Model-Informed AI for Tacrolimus Concentration Prediction Using Outpatient Data

Gyu Seong Lee[1], On Hu Park[2], Mi Kyong Shim[3]*, Dae Keun Park[1]*,

[1]Department of AI Healthcare Convergence, General Graduate School, Cha University, Pocheon, Gyeonggido, Republic of Korea

[2]Division of Healthcare Sciences, Cha University, Pocheon, Gyeonggido, Republic of Korea

[3]College of Pharmacy, Cha University, Pocheon, Gyeonggido, Republic of Korea

*** Correspondence:**
Mi Kyong Shim
pharmsuni@cha.ac.kr
Dae Keun Park
dkpark@cha.ac.kr

Keywords: Model-Informed Artificial Intelligence, Outpatients, Personalized Medicine, Population Pharmacokinetics, Tacrolimus

Abstract

Background: Accurate prediction of tacrolimus exposure is essential for individualized dosing because tacrolimus exhibits substantial inter-individual variability and a narrow therapeutic window in outpatient care. While population pharmacokinetic (PopPK) models provide mechanistic interpretability, they may not fully capture complex patient heterogeneity and longitudinal patterns observed in real-world data. **Methods:** We developed a model-informed artificial intelligence (AI) framework that embeds PopPK-predicted concentrations as input features for data-driven models. Retrospective outpatient electronic medical records from Seoul National University Hospital were collected for tacrolimus-treated patients with glomerulonephritis. After excluding four patients with excessively repeated measurements or anomalous concentration profiles, 141 patients were included in the final analytic cohort. The PopPK model was developed in all 141 patients; the sequence-based AI models, which require repeated within-patient troughs, were developed in the subset with at least five measurements (n = 109). The dataset was split at the patient level into training, validation, and test sets at a 60:20:20 ratio. Four one-compartment PopPK models—fixed-effect, random-effect, covariate-extended, and covariate + random-effect models—were built using nlmixr2, with ka fixed to 4.5 h^{-1}, and simulated via RxODE to generate PopPK-derived predictions. AI models (RNN, LSTM, and LightGBM) were trained with DVI embedded as an additional feature. Model performance was evaluated using independent patient-level validation and test sets. Performance was assessed using RMSE and MAE; differences between models were assessed using a patient-clustered bootstrap, with F-tests as an exploratory comparison of prediction-error variance, and interpretability was evaluated using SHAP. **Results:** The final PopPK model yielded an RMSE of 3.261 ng/mL, an MAE of 2.349 ng/mL. Embedding PopPK predictions into all AI models improved predictive performance relative to the PopPK baseline model. LightGBM showed the best overall performance,

with the lowest RMSE of 2.606 ng/mL, the lowest MAE of 1.812 ng/mL. The RNN-based and LSTM-based embeddings also showed improved predictive accuracy relative to the PopPK baseline model. These improvements over PopPK were statistically significant by patient-clustered bootstrap (all paired RMSE-difference 95% confidence intervals excluded zero). SHAP analysis identified the previous prediction error, ERR_LAG1, as the most influential feature across AI models, indicating that longitudinal error correction substantially contributed to predictive improvement. PopPK-derived prediction, DVI, remained among the top-ranked predictors in LightGBM and RNN, supporting the added value of mechanistic PopPK information. **Conclusion:** Embedding PopPK-derived predictions into AI models improves tacrolimus concentration prediction in outpatients while retaining clinically meaningful interpretability. This model-informed AI approach combines mechanistic pharmacokinetic information with data-driven temporal error correction and may provide a scalable framework for personalized therapeutic drug monitoring using real-world outpatient data.

1. Introduction

Accurate prediction of drug concentrations is a cornerstone of precision dosing and therapeutic drug monitoring (TDM), particularly for drugs with a narrow therapeutic index (Cremers et al., 2016; Pérez-Blanco and Lanao, 2022; Gona et al., 2023). Tacrolimus is widely used for immunosuppression in transplantation and for immune-mediated diseases, yet it exhibits substantial inter-individual variability and a narrow therapeutic window (Campagne et al., 2019). In outpatient care, adherence variability, concomitant medications, and time-varying physiological factors such as renal function and inflammatory status further complicate exposure prediction and dose optimization (Francke et al., 2021; Herblum et al., 2021). Subtherapeutic exposure may lead to inadequate immunosuppression, whereas supratherapeutic exposure increases the risk of toxicity, including nephrotoxicity, underscoring the clinical need for robust concentration prediction in real-world outpatient settings (Bentata, 2020; Francke et al., 2021).

Population pharmacokinetic (PopPK) modeling has long served as a mechanistic framework for individualized dosing by estimating patient-specific pharmacokinetic parameters (e.g., clearance and volume) from sparse concentration measurements and enabling Bayesian forecasting (Campagne et al., 2019; Minichmayr et al., 2024). Tacrolimus PopPK studies have identified clinically relevant covariates, including body size and renal function markers, as contributors to pharmacokinetic variability (Campagne et al., 2019). However, conventional compartmental models may be constrained by structural assumptions and limited covariate sets, and they may not fully capture complex nonlinear interactions, longitudinal patterns, and real-world data irregularities such as time-varying covariates and unmeasured confounding (Maier et al., 2022; Minichmayr et al., 2024).

Recent advances in artificial intelligence (AI) and machine learning (ML) have introduced data-driven approaches that can complement mechanistic models for drug concentration prediction. Sequence models such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks can learn temporal dependencies from longitudinal data, while tree-based ensemble methods such as LightGBM can capture nonlinear relationships among structured clinical variables (Poweleit et al., 2023; Wang et al., 2024; Jian et al., 2025). Despite promising performance, purely data-driven models can be limited by reduced interpretability and potential instability in smaller real-world cohorts (Barrett, 2024).

To bridge mechanistic interpretability and data-driven flexibility, model-informed AI approaches integrate pharmacokinetic knowledge into AI models (Li et al., 2023; Barrett, 2024; Wang et al., 2024). A practical strategy is to embed PopPK-derived predictions or parameters as input features,

allowing AI models to learn residual patterns beyond what mechanistic models explain while retaining clinically meaningful constraints (Stankevičiūtė et al., 2023).

Therefore, this study aimed to develop and evaluate a PopPK-embedded AI framework for tacrolimus concentration prediction using real-world outpatient EMR data. We constructed four one-compartment PopPK models and three PopPK-embedded AI models (RNN, LSTM, and LightGBM), compared predictive performance using independent patient-level evaluation sets, assessed differences in prediction-error dispersion, and evaluated interpretability using Shapley additive explanations (SHAP). Although the combination of population pharmacokinetic modeling and machine learning has been investigated for tacrolimus, most prior studies have focused on transplant recipients with relatively dense sampling, and data-driven approaches in outpatient glomerulonephritis cohorts using sparse, irregularly sampled real-world electronic medical record data remain scarce. In the present study, we combined a mechanistic PopPK prior (the population prediction, embedded as the feature DVI) with a data-driven temporal correction based on the lagged prediction error, used SHAP analysis to decompose the relative contributions of mechanistic and temporal features, and adopted strict patient-level data splitting to prevent leakage of within-patient information.

A central design consideration was that the framework should be applicable to a new outpatient directly from routinely available covariates, without first fitting that individual's pharmacokinetic parameters. For a patient who has not yet provided a measured trough, the individual random effects cannot be updated, so an empirical-Bayes individual prediction (IPRED) is unavailable at this cold-start point. Rather than adopting sequential Bayesian re-estimation of individual parameters as troughs accumulate, we deliberately built the framework around the population (typical-value) prediction (embedded as DVI, with random effects set to zero), which is computable for any patient, and supplied data-driven individualization through the lagged prediction error (ERR_LAG1)—a lightweight correction that carries forward the deviation of a patient's most recent prior trough from the population prediction and updates automatically as clinic visits accumulate, without per-patient model fitting. This pairing preserves portability to previously unseen patients—for whom the visit-to-visit correction becomes available from the second observed trough onward—while providing a data-driven signal intended to recover patient-specific offsets that the population prior alone cannot represent.

2. Materials and Methods

2.1 Study design, ethics, and study population

This retrospective observational study was conducted using electronic medical record (EMR) data from outpatients diagnosed with glomerulonephritis and treated with oral tacrolimus at Seoul National University Hospital. The study design and data source were consistent with those used in a previously approved population pharmacokinetic investigation conducted at the same institution. The study was performed in accordance with the Declaration of Helsinki and institutional ethical standards. Institutional review and approval were obtained for retrospective analysis of anonymized EMR data, and the requirement for written informed consent was waived due to the non-interventional nature of the study.

Eligible patients were adults (≥18 years) with a diagnosis of glomerulonephritis who received oral tacrolimus capsules administered twice daily and underwent routine therapeutic drug monitoring

(TDM) in the outpatient setting. Patients who had not reached steady-state tacrolimus concentrations, according to prior pharmacokinetic study criteria at the same institution, were excluded. To ensure sufficient longitudinal information, patients with fewer than three tacrolimus trough concentration measurements were also excluded from the pharmacokinetic analysis. A more stringent measurement requirement was subsequently applied when constructing the sequence-based artificial intelligence cohort, because the recurrent and lagged-error features depend on repeated within-patient observations.

A total of 145 patients were initially identified. Four patients were excluded because of excessively repeated measurements or anomalous concentration profiles, resulting in a final analytic cohort of 141 patients on which the population pharmacokinetic (PopPK) model was developed. Because the artificial intelligence (AI) models rely on sequence and lagged prediction-error features that require repeated trough measurements per patient, the AI models were developed on the subset of 109 patients with at least five trough measurements. The 32 patients excluded from the AI cohort had only one to four troughs (median 2), compared with 5 to 56 troughs (median 14) among the 109 retained patients. Clinical, demographic, laboratory, and dosing information were retrospectively extracted from the EMR system. Missing body weight values were imputed using age-binned mean weights to ensure covariate completeness.

