



**HAL**  
open science

## A Machine Learning Approach to Predict Interdose Vancomycin Exposure

Mehdi Bououda, David W Uster, Egor Sidorov, Marc Labriffe, Pierre Marquet, Sebastian G Wicha, Jean-Baptiste Woillard

► **To cite this version:**

Mehdi Bououda, David W Uster, Egor Sidorov, Marc Labriffe, Pierre Marquet, et al.. A Machine Learning Approach to Predict Interdose Vancomycin Exposure. *Pharmaceutical Research*, 2022, 39 (4), pp.721-731. 10.1007/s11095-022-03252-8 . hal-04414352

**HAL Id: hal-04414352**

**<https://unilim.hal.science/hal-04414352>**

Submitted on 29 Jan 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# A machine learning approach to predict interdose vancomycin exposure

Mehdi Bououda<sup>a</sup>, David W. Uster<sup>b</sup>, Egor Sidorov<sup>c</sup>, Marc Labriffe<sup>a,c</sup>, Pierre Marquet<sup>a,c</sup>, Sebastian G. Wicha<sup>b</sup>, Jean-Baptiste Woillard<sup>a,c</sup>

- a) P&T, UMR1248 Université de Limoges, INSERM, Limoges, France.
- b) Dept. of Clinical Pharmacy, Institute of Pharmacy, University of Hamburg, Hamburg, Germany
- c) CHU Limoges, Service de pharmacologie, toxicologie et pharmacovigilance, Limoges, France

Mail address for each author : [mehdi.bououda@etu.unilim.fr](mailto:mehdi.bououda@etu.unilim.fr), [david.uster@uni-hamburg.de](mailto:david.uster@uni-hamburg.de),  
[egor.sidorov@etu.unilim.fr](mailto:egor.sidorov@etu.unilim.fr), [marc.labriffe@unilim.fr](mailto:marc.labriffe@unilim.fr), [pierre.marquet@unilim.fr](mailto:pierre.marquet@unilim.fr),  
[Sebastian.wicha@uni-hamburg.de](mailto:Sebastian.wicha@uni-hamburg.de), [jean-baptiste.woillard@unilim.fr](mailto:jean-baptiste.woillard@unilim.fr)

Corresponding author:

Jean-Baptiste Woillard, PharmD, PhD

ORCID: 0000-0003-1695-0695

Address: Service de pharmacologie, toxicologie et pharmacovigilance, CBRS, CHU Limoges, 2 rue du Pr Descottes, 87000 Limoges Cedex, France

Phone: +33 5 55 05 61 40

Fax: +33 5 55 05 61 62

Email: [jean-baptiste.woillard@unilim.fr](mailto:jean-baptiste.woillard@unilim.fr)

Keywords: machine learning; vancomycin; population pharmacokinetics; Model informed precision dosing; simulations

## Abstract

Introduction: Estimation of vancomycin area under the curve (AUC) is challenging in the case of discontinuous administration. Machine learning approaches are increasingly used and can be an alternative to population pharmacokinetic (POPPK) approaches for AUC estimation.

The objectives were to train XGBoost algorithms based on simulations performed in a previous POPPK study to predict vancomycin AUC from early concentrations and a few features (i.e. patient information) and to evaluate them in a real-life external dataset in comparison to POPPK.

Patients and Methods: Six thousand simulations performed from 6 different POPPK models were split into training and test sets. XGBoost algorithms were trained to predict trapezoidal rule AUC a priori or based on 2, 4 or 6 samples and were evaluated by resampling in the training set and validated in the test set. Finally, the 2-sample algorithm was externally evaluated on 28 real patients and compared to a state-of-the-art POPPK model-based averaging approach.

Results: The trained algorithms showed excellent performances in the test set with relative mean prediction error (MPE)/ imprecision (RMSE) of the reference AUC=3.3/18.9, 2.8/17.4, 1.3/13.7 % for the 2, 4 and 6 samples algorithms respectively. Validation in real patient showed flexibility in sampling time post-treatment initiation and excellent performances MPE/RMSE<1.5/12 % for the 2 samples algorithm in comparison to different POPPK approaches.

Conclusions: The Xgboost algorithm trained from simulation and evaluated in real patients allow accurate and precise prediction of vancomycin AUC. It can be used in combination with POPPK models to increase the confidence in AUC estimation.

# 1. Introduction

Vancomycin belongs to the glycopeptides and is largely used to treat Gram-positive bacteria resistant to methicillin. It is characterized by a narrow therapeutic range, in particular in special patient populations (renal injury, critically ill) highlighting the need for therapeutic drug monitoring (TDM).

