Machine learning-based model selection and averaging outperform single-model approaches for a priori vancomycin precision dosing

Os, W. van; O'Jeanson, A.; Troisi, C.; Liu, C.; Brooks, J.T.; Hughes, J.H.; ... ; Keizer, R.J.

CPT: Pharmacometrics & Systems Pharmacology

ARTICLE OPEN ACCESS

Machine Learning-Based Model Selection and Averaging Outperform Single-Model Approaches for a Priori Vancomycin Precision Dosing

Wisse van Os[1,2] | Amaury O'Jeanson[3] | Carla Troisi[4] | Chun Liu[4] | Jordan T. Brooks[5] | Jasmine H. Hughes[5] | Dominic M. H. Tong[5] | Ron J. Keizer[5]

[1]Division of Systems Pharmacology & Pharmacy, Leiden Academic Centre for Drug Research, Leiden University, Leiden, the Netherlands | [2]Department of Clinical Pharmacology, Medical University of Vienna, Vienna, Austria | [3]Department of Pharmacy, Uppsala University, Uppsala, Sweden | [4]Department of Medical and Surgical Sciences, Alma Mater Studiorum-University of Bologna, Bologna, Italy | [5]InsightRX, San Francisco, California, USA

Correspondence: Wisse van Os (w.van.os@lacdr.leidenuniv.nl)

Received: 19 April 2025 | **Revised:** 29 June 2025 | **Accepted:** 11 July 2025

Funding: This work was supported by the European Union's H2020 Marie Skłodowska-Curie Actions Research and Innovation Program (grant number 861323).

Keywords: machine learning | model-informed precision dosing (MIPD) | multi-label classification | population pharmacokinetics | therapeutic drug monitoring (TDM) | vancomycin | XGBoost

ABSTRACT

Selecting an appropriate population pharmacokinetic (PK) model for individual patients in model-informed precision dosing (MIPD) can be challenging, particularly in the absence of therapeutic drug monitoring (TDM) samples. We developed a machine learning (ML) model to guide individualized PK model selection for a priori MIPD of vancomycin based on routinely recorded patient characteristics. This retrospective analysis included 343,636 vancomycin TDM records, each from a distinct adult patient across 156 healthcare centers, along with a priori predictions from six PK models. A multi-label classification approach was applied, labeling PK model predictions based on whether they fell within 80%–125% of observed TDM values. Various modeling strategies were evaluated using XGBoost as the base algorithm, with binary relevance selected for the final model. At the prediction stage, PK models were ranked and averaged for each patient based on ML-predicted probabilities that predictions would fall within 80%–125% of the observed concentration. Selecting the highest ranked PK model for each patient and ML-based model averaging outperformed all single PK models, body mass index-based selection, and naive averaging. On a population level, these ML approaches resulted in more accurate predictions, a higher proportion of predictions within 80%–125% of observed vancomycin concentrations, and no systematic bias. Predictive performance declined with lower ML-assigned rankings, and selecting the lowest-ranked PK model for each patient resulted in worse performance than the worst-performing single PK model. By guiding the selection of appropriate models and avoiding less suitable ones, ML approaches for a priori MIPD may improve early dosing decisions.

1 | Introduction

Model-informed precision dosing (MIPD) aims to personalize drug dosing to achieve plasma drug exposure associated with efficacy and reduced risk of toxicity. It commonly relies on population pharmacokinetic (PK) models to predict individual drug exposures based on patient characteristics and, when available, plasma drug concentrations obtained through therapeutic drug monitoring (TDM) using Bayesian forecasting [1]. For many drugs, multiple PK models exist, and

Summary

- What is the current knowledge on the topic?
  - Pharmacokinetic (PK) model selection is a key challenge in model-informed precision dosing (MIPD), especially in the absence of therapeutic drug monitoring (TDM) data. Current practice often relies on patient similarity to model development populations or external validation studies.
- What question did this study address?
  - Can machine learning (ML) guide individualized PK model selection and help avoid inappropriate models for a priori vancomycin MIPD, using only routinely recorded patient characteristics?
- What does this study add to our knowledge?
  - ML-based ranking and averaging of PK models improved a priori prediction accuracy compared to conventional model selection strategies for vancomycin MIPD and helped avoid inaccurate models. This study highlights the potential of ML models trained with real-world data to guide model selection in MIPD.
- How might this change drug discovery, development, and/or therapeutics?
  - Integrating ML-guided model selection into MIPD tools could improve early dosing decisions before TDM data become available.

MIPD software tools typically incorporate several of these. Predictions from different models can vary substantially for a given patient, so the choice of model at the point-of-care can impact dosing decisions [2–4]. Selecting an appropriate model and avoiding inappropriate ones is therefore essential to fully realize the potential of MIPD [5].

Algorithms for model selection and averaging have been shown to improve predictive performance in Bayesian forecasting but depend on the availability of TDM samples [6, 7]. In the absence of TDM data (i.e., for a priori MIPD), practitioners typically select a model based on patient similarity to the model development population [8], results from external validation studies [4, 9–12], or prior experience. However, the best model for an individual patient is not always the one developed using the most similar population or the one that performs best across a validation cohort [5, 13]. Patients may also belong to populations underrepresented in PK studies [14], or to multiple subpopulations simultaneously [15]. Moreover, the suitability of PK models may be influenced by multiple patient characteristics, complicating generalizations based on single variables. For example, optimal vancomycin model selection was recently shown to vary across patient groups defined by combinations of age and BMI [16]. Since conducting prospective validation studies for highly specific populations is impractical, selecting the right model for the right patient remains complex. Machine learning (ML) appears well-suited to address this challenge. The ability of ML to identify complex patterns in data has driven growing interest in its applications in pharmacometrics and MIPD [1, 17–22].

This study aimed to develop an ML model capable of ranking and averaging PK models in the absence of TDM samples, based only on routinely recorded patient characteristics. Vancomycin was selected as a case study due to its narrow therapeutic window and the well-established recommendations for MIPD in its dosing guidelines [23]. Additionally, we had access to a large dataset of vancomycin TDM records to train and evaluate the proposed ML approach.

2 | Methods

2.1 | TDM Records and Predictions

This analysis was based on retrospective data from adult patients receiving vancomycin at healthcare centers across the United States between January 1, 2020, and January 30, 2024. Data were entered by users or extracted from electronic health records into the MIPD clinical decision support software InsightRX Nova and included patient characteristics, dosing information, and TDM values. As the data were collected as part of routine clinical care at contracted healthcare entities and were fully deidentified prior to analysis, the study is exempt from US federal Human Research Subjects regulations under 45 CFR 46.102(f). Records were excluded if key data were missing, or the likelihood of erroneous entries was high. The following exclusion criteria were applied: age <18 or >110 years, weight <30 or >400 kg, height <120 or >220 cm, serum creatinine (SCr) ≤ 0 or >20 mg/dL, vancomycin doses <0.1 g or >6 g, TDM values >50 mg/L, or TDM samples collected more than 72 h after the last recorded dose. Ambiguous records, such as entries with the same patient identifier and timestamp but conflicting TDM values, were also excluded. Of the remaining records, only the first recorded TDM sample for each patient was included to balance the influence of individual patients' data.

