**Propofol pharmacokinetic modelling –**

**Comparison of NONlinear Mixed Effects Modelling (NONMEM)**

**and Artificial Intelligence based Modelling**

Sascha Kreuer1, 2, Lukas M. Müller-Wirtz1, Christiane Dings2,3, Felix Maurer1, Tobias Hüppe1, Pia Zürn1, Steven L. Shafer4, Daniel I. Sessler5, Thorsten Lehr2,3, Thomas Volk1, Christian Schorr6

1. Department of Anaesthesiology, Intensive Care and Pain Therapy, Saarland University Medical Center and Saarland University Faculty of Medicine, 66482 Homburg / Saar, Germany
2. Saarmetrics GmbH, Saarland University, Saarbrücken, Germany
3. Clinical Pharmacy, Saarland University, Saarbrücken, Germany
4. Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University, Palo Alto, CA, USA
5. Department of Outcomes Research, Anesthesiology Institute, Cleveland Clinic, Cleveland, Ohio; USA
6. German Research Centre for Artificial Intelligence , Saarland Informatics Campus D 3\_2, Saarbrücken, Germany

The Department of Anaesthesiology, Intensive Care and Pain Therapy, Saarland University Medical Center and Saarland University Faculty of Medicine is part of the OUTCOMES RESEARCH Consortium, Cleveland, Ohio, USA.

**Corresponding Author:**

Prof. Dr. S. Kreuer

Department of Anaesthesiology, Intensive Care and Pain Therapy

Saarland University Medical Center and Saarland University Faculty of Medicine

66421 Homburg / Saar, Germany

Phone: +4968411622443

Email: sascha.kreuer@uks.eu

**Acknowledgements**

The authors would like to thank Brett Ezarsky and Matthew Connor (both C3.AI, Redwood City, CA, USA) for technical help.

**Funding**

The equipment for clinical measurements and determination of propofol blood concentrations was provided by B.Braun Melsungen (Melsungen, Germany).

**Details of author’s contributions**

Sascha Kreuer: NONMEM, Statistical analysis, Manuscript preparation

Lukas M. Müller-Wirtz: Random Forest modelling, Manuscript preparation

Christiane Dings: NONMEM modelling, Data Set Manuscript preparation

Felix Maurer: Measurement propofol blood concentration, Quality Management measurement blood concentration

Tobias Hüppe: Patient anaesthetize, Clinical measurement

Pia Zürn: Data management

Steven L. Shafer: Check of the NONMEM modelling, Preparation NONMEM parts of the manuscript

Daniel I. Sessler: Manuscript preparation

Thorsten Lehr: Manuscript preparation

Thomas Volk: Manuscript preparation

Christian Schorr: Data preprocessing, AI modelling, Preparation AI parts of the manuscript

**Keywords**

Propofol

Pharmacokinetic modelling

NONMEM

Artificial Intelligence (AI) / Machine learning (ML)

**Abstract**

**Background:** For propofol an intravenous anesthetic, many classical models based on compartments structure are available. A future and important approach for pharmacokinetic modeling is the use of artificial intelligence (AI) in the form of machine learning (ML). We therefore tested the hypothesis AI/ML modeling leads to better predictions than NONlinear Mixed Effects modelling (NONMEM).

**Methods:** Data of 48 adults having general anesthesia who had medical indications for an arterial catheter were included. Anesthesia was provided with propofol and remifentanil. Arterial blood was sampled every 30 min and propofol plasma concentrations were measured by liquid chromatography. All data were recorded digitally. Propofol pharmacokinetic modeling was performed using NONMEM (ICON Development Solutions, Ellicott City, MD, USA) and Pirana Modeling Workbench (Certara Overlook Center, Princeton, NJ, USA) and C3.AI Ex Machina (C3.AI, Redwood City, CA, USA) for random forest (RF) AI/ML modelling. Different statistical approaches were used to compare the goodness of fit of both models.

