Machine learning for covariate imputation – application in a real-world scenario

Verena Schöning[1,#], Claudia Suenderhauf[2,#], Laura Hermann[1,3], Stephan Krähenbühl[4], Manuel Haschke[1] and Felix Hammann[1]

(1) Division of Clinical Pharmacology and Toxicology, Department of General Internal Medicine, Inselspital, Bern University Hospital, Switzerland, (2) Schwerpunkt Alterspsychiatrie, Psychiatrie Baselland, Liestal, Switzerland, (3) Medical Clinic, Zug Cantonal Hospital, Baar, Switzerland, (4) Clinical Pharmacology & Toxicology, University Hospital Basel, Switzerland
Corresponding author: felix.hammann@insel.ch, presenting author: verena.schoening@insel.ch #these authors equally contributed to this work and share first authorship.

Objective

Population pharmacokinetic (PopPK) models describe the changes in drug concentration across diverse patient populations, leveraging covariate effects using a non-linear mixed-effects (NLME) modeling approach. This aims to describe the variability of drug exposure within populations by considering patient-specific covariates, e.g., age, sex, weight, disease state, or organ function. However, **retrospective studies** in particular are susceptible to **missing covariate information**. While several approaches for data imputation in pharmacometrics exist, **machine learning (ML) paradigms are not yet widely used** but can offer **interesting additions to the analytical toolbox**. The major advantage of ML imputation methods is the leverage of all existing covariates to assume missing data values. For example, estimation of renal function depends on several parameters such as sex, age, weight, etc. Trained ML models can capture the relationships between the parameters and use them to impute missing renal function, instead of merely replacing the value with the mean renal function of the whole study population. Therefore, even if **some covariates are not part of the final PopPK dataset, they can improve the imputation of relevant covariates**. Two popular ML techniques for imputation are Random Forest (RF) and k-nearest neighbor (kNN). We decided to use retrospective data from three aminoglycosides (gentamicin, amikacin, and tobramycin) for which covariate information was missing for a considerable share of patients. We built and compared PopPK models of differently imputed datasets (RF, kNN) with a complete case dataset (CCD) and findings from published literature. Additionally, a sensitivity analysis was conducted to assess the robustness of the approach.