For each included TDM record, a priori predictions were generated using six parametric population PK models available on InsightRX Nova. These predictions were based on the patient's dosing history, covariates, and population point estimates of the PK parameters, and did not incorporate interindividual or residual variability. The models included the Buelga [24], Carreno [25], Colin [26], Goti/Tong [27, 28], Hughes [29], and Thomson [30] models. Selection was based on performance in validation studies and implementation in other MIPD software platforms. The Goti model, which performed best in an external validation [4], was modified by removing a rule of rounding SCr to 1 mg/dL for patients aged ≥ 65 years with SCr values <1 mg/dL, which was demonstrated to improve performance [28]. The Carreno model, developed for obese patients, outperformed other obesity-specific models in a limited sampling setting and has performed well in internal InsightRX validations, although the Hughes model has recently demonstrated superior performance [25, 29]. The Colin model, developed by pooling datasets from various studies, performs well across general, obese, and younger adult populations [13, 16, 29]. The Buelga model, developed in adults with hematological malignancies and used in the DoseMe platform [31], is the only one-compartment model included. The Thomson model, used here with a capped creatinine clearance of 150 mL/min, has shown strong internal performance [16]. A more detailed overview of the models is provided in Table 1.

TABLE 1 | Overview of the included PK models.

PK model	*n* subjects/*n* samples	Population	Structure	Covariates
Buelga [24]	215/1004	Adults with hematological malignancies	1 CMT, linear	CrCl, weight
Carreno [25]	12/71	Obese adult patients	2 CMT, linear	CrCl
Colin [26]	2554/8300	Pooled model (neonates, children, adults; healthy, ICU, obese)	2 CMT, linear	SCr, weight, PMA
Goti/Tong [27, 28]	1812/2765	Adult patients, including ICU and hemodialysis patients	2 CMT, linear	CrCl[a], weight, hemodialysis
Hughes [29]	83/272	Obese adult patients	2 CMT, linear	CrCl, fat-free mass
Thomson [30] (modified)	398/1557	Adult patients	2 CMT, linear	CrCl (capped[b]), weight

Abbreviations: CMT, compartment; CrCl, creatinine clearance; ICU, intensive care unit; PMA, postmenstrual age; SCr, serum creatinine.
[a]The Goti model was modified by removing a rule of rounding SCr to 1 mg/dL for patients with age ≥ 65 years and SCr < 1 mg/dL, described in Tong et al. [28].
[b]The Thomson model was modified by capping CrCl at 150 mL/min.

2.2 | Features

Eight patient characteristics were initially selected as features (i.e., predictors) for ML model development. These included age, weight, height, sex, and SCr, which were available from the TDM records, as well as three derived features: body mass index (BMI), body surface area, and estimated glomerular filtration rate (eGFR) calculated using the 2021 CKD-EPI creatinine formula [32]. Note that different equations for eGFR may have been used for the PK predictions, depending on the covariate specifications of the model.

At various stages of model development, we evaluated whether including additional features improved performance. These features included user-supplied tags (e.g., whether the patient was in the intensive care unit), binarized versions of patient characteristics (e.g., BMI ≥ 30 or < 18.5), and indicators specifying whether a patient characteristic was below, above, and outside the ranges reported in the population PK model publications.

2.3 | Machine Learning Model Development

The ML task was framed as a multi-label classification problem, with the labels representing the included PK models. For each TDM record, the predictions from each of the six PK models were classified based on whether they fell within 80%–125% of the observed TDM value (Figure 1). This range was selected because it treats underpredictions and overpredictions symmetrically on the log scale and approximately reflects the width of commonly targeted vancomycin exposure or concentration ranges [11].

We used the utiml R package for ML model development, which is specifically designed for multi-label classification tasks [33]. Unlike standard classification problems, multi-label problems allow each instance to be associated with multiple labels. The utiml package offers several transformation strategies to make such problems compatible with standard classification algorithms, for example by training independent binary classifiers for each label, comparing labels pairwise, predicting labels sequentially, or treating each label set as a unique class. These transformations also influence the model output, producing either label-wise probabilities or direct label rankings. Eight transformation strategies (listed in Table S1) were selected for comparison based on their suitability for our dataset, which included a large number of observations and a relatively small, nonhierarchical label space including six labels. More details on transformation approaches are available through Rivolli and De Carvalho [33]. XGBoost was chosen as the base algorithm [34], as it was considered the most appropriate among those available in the utiml package due to its strong performance on tabular data, ability to handle correlated features, computational efficiency on large datasets, and previous performance in similar contexts [21, 35–39].

The dataset was divided into training, validation, and test sets using a 70%–15%–15% split. The training set was used for hyperparameter tuning and model training. Hyperparameter tuning was performed using 10-fold cross-validation with a random search strategy across 100 iterations [40]. The optimized hyperparameters were the learning rate (eta, ranging from 0.01 to 0.3), the number of boosting iterations (nrounds, from 50 to 200 in steps of 50), maximum tree depth (max_depth, integer values from 3 to 9), the proportion of subsampled training rows (subsample, from 0.6 to 1), and the proportion of columns subsampled per tree (colsample_bytree, from 0.6 to 1). Mean average precision was chosen as the evaluation metric during hyperparameter optimization because it captures ranking quality and balances precision and recall.

The validation set was used to compare ML models developed using different multi-label transformation strategies. For these comparisons, we focused on metrics that reflect how ML-based ranking and averaging would perform if applied in MIPD, thereby ensuring that model evaluation was aligned with the study's aim of improving predictive performance in this context. The primary metric was the percentage of predictions within 80%–125% of the observed TDM value, capturing both accuracy and clinical relevance. Additional evaluation metrics are described in the *Performance evaluation* section. The test set was reserved for evaluating the final model.

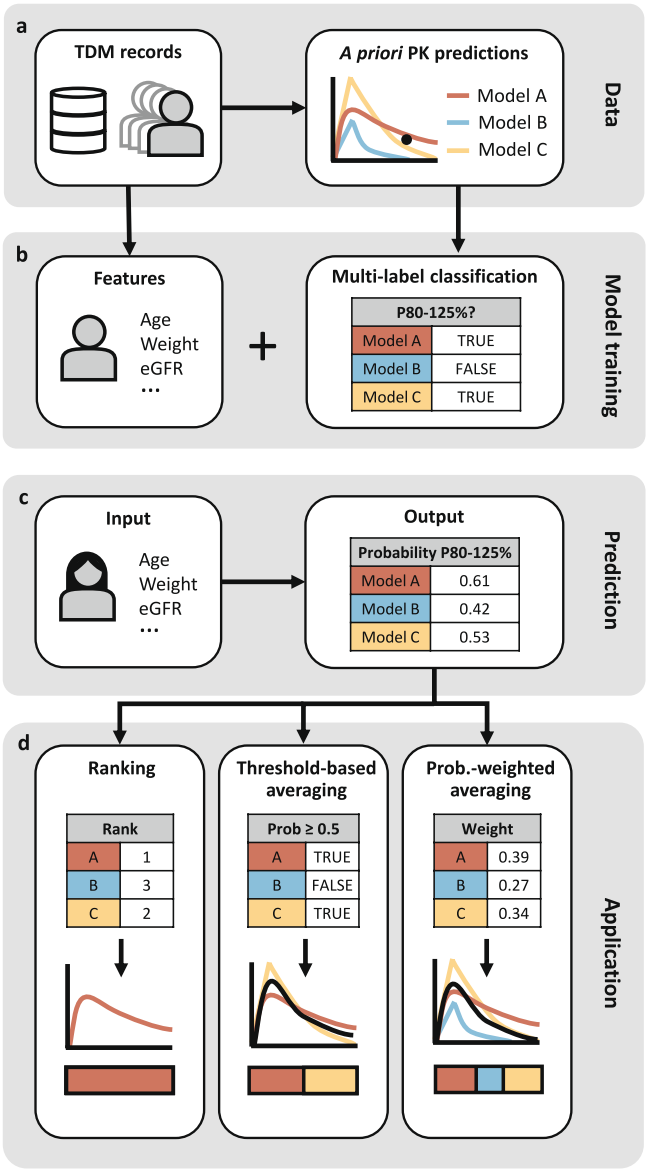

FIGURE 1 | Schematic overview of the development and application of the multi-label classification ML model, illustrated with three hypothetical population PK models A, B and C. (a) A priori predictions are generated using patient covariates and dosing information extracted from TDM records and population point estimates of PK model parameters. (b) For ML model training, patient characteristics serve as features, and labels are assigned based on whether PK model predictions fall within 80%–125% of the TDM observation. (c) At the prediction stage, the ML model outputs probabilities for each PK model predicting within the 80%–125% range, using patient characteristics alone. (d) Predicted probabilities are used to rank or average PK models, with model averaging illustrated by black lines.

