Reinforcement Learning for Adaptive Tacrolimus Dosing with Drug-Drug Interaction Management

Apoorva Srinivasan [1] **Jake Patock** [1 2] **Jacob Berkowitz** [1 2] **Nicholas P. Tatonetti** [1]

Abstract

Tacrolimus is a widely used immunosuppressive drug routinely given to patients following solid organ transplantation. However, its narrow therapeutic window (4–15 ng/mL), high inter-patient pharmacokinetic variability, and frequent exposure to CYP3A mediated co-medications can make dosing challenging. We formulate tacrolimus dosing with drug-drug interaction (DDI) management as a Partially Observable Markov Decision Process (POMDP) and train a LSTM-augmented Proximal Policy Optimization (LSTM-PPO) model in a two-compartment mechanistic PK simulator covering four clinically relevant CYP3A4 inhibitors. The agent achieves 93.1% time in therapeutic window (TITW), a 13.9 percentage-point improvement over a proportional therapeutic drug monitoring controller. Against a memoryless ablation, the agent reduces toxic concentration events eight-fold and rejection risk six-fold, demonstrating that memory prevents extreme concentration excurcions in DDI patients. In EHR-grounded counterfactual simulation of 6,394 Cedars Sinai tacrolimus patients, the policy achieved 90.2% simulated TITW versus 84.5% observed clinical TITW, with positive differences among reported demographics and clinical strata. These results support memory-augmented RL in mechanistic PK simulators can optimize tacrolimus dosing under partially observed DDI dynamics and patient characteristics.

[1]Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA [2]Cedars-Sinai Health Sciences University, Los Angeles, CA, USA. Correspondence to: Nicholas P. Tatonetti <nicholas.tatonetti@cshs.org>, Apoorva Srinivasan <apoorva.srinivasan@cshs.org>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Tacrolimus is a calcineurin inhibitor widely used in solid organ transplantation. By inhibiting calcineurin-mediated T-cell activation, tacrolimus prevents immune-mediated rejection of transplanted organs (Liu et al., 1991). However, dosing is notoriously difficult due to the drug's narrow therapeutic window (typically 4–15 ng/mL) and high pharmacokinetic (PK) variability across and within patients (Henkel et al., 2023). Thus, it requires frequent monitoring of the whole-blood concentration to achieve therapeutic levels. Even small changes in systemic exposure can be clinically significant. Concentrations below the therapeutic window risks both acute and chronic graft rejection; concentrations above it can cause nephrotoxicity, neurotoxicity, diabetes, hypertension, and increased susceptibility to opportunistic infections (Kaye et al., 2024; Larsson et al., 2026). Balancing these competing risks requires highly individualized and adaptive dosing adjustments.

The current standard of care is therapeutic drug monitoring (TDM). It involves measuring pre-dose trough concentrations to guide dosing adjustments (Tolou-Ghamari & Palizban, 2025). Monitoring frequency is highest during early post-transplant phase which could include daily measurements and it tapers down to quarterly during maintenance phase (Kasiske et al., 2000). TDM has several drawbacks which makes it less than ideal for drug dosing. First, it is inherently reactive and lagged. Second, the pre-dose trough is merely a surrogate marker for total drug exposure (the AUC). Lastly, the dosing adjustments in TDM are empirical, relying on simple rules like proportional controller which does not take into account the complex non-linear dynamics of the patient states.

A particularly under-addressed challenge is the management of drug–drug interactions (DDIs). Tacrolimus is almost exclusively metabolized by CYP3A enzymes (principally CYP3A4 and CYP3A5) in the liver and intestinal wall (Van Gelder, 2002). Consequently, co-administered drugs that inhibit these enzymes cause tacrolimus to accumulate, with trough increases ranging from 1.5 to 10-fold or greater (Jiang et al., 2025). Transplant patients are on tacrolimus for life which means DDI exposure is a lifelong risk. Common inhibitors include antifungals (fluconazole,

voriconazole), calcium channel blockers (verapamil, diltiazem), and macrolide antibiotics (erythromycin) (Kriegl et al., 2022; Chow & Jusko, 2004a; Paterson & Singh, 1997). Current clinical guidance for DDI management consists largely of qualitative recommendations without systematic computational support or validation (Lempers et al., 2015).

Reinforcement learning (RL) offers a practical approach for sequential dosing under uncertainty, optimizing clinical objectives without requiring explicit specification of all patient parameters. With Long Short-Term Memory (LSTM), RL agent maintains memory of a patient's trough trajectories, enabling implicit inference of hidden pharmacokinetic state including CYP3A4 inhibition level and patient-specific characteristics.

Previous work on RL for drug dosing has shown promise for anticoagulants (Anzabi Zadeh et al., 2023), vasopressors (Komorowski et al., 2018), and insulin dosing (Zhu et al., 2021). This gap exists because drug-drug interactions (DDIs) in tacrolimus present fundamentally different challenges. Unlike other applications, the pharmacokinetic disruption here is highly dynamic rather than static. An effective dosing policy must account for multiple overlapping inhibitor mechanisms, all of which unfold and resolve over a timescale of days to weeks.

In this work, our main contributions are:

1. We formulate tacrolimus dose adjustment under CYP3A-mediated drug-drug interactions as a partially observable sequential decision problem, where patient-specific pharmacokinetics and characteristics are latent and must be inferred from observed trough trajectories.
2. We develop a literature-calibrated PK simulator for tacrolimus dosing under multiple clinically relevant CYP3A4 inhibitors, and use it to train and evaluate recurrent reinforcement learning policies against rule-based, model-based, and memory-less baselines.
3. We conduct an EHR-informed counterfactual analysis on 6,394 real patients from the Cedars-Sinai OMOP cohort, across the major transplant types, demonstrating consistent improvement across demographic and clinical subgroups.

2. Methods

2.1. Mechanistic PK Simulator Environment

We developed a two-compartment pharmacokinetic simulator with a mechanistic CYP3A4 drug-drug interaction layer to serve as the training environment for the RL agent. The model tracks drug through a gut absorption compartment and central and peripheral disposition compartments, governed by well-established ordinary differential equations (Appendix Section A) (Pang & Rowland, 1977; Rowland et al., 1973). Drug-drug interactions are modeled as time-varying suppression of hepatic CYP3A4 activity, which scales apparent clearance (CL/F), accounting for both reversible competitive inhibition and mechanism-based inactivation (MBI) (Venkatakrishnan et al., 2007). Four clinically relevant CYP3A4 inhibitors were modeled spanning a range of inhibition mechanisms and magnitudes: fluconazole (He et al., 2020), verapamil (**?**), clarithromycin (Cheung & Senior, 2016) and erythromycin (Chiang et al., 2019). Drug-specific inhibition parameters were extracted and the models validated from published PK models and in vitro studies (Appendix Table 3 and Appendix Table 4).

