A Generative and Causal Pharmacokinetic Model for Factor VIII in Hemophilia A: A Machine Learning Framework for Continuous Model Refinement

Alexander Janssen[1,*], Louk Smalbil[2], Frank C. Bennis[3,4], Marjon H. Cnossen[5], Ron A. A. Mathôt[1,*] and for the OPTI-CLOT study group and SYMPHONY consortium

In rare diseases, such as hemophilia A, the development of accurate population pharmacokinetic (PK) models is often hindered by the limited availability of data. Most PK models are specific to a single recombinant factor VIII (rFVIII) concentrate or measurement assay, and are generally unsuited for answering counterfactual ("what-if") queries. Ideally, data from multiple hemophilia treatment centers are combined but this is generally difficult as patient data are kept private. In this work, we utilize causal inference techniques to produce a hybrid machine learning (ML) PK model that corrects for differences between rFVIII concentrates and measurement assays. Next, we augment this model with a generative model that can simulate realistic virtual patients as well as impute missing data. This model can be shared instead of actual patient data, resolving privacy issues. The hybrid ML-PK model was trained on chromogenic assay data of lonoctocog alfa and predictive performance was then evaluated on an external data set of patients who received octocog alfa with FVIII levels measured using the one-stage assay. The model presented higher accuracy compared with three previous PK models developed on data similar to the external data set (root mean squared error = 14.6 IU/dL vs. mean of 17.7 IU/dL). Finally, we show that the generative model can be used to accurately impute missing data (< 18% error). In conclusion, the proposed approach introduces interesting new possibilities for model development. In the context of rare disease, the introduction of generative models facilitates sharing of synthetic data, enabling the iterative improvement of population PK models.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Population pharmacokinetic (PK) models are a useful tool to personalize treatment. In hemophilia A, there are many models to describe the PK of different factor VIII (FVIII) concentrates. Ideally, there exists a unified PK model that offers accurate predictions for all FVIII concentrates.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ Can we combine generative modeling and techniques from causal inference to create a hybrid machine-learning (ML)-PK model that can accurately predict the PK of drug B when it was trained on drug A?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ The ML-PK model more accurately predicts concentrations of drug B compared with models specifically trained on such data. This might be attributable to the implementation of causal covariates. The generative model can accurately impute missing data.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ The proposed ML-PK model is interpretable and training is simple. By sharing generative models, a synthetic copy of otherwise sensitive data can still be made available. The framework enables the continuous refinement of population PK models on new data.

[1]Department of Clinical Pharmacology, Hospital Pharmacy, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands; [2]Quantitative Data Analytics Group, Department of Computer Science, VU Amsterdam, Amsterdam, The Netherlands; [3]Follow Me & Emma Neuroscience Group, Emma Children's Hospital, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands; [4]Amsterdam Reproduction and Development, Amsterdam, The Netherlands; [5]Department of Pediatric Hematology, Erasmus MC Sophia Children's Hospital, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands. *Correspondence: Alexander Janssen (a.janssen@amsterdamumc.nl), Ron A. A. Mathôt (r.mathot@amsterdamumc.nl)

OPTI-CLOT study group and SYMPHONY consortium are presented in the Acknowledgments.

Received September 29, 2023; accepted January 25, 2024. doi:10.1002/cpt.3203

Hemophilia A is an X-linked recessive bleeding disorder caused by a deficiency or dysfunction of the blood clotting factor VIII (FVIII). Severe hemophilia A (endogenous FVIII activity level < 1% or < 1 IU/dL) are at increased risk of prolonged bleeding, significant morbidity, and reduced quality of life. Personalized prophylaxis involving the administration of exogenous FVIII is the cornerstone of hemophilia A treatment. The pharmacokinetic (PK) properties of FVIII play a crucial role in the determination of the optimal dosing regimen for the prevention of spontaneous bleeding. However, the significant interindividual variability in the PK of FVIII makes accurately predicting FVIII concentration-time profiles challenging.[1,2]

Population PK modeling has emerged as a valuable tool for characterizing the PK of drugs in heterogeneous patient populations. Several of such models have already been developed for the wide range of recombinant FVIII (rFVIII) concentrates currently used in clinical practice.[3] However, most have been developed for a specific brand of rFVIII concentrate on relatively small patient populations. This might pose problems, as differences in covariate implementations, potential biases in small or single center data sets, varying PK for different rFVIII formulations, or the FVIII assay type/reagents can all potentially affect model accuracy. External validation studies have indeed shown that model parameters frequently need to be adjusted when attempting predictions on new data.[3–6] Ideally, population PK models correct for these sources of variability, but this requires larger scale data sets rarely available in part due to data confidentiality.