**Results:** The propofol concentration of 450 blood samples were measured, in each patient we take 9 ± 3 (5-19) samples. The propofol plasma concentration was between 0.2 and 8.0 µg/ml. A 3-compartment model was not significantly better than a 2 compartment model (NONMEM). Of demographic data, only weight on clearance improves the model. For RF the combination of two factors improves the predictions: weight / height are equivalent to height / age. The R² values for the line of identity between measured and calculated propofol plasma concentrations were R²=0.67 (NONMEM) and 0.93 (RF). The overall MDAPE were 18.1 % vs. 6.4 %; pseudo R² McFadden were 0.19 vs. 0.27.

**Conclusion:** In our patients, the prediction of measured propofol plasma concentration was better with RF AI/ML modelling than with a classical NONMEM model. Therefore, RF AI/ML modeling is a promising approach for propofol pharmacokinetic modelling. However, further steps are necessary for clinical application.

**Introduction**

In recent decades, numerous pharmacokinetic models have been developed to accurately dose propofol administration during anaesthesia [Struys 2016]. Complex pharmacological relationships for propofol were introduced in the target-controlled infusion (TCI) system in 1996. By programming syringe pumps with various pharmacokinetic data, a desired plasma target concentration could be maintained in specific patients. Because of concerns about inter-individual pharmacokinetic variability, TCI systems have yet to be approved by the FDA in the United States.

Pharmacokinetic models developed by Marsh [MARSH] and Schnider [Schnider] have been integrated in commercial propofol TCI syringe pumps. Recently, Eleveld used a large diverse population to develop a pharmacokinetic/pharmacodynamic model for predicting propofol concentrations and the EEG based bispectral index (BIS) during TCI anesthesia and sedation [Eleveld]. Specifically, Eleveld used data from 1,033 patients in the Open TCI Initiative which started in 2008 and combines data from many investigators. All 3 models were calculated with NONlinear Mixed Effects Modelling (NONMEM) including 3 compartments. The differences based in the number and selection of cofactors. However, the median absolute performance error (MDAPE) as a measure of accuracy of the 3 models is comparable [Hüppe].

A new approach to pharmacokinetic modeling is the use of artificial intelligence (AI) in the form of so-called machine learning (ML) [McComb2021]. ML is a field of research with the goal of developing predictive algorithms able to learn and improve through the use of data. Training data is used to build a model in order to make predictions or decisions without being explicitly programmed how to do so.

In medicine, ML is successfully used in detecting lung cancer in CT scans [Ardila 2019] or skin cancer from photographs using deep convolutional neural networks for example. It is also used to assess the risk of sudden cardiac death or other heart diseases based on **electrocardiograms [Siontis 2021].**

The general impact of introducing AI to clinical pharmacology is discussed by Gambus et al. [Gambus 2018]. They also emphasize that the current standard of manual pharmacometric modelling will be replaced by AI approaches in the foreseeable future. A broad summary of the use of AI in anesthesiology ML and ensuing ethical questions is given in [Hashimoto 2021]. Another high-level overview on the subject can be found in [Connor 2019, Sassenscheid 2020]

For our Propofol pharmacokinetic modelling, we use the random forest (RF) algorithm, which is suitable for both for classification and regression tasks. It needs less training data than deep learning with neural networks while still providing good results. It also has only a few hyper-parameters to fine-tune which allows for quick optimization.

We evaluated propofol administration and plasma concentrations in 48 patients. Specifically, we tested the hypothesis that a RF pharmacokinetic model provides better predictions (higher pseudo R² McFadden and smaller median MDAPE and smaller median performance error, MDPE) of arterial propofol plasma concentration than a classical NONMEM compartment model.

**Methods**

The study was approved by the regional ethics committee (Ärztekammer des Saarlandes, Saarbrücken, Germany: Ref: 39/16) and written informed consent was obtained from participating patients at least a day before surgery. Data of 50 adults having general anaesthesia who had medical indications for an arterial catheter, were American Society of Anesthesiologists (ASA) physical status II-III, had a body mass index (BMI) less than 35 kg/m², and an expected duration of surgery >1 hour were included. Exclusion criteria were contraindications for propofol or remifentanil, cardiac or lung surgery, lung disease, pregnancy or breastfeeding, renal replacement therapy, known HIV or hepatitis C infections.