The last international recommendation suggest the use of its area under the curve (AUC) for TDM with an AUC/MIC target 400-600 [1]. In the case of a continuous infusion, AUC is easily calculated by multiplication of the concentration at steady state (mg/l) with the interdose interval (hours) while in the case of discontinuous administration, the AUC cannot be directly calculated and the trapezoidal rule based on many samples remains theoretically the gold standard. However, trapezoidal rule AUC is not compatible with a routine care practice, as it requires many samples. An alternative is to use population pharmacokinetic (POPPK) models and maximum a posteriori Bayesian estimators (MAP-BE) that allows to estimate individual parameters and exposure based on a few covariates and a limited sampling strategy. Recently, a model averaging approach (MAA), where individual model predictions are averaged across a number of candidate models weighted by the individual model fit, has shown more consistent performance in simulated and real-life clinical data in comparison to the classical single-modeling by automatically weighting the candidate models for each patient [2]. Another approach is the model-selection algorithm (MSA) where the best performing model in an individual patient is selected automatically and the other candidate models are discarded. Based on simulated data, Uster et al demonstrated an adequate predictive performance of the approaches using the relative mean prediction error (rMPE) and relative root mean square error (rRMSE) across the 6000 simulated patients [2]. The MAA and MSA approaches are both implemented in TDMx (<http://www.tdmx.eu/>).

Up to now, machine learning (ML) approaches have not been extensively used in pharmacometrics. We recently developed Xgboost algorithms [3] that allowed to accurately estimate the interdose AUC

for two immunosuppressants, trained either from MAP-BE estimated AUC or from POPPK simulations [4–6]. Interestingly, predictions using these algorithms were numerically better than those obtained using MAP-BE and offered simplifications of the limited sampling strategy with two samples required only (vs 3 for POPPK approaches).

In this context, the objectives of this work were to train Xgboost algorithms, based on simulations designed in the model averaging study, allowing predicting vancomycin AUC from early concentrations and a few features and to evaluate them in a real-life external dataset in comparison to MAA and MSA approaches.

## 2. Materials and methods

### 2.1. Data

Simulated data from the previous model averaging study [2], which aimed to validate two automated POPPK approaches in their (Bayesian) forecasting performance, were used to develop and validate the Xgboost algorithms. Using such simulated data allowed to calculate the reference AUC using the trapezoidal rules method but also to compare the performances of the developed algorithms to the multi-model approaches. Briefly, simulations have been performed from 6 POPPK models from heterogeneous populations including extremely obese [7], critically ill [8], hospitalized [9], and those with sepsis [10], trauma [11] and post-heart surgery [12]. For each of the 6 POPPK models, 1000 individuals receiving 1000 mg every 12 hours through 1h infusion from time 0 to time 36h and 4 covariates (creatinine, age, weight, body size with correlation between the last 3 variables) were simulated [2]. AUC<sub>24-36h</sub> were calculated for each simulated profile and 4 scenarios of different combination of concentration time (peak and trough) were evaluated as performed in the original article: “a priori”, “one occasion”, “two occasions” and “general fit”. Briefly, the a priori scenario only included demographic and lab results, the one occasion scenario included 2 concentrations sampled

at 13.25h and 23.75h post treatment start (2 samples), the two occasion included the same as one occasion + concentrations sampled at 1.25h and 11.75h post administration (4 samples) and the general fit included the same as two occasions + concentrations sampled at 25.25h and 35.75h post administration (6 samples). A table of the sampling scheme in each scenario is provided in Supplemental data (Supplemental Table 1).

## 2.2. Development of the Xgboost algorithms

All the analysis were performed in R version 4.1.0 using the Tidyverse [13] and Tidymodels [14] frameworks.

### 2.2.1. Exploratory data analysis

The linear association between AUC<sub>24-36h</sub> and features were explored using a correlation matrix.

### 2.2.2. Preprocessing and features engineering

As the data were simulated, there was no missing data. No preprocessing was applied as Xgboost does not require data normalization. A new feature was calculated (difference in concentrations) corresponding to the difference between the peak and trough value. Data splitting was performed by random selection of patients in a training (75%) and a test set (25%).