In order to adjust for variability between subpopulations, it can be useful to consider causal inference techniques during model development. Explicit use of these techniques has been lacking from the pharmacometrics literature,[7] although model components are informally judged based on biological plausibility. In addition, counterfactual analysis is used extensively in practice, for example, when simulating individual drug exposure following alternative (i.e., "unseen") dosing schedules. However, more complex queries, such as "what if the patient received a different drug," are not necessarily supported by most models. To answer such questions, population PK models should ideally incorporate notions of causality. As an example, von Willebrand factor (VWF) levels are well known to be an important determinant of FVIII clearance, but are rarely included as a covariate.[3] One prominent reason is that VWF levels are seldom measured, and thus frequently unavailable during model development. Alternatively, covariates such as patient age or blood group – which are correlated to VWF – are included. It is, however, likely that these variables have no independent causal effect, but rather that their effects are mediated through VWF.[2,8,9] As a result, interventions affecting VWF levels, such as hemostatic challenges sustained during surgery, are not described by the model, resulting in incorrect predictions.[10,11]

An important component of causal inference involves detailing variable dependencies in a directed acyclic graph (DAG). In a DAG, nodes (variables) are connected via edges, which describe the presence and direction of causal relationships:

$$X \rightarrow Z \rightarrow Y \tag{1}$$

Here, variable X affects variable Z which in turn affects Y. This is analogous to our previous example of age or blood group (X) being related to VWF levels (Z) which has a causal effect on FVIII clearance (Y). When we only implement the effect of X on Y, any effects on Z are not represented by the model. The DAG facilitates the identification of problematic variables and confounders affecting the predictions.

A DAG incorporates known information about causal effects with domain-specific assumptions to describe the data-generating process. Expanding on this view, we can create models that reproduce the observed data based on the relationships in the graph. By supporting population PK models with generative models, it is possible to impute missing data, answer counterfactual queries, or generate realistic virtual patients with corresponding drug exposures. In addition, it is possible to share generative models instead of real patient data, avoiding issues with data privacy. Similarly, we can combine multiple PK models into a model ensemble and weight the predictions for any new patients by their similarity to virtual ones from corresponding generative models. This would offer an interesting new