2.2 Data preprocessing and feature engineering

All dosing administrations and concentration measurements were organized into a patient-level event-table structure. Because outpatient EMR data are irregularly sampled, each patient timeline was discretized into a 12-hour grid. Time points without dosing or concentration measurements were explicitly represented to enable consistent longitudinal feature computation across patients.

From dosing histories and timestamps, engineered features were derived, including cumulative dose and time since last dose. For time-varying clinical covariates (e.g., laboratory values), missing values were handled using linear interpolation combined with last observation carried forward, yielding continuous covariate trajectories aligned to the discretized time grid.

To mitigate the influence of extreme values, continuous variables were winsorized and normalized prior to model training, followed by Z-score standardization. Categorical variables (e.g., sex) were one-hot encoded. Patient identifiers were excluded from model inputs and used only for patient-level data splitting to prevent information leakage.

2.3 Population pharmacokinetic (PopPK) modeling

Tacrolimus population pharmacokinetic modeling was performed using nlmixr2. Concentration–time data were described using a one-compartment model with first-order oral absorption and first-order elimination. This model structure was selected as a pragmatic approach for sparse, trough-based outpatient TDM data. The absorption rate constant (ka) was fixed at 4.5 h^{-1} based on prior literature and to stabilize parameter estimation under sparse sampling.

Four PopPK model specifications were evaluated:
(1) a fixed-effect model estimating population mean clearance and volume of distribution;
(2) a random-effect model incorporating inter-individual variability;
(3) a covariate-extended model evaluating clinically relevant covariates using a stepwise forward–

backward selection procedure; and
(4) a covariate + random-effect model combining the selected covariates with inter-individual variability.

Renal function markers, body weight, and serum albumin were retained as clinically relevant covariates influencing clearance and/or volume of distribution during stepwise selection; parameter-level Wald significance is reported in Table 3. The covariate + random-effect model was selected as the final PopPK model because it incorporated both clinically relevant covariates and inter-individual variability; its residual error was comparable to that of the random-effect model, but it additionally provided clinically interpretable covariate relationships. This modeling strategy provided the mechanistic basis for subsequent AI embedding.

2.4 Model evaluation of PopPK

Model adequacy was assessed using standard goodness-of-fit diagnostics, including comparisons between observed and model-predicted concentrations and evaluation of residual distributions. These diagnostics were used to confirm the suitability of the final PopPK model for downstream integration with AI models.

2.5 PopPK-derived prediction embedding

For each PopPK model, concentration–time profiles were simulated using RxODE with the estimated fixed-effect (population) parameters, with random effects set to zero, and patient-specific covariate and dosing event tables. The simulated concentration at each grid time point was defined as the PopPK-derived prediction, referred to as the drug variability index (DVI). DVI was incorporated as an additional input feature into AI models to implement a model-informed embedding strategy, allowing data-driven models to learn residual patterns beyond the mechanistic PopPK structure.

In addition, lagged prediction-error information was used to capture longitudinal deviation from prior predictions. For each observed trough, the prediction error was defined as the difference between the observed concentration and the DVI (observed concentration minus population prediction). ERR_LAG1 was then constructed, within each patient trajectory, by carrying the last observed prediction error forward and shifting it by one time step (a per-patient forward-fill followed by a one-step lag), so that each record was assigned the residual of the most recent preceding observed trough. The first record of each patient was set to zero. By construction, ERR_LAG1 uses only the previous observed trough's residual and never the current or any future observation, thereby precluding information leakage.

Two considerations motivated this construction. First, the embedded PopPK feature (DVI) was defined as the population (typical-value) prediction, with random effects set to zero, rather than as an individual, empirical-Bayes prediction (IPRED). An IPRED requires the patient's own measured troughs to update the individual random effects and therefore cannot be formed for a new outpatient at the first visit, before any concentration is available; embedding IPRED would require per-patient Bayesian re-estimation and would leave this cold-start case uncovered, which is incompatible with the intended applicability of the model to previously unseen patients. Individual-level information is instead injected through ERR_LAG1, a lightweight sequential correction that becomes available for

any patient once at least one prior trough exists—that is, from the second monitored visit onward, the first record being set to zero and providing no individualization. Second, ERR_LAG1 carries the deviation observed at the most recent preceding trough, which in this cohort was separated from the current trough by a median of 28 days (interquartile range, 16–42 days; across 1,707 consecutive-trough intervals, 98% were ≥ 7 days apart and only 0.2% were ≤ 24 hours apart). ERR_LAG1 therefore constitutes a visit-to-visit correction on the scale of the outpatient therapeutic drug monitoring interval, rather than a within-day update; whether a residual of this age remains informative is an empirical question addressed in the Results and Discussion.

An overview of the overall workflow, from EMR preprocessing to PopPK simulation, DVI embedding, AI training, and SHAP interpretability, is summarized in Figure 1.

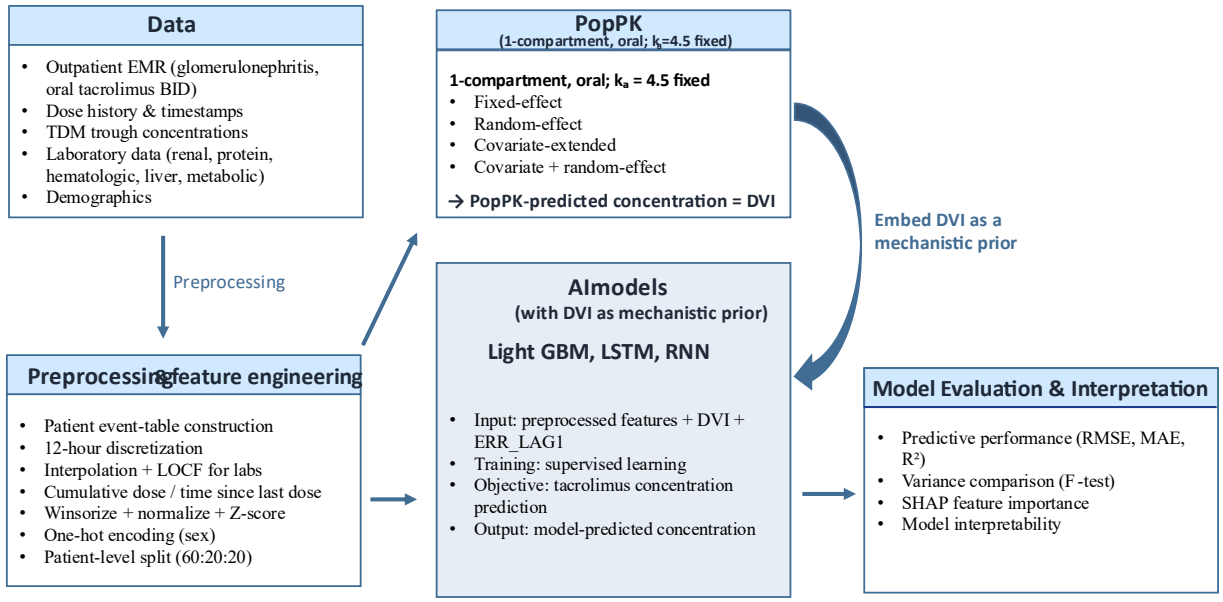

Figure 1. Graphical Abstract of the Model-informed AI framework for tacrolimus trough concentration prediction in outpatients.

Outpatient EMR data, including dosing history, demographics, and time-varying laboratory covariates, were organized into patient-level longitudinal event tables and discretized on a 12-hour grid. Four one-compartment PopPK models generated PopPK-derived predictions for each time point. The PopPK-predicted concentration, defined as DVI, was embedded as an additional feature into AI models, including LightGBM, LSTM, and RNN. Lagged prediction-error information was also used to support longitudinal error correction. Models were trained and evaluated using patient-level training, validation, and test splits. Performance was evaluated using RMSE, MAE, and variance reduction in prediction errors. Interpretability was assessed using SHAP.
Abbreviations: AI, artificial intelligence; EMR, electronic medical record; PopPK, population pharmacokinetics; DVI, drug variability index; LightGBM, Light Gradient Boosting Machine; LSTM, long short-term memory; RNN, recurrent neural network; RMSE, root mean square error; MAE, mean absolute error; SHAP, Shapley additive explanations.

2.6 Development of PopPK-embedded AI models

Three AI models were developed to predict tacrolimus concentrations at each time point: recurrent neural networks (RNN), long short-term memory (LSTM) networks, and Light Gradient Boosting Machine (LightGBM).

RNN models consisted of a single recurrent layer with a hidden size of 32, whereas LSTM models consisted of two recurrent layers with a hidden size of 64. Both sequence-based models were implemented using PyTorch and trained on multivariate time-series inputs, including dosing-derived features, time-varying clinical covariates, and embedded DVI.

LightGBM models were trained on engineered structured features, including cumulative dose, time since last dose, laboratory covariates, and DVI. Time points were treated as independent samples without explicit recurrent structure. The input features used in the PopPK-embedded AI models are summarized in Table 1.

