep awareness from hierarchical agents

As discussed in Sections 5.3 and 5.4.6, hierarchical agent performance in Experiment Sets 1 and 2 was likely influenced by finite-horizon exploitation, an unintended consequence of providing the manager agent with timestep information during training. To more accurately characterize how AMR information degradation affects hierarchical policy learning, we plan to remove timestep awareness from the manager's observation, and instead handle episode-boundary macro-action truncation through boundary-aware clipping in the wrapped environment. This modification would allow cleaner isolation of the effect of information quality on learned policy quality, and provide empirical grounding for distinguishing genuine long-horizon policy learning from fixed-horizon artifacts.

5.5.2 Multi-agent, multi-locale environments

We plan to extend the framework to support multiple agents operating across geographically distinct locales. In this configuration, each locale would represent a separate community with its own AMR dynamics,

and patients could migrate between locales, potentially spreading resistant infections. Individual agents would represent clinicians or healthcare facilities with limited visibility into peers' prescribing practices — observing only their own patient encounters and local antibiogram data. This extension would enable investigation of emergent dynamics in decentralized prescribing ecosystems, including how local optimization by individual agents affects community-level resistance, whether coordination mechanisms could improve collective outcomes, and how informational asymmetries across locales shape optimal strategies.

5.5.3 Non-stationary dynamics

Real-world patient populations and resistance mechanisms evolve over time, yet our current model treats both as stationary. Extending the simulator to support non-stationary dynamics would enable investigation of clinically important questions: How frequently should prescribing guidelines be updated to remain effective? What signals indicate that a guideline has become outdated? How much avoidable harm accumulates when policies fail to adapt to shifting patient demographics or pathogen ecology? We plan to address this by implementing time-varying patient attribute distributions to model demographic shifts, seasonal variation in infection prevalence, and emergence of high-risk subpopulations. We also plan to extend the leaky-balloon resistance dynamics to support regime-dependent behavior — for instance, modeling scenarios where sustained intensive antibiotic use creates irreversible changes in resistance mechanisms, such that the balloon accumulates a permanent baseline pressure that cannot be fully relieved even with extended periods of antibiotic restriction.

5.5.4 Upper bounds on personalized antibiogram prediction

Recent literature has explored machine learning approaches to predict patient-specific antibiotic susceptibilities from electronic medical record data, with the goal of improving empiric therapy selection (Al Mazrouei et al. 2025). While this body of work has grown substantially and spans a range of clinical settings, predictive models have generally achieved moderate discriminative performance (Sakagianni et al. 2023), leaving open a more fundamental question: even with perfect patient-level susceptibility prediction, how much collective clinical benefit could actually be achieved at the population level?

Our simulation framework is well positioned to address this question. By extending the patient model to include individual-level antibiotic susceptibility profiles, we can compare outcomes when RL agents have access to population-level antibiogram data alone versus perfect oracle knowledge of each patient's true susceptibility — directly quantifying the theoretical ceiling on collective clinical benefit from personalized prediction under realistic conditions, including noisy observations, delayed resistance data, and heterogeneous patient populations. If this ceiling is high, it would justify continued investment in improving model performance despite currently modest discriminative metrics. If even perfect personalized prediction yields only marginal improvement over population-level antibiograms, it would suggest that resources might be more productively directed toward other stewardship interventions. We could further examine how this theoretical benefit varies across clinical contexts — including baseline resistance prevalence, population heterogeneity, and available antibiotic options — to identify settings where personalized prediction tools would provide the greatest value.

5.6 Conclusions

This study demonstrates how the `abx_amr_simulator` library can serve as a quantitative testbed for examining how information deficiencies and structural uncertainty shape antibiotic prescribing strategy performance under controlled conditions. Across four experiment sets of increasing complexity, we showed that hierarchical reinforcement learning agents can learn conservative, clinically effective prescribing policies — outperforming both value-iteration benchmarks and fixed prescribing rules across clinical and stewardship metrics — without requiring explicit antimicrobial resistance penalties in the reward function. The framework supports directional analysis of factors including delayed or noisy antibiogram data, biased patient-risk

perception, and differential access to patient-level clinical information, and their effects on the clinical-stewardship trade-off inherent in antibiotic prescribing. While findings are not intended as direct clinical guidance, they provide empirically grounded, hypothesis-generating insights that can inform stewardship program design, surveillance investment priorities, and the value of improving clinical risk assessment tools.

References

Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. "Optuna: A Next-Generation Hyperparameter Optimization Framework." In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–31.

Al Mazrouei, Nadia, Asim Ahmed Elnour, Safaa Badi, Fahad T. Alsulami, Ali Awadallah Mohamed Saeed, Khalid Awad Al-Kubaisi, Vineetha Menon, et al. 2025. "The Impact of Artificial Intelligence on the Prescribing, Selection, Resistance, and Stewardship of Antimicrobials: A Scoping Review." *BMC Infectious Diseases*, December. https://doi.org/10.1186/s12879-025-12320-4.

America (IDSA), Infectious Diseases Society of. 2011. "Combating Antimicrobial Resistance: Policy Recommendations to Save Lives." *Clinical Infectious Diseases* 52 (suppl_5): S397–428.

Barlam, Tamar F, Sara E Cosgrove, Lilian M Abbo, Conan MacDougall, Audrey N Schuetz, Edward J Septimus, Arjun Srinivasan, et al. 2016. "Implementing an Antibiotic Stewardship Program: Guidelines by the Infectious Diseases Society of America and the Society for Healthcare Epidemiology of America." *Clinical Infectious Diseases* 62 (10): e51–77.