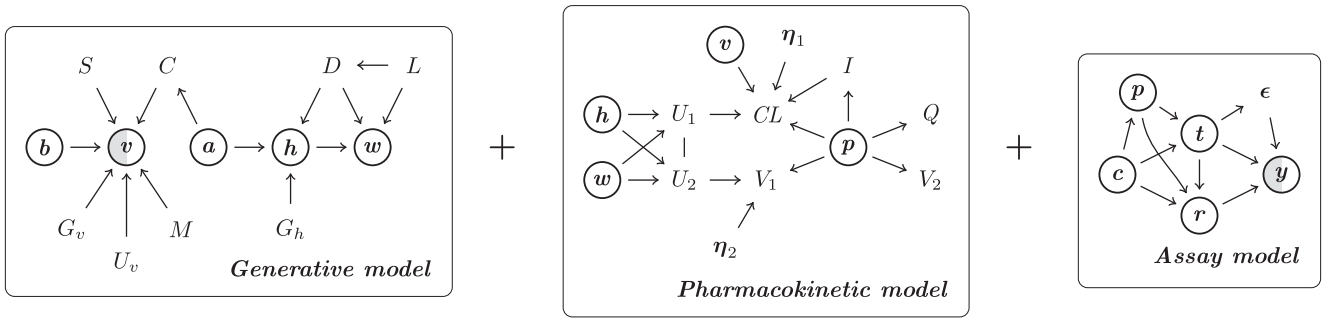

Figure 1 Directed acyclic graphs describing covariate relationships. Observed variables are denoted by circles, variables not in a circle indicate unmeasured or latent variables. Partially filled nodes indicate partially observed variables. Edges without an arrow represent relationship with unknown direction. DAGs were separated per model to facilitate presentation of the graph. a, age; b, blood group; C, co-morbidities; c, treatment center; CL, clearance; D, diet; DAGs, directed acyclic graphs; G_h, genetic factors related to height; G_v, genetic factors related to VWF; h, height; I, product-specific inhibitor; L, lifestyle; M, co-medication; p, rFVIII concentrate; r, assay reagent; S, stress; t, assay type; U, latent variable; U_v, unknown factors related to VWF; v, VWF; w, weight; y, observation; ε, residual variance; η, random effect estimate representing unobserved effects; VWF, von Willebrand factor.

approach to the development of population PK models and is especially relevant in the context of rare diseases.

The contributions of the current work are three-fold: (1) to learn the causal graph describing the sources of variability relevant for treatment using rFVIII concentrates, (2) to develop a generative model based on this graph, and (3) to perform a first step in the development of a PK model that accurately predicts FVIII levels in counterfactual scenarios. Novel machine-learning (ML) algorithms are used to simplify the process of model development and to facilitate others to train the model on new patient populations. Additionally, we use interpretable algorithms to promote causal interpretation and evaluation of the model. This work describes an initial use case for hemophilia A, but the proposed framework of combining causal inference, generative models, and ML-based population PK modeling can of course be applied to other problems.

METHODS

Causal graph

Causal relationships between all relevant variables were described using a DAG and was informed based on previous literature on the PK of FVIII and consultations with (pediatric) hematologists (see **Figure 1**). Correctness of the proposed DAG was evaluated by fitting models for alternative hypotheses and comparing model performance. In the generative model, VWF levels were affected by multiple factors, including patient blood group and age (the latter mediated through the presence of comorbidities). It was assumed that these factors had no independent causal effect on FVIII PK. To test this assumption, an alternative model was fit with age and blood group as covariates (removing VWF) and compared with a model where age and blood group were added after learning the effect of VWF.

Next, the effect of patient weight and/or height on FVIII clearance (CL) and volume of distribution (V_1) acts through unobserved factors U, which could, for example, represent plasma volume. We hypothesized that the variability in this latent factor is more closely correlated to fat-free mass (FFM), and thus compared models using an estimate of FFM[12] to those with weight and/or height as covariates.

We assumed that the variability of intercompartmental clearance (Q) and peripheral volume of distribution (V_2) was relatively low such that covariates were less important for these parameters. However, the specific rFVIII concentrate administered was chosen to affect all PK parameters, of which the effects are likely attributable to differences in molecular structure. Models were also fit including the effect of FFM on Q and V_2.

Finally, the type of assay (one-stage or chromogenic), the assay reagents used, and specific rFVIII concentrates were identified to affect FVIII measurements in the assay model. As an example of the latter effect, lonoctocog alfa levels are known to be underestimated by roughly twofold when using the one-stage assay.[13] We first fit an assay conversion for octocog alfa chromogenic levels to one-stage levels using an exponential model, and then estimated an additional proportional effect for lonoctocog alfa.

Population PK model

A population PK model was constructed using deep compartment models (DCMs), a hybrid ML/PK technique that learns covariate effects directly from data.[14] A specific neural network architecture was used such that model output was interpretable. Additionally, a deep ensemble was fit in order to approximate model uncertainty with respect to the learned effects.[15] After fitting the fixed effects model, random effects model parameters for Bayesian forecasting were estimated by optimizing the first-order conditional estimation method with interaction (FOCEI) objective function.[16] More information on model architecture and training approach is outlined in **Supplementary Material S1** section 1.