2.2. Patient Variability

We sampled virtual patients from validated population distributions for each simulated episode. CYP3A5 genotypes (**1/*1*, **1/*3*, **3/*3*) were drawn from population frequencies and applies discrete clearance multipliers, accounting for 40–50% of total PK variability (Hesselink et al., 2003; Thölking et al., 2014). Residual inter-individual variability is captured by a log-normal clearance multiplier (CV 50%), with weight-scaled volume of distribution (Staatz & Tett, 2004). Stochastic episode-level variability includes bioavailability noise (V. et al., 2024), assay imprecision (Christians et al., 2015; Koster et al., 2013), food effects (Bekersky et al., 2001), and non-adherence (Prendergast & Gaston, 2010; Dew et al., 2007), collectively producing trough variability consistent with published transplant cohorts (Appendix A.5).

2.3. POMDP Formulation

We formulate the tacrolimus dosing problem as a Partially Observable Markov Decision Process (POMDP). At each timestep, the agent observes 8-dimensional vector: normalized trough concentration, DDI active flag, normalized episode day, dose statistics (mean/SD over last 5 steps), last dose, days since DDI start, and trough velocity (delta from previous trough).

The hidden state includes DDI drug identity, drug-specific potency parameters (K_i, k_{inact}, K_I), patient pharmacokinetic parameters (CL, V_d), CYP3A5 genotype, missed doses, food effects, and true CYP3A4 enzyme activity. This is never observed by the agent and it has to be inferred. This set up reflects clinical reality, where clinicians observe only trough concentrations and medication lists without access to underlying enzymatic state.

The action space consists of 7 discrete dose levels expressed as fractions of the 2.0 mg BID baseline dose: $\{0.0, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50\}$, corresponding to actual doses of $\{0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$ mg. The agent selects one dose per 12 h dosing interval (two

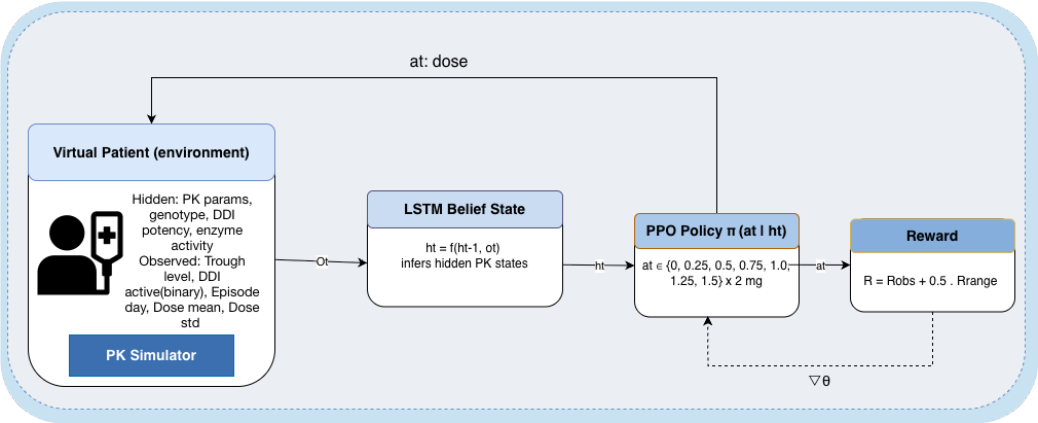

Figure 1. POMDP decision loop.

decisions per simulated day). Each step thus corresponds to a 12 h interval, and episodes run for 45 days (90 steps).

2.4. Reward Function

The reward function was designed with the treatment goal of maintaining the tacrolimus trough within their target ranges while avoiding severe toxicities. It encodes two priorities– to keep each trough measurement within the therapeutic window and to sustain control across the full episode. It combines an immediate trough signal with a cumulative episode signal: $R_{\text{total}} = R_{\text{obs}} + 0.5 \times R_{\text{range}}$. R_{obs} rewards troughs within the therapeutic window [4–15 ng/mL] with a score scaled toward the clinical optimum (9.5 ng/mL), applies graded penalties in the sub-toxic out-of-range zones [3–4 and 15–20 ng/mL], and a fixed penalty of -3.0 at or beyond the critical thresholds ($C \leq 3$ or $C \geq 20$ ng/mL), treating severe safety violations as categorical rather than graded errors. $R_{\text{range}} = 2 \times (\text{steps in range}/\text{total steps}) - 1$ incentivizes the agent to maintain sustained therapeutic control across the episode rather than reactive recovery after excursions.

2.5. LSTM-PPO Architecture

We train a Proximal Policy Optimization (PPO) agent augmented with Long Short-Term Memory (LSTM), implemented via Stable-Baselines3-Contrib RecurrentPPO. The LSTM hidden state (h_t, c_t) functions as a learned belief state over unobserved pharmacokinetic parameters, implicitly inferring patient-specific characteristics and enzyme activity from trough trajectory. The actor outputs a probability distribution over 7 discrete dose levels; the critic estimates expected cumulative reward $V(s_t)$. Full details in Appendix Table 6.

2.6. Baseline Methods

We compare against five baseline methods. *Constant Dose* (2.0 mg BID uniformly), *Random* (uniform sampling), *Proportional Controller* ($K_p = 0.5$), *Model Predictive Control* (forward PK simulation with population parameters), and *PPO Vanilla (MLP)* (feedforward policy without LSTM). Full details in Appendix Section B.

2.7. Evaluation and EHR-Based Counterfactual Simulation

Primary evaluation used 500,000 simulated steps per method, reporting mean TITW, toxic events ($C \geq 20$ ng/mL), and rejection risk ($C \leq 3$ ng/mL) with 95% bootstrap CIs.

The EHR-informed counterfactual simulation used 6,394 tacrolimus patients from the Cedars-Sinai OMOP Clinical Data Warehouse (CDM v5.4; snapshot 2024-12-10; coverage 2015–2024), comprising 312,798 trough measurements over a median follow-up of 2.3 years. Inclusion criteria required ≥3 tacrolimus measurements, adult age (≥18), and valid trough concentrations (0–100 ng/mL). Patients on concurrent cyclosporine were excluded. DDI exposure was defined as a concurrent prescription of a CYP3A4 inhibitor overlapping the measurement date by ≥3 days.

3. Results

3.1. Main Performance Comparison

Table 1 presents performance metrics for all methods. The LSTM-PPO agent achieves 93.1% TITW (95% CI: 91.5–94.7%), outperforming all baseline methods.

MPC, despite having direct access to the PK model equations, achieves only 74.4% TITW, showing that accurate inference of patient-specific parameters is required, which the LSTM learns implicitly through trajectory observation.

Table 1. Performance metrics over simulated evaluation episodes. TITW is the percentage of evaluated timepoints with trough concentration within the therapeutic window (4–15 ng/mL). Toxic Episodes and Rejection-risk Episodes report the percentage of episodes with at least one trough concentration $\geq$ 20 ng/mL or $\leq$ 3 ng/mL, respectively.