### 2.2.3. Training of the algorithm

Hyperparameters were tuned in the training set using 10-fold cross validation (grid of 30 semi-random combinations of hyperparameters) by selecting the ones associated with the lowest RMSE between estimated and reference AUC. The principle of gradient boosting, the reasons that led us to use the Xgboost algorithm and the detail of hyperparameters tuned are presented in the Supplemental method. The relative importance was determined using random permutations and variable importance plots were drawn. The retained algorithms were evaluated using additional 10-fold cross-validations to assess the mean  $\pm$  standard deviation (SD) of RMSE and r square ( $r^2$ ) in the training set (allowing to prevent overfitting). Scatter plots of estimated vs. reference AUC were

drawn. Finally, AUC24-36h prediction was performed in the test set and RMSE and r2 as well as rMPE, rRMSE, number and contingency table of prediction and observation with AUC24\_36h <200 mg\*h/L, between 200 and 400 mg\*h/L and >400 mg\*h/L. The code used for the development of the Xgboost algorithm is provided in Supplemental data.

### 2.3. Evaluation of the simulations

The analysis was performed on the whole simulations (n=6000) with a data splitting stratified by simulated populations (800 for training and 200 for testing in the 6 simulated populations). One algorithm was trained and evaluated for each of the 4 strategies (a priori, 1 occasion, 2 occasions and general fit) and rMPE or rRMSE were compared overall and split by study. In a second time, algorithms were trained in each simulated population and performances were investigated in the 5 other simulated populations for each of the 3 strategies (excluding a priori) leading to 18 algorithms trained (3 strategies \* 6 simulated populations).

### 2.4. Evaluation in real patient

Finally, the algorithm was evaluated in 28 anonymized real patients extracted from the PKJust website (<https://pharmaco.chu-limoges.fr/>), after complete de-identification of the data, in accordance with the European GDPR regulation. This program is used for routine care vancomycin and aminoglycoside dose adjustment in the Limoges University Hospital. Patient having a dose adjustment request between 01/01/2014 and 31/12/2019 with a twice daily administration of vancomycin through 1h infusion were selected. Two samples from the same dosing interval were extracted with one being drawn maximum 30 min after the infusion (i.e. peak). The time since initiation of the vancomycin was not used as criterion, to evaluate the flexibility of the ML algorithm to predict based on different occasion as supplied during training of the algorithm. As the “true” reference AUC was not available, the AUC estimations were compared to those of the PKJust program and those of the MAA and MSA approaches implemented in TDMx platform

(<http://www.tdmx.eu/>) by calculation of  $\sigma$  and relative MPE/RMSE and scatter plot and Bland Altman plot were drawn.

## 3. Results

### 3.1. Data

The summary of the dataset simulated for the whole simulations (n=6000) and the characteristics of the patients used for external evaluation are presented in Table 1. The correlation matrix of simulated features is presented in Supplemental Figure 1.

### 3.2. Xgboost algorithms in the whole dataset

Performances of the general model based on the overall dataset split by simulated population and obtained in the test set are presented in Figure 1 (without the a priori scenario). The overall performances in the training set using cross-validation and in the test set are presented in Table 2. The performances were excellent for the a posteriori scenario with rRMSE values lower than 22.5% and rMPE lower than 4.1% (corresponding to the worse performances obtained for 1 occasion in the test set for simulations from post heart surgery patients ; Table 3). As expected the performances were drastically less accurate and precise for the a priori scenario (highest rMPE/rRMSE for simulations from critically ill patients = 36.5%/54.5%). Interestingly, the use of the covariate “simulated population” (that require to a priori classify the patient in one of the 6 indication between extremely obese, critically ill, hospitalized, sepsis, trauma and post-heart surgery) largely improved the estimation (rMPE/rRMSE in the test set = 6.2/28.1%).

The contingency table of the test set showing the number of patients with AUC<200, between 200 and 400 and >400 observed and predicted is presented in the Figure 2 for the 3 a posteriori scenario.

The variable importance plot (Figure 3 for the 3 a posteriori scenario) showed that the concentrations observed were the most important features to predict AUC<sub>24-36h</sub> and that among



the concentrations, the current through (C23.75 or C35.75) and peak (C25.25) concentrations were the most important (in comparison to past trough or peak concentrations or other features). After the concentrations, the serum creatinine and simulated population identifier were the most important clinical features. The scatter plot in the test set of the predicted AUC as function of observed AUC for the 3 a posteriori scenario is presented in the supplemental Figure 2.

### 3.3. Xgboost algorithms in simulated populations

The rMPE and rRMSE of the algorithms trained in one simulated population but used to determine the AUC in the others is presented in Table 4 for the 3 a posteriori scenarios. The algorithms developed in the hospitalized, sepsis and critically ill simulated populations were, in mean, the ones associated with the lowest rMPE and rRMSE when evaluated in the other simulated populations (evaluating generalisability) for the 3 scenarios studied while the algorithm trained in post heart surgery was performing the worst (poor generalisability).