2.4 | Processing ML Predictions

Depending on the transformation strategy, the model output was either a direct ranking of PK models or predicted probabilities indicating the likelihood that each PK model would produce a prediction within 80%–125% of the observed concentration, which could also be used for ranking. In addition, two model-averaging approaches were applied for the transformation strategies that output probabilities. The first method involved averaging the PK model predictions using the normalized ML-predicted probabilities as weights:

$$\mathrm{PRED}_{i,\mathrm{avg\_prob}} = \sum_{m=1}^{M}\left(\frac{P_{i,m}\cdot \mathrm{PRED}_{i,m}}{\sum_{m=1}^{M} P_{i,m}}\right) \quad (1)$$

where M represents the number of PK models, $P_{i,m}$ represents the predicted probability for patient i and model m, and $\mathrm{PRED}_{i,m}$ is the corresponding prediction.

The second method involved averaging predictions from all PK models with predicted probabilities ≥ 0.5 for that patient (i.e., PK models expected to produce predictions within 80%–125% of the true concentration based on a bipartition rule):

$$\mathrm{PRED}_{i,\mathrm{avg\_threshold}} = \frac{\sum_{m=1}^{M} \mathbb{I}\left(P_{i,m}\geq 0.5\right)\cdot \mathrm{PRED}_{i,m}}{\sum_{m=1}^{M} \mathbb{I}\left(P_{i,m}\geq 0.5\right)} \quad (2)$$

where $I\left(P_{i,m}\geq 0.5\right)$ is an indicator function that equals 1 if $P_{i,m}\geq 0.5$ and 0 otherwise. If no models met this threshold, the prediction from the PK model with the highest predicted probability was selected instead.

2.5 | Performance Evaluation

The performance of ML-based model ranking and averaging was compared with that of individual PK models, as well as naive (i.e., unweighted) model averaging:

$$\mathrm{PRED}_{i,\mathrm{avg\_naive}} = \sum_{m=1}^{M}\frac{1}{M}\cdot \mathrm{PRED}_{i,m} \quad (3)$$

Additionally, rules-based selection approaches using a BMI cutoff of 40 $\mathrm{kg/m^2}$ were assessed. The Carreno and Hughes models were developed using data from patients with obesity [25, 29], and switching between PK models based on BMI is common practice in MIPD. In most hospitals using InsightRX Nova, model switching based on this cutoff is implemented by default, although the models involved vary by site. The true best PK model for each TDM observation was retrospectively identified as the model with the lowest absolute prediction error.

All ML and PK approaches were evaluated according to the following metrics:

i. the percentage of predictions within 80%–125% of the TDM observations ($P_{80\%-125\%}$):

$$P_{80\%-125\%,k} = \frac{\sum_{i=1}^{N} \mathbb{I}\left(0.8\cdot \mathrm{TDM}_i \leq \mathrm{PRED}_{i,k} \leq 1.25\cdot \mathrm{TDM}_i\right)}{N}\cdot 100\,\% \quad (4)$$

ii. the root mean squared error (RMSE):

$$\mathrm{RMSE}_k = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{TDM}_i - \mathrm{PRED}_{i,k}\right)^2} \quad (5)$$

iii. the mean absolute error (MAE):

$$\mathrm{MAE}_k = \frac{1}{N}\sum_{i=1}^{N} \mid \mathrm{TDM}_i - \mathrm{PRED}_{i,k} \mid \quad (6)$$

iv. the distribution of prediction errors (PE), with a focus on the median PE to detect bias:

$$\mathrm{PE}_{i,k} = \frac{\mathrm{PRED}_{i,k} - \mathrm{TDM}_i}{\mathrm{TDM}_i} \cdot 100\,\% \quad (7)$$

In these formulas, N is the total number of TDM observations in the validation set (for ML model development) or test set (for final model evaluation), i represents the observation number, and k represents the approach to PK model selection (e.g., single PK models, naive averaging, or the ML-predicted best-ranked PK model).

2.6 | Software

Population PK model predictions were made using the R (v4.1) package PKPDsim (v1.4.0) [41]. All other analyses were performed in R (v4.4.2) [42], using the utiml (v0.1.7) [33], caret (v7.0-1) [43], and xgboost (v1.7.8.1) [34] packages for ML model development and the tidyverse collection (v2.0.0) [44] for data wrangling and visualization. The script used to develop the ML model is provided as Supporting Information.

TABLE 2 | Overview of patient characteristics ($n = 343{,}636$) and TDM values.

Characteristic	Values
Age (years)	64.5 (52.0–75.0) [22.3–94.0]
Weight (kg)	83.0 (68.0–102) [41.3–184]
Height (cm)	170 (163–180) [147–194]
Sex (% male)	57.1
SCr (mg/dL)	0.960 (0.740–1.30) [0.370–5.75]
eGFR (mL/min/1.73 m^2)[a]	79.7 (53.2–99.8) [9.12–136]
BMI (kg/m^2)	28.3 (23.7–34.4) [15.7–61.7]
BSA (m^2)	1.96 (1.76–2.18) [1.36–2.86]
TDM observation (mg/L)	13.4 (9.70–18.2) [3.90–38.0]

Note: Data are presented as median (IQR) and [1st–99th percentile] or percentages.
Abbreviations: BMI, body mass index; BSA, body surface area; eGFR, estimated glomerular filtration rate; SCr, serum creatinine; TDM, therapeutic drug monitoring.
[a]Calculated using the 2021 CKD-EPI creatinine formula [32].

3 | Results

3.1 | Data

A total of 343,636 TDM records from distinct patients across 156 centers in the United States were included from an initial dataset of 974,411 records from 397,223 patients. Details of the included records are presented in Table 2. Label frequencies (i.e., the number of times each PK model's a priori prediction fell within 80%–125% of the observed TDM value) were well balanced (Table 3). Among the six population PK models, the Hughes model had the highest percentage of positive labels (49.4%) and the Buelga model the lowest (35.8%). All 64 possible label sets (i.e., unique combinations of PK models with predictions within 80–125%) were observed in the training data. The mean number of positive labels per TDM record (i.e., the cardinality) was 2.65.

3.2 | Machine Learning Model

The Binary Relevance (BR) approach, in which independent binary classifiers are trained for each label, was selected from the eight evaluated transformation strategies, as it performed similarly to or better than more complex and computationally expensive alternatives (Table S1). Two strategies that produced direct rankings slightly outperformed BR on the ranking task (Calibrated Label Ranking and Ranking by Pairwise Comparison), but given the minimal performance difference, BR was selected as it outputs probabilities that offer additional insight into relative predicted performance and allow for model averaging. Variation in mean average precision across the tested hyperparameter space was limited (Figure S1), with model performance most sensitive to the learning rate (eta) and maximum tree depth (max_depth). The mean average precision of the final model was 0.738 on the training set, 0.727 on the validation set, and 0.725 on the test set, indicating minimal overfitting. Across the binary classifiers trained for each label within the multi-label classification model, age, BMI, eGFR, and SCr were the

TABLE 3 | Frequencies of a priori predictions within 80%–125% of the observed TDM value for each PK model and corresponding label imbalance ratios.