Method	TITW (%)	Toxic Episodes (%)	Rejection-risk Episodes (%)
LSTM-PPO	**93.1 (91.5–94.7)**	**0.06 (0.04–0.07)**	**2.51 (2.41–2.61)**
PPO Vanilla	92.6 (91.0–94.2)	0.47 (0.43–0.51)	15.32 (15.08–15.55)
Proportional	79.2 (76.7–81.7)	0.83 (0.77–0.89)	6.08 (5.92–6.23)
MPC	74.4 (71.7–77.1)	0.63 (0.58–0.68)	9.23 (9.04–9.42)
Const. Dose	58.1 (55.0–61.2)	9.93 (9.74–10.13)	14.60 (14.36–14.83)
Random	56.8 (53.7–59.9)	9.60 (9.41–9.79)	16.89 (16.65–17.13)

Comparing LSTM-PPO against PPO Vanilla isolates the recurrent memory mechanism's contribution. The two methods achieve statistically indistinguishable TITW (93.1% vs. 92.6%), suggesting a memoryless policy can match *average* performance through an implicitly conservative dosing strategy. However, there is a six-fold difference in rejection risk (2.51% vs. 15.32%) and an eight-fold difference (0.06% vs. 0.47%) in toxic event rate. This implies that memory specifically prevents the tail failures in DDI or unusual PK phenotypes rather than improving average performance.

3.2. EHR-Informed Counterfactual Simulation

To assess the policy on a realistic patient population, we ran a counterfactual simulation using the Cedars-Sinai OMOP cohort which included 6,394 tacrolimus patients with a median of 49 measurements per patient. The cohort comprised of kidney (68%), liver (19%), heart (8%), and lung (5%) transplant recipients. Clinical TITW across the full cohort was 84.5% (CI: 83.9–85.2%). The RL agent achieved 90.2% , an improvement of 5.7 percentage points.

Table 2 presents counterfactual results stratified by patient characteristics and DDI exposure. The simulated improvement over observed clinical TITW is positive across all demographic strata. The largest apparent gaps occur in subgroups defined by poor or variable baseline control: patients with poor baseline control (clinical TITW <50%) show a 49.0pp gap, and high-variability patients a 9.1pp gap (vs 2.3pp for low-variability). These strata are, however, defined by low or unstable observed TITW, so a substantial portion of the apparent gap reflects regression to the mean rather than policy performance alone; we therefore read them as upper bounds. Across DDI conditions the policy maintains >91% simulated TITW, with the strongest mechanism-based inhibitors (erythromycin, clarithromycin) showing the largest gaps.

Table 2. EHR-grounded counterfactual results. *Top:* Subgroup analysis (n = 6,394). *Bottom:* DDI stratification. Clin. is observed clinical TITW; RL is simulated TITW under the learned policy; Δpp is their difference in percentage points, computed from unrounded values.

Subgroup Analysis				
Subgroup	n	Clin.	RL	Δpp
Age <50y	2,024	86.1	90.1	+4.0
Age 50–65y	2,345	85.3	90.1	+4.7
Age $\geq$65y	2,025	82.0	90.6	+8.5
Male	3,919	84.5	90.0	+5.5
Female	2,475	84.5	90.6	+6.1
Poor (<50%)	173	38.3	87.4	+49.0
Mod. (50–75%)	898	67.7	90.9	+23.2
Good (>75%)	5,323	88.9	90.2	+1.4
Low var.	3,198	87.2	89.5	+2.3
High var.	3,196	81.8	91.0	+9.1
DDI Stratification				
DDI Drug	n	Clin.	RL	Δpp
Verapamil	1,912	85.0	91.8	+6.8
Fluconazole	282	86.7	91.1	+4.4
Erythromycin	164	81.6	91.3	+9.7
Clarithromycin	61	81.8	93.8	+12.0

4. Discussion

We developed a memory-augmented RL framework for tacrolimus dose adjustment under CYP3A-mediated drug-drug interactions, evaluated both on simulated episodes and on an EHR-grounded counterfactual simulation of 6,394 tacrolimus-treated patients. In simulation, the agent achieves 93.1% TITW with eight-fold toxic event reduction and six-fold rejection risk reduction. The LSTM maintains memory of trough trajectories, which allows inference of hidden pharmacokinetic state and anticipation of delayed DDI dynamics. Memoryless policies achieve similar average TITW through conservative dosing but fail for DDI and unusual-PK patients.

In the EHR-grounded counterfactual simulation, the policy achieved 5.7 percentage-point higher simulated TITW than observed clinical TITW, with the largest apparent gaps in subgroups defined by poor or variable baseline control. It should be notes that this strata are most susceptible to regression to the mean and should therefore be read as upper bounds rather than as policy effect alone.

There are several limitations. First, the agent is trained and evaluated in a simulator that necessarily simplifies real patient complexity. Second, the modeled DDI drug set is incomplete; several other drugs such as ritonavir and antiretroviral combinations have not been considered here. Third, the OMOP data lack observed administered doses, precluding direct comparison against real-world dosing decision. Most importantly, the EHR comparison is between observed and simulated quantities: the clinical arm uses observed EHR troughs, whereas the RL arm uses troughs generated by our own simulator under the learned policy. The agent's troughs are therefore just model predictions, so the sim-to-real gap is untested.

5. Conclusion

Memory-augmented reinforcement learning in mechanistic PK simulators offers a principled approach to precision tacrolimus dosing under partially observed DDI dynamics. The recurrent policy's greatest strength is its ability to prevent the extreme concentration excursions that represent the most significant safety gaps in current practice. Realizing this benefit clinically will require prospective validation against real patient responses, alongside more thorough ablations and an expanded DDI drug set, before any integration into clinical workflows.

References

Anderson, B. J. and Holford, N. H. G. Mechanism-based concepts of size and maturity in pharmacokinetics. *Annual Review of Pharmacology and Toxicology*, 48:303–332, 2008. doi: 10.1146/annurev.pharmtox.48.113006.094708.

Anzabi Zadeh, S., Street, W. N., and Thomas, B. W. Optimizing warfarin dosing using deep reinforcement learning. *Journal of Biomedical Informatics*, 137:104267, January 2023. ISSN 1532-0480. doi: 10.1016/j.jbi.2022.104267.

Arroyo-Currás, N., Ortega, G., Copp, D. A., et al. High-precision control of plasma drug levels using feedback-controlled dosing. *ACS Pharmacology & Translational Science*, 1(2):110–118, 2018. doi: 10.1021/acsptsci.8b00033.

Bekersky, I., Dressler, D., and Mekki, Q. A. Effect of low- and high-fat meals on tacrolimus absorption following 5 mg single oral doses to healthy human subjects. *Journal of Clinical Pharmacology*, 41(2):176–182, February 2001. ISSN 0091-2700. doi: 10.1177/00912700122009999.