### 3.4. Comparison with the model averaging approach

Comparison of the performances in the training (using cross-validation), the test set vs. the MAA for each scenario and each simulated population using the model developed on the whole population is performed in Table 2 and split by models in Table 3.

#### 3.4.1. A priori

The a priori performances of the Xgboost algorithm were very close to the ones of the MAA approach overall (Table 2) or split by simulated population (Table3).

#### 3.4.2. A posteriori

The performances based on observed samples were excellent and very similar between the ML algorithms and the model averaging approach with mean rMPE lower than 5% and rRMSE <20% whatever the scenario was for the training, test and MAA approaches. Interestingly, as shown in Table 3, some of the bad predictions with ML approaches were better using MAA and vice versa (e.g

rRMSE for general fit = 22.4% for ML vs 7.71% for MAA in the extremely obese subpopulation; rMPE for 2 occasions = 2.27 % for ML vs 7.88% for MAA in the critically ill subpopulation) showing complementarity between the 2 approaches.

### 3.5. Evaluation in real-life patients

Performances were excellent with a rMPE /rRMSE = 0.97 / 12.7 % corresponding to a MPE / RMSE = -3.85 / 28.5 mg/L when compared to PKJust , a rMPE /rRMSE = 0.8 / 11.9 % corresponding to a MPE / RMSE = 0.17 / 24.2 mg/L when compared to MSA and rMPE /rRMSE = 1.2 / 11.8 % corresponding to a MPE / RMSE = 1.38 / 24.2 mg/L when compared to MAA. There were only 2 out of the 28 patients (7.3%) that exhibited a relative MPE value > 20% between both PKJust, MAA or MSA and the ML approach. The scatter plot and Bland Altman for ML vs PKJust, MSA or MAA are presented in Figure 4.

## 4. Discussion

In this study, we trained Xgboost algorithms on early vancomycin simulated concentrations obtained from different POPPK models to predict the vancomycin AUC<sub>24-36h</sub>. The performances were as good as the ones of the recently published model averaging approach that performed slightly better than individual POPPK models [2]. Finally, the performances obtained using ML algorithms in real patients with sparse and various sampling times were comparable to those obtained using PKJust or the MAA and MSA.

The choice of using the simulations performed by Uster et al to train our ML algorithms was to perform a direct comparison with the model averaging approach. However this added the constraint to not use in our algorithm the dose or the deviation to theoretical sampling time as predictors as previously done [6] because in the Uster et al simulations, the dose administered was fixed to 1g for every simulated patient (whatever the renal function was) and no uncertainty was added on

sampling times. Additionally, the simulations in Uster et al being drawn up to 36h, we have been obliged therefore to predict the AUC<sub>24-36h</sub> that is not the AUC usually measured in routine practice because the steady state is not reached for most of the patients. Indeed, in patients with a normal renal function, steady state is reached around 48h after the beginning of the vancomycin. The reason for that is that POPPK guided dosing does not require steady state doses, which enables early adjustments of the antibiotic therapy leading to a decrease in death rate [15]. Nonetheless, the algorithm developed was evaluated in real patients sampled after 36h leading to accurate estimations of the interdose AUC.

Overall, the performances obtained with the ML algorithms on the whole population or split by simulated populations for the a priori scenario were similar than the ones of Uster et al. The use of the simulated population covariate improved largely the performance. However, that implies an a priori classification of the patient into a distinct population, ignoring atypical patients and increasing the risk of bias for them. Additionally, patients could belong to multiple distinct population as for example being both obese and critically ill.

The a posteriori results obtained with the ML algorithms on the whole population or split by simulated populations were very comparable to the ones of Uster et al. That confirms the results of our ML algorithms as the MAA seems to be a promising pharmacometric approach, taking into account the diversity of the population through the use of different priors. We observed that using the ML approaches, all the relative prediction errors were positive (Table 3) without a clear explanation for that. Interestingly, the variable indicative of the simulated population from which the model was produced exhibits a lower importance in comparison to the concentrations in the ML algorithms. In the context of MAP-BE, this has been initially described years ago by Sheiner and Beal for digoxin that addition of covariate does not carry as much information as one plasma concentration [16].