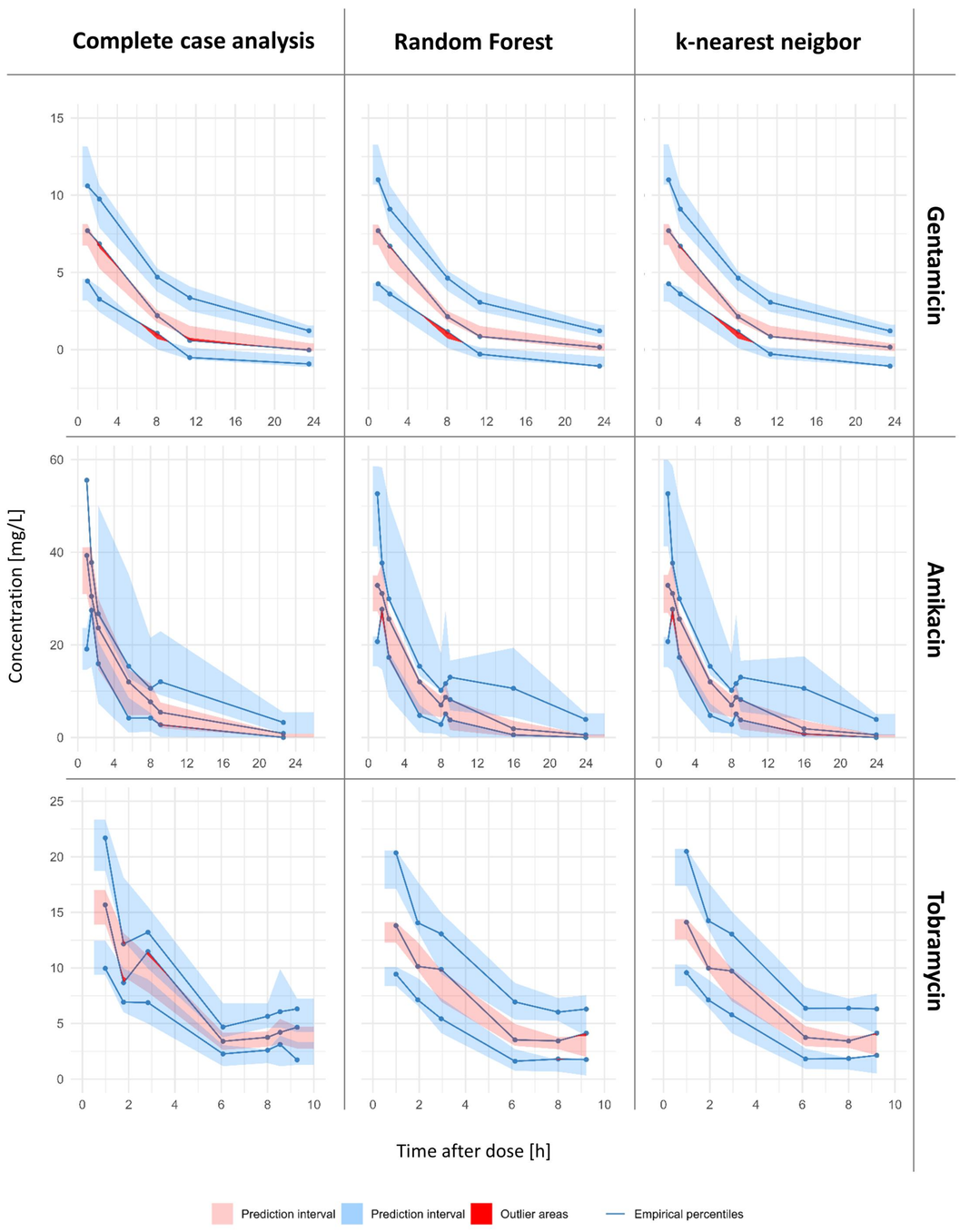

Figure 1: Comparative visual predictive checks (VPCs)

Methods

We conducted a retrospective, single-center study based on data collected at the University Hospital Basel from **routine therapeutic drug monitoring (TDM)** of **gentamicin, amikacin, and tobramycin** between January 1st 2014 to December 31st 2017. Data were collected from electronic health records (EHRs), and included demographic information, laboratory values (including plasma concentrations of the studied aminoglycosides), dosing and timing of administered aminoglycosides, diagnosis, and outcomes. **Missing covariate information was imputed** using two different ML algorithms: **Random Forest (RF)** and **k-nearest neighbor (kNN)**. Additionally, we created a CCD, where occasions or patients with missing covariate information were excluded (**Table 1**). We then used NLME modelling to build **PopPK models for each dataset** and validated the models with non-parametric bootstrap analyses. Furthermore, we conducted a sensitivity analysis to assess the robustness of the imputation by shifting the imputed values to 80% and 120% of the original values. Lastly, we compared the parameter estimates with previously published PopPK studies.

Results

We included 189 occasions with 300 plasma concentrations for gentamicin, 72 occasions with 132 plasma concentrations for amikacin, and 141 occasions with 280 concentrations for tobramycin in the analysis. Gentamicin had the highest share of most complete cases (82.7%), followed by amikacin (68.2%) and tobramycin (40.0%). After considering different structural models, we found that **one-compartment models** with **a linear elimination** described the data of all three aminoglycosides after intravenous administration best with **estimated GFR on clearance** and **weight on volume of distribution**. Overall, the point estimates from both imputation methods for gentamicin and amikacin are comparable and, in most cases, align closely with those from the CCD. For gentamicin, even though the point estimates align, the percentage relative standard errors (RSE%) of CL_{IOV} and V_{IIV} are notably lower in the models with imputed data compared to the CCD. In the amikacin models, the RSE% for V_{IIV} and CL_{IIV} are also lowest in the imputed models. For tobramycin, the V_{IOV} are decreased in the RF model. Furthermore, the V_{IIV} and CL_{IOV} are lower in both imputed models. Re-estimation of the PopPK models with the shifted imputation values resulted in minor changes in parameter estimates. However, no clinically meaningful differences were observed.

Laboratory value (n=428)	Normal range	Missing values (%)	CCD (median [IQR])	RF (median [IQR])	kNN (median [IQR])
Albumin (g/L)	35-52	166 (38.8%)	24.0 [20.0, 28.0]	23.0 [21.0, 26.1] *	23.28 [21.0, 27.0]
Albumin corrected calcium (mmol/L)	2.10-2.65	161 (37.6%)	2.52 [2.45, 2.6]	2.51 [2.5, 2.6]	2.52 [2.48, 2.57]
Creatine kinase (U/L)	0-170	166 (38.8%)	34.0 [19.0, 62.8]	42.4 [27.0, 70.6] **	39.0 [25.0, 9.4] *
Creatinine (µmol/L)	42-80	163 (38.1%)	79.0 [66.0, 106.0]	80.2 [68.6, 97.7]	76.6 [68.0, 96.3]
C-reactive protein (mg/L)	<10.0	166 (38.8%)	57.6 [15.9, 103.4]	71.0 [37.5, 110.6] *	68.5 [31.1, 106.7] *
Cystatin C (mg/L)	0.53-0.95	409 (95.6%)	2.33 [1.88, 2.33]	1.93 [1.87, 2.01]**	1.91 [1.85, 1.96]**
Leucocytes (x10S9/L)	3.5-10.0	165 (38.6%)	6.8 [4.3, 9.8]	7.6 [5.7, 9.4] *	7.6 [5.7, 9.4] *
Procalcitonin (ng/mL)	<0.5	320 (75.8%)	0.5 [0.2, 0.9]	0.6 [0.4, 1.1]**	0.6 [0.4, 1.2]**
Urea nitrogen (mmol/L)	1.8-7.1	150 (35.0%)	4.9 [3.3, 7.6]	5.5 [4.0, 6.9]	4.8 [3.6, 6.7]

Table 1: Summary information of laboratory values imputation.

CCD: complete case dataset, RF: random forest imputation dataset, kNN: k-nearest neighbor imputation dataset; significance was tested with a Wilcoxon rank sum test in comparison to CCD, *p<0.05, **p<0.001

Conclusions

The resulting parameter point estimates from RF and kNN showed no bias when compared with CCD while reducing random error, reflecting the **ML imputation techniques' propensity to generate supporting data points by picking up on the underlying distributions of covariates**.