Anesthesia was provided with propofol target controlled Infusion (TCI) using the Marsh Model [Marsh] and remifentanil TCI Minto model [Minto], both in plasma mode (Perfusor Space TCI, BBraun Melsungen). Target concentrations of each drug were selected to optimize clinical conditions by anesthesiologists who were not involved in the study. Dosages were recorded digitally.

**Blood samples**

2-ml of arterial blood was sampled every 30 min, and 5-10 min after changing the propofol TCI target concentration. Collection times were documented with digital time stamps. Samples were prepared by solid-phase extraction. For each patient, a calibration curve with nine blank plasma samples (lyophilized drug-free serum, Bio-Rad, Munich, Germany) spiked with propofol was prepared with a certified reference standard (Sigma-Aldrich, Steinheim, Germany). Separation was carried out on an Agilent 1260 Infinity series LC system. Detection was performed on an atmospheric pressure ionization-electrospray coupled mass selective detector model G6130BA (Agilent, Waldbronn, Germany). All calibrators, the patient samples as well as three quality control standards with a known concentration of 1, 4 and 6 µg/ml propofol were measured as triplicates [Maurer 2018].

**Dataset**

The same data set was created for both modeling procedures and contained data of 48 patients. For the NONMEM analysis the data set was extended by the value “MDV” which indicated missing measured propofol plasma concentrations. The data set contained the following parameters:

* Independent variables: time [min], propofol infusion rates [mg/min]
* Covariate for testing: weight [kg], height [cm], age [years], sex [1 or 0]
* Dependent variable: measured propofol plasma concentrations [µg/ml]

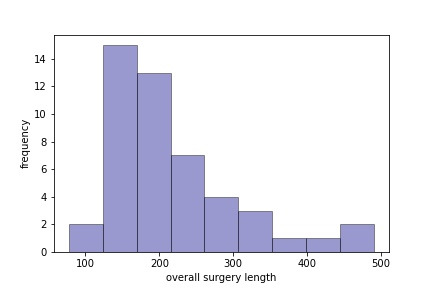


Fig x: Distribution of surgery durations

Since the propofol infusion rate is predetermined by the Marsh model integrated into the Perfusor Space TCI all administered infusions are independent of each other. Thus the standard RF algorithm and neural networks can be used for modelling since the data does not actually represent a time series with interdependent variables.

**Additional data set preparation for AI modeling**

A main problem for AI modeling with RF and NN is the requirement of a fixed input dimension. Since the surgeries are of different durations, we need to preprocess the data prior to model training. For each of the 48 patients, the corresponding data set was therefore split into separate blocks between blood concentration measurements. This leads to a total of 449 data entries of different length. To preserve the information of the ordering of data blocks belonging to the same surgery we added a parameter previous Propofol concentration [µg/ml], which starts with a value of 0 for the first data block and uses the measured Propofol concentration of the previous data block for all other data blocks. This splitting also has the benefit of virtually increasing the amount of data for model training. Especially NN require large amounts of training data therefore 449 data points of short but uniform dimension are better suited to AI modelling than 48 data points of higher but varying dimensions.

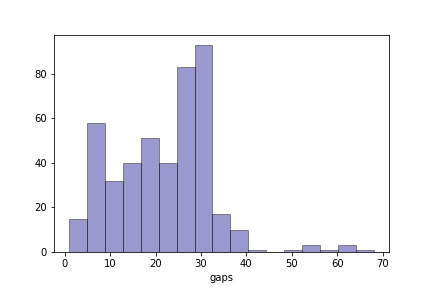


Figure x: Distribution of gap durations

We evaluated two approaches to solve the problem of varying gaps lengths: Propofol infusion summation and rescaling with interpolation.

**Propofol infusion summation**

We sum over all propofol infusion rates [mg/min] and use the result as a feature [mg]. In order to include the time spent between two consecutive blood samples, we also introduce an additional feature called gap duration [min]. Although this approach simplifies the model and loses most of the temporal resolution it allows to include all data in the model by circumventing the varying gap durations.