Another question was about the performances of ML algorithms trained in a given simulated population on the other simulations i.e. does the population used for training the algorithm highly influence the performances in other patient populations. We observed that it influences the performances and the results presented in Table 4 showed that the ML algorithms trained from the simulations from sepsis [10] and hospitalised patients [9] were associated with the best performances that can be interpreted as they are the most generalizable algorithms. This is insofar remarkable and unexpected, as the population in the sepsis patients model was critically ill and vancomycin was administered as continuous infusion. Hence, more studies might be needed to confirm this. In any cases, the lowest performance of the ML algorithms trained from the simulations of the other models only means that they are less generalizable for other populations. In routine care, one can imagine to choose *a priori* a ML algorithm developed from a given simulated population based on the patient characteristics. However, as the general model developed on the overall population exhibited very good performances in every simulated populations, its use is more flexible than choosing one or the other ML algorithm developed in a given simulated population depending on the patient characteristics.

As expected, the one occasion gave the lowest performances in comparison to the general fit. However, the loss in performances is largely balanced by its usability in routine clinical cares as it only requires 2 samples vs 6 samples for the general fit.

The ML algorithm developed theoretically requires very strict adherence to sampling times and is largely less flexible than MAP-BE for clinical routine. That led us to investigate its performances in real patient data sampled at time different from the ones used for the train of the algorithm (peak = 13.25h and trough = 23.75h for the one occasion scenario). The only restriction was to accept samples drawn max 30min after the infusion for peak, and peak and trough drawn during the same infusion. The first criteria can be easily explained by the fact that as the ML algorithm was trained on concentration at peak, any deviation from the peak will decrease the concentration leading to a

under prediction of the true AUC. The second criteria was to avoid intra-individual variability from one infusion to another. Of note, the maximal deviation between the trough concentration and the next infusion was 4 hours that give information about the flexibility of our ML algorithm with trough sampling. The performances were very similar whatever the comparator was (MSA or MAA) with rMPE lower than 1.5% and relative imprecision < 12%. There were 2 patients which exhibited a rMPE>20% but a careful examination of these patient does not highlighted specific particularities (#1 72 years-old woman with serum creatinine = 93 $\mu$ M, receiving a dose of 400mg bid, after 6 doses, trough and peak concentrations = 7.5 and 18 mg/L respectively, AUC MSA/MAA vs ML = 144 vs 194 mg\*h/l; #2 68 years-old man with serum creatinine = 51 $\mu$ M, receiving a dose of 1500mg bid, after 5 doses, trough and peak concentrations = 7.6 and 26.2 mg/L respectively, AUC MSA/MAA vs ML = 205 vs 160 mg\*h/l).

One can ask the added value of the ML in addition to standard MAA. These results show the complementarity of these approaches and highlights a possible use of ML approaches in complement to POPPK and MAP-BE approach: in the case of important differences in the AUC estimation between both approaches, the results should be interpreted very carefully while in case of agreement, the confidence in AUC estimation can be increased. Additionally, the Xgboost algorithm do not rely on a prespecified distribution while parametric POPPK approach is constraint to normality and it could learn nonlinear process. Theoretically, the ML algorithm should perform better on atypical patient and decrease the bias once a similar profile has been learn during the development of the algorithm. The current algorithm, has been developed from simulations from a parametric POPPK model and thus suffers from the same limitations. However, the algorithm might be improved by feeding it with new real observed data which would theoretically leave the normal distribution and improve the performances.

The best way to develop such algorithms is to use real experimental data. However in most of the cases, no such large amount of experimental data required to develop ML algorithm ( ie data driven

approach) is available. The comparison of the performances of ML algorithms developed from simulations to the one developed from abbreviated PK profiles has already been performed in our previous studies with tacrolimus without clear differences in performances [6,17]. The comparison to algorithms developed from full real PK profiles has to be performed in future studies. Nevertheless, a limit of the generalization of this approach that learn from simulation derived from a POPPK model is (i) the availability of a POPPK model and (ii) that this model has to be well specified in order to obtain “realistic” simulations. A way to improve the simulations is to apply some filters to remove very unrealistic ones. An additional limitation (but also applicable to POPPK) is that the ML algorithm developed from these simulations is probably not extrapolatable in other patient groups not covered by the initial model. Finally, the evaluation/validation in real data is very important for algorithms developed from simulated data. The limits of the ML algorithm when compared to the MAP-BE approaches were that it does not allow to draw simulations for probability of target attainment, to evaluate different administration scheme or to estimate when the steady state will be attained. Its goal is only to estimate the interdose AUC with the lowest error. However, the ML algorithms in the present article were trained for a 12h interdose and other interdoses (6, 8 or 24h) would require to train algorithms for each interdose that limits the use of this approach in routine clinical care for drug with many dosing schemes (including vancomycin and other antibiotics as beta lactams). Another important limit is the black box phenomenon observed with such complicated algorithm [18]. The variable importance plot allows however to better understand important features.

