ORIGINAL CLINICAL RESEARCH REPORT

The Performance of an Artificial Neural Network Model in Predicting the Early Distribution Kinetics of Propofol in Morbidly Obese and Lean Subjects

Jerry Ingrande, MD, MS,* Rodney A. Gabriel, MD, MAS,*† Julian McAuley, PhD,‡ Karolina Krasinska, MSc,§ Allis Chien, PhD,§ and Hendrikus J. M. Lemmens, MD, PhD‖

BACKGROUND: Induction of anesthesia is a phase characterized by rapid changes in both drug concentration and drug effect. Conventional mammillary compartmental models are limited in their ability to accurately describe the early drug distribution kinetics. Recirculatory models have been used to account for intravascular mixing after drug administration. However, these models themselves may be prone to misspecification. Artificial neural networks offer an advantage in that they are flexible and not limited to a specific structure and, therefore, may be superior in modeling complex nonlinear systems. They have been used successfully in the past to model steady-state or near steady-state kinetics, but never have they been used to model induction-phase kinetics using a high-resolution pharmacokinetic dataset. This study is the first to use an artificial neural network to model early- and late-phase kinetics of a drug.

METHODS: Twenty morbidly obese and 10 lean subjects were each administered propofol for induction of anesthesia at a rate of 100 mg/kg/h based on lean body weight and total body weight for obese and lean subjects, respectively. High-resolution plasma samples were collected during the induction phase of anesthesia, with the last sample taken at 16 hours after propofol administration for a total of 47 samples per subject. Traditional mammillary compartment models, recirculatory models, and a gated recurrent unit neural network were constructed to model the propofol pharmacokinetics. Model performance was compared.

RESULTS: A 4-compartment model, a recirculatory model, and a gated recurrent unit neural network were assessed. The final recirculatory model (mean prediction error: 0.348; mean square error: 23.92) and gated recurrent unit neural network that incorporated ensemble learning (mean prediction error: 0.161; mean square error: 20.83) had similar performance. Each of these models overpredicted propofol concentrations during the induction and elimination phases. Both models had superior performance compared to the 4-compartment model (mean prediction error: 0.108; mean square error: 31.61), which suffered from overprediction bias during the first 5 minutes followed by under-prediction bias after 5 minutes.

CONCLUSIONS: A recirculatory model and gated recurrent unit artificial neural network that incorporated ensemble learning both had similar performance and were both superior to a compartmental model in describing our high-resolution pharmacokinetic data of propofol. The potential of neural networks in pharmacokinetic modeling is encouraging but may be limited by the amount of training data available for these models. (Anesth Analg XXX;XXX:00–00)

KEY POINTS

- **Question:** Can artificial neural networks accurately predict the early induction kinetics of propofol?
- **Findings:** A recirculatory model and artificial neural network had similar performance in describing the early- and late-phase kinetics of propofol.
- **Meaning:** The performance of artificial neural networks may be limited by the amount of training data available.

From the *Department of Anesthesiology and †Division of Biomedical Informatics, Department of Medicine, University of California, San Diego School of Medicine, La Jolla, California; ‡Department of Computer Science and Engineering, University of California, San Diego, La Jolla, California; §Stanford University Mass Spectrometry Laboratory, Stanford, California; and ‖Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Stanford, California.

Accepted for publication April 17, 2020.

Funding: Departmental (Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine).

The authors declare no conflicts of interest.

DOI: 10.1213/ANE.0000000000004897

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (www.anesthesia-analgesia.org).

Clinical trials registry number: NCT01591148.

Summary statement: A gated recurrent unit neural network had similar performance to a recirculatory model, and both had superior performance to a compartmental model in predicting the early distribution kinetics of propofol.

Reprints will not be available from the authors.

Address correspondence to Jerry Ingrande, MD, MS, Department of Anesthesiology, University of California, San Diego School of Medicine, 402 Dickinson St, MPF Bldg, Mail Code 0801, Hillcrest, CA 92103. Address e-mail to jingrande@ucsd.edu.

GLOSSARY

ANN = artificial neural network; **ASA** = American Society of Anesthesiologists; **BMI** = body mass index; **ENT** = ear, nose, and throat; **GC-MS** = gas chromatography-mass spectroscopy; **GRU** = gated recurrent unit; **ID** = internal diameter; **IRB** = institutional review board; **IV** = intravenous; **LBW** = lean body weight; **LL** = log-likelihood; **LSTM** = long short-term memory; **MEM** = mixed-effects model; **MPE** = mean prediction error; **MSE** = mean square prediction error; **OBJ** = objective function; **PD** = pharmacodynamic; **PK** = pharmacokinetic; **SRM** = selected reaction monitoring mode; **TBW** = total body weight

The limitations of traditional compartmental models to accurately describe early distribution kinetics have been well described.[1] These models assume complete, instantaneous intravascular mixing and a steady decline in drug concentration thereafter.[2] Physiologically, we understand that drug is distributed to a system of organs and tissues, and that this distribution is governed by blood flow to these tissues and the relative affinity of these tissues to drug. Compartmental models ignore this fact.

In addition to the structural constraints of a compartmental model, further model misspecification may be introduced when models are built on sparsely sampled datasets.[3] Infrequent blood sampling after drug administration—together with the failed assumption of instantaneous mixing—leads to overestimations of central volume, ultimately leading to drug overdose and supratherapeutic effect.[4]

It stands to reason that physiologically based models, or recirculatory models, would perform better than compartmental models. However, Masui et al[5] found that, for both bolus and short infusions of propofol, performance of compartmental models and a physiologically based recirculatory model all overestimated propofol plasma concentrations.