**Rescaling with interpolation**

Analysis of figure x (distribution of gap durations) yields a maximum at 30 min with most of the values lying in an interval between 30 -t and 30 + t min. In order to keep as much of the temporal information as possible we introduce a tolerance parameter t and a target gap duration d. All data blocks shorter than *d - t* min and longer than *d + t* min are discarded. The remaining data blocks are then rescaled to a virtual gap dimension of t using linear interpolation:

Add formula or pseudo-code

We chose t = 5 and d = 30 to make a good compromise between preprocessing distortion and data availability. Fig. x shows the data blocks between 25 and 35 min gap duration resulting from this choice of parameters.

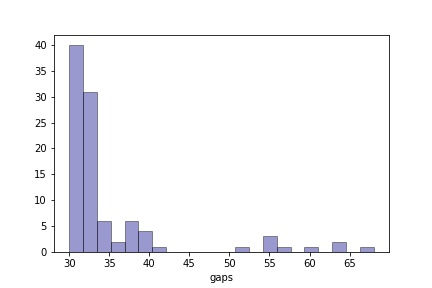


Figure x: Distribution of gap durations after applying tolerance filter (update required)

After applying rescaling all data blocks are of equal dimension and ready to be used in the AI training step.

**Transformed Data sets**

**Data set 1: Propofol infusion summation**

Features:

* weight [kg]
* height [cm]
* age [years]
* sex [1 or 0]
* administered propofol [mg]
* time elapsed since surgery start [min]
* previously measured propofol plasma concentration [µg/ml]

Target:

* measured propofol plasma concentration [µg/ml]

**Data set 2: Rescaling with interpolation**

Features:

* weight [kg]
* height [cm]
* age [years]
* sex [1 or 0]
* rescaled propofol infusion rates for minutes 1 - 30 [mg/min]
* tolerance t [unitless]
* time elapsed since surgery start [min]
* previously measured propofol plasma concentration [µg/ml]

Target:

* measured propofol plasma concentration [µg/ml]

**Explainabiliy of AI modelling**

AI predictions are often black boxes, that do not explain, how their prediction where arrived at. This is a crucial point for the acceptance of their use, especially in clinical practice. Therefore we include three methods to increase the interpretability of the AI models used. The first is permutation-based feature importance [PFI], which XXX. A similar approach is used in feature importance computed with SHAP values [SHAP]

A more advanced method is LIME (Local interpretable model-agnostic explanations) [LIME].

**Non-linear mixed effects modeling**

Propofol pharmacokinetic modeling was performed using the NONMEM® software package (version VI 2.0; ICON Development Solutions, Ellicott City, MD, USA) and Pirana Modeling Workbench (Certara Overlook Center, Suite 101, Princeton, NJ, USA). Model selection was based on changes in the NONMEM objective function (OFV) -2 log likelihood. One model was declared superior to the other model when the objective function value was reduced by 3.84 (P<0.05). First different numbers of compartments were tested without demographic data. In a second step the influence of demographic data were analyzed. Population prediction (PRED) and individual prediction (IPRED) were calculated. IPRED was not further evaluated.

**Random forest (RF)**

The C3.AI Ex Machina (C3.AI, 1300 Seaport Boulevard, Redwood City, CA, USA) implementation was used for the actual RF modelling. The software is a closed, code free drag and drop interface for AI modeling. The R² value of linear regression between observed and predicted values is used as objective function. In each calculation, 33% of the observed values are re-randomized and excluded from the model development. Model selection was not based on the R² value of linear regression. Instead, the pseudo R² McFadden has been used.

The patient ID is not a parameter for the model calculation. Three different model parameters can be set:

* Maximum tree depth: represents the depth of each tree in the forest
* Maximum number of bins: is used for discretizing continuous features and for choosing how to split on features at each node.
* Number of trees: total number of decision trees that are created

First different parameter values were tested without demographic data. In a second step the influence of demographic data was analyzed.

**Time courses RF**

Normally, the C3.AI surface can only calculate values at the time when an observed value is available. Therefore, in order to predict the propofol concentrations minute by minute, a simulation was developed within the C3.AI interface. This was implemented by creating different data sets from the original one. In the process, the measurement times were shifted and the values were replaced by placeholders. The calculation was then performed with the original RF model.