The model was fit on data from two clinical trials evaluating the effectiveness of lonoctocog alfa (Afstyla) during prophylactic treatment, kindly provided by CSL Behring GmbH. The data set included information on the country of residence, age, body weight, height, and VWF:Ag levels of 103 patients with severe hemophilia A followed over a combined total of 133 visits. Dense PK profiles (median of 12 FVIII measurements per visit) were collected for each of the individuals. A two-compartment model was used and random effects were estimated for the CL and V_1 parameters. Combined additive and proportional residual error were assumed. Covariates were selected based on direct causal relationships in the DAG, avoiding confounders.

A subset of the patients also received octocog alfa (Advate, $n = 27$). This enabled us to learn a conversion from lonoctocog alfa PK parameters to octocog alfa parameters. It was assumed that any disparities in PK followed from differences in the specific concentrate administered, rather than the effect of the covariates. First, individual estimates of the PK parameters were obtained based on the lonoctocog alfa data. A Bayesian model was then used to obtain posterior distributions over the proportional change in these parameters when predicting octocog alfa levels.

Finally, because both the one-stage and chromogenic assay were used to measure FVIII levels, an assay conversion model could be developed for both lonoctocog alfa and octocog alfa. An exponential model was used to transform chromogenic assay measurements to corresponding one-stage assay measurements.

Generative models

We make the distinction between two different types of generative models: those with a covariate-focus and those with a data set focus. The former attempts to describe covariate relationships shared between data sets and is suited for data imputation and for estimating downstream effects of "do expressions" (e.g., estimating the increase in height and weight of a child aging 2 years) following from the causal graph. In contrast, generative models with a data set focus aim to produce virtual patients similar to the real patients. These models do not necessarily rely on a DAG are not suited for data imputation.

Covariate-focus generative model

Public data sets were collected in order to describe the relationships between each of the covariates. Information on the relationship between body weight, height, and age was obtained for 1,635 men from the National Health and Nutrition Examination Survey (NHANES) data set.[17] Publicly available data on VWF:Ag were extracted from several publications using WebPlotDigitizer (Rohatgi A., version 4.6).[8,18] A total of 870 VWF:Ag levels with corresponding patient age and blood group were available. Depending on the complexity of the relationships, different probabilistic ML models were fit based on the DAG to learn each of the conditional distributions. Heteroscedastic noise was assumed in all models. More details can be found in **Supplementary Material** S1 section 2.

Data set specific generative model

A generative model was developed for the data from the lonoctocog alfa data set. To this end, neural spline models were fit to learn the joint distribution over patient age, weight, height, and VWF levels. A large, curated data set of virtual patients is shared alongside model code.

Model evaluation

Accuracy of the generative model with covariate-focus was evaluated using the lonoctocog alfa data in two scenarios: (1) data on VWF levels were missing and (2) only data on patient age was available. The first scenario represents data frequently unavailable in the clinical setting,

whereas scenario two reflects an extreme setting where none of the covariates used in the PK model are available. Two approaches for data generation were compared. In the first approach, data were generated *a priori* based on the median of the prior distributions. Because data on blood group was unavailable in the lonoctocog alfa data set, predictions were compared assuming that all patients either had blood group O or non-O. In the second approach, a Bayesian model was implemented to produce posterior distributions of the missing covariates and random effect parameters based on observed FVIII levels. Here, the prior distribution for VWF:Ag was implemented as a mixture distribution indexed by blood group. As a result, the model also obtains a posterior probability of the patient having blood group O. Again, posterior median was collected. Accuracy of the generated covariates was evaluated using the mean absolute percentage error (MAPE).

Performance of the predictive model was validated on an external dataset of FVIII PK profiles collected for patients with moderate and severe hemophilia A ($n = 40$) during the OPTI-CLOT clinical trial.[19] Only data from patients who received octocog alfa and turoctocog alfa (NovoEight; similar PK as octocog alfa[20]) were used. The data set contained information on patient age, weight, height, blood group, and VWF:Ag levels. VWF levels were available for 16 patients. Missing values were imputed using the generative model using the *a priori* approach. A median of 3 FVIII measurements were available per patient, collected roughly 4, 24, and 48 hours after dose. The one-stage assay was used to measure FVIII levels. Predictions from the PK model were thus converted from chromogenic to one-stage levels using the assay conversion model. Model performance was compared with four representative PK models trained on one-stage assay data of octocog alfa, with two models also trained on other concentrates.[1,21–23] Predictive performance was represented by the root mean squared error (RMSE), mean error (ME), and coefficient of determination (R^2).