Cheung, K. K.-T. and Senior, P. A. Tacrolimus toxicity in islet transplantation due to interaction with macrolides. *Clinical Diabetes and Endocrinology*, 2:2, 2016. doi: 10.1186/s40842-016-0019-7.

Chiang, L.-H., Wu, T.-H., Tsai, T.-C., and Lee, W.-C. Coadministration of erythromycin to increase tacrolimus concentrations in liver transplant recipients. *Transplantation Proceedings*, 51(5):1439–1441, 2019. doi: 10.1016/j.transproceed.2019.01.145.

Chow, F.-S. and Jusko, W. J. Immunosuppressive interactions among calcium channel antagonists and selected corticosteroids and macrolides using human whole blood lymphocytes. *Drug Metabolism and Pharmacokinetics*, 19(6):413–421, 2004a. doi: 10.2133/dmpk.19.413.

Chow, F.-S. and Jusko, W. J. Immunosuppressive interactions among calcium channel antagonists and selected corticosteroids and macrolides using human whole blood lymphocytes. *Drug Metabolism and Pharmacokinetics*, 19(6):413–421, December 2004b. ISSN 1347-4367. doi: 10.2133/dmpk.19.413.

Christians, U., Vinks, A. A., Langman, L. J., Clarke, W., Wallemacq, P., van Gelder, T., Renjen, V., Marquet, P., and Meyer, E. J. Impact of Laboratory Practices on Interlaboratory Variability in Therapeutic Drug Monitoring of Immunosuppressive Drugs. *Therapeutic Drug Monitoring*, 37(6):718–724, December 2015. ISSN 1536-3694. doi: 10.1097/FTD.0000000000000205.

Dew, M. A., DiMartini, A. F., De Vito Dabbs, A., Myaskovsky, L., Steel, J., Unruh, M., Switzer, G. E., Zomak, R., Kormos, R. L., and Greenhouse, J. B. Rates and Risk Factors for Nonadherence to the Medical Regimen After Adult Solid Organ Transplantation. *Transplantation*, 83(7):858–873, April 2007. ISSN 0041-1337. doi: 10.1097/01.tp.0000258599.65257.a6. URL https://journals.lww.com/00007890-200704150-00005.

Francke, M. I., Visser, W. J., Severs, D., de Mik-van Egmond, A. M. E., Hesselink, D. A., and De Winter, B. C. M. Body composition is associated with tacrolimus pharmacokinetics in kidney transplant recipients. *European Journal of Clinical Pharmacology*, 78(8):1273–1287, August 2022. doi: 10.1007/s00228-022-03323-0.

He, J., Yu, Y., Yin, C., Liu, H., Zou, H., Ma, J., Yang, W., Liu, Y., Zhong, L., and Chen, X. Clinically significant drug-drug interaction between tacrolimus and fluconazole in stable renal transplant recipient and literature review. *Journal of Clinical Pharmacy and Therapeutics*, 45(2):264–269, April 2020. ISSN 1365-2710. doi: 10.1111/jcpt.13075.

Henkel, L., Jehn, U., Thölking, G., and Reuter, S. Tacrolimus-why pharmacokinetics matter in the clinic. *Frontiers in Transplantation*, 2:1160752, 2023. ISSN 2813-2440. doi: 10.3389/frtra.2023.1160752.

Hesselink, D. A., van Schaik, R. H. N., van der Heiden, I. P., van der Werf, M., Gregoor, P. J. H. S., Lindemans, J., Weimar, W., and van Gelder, T. Genetic polymorphisms of the CYP3A4, CYP3A5, and MDR-1 genes and pharmacokinetics of the calcineurin inhibitors cyclosporine and tacrolimus. *Clinical Pharmacology and Therapeutics*, 74(3):245–254, September 2003. ISSN 0009-9236. doi: 10.1016/S0009-9236(03)00168-1.

Hovorka, R., Canonico, V., Chassin, L. J., et al. Nonlinear model predictive control of glucose concentration in sub-

jects with type 1 diabetes. *Physiological Measurement*, 25(4):905–920, 2004. doi: 10.1088/0967-3334/25/4/010.

Jiang, Y., Wang, L., Liu, J., Jiang, L., Li, K., Lv, Z., Zhou, S., and Shao, F. Physiologically Based Pharmacokinetic Modeling to Evaluate Drug-Drug Interactions of Tacrolimus With Ritonavir, a CYP3A Irreversible Inhibitor: Applications for Dosing Optimization in Transplant Patients. *Clinical Pharmacology in Drug Development*, 14(10):797–808, October 2025. ISSN 2160-7648. doi: 10.1002/cpdd.1572.

Kasiske, B. L., Vazquez, M. A., Harmon, W. E., et al. Recommendations for the outpatient surveillance of renal transplant recipients. *Journal of the American Society of Nephrology*, 11(Suppl 15):S1–S86, 2000.

Kaye, A. D., Shah, S. S., Johnson, C. D., et al. Tacrolimus- and mycophenolate-mediated toxicity: Clinical considerations and options in management of post-transplant patients. *Current Issues in Molecular Biology*, 47(1):2, 2024. doi: 10.3390/cimb47010002.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, November 2018. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-018-0213-5. URL https://www.nature.com/articles/s41591-018-0213-5.

Koster, R. A., Alffenaar, J.-W. C., Greijdanus, B., and Uges, D. R. A. Fast LC-MS/MS analysis of tacrolimus, sirolimus, everolimus and cyclosporin A in dried blood spots and the influence of the hematocrit and immunosuppressant concentration on recovery. *Talanta*, 115:47–54, October 2013. ISSN 1873-3573. doi: 10.1016/j.talanta.2013.04.027.

Kriegl, L., Boyer, J., Egger, M., and Hoenigl, M. Antifungal stewardship in solid organ transplantation. *Transplant Infectious Disease*, 24:e13855, 2022. doi: 10.1111/tid.13855.

Larsson, A., Saldeen, J., Cedernaes, J., et al. Decreasing tacrolimus concentrations in routine therapeutic drug monitoring data indicate adherence to updated therapeutic goals. *Biomedicines*, 14(1):94, 2026. doi: 10.3390/biomedicines14010094.

Lempers, V. J. C., Martial, L. C., Schreuder, M. F., et al. Drug-interactions of azole antifungals with selected immunosuppressants in transplant patients: strategies for optimal management in clinical practice. *Current Opinion in Pharmacology*, 24:38–44, 2015. doi: 10.1016/j.coph.2015.07.002.

Liu, J., Farmer, J. D., Lane, W. S., Friedman, J., Weissman, I., and Schreiber, S. L. Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP-FK506 complexes. *Cell*, 66(4):807–815, August 1991. ISSN 0092-8674. doi: 10.1016/0092-8674(91)90124-h.