Even the most robust physiologically based pharmacokinetic (PK) models and recirculatory models capable of modeling early distribution kinetics run the risk of structural misidentification. The available data and the information embedded in it can be insufficient to estimate the parameters in such a model.[6,7] Artificial neural networks (ANNs) offer the advantage that they are devoid of such constraints.[8] They are not confined to a specific structural model and, therefore, are not as prone to model misspecification.[9]

These systems have been used in the past for PK modeling.[9–11] However, these studies used neural networks to model steady-state or near steady-state conditions using sparsely sampled datasets. The objective of this study is to compare the performance of a compartmental model, recirculatory model, and an ANN to describe propofol PK from a frequently sampled dataset. We hypothesize that the ANN will have better performance because of its ability to model complex nonlinear systems without assuming a particular structure.

METHODS

Subject Selection and Informed Consent

This study was approved and regulated by the Institutional Review Board (IRB) at Stanford University (IRB no. 16509) and was registered in ClinicalTrials.gov (NCT01591148, principal investigator: Jerry Ingrande, MD, MS; date of registration: May 1, 2012). Written, informed consent was obtained from all subjects before enrollment. Thirty subjects were enrolled (20 morbidly obese, body mass index ≥40; 10 lean, body mass index <25). Inclusion criteria included subjects of adult age (≥18 years) with an American Society of Anesthesiologists (ASA) physical status I, II, or III undergoing elective surgery requiring general anesthesia. Obese patients underwent elective laparoscopic sleeve gastrectomies, gastric bandings, and gastric bypasses. Lean patients underwent a variety of elective general, plastic, gynecologic, and ear, nose, and throat (ENT) cases. Patients with evidence of hepatic, renal, cardiovascular, pulmonary, or major psychiatric illness were excluded from the study. Patients with a history of difficult intubation or who were taking concomitant medications that may alter the pharmacodynamics of propofol (eg, sedatives, opioids, or other medications) were also excluded.

Preinduction

Immediately before surgery, total body weight (TBW), lean body weight (LBW), and percent body fat of the subject were determined using a Tanita body impedance scale (Tanita Corp, Tokyo, Japan). An 18 or 20 G peripheral intravenous (IV) catheter was placed in the left or right upper extremity antecubital vein. Per study protocol, a 20 G arterial catheter was inserted into the subject's left or right radial artery after infiltration with 2% lidocaine. Subjects were then transported to the operating room. Standard ASA monitors were applied to the subject. Noninvasive bioimpedance cardiac output measurements were obtained using a NICCOMO cardiac output analyzer (medisde, Ilmenau, Germany). Hundred percentage oxygen was applied to the subject via facemask. No subject received premedication before induction. Before induction, each subject was asked to hold a weighted 20 mL saline-filled syringe between thumb and index finger in the hand opposite the IV and was instructed not to drop it.

Induction of Anesthesia and Maintenance

The 20 morbidly obese subjects received a propofol infusion of 100 mg·kg^{-1}·hour^{-1} based on LBW. The 10 lean subjects received a propofol infusion of 100 mg·kg^{-1}·hour^{-1} based on TBW. The dose regimen of 100 mg·kg^{-1}·hour^{-1} was chosen because this will result in relatively fast induction times (1–2 minutes) and not expose the morbidly obese subjects to risk associated with prolonged induction. LBW was chosen in the morbidly obese group because, in our preliminary study, LBW was found to be a more appropriate dosing scalar than TBW.[12] Following loss of consciousness—defined as drop of the weighted syringe, the propofol infusion was stopped, and no further propofol was administered. Each subject then received a bolus of fentanyl (200 µg) and succinylcholine (1 mg·kg^{-1} TBW) before tracheal intubation. Anesthesia was maintained with sevoflurane, oxygen, and air. All physiologic data from the anesthesia monitors and infusion parameters were recorded via a computer running the RUGLOOP II application (Demed, Temse, Belgium).

Plasma Collection

Arterial samples were collected from an arterial line catheter placed in the subjects' left or right radial artery. Each catheter was connected to a closed sampling port. Arterial blood samples (3–4 mL) for propofol plasma concentration determination were drawn at time 0, every 5 seconds to 2 minutes, every 0.5 minutes to 4 minutes, at minute 5, every 2 minutes to minute 15, every 15 minutes to 1 hour, every 60 minutes to 6 hours, and every 120 minutes to 16 hours. Continuous blood flow from the arterial catheter was achieved by connecting the catheter to a negative pressure syringe-pump, which generated a flow of 50 mL·minute^{-1} as described.[2] Samples were obtained at a sampling port proximal to the arterial catheter (dead space 1 mL) by 2 anesthesiologists dedicated only to sampling. Following collection, samples were immediately placed on ice and centrifuged. Separated plasma was removed and stored at –80°C until analysis.

Propofol Plasma Concentration Analysis

Gas chromatography-mass spectroscopy (GC-MS)/MS analysis was performed on a Bruker Scion TQ gas chromatographer coupled with the triple mass spectrometer (Bruker Corporation, Fremont, CA). The instrument was fitted with a Bruker BR-5 column (30 m × 250 µm internal diameter [ID] × 25 µm film thickness). The GC was equipped with split/splitless injector and was operated at the split ratio 10:1 for 1 µL sample injection volume and He carrier gas at 1.1 mL·minute^{-1}. The method used an isocratic oven program (195°C) to achieve a cycle time of 2.3 minutes injection-to-injection, with inlet temperature of 300°C and MS source temperature of 230°C.

Mass spectra were obtained using electron impact ionization mode with electron impact energy of 70 eV. The mass spectrometer was operated in selected reaction monitoring mode (SRM). Two SRM transitions were used for each propofol and the internal standard: 163.2 > 117.0, 178.0 > 163.0 and 195.3 > 177.0, 177.0 > 125.0, respectively. The linear calibration curve range extended from 1 to 4000 ng·mL^{-1}, and lower limits of quantitation were 1 pg·mL^{-1} in extracted plasma.