Table 1. Input features used in PopPK-embedded AI models for tacrolimus concentration prediction

Feature category	**Feature name**	**Description / unit**	**Time-varying**	**Engineering / handling**
Mechanistic (PopPK)	**DVI**	PopPK-predicted concentration at time *t* (ng/mL)	Yes	Simulated via RxODE from PopPK; embedded as feature
Model-informed (temporal)	**ERR_LAG1**	Lagged PopPK prediction error (observed - DVI) from the most recent prior trough (ng/mL)	Yes	Per-patient forward-fill then one-step lag; first record = 0; past-only (no leakage)
Dose-derived	**CumulativeDose**	Cumulative tacrolimus dose up to time *t*	Yes	Derived from dosing history
Dose-derived	**TimeSinceLastDose**	Time since last dose (h)	Yes	Derived from dosing timestamps
Dose-derived (optional, if used)	**LastDoseAmount**	Last administered dose amount	Yes	From dosing records
Demographics	**Age**	Age (years)	No	Z-score standardization

Demographics	**Sex_Male**	Sex (one-hot)	No	One-hot encoded
Body size	**BodyWeight (BW)**	Body weight (kg)	Yes*	Missing BW imputed by age-binned mean (5-year bins); interpolated if needed
Renal function	**SerumCreatinine (SCr)**	Serum creatinine (mg/dL)	Yes	Interpolation + LOCF on 12-h grid
Renal function	**eGFR**	Estimated GFR (mL/min)	Yes	Interpolation + LOCF
Renal function	**BUN**	Blood urea nitrogen (mg/dL)	Yes	Interpolation + LOCF
Protein / nutrition	**Albumin (ALB)**	Serum albumin (g/dL)	Yes	Interpolation + LOCF
Protein / nutrition	**TotalProtein**	Total protein (g/dL)	Yes	Interpolation + LOCF
Hematology	**WBC**	White blood cell count (K/μL)	Yes	Interpolation + LOCF
Hematology	**Hemoglobin (Hgb)**	Hemoglobin (g/dL)	Yes	Interpolation + LOCF
Hematology	**Hematocrit (Hct)**	Hematocrit (%)	Yes	Interpolation + LOCF
Hematology	**Platelet (PLT)**	Platelet count (K/μL)	Yes	Interpolation + LOCF
Inflammation/others	**UricAcid**	Uric acid (mg/dL)	Yes	Interpolation + LOCF
Liver function	**AST**	AST (IU/L)	Yes	Interpolation + LOCF

Liver function	**ALT**	ALT (IU/L)	Yes	Interpolation + LOCF
Liver function	**TotalBilirubin (TBil)**	Total bilirubin (mg/dL)	Yes	Interpolation + LOCF
Metabolic	**TotalCholesterol**	Total cholesterol (mg/dL)	Yes	Interpolation + LOCF
Disease severity	**UPCR**	Urine protein/creatinine ratio	Yes	Interpolation + LOCF
Time index	**TimeIndex**	Time point indicator on discretized grid	Yes	Implicit in sequence models; optional for LightGBM

This table summarizes the clinical, demographic, laboratory, dosing-derived, and model-informed features used as inputs for the PopPK-embedded AI models. DVI represents the PopPK-derived predicted tacrolimus concentration generated by RxODE simulations. Time-varying laboratory covariates were aligned to a 12-hour discretized grid using interpolation and last observation carried forward. Continuous variables were winsorized and standardized before model training. Abbreviations: AI, artificial intelligence; PopPK, population pharmacokinetics; BW, body weight; DVI, drug variability index; LOCF, last observation carried forward; SCr, serum creatinine; eGFR, estimated glomerular filtration rate; BUN, blood urea nitrogen; ALB, albumin; WBC, white blood cell count; Hgb, hemoglobin; Hct, hematocrit; PLT, platelet count; AST, aspartate aminotransferase; ALT, alanine aminotransferase; TBil, total bilirubin; UPCR, urine protein/creatinine ratio.

All continuous variables were winsorized and normalized to reduce the influence of extreme values, followed by Z-score standardization prior to model training. Time-varying covariates were aligned to a 12-hour discretized grid and imputed using linear interpolation combined with last observation carried forward (LOCF). Patient identifiers were excluded from model inputs and used only for patient-level training, validation, and test splitting to prevent leakage.

2.7 Training–validation–test split and performance evaluation

To prevent information leakage across repeated measurements within patients, all data were split at the patient level into training, validation, and test sets in an approximate 60:20:20 ratio. The PopPK cohort (141 patients) was split into 84 training, 28 validation, and 29 test patients. The AI cohort (109 patients with at least five troughs) was split into 66 training, 22 validation, and 21 test patients; the AI test set comprised 21 patients and 332 sequence-endpoint observations. To ensure a common basis for comparison, the PopPK baseline was evaluated as the covariate + random-effect population prediction on the same 332 test endpoints used to evaluate the AI models. Because the PopPK model was developed on the full 141-patient cohort, some AI test patients may also have contributed to PopPK model fitting; however, DVI is a population (typical-value) prediction with random effects set

to zero and does not use any individual patient's observed concentrations, so this does not introduce information leakage into the AI evaluation. Model performance was evaluated using root mean square error (RMSE), mean absolute error (MAE), all computed on the nanogram-per-milliliter concentration scale. The four PopPK models were developed on the full analytic cohort of 141 patients (patient-level training, validation, and test sets of 84, 28, and 29 patients, respectively; Table 2). Because the PopPK-embedded AI models depend on longitudinal sequences to derive lagged features—most importantly the previous prediction error (ERR_LAG1), which requires a sufficient history of prior troughs—these models were developed and evaluated on a subcohort of patients with at least five trough measurements. This subcohort comprised 109 patients, partitioned at the patient level into training, validation, and test sets of 66, 22, and 21 patients, respectively (approximately 60:20:20, consistent with the ratio used for the PopPK cohort). The independent AI test set thus comprised 21 patients contributing 332 trough prediction endpoints, on which the predictive performance reported in Section 3.4 was evaluated. The independent validation and test cohorts were used to assess generalizability of the AI models relative to the final PopPK model. To quantify uncertainty while respecting the clustered structure of the data, 95% confidence intervals for each metric were obtained by patient-clustered bootstrap resampling (B = 2000 replicates), in which whole patients were resampled with replacement. As the primary inferential test, the between-model difference in RMSE (PopPK minus AI) was compared using the same patient-clustered bootstrap, and an improvement was considered statistically significant when the 95% confidence interval for the paired RMSE difference excluded zero.

As an exploratory secondary analysis, we assessed whether the AI models reduced the dispersion of prediction errors relative to the covariate + random-effect PopPK model using F-tests based on prediction-error variance estimates. These F-tests treat the prediction errors as independent observations; because the 332 test windows are derived from only 21 patients and are therefore not fully independent, the associated p-values are reported as exploratory only and should not be interpreted as formal confirmatory tests. Accordingly, the patient-clustered bootstrap comparison of RMSE differences described above served as the primary inferential basis.

2.8 Model interpretability

Model interpretability was evaluated using Shapley additive explanations (SHAP). For LightGBM, TreeExplainer was used to compute feature-level SHAP values. For sequence-based models, GradientExplainer was applied, and SHAP values were aggregated across time points to obtain global feature importance rankings. These analyses were used to identify key predictors and to assess consistency with established pharmacokinetic knowledge.

3. Results

3.1 Patient characteristics

A total of 145 patients with glomerulonephritis treated with oral tacrolimus were initially identified for this retrospective analysis. Four patients were subsequently excluded due to excessively repeated measurements or anomalous concentration profiles, resulting in a final cohort of 141 patients. The PopPK model was developed using all 141 patients, which were split at the patient level into training (n = 84, 60%), validation (n = 28, 20%), and test (n = 29, 20%) sets. Because the sequence-based AI models required a sufficient run of longitudinal troughs to construct the lagged prediction-error and sequence features, the AI cohort was restricted to the 109 patients with at least five trough

measurements (median 14, range 5–56); the 32 patients excluded from this cohort had only one to four troughs (median 2). These 109 patients were likewise split at the patient level into training (n = 66), validation (n = 22), and test (n = 21) sets, and the AI test set comprised 21 patients contributing 332 sequence-endpoint observations. Baseline demographic, clinical, and laboratory characteristics of each subset are summarized in Table 2.

The mean age was approximately 37–42 years across subsets, and male patients accounted for approximately 34–39%. The mean body weight was 63–64 kg. The mean duration of tacrolimus therapy ranged from 523 to 671 days, reflecting heterogeneous and long-term real-world outpatient exposure.

Renal function indices showed substantial inter-individual variability. The mean serum creatinine concentration ranged from 1.123 to 1.582 mg/dL, corresponding to estimated glomerular filtration rates of 73.618–83.542 mL/min. Hematologic parameters, including hematocrit and hemoglobin, were within expected outpatient ranges but exhibited broad dispersion. **Tacrolimus trough concentrations also demonstrated marked variability, with subset means ranging from 3.472 to 3.914 ng/mL and individual values ranging from 0.520 to 13.300 ng/mL, underscoring the challenge of exposure control in this population.**

The etiologies of glomerulonephritis were heterogeneous. Lupus nephritis, IgA nephropathy, and minimal change disease were the most common diagnoses, followed by other immune-mediated and proliferative glomerular diseases.

Table 2. Baseline characteristics of the training, validation, and test datasets.

Characteristic	Training data set (60%)	Validation data set (20%)	Test data set (20%)
Number of patients	84	28	29
Sex, male n (%)	33 (39.3%)	10 (35.7%)	10 (34.5%)
Age, years	37.722 ± 13.375 (17.857–69.827)	36.787 ± 14.938 (18.917–75.486)	42.120 ± 13.509 (17.333–64.375)
Body weight, kg	64.436 ± 11.567 (38.872–114.950)	62.943 ± 9.567 (38.900–84.013)	64.500 ± 11.389 (42.300–87.033)
Period on tacrolimus, days	581.341 ± 379.540 (12.000–1687.004)	670.716 ± 579.407 (14.000–2225.004)	523.001 ± 473.209 (28.004–2165.004)
Tacrolimus trough concentration, ng/mL	3.826 ± 2.384 (0.520–13.300)	3.914 ± 2.113 (0.600–8.655)	3.472 ± 1.783 (0.600–6.840)