Model code

Models were implemented in the Julia programming language (version 1.8.3) with the DifferentialEquations.jl package as a main dependency.[24] All relevant model code (including generative models) is available at https://github.com/Janssena/DeepFVIII.jl.

RESULTS

An overview of the patient characteristics for the lonoctocog alfa data set and the OPTI-CLOT data set are shown in **Table 1**. Importantly, data on VWF levels were missing for more than half of patients (24/40) in the test data set.

A deep ensemble of DCMs was fit to predict lonoctocog alfa levels measured using the chromogenic assay. The final model included the effect of FFM on CL and V_1 and the effect of VWF on CL. The DAG is shown in **Figure 1**. The validation set RMSE of median typical predictions from the deep ensemble was 11.0 ± 1.1 IU/dL. Coefficient of variation of random effects on CL and V_1 were 23% and 18%, respectively ($CV(\%) = \sqrt{\exp(\omega^2) - 1} \times 100$). Estimated standard deviation of additive error was 1.3 IU/dL and the estimate of proportional error was 8.4%.

Learned functions could be visualized and matched expectations about the causal effect of the covariates (see **Figure 2**). Investigations on alternative hypotheses supported the proposed final model (see **Table S1**).

Next, the conversion model was created to adjust individual lonoctocog alfa PK parameters to octocog alfa PK parameters. Estimated CL of octocog alfa was increased by 15% (95% credible interval (CrI): 13–17), V_1 was decreased by 19% (95% CrI: 16–23), Q was decreased by 74% (95% CrI: 49–83), and V_2 was 223% higher (95% CrI: 193–253). The learned correction factors led to very accurate predictions using the random effect estimates for rFVIII-SingleChain in all but one patient (see **Figure S9**). The conversion of chromogenic assay levels to one-stage assay levels was represented by the following equation:

$$\text{osa} = \max\left(0, \frac{-3.07 + 4.76 \cdot \text{csa}^{0.66}}{2.10^{\text{Lonoctocog alfa}}}\right) \quad (2)$$

Table 1 Patient characteristics

	Training data		Test data		
	Lonoctocog alfa	Octocog alfa	Overall	Octocog alfa	Turoctocog alfa
n	103	27	40	19	21
Age in years median [range]	26 [1–60]	32 [19–60]	49 [18–77]	48 [18–77]	49 [21–77]
Height in cm median [range]	172 [84–194]	178 [163–190]	182 [148–198]	183 [143–195]	179 [170–198]
Weight in kg median [range]	68 [12–112]	77 [59–100]	89 [61–134]	88 [61–133]	95 [63–134]
BMI median [range]	21 [13–37]	25 [19–30]	27 [19–43]	27 [19–36]	27 [21–43]
Fat-free mass in kg median [range]	55 [9.6–75]	59 [50–72]	66 [44–85]	66 [44–85]	67 [52–78]
Blood group O	missing	missing	63%	53%	71%
VWF:Ag median [range] (% missing)	114 [42.7–296] (0%)	125 [73–242] (0%)	115 [73–225] (60%)	141 [108–222] (63%)	106 [73–225] (57%)
Number of FVIII measurements (median)	1,465 (12)	292 (11)	125 (3)	57 (3)	68 (3)
Assay	One-stage + chromogenic		One-stage		
Reagent	Pathromtin SL + Coamatic test kit		Treatment center specific (unspecified)		

BMI, body mass index; FVIII, factor VIII; VWF:Ag, von Willebrand factor antigen.