Statistical Analysis and Compartmental/Recirculatory PK Model Construction

All statistical computations were performed using the R software package.[13] A population pharmacokinetic/pharmacodynamic (PK/PD) model was developed using mixed-effects modeling using NONMEM 7.2 and 7.3 software (ICON Development Solutions, Hanover, MD) and PLTTools (PLTsoft, San Francisco, CA).

PK models were constructed using the ADVAN 13 subroutine in NONMEM. Basic structural models were first evaluated (compartmental and recirculatory) by performing naïve-pooled analyses, combining all PK observations from all subjects. Next, mixed-effects models (MEM) were constructed. Population variability was modeled as a random effect for each PK variable using the model:

$$P_i = P_{TV} x\, e^n,$$

where P_i was defined as the individual parameter value, P_{TV} is the typical value of the parameter in the population, and n defined as the random variable.

A constant coefficient of variation model was used to describe intraindividual error according to the equation:

$$C_{ij} = C_{\text{predicted},ij} \times (1 + \varepsilon_{ij}),$$

where C_{ij} was defined as the jth plasma concentration in the ith subject, $C_{\text{predicted,ij}}$ is the jth predicted plasma concentration in the ith subject, and ε_{ij} is the coefficient of proportional residual error.

Linear and logistic regression was used to analyze the relationship between PK parameters and continuous and categorical covariates, respectively. Covariates that appeared to have a significant relationship with any PK parameter were introduced into the model in a forward, stepwise manner. Forward covariate selection was used until there was no further improvement in the model. Backward covariate selection was used to obtain the best-fit and most parsimonious model. Observation of the objective function (OBJ; minus

twice the log-likelihood [−2 LL]) was used to facilitate model selection, with a significant minimization ($P < .01$) used to facilitate model selection. Predictive performance and model validation was then performed by visual predictive checks of observed versus predicted concentrations for both population and individual fits and via analysis of weighted residuals versus predicted concentration and time.

Model bias was estimated by calculating the mean prediction error (MPE) according to the equation[9,10]

$$\text{MPE} = \left(\frac{1}{n} \right) \sum_{i=1}^{n} (\text{predicted } Cpij - \text{measured } Cpij)$$

A value of 0 indicates zero bias. Model precision was assessed by calculating the mean square prediction error (MSE)[9,10]:

$$\text{MSE} = \left(\frac{1}{n} \right) \sum (\text{predicted } Cpij - \text{measured } Cpij)^2$$

A value of 0 indicates perfect precision.

Goodness of fit plots—including the ratio of observed to predicted concentrations versus time, population predicted versus observed concentrations, and overall model fit were performed for model evaluation. If these plots indicated an unacceptable amount of bias, the model was excluded regardless of the OBJ value.

After identification of the final model, LL plots were performed for each model parameter estimate (THETA). If a parameter that was included in the model could not be reliably estimated from the data, the model was assumed to be overfit and the model was rejected.

Bootstrap analysis of the final model was performed for internal validation. The bootstrap was performed by first creating 1000 new datasets of the same length as the original dataset by resampling subjects at random from the original dataset. The final model was fit to each dataset, and the distribution of the parameter estimates (THETAs) examined, ultimately providing mean, median, and percentiles of each of the parameter estimates in the model.

Prediction-corrected visual predictive checks were performed to graphically assess whether simulations of each model can reliably predict the central trend (50th percentile) and variability (5th and 95th percentiles) in the observed data. Each model was used to create 1000 new datasets each containing simulated propofol concentrations. The 5th, 50th, and 95th percentiles of the simulated data were compared to the same percentiles of the actual observed data.

ANN Model Construction

The model architectures for ANNs included long short-term memory (LSTM) and gated recurrent unit (GRU)—both of which are forms of recurrent neural networks, which are beneficial in predicting sequential data. LSTM is a special ANN that memorizes long-term dependencies by maintaining an internal state variable which is passed from one node to another.[14,15] GRU is similar to LSTM but has fewer parameters required.[16,17] For GRUs, the model's hidden state (representing the model's current latent context) at each step is fed back into the model at the next step. Each step corresponded to a single observation. In addition to the hidden state, features corresponding to the current state, as well as static features including patient covariates, were passed in the model. This specific architectural choice has been shown to be useful when generating language (eg, words or characters) over several steps.[18]

With both LSTM and GRU models, we experimented with the number of layers and number of hidden units, with a combination of values chosen between 1–4 and 5–20, respectively. For the final ANN architecture, the input layer consisted of 9 nodes for each time step (gender, LBW, TBW, age, total propofol dose administered [mg], rate of propofol administration [mg/min]), 2 hidden layers (10 nodes per layer), and 47 outputs, each node representing propofol concentration at a distinct time point. Model weights were calculated with a gradient descent optimization algorithm using the Adam optimizer in Tensorflow.[19] Gradient descent was terminated once performance no longer improved on the validation set. Selection of an optimal model without overfitting was performed using the training and validation sets. The model is trained to find optimal weights on the training set during one training epoch; then this model is applied to the validation set, in which the validation error was then calculated—defined as the mean square error (Equation 4). Validation error tends to decrease as the train epochs repeat; however, the error begins to increase when overfitting occurs. At the training epoch, just before this occurred, was when the final model was chosen. The final model chosen was then applied on the test set in which the test error is reported.

We performed cross-validation to calculate a mean test error for the ANN model. The cross-validation setup consisted of 22 data points (each data point referring to the set of PK data from 1 subject) and 11 folds. Thus, each fold consisted of 2 subjects. Nine folds were used as a training set and 1 fold for validation of the model. The validation set was used to select the best model, and training was terminated once the validation error did not improve further. The final fold was used as a testing set, where the final model was applied to obtain a test error. Folds were reassigned as training, validation, and test iteratively until all folds served as a test set. In this way, each

data point appears in a test fold exactly once, such that we can accumulate a test error that accounts for all of the data points.