**Cross validation NONMEM and RF**

A cross validation was performed with both final models (NONMEM and RF). For this purpose 9 times 5 patients and once 3 patients were deleted from the data set and the final models were recalculated with the final parameter settings. Thereafter, the excluded patients were tested on the appropriate model. The data was then copied to a new output file.

**Statistics**

All NONMEM and RF output files were analyzed in the same Excel sheet (Excel 2016, Microsoft Corporation, One Microsoft Way, Redmond, WA, USA). The following parameters were calculated:

* Pseudo R² McFadden: the value is defined as 1-(log likelihood of the fitted model / log likelihood of the 0 model) [McFadden].Here, the 0 model is calculated by setting all predicted concentrations to the mean value of the measured concentrations. McFadden's pseudo R2 ranging from 0.2 to 0.4 indicates very good model fit [Mcfadden 2].
* MDPE: median performance error; Median of the percentage difference between the predicted and the measured propofol plasma concentrations
* MDAPE: Median absolute performance error; Median of the absolute percentage difference between the predicted and the measured propofol plasma concentrations
* R² for the goodness of fit to the line of identity: linear regression analysis with steepness a=1 and axis section Y0=0 between predicted and measured propofol plasma concentrations
* Likelihood measures the goodness of data for observed data of predicted values. The value as such has no informative value. It rather serves the comparison of models with the same data set, the higher the value the better the model describes the measured values.
* Log likelihood is the natural logarithm of the likelihood. The closer the value is to zero, the better the model describes the measured values

MDPE and MDAPE were calculated for each individual patient for population NONMEM and RF predictions. Best, median, and worst fit of individual patients were determined using the MDAPE.

Graphics and Bland & Altman plots were generated using SigmaPlot (version 12.5; Systat Software, Inc., Richmond, CA, USA).

**Results**

**Patients and measured propofol concentrations**

Two of the 50 patients had to be excluded due to technical problems with data recording. The final analysis therefore included 14 women and 34 men who were 70 ± 11 yr old, weighted 78 ± 14 kg, and were 171 ± 10 cm tall. The mean time per case was 3.5 ± 1.5 hours, for a total of 169 hours.

The propofol concentration of 450 blood samples were measured, in each patient we take 9 ± 3 (5-19) samples. The propofol plasma concentration was between 0.2 and 8.0 µg/ml.

**NONMEM**

A 3-compartment model (NONMEM OFV = -67.2) was not significantly better than a 2 compartment model (OFV = -65.9). Of demographic data, only weight on clearance improves the model; 3-compartments OFV = -88.0 vs 2 compartment OFV = -86.6. This was also confirmed in the other statistic tests (Table 1).

**Random Forest**

There is a correlation between the Pseudo R² McFadden values and the parameters of the RF model. The value is between 0.17 and 0.25 without demographic data (Table 2 a). The inclusion of the demographic data leads to an increase of the value to 0.27. At least 2 factors are necessary for this. Weight and height are equivalent to height and age. As the final model, number 5 was selected with height and weight. The systematic study of the influence of demographic data on model 5 is shown in Table 2 b. Model 9 with all 4 investigated factors is used as a reference, here no higher pseudo R² McFadden values is shown compared to the final model.

**Neural networks**

XXX

**Comparison**

The goodness of fit plots of both final models are shown in Fig. 1. The R² values for the line of identity between measured and calculated propofol plasma concentrations were R²=0.67 (NONMEM) and 0.93 (RF) (Fig. 1a). The overall MDAPE were 18.1 % vs. 6.4 %; Pseudo R² McFadden were 0.19 vs. 0.27 (Table 1 and 2b).

The Bland & Altman plots indicates lower values of BIAS and limits of agreement for the RF model (Left part of Fig. 2b) compared to the NONMEM model (Left part of Fig. 2a). However, plot shows a systematic error: low concentrations are overestimated and high concentrations are underestimated.