Urine protein/creatinine ratio	1.825 ± 1.632 (0.086–6.391)	1.686 ± 1.306 (0.045–4.891)	1.805 ± 2.207 (0.051–11.163)
White blood cell count, K/μL	9.332 ± 2.443 (5.195–15.825)	10.038 ± 3.098 (3.815–16.463)	9.035 ± 2.158 (4.768–14.416)
Hemoglobin, g/dL	13.408 ± 1.565 (8.250–16.820)	13.755 ± 1.708 (9.842–17.533)	13.162 ± 1.590 (10.425–16.400)
Hematocrit, %	40.182 ± 4.246 (24.583–48.800)	41.418 ± 4.566 (29.584–51.250)	39.361 ± 4.009 (31.800–47.429)
Platelet count, K/μL	258.458 ± 60.516 (120.375–404.000)	254.015 ± 54.619 (103.976–360.737)	229.544 ± 53.557 (156.125–392.750)
Blood urea nitrogen, mg/dL	19.872 ± 8.295 (7.250–50.500)	21.016 ± 13.026 (9.250–61.474)	26.673 ± 21.980 (8.375–108.000)
Serum creatinine, mg/dL	1.123 ± 0.596 (0.476–4.462)	1.182 ± 0.679 (0.622–3.660)	1.582 ± 1.931 (0.541–10.851)
Estimated GFR, mL/min	83.542 ± 28.649 (10.054–132.129)	81.730 ± 32.640 (20.237–129.774)	73.618 ± 36.396 (4.498–138.700)
Total cholesterol, mg/dL	201.558 ± 45.044 (118.296–363.100)	199.346 ± 35.745 (123.000–286.800)	193.047 ± 32.218 (149.067–273.099)
Total protein, g/dL	6.409 ± 0.657 (4.600–8.145)	6.421 ± 0.592 (5.537–8.000)	6.435 ± 0.556 (5.151–7.433)
Serum albumin, g/dL	3.847 ± 0.401 (2.725–4.487)	3.890 ± 0.321 (3.260–4.556)	3.867 ± 0.363 (2.900–4.407)
Uric acid, mg/dL	6.447 ± 1.622 (2.475–10.525)	6.470 ± 2.105 (2.400–14.250)	6.467 ± 2.092 (2.717–10.275)
Total bilirubin, mg/dL	0.732 ± 0.258 (0.397–2.000)	0.755 ± 0.256 (0.326–1.425)	0.769 ± 0.270 (0.300–1.412)
AST, IU/L	20.462 ± 6.842 (12.045–50.562)	22.686 ± 9.223 (14.786–54.750)	23.210 ± 22.190 (10.606–134.500)

ALT, IU/L	23.492 ± 13.269 (3.667–68.333)	31.664 ± 32.339 (10.667–160.000)	29.922 ± 37.048 (7.800–212.500)

This table presents demographic, clinical, laboratory, and tacrolimus treatment characteristics of the final analytic cohort after patient-level splitting into training, validation, and test datasets at a 60:20:20 ratio. Continuous variables are presented as mean ± standard deviation with ranges in parentheses, unless otherwise specified. Categorical variables are presented as number and percentage. AI models used a subcohort of 109 patients with ≥5 troughs (66/22/21). Abbreviations: eGFR, estimated glomerular filtration rate; AST, aspartate aminotransferase; ALT, alanine aminotransferase.

3.2 Population pharmacokinetic (PopPK) modeling

3.2.1 Model structure

Tacrolimus concentration–time data were described using a one-compartment model with first-order absorption and elimination, consistent with conventional sparse trough-based PopPK analyses. The absorption rate constant (ka) was fixed at 4.5 h^{-1} based on prior literature. Clearance and volume of distribution were estimated under four model specifications: a fixed-effect model, a random-effect model, a covariate-extended model, and a covariate + random-effect model.

3.2.2 Parameter estimates

Estimated pharmacokinetic parameters are summarized in Table 3. The fixed-effect model estimated population mean clearance and volume of distribution without inter-individual variability. The random-effect model incorporated inter-individual variability on clearance. In the covariate-extended model, serum creatinine, body weight, and serum albumin were retained as covariates influencing clearance and/or volume of distribution; among these, only the serum creatinine effect on clearance reached statistical significance (Wald $p < 0.01$, Table 3), and this effect was no longer statistically significant in the final covariate + random-effect model once inter-individual variability on clearance was added.

The covariate + random-effect model combined these covariate effects with inter-individual variability on clearance; although its residual error was comparable to that of the random-effect model, it was selected as the final PopPK model for comparison with the PopPK-embedded AI models because it additionally provided clinically interpretable covariate relationships. Overall, simultaneous incorporation of clinically relevant covariates and inter-individual variability reduced unexplained variability and provided a stable mechanistic foundation for subsequent AI embedding.

Table 3. Population pharmacokinetic parameter estimates of tacrolimus

Parameter	Fixed-effect model	Random-effect model	Covariate-extended model	Covariate + Random model
ka (h^{-1})	4.5 (fixed)	4.5 (fixed)	4.5 (fixed)	4.5 (fixed)

CL (log-scale, tcl)	-3.474***	-3.520***	-3.441***	-3.454***
V (log-scale, tv)	0.728	0.246	1.278	0.828
SCRR effect on CL (tcl_scrr)	-	-	-0.138**	-0.103
SCRR effect on V (tv_scrr)	-	-	-0.321	-0.180
ALBU effect on V (tv_albu)	-	-	0.191	0.065
BWTT effect on V (tv_bwtt)	-	-	-0.059	-0.089
IIV on CL, ω^2 (%CV)	-	0.181 (44.6%)	-	0.183 (44.8%)
IIV on V (%CV)	-	-	-	-
Residual error (log.sd)	0.977	0.500	0.965	0.498

This table summarizes estimated pharmacokinetic parameters across four one-compartment PopPK model specifications: fixed-effect, random-effect, covariate-extended, and covariate + random-effect models. The absorption rate constant, ka, was fixed at 4.5 h^{-1} across all models. Clearance and volume of distribution are presented on the log scale. The covariate + random-effect model combined clinically relevant covariates with inter-individual variability and was selected as the final PopPK model for comparison with the AI models; its residual error was comparable to that of the random-effect model, but it additionally provided clinically interpretable covariate relationships.

Significance of fixed-effect and covariate parameters was assessed by Wald test (t = estimate/SE): *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Inter-individual variability (ω^2 on CL) and residual error are variance parameters and were not significance-tested; ka was fixed at 4.5 h^{-1}.
Abbreviations: PopPK, population pharmacokinetics; ka, absorption rate constant; CL, clearance; V, volume of distribution; SCRR, serum creatinine; ALBU, serum albumin; BWTT, body weight; IIV, inter-individual variability.

3.2.3 Goodness-of-fit and visual predictive checks

The observed tacrolimus concentrations were more closely aligned with individual predictions (IPRED) than with population predictions (PRED), and both the population residuals (RES = observed - PRED) and the conditional weighted residuals (CWRES) were generally centered around zero over time, supporting the adequacy of the final PopPK model. Notably, these individual-level diagnostics (IPRED and CWRES) can be obtained only after a patient has contributed measured troughs; for a newly presenting transplant patient, only the population-level PRED and RES are available. This gap motivates augmenting the population prediction with a data-driven, error-based correction, which is the basis of the PopPK-embedded AI framework.

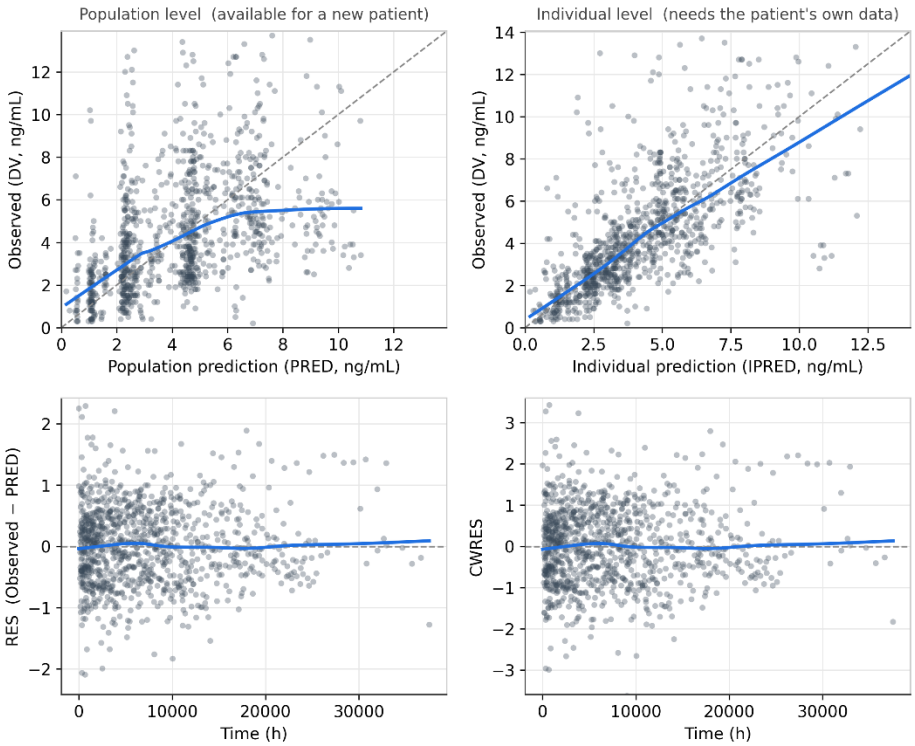

Figure 2. Goodness-of-fit diagnostic plots of the final covariate + random-effect PopPK model (estimation dataset).

The upper-left panel shows observed tacrolimus concentrations (DV) versus population predictions (PRED), and the upper-right panel shows observed concentrations versus individual predictions (IPRED); in both, the dashed diagonal line represents the line of identity. The lower-left panel shows population residuals (RES = observed - population prediction) versus time, and the lower-right panel shows conditional weighted residuals (CWRES) versus time. Observed concentrations aligned more closely with individual than with population predictions, and both RES and CWRES were centered around zero over time, supporting the adequacy of the final PopPK model. Because the individual-level diagnostics (IPRED and CWRES) require each patient's own measured troughs, a newly presenting patient can be served only by the population-level PRED and RES, motivating their combination with a data-driven error correction (the PopPK-embedded AI framework).

Abbreviations: PopPK, population pharmacokinetics; DV, observed tacrolimus concentration; PRED, population prediction; IPRED, individual prediction; RES, residual (observed - population prediction); CWRES, conditional weighted residuals.

3.3 AI modeling

Three PopPK-embedded AI models—LightGBM, LSTM, and RNN—were developed by incorporating the PopPK-predicted concentration, DVI, as an additional input feature. All models were trained using patient-level training data and evaluated on independent validation and test cohorts. These models were developed on the AI subcohort of 109 patients with at least five trough measurements (patient-level training, validation, and test sets of 66, 22, and 21 patients); the independent test set comprised 21 patients and 332 trough prediction endpoints.