To make our model even more competitive with the MEM, we treated the MEM model's output as additional input features into the GRU. This helps with model convergence, allowing our model to improve on the already good performance of the existing MEM model. We also used a form of ensembling to combine our ANN with simple feature transforms as well as the output of the MEM model. In a single training step, the ANN learned to weigh the results of the existing MEM model and incorporate these into the final ANN. These modifications were to ensure that our model was able to quickly converge to a strong solution, which prevented our model from overfitting. Ensembling is a commonly used technique in machine learning where multiple classifiers are used to obtain performance superior to a single classifier method.[20]

RESULTS

PK analysis was performed with data pooled from 30 subjects each contributing 47 PK observations for a total of 1410 observations. Six subjects were excluded from the final analysis because of propofol infusion problems. The final analysis included 24 subjects contributing 47 observations each. The demographics of these subjects are shown in Table 1.

PK models were refined using NONMEM 7.2 and 7.3 according to −2 LL and the standard errors of the parameter estimates. One-, 2-, 3-, and 4-compartment MEM were constructed. OBJ, MPE, and MSE are reported (Supplemental Digital Content, Table 1, http://links.lww.com/AA/D89). Of the compartmental models, a 4-compartment model best fit the data (OBJ: 497.92; MPE: 0.108; MSE: 31.61). Analysis of covariates versus model parameters did not demonstrate any significant relationships. Covariates analyzed included TBW, LBW, age, cardiac output, and gender. The base 4-compartment model was accepted as the final compartment model.

A recirculatory model using a proportional error model was constructed (OBJ: 160.09; MPE: −0.537; MSE: 22.41; Supplemental Digital Content, Figure 1, http://links.lww.com/AA/D89). Subsequently, the same structural model was evaluated however, using a combined additive and proportional error model.[21] The combined error model had better performance compared to the proportional error model (OBJ: 71.39; MPE: 0.348; MSE: 23.92). Analysis of the relationships between model parameters versus measured covariates (TBW, LBW, age, cardiac output, and gender) revealed a positive linear relationship between V4 and V5 with age >65 years. A separate estimate for V4 and V5 for ages >65 years were included and subsequently removed from the basic model in a forward and backward manner. The addition of age as a covariate for both V4 and V5 did not improve model fit (OBJ: 95.74; MPE: −0.308; MSE: 23.53). The combined error recirculatory model, without covariates, was therefore accepted as the final model. LL plots confirmed that all parameters could be reliably estimated in the final model (Supplemental Digital Content, Figure 2, http://links.lww.com/AA/D89). PK parameters are outlined in Table 2.

ANNs were constructed and compared to the mixed models. An LSTM and a GRU model were constructed (Figure 1). Both models had similar performance, but the GRU was accepted as the best model due to its more parsimonious structure. Tensorflow code for the GRU model can be found at http://jmcauley.ucsd.edu/propofol_dec19.html. The GRU model had lower bias and similar precision as the optimized recirculatory model (MPE: 0.161; MSE: 20.83).

Plots of observed versus predicted concentrations over time showed that during the first 20 minutes, the combined error recirculatory model and GRU models both showed a constant overprediction bias (Figure 2A, b and d). The 4-compartment model showed an initial overprediction bias followed by an under-prediction bias (Figure 2A, c). All models with the exception of the 4-compartment model overpredicted the observed concentration (Figure 2A). The overprediction bias of the recirculatory models and GRU model remained after 20 minutes (Figure 2B). There was a high under-prediction bias in the 4-compartment model (Figure 2B, c).

Direct comparison of observed versus predicted concentrations demonstrated a minimal overprediction bias at higher concentrations in the combined error recirculatory, 4-compartment, and GRU models (Figure 3B–D, respectively).

Each model was plotted against the raw data as shown in Figure 4. The 4-compartment model suffers from under-prediction bias after 5 minutes (Figure 4, inset).

Table 1. Demographics

Group	N	Gender (M/F)	Age (y)	TBW (kg)	LBW (kg)	BMI (kg/m²)	Cardiac Output (L/min)
Obese	17	3/14	42.9 (13.4)	129.9 (20.4)	61.8 (12.9)	46.2 (5.9)	9.6 (3.0)
Lean	7	3/4	53.6 (10.3)	72.7 (4.8)	51.2 (5.6)	23.5 (1.1)	8.0 (0.96)
Population	24	6/18	46.1 (14.0)	113.3 (31.3)	58.7 (12.2)	39.5 (11.6)	9.2 (2.7)

Data presented as mean (standard deviation).
Abbreviations: BMI, body mass index; LBW, lean body weight; TBW, total body weight.

Table 2. Pharmacokinetic Parameters for the 4-Compartment and Final Recirculatory Models of Propofol

4-Compartment Model			
Parameter	**Parameter Name**	**Parameter Estimate**	**5th and 95th Percentiles of Parameter Estimates**
V1	Central volume	0.63	0.14, 0.70
V2	Peripheral volume 1	0.53	0.17, 0.91
V3	Peripheral volume 2	3.61	0.60, 9.23
V4	Peripheral volume 3	122	17.9, 426
CL1	Elimination clearance	0.53	0.51, 0.58
CL2	Distribution clearance 2	335	0.19, 579
CL3	Distribution clearance 3	0.24	0.11, 0.42
CL4	Distribution clearance 4	0.20	0.12, 0.35
Final Recirculatory Model			
Parameter	**Parameter Name**	**Parameter Estimate**	**5th and 95th Percentiles of Parameter Estimates**
Q	Plasma flow	2.10	1.32; 2.95
CL1	Elimination clearance	2.51	2.34, 2.75
Q4	Intertissue clearance (fast)	0.53	0.32, 0.76
Q5	Intertissue clearance (slow)	0.50	0.32, 0.71
V1	Central volume	0.94	0.43, 1.82
V2	Lung volume	0.38	0.27, 0.53
V3	Arterial volume	0.49	0.24, 0.85
V4	Peripheral volume (fast)	2.36	1.06, 4.07
V5	Peripheral volume (slow)	120	93.3, 150

Flow in liters per minute; volumes in liters; clearances presented as liters per minute.