The results of the cross validation are different. With the NONMEM model there is a small deterioration of the results likelihood 113.4 vs, 112.5 and pseudo R² McFadden 0.18 vs. 1.9 (Table 1, right part of Fig. 2a). The RF model, on the other hand, shows clear difference likelihood 125.5 vs. 116.9 pseudo R² McFadden 0.22 vs. 0.27 (Table 2b, right part of Fig. 2b). Never the less, the results of the random forest model are still better than with NONMEM.

**Patient’s individual time courses**

Minute-by-minute courses of predicted concentrations using both final models and measured concentrations at respective time points from 6 patients are shown in Fig. 3. Patients were selected based on lowest, median, and highest MDAPE for NONMEM (left side of Fig 3) and RF (right side of Fig 3).

**Discussion**

In our patients, the prediction of actual measured propofol plasma concentration was better with RF AI/ML modelling than a classical NONMEM compartment model as indicated by the different used statistical approaches.

**Statistics**

The C3.AI software optimizes the model using linear regression. Our data set consists of repeated measurements within the same subjects, therefore linear regression is not permitted [Schober 2021]. Nevertheless, we show the values for the different RF and NONMEM models. Since the value is common and can therefore be generally classified. The pseudo McFadden R² value was used for the model comparison because it is appropriate for time series. However, it can only be used to compare models with the same data set. Therefore, both methods are not suitable for selecting the best, median, and worst individual patient courses. Thus, we used the MDAPE for this purpose.

**NONMEM**

For decades nonlinear mixed effects modelling with NONMEM is the gold standard for population pharmacokinetic and dynamic modeling. Thereby patterns over time are described non-linearly by using ordinary differential equations. Mixed-effects mean the combination of fixed and random effects. Fixed effects are the structural model describes the dependent variable at the observed time points of the independent variable (time) as a function of input (dose), pharmacometric parameters (e.g. volumes, clearance) and covariates (e.g. demographic data). Random effects describe the stochastical part of the model including inter-, intraindividual, and residual variability.

**NONMEM propofol models**

Eleveld was able to incorporate more than 15,000 propofol concentrations from more than 1,000 patients from 30 studies, making the results both robust and generalizable [Eleveld]. Eleveld's model includes covariates such as age, weight, height, and sex to predict arterial propofol concentrations. Marsh's model includes only weight as a covariate of volume of distribution [Marsh]. Schnider's model uses a fixed central volume of distribution (4.3 L) [Schnider]. In addition, it reliably calculates lean body mass based on age, weight, height and gender.

In contrast, in our patients, 2 compartments are as good as 3 compartments. This is due to the collection times of the blood samples. We did not collect samples immediately after propofol bolus administration. Therefore, the central compartment transitions to the fast peripheral compartment [XXX]. The analysis of demographic data shows that only weight on clearance improves the model significantly. Could be due to low variability in demographic data. However, the MDAPE is better than with the Marsh (25%), Schnider (26 %) and Eleveld (22 %) model [Hüppe BJA].

**Random forest**

A very simple and easy to understand model is a decision tree [Breiman 1984]. It resembles a flowchart where each node constitutes a conditional statement on one of the input variables. The possible results of these statements represent branches leasing either to other nodes with a further conditional statements or to a leaf with a class label. So if a decision tree is traversed from the top along its branches, it computes the classification of the input data. A single decision tree is not a very effective predictor since its performance depends on the formulation of the conditional statements. RF is an ensemble method consisting of a large number of decision trees [Breiman 2001]. Each tree is built randomly using only a part of the available input variables, so that the predictions between each tree are different. Given a sufficiently amount of decision tress - a forest - the averaged prediction of all trees is vastly superior to using only a single decision tree. RF can be used both for classification and regression tasks. Due to the comparatively small amount of hyperparameters and its low training data requirements, RF was chosen as a preliminary modelling algorithm in favor of more complex and data hungry methods like neural networks.

RF tends to overfit on training data, slightly worse results on validation data (excluded from training) are to be expected as a result of smaller over fitting and better generalization. The training set for each tree consist of two-thirds randomized of the original data. In addition, we performed a x-fold cross validation to gain a more robust and general model.