Bertollo, Leandro G, Diego S Lutkemeyer, and Anna S Levin. 2018. "Are Antimicrobial Stewardship Programs Effective Strategies for Preventing Antibiotic Resistance? A Systematic Review." *American Journal of Infection Control* 46 (7): 824–36.

Harandi, Hamid, Maryam Shafaati, Mohammadreza Salehi, Mohammad Mahdi Roozbahani, Keyhan Mohammadi, Samaneh Akbarpour, Ramin Rahimnia, Gholamreza Hassanpour, Yasin Rahmani, and Arash Seifi. 2025. "Artificial Intelligence-Driven Approaches in Antibiotic Stewardship Programs and Optimizing Prescription Practices: A Systematic Review." *Artificial Intelligence in Medicine* 162 (April): 103089. https://doi.org/10.1016/j.artmed.2025.103089.

Haredasht, Fateme Nateghi, Manoj V. Maddali, Stephen P. Ma, Amy Chang, Grace Y. E. Kim, Niaz Banaei, Stanley Deresinski, Mary K. Goldstein, Steven M. Asch, and Jonathan H. Chen. 2025. "Enhancing Antibiotic Stewardship: A Machine Learning Approach to Predicting Antibiotic Resistance in Inpatient Care." *AMIA Annual Symposium Proceedings* 2024 (May): 857–64. https://pmc.ncbi.nlm.nih.gov/articles/PMC12099390/.

King, Eshan S., Anna E. Stacy, Davis T. Weaver, Jeff Maltas, Rowan Barker-Clarke, Emily Dolson, and Jacob G. Scott. 2025. "Fitness Seascapes Are Necessary for Realistic Modeling of the Evolutionary Response to Drug Therapy." *Science Advances* 11 (24): eadv1268. https://doi.org/10.1126/sciadv.adv1268.

Laxminarayan, Ramanan, Adriano Duse, Chand Wattal, Anita KM Zaidi, Heiman FL Wertheim, Nithima Sumpradit, Erika Vlieghe, et al. 2013. "Antibiotic Resistance—the Need for Global Solutions." *The Lancet Infectious Diseases* 13 (12): 1057–98.

Lee, Joyce. 2026. "abx_amr_simulator (Version 0.1.0)." Zenodo. https://doi.org/10.5281/ZENODO.18969700.

Leth, Frank van, and Constance Schultsz. 2023. "Unbiased Antimicrobial Resistance Prevalence Estimates Through Population-Based Surveillance." *Clinical Microbiology and Infection* 29 (4): 429–33.

Murray, Christopher JL, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, Authia Gray, Chieh Han, et al. 2022. "Global Burden of Bacterial Antimicrobial Resistance in 2019: A Systematic Analysis." *The Lancet* 399 (10325): 629–55.

Organization, World Health. 2022. "Global Antimicrobial Resistance and Use Surveillance System (GLASS) Report 2022." World Health Organization.

Sakagianni, Aikaterini, Christina Koufopoulou, Georgios Feretzakis, Dimitris Kalles, Vassilios S Verykios, and Pavlos Myrianthefs. 2023. "Using Machine Learning to Predict Antimicrobial Resistance a Literature Review." *Antibiotics* 12 (3): 452.

Schweitzer, Valentijn A, Inger van Heijl, Cornelis H van Werkhoven, Jasmin Islam, Kelly D Hendriks-Spoor, Julia Bielicki, Marc JM Bonten, et al. 2019. "The Quality of Studies Evaluating Antimicrobial

Stewardship Interventions: A Systematic Review." *Clinical Microbiology and Infection* 25 (5): 555–61.

Sutton, Kent F, and Lucas W Ashley. 2024. "Antimicrobial Resistance in the United States: Origins and Future Directions." *Epidemiology & Infection* 152: e33.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. Second edition. Adaptive Computation and Machine Learning. Cambridge, MA, USA: MIT Press. http://incompleteideas.net/book/the-book-2nd.html.

Truong, William R, Levita Hidayat, Michael A Bolaris, Lee Nguyen, and Jason Yamaki. 2021. "The Antibiogram: Key Considerations for Its Development and Utilization." *JAC-Antimicrobial Resistance* 3 (2): dlab060.

Van Boeckel, Thomas P, Charles Brower, Marius Gilbert, Bryan T Grenfell, Simon A Levin, Timothy P Robinson, Aude Teillant, and Ramanan Laxminarayan. 2015. "Global Trends in Antimicrobial Use in Food Animals." *Proceedings of the National Academy of Sciences* 112 (18): 5649–54.

Wang, Xinyu, Luyuan Nong, Gioia Schaar, Belinda Koenders, Martijs Jonker, Wim de Leeuw, and Benno H Ter Kuile. 2025. "Collateral Sensitivity and Cross-Resistance in Six Species of Bacteria Exposed to Six Classes of Antibiotics." *Microbiology Spectrum*, e00983–25.

Weaver, Davis T, Eshan S King, Jeff Maltas, and Jacob G Scott. 2024. "Reinforcement Learning Informs Optimal Treatment Strategies to Limit Antibiotic Resistance." *Proceedings of the National Academy of Sciences* 121 (16): e2303165121.

WHO. 2015. "Global Action Plan on Antimicrobial Resistance." Geneva: World Health Organization. https://iris.who.int/handle/10665/193736.