Prediction-corrected visual predictive checks demonstrated that 6.8% of the data fell outside of the 5% and 95% percentile of observations in the 4-compartment model (Supplemental Digital Content, Figure 3, http://links.lww.com/AA/D89). Seven percent of the data fell outside of these percentiles in our final, combined error recirculatory model (Supplemental Digital Content, Figure 4, http://links.lww.com/AA/D89).

DISCUSSION

ANNs have been praised for their ability to model complex nonlinear data and have been proposed as a possible replacement for MEM.[8] This study aimed to evaluate the performance of a conventional mammillary compartmental model, a recirculatory model, and an ANN in characterizing the PK of propofol using a frequently sampled prospectively collected dataset. A GRU model had comparable performance to the recirculatory model, with both having better performance compared to the 4-compartment model.

Knowledge of an anesthetic induction agent kinetics during induction is necessary to achieve a safe and therapeutic peak concentration. Unfortunately, compartmental PK models are limited in their ability

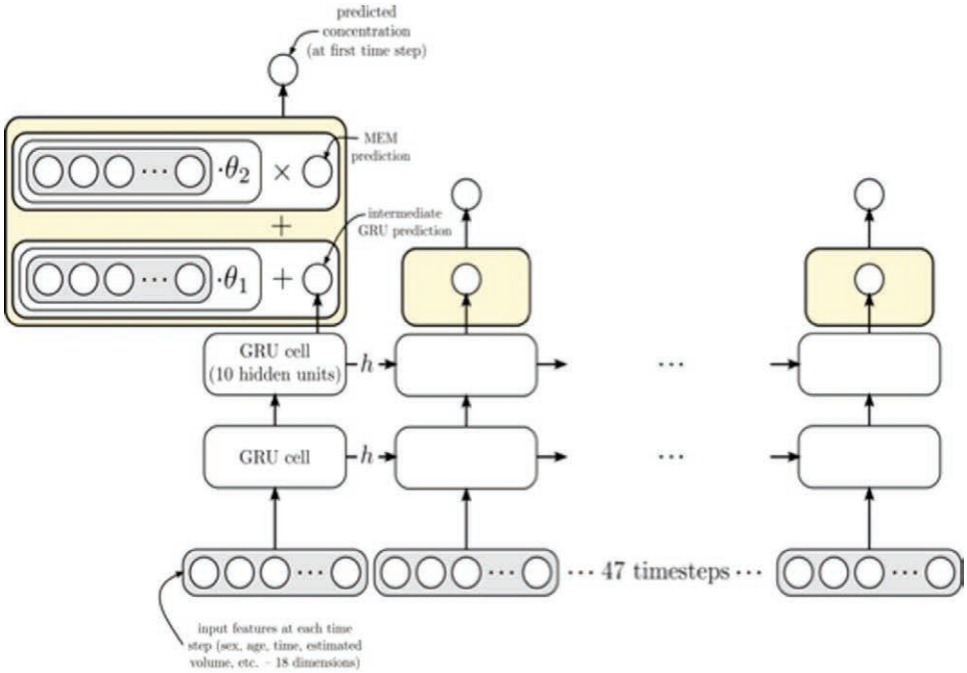

Figure 1. Schematic of the GRU neural network model. The input layer consisted of 9 nodes for each time step and 2 hidden layers with 10 nodes per layer. There were a total of 47 outputs, with each representing propofol concentration at the measured time-points. GRU indicates gated recurrent unit; MEM, mixed-effects model.