The initial propofol concentration peak after bolus administration is not reflected by the RF model, because we did not measure the concentration at this early time. In principle, the RF model tends to underestimate high and overestimate low concentrations especially at the beginning and the end of propofol dosing which is actually a sign that it generalizes well and does not overfit to the training data.

One disadvantage of the C3.AI software used, is that AI models cannot be exported and therefore not be used outside of it.

**Neural network**

**XXX**

**NOMEM vs. Random forest vs. Neural network**

The pseudo R² McFadden values and MDAPE are better for using RF than for NONMEM modelling. For cross validation, the difference from the final model is smaller for NONMEM than for RF. However, even here the statistical results are still better with RF. Presumably changes in concentration where no measurement was performed are better detected by NONMEM. With RF, there are sometimes erratic changes in predicted concentrations.

**Comparison with the literature**

Regarding propofol and remifentanil modelling with AI, there have been several studies so far. [Poynton 2009] use both a non-linear mixed effects model as well assupport vector machines and neural nets to model remifentanil in healthy person. …

The pharmacodynamics interaction of Propofol and Reminfentanil indicated by bispectral index was modelled with long short-term memory neural networks [Lee 2018]. Their results state that their AI-based approach is more accurate than the traditional model. …

[Schamberg 2020] use a reinforcement learning approach to predict the medication regimen of propofol during surgery. They train a deep neural network to map an observed anesthetic state to a probability of infusing a fixed propofol dosage. A deterministic policy then transforms the probability of infusion to a continuous infusion rate. The algorithm is trained and tested on simulated data only, generated using standard pharmacokinetic models. The conclusion drawn from their numerical experiments is that their approach performs much better than a standard PID controller, but has to be evaluated on real patient data, too.

Inegrande et el. studied the prediction of the early distribution kinetics of Propofol in morbidly obese and lean subjects comparing a machine learning (ML) incorporated ensemble learning [Ingrande 2020]. Both of the later models were found to be superior to the compartmental model. A limiting factor for wide-spread use could be the small amount of training data available to train the neural network, though.

Differences to our approach / our data?

Most AI algorithms are black boxes, meaning that input and predictions are known, but not the how the algorithm comes to a certain prediction. The actual model turning input into predictions is therefore unknown, which is a crucial issue when using AI as a basis for life and death decisions. Compared to neural networks, the RF algorithm we use is more explainable. It consists of a number of decision trees with different node splits which then vote on the prediction. Each tree can be interpreted easily by a human, but the entire forest is more difficult due to the large number of individual decision trees. Using methods to compute the relative importance a RF model assigns to each input variable allows for some insight into the models behaviour, rendering it more transparent and thus trustworthy to the clinician.

**Further future development steps**

**Technical:** For a possible future use of pharmacometric AI models, it is necessary to implement the models in an open format, like Python, that allows a standalone use of the resulting model outside a closed proprietary software. Instead of using R² as the objective function another test like likelihood or log likelihood could be investigated. Switching from Random Forest to neural network based approaches, for example using RNNs, is another step which could further improve the predicted model, but requiring

**Propofol model:** A larger data set with a wider range of demographics is needed to develop and test AI propofol models. Additionally, components of population and physiologically based models could be integrated into the AI model as additional features.

**Ethical:** Predictions from AI models have to be transparent and easily interpretable by humans. At the moment most algorithms - and especially neural networks - are essentially black boxes, which do not explain how they come from input data to actual prediction. In medicine this is a very critical point. Future work needs to include interpretation strategies like local surrogate methods [] to make the modelling understandable for physicians.

**Conclusion**

AI modeling using a random forest model is a promising approach to predict propofol plasma concentrations.

**References**

Struys MM, De Smet T, Glen JI, Vereecke HE, Absalom AR, Schnider TW. The history of target-controlled infusion. Anesth Analg 2016; 122: 56–69

Marsh B, White M, Morton N, Kenny GN. Pharmacokinetic model driven infusion of propofol in children. Br J Anaesth 1991; 67: 41–48.

Schnider TW, Minto CF, Gambus PL, et al. The influence of method of administration and covariates on the pharmacokinetics of propofol in adult volunteers. Anesthesiology 1998; 88: 1170–1182.