PK model	Positive label frequency, *n* (%)	Label imbalance ratio[a]
Buelga	122,988 (35.8%)	1.38
Carreno	143,639 (41.8%)	1.18
Colin	161,353 (47.0%)	1.05
Goti/Tong	155,683 (45.3%)	1.09
Hughes	169,921 (49.4%)	1.00 (majority)
Thomson (modified)	156,747 (45.6%)	1.08

[a]Calculated as the ratio between the frequencies of the majority and considered labels.

most influential features (Figure 2). Adding additional features did not improve performance on the validation set.

For all PK models, ML-predicted probabilities that a priori predictions would fall within 80%–125% of the true drug concentration generally increased with age, while probabilities declined at eGFR values above 100 mL/min/1.73 m^2 and with low SCr levels (Figure 3). Probabilities for the Buelga, Colin, Goti/Tong, and modified Thomson models decreased for patients with obesity (BMI $\geq$30 kg/m^2), whereas probabilities for the Hughes and Carreno models remained more stable at higher BMIs. Similar trends were observed for weight. The Carreno model was the only one for which probabilities were typically higher for females than males across the test set, although population-level differences between the sexes were generally small for all PK models.

When transforming the predicted probabilities into rankings, the Colin and Hughes models were most frequently ranked as the best-performing models (30.7% and 30.1% of test set records, respectively), followed by the modified Thomson (17.8%), Carreno (12.0%), Goti/Tong (9.3%) and Buelga (0.2%) models (Figure 4). The Buelga (52.6%) and Carreno (36.8%) models were most often ranked last. In contrast, the modified Thomson model was ranked last in only 0.1% of test set records.

3.3 | Performance Evaluation

Across the test set population, the Hughes model was the best-performing single model, while the Buelga model performed worst (Figure 5). Switching from a general population model to the Carreno model for morbidly obese patients improved population-level performance compared to using either model individually. However, the Hughes model still outperformed BMI-based selection involving the Hughes model on most metrics.

Individualized ML-based model selection and averaging outperformed all single PK models, BMI-based model selection, and naive averaging in terms of RMSE, MAE, and $P_{80\%-125\%}$, and demonstrated no systematic bias. Furthermore, the population-level performance of the individually predicted rankings followed the expected trend: selecting the best-ranked model for each patient outperformed selecting the second-ranked model, and so on. Selecting the lowest-ranked model for each patient performed the worst out of all evaluated approaches.

Given that the Buelga model is a one-compartment model, while vancomycin is generally described by two-compartment kinetics, and that clearance in its development dataset was shown to be significantly higher than in other populations [26], we also evaluated naive averaging with the Buelga model excluded. This improved performance, but ML-based model selection and averaging still outperformed this modified naive approach (Figure 5). Since the Buelga model was generally predicted to rank low (Figure 4), completely excluding it from the analysis had minimal impact on the performance of ML-based strategies (Figure S2).

The performance of the individual binary classification models trained for each label within the multi-label BR framework was modest, with a macro-averaged AUC-ROC of 0.611 across all labels.

4 | Discussion

Selecting the most appropriate PK model at the point-of-care in MIPD is challenging, particularly in the absence of TDM data. In this study, we developed a multi-label classification ML approach for a priori PK model ranking and averaging in MIPD, based only on routinely recorded patient characteristics. We demonstrated the approach using a large, real-world dataset of adult patients receiving vancomycin.

ML-based PK model selection and averaging outperformed single PK models, BMI-based selection strategies, and naive averaging. On a population level, the ML approaches resulted in more accurate PK predictions without systematic bias, although improvements over some approaches were modest. Selecting the second-highest ranked model for each patient also outperformed most other approaches. Importantly, population-level performance declined at lower ranks, and selecting the

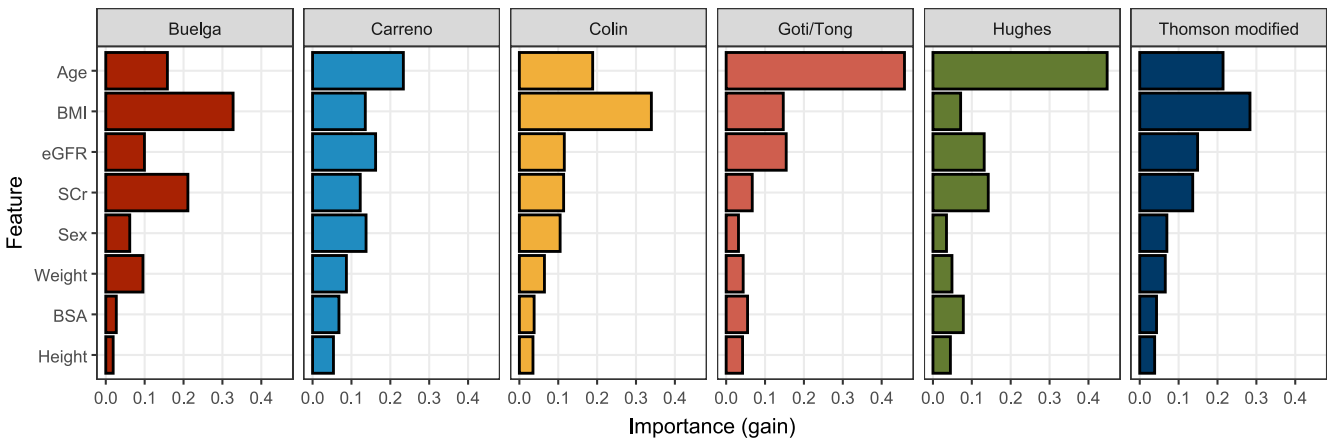

FIGURE 2 | Feature importance for the binary classifiers trained for each vancomycin PK model in the multi-label classification model. BMI, body mass index; BSA, body surface area; eGFR, estimated glomerular filtration rate; SCr, serum creatinine.