In addition to DVI, lagged prediction-error information was included to represent longitudinal deviation from prior model predictions. This structure allowed the AI models to combine mechanistic PopPK-derived information with data-driven temporal error correction.

3.4 Comparison of predictive performance

3.4.1 Performance metrics

Predictive performance was assessed using RMSE and MAE. The quantitative predictive performance of the final PopPK model and PopPK-embedded AI models is summarized in Table 4.

The final PopPK model yielded an RMSE of 3.261 ng/mL, an MAE of 2.349 ng/mL. Embedding PopPK predictions improved predictive performance for all three AI models relative to the PopPK baseline model. Among them, LightGBM demonstrated the best overall performance, with the lowest RMSE (2.606 ng/mL), the lowest MAE (1.812 ng/mL). The RNN embedding also improved prediction accuracy relative to the PopPK baseline, with an RMSE of 2.681 ng/mL, an MAE of 1.823 ng/mL. LSTM showed an RMSE of 2.767 ng/mL, an MAE of 1.931 ng/mL, indicating improvement relative to the PopPK baseline, though a smaller gain than LightGBM and RNN. Uncertainty was quantified by patient-clustered bootstrap resampling (B = 2000). On the test set, the RMSE (95% CI) was 3.261 (2.490–3.910) ng/mL for the PopPK model, 2.606 (2.125–3.101) ng/mL for LightGBM, 2.681 (2.104–3.324) ng/mL for RNN, and 2.767 (2.127–3.438) ng/mL for LSTM. In the paired analysis, all three AI models significantly reduced RMSE relative to the PopPK model, with mean improvements (ΔRMSE = PopPK - AI) of +0.655 (95% CI 0.256–0.972) ng/mL for LightGBM, +0.580 (0.221–0.945) ng/mL for RNN, and +0.495 (0.181–0.800) ng/mL for LSTM; all intervals excluded zero.

The relationship between observed and predicted tacrolimus concentrations for the final PopPK model and PopPK-embedded AI models is shown in Figure 3. Relative to the final PopPK baseline model, the PopPK-embedded AI models showed tighter clustering around the identity line and improved predictive performance. Among the AI models, LightGBM showed the lowest RMSE, indicating the best overall agreement between observed and predicted concentrations.

Table 4. Predictive performance of the final PopPK model and PopPK-embedded AI models

Model	RMSE (95% CI)	MAE (95% CI)
PopPK	3.261 (2.490–3.910)	2.349 (1.756–2.898)
LightGBM	2.606 (2.125–3.101)	1.812 (1.511–2.127)

LSTM	2.767 (2.127–3.438)	1.931 (1.515–2.381)
RNN	2.681 (2.104–3.324)	1.823 (1.461–2.288)

This table compares the predictive performance of the final PopPK model and three PopPK-embedded AI models on the independent test set (21 patients, 332 sequence-endpoint observations). Performance was evaluated using RMSE and MAE, all computed on the ng/mL concentration scale; lower RMSE and MAE indicate better predictive accuracy. Point estimates are shown with 95% confidence intervals obtained by patient-clustered bootstrap resampling (B = 2000 replicates, resampling whole patients). LightGBM showed the lowest RMSE and MAE, indicating the best overall predictive performance. The statistical significance of each AI model's improvement over the PopPK model is established by the paired difference in RMSE (Table 5).
Abbreviations: PopPK, population pharmacokinetics; AI, artificial intelligence; RMSE, root mean square error; MAE, mean absolute error; LightGBM, Light Gradient Boosting Machine; LSTM, long short-term memory; RNN, recurrent neural network.

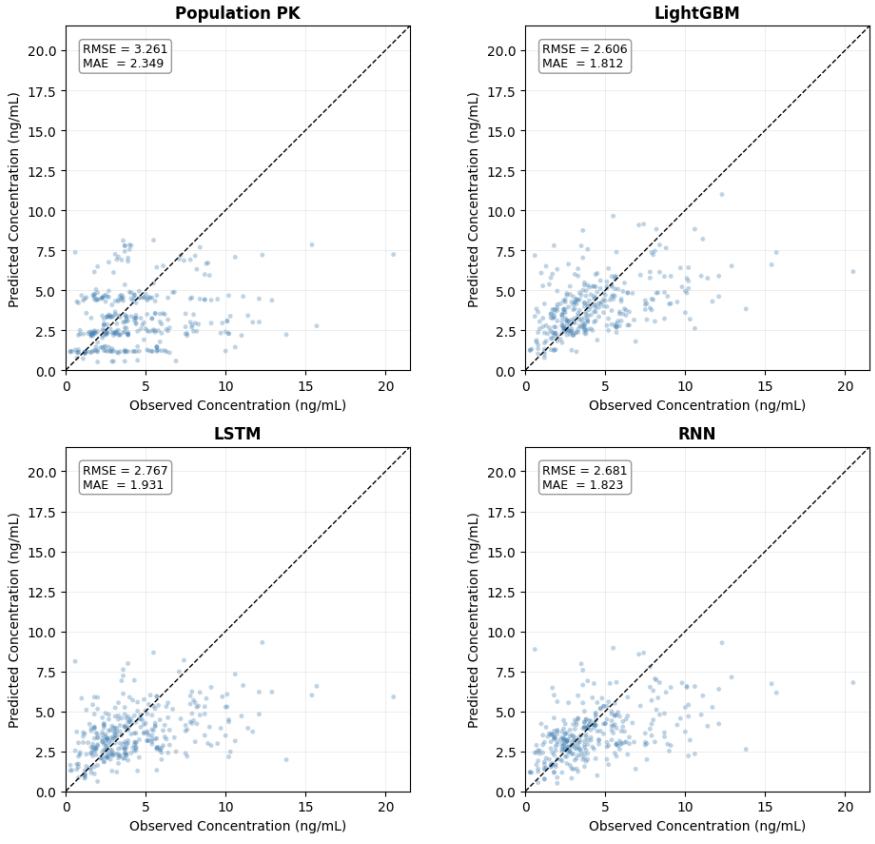

Figure 3. Observed versus predicted tacrolimus concentrations for the final PopPK model and PopPK-embedded AI models.

Observed tacrolimus concentrations were plotted against model-predicted concentrations for the PopPK model, LightGBM, LSTM, and RNN. The dashed diagonal line represents the line of identity, where observed and predicted concentrations are equal. Relative to the PopPK baseline model, the PopPK-embedded AI models showed improved predictive performance, as indicated by lower RMSE and MAE values. LightGBM showed the best overall agreement

with observed concentrations among the AI models. However, underprediction was still observed in some higher concentration ranges.
Abbreviations: PopPK, population pharmacokinetics; AI, artificial intelligence; LightGBM, Light Gradient Boosting Machine; LSTM, long short-term memory; RNN, recurrent neural network; RMSE, root mean square error; MAE, mean absolute error.

3.4.2 Variance reduction

The primary comparison of predictive accuracy was based on the paired, patient-clustered bootstrap analysis of RMSE described above, in which all three AI models significantly outperformed the covariate + random-effect PopPK model (all ΔRMSE 95% CIs excluding zero; Table 5). As an exploratory, supportive analysis, we also compared prediction-error variance using an F-test, which was consistent with the primary result: LightGBM ($F = 1.408$, $p = 9.6 \times 10^{-4}$), RNN ($F = 1.415$, $p = 8.3 \times 10^{-4}$), and LSTM ($F = 1.320$, $p = 5.8 \times 10^{-3}$) each showed a reduction in error variance relative to the PopPK model. Because this F-test assumes independent, normally distributed residuals and does not account for within-patient clustering, it should be interpreted as supportive rather than confirmatory.

Table 5. Improvement in predictive accuracy of the PopPK-embedded AI models relative to the PopPK baseline model: paired difference in RMSE (patient-clustered bootstrap), with exploratory variance-ratio F-tests.

Comparison	ΔRMSE (PopPK - AI)	95% CI	F	p-value
LightGBM vs PopPK	0.655	0.256–0.972	1.408	9.6×10^{-4}
LSTM vs PopPK	0.495	0.181–0.800	1.320	5.8×10^{-3}
RNN vs PopPK	0.580	0.221–0.945	1.415	8.3×10^{-4}

ΔRMSE = RMSE(PopPK) - RMSE(AI); values > 0 indicate a smaller prediction error for the PopPK-embedded AI model than for the PopPK baseline, and all three 95% confidence intervals (patient-clustered bootstrap, B = 2000; 21 patients) excluded zero, indicating a statistically significant improvement. The F-statistics and p-values are exploratory only, as the F-test treats the 332 sequence-endpoint observations as independent and does not account for within-patient clustering. The F-ratio was defined as the variance of the PopPK prediction errors divided by that of the AI prediction errors (F = Var_PopPK / Var_AI); because the two models' errors are paired at the same observations, this ratio does not account for that pairing.
Abbreviations: PopPK, population pharmacokinetics; AI, artificial intelligence; RMSE, root mean square error; CI, confidence interval; LightGBM, Light Gradient Boosting Machine; LSTM, long short-term memory; RNN, recurrent neural network.

3.5 Model interpretability (SHAP analysis)

SHAP analysis demonstrated that the previous prediction error, ERR_LAG1, was the most influential feature across all three AI models (Figures 4–6). **This finding indicates that longitudinal error correction was a dominant driver of the predictive improvement obtained relative to the PopPK baseline.**

In the LightGBM model, ERR_LAG1 exhibited substantially higher mean absolute SHAP values than most other variables, followed by total cholesterol, body weight, and the PopPK-derived prediction DVI. In the LSTM model, ERR_LAG1 was similarly dominant, with total cholesterol, body weight, and recent dosing history among the next most important features. The RNN model also ranked ERR_LAG1 first, followed by ALT, total cholesterol, and DVI.