Mayhew, B. S., Jones, D. R., and Hall, S. D. An in vitro model for predicting in vivo inhibition of cytochrome P450 3A4 by metabolic intermediate complex formation. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, 28(9):1031–1037, September 2000. ISSN 0090-9556.

Pang, K. S. and Rowland, M. Hepatic clearance of drugs. I. Theoretical considerations of a "well-stirred" model and a "parallel tube" model. Influence of hepatic blood flow, plasma and blood cell binding, and the hepatocellular enzymatic activity on hepatic drug clearance. *Journal of Pharmacokinetics and Biopharmaceutics*, 5(6):625–653, December 1977. ISSN 0090-466X. doi: 10.1007/BF01059688. URL http://link.springer.com/10.1007/BF01059688.

Paterson, D. L. and Singh, N. Interactions between tacrolimus and antimicrobial agents. *Clinical Infectious Diseases*, 25(6):1430–1440, 1997. doi: 10.1086/516138.

Prendergast, M. B. and Gaston, R. S. Optimizing medication adherence: an ongoing opportunity to improve outcomes after kidney transplantation. *Clinical journal of the American Society of Nephrology: CJASN*, 5(7):1305–1311, July 2010. ISSN 1555-905X. doi: 10.2215/CJN.07241009.

Rowland, M., Benet, L. Z., and Graham, G. G. Clearance concepts in pharmacokinetics. *Journal of Pharmacokinetics and Biopharmaceutics*, 1(2):123–136, April 1973. ISSN 0090-466X. doi: 10.1007/BF01059626.

Sheppard, L. C. Computer control of the infusion of vasoactive drugs. *Annals of Biomedical Engineering*, 8(5–6):431–434, 1980. doi: 10.1007/BF02363444.

Silverman, R. B. [10] Mechanism-based enzyme inactivators. In *Methods in Enzymology*, volume 249, pp. 240–283. Elsevier, 1995. ISBN 978-0-12-182150-0. doi: 10.1016/0076-6879(95)49038-8. URL https://linkinghub.elsevier.com/retrieve/pii/0076687995490388.

Staatz, C. E. and Tett, S. E. Clinical Pharmacokinetics and Pharmacodynamics of Tacrolimus in Solid Organ Transplantation:. *Clinical Pharmacokinetics*, 43(10):623–653, 2004. ISSN 0312-5963. doi: 10.2165/00003088-200443100-00001. URL http://link.springer.com/10.2165/00003088-200443100-00001.

Templeton, I. E., Thummel, K. E., Kharasch, E. D., Kunze, K. L., Hoffer, C., Nelson, W. L., and Isoherranen, N. Contribution of itraconazole metabolites to inhibition of CYP3A4 in vivo. *Clinical Pharmacology & Therapeutics*, 83(1):77–85, January 2008. doi: 10.1038/sj.clpt.6100230.

Thölking, G., Fortmann, C., Koch, R., Gerth, H. U., Pabst, D., Pavenstädt, H., Kabar, I., Hüsing, A., Wolters, H., Reuter, S., and Suwelack, B. The tacrolimus metabolism rate influences renal function after kidney transplantation. *PloS One*, 9(10):e111128, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0111128.

Tolou-Ghamari, Z. and Palizban, A.-A. Tacrolimus pharmacotherapy: Infectious complications and toxicity in organ transplant recipients; an updated review. *Current Drug Research Reviews*, 17:301–310, 2025. doi: 10.2174/0125899775259326231212073240.

V., S., M., R., M. S., S., and S., P. Understanding the factors influencing pharmacokinetics of tacrolimus. *International Journal of Research in Medical Sciences*, 12(5):1769–1775, April 2024. ISSN 2320-6012, 2320-6071. doi: 10.18203/2320-6012.ijrms20241273. URL https://www.msjonline.org/index.php/ijrms/article/view/13384.

Van Gelder, T. Drug Interactions with Tacrolimus:. *Drug Safety*, 25(10):707–712, 2002. ISSN 0114-5916. doi: 10.2165/00002018-200225100-00003. URL http://link.springer.com/10.2165/00002018-200225100-00003.

Venkatakrishnan, K., Obach, R. S., and Rostami-Hodjegan, A. Mechanism-based inactivation of human cytochrome P450 enzymes: strategies for diagnosis and drug–drug interaction risk assessment. *Xenobiotica*, 37(10-11):1225–1256, November 2007. ISSN 0049-8254, 1366-5928. doi: 10.1080/00498250701670945. URL http://www.tandfonline.com/doi/full/10.1080/00498250701670945.

Zhu, T., Li, K., Herrero, P., and Georgiou, P. Basal Glucose Control in Type 1 Diabetes Using Deep Reinforcement Learning: An *In Silico* Validation. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1223–1232, April 2021. ISSN 2168-2194, 2168-2208. doi: 10.1109/JBHI.2020.3014556. URL https://ieeexplore.ieee.org/document/9159862/.

A. Pharmacokinetic Simulator Equations

A.1. Two-Compartment Model with First-Order Absorption

Tacrolimus disposition is described by a two-compartment model (central and peripheral) with first-order oral absorption from a gut compartment. Drug amounts are tracked in μg and concentrations in ng/mL. All clearances and volumes are *apparent* (divided by the oral bioavailability F), as is standard for oral popPK parameterizations.

Gut compartment.

$$\frac{dA_{\text{gut}}}{dt} = -k_a\, A_{\text{gut}} \tag{1}$$

where k_a is the first-order absorption rate constant and A_{gut} is the drug amount in the gut (μg).

Central and peripheral compartments.

$$\frac{dA_c}{dt} = k_a\, A_{\text{gut}} - (k_{10} + k_{12})\, A_c + k_{21}\, A_p \tag{2}$$

$$\frac{dA_p}{dt} = k_{12}\, A_c - k_{21}\, A_p \tag{3}$$

where A_c and A_p are the amounts in the central and peripheral compartments (μg), and the micro-rate constants are

$$k_{10} = \frac{\text{CL}/F}{V_1/F}, \qquad k_{12} = \frac{Q/F}{V_1/F}, \qquad k_{21} = \frac{Q/F}{V_2/F}. \tag{4}$$

Here CL/F is apparent clearance, Q/F apparent inter-compartmental clearance, and V_1/F, V_2/F the apparent central and peripheral volumes.