Beyond this study, there are many applications of ML in the field of pharmacometrics in complement or substitution of POPPK modeling. The first one is the prediction of the first dose using these type of algorithms. ML could better learn nonlinear associations between first dose requirement of a drug and factors of variability. The second application is the use of such approach as presented in the present article, i.e. in substitution to POPPK modeling to predict a PK parameter (e.g. CL) or an exposure metrics (here AUC). Finally, there are applications that mix both approaches to improve the individual predictions in antibiotic clearances [19]. Another good example is the work performed by

Hughes and Keizer in which the authors developed a ML algorithm in addition to POPPK to choose if flatten priors have to be used for vancomycin MIPD based on an individual observation [20].

## 5. Conclusions

The Xgboost algorithm trained in the present study allows accurate AUC estimation for a 12h interdose in simulated and real patients with flexibility on the time post infusion. Further validation of this proof of concept study in larger populations is needed. The place of such algorithms is not well defined but it seems complementary to standard POPPK approaches and could be used in addition to them to increase the confidence in the results.

## References

1. Wicha SG, Märtson A-G, Nielsen EI, Koch BCP, Friberg LE, Alffenaar J-W, et al. From Therapeutic Drug Monitoring to Model-Informed Precision Dosing for Antibiotics. *Clin Pharmacol Ther.* 2021;109:928–41.
2. Uster DW, Stocker SL, Carland JE, Brett J, Marriott DJE, Day RO, et al. A Model Averaging/Selection Approach Improves the Predictive Performance of Model-Informed Precision Dosing: Vancomycin as a Case Study. *Clin Pharmacol Ther.* 2021;109:175–83.
3. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16.* 2016;785–94.
4. Woillard J-B, Labriffe M, Debord J, Marquet P. Mycophenolic acid exposure prediction using machine learning. *Clinical Pharmacology & Therapeutics.* 2021;
5. Woillard J-B, Labriffe M, Debord J, Marquet P. Tacrolimus exposure prediction using machine learning. *Clin Pharmacol Ther.* 2020;
6. Woillard J-B, Labriffe M, Prémaud A, Marquet P. Estimation of drug exposure by machine learning based on simulations from published pharmacokinetic models: The example of tacrolimus. *Pharmacol Res.* 2021;167:105578.
7. Adane ED, Herald M, Koura F. Pharmacokinetics of vancomycin in extremely obese patients with suspected or confirmed *Staphylococcus aureus* infections. *Pharmacotherapy.* 2015;35:127–39.
8. Revilla N, Martín-Suárez A, Pérez MP, González FM, Fernández de Gatta MDM. Vancomycin dosing assessment in intensive care unit patients based on a population pharmacokinetic/pharmacodynamic simulation. *Br J Clin Pharmacol.* 2010;70:201–12.
9. Thomson AH, Staatz CE, Tobin CM, Gall M, Lovering AM. Development and evaluation of vancomycin dosage guidelines designed to achieve new target concentrations. *J Antimicrob Chemother.* 2009;63:1050–7.
10. Roberts JA, Taccone FS, Udy AA, Vincent J-L, Jacobs F, Lipman J. Vancomycin dosing in critically ill patients: robust methods for improved continuous-infusion regimens. *Antimicrob Agents Chemother.* 2011;55:2704–9.
11. Medellín-Garibay SE, Ortiz-Martín B, Rueda-Naharro A, García B, Romano-Moreno S, Barcia E. Pharmacokinetics of vancomycin and dosing recommendations for trauma patients. *J Antimicrob Chemother.* 2016;71:471–9.
12. Mangin O, Urien S, Mainardi J-L, Fagon J-Y, Faisy C. Vancomycin pharmacokinetic and pharmacodynamic models for critically ill patients with post-sternotomy mediastinitis. *Clin Pharmacokinet.* 2014;53:849–61.
13. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. *Journal of Open Source Software.* 2019;4:1686.
14. Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. [Internet]. 2020. Available from: <https://www.tidymodels.org>



15. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit Care Med.* 2006;34:1589–96.
16. Sheiner LB, Beal S, Rosenberg B, Marathe VV. Forecasting individual pharmacokinetics. *Clin Pharmacol Ther.* 1979;26:294–305.
17. Woillard J-B, Labriffe M, Debord J, Marquet P. Tacrolimus Exposure Prediction Using Machine Learning. *Clin Pharmacol Ther.* 2021;110:361–9.
18. Badillo S, Banfai B, Birzele F, Davydov II, Hutchinson L, Kam-Thong T, et al. An Introduction to Machine Learning. *Clinical Pharmacology & Therapeutics.* 2020;107:871–85.
19. Tang B-H, Guan Z, Allegaert K, Wu Y-E, Manolis E, Leroux S, et al. Drug Clearance in Neonates: A Combination of Population Pharmacokinetic Modelling and Machine Learning Approaches to Improve Individual Prediction. *Clin Pharmacokinet.* 2021;60:1435–48.
20. Hughes JH, Keizer RJ. A hybrid machine learning/pharmacokinetic approach outperforms maximum a posteriori Bayesian estimation by selectively flattening model priors. *CPT Pharmacometrics Syst Pharmacol.* 2021;10:1150–60.