Importantly, DVI ranked within the top five features for both LightGBM and RNN, confirming that mechanistic PopPK information contributed meaningfully to the AI models beyond error correction alone. Across model architectures, metabolic markers, body-size variables, and selected laboratory covariates emerged as secondary predictors. **These variables should be interpreted cautiously as markers of patient-level physiological or disease-status variability rather than as direct causal determinants of tacrolimus pharmacokinetics.**

To facilitate direct comparison of key predictors among models, the top-ranked SHAP features for each AI model are summarized in Table 6. This table provides a concise overview of shared and model-specific drivers of tacrolimus concentration prediction, complementing the detailed SHAP visualizations shown in Figures 4–6. **Overall, these findings indicate that the PopPK-embedded AI framework leveraged both data-driven temporal error correction and mechanistic PopPK-derived information to refine tacrolimus exposure prediction.**

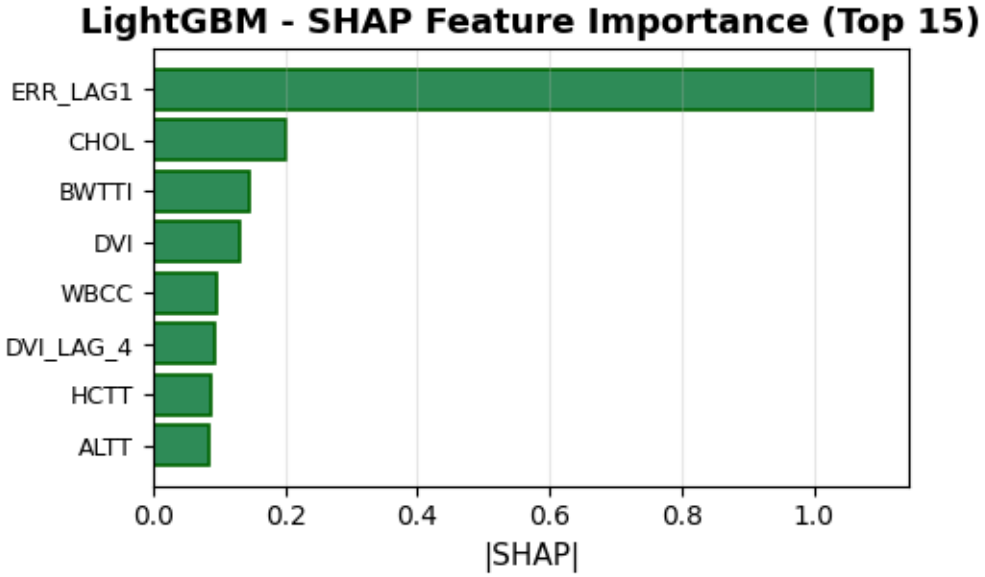

Figure 4. SHAP feature importance for the LightGBM model.

The bar plot shows the top-ranked features contributing to tacrolimus concentration prediction in the LightGBM model based on mean absolute SHAP values. ERR_LAG1 showed the largest contribution, indicating that the previous prediction error was the dominant driver of LightGBM-based prediction. DVI was also ranked among the top features, supporting the contribution of PopPK-derived mechanistic information. Other influential variables included total

cholesterol, body weight, white blood cell count, hematocrit, and ALT, suggesting that patient-level physiological and laboratory features contributed to residual prediction beyond the PopPK model.
Abbreviations: SHAP, Shapley additive explanations; LightGBM, Light Gradient Boosting Machine; ERR_LAG1, previous prediction error; CHOL, total cholesterol; BWTTI, body weight; DVI, drug variability index; WBCC, white blood cell count; HCTT, hematocrit; ALTT, alanine aminotransferase; PopPK, population pharmacokinetics.

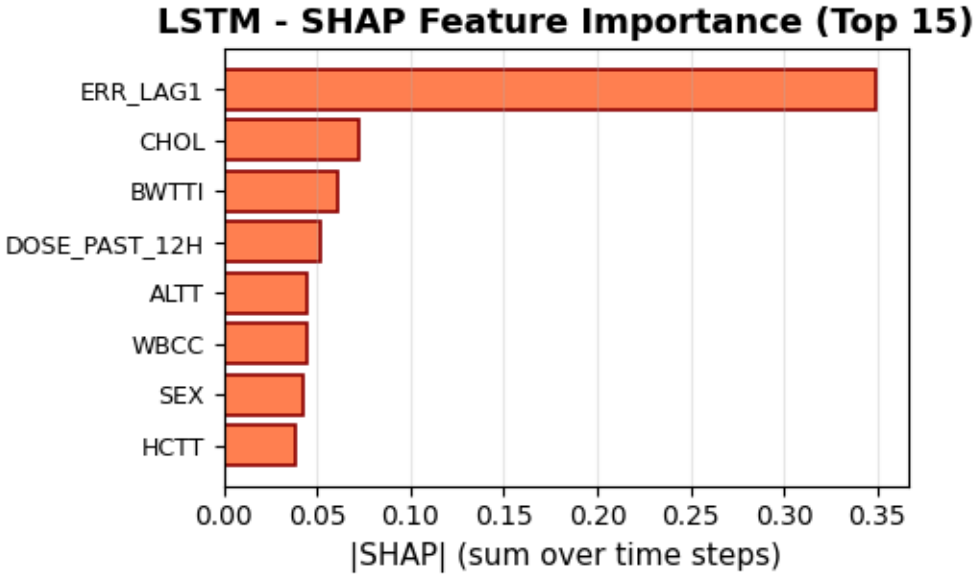

Figure 5. SHAP feature importance for the LSTM model.

The bar plot shows the top-ranked features contributing to tacrolimus concentration prediction in the LSTM model based on SHAP values aggregated over time steps. ERR_LAG1 showed the largest contribution, indicating that previous prediction error was the dominant predictor in the LSTM model. Other influential variables included total cholesterol, body weight, recent dose exposure within the previous 12 hours, ALT, white blood cell count, sex, and hematocrit. These findings suggest that the LSTM model incorporated longitudinal error-correction information, recent dosing history, and time-varying patient characteristics when predicting tacrolimus concentrations.
Abbreviations: SHAP, Shapley additive explanations; LSTM, long short-term memory; ERR_LAG1, previous prediction error; CHOL, total cholesterol; BWTTI, body weight; DOSE_PAST_12H, dose administered within the previous 12 hours; ALTT, alanine aminotransferase; WBCC, white blood cell count; HCTT, hematocrit.

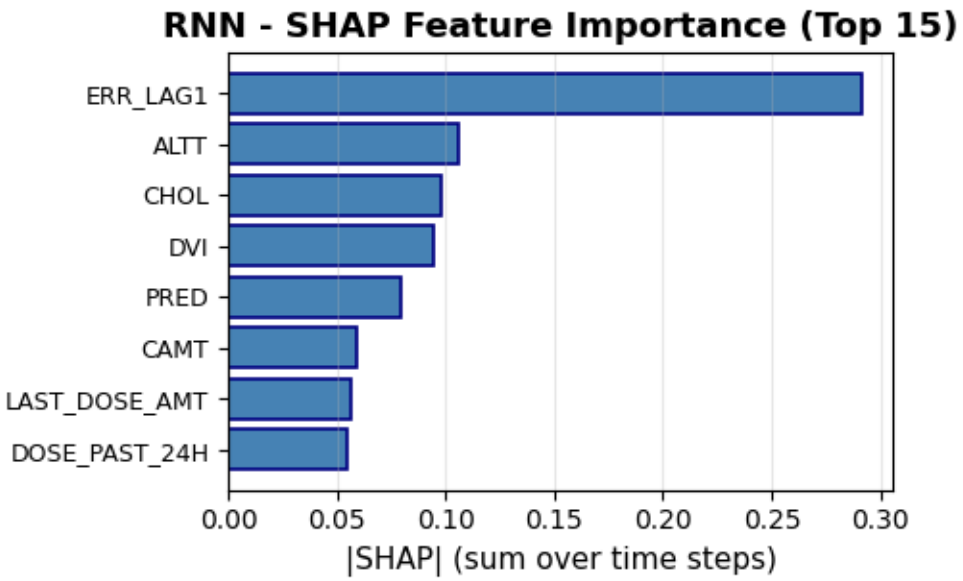

Figure 6. SHAP feature importance for the RNN model.

The bar plot shows the top-ranked features contributing to tacrolimus concentration prediction in the RNN model based on SHAP values aggregated over time steps. ERR_LAG1 showed the largest contribution, indicating that previous prediction error was the dominant predictor in the RNN model.DVI was also among the top-ranked features, supporting the contribution of PopPK-derived mechanistic prediction information. Prednisolone (PRED), a concomitant medication, likewise ranked highly, consistent with its known interaction with tacrolimus metabolism. Other influential variables included ALT, total cholesterol, cumulative amount, last dose amount, and dose administered within the previous 24 hours, suggesting that the RNN model incorporated longitudinal error-correction information, recent dosing history, and time-varying clinical covariates when predicting tacrolimus concentrations.

Abbreviations: SHAP, Shapley additive explanations; RNN, recurrent neural network; ERR_LAG1, previous prediction error; ALTT, alanine aminotransferase; CHOL, total cholesterol; DVI, drug variability index; PRED, prednisolone (concomitant medication); CAMT, cumulative amount; LAST_DOSE_AMT, last dose amount; DOSE_PAST_24H, dose administered within the previous 24 hours; PopPK, population pharmacokinetics.

Table 6. Top-ranked SHAP features across PopPK-embedded AI models.

Rank	LightGBM	LSTM	RNN
1	ERR_LAG1	ERR_LAG1	ERR_LAG1
2	CHOL	CHOL	ALTT
3	BWTTI	BWTTI	CHOL
4	DVI	DOSE_PAST_12H	DVI

5	WBCC	ALTT	PRED

This table summarizes the five most influential features identified by SHAP analysis for each AI model. ERR_LAG1 was the top-ranked feature across all three AI models, suggesting that longitudinal error correction was a major contributor to predictive improvement. DVI remained among the top-ranked features in LightGBM and RNN, supporting the contribution of PopPK-derived mechanistic information.
Abbreviations: SHAP, Shapley additive explanations; AI, artificial intelligence; PopPK, population pharmacokinetics; ERR_LAG1, previous prediction error; DVI, drug variability index; CHOL, total cholesterol; BWTTI, body weight; ALTT, alanine aminotransferase; WBCC, white blood cell count; PRED, prednisolone (concomitant medication).