Analytic solution over a dosing interval. The disposition subsystem (2)–(3) is linear and is solved exactly over each $\Delta t = 12$ h dosing interval via the matrix exponential. Writing $\mathbf{x} = (A_c, A_p)^\top$,

$$\frac{d\mathbf{x}}{dt} = \mathbf{M}\,\mathbf{x} + k_a A_{\text{gut}}(t)\,\mathbf{b}, \qquad \mathbf{M} = \begin{pmatrix} -(k_{10}+k_{12}) & k_{21} \\ k_{12} & -k_{21} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{5}$$

The gut compartment is propagated analytically by (1), giving the amount absorbed over the interval,

$$\Delta A_{\text{abs}} = A_{\text{gut}}(t) - A_{\text{gut}}(t+\Delta t) = A_{\text{gut}}(t)\big(1 - e^{-k_a \Delta t}\big). \tag{6}$$

The exact update for the disposition states is

$$\mathbf{x}(t+\Delta t) = e^{\mathbf{M}\Delta t}\,\mathbf{x}(t) + \int_0^{\Delta t} e^{\mathbf{M}(\Delta t - s)}\, k_a A_{\text{gut}}(t)\, e^{-k_a s}\,\mathbf{b}\, ds. \tag{7}$$

Because k_a is large relative to the disposition rate constants (absorption is $> 99\%$ complete within the first hour of each interval), we approximate the absorbed mass as a single input delivered at the interval midpoint, $s = \Delta t/2$:

$$\mathbf{x}(t+\Delta t) \approx e^{\mathbf{M}\Delta t}\,\mathbf{x}(t) + e^{\mathbf{M}\Delta t/2}\,\Delta A_{\text{abs}}\,\mathbf{b}. \tag{8}$$

Pre-dose trough. The observed pre-dose trough concentration is

$$C_{\text{trough}}(t) = \frac{A_c(t)}{V_1/F} \quad \text{(ng/mL)}. \tag{9}$$

Body-size (allometric) scaling. Patient-specific disposition parameters are scaled from the population typical values by allometric body-weight scaling, with exponent 1 for volumes and 0.75 for clearances (Anderson & Holford, 2008):

$$(V_1/F)_i = (V_1/F)_{\text{pop}} \left(\frac{W_i}{70}\right)^{1.0}, \qquad (V_2/F)_i = (V_2/F)_{\text{pop}} \left(\frac{W_i}{70}\right)^{1.0}, \tag{10}$$

$$(Q/F)_i = (Q/F)_{\text{pop}} \left(\frac{W_i}{70}\right)^{0.75}, \qquad (\text{CL}/F)_i = (\text{CL}/F)_{\text{pop}} \left(\frac{W_i}{70}\right)^{0.75} \eta_i, \tag{11}$$

where W_i is body weight (kg) and η_i is the dimensionless patient-level clearance multiplier (CYP3A5 genotype $\times$ log-normal inter-individual variability $\times$ time-varying covariates and DDI effect; see Eq. (20)). The absorption rate k_a is not weight-scaled.

A.2. DDI Enzyme Dynamics

CYP3A4 mediates $\approx 95\%$ of tacrolimus hepatic metabolism. The active hepatic CYP3A4 fraction $E(t) \in [0, 1]$ evolves during and after exposure to a co-administered inhibitor. Each inhibitor is treated as acting through *either* mechanism-based inactivation (MBI; $k_{\text{inact}} > 0$) *or* reversible competitive inhibition ($k_{\text{inact}} = 0$), but not both simultaneously. Over a given exposure window the inhibitor plasma concentration $[I]$ is held constant at its steady-state value, so the inhibition driving terms are time-invariant during exposure.

Mechanism-based inactivation. For MBI inhibitors (verapamil,erythromycin, clarithromycin), enzyme is irreversibly inactivated while being continuously resynthesized at the natural turnover rate k_{deg} (Silverman, 1995; Mayhew et al., 2000). Defining the pseudo-first-order inactivation rate

$$\lambda = \frac{k_{\text{inact}}\,[I]}{K_I + [I]}, \tag{12}$$

the steady-state and time course of the active fraction are

$$E_{\text{ss}} = \frac{k_{\text{deg}}}{k_{\text{deg}} + \lambda}, \tag{13}$$

$$k_{\text{onset}} = k_{\text{deg}} + \lambda, \tag{14}$$

$$E(t) = E_{\text{ss}} + \big(1 - E_{\text{ss}}\big)\, e^{-k_{\text{onset}}\, t}, \tag{15}$$

where t is time since the start of exposure, k_{inact} (h^{-1}) is the maximal inactivation rate, K_I (μM) the inhibitor concentration giving half-maximal inactivation, $[I]$ (μM) the inhibitor plasma concentration, and k_{deg} (h^{-1}) is the natural first-order turnover rate constant of CYP3A4, governing the basal degradation and resynthesis of active enzyme in the absence of inhibitor. The balance between k_{inact} and k_{deg} in (13)–(14) follows the standard turnover formulation (Mayhew et al., 2000).

Reversible competitive inhibition. For purely reversible inhibitors (fluconazole), the steady-state active fraction follows the competitive-inhibition relation

$$E_{\text{ss}} = \frac{1}{1 + [I]/K_i}, \tag{16}$$

$$E(t) = E_{\text{ss}} + \big(1 - E_{\text{ss}}\big)\, e^{-k_{\text{on}}^{\text{rev}}\, t}, \tag{17}$$

where K_i (μM) is the reversible inhibition constant. The approach to steady state is governed not by enzyme turnover but by accumulation of the inhibitor to its steady-state plasma concentration. Note that this onset constant deliberately does *not* include k_{deg}: reversible binding equilibrates without destroying enzyme, so turnover does not set the onset time scale.

Recovery. Following inhibitor discontinuation, the active fraction returns toward baseline by first-order *de novo* resynthesis:

$$E(t_{\text{rec}}) = 1 - \big(1 - E_{\text{end}}\big)\, e^{-k_{\text{deg}}\, t_{\text{rec}}}, \tag{18}$$

where E_{end} is the active fraction at the moment of discontinuation (carried forward continuously from (15) or (17), so there is no discontinuity at the switch) and t_{rec} is time since discontinuation. For MBI, recovery is resynthesis-limited and takes $\approx$ 24–72 h depending on E_{end}. For reversible inhibitors, the same first-order form approximates washout governed by inhibitor elimination (Section A.3).

Effective clearance modulation. The net effect on systemic clearance combines hepatic CYP3A4 suppression with the CYP3A4-independent elimination fraction ($\approx 5\%$):

$$\eta_{\text{DDI}}(t) = 0.95\, E(t) + 0.05. \tag{19}$$

The full patient clearance multiplier entering (11) is the product of the genotype effect, log-normal inter-individual variability, and the time-varying physiological and DDI covariates:

$$\eta_i = \eta_{\text{genotype}} \times \eta_{\text{IIV}} \times \eta_{\text{drift}}(t) \times \eta_{\text{steroid}}(t) \times \eta_{\text{illness}}(t) \times \eta_{\text{DDI}}(t). \tag{20}$$

A.3. Modeling Assumptions and Simplifications

We note the following simplications we made relative to a full-physiological PK implementation for transparency:

1. **Constant inhibitor exposure.** $[I]$ is held at its steady-state plasma value over each exposure window, so λ (Eq. 12) and E_{ss} are time-invariant during exposure. We do not model inhibitor accumulation and elimination kinetics explicitly.
2. **Single mechanism per inhibitor.** Each inhibitor is assigned to either MBI or reversible inhibition; combined reversible-plus-MBI action is not modeled, and the calibrated $k_{\text{inact}}/K_I/K_i$ values absorb this approximation.
3. **Reversible washout.** Post-discontinuation recovery for reversible inhibitors uses the resynthesis form (18) as a phenomenological approximation; in reality washout is governed by inhibitor elimination rather than enzyme turnover.