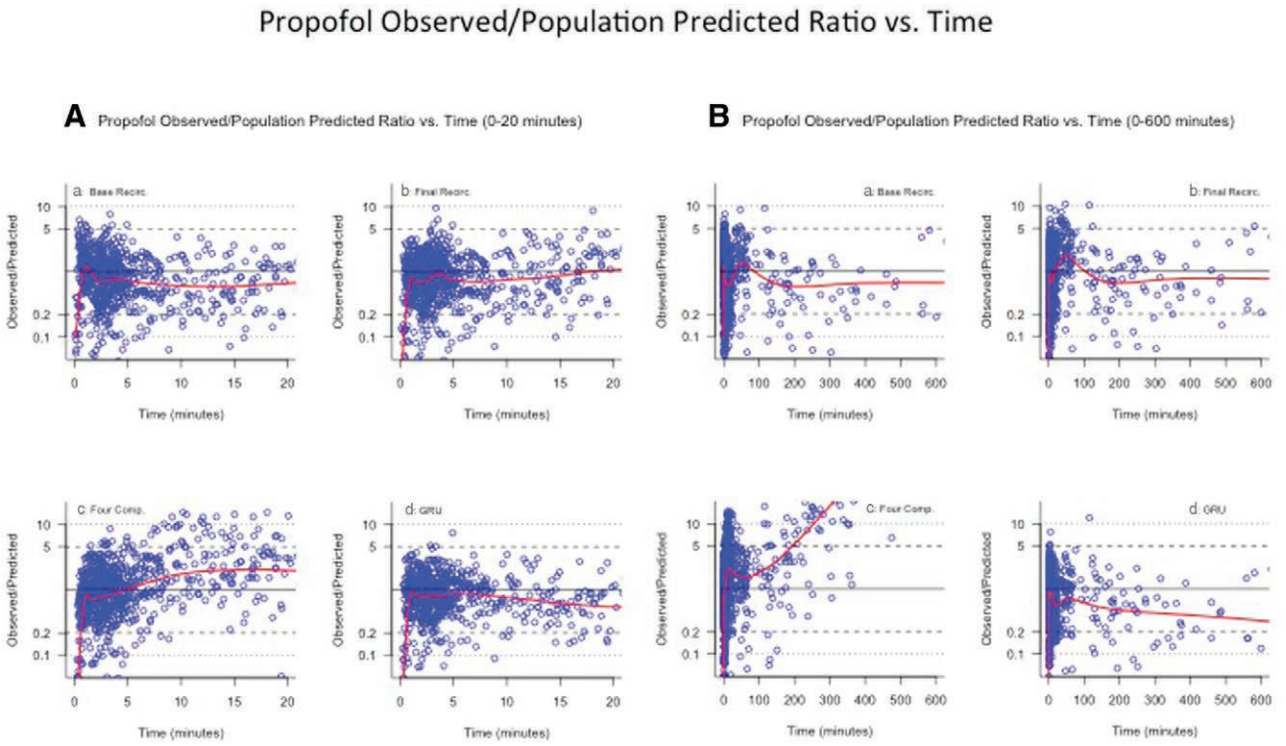

Figure 2. A, Propofol observed/population predicted concentration ratio versus time (0–20 min). There was a consistent overprediction bias in the final, combined error recirculatory model (b) and GRU model (d). The base recirculatory model demonstrated an initial under-prediction, followed by an overprediction bias (a), as opposed to the 4-compartment model, which initially overpredicted propofol concentrations before under-predicting concentrations after 5 min (c). Blue circles: ratio of observed versus predicted propofol concentrations at each time point. Red line: smoothed fit of the regression. B, Propofol observed/population predicted concentration ratio versus time (0–600 min). There is a persistence in model overprediction seen in the both recirculatory and GRU models (a, b, d). There was an unacceptably high under-prediction bias in the 4-compartment model in propofol concentrations measured after 100 min (c). Blue circles: ratio of observed versus predicted propofol concentrations at each time point. Red line: smoothed fit of the regression. GRU indicates gated recurrent unit.

to estimate induction kinetics.[1] These models ignore intravascular mixing and assume complete mixing of drug the moment it is administered. The consequences of this assumption, particularly on overestimating central volume of distribution, are well described.[1,4,22]

Furthermore, unless frequent, high-frequency sampling is performed, the problem of overestimation is intensified.[23] Estimations of the central volume of distribution are directly related to the blood-sampling schedule.[24] Models that rely on sparse sampling after drug administration will fail to capture peak plasma concentrations. This may result in overestimation of central volume. Such inaccuracies of a PK model may be masked during the maintenance phase of anesthesia.[25] Miscalculations in drug administration due to errors in PK models are enhanced during the induction phase, a time at which plasma and effect-site concentrations are changing rapidly. Models derived from infrequent blood sampling not only overestimated the central volume, but also the volume and clearance from the rapidly equilibrating tissue.[26]

The fact that the recirculatory model had better performance than the compartmental model is not surprising considering the limitations of a compartment model.[1] A recirculatory model characterizes the delay between drug administration and the site of sampling. We understand this delay to be secondary to transit of drug through the peripheral venous system and heart/lungs/great vessels.[27] Failure to account for this delay resulted in an underestimation of plasma concentration during the first minutes of administration (Figures 2 and 4).

When we analyzed model performance after induction, we found that compartmental model performance was poor (Figures 2 and 4). We were unable to describe a compartmental model that performed well in predicting early and late kinetics. This could only occur if model parameter estimates were constrained to specific values. However, this introduced model overfitting.

Although our recirculatory model demonstrated better performance compared to the compartmental model, we hypothesized that the ANN would have even better performance because of the theoretical advantage of being able to model complex nonlinear systems without assuming a structural model. ANNs are deep feed-forward networks where all layers share the same weights.

The main goal of ANNs is to establish and learn long-term input dependencies. However, they are