4. Discussion

In this study, we demonstrated that embedding PopPK-derived predictions into AI models improved tacrolimus concentration prediction in a real-world outpatient cohort. Across all evaluated AI architectures, embedding PopPK-derived predictions yielded lower RMSE and MAE relative to the final PopPK baseline model. LightGBM achieved the best overall performance, with an RMSE reduction from 3.261 to 2.606 ng/mL. These findings indicate that PopPK-embedded AI models improved predictive accuracy. The population PopPK prediction alone showed limited predictive accuracy on the test set, which is expected because it is a population (typical-value) prediction with the random effects set to zero and therefore cannot capture the between-patient variability that dominates sparse, irregularly sampled outpatient trough data. This reflects a limitation of population prediction in real-world monitoring rather than a failure of the structural model itself, and it is precisely this residual, patient-specific variability that the model-informed features—the PopPK prediction (DVI) together with the lagged prediction error (ERR_LAG1)—were designed to recover.

This pairing reflects a deliberate design choice. Our objective was a framework applicable to a new outpatient directly from routinely available covariates, without first fitting that patient's individual pharmacokinetic parameters; because a patient presenting without prior troughs has no measured trough concentrations from which to update the individual random effects, an empirical-Bayes individual prediction (IPRED) cannot be formed at that point. An individualized IPRED would show closer apparent agreement with observations in-sample—as reflected in the tighter alignment of observations with IPRED than with the population prediction on the estimation dataset (Figure 2)—but that alignment is itself conditioned on the individual estimates a new patient lacks and therefore does not reflect what would be available for a patient without prior measured troughs. We accordingly embedded the population prediction (DVI) and recovered patient-specific deviation through the lagged prediction error rather than through per-patient Bayesian fitting. A boundary condition follows directly: at a patient's first observed trough ERR_LAG1 is set to zero and the prediction reduces to the population level, with data-driven individualization emerging from the second trough onward. The resulting task is thus visit-to-visit forecasting of the next trough from strictly past information, on the approximately monthly scale of outpatient monitoring noted in the Methods—a genuine prediction problem portable to previously unseen patients rather than reconstruction of the within-dose concentration–time profile. That ERR_LAG1 remained the most influential feature despite carrying a deviation that is, on average, several weeks old may indicate that a patient's offset from the population prediction is relatively persistent across visits, although we did not formally quantify its temporal autocorrelation; the sensitivity of this correction to the elapsed time since the last observation is considered further among the study limitations.

The performance differences observed among PopPK models highlight well-recognized trade-offs inherent to mechanistic modeling. The fixed-effect model provided a simple population-level description but did not account for inter-individual variability. The random-effect model improved individual-level fit by incorporating inter-individual variability, whereas the covariate-extended model explained part of the variability using clinically relevant covariates such as renal function, body weight, and serum albumin. The covariate + random-effect combined covariate effects with inter-individual variability and was therefore selected as the final model; although its residual error was similar to that of the random-effect model, it additionally captured clinically interpretable covariate relationships (renal function, body weight, serum albumin) needed to describe tacrolimus exposure in this outpatient cohort. Nevertheless, even the most refined PopPK model remained constrained in its ability to capture residual nonlinearities, temporal deviations, and complex interactions characteristic of heterogeneous outpatient EMR data.

Among the AI-based approaches, LightGBM achieved the most favorable balance between predictive accuracy and stability. RNN showed slightly better quantitative performance than LSTM based on RMSE and MAE, although both recurrent embeddings improved prediction accuracy relative to the PopPK baseline model. The performance difference between RNN and LSTM may reflect differences in model architecture, sample size, and the irregularity of outpatient TDM data.

Although LightGBM treated individual time points as independent observations, longitudinal information was implicitly encoded through engineered features such as cumulative dose, time since last dose, interpolated time-varying covariates, and lagged prediction-error information. These representations may provide sufficient temporal context for outpatient therapeutic drug monitoring (TDM) data with irregular sampling, thereby supporting strong generalization despite the absence of an explicit sequence-based structure. Future work may explore hybrid modeling strategies that combine explicit temporal architectures with gradient-boosted decision trees.

Interpretability remains a major barrier to the clinical adoption of AI in pharmacology. In contrast to the initial expectation that DVI would be the dominant predictor, SHAP analysis identified ERR_LAG1 as the most influential feature across all AI models. This finding suggests that prior deviations between observed and predicted concentrations provide substantial information for subsequent tacrolimus concentration prediction. At the same time, DVI remained among the top-ranked predictors in LightGBM and RNN, supporting the continued value of PopPK-derived mechanistic information as a clinically meaningful prior. Therefore, the proposed framework should be interpreted not as a purely PopPK-driven model, but as a model-informed AI approach that integrates mechanistic prediction with longitudinal error correction.

Several limitations should be acknowledged. First, the study cohort was derived from a single institution and focused on outpatients with glomerulonephritis, which may limit generalizability. External validation in broader outpatient populations, including transplant recipients, is warranted. Second, pharmacogenetic determinants, such as CYP3A5 genotype, were not available and may further enhance predictive performance if incorporated. Third, model evaluation focused on predictive accuracy rather than downstream clinical outcomes; future studies should investigate whether improved concentration prediction translates into superior dosing decisions and patient outcomes through prospective or simulation-based analyses. Fourth, the test set was small, comprising only 21 patients and 332 sequence-endpoint observations, and the resulting performance estimates carried wide confidence intervals; the improvement over the PopPK model is most robustly supported by the paired reduction in RMSE, which was statistically significant for all three models. Model development and evaluation also relied on a single patient-level train–validation–test split

without repeated resampling or nested cross-validation. Fifth, because ERR_LAG1 was a highly influential predictor, the model may be particularly useful in patients with prior TDM records but may have reduced applicability for initial tacrolimus concentration prediction in patients without previous observations. Moreover, ERR_LAG1 carries forward the most recent observed prediction error without any staleness cap, so a residual derived from a temporally distant prior trough is weighted identically to a recent one; incorporating a time-decay or recency weighting is a natural direction for refinement. Future studies should separately evaluate model performance for first-time prediction and sequential prediction scenarios, and should assess the sensitivity of the temporal correction to the elapsed time since the last observation.

Although the EMR data used in this study were collected more than a decade ago, this factor is unlikely to materially compromise the validity of the findings. Tacrolimus is a well-established drug with stable pharmacokinetic properties, and the primary determinants of its exposure, including renal function, body size, hematologic parameters, and concomitant clinical status, remain relevant. Importantly, the central contribution of this study lies in the proposed model-informed AI framework rather than in estimating time-sensitive clinical effect sizes. Evaluation on a heterogeneous, imperfect real-world dataset nonetheless illustrates the applicability of the approach to routine outpatient EMR data.

Despite these limitations, our findings provide evidence that PopPK-embedded AI represents a practical and scalable strategy for improving tacrolimus concentration prediction while preserving clinically meaningful interpretability. By combining PopPK-derived predictions with longitudinal error correction and structured EMR features, this framework may be extended to other drugs requiring therapeutic drug monitoring and to diverse real-world outpatient settings.

5. Conclusion

Embedding PopPK-derived predictions into AI models improved tacrolimus concentration prediction in outpatients relative to the PopPK-alone baseline, with LightGBM showing the best overall predictive performance and statistically significant reduction in prediction-error relative to conventional PopPK, established by the paired difference in RMSE (patient-clustered bootstrap). SHAP analyses indicated that longitudinal error correction, represented by ERR_LAG1, was the dominant contributor to AI model performance, while PopPK-derived prediction remained an important mechanistic feature in selected models. These findings support model-informed AI embedding as a promising framework that integrates mechanistic pharmacokinetic knowledge with data-driven learning for precision dosing and real-world outpatient therapeutic drug monitoring.

Data Availability Statement

The datasets presented in this article are not readily available because of institutional and ethical restrictions related to patient privacy. Requests to access the datasets should be directed to the corresponding author, subject to approval by the relevant ethics committee and institutional policies.

Ethics Statement

This study was conducted in accordance with the Declaration of Helsinki and applicable institutional guidelines. The protocol for retrospective analysis of de-identified outpatient EMR data was reviewed and approved by the Institutional Review Board (IRB) of Seoul National University Hospital. The requirement for written informed consent was waived due to the non-interventional design and the use of anonymized retrospective data.

Author Contributions

GL conceptualized the study, developed and implemented the artificial intelligence models, performed data preprocessing, model validation, and statistical analyses, and drafted the original manuscript. MS and DP jointly served as corresponding authors and supervised the overall study. MS led data acquisition and curation, developed the initial population pharmacokinetic (PopPK) models, and provided clinical interpretation. DP provided methodological supervision, reviewed and refined the analytical framework, and advised on model evaluation and interpretation. MS and DP critically revised the manuscript for important intellectual content. All authors reviewed the final version of the manuscript, approved it for submission, and agreed to be accountable for all aspects of the work.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

The authors received no specific funding for this work.

Acknowledgments

The authors sincerely thank Professor Jung Mi Oh of the College of Pharmacy, Seoul National University, for her valuable guidance as an academic advisor. The authors also thank Professor Yon Su Kim of the Department of Nephrology, Seoul National University College of Medicine, for his support and assistance with data access and use.

Supplementary Material

Supplementary material for this article can be found online at the journal website.

Tables

Tables
Table 1. Input features used in PopPK-embedded AI models for tacrolimus concentration prediction.

Table 2. Baseline characteristics of the training, validation, and test datasets.
Table 3. Population pharmacokinetic parameter estimates of tacrolimus.
Table 4. Predictive performance of the final PopPK model and PopPK-embedded AI models.
Table 5. Improvement in predictive accuracy of the PopPK-embedded AI models relative to the PopPK baseline model: paired difference in RMSE (patient-clustered bootstrap), with exploratory variance-ratio F-tests.

Table 6. Top-ranked SHAP features across PopPK-embedded AI models.