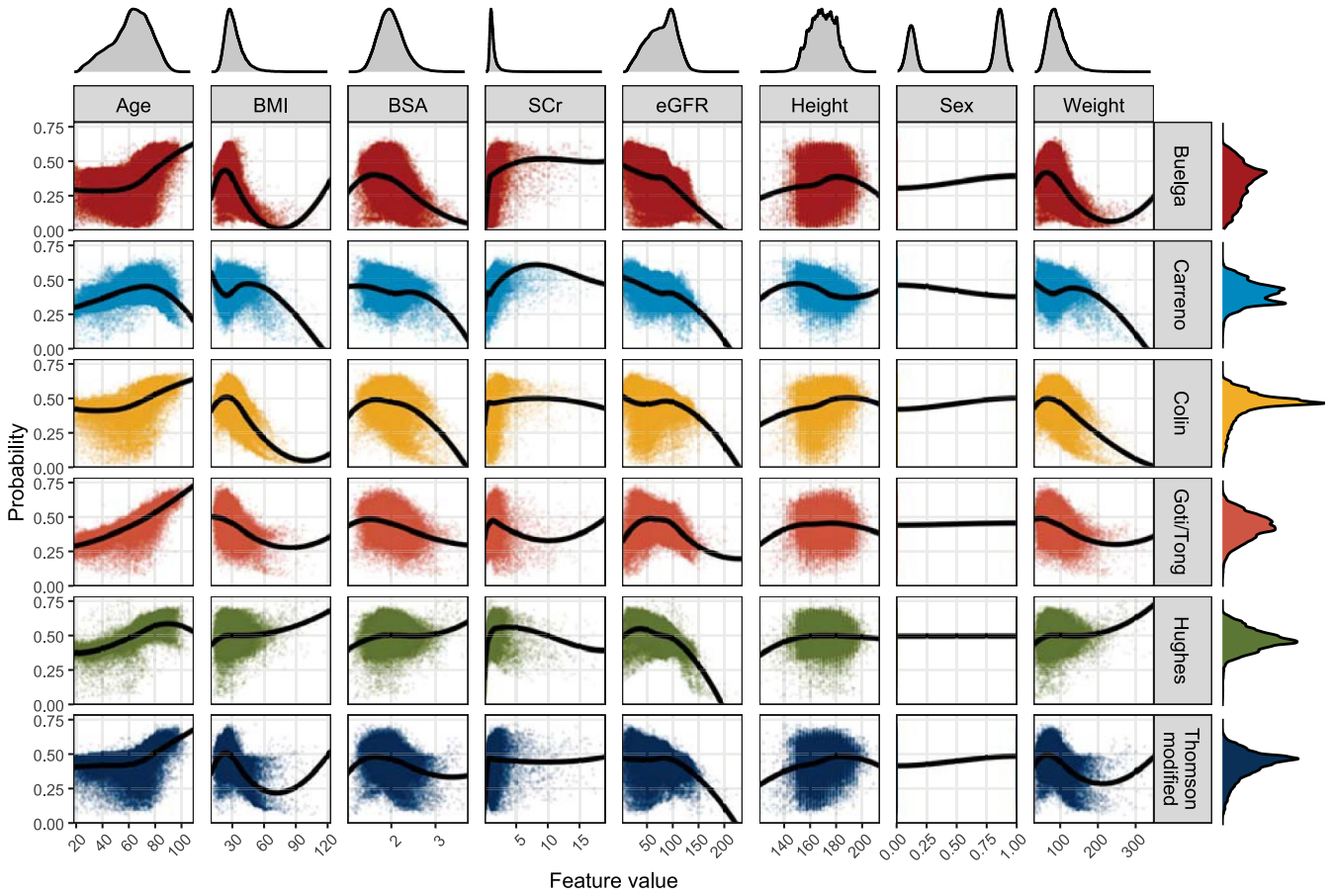

FIGURE 3 | Correlations between patient characteristics used as features in the ML model and the predicted probabilities that a priori predictions from the PK models fall within 80%–125% of the observed drug concentrations. The marginal distributions of feature values and probabilities are shown along the axes. Black lines indicate LOESS regression. Data shown are from the test set only. Age is expressed in years; BMI in kg/m^2; BSA in m^2; SCr in mg/dL; eGFR in $mL/min/1.73\,m^2$; height in cm; and weight in kg. For sex, a value of 1 corresponds to male. BMI, body mass index; BSA, body surface area; eGFR, estimated glomerular filtration rate; SCr, serum creatinine.

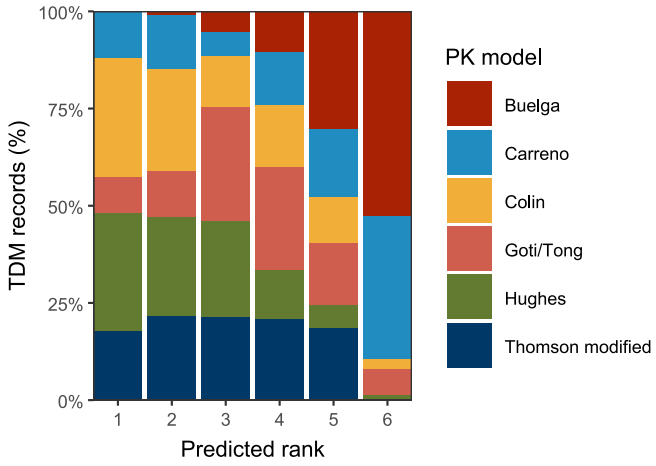

FIGURE 4 | Distribution of the predicted ranks for each PK model across the test set. TDM, therapeutic drug monitoring.

lowest-ranked model for each patient performed worse than any single PK model. This highlights the ML model's ability not only to identify suitable PK models but also to help avoid those likely to perform poorly for specific patients, reducing the risk of treatment failure or toxicity.

In addition to a ranking, the ML-predicted probabilities offer insights into the expected comparative accuracy of PK models, which can support clinical decision-making when integrating ML approaches into MIPD workflows. For example, if the top-ranked models are predicted to perform comparably, users may rely on other selection criteria, such as patient similarity to the model development population. Low predicted probabilities for all PK models could encourage earlier TDM sample collection to generate MAP Bayesian predictions, or prompt users to reduce reliance on Bayesian priors for that patient [35].

The ML-predicted probabilities also enabled model averaging, although this did not provide a meaningful performance improvement compared to ML-based model selection. Model selection further presents several implementation advantages over model averaging: it provides interpretable PK parameters, enables simulation of concentration–time profiles and AUCs, and is less of a "black box" to the user.

An advantage of the BR approach is that independent classifiers are trained for each PK model. This enables selective exclusion of models in downstream ranking and averaging, offering flexibility to adapt to different clinical contexts.

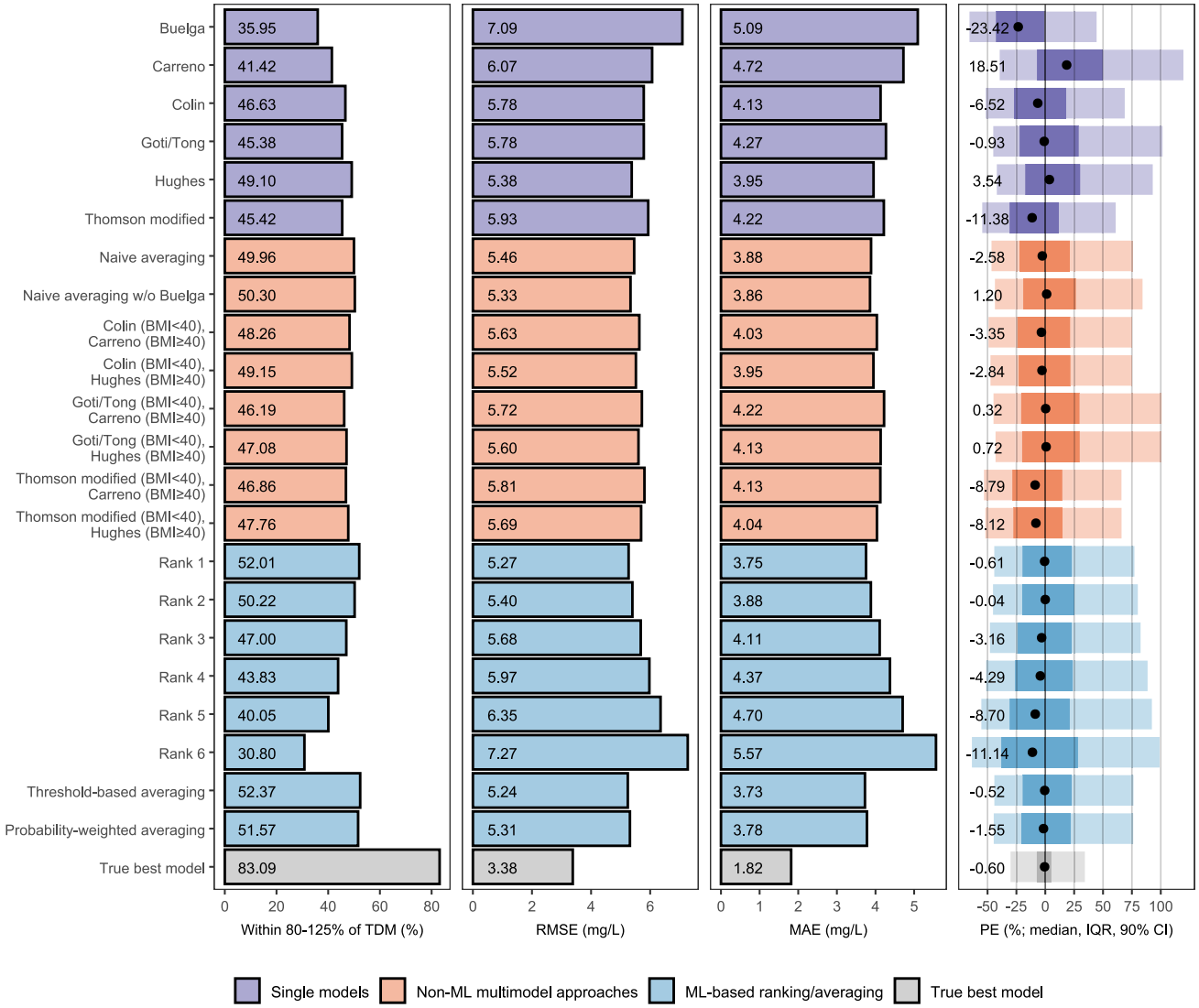

FIGURE 5 | Performance metrics of ML-based model ranking and averaging compared to single PK models, BMI-based model selection and naive averaging (with and without the Buelga model). The 'True best' model refers to the retrospectively identified best-performing PK model for each observation. CI, confidence interval; IQR, interquartile range; MAE, mean absolute error; ML, machine learning; PE, prediction error; RMSE, root mean squared error; TDM, therapeutic drug monitoring.