A.4. Patient Variability Parameters

Inter-individual pharmacokinetic variability is captured through the following distributions (Table 3 and Table 4).

Table 3. PK Model Parameters and Variability.

Parameter	Symbol	Distribution	Reference
Absorption rate	k_a	3.58 h^{-1} (fixed)	(Francke et al., 2022)
Volume of distribution	V_d	1.0 × Weight L/kg	(Staatz & Tett, 2004)
Baseline clearance	$CL_{int,0}$	LogNormal(22, CV=50%) L/h	(Staatz & Tett, 2004)
CYP3A5 *1/*1	CL multiplier	1.40×	(Hesselink et al., 2003)
CYP3A5 *1/*3	CL multiplier	1.21×	(Hesselink et al., 2003)
CYP3A5 *3/*3	CL multiplier	1.00× (reference)	(Hesselink et al., 2003)
Bioavailability	F	0.25 ± 0.10	(Staatz & Tett, 2004)
Unbound fraction	f_u	0.01 (fixed)	(Staatz & Tett, 2004)

Table 4. CYP3A4 Inhibitor Pharmacodynamic Parameters.

Inhibitor	K_i (μM)	k_{inact} (h^{-1})	K_I (μM)
Fluconazole	13-24	—	—
Verapamil	1.8	0.05	15.0
Erythromycin	10	0.012	8.5
Clarithromycin	2.8	0.15	6.2

Parameters extracted from (Templeton et al., 2008; Venkatakrishnan et al., 2007).

A.5. Stochastic Variability Sources

Beyond inter-individual PK variability, the simulator incorporates episode-level stochastic noise to capture real-world measurement and adherence variability:

- **Assay imprecision**: Trough measurements are subject to CV=15% analytical error (Christians et al., 2015; Koster et al., 2013)
- **Food effects**: Random meal timing introduces bioavailability variation of ±20% (Bekersky et al., 2001)
- **Non-adherence**: 5% probability of missed doses per day (Prendergast & Gaston, 2010; Dew et al., 2007)
- **Hematocrit variation**: Whole blood trough concentrations vary with hematocrit (CV=8%) (Koster et al., 2013)

These combined sources produce trough coefficient of variation consistent with clinical transplant cohorts (CV ≈ 30–45%).

B. Drug-Drug Interaction Model Validation

Table 5. DDI Fold-Change Validation: PK Model vs. Literature

Inhibitor	Mechanism	Literature Range	PK Predicted	Reference
Fluconazole	Reversible	1.5–3.0×	2.93×	(He et al., 2020)
Verapamil	Mild to Moderate MBI	2.0–3.0×	1.56×	(Chow & Jusko, 2004b)
Erythromycin	Mild to Moderate MBI	1.4–4.6×	4.36×	(Chiang et al., 2019)
Clarithromycin	Strong MBI	2.0–6.0×	5.51×	(Cheung & Senior, 2016)

MBI: Mechanism-based inactivation. Fold-change represents ratio of tacrolimus trough concentration with DDI vs. without DDI. PK predictions based on population simulation (n=100 virtual patients per drug). Literature ranges compiled from clinical studies and case reports.

C. RL Agent Architecture and Training

Table 6 presents the complete LSTM-PPO network architecture and training hyperparameters used in this work.

Table 6. LSTM-PPO Architecture and Training Hyperparameters.

Component	Details
Network architecture	
Input dimension	8
MLP layer 1	Linear(8→128) + ReLU
MLP layer 2	Linear(128→128) + ReLU
LSTM hidden units	128
LSTM layers	1
Actor head	Linear(128→7) + Softmax
Critic head	Linear(128→1)
Shared LSTM	`shared_lstm=True`
Critic LSTM	`enable_critic_lstm=False`
PPO hyperparameters	
Clip ratio ε	0.2
GAE λ	0.95
Entropy coefficient	0.01
Value function coefficient	0.5
Learning rate	3×10^{-4} (Adam)
Batch size	64
Epochs per rollout	10
n_{steps}	128
Training timesteps	500,000
Framework	
Implementation	Stable-Baselines3-Contrib
Action space	7 discrete dose levels
Dose levels (mg BID)	{0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0}
Episode length	45 days (90 steps)

D. Baseline Methods

We compare LSTM-PPO against five baseline methods spanning rule-based, control-theoretic, model-based, and learning-based approaches.

D.1. Constant Dose

The Constant Dose baseline administers a fixed dose of 2.0 mg twice daily (BID) to all patients regardless of observed trough concentrations, weight, or other patient characteristics. This represents a naive approach with no personalization

or feedback control. The 2.0 mg BID dose (4.0 mg total daily dose) was selected as it is common starting dose in clinical practice for average adult patients.

D.2. Random

The Random baseline selects doses uniformly at random from the seven discrete dose levels available to the RL agent: 0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0 mg BID. This baseline establishes a lower bound on performance and serves as a sanity check that all methods are learning meaningful dosing strategies rather than random behavior.

D.3. Proportional Controller

The Proportional (P) Controller implements a well-known feedback control strategy that adjusts dosing based on the error between current trough concentration and target (Sheppard, 1980):

$$d_{t+1} = d_t + K_p \cdot (C_{\text{target}} - C_{\text{trough},t}) \tag{21}$$

where d_t is the current dose, $C_{\text{trough},t}$ is the observed trough concentration, $C_{\text{target}} = 9.5$ ng/mL is the center of the therapeutic window, and $K_p = 0.5$ is the proportional gain. The proportional gain was tuned empirically to balance responsiveness with stability. Doses are clipped to the range [0.5, 3.0] mg BID and rounded to the nearest available discrete dose level. The controller initializes with a weight-normalized dose: $d_0 = 0.1 \times$ weight (kg) mg/day, converted to BID dosing.

This baseline represents a simple, interpretable feedback strategy commonly employed in clinical settings for drugs with narrow therapeutic windows (Arroyo-Currás et al., 2018).

D.4. Model Predictive Control (MPC)

The MPC baseline leverages the PK simulator directly to optimize dosing via forward simulation (Hovorka et al., 2004). At each decision point, MPC solves the following optimization problem:

$$d_t^* = \arg \max_{d_t,\ldots,d_{t+H-1}} \sum_{k=t}^{t+H-1} R(C_k) \tag{22}$$

where $H = 7$ days is the planning horizon, C_k is the predicted trough concentration at time step k, and $R(\cdot)$ is the same reward function used for RL training (combining immediate concentration reward and time-in-window reward).