Eleveld DJ, Colin P, Absalom AR, Struys MMRF. Pharmacokinetic–pharmacodynamic model for propofol for broad application in anaesthesia and sedation. Br J Anaesth 2018; 120: 942-959.

McComb M, Bies R, Ramanathan M. Machine learning in pharmacometrics: Opportunities and challenges. Br J Clin Pharmacol. 2021;1–18.

Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001). <https://doi.org/10.1023>

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1984). Classification And Regression Trees (1st ed.). Routledge. https://doi.org/10.1201/9781315139470

Cummins, M R, Choi, B, Kim, Y S, Park, I S. Machine Learning Methods Applied to Pharmacokinetic Modelling of Remifentanil in Healthy Volunteers: A Multi-Method Comparison, December 2009, The Journal of international medical research 37(6):1680-91, DOI:10.1177/147323000903700603

Sibieude, E, Khandelwal, A, Hesthaven, J S, Girard, P, Terranova, N. Fast screening of covariates in population models empowered by machine learning, Journal of Pharmacokinetics and Pharmacodynamics (2021) 48:597–609, <https://doi.org/10.1007/s10928-021-09757-w>

Sassenscheidt J, Jungwirth B, Kubitz JC. Machine learning in anesthesiology. Anaesthesist. 2020;69:535–43.

Poynton MR, Choi BM, Kim YM, Park IS, Noh GJ, Hong SO, et al. Machine learning methods applied to pharmacokinetic modelling of remifentanil in healthy volunteers: A multi-method comparison. J Int Med Res. 2009;37:1680–91.

Lee H-C, Ryu H-G, Chung E-J, Jung C-W. Prediction of Bispectral Index during Target-controlled Infusion of Propofol and Remifentanil. Anesthesiology. 2018;128:492–501.

Gambus P, Shafer SL. Artificial intelligence for everyone. Anesthesiology. 2018;128:431–3.

Connor CW. Artificial Intelligence and Machine Learning in Anesthesiology. Anesthesiology. 2019;131:1346–59.

Ardila, D., Kiraly, A.P., Bharadwaj, S. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 25, 954–961 (2019). <https://doi.org/10.1038/s41591-019-0447-x>

Ingrande, Jerry; Gabriel, Rodney A.; McAuley, Julian; Krasinska, Karolina; Chien, Allis; Lemmens, Hendrikus J. M. The Performance of an Artificial Neural Network Model in Predicting the Early Distribution Kinetics of Propofol in Morbidly Obese and Lean Subjects, Anesthesia & Analgesia: November 2020 - Volume 131 - Issue 5 - p 1500-1509, doi: 10.1213/ANE.0000000000004897

Mahdi Mahfouf, Catarina S. Nunes, Derek A. Linkens, John E. Peacock, Modelling and multivariable control in anaesthesia using neural-fuzzy paradigms: Part II. Closed-loop control of simultaneous administration of propofol and remifentanil, Artificial Intelligence in Medicine, Volume 35, Issue 3, 2005, Pages 207-213, ISSN 0933-3657, <https://doi.org/10.1016/j.artmed.2004.12.005>.

Lowery, C., Faisal, A.A.: Towards efficient, personalized anesthesia using continuous reinforcement learning for propofol infusion control. In: 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 1414–1417. IEEE (2013)

[Gabe Schamberg](https://arxiv.org/search/cs?searchtype=author&query=Schamberg%2C+G), [Marcus Badgeley](https://arxiv.org/search/cs?searchtype=author&query=Badgeley%2C+M), [Emery N. Brown](https://arxiv.org/search/cs?searchtype=author&query=Brown%2C+E+N) . Controlling Level of Unconsciousness by Titrating Propofol with Deep Reinforcement Learning, International Conference on Artificial Intelligence in Medicine 2020

Siontis, K.C., Noseworthy, P.A., Attia, Z.I. *et al.* Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat Rev Cardiol* **18,** 465–478 (2021). <https://doi.org/10.1038/s41569-020-00503-2>

Schober P, Vetter T: Linear Regression in Medical Research. Anesth Analg . 2021 Jan;132(1):108-109