Several notable trends between the ML-predicted probabilities and patient characteristics were observed. In contrast to the other PK models, probabilities for the Carreno and Hughes models did not decrease for patients with obesity, which is not surprising as both were developed using data only from patients with BMIs $\geq 40\,kg/m^2$ [25, 29]. Interestingly, however, predicted probabilities—particularly for the Hughes model—were not necessarily lower for nonobese patients. This aligns with a previous study based on a partly overlapping dataset, which found that the Hughes model performed well across a wide range of BMIs [16]. One possible explanation is that this model uses fat-free mass as a covariate, which may allow better generalization to nonobese patients for this hydrophilic drug. The same study also reported lower accuracy of vancomycin PK model predictions in younger adults [16]. In the present study, we also observed that predicted probabilities increased with age and declined at eGFR estimates above ~100 mL/min/1.73 m^2, which is more common in younger adults. This may reflect the underrepresentation of young adults, who may have distinct physiology and comorbidities, in PK studies and highlights the need for vancomycin PK models that better serve this population [16]. Age and BMI were also among the most influential features in the ML model, supporting the relevance of evaluating vancomycin model performance across these characteristics.

There was a substantial gap between the performance of the ML approaches and the performance of the retrospectively identified best model (i.e., the hypothetical scenario where the best-performing PK model is always known a priori). In addition, the ability of the individual binary classification models within the multi-label framework to distinguish between positive and negative instances was limited. These outcomes suggest that routinely recorded patient characteristics explain only a limited part of the variability in vancomycin exposure and thus the comparative performance of vancomycin PK models. This aligns with the observed improvements in the predictive performance of vancomycin PK models when MAP Bayesian forecasting

based on TDM samples is applied compared to a priori predictions [7, 13, 16, 45]. A priori model selection and averaging approaches, such as the one presented here, should therefore be viewed as tools to support dosing decisions when TDM samples are not (yet) available rather than as replacements for Bayesian forecasting.

Our results echo findings from two recent studies that explored how ML approaches can support a priori model selection and averaging based on patient characteristics, albeit with different methodologies. Soeorg et al. developed subgroup identification-based approaches using classification trees for vancomycin models in neonates and infants, which modestly outperformed single best PK models in terms of the proportion of predictions within 20% of the observed values but not within 60% [46]. Similarly, Agema et al. demonstrated that model ensembling based on PK model-specific decision trees increased the proportion of patients in the target range for vancomycin and imatinib compared to the single best models [47]. Both studies were limited by dataset size, with the Soeorg and Agema studies including 208 and 91 vancomycin treatment episodes, respectively, from a single center for model training. In contrast, our training set comprised 240,545 patients from 155 centers across the United States. These aspects reduce the risk of overfitting and increase confidence that the ML model will generalize to external data, as supported by the similar performance on the training, validation, and test sets.

Our results also relate to a simulation study by Chan et al., who introduced Synthetic Model Combination (SMC), an ML-based model ensembling strategy, using vancomycin as a case study [48]. SMC assumes that models developed on populations more similar to a given patient will perform better for that patient. Some of our findings, such as the observed trends in probabilities with age and BMI, lend support to this idea. However, our results also suggest that routinely recorded patient characteristics explain only a limited part of the variability in comparative vancomycin PK model performance and that models may generalize beyond their original development populations, potentially limiting the clinical applicability of SMC-like methods.

This study has several limitations. First, this was a retrospective analysis, and while implausible data were removed, remaining data entry errors could not be identified or corrected. Moreover, patient characteristics beyond those included in our ML model, such as critical illness or concomitant administration of diuretics [49, 50], are known to influence vancomycin PK and thus the relative suitability of different PK models. Although the MIPD software used in this study allows users to input optional tags for such factors, these fields are not required for obtaining predictions and are therefore likely recorded inconsistently. This may explain why incorporating additional features did not improve model performance (data not shown). Future analyses could evaluate whether training an ML model on a prospective dataset with a broader range of potentially relevant patient characteristics improves predictive performance, despite its inevitably smaller size.

Second, the selected 80%–125% range used for labeling PK predictions is arbitrary. It was chosen because it treats prediction errors symmetrically on the log scale and approximately corresponds to the width of common vancomycin exposure target ranges [11]. We also explored single-label multi-class classification, where for each record the assigned label is the best-performing PK model. This approach removes the need to define a range of clinically acceptable prediction errors but performs worse than the multi-label classification approach in our dataset (results not shown). This may be due to inherent drawbacks of single-label classification in this context, including the loss of information about other PK models that perform similarly to the best model and the forced assignment of a positive label even when all PK model predictions are poor.

Third, although ML-based model selection and averaging improved prediction accuracy, we did not evaluate the impact on dosing decisions. Vancomycin doses are often rounded to the nearest 250 mg increment, meaning different PK model predictions may lead to identical dose recommendations, as demonstrated in a recent simulation study on vancomycin dosing in neonates [3].

In conclusion, our results demonstrate the potential of ML, specifically multi-label classification, for guiding individualized PK model selection in a priori MIPD of vancomycin in adults. While Bayesian forecasting remains essential once TDM samples become available, integrating ML-based model selection into MIPD workflows could enable making informed dosing decisions at an early stage to improve target attainment and clinical outcomes.

Author Contributions

W.O. wrote the manuscript. W.O., J.T.B., J.H.H., D.M.H.T., and R.J.K. designed the research. W.O., A.O, C.T., and C.L. performed the research. W.O. analyzed the data.

Conflicts of Interest

J.H.H., D.M.H.T., J.T.B. and R.J.K. are employees and stockholders of InsightRX. All other authors declared no competing interests for this work.

References

1. I. K. Minichmayr, E. Dreesen, M. Centanni, et al., "Model-Informed Precision Dosing: State of the Art and Future Perspectives," *Advanced Drug Delivery Reviews* 215 (2024): 115421, https://doi.org/10.1016/j.addr.2024.115421.

2. P. Gandia, S. Chaiben, N. Fabre, and D. Concordet, "Vancomycin Population Pharmacokinetic Models: Uncovering Pharmacodynamic Divergence Amid Clinicobiological Resemblance," *CPT: Pharmacometrics & Systems Pharmacology* 14 (2025): 142–151, https://doi.org/10.1002/psp4.13253.

3. M. El Hassani, M. Blouin, and A. Marsot, "A Simulation Study to Assess the Influence of Population Pharmacokinetic Model Selection on Initial Dosing Recommendations of Vancomycin in Neonates," *British Journal of Clinical Pharmacology* 91 (2025): 1223–1232, https://doi.org/10.1111/bcp.16345.

4. A. Broeker, M. Nardecchia, K. P. Klinker, et al., "Towards Precision Dosing of Vancomycin: A Systematic Evaluation of Pharmacometric Models for Bayesian Forecasting," *Clinical Microbiology and Infection* 25 (2019): 1286.e1–1286.e7, https://doi.org/10.1016/j.cmi.2019.02.029.