Figure legends

Figure 1. Graphical abstract of the model-informed AI framework for tacrolimus trough concentration prediction in outpatients.

Outpatient electronic medical record data, including dosing history, demographic characteristics, and time-varying laboratory covariates, were organized into patient-level longitudinal event tables and discretized on a 12-hour grid. Four one-compartment population pharmacokinetic models were developed to generate PopPK-derived concentration predictions at each time point. The PopPK-predicted concentration, defined as the drug variability index, was embedded as an additional feature into artificial intelligence models, including LightGBM, LSTM, and RNN. Lagged prediction-error information was also incorporated to support longitudinal error correction. Models were trained and evaluated using patient-level training, validation, and test splits. Predictive performance was evaluated using RMSE, MAE, and variance reduction in prediction errors. Model interpretability was assessed using SHAP analysis.

Abbreviations: AI, artificial intelligence; EMR, electronic medical record; PopPK, population pharmacokinetics; DVI, drug variability index; LightGBM, Light Gradient Boosting Machine; LSTM, long short-term memory; RNN, recurrent neural network; RMSE, root mean square error; MAE, mean absolute error; SHAP, Shapley additive explanations.

Figure 2. Goodness-of-fit diagnostic plots of the final PopPK model (estimation dataset)

The upper-left panel shows observed tacrolimus concentrations (DV) versus population predictions (PRED), and the upper-right panel shows observed concentrations versus individual predictions (IPRED); in both, the dashed diagonal line represents the line of identity. The lower-left panel shows population residuals (RES = observed - population prediction) versus time, and the lower-right panel shows conditional weighted residuals (CWRES) versus time. Observed concentrations aligned more closely with individual than with population predictions, and both RES and CWRES were centered around zero over time, supporting the adequacy of the final PopPK model. Because the individual-level diagnostics (IPRED and CWRES) require each patient's own measured troughs, a newly presenting patient can be served only by the population-level PRED and RES, motivating their combination with a data-driven error correction (the PopPK-embedded AI framework).

Abbreviations: PopPK, population pharmacokinetics; DV, observed tacrolimus concentration; PRED, population-predicted; IPRED, individual prediction; RES, residual (observed - population prediction); CWRES, conditional weighted residuals.

Figure 3. Observed versus predicted tacrolimus concentrations for the final PopPK model and PopPK-embedded AI models.

Observed tacrolimus concentrations were plotted against model-predicted concentrations for the final PopPK model, LightGBM, LSTM, and RNN. The dashed diagonal line represents the line of identity, where observed and predicted concentrations are equal. Relative to the PopPK baseline model, the PopPK-embedded AI models showed improved predictive performance, as indicated by lower RMSE and MAE values. Among the AI models, LightGBM showed the best overall agreement with observed concentrations. Underprediction was still observed in some higher concentration ranges.

Abbreviations: PopPK, population pharmacokinetics; AI, artificial intelligence; LightGBM, Light Gradient Boosting Machine; LSTM, long short-term memory; RNN, recurrent neural network; RMSE, root mean square error; MAE, mean absolute error.

Figure 4. SHAP feature importance for the LightGBM model.

The bar plot shows the top-ranked features contributing to tacrolimus concentration prediction in the LightGBM model based on mean absolute SHAP values. ERR_LAG1 showed the largest contribution, indicating that the previous prediction error was the dominant driver of LightGBM-based prediction. DVI was also ranked among the top features, supporting the contribution of PopPK-derived mechanistic information. Other influential variables included total cholesterol, body weight, white blood cell count, hematocrit, and ALT, suggesting that patient-level physiological and laboratory features contributed to residual prediction beyond what the embedded PopPK component explained.

Abbreviations: SHAP, Shapley additive explanations; LightGBM, Light Gradient Boosting Machine; ERR_LAG1, previous prediction error; CHOL, total cholesterol; BWTTI, body weight; DVI, drug variability index; WBCC, white blood cell count; HCTT, hematocrit; ALTT, alanine aminotransferase; PopPK, population pharmacokinetics.

Figure 5. SHAP feature importance for the LSTM model.

The bar plot shows the top-ranked features contributing to tacrolimus concentration prediction in the LSTM model based on SHAP values aggregated over time steps. ERR_LAG1 showed the largest contribution, indicating that previous prediction error was the dominant predictor in the LSTM model. Other influential variables included total cholesterol, body weight, recent dose exposure within the previous 12 hours, ALT, white blood cell count, sex, and hematocrit. These findings suggest that the LSTM model incorporated longitudinal error-correction information, recent dosing history, and time-varying patient characteristics when predicting tacrolimus concentrations.

Abbreviations: SHAP, Shapley additive explanations; LSTM, long short-term memory; ERR_LAG1, previous prediction error; CHOL, total cholesterol; BWTTI, body weight; DOSE_PAST_12H, dose administered within the previous 12 hours; ALTT, alanine aminotransferase; WBCC, white blood cell count; HCTT, hematocrit.

Figure 6. SHAP feature importance for the RNN model.

The bar plot shows the top-ranked features contributing to tacrolimus concentration prediction in the RNN model based on SHAP values aggregated over time steps. ERR_LAG1 showed the largest contribution, indicating that previous prediction error was the dominant predictor in the RNN model. DVI and PRED were also among the top-ranked features, supporting the contribution of PopPK-derived mechanistic prediction information. Other influential variables included ALT, total cholesterol, cumulative amount, last dose amount, and dose administered within the previous 24 hours, suggesting that the RNN model incorporated longitudinal error-correction information, recent dosing history, and time-varying clinical covariates when predicting tacrolimus concentrations.

Abbreviations: SHAP, Shapley additive explanations; RNN, recurrent neural network; ERR_LAG1, previous prediction error; ALTT, alanine aminotransferase; CHOL, total cholesterol; DVI, drug variability index; PRED, prednisolone (concomitant medication); CAMT, cumulative amount; LAST_DOSE_AMT, last dose amount; DOSE_PAST_24H, dose administered within the previous 24 hours; PopPK, population pharmacokinetics.

References

Barrett, J.S. (2024). Artificial Intelligence Opportunities to Guide Precision Dosing Strategies. *The Journal of Pediatric Pharmacology and Therapeutics* 29**,** 434-440.

Bentata, Y. (2020). Tacrolimus: 20 years of use in adult kidney transplantation. What we should know about its nephrotoxicity. *Artificial organs* 44**,** 140-152.

Campagne, O., Mager, D.E., and Tornatore, K.M. (2019). Population Pharmacokinetics of Tacrolimus in Transplant Recipients: What Did We Learn About Sources of Interindividual Variabilities? *J Clin Pharmacol* 59**,** 309-325.

Cremers, S., Guha, N., and Shine, B. (2016). Therapeutic drug monitoring in the era of precision medicine: opportunities! *Br J Clin Pharmacol* 82**,** 900-902.

Francke, M.I., Andrews, L.M., Le, H.L., Van De Wetering, J., Clahsen-Van Groningen, M.C., Van Gelder, T., Van Schaik, R.H., Van Der Holt, B., De Winter, B.C., and Hesselink, D.A. (2021). Avoiding tacrolimus underexposure and overexposure with a dosing algorithm for renal transplant recipients: a single arm prospective intervention trial. *Clinical Pharmacology & Therapeutics* 110**,** 169-178.

Gona, S.R., Rosenberg, J., Fyffe-Freil, R.C., Kozakiewicz, J.M., and Money, M.E. (2023). Failure of current digoxin monitoring for toxicity: new monitoring recommendations to maintain therapeutic levels for efficacy. *Frontiers in Cardiovascular Medicine* 10**,** 1179892.

Herblum, J., Dacouris, N., Huang, M., Zaltzman, J., Prasad, G.R., Nash, M., and Chen, L. (2021). Retrospective analysis of tacrolimus intrapatient variability as a measure of medication adherence. *Canadian Journal of Kidney Health and Disease* 8**,** 20543581211021742.

Jian, F., Chen, X., Wang, M., Li, X., Jian, H., Ji, R., Liang, L., Yu, Z., and Chen, Y. (2025). Machine learning-based prediction model for teicoplanin plasma concentrations in adults with liver disease using real-world data. *Frontiers in Pharmacology* 16**,** 1703976.

Li, Z.R., Li, R.D., Niu, W.J., Zheng, X.Y., Wang, Z.X., Zhong, M.K., and Qiu, X.Y. (2023). Population pharmacokinetic modeling combined with machine learning approach improved tacrolimus trough concentration prediction in Chinese adult liver transplant recipients. *The Journal of Clinical Pharmacology* 63**,** 314-325.

Maier, C., De Wiljes, J., Hartung, N., Kloft, C., and Huisinga, W. (2022). A continued learning approach for model-informed precision dosing: updating models in clinical practice. *CPT: Pharmacometrics & Systems Pharmacology* 11**,** 185-198.

Minichmayr, I., Dreesen, E., Centanni, M., Wang, Z., Hoffert, Y., Friberg, L.E., and Wicha, S. (2024). Model-informed precision dosing: State of the art and future perspectives. *Advanced drug delivery reviews* 215**,** 115421.

Pérez-Blanco, J.S., and Lanao, J.M. (2022). Model-informed precision dosing (MIPD). Pharmaceutics 14, 2731.

Poweleit, E.A., Vinks, A.A., and Mizuno, T. (2023). Artificial intelligence and machine learning approaches to facilitate therapeutic drug management and model-informed precision dosing. *Therapeutic drug monitoring* 45**,** 143-150.

Stankevičiūtė, K., Woillard, J.-B., Peck, R.W., Marquet, P., and Van Der Schaar, M. (2023). Bridging the worlds of pharmacometrics and machine learning. *Clinical Pharmacokinetics* 62**,** 1551-1565.

Wang, Y.-P., Lu, X.-L., Shao, K., Shi, H.-Q., Zhou, P.-J., and Chen, B. (2024). Improving prediction of tacrolimus concentration using a combination of population pharmacokinetic modeling and machine learning in chinese renal transplant recipients. *Frontiers in Pharmacology* 15**,** 1389271.