## Acknowledgements:

The authors thank the patients whose data were used in this study.

## Funding section:

No funding was received for this study

## Conflict of interest:

The authors have no conflicts of interest to declare for this work

## Author contributions:

MB, JBW, designed the work, MB, JBW, ES, ML, PM, DWU, SGW analysis and interpret the data, MB, JBW, DWU, SGW wrote the draft.

## Figure legends

Figure 1: Relative Mean Prediction Error and Root Mean Square Error of the general model obtained in the training set using 10 fold cross-validation split by models used to simulate the respective population for the 3 a posteriori scenario. 1\_Occas = 1 occasion scenario (2 samples), 2\_Occas = 2 occasions scenario (4 samples) and General model fit (6 samples).

Figure 2: Contingency table of prediction and observation (truth) with  $AUC_{24-36} < 200$ , between 200 and 400 and  $> 400$  for the 3 a posteriori scenario (A= 1 occasion (2 samples), B = 2 occasions (4 samples) and C = general fit (6 samples)).

Figure 3: Variable importance plot for the prediction of  $AUC_{24-36}$  for the 3 a posteriori scenario (A= 1 occasion (2 samples), B = 2 occasions (4 samples) and C = general fit (6 samples)).  $C_X$  is the concentration measured at time X, MODN is the categorical covariate for simulated population (from 1 to 6), SCR is serum creatinine,  $C_{diff}$  is the measured difference between the peak and trough, TBW is the body weight.

Figure 4: Scatter plot of ML vs the model averaging algorithm (MAA) (A) and model selection algorithm (MSA) (B) AUC estimations and Bland Altman curve presenting the difference as function of the mean between ML and the MAA (C) or MSA (D) AUC estimations.

Table 1: Summary of simulations characteristics and real patient used for the validation

Variable	Simulations (n=6000)	Real patients (n=28)
Reference AUC <sub>24-36h</sub> (mg*h/L)	216.38 [167.12, 281.24]	NA
Body weight (Kg)	72.1 [66.6, 77.6]	71.5 [36.2, 120]
Body height (m)	1.70 [1.67, 1.74]	NA
Age (years)	57 [51, 64]	67 [32, 95]
Sex (%)		
Male	2910 (48.5)	11 (39.3)
Female	3090 (51.5)	17 (60.7)
Serum creatinine (µM)	81.59 [67.19, 99.51]	83.50 [24.00, 175.00]
Concentration at 1.25 h (mg/L)	18.12 [11.54, 26.11]	NA
Concentration at 11.75 h (mg/L)	6.80 [3.93, 9.95]	NA
Concentration at 13.25 h (mg/L)	24.47 [17.11, 33.34]	NA
Concentration at 23.75 h (mg/L)	9.95 [6.19, 14.40]	NA
Concentration at 25.25 h (mg/L)	27.54 [20.16, 36.85]	NA
Concentration at 35.75 h (mg/L)	12.03 [7.59, 17.33]	NA
Difference between concentration at 23.75h and 13.25h	-14.00 [-22.03, -6.68]	17.45 [6.70, 29.90]
Difference between concentration at 11.75h and 1.25h	-11.01 [-18.73, -4.48]	NA
Difference between concentration at 35.75h and 25.25h	-15.50 [-23.75, -7.90]	NA
Dose (mg)	1000[1000-1000]	1000.00 [400.00, 1500.00]
Number of dose received between begin of infusion and sampling	NA	6.00 [1.00, 11.00]
Concentration at trough (mg/L)	NA	11.45 [2.20, 26.10]
Concentration at peak (mg/L)	NA	30.30 [13.50, 52.00]
Difference between peak and trough	NA	17.45 [6.70, 29.90]

Median (Interquartile Range) for simulations and median (min-max) for real data, AUC is area under the curve; Reference AUC24-36h is the one obtained using trapezoidal rule on simulations

Table 2: Performances of the general algorithm developed in the overall population in the training set (10 fold cross-validation) and in the test set.