5. R. J. Keizer, R. Ter Heine, A. Frymoyer, L. J. Lesko, R. Mangat, and S. Goswami, "Model-Informed Precision Dosing at the Bedside: Scientific

Challenges and Opportunities," *CPT: Pharmacometrics & Systems Pharmacology* 7 (2018): 785–787, https://doi.org/10.1002/psp4.12353.

6. L. M. Schatz, S. Greppmair, A. K. Kunzelmann, et al., "Predictive Performance of Multi-Model Approaches for Model-Informed Precision Dosing of Piperacillin in Critically Ill Patients," *International Journal of Antimicrobial Agents* 64 (2024): 107305, https://doi.org/10.1016/j.ijantimicag.2024.107305.

7. D. W. Uster, S. L. Stocker, J. E. Carland, et al., "A Model Averaging/Selection Approach Improves the Predictive Performance of Model-Informed Precision Dosing: Vancomycin as a Case Study," *Clinical Pharmacology and Therapeutics* 109 (2021): 175–183, https://doi.org/10.1002/cpt.2065.

8. Z. L. Taylor, E. A. Poweleit, K. Paice, et al., "Tutorial on Model Selection and Validation of Model Input Into Precision Dosing Software for Model-Informed Precision Dosing," *CPT: Pharmacometrics & Systems Pharmacology* 12 (2023): 1827–1845, https://doi.org/10.1002/psp4.13056.

9. T. Guo, R. M. van Hest, L. F. Roggeveen, et al., "External Evaluation of Population Pharmacokinetic Models of Vancomycin in Large Cohorts of Intensive Care Unit Patients," *Antimicrobial Agents and Chemotherapy* 63 (2019): e02543-18, https://doi.org/10.1128/AAC.02543-18.

10. S. Greppmair, A. Brinkmann, A. Roehr, et al., "Towards Model-Informed Precision Dosing of Piperacillin: Multicenter Systematic External Evaluation of Pharmacokinetic Models in Critically Ill Adults With a Focus on Bayesian Forecasting," *Intensive Care Medicine* 49 (2023): 966–976, https://doi.org/10.1007/s00134-023-07154-0.

11. R. ter Heine, R. J. Keizer, K. van Steeg, et al., "Prospective Validation of a Model-Informed Precision Dosing Tool for Vancomycin in Intensive Care Patients," *British Journal of Clinical Pharmacology* 86 (2020): 2497–2506, https://doi.org/10.1111/bcp.14360.

12. M. El Hassani and A. Marsot, "External Evaluation of Population Pharmacokinetic Models for Precision Dosing: Current State and Knowledge Gaps," *Clinical Pharmacokinetics* 62 (2023): 533–540, https://doi.org/10.1007/s40262-023-01233-7.

13. P. J. Colin, D. J. Eleveld, A. Hart, and A. H. Thomson, "Do Vancomycin Pharmacokinetics Differ Between Obese and Non-Obese Patients? Comparison of a General-Purpose and Four Obesity-Specific Pharmacokinetic Models," *Therapeutic Drug Monitoring* 43 (2021): 126–130, https://doi.org/10.1097/FTD.0000000000000832.

14. J. H. Hughes, K. H. Woo, R. J. Keizer, and S. Goswami, "Clinical Decision Support for Precision Dosing: Opportunities for Enhanced Equity and Inclusion in Health Care," *Clinical Pharmacology and Therapeutics* 113 (2023): 565–574, https://doi.org/10.1002/cpt.2799.

15. A. Alsultan, S. A. Dasuqi, A. Almohaizeie, et al., "External Validation of Obese/Critically Ill Vancomycin Population Pharmacokinetic Models in Critically Ill Patients Who Are Obese," *Journal of Clinical Pharmacology* 64 (2024): 353–361, https://doi.org/10.1002/jcph.2375.

16. M.-S. A. Hughes, T. Lee, J. D. Faldasz, and J. H. Hughes, "Impacts of Age and BMI on Vancomycin Model Choice in a Bayesian Software: Lessons From a Very Large Multi-Site Retrospective Study," *Pharmacotherapy* 44 (2024): 794–802, https://doi.org/10.1002/phar.4613.

17. K. Stankevičiūtė, J.-B. Woillard, R. W. Peck, P. Marquet, and M. van der Schaar, "Bridging the Worlds of Pharmacometrics and Machine Learning," *Clinical Pharmacokinetics* 62 (2023): 1551–1565, https://doi.org/10.1007/s40262-023-01310-x.

18. E. A. Poweleit, A. A. Vinks, and T. Mizuno, "Artificial Intelligence and Machine Learning Approaches to Facilitate Therapeutic Drug Management and Model-Informed Precision Dosing," *Therapeutic Drug Monitoring* 45 (2023): 143–150, https://doi.org/10.1097/FTD.0000000000001078.

19. Z. Huang, P. Denti, H. Mistry, and F. Kloprogge, "Machine Learning and Artificial Intelligence in PK-PD Modeling: Fad, Friend, or Foe?," *Clinical Pharmacology and Therapeutics* 115 (2024): 652–654, https://doi.org/10.1002/cpt.3165.

20. M. McComb, R. Bies, and M. Ramanathan, "Machine Learning in Pharmacometrics: Opportunities and Challenges," *British Journal of Clinical Pharmacology* 88 (2022): 1482–1499, https://doi.org/10.1111/bcp.14801.

21. L. Keutzer, H. You, A. Farnoud, et al., "Machine Learning and Pharmacometrics for Prediction of Pharmacokinetic Data: Differences, Similarities and Challenges Illustrated With Rifampicin," *Pharmaceutics* 14 (2022): 1530, https://doi.org/10.3390/pharmaceutics14081530.

22. I. K. Minichmayr, T. Mizuno, S. Goswami, R. W. Peck, T. M. Polasek, and the American Society of Clinical Pharmacology and Therapeutics Precision Dosing Community, "Recent Advances Addressing the Challenges of Precision Dosing," *Clinical Pharmacology and Therapeutics* 116 (2024): 527–530, https://doi.org/10.1002/cpt.3365.

23. M. J. Rybak, J. Le, T. P. Lodise, et al., "Therapeutic Monitoring of Vancomycin for Serious Methicillin-Resistant *Staphylococcus aureus* Infections: A Revised Consensus Guideline and Review by the American Society of Health-System Pharmacists, the Infectious Diseases Society of America, the Pediatric Infectious Diseases Society, and the Society of Infectious Diseases Pharmacists," *American Journal of Health-System Pharmacy* 77 (2020): 835–864, https://doi.org/10.1093/ajhp/zxaa036.

24. D. S. Buelga, M. del Mar Fernandez de Gatta, E. V. Herrera, A. Dominguez-Gil, and M. J. García, "Population Pharmacokinetic Analysis of Vancomycin in Patients With Hematological Malignancies," *Antimicrobial Agents and Chemotherapy* 49 (2005): 4934–4941, https://doi.org/10.1128/AAC.49.12.4934-4941.2005.

25. J. J. Carreno, B. Lomaestro, J. Tietjan, and T. P. Lodise, "Pilot Study of a Bayesian Approach to Estimate Vancomycin Exposure in Obese Patients With Limited Pharmacokinetic Sampling," *Antimicrobial Agents and Chemotherapy* 61 (2017): e02478-16, https://doi.org/10.1128/AAC.02478-16.

26. P. J. Colin, K. Allegaert, A. H. Thomson, et al., "Vancomycin Pharmacokinetics Throughout Life: Results From a Pooled Population Analysis and Evaluation of Current Dosing Recommendations," *Clinical Pharmacokinetics* 58 (2019): 767–780, https://doi.org/10.1007/s40262-018-0727-5.