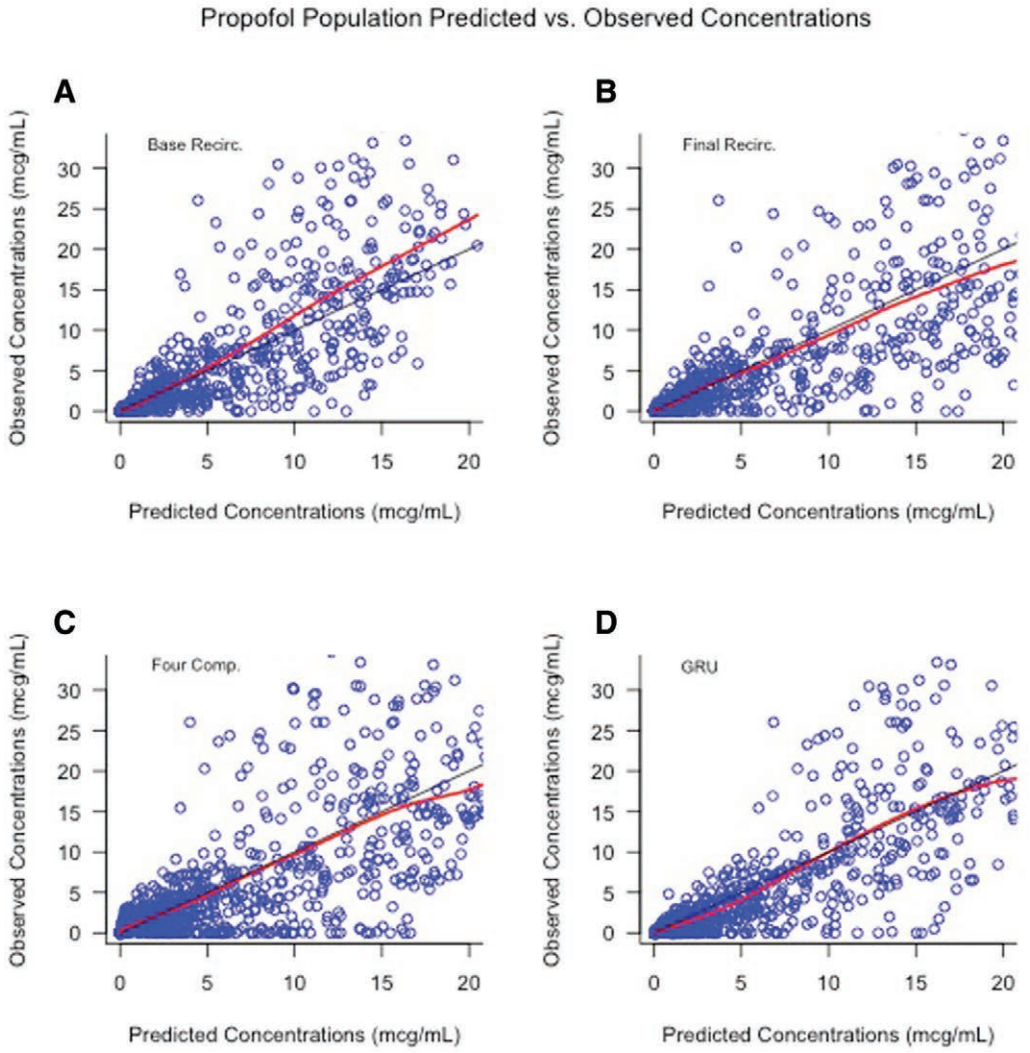

Figure 3. Propofol population predicted versus observed concentrations. The under-prediction bias at higher concentrations seen in the base recirculatory model (A) was corrected in the final, combined error model (B). There was, in general, good agreement between observed and population predicted concentrations in the final recirculatory model (B), 4-compartment model (C), and GRU model (D). There is a slight overprediction bias in the 4-compartment and GRU models at higher concentrations of propofol. Blue circles: observed concentrations. Black line: line of identity. Red line: smoothed fit of the regression. GRU indicates gated recurrent unit.

limited by their inability to store this long-term information.[28] LSTM networks correct for this. They are therefore more effective than conventional ANNs when there are multiple layers for each time step.[14] GRUs are a variation of the LSTM but with fewer parameters. They combine the cell state and hidden units resulting in a model that is simpler than the standard LSTM model.

In this study, a GRU model modestly overestimated plasma concentrations during the first 20 minutes (Figure 2A, d). However, the overestimation bias in the GRU model was consistent, in contrast to the recirculatory model where there was oscillation between over and under-prediction bias (Figure 2A and B).

We expected the GRU to have the best performance of all the models compared in this study, as the strength of these models is their lack of confinement to a specific structure. However, we did not see this. We presume that this is secondary to the small number of samples available to train the GRU. Although we did not perform an a priori sample size calculation, the small uncertainty in the parameter estimates demonstrates that the sample size was justified. Our data is large for a prospective PK study; however, it is small for training ANNs. A study comparing the performance of LTSM neural network to a response surface model in predicting bispectral index values during infusions of remifentanil and propofol demonstrated improved accuracy with the LSTM model.[29] This study was comprised of over 2 million data points, while ours included 1128.

While a structure-less model may reduce bias, unacceptably high variance may result, especially when training datasets are small.[30] For this reason, we

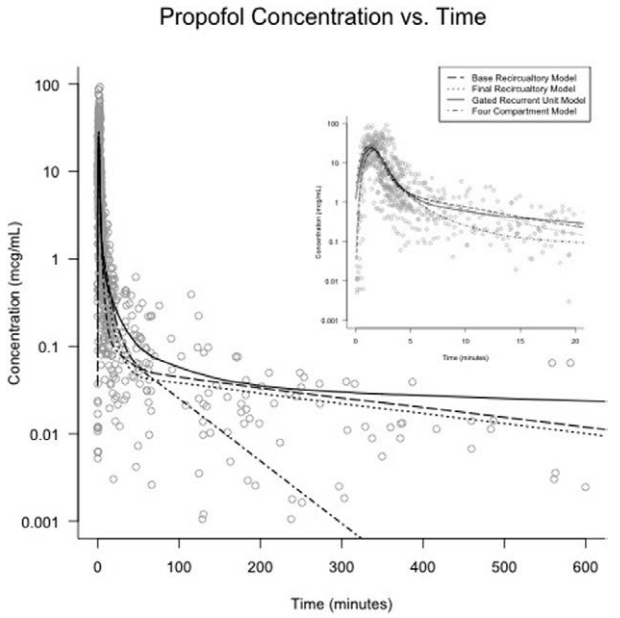

Figure 4. Final model fit versus observed concentrations. All 4 models have similar performance during the early drug distribution phase (inset). There is an under-prediction bias in the 4-compartment model after 5 min (inset). The under-prediction bias seen in the 4-compartment model worsens after 100 min. Grey circles: observed concentrations. Dashed line: smoothed base recirculatory model fit. Dotted line: smoothed final recirculatory model fit. Solid line: smoothed GRU model fit. Broken line: smoothed 4-compartment model fit. GRU indicates gated recurrent unit.

performed ensemble learning. In this way, the model learns to "weight" certain time-points during which the MEM is a more reliable predictor and put weight on the GRU component only when its predictions are better. This method provided 2 major benefits to the model: (1) benefit from the MEMs predictions in cases where that model is accurate (but learn to ignore it otherwise); and (2) arrival at an accurate solution quickly since the model's inputs already include a strong classifier, which can prevent overfitting. Given enough data, it is likely that we could discard this ensembling component, though we found it useful when trying to learn a complex model with a relatively small number of samples.

None of our models have been prospectively validated to assess their clinical performance. However, we performed simulations of a standard induction dose of propofol (2 mg/kg given over 10 seconds) and compared these to a clinically validated model (Supplemental Digital Content, Figure 5, http://links.lww.com/AA/D89).[31] Concentration-time profiles between the 4-compartment and GRU models as well as the model published by Schnider et al[31] were similar. All 3 models demonstrated peak concentrations that were higher than the recirculatory model.