		A priori	1 Occasion	2 Occasion	General Model Fit
<b>Training</b>	RMSE $\pm$ SD (mg*h/L) *	76.0 $\pm$ 1.040	40.0 $\pm$ 0.438	37.6 $\pm$ 0.394	30.0 $\pm$ 0.368
	R <sup>2</sup> $\pm$ SD *	0.223 $\pm$ 0.0176	0.785 $\pm$ 0.0064	0.809 $\pm$ 0.005	0.879 $\pm$ 0.002
	Relative MPE (%)	11.0	3.2	2.7	1.7
	Relative RMSE (%)	39.6	19.1	17.6	13.8
<b>Testing</b>	RMSE (mg*h/L)	76.5	41.2	38.8	30.8
	R <sup>2</sup>	0.226	0.776	0.800	0.874
	Relative MPE (%)	9.18	3.3	2.8	1.3
	Relative RMSE (%)	37.1	18.9	17.4	13.7
<b>Model averaging [2]</b>	Relative MPE (%)	10.87	2.8	2.7	1.6
	Relative RMSE (%)	36.30	18.8	16.7	12.8

\* RMSE & standard deviation obtained after 10-fold cross-validation. MPE is Mean Prediction error and RMSE is root mean square error (i.e. imprecision)

Table 3: Comparison of the performances for each scenario and simulated population to the model averaging approach

Scenario	Sub-population	Testing		Model averaging [2]	
		Relative MPE (%)	Relative RMSE (%)	Relative MPE (%)	Relative RMSE (%)
A priori	Extremely obese	-2.14	29.6	-0.48	28.49
	Post heart surgery	-14.9	27.2	-13.86	26.95
	Trauma	-6.24	24.6	-6.73	24.73
	Critically ill	36.5	54.5	37.11	54.08
	Sepsis patients	26.6	44.1	30.48	49.46
	Hospitalised	18.4	35.4	18.72	34.10
1 Occasion	Extremely obese	2.03	12.3	0.52	14.17
	Post heart surgery	4.55	21.3	3.01	23.71
	Trauma	3.94	20.5	-4.48	17.36
	Critically ill	2.27	18.7	8.22	22.19
	Sepsis patients	5.12	21.2	3.87	19.77
	Hospitalised	2.34	14.6	5.86	15.44
2 Occasion	Extremely obese	1.44	10.6	-0.81	9.93
	Post heart surgery	5.52	20.9	4.82	23.02
	Trauma	3.09	18.6	-2.55	16.07
	Critically ill	2.27	18.7	7.88	20.76
	Sepsis patients	5.12	21.2	2.65	17.48
	Hospitalised	2.34	14.6	4.00	12.93
General Model Fit	Extremely obese	0.90	22.4	-2.08	7.71
	Post heart surgery	3.63	16.7	5.21	19.09
	Trauma	2.76	14.7	-1.72	12.22
	Critically ill	1.44	15.1	4.68	15.31
	Sepsis patients	2.90	14.8	1.49	12.97
	Hospitalised	1.63	10.5	2.29	9.43

MPE is Mean Prediction error and RMSE is root mean square error (i.e. imprecision)

Table 4: Mean of relative MPE and relative RMSE split by scenario for each algorithm trained in a given simulated population and estimated on the 5000 other patients.

<b>Population</b>	<b>rMPE</b>	<b>rRMSE</b>
Extremely obese_1_OCC	13.7	26.3
Post heart surgery_1_OCC	21.9	31.9
Trauma_1_OCC	16.9	29.3
Critically ill_1_OCC	- 5.2	25.0
Sepsis patients_1_OCC	5.1	21.0
Hospitalised_1_OCC	1.6	23.7
Extremely obese_2_OCC	15.8	25.6
Post heart surgery_2_OCC	17.2	33.0
Trauma_2_OCC	17.6	31.2
Critically ill_2_OCC	- 5.0	22.9
Sepsis patients_2_OCC	5.0	23.1
Hospitalised_2_OCC	2.9	22.8
Extremely obese_GF	7.5	20.3
Post heart surgery_GF	21.2	37.0
Trauma_GF	11.3	21.7
Critically ill_GF	-1.8	17.6
Sepsis patients_GF	3.2	17.2
Hospitalised_GF	3.7	17.8

rMPE is relative Mean Prediction error and rRMSE is relative Root Mean Square Error (i.e. imprecision), subpopulation\_1\_OCC is for 1 occasion scenario (2 samples), subpopulation\_2\_OCC is for 1 occasion scenario (4 samples) and GF is for General model fit (6 samples).