27. V. Goti, A. Chaturvedula, M. J. Fossler, S. Mok, and J. T. Jacob, "Hospitalized Patients With and Without Hemodialysis Have Markedly Different Vancomycin Pharmacokinetics: A Population Pharmacokinetic Model-Based Analysis," *Therapeutic Drug Monitoring* 40 (2018): 212–221, https://doi.org/10.1097/FTD.0000000000000490.

28. D. M. H. Tong, J. H. Hughes, and R. J. Keizer, "Use of Age-Adjusted Serum Creatinine in a Vancomycin Pharmacokinetic Model Decreases Predictive Performance in Elderly Patients," *Therapeutic Drug Monitoring* 43 (2021): 139–140, https://doi.org/10.1097/FTD.0000000000000819.

29. M.-S. A. Hughes, J. H. Hughes, J. Endicott, M. Langton, J. W. Ahern, and R. J. Keizer, "Developing Parametric and Nonparametric Models for Model-Informed Precision Dosing: A Quality Improvement Effort in Vancomycin for Patients With Obesity," *Therapeutic Drug Monitoring* 46 (2024): 575–583, https://doi.org/10.1097/FTD.0000000000001214.

30. A. H. Thomson, C. E. Staatz, C. M. Tobin, M. Gall, and A. M. Lovering, "Development and Evaluation of Vancomycin Dosage Guidelines Designed to Achieve New Target Concentrations," *Journal of Antimicrobial Chemotherapy* 63 (2009): 1050–1057, https://doi.org/10.1093/jac/dkp085.

31. R. V. Shingde, S. E. Reuter, G. G. Graham, et al., "Assessing the Accuracy of Two Bayesian Forecasting Programs in Estimating Vancomycin Drug Exposure," *Journal of Antimicrobial Chemotherapy* 75 (2020): 3293–3302, https://doi.org/10.1093/jac/dkaa320.

32. L. A. Inker, N. D. Eneanya, J. Coresh, et al., "New Creatinine- and Cystatin C–Based Equations to Estimate GFR Without Race," *New*

England Journal of Medicine 385 (2021): 1737–1749, https://doi.org/10.1056/NEJMoa2102953.

33. A. Rivolli and A. C. P. L. F. De Carvalho, "The Utiml Package: Multi-Label Classification in R," *R Journal* 10 (2018): 24–37, https://doi.org/10.32614/RJ-2018-041.

34. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, 2016), 785–794.

35. J. H. Hughes and R. J. Keizer, "A Hybrid Machine Learning/Pharmacokinetic Approach Outperforms Maximum a Posteriori Bayesian Estimation by Selectively Flattening Model Priors," *CPT: Pharmacometrics & Systems Pharmacology* 10 (2021): 1150–1160, https://doi.org/10.1002/psp4.12684.

36. C. Codde, F. Rivals, A. Destere, et al., "A Machine Learning Approach to Predict Daptomycin Exposure From Two Concentrations Based on Monte Carlo Simulations," *Antimicrobial Agents and Chemotherapy* 68 (2024): e0141523, https://doi.org/10.1128/aac.01415-23.

37. L. Ponthier, P. Ensuque, A. Destere, et al., "Optimization of Vancomycin Initial Dose in Term and Preterm Neonates by Machine Learning," *Pharmaceutical Research* 39 (2022): 2497–2506, https://doi.org/10.1007/s11095-022-03351-6.

38. Q.-Y. Li, B. H. Tang, Y. E. Wu, et al., "Machine Learning: A New Approach for Dose Individualization," *Clinical Pharmacology and Therapeutics* 115 (2024): 727–744, https://doi.org/10.1002/cpt.3049.

39. J.-B. Woillard, M. Labriffe, J. Debord, and P. Marquet, "Tacrolimus Exposure Prediction Using Machine Learning," *Clinical Pharmacology and Therapeutics* 110 (2021): 361–369, https://doi.org/10.1002/cpt.2123.

40. J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research* 13 (2012): 281–305.

41. R. Keizer, J. Hughes, D. Tong, K. Woo, and J. Brooks, "PKPDsim: Tools for Performing Pharmacokinetic-Pharmacodynamic Simulations," R Package Version 1.4.0, InsightRX, 2025.

42. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2024).

43. M. Kuhn, "Building Predictive Models in R Using the Caret Package," *Journal of Statistical Software* 28 (2008): 1–26, https://doi.org/10.18637/jss.v028.i05.

44. H. Wickham, M. Averick, J. Bryan, et al., "Welcome to the Tidyverse," *Journal of Open Source Software* 4, no. 43 (2019): 1686, https://doi.org/10.21105/joss.01686.

45. J. H. Hughes, D. M. H. Tong, S. S. Lucas, J. D. Faldasz, S. Goswami, and R. J. Keizer, "Continuous Learning in Model-Informed Precision Dosing: A Case Study in Pediatric Dosing of Vancomycin," *Clinical Pharmacology and Therapeutics* 109 (2021): 233–242, https://doi.org/10.1002/cpt.2088.

46. H. Soeorg, R. Kalamees, I. Lutsar, and T. Metsvaht, "Subgroup Identification-Based Model Selection to Improve the Predictive Performance of Individualized Dosing," *Journal of Pharmacokinetics and Pharmacodynamics* 51 (2024): 253–263, https://doi.org/10.1007/s10928-024-09909-8.

47. B. C. Agema, T. Kocher, A. B. Öztürk, et al., "Selecting the Best Pharmacokinetic Models for a Priori Model-Informed Precision Dosing With Model Ensembling," *Clinical Pharmacokinetics* 63 (2024): 1449–1461, https://doi.org/10.1007/s40262-024-01425-9.

48. A. Chan, R. Peck, M. Gibbs, and M. van der Schaar, "Synthetic Model Combination: A New Machine-Learning Method for Pharmacometric Model Ensembling," *CPT: Pharmacometrics & Systems Pharmacology* 12 (2023): 953–962, https://doi.org/10.1002/psp4.12965.

49. S. I. Blot, F. Pea, and J. Lipman, "The Effect of Pathophysiology on Pharmacokinetics in the Critically Ill Patient – Concepts Appraised by the Example of Antimicrobial Agents," *Advanced Drug Delivery Reviews* 77 (2014): 3–11, https://doi.org/10.1016/j.addr.2014.07.006.

50. S. E. Medellín-Garibay, B. Ortiz-Martín, A. Rueda-Naharro, B. García, S. Romano-Moreno, and E. Barcia, "Pharmacokinetics of Vancomycin and Dosing Recommendations for Trauma Patients," *Journal of Antimicrobial Chemotherapy* 71 (2016): 471–479, https://doi.org/10.1093/jac/dkv372.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1.** Supplementary text. **Figure S1.** Relationship between XGBoost hyperparameter values and performance of the Binary Relevance model. Mean average precision is shown for 100 randomly sampled hyperparameter combinations. The red dot indicates the combination selected for the final model. Blue lines represent LOESS smoothers. Hyperparameter definitions: colsample_bytree, proportion of columns subsampled for each tree; eta, learning rate; max_depth, maximum tree depth; nrounds, number of boosting iterations; subsample, subsample proportion of training rows. **Figure S2.** Performance metrics for all evaluated approaches when the Buelga model was excluded from the analysis. Results are shown for the test set. Abbreviations: TDM, therapeutic drug monitoring; ML, machine learning; RMSE, root mean squared error; MAE, mean absolute error; PE, prediction error; IQR, interquartile range; CI, confidence interval. **Table S1.** Performance of evaluated transformation strategies for converting multi-label classification tasks into subtasks suitable for classification algorithms on the validation dataset. Two strategies (CLR and RPC) output a direct ranking of PK models. In cases of tied ranks, a random tie-break was applied.