ANNs have yet to be used to model early induction kinetics, a time when drug concentration is changing rapidly. Though these models show promise in modeling complex nonlinear systems, their utility in modeling complex PK data may be limited by the size of the data available to train the model. This study demonstrated similar performance between a recirculatory model and GRU neural network. However, superior performance of the neural network may be seen with a larger dataset. ■

ACKNOWLEDGMENTS

The authors thank Vincent Coates Foundation, Los Altos, CA; Yuxi Wu, BA, Stanford University Mass Spectrometry Laboratory, Stanford, CA.

DISCLOSURES

Name: Jerry Ingrande, MD, MS.
Contribution: This author helped design the study, collect and analyze the data, and prepare the manuscript.
Name: Rodney A. Gabriel, MD, MAS.
Contribution: This author helped analyze the data and prepare the manuscript.
Name: Julian McAuley, PhD.
Contribution: This author helped analyze the data and prepare the manuscript.
Name: Karolina Krasinska, MSc.
Contribution: This author helped analyze the data and prepare the manuscript.
Name: Allis Chien, PhD.
Contribution: This author helped analyze the data and prepare the manuscript.
Name: Hendrikus J. M. Lemmens, MD, PhD.
Contribution: This author helped design the study, collect and analyze the data, and prepare the manuscript.
This manuscript was handled by: Ken B. Johnson, MD.

REFERENCES

1. Fisher DM. (Almost) everything you learned about pharmacokinetics was (somewhat) wrong! *Anesth Analg*. 1996;83:901–903.
2. Masui K, Kira M, Kazama T, Hagihira S, Mortier EP, Struys MM. Early phase pharmacokinetics but not pharmacodynamics are influenced by propofol infusion rate. *Anesthesiology*. 2009;111:805–817.
3. Park K, Verotta D, Blaschke TF, Sheiner LB. A semiparametric method for describing noisy population pharmacokinetic data. *J Pharmacokinet Biopharm*. 1997;25:615–642.
4. Krejcie TC, Avram MJ. Recirculatory pharmacokinetic modeling: what goes around, comes around. *Anesth Analg*. 2012;115:223–226.
5. Masui K, Upton RN, Doufas AG, et al. The performance of compartmental and physiologically based recirculatory pharmacokinetic models for propofol: a comparison using bolus, continuous, and target-controlled infusion data. *Anesth Analg*. 2010;111:368–379.
6. Tsamandouras N, Rostami-Hodjegan A, Aarons L. Combining the 'bottom up' and 'top down' approaches in pharmacokinetic modelling: fitting PBPK models to observed clinical data. *Br J Clin Pharmacol*. 2015;79:48–55.
7. Villaverde AF, Barreiro A, Papachristodoulou A. Structural identifiability of dynamic systems biology models. *PLoS Comput Biol*. 2016;12:e1005153.
8. Gambus P, Shafer SL. Artificial intelligence for everyone. *Anesthesiology*. 2018;128:431–433.
9. Ng CM. Comparison of neural network, Bayesian, and multiple stepwise regression-based limited sampling models to estimate area under the curve. *Pharmacotherapy*. 2003;23:1044–1051.
10. Chow HH, Tolle KM, Roe DJ, Elsberry V, Chen H. Application of neural networks to population pharmacokinetic data analysis. *J Pharm Sci*. 1997;86:840–845.

11. Brier ME, Aronoff GR. Application of artificial neural networks to clinical pharmacology. *Int J Clin Pharmacol Ther*. 1996;34:510–514.
12. Ingrande J, Brodsky JB, Lemmens HJ. Lean body weight scalar for the anesthetic induction dose of propofol in morbidly obese subjects. *Anesth Analg*. 2011;113:57–62.
13. R Core Team 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/. Accessed March 13, 2014.
14. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9:1735–1780.
15. Gers FA, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM. *Neural Comput*. 2000;12:2451–2471.
16. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv*. 2014;1412.3555v1.
17. Cho K, Van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: encoder-decoder approaches. *arXiv*. 2014;1409.1259v2.
18. Ni J, Lipton ZC, Vikram S, McAuley J. Estimating reactions and recommending products with generative models of reviews. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Taipei, Taiwan, November 27–December 1, 2017; 2017:783–791.
19. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv*. 2014;1412.6980v9.
20. Dietterich T. Machine-learning research: four current directions. *AI Magazine*. 1997;18:97–136.
21. Proost JH. Combined proportional and additive residual error models in population pharmacokinetic modelling. *Eur J Pharm Sci*. 2017;109S:S78–S82.
22. Krejcie TC, Avram MJ. What determines anesthetic induction dose? It's the front-end kinetics, doctor! *Anesth Analg*. 1999;89:541–544.
23. Weiss M. Modelling of initial distribution of drugs following intravenous bolus injection. *Eur J Clin Pharmacol*. 1983;24:121–126.
24. Avram MJ, Henthorn TK, Shanks CA, Krejcie TC. The initial rate of change in distribution volume is the sum of intercompartmental clearances. *J Pharm Sci*. 1986;75:919–920.
25. Echevarría GC, Elgueta MF, Donoso MT, Bugedo DA, Cortínez LI, Muñoz HR. The effective effect-site propofol concentration for induction and intubation with two pharmacokinetic models in morbidly obese patients using total body weight. *Anesth Analg*. 2012;115:823–829.
26. Avram MJ, Krejcie TC. Using front-end kinetics to optimize target-controlled drug infusions. *Anesthesiology*. 2003;99:1078–1086.
27. Upton RN, Ludbrook G. A physiologically based, recirculatory model of the kinetics and dynamics of propofol in man. *Anesthesiology*. 2005;103:344–352.
28. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans Neural Netw*. 1994;5:157–166.
29. Lee HC, Ryu HG, Chung EJ, Jung CW. Prediction of bispectral index during target-controlled infusion of propofol and remifentanil: a deep learning approach. *Anesthesiology*. 2018;128:492–501.
30. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. *Neural Comput*. 1992;4:1–58.
31. 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.