PK Simulation: Forward predictions use the PK model equations (Appendix A) with *population-average* pharmacokinetic parameters rather than patient-specific values.

Key Limitation: MPC has privileged access to the true PK model structure and equations, which RL does not. However, MPC does not know patient-specific parameters or current DDI status. While MPC observes the DDI active flag, it does not know which specific inhibitor is being administered or its potency, and therefore cannot adjust its population-average parameters accordingly. This baseline tests whether direct model-based optimization with population parameters can outperform learned policies that implicitly infer patient-specific parameters from observed trajectories.

D.5. PPO Vanilla (MLP)

PPO Vanilla uses the same Proximal Policy Optimization algorithm and training configuration as LSTM-PPO, but replaces the recurrent LSTM architecture with a feedforward multi-layer perceptron (MLP). This helps us distinguish the contribution of recurrent memory to performance.

Architecture:

- **Input:** Same 8-dimensional observation vector as LSTM-PPO
- **Policy Network (Actor):** MLP with two hidden layers of [128, 128] units, ReLU activations, output layer with 7 units (one per dose level) and softmax activation

- **Value Network (Critic):** MLP with two hidden layers of [128, 128] units, ReLU activations, single output unit for state value $V(s)$

Training: Identical hyperparameters to LSTM-PPO (learning rate, batch size, GAE-λ, clipping, entropy coefficient) to ensure fair comparison. Training runs for the same number of environment steps (500,000) across multiple patients.

Key Difference: PPO Vanilla is *memoryless*—it cannot integrate information across multiple time steps. Each dosing decision is made based solely on the current observation, without access to the historical trough trajectory that would enable inference of patient-specific PK parameters or DDI dynamics. This makes PPO Vanilla a reactive rather than anticipatory policy.

E. EHR-Based Counterfactual Simulation: Detailed Results

Figure 2 presents a comprehensive view of the EHR-grounded counterfactual simulation on 6,394 real patients from the Cedars-Sinai OMOP cohort.

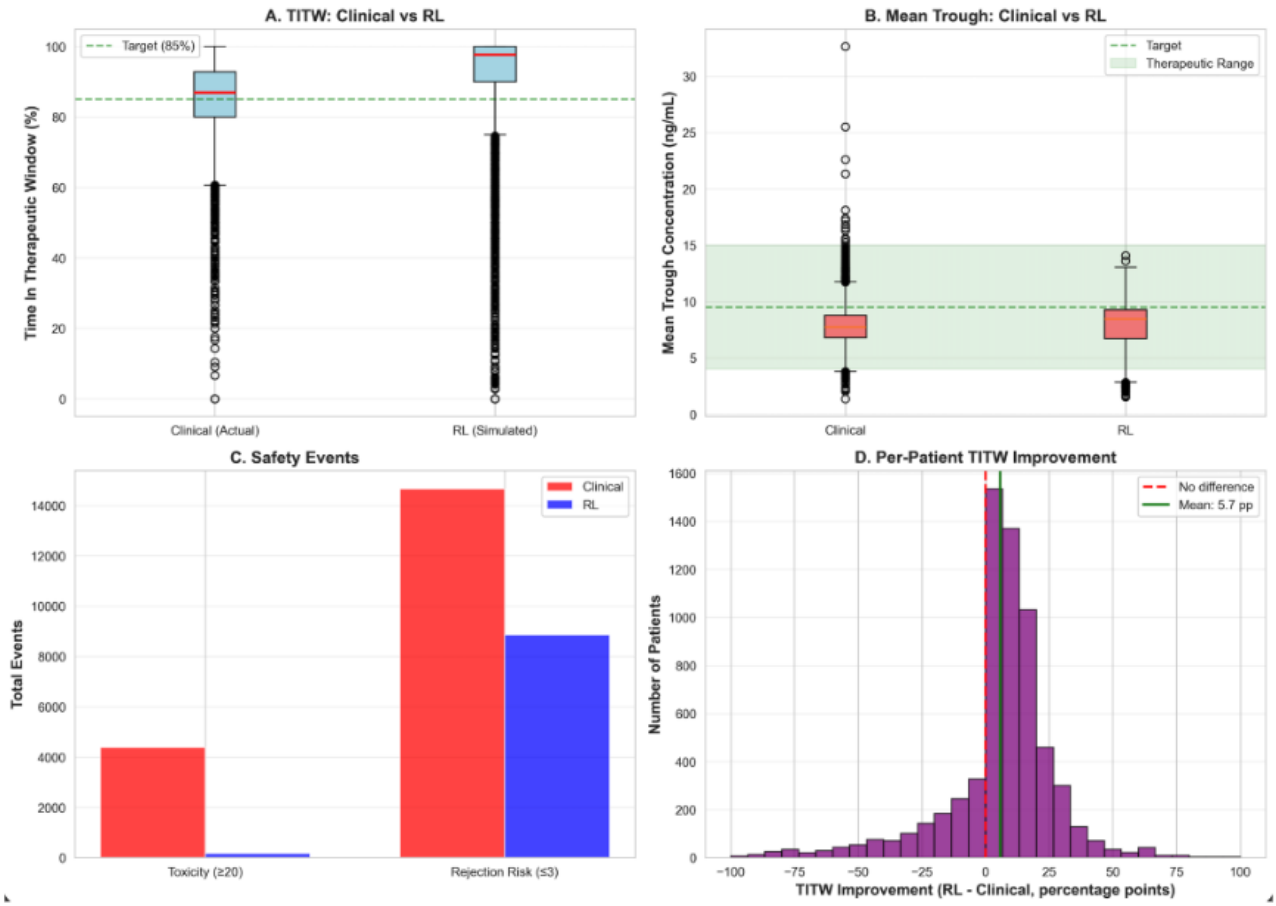

Figure 2. EHR-grounded counterfactual simulation comparing observed EHR trough trajectories with simulated trough trajectories under the learned RL policy across 6,394 patients. **(A)** Time in therapeutic window (TITW) comparison showing higher median TITW with reduced variability under the simulated policy. The target TITW of 85% (dashed line) is exceeded by the RL agent. **(B)** Mean trough concentration distributions for both approaches, demonstrating tighter control around the therapeutic range (4–15 ng/mL, shaded region) with the RL agent. **(C)** Safety events across the full cohort: the RL agent reduces both toxicity events ($C \geq 20$ ng/mL; red bars) and rejection risk events ($C \leq 3$ ng/mL; purple bars). **(D)** Per-patient distribution of the difference between simulated RL and observed clinical TITW, showing a positive gap for 90.2% of patients (5,768/6,394), with a mean of +5.7 percentage points. The distribution is right-skewed; the largest gaps fall in patients with poor baseline control, a subgroup defined by low observed TITW and therefore partly attributable to regression to the mean.