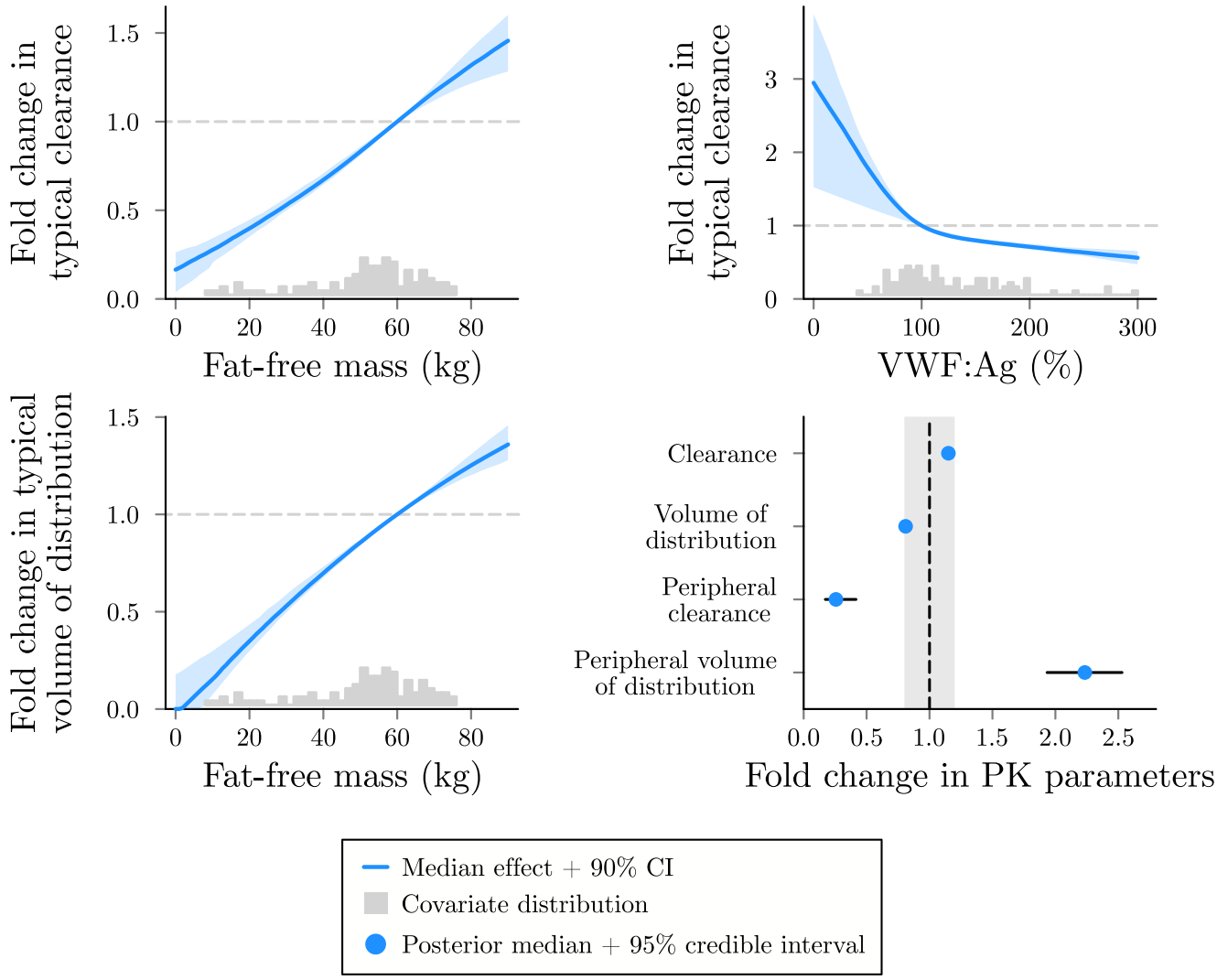

Figure 2 Visualizations of learned covariate effects. Each line depicts the median effect over the predictions from the deep ensemble, along with its 90% CI. Histograms represent the distribution of the observed covariates. In the bottom right, the median and its 95% credible interval from the posterior distributions of the difference in PK parameters between lonoctocog alfa and rFVIII are shown. The shaded area covers a <20% change in the PK parameter value. CI, confidence interval; PK, pharmacokinetic; rFVIII, recombinant factor VIII; VWF, von Willebrand factor.

After applying the PK and the assay conversion, test error on the external data set was slightly higher compared with accuracy on the train set (RMSE = 14.6 IU/dL, $R^2 = 0.90$). The RMSE of typical predictions from our model was lower compared with three of the previously published models[1,21,22] (mean RMSE = 17.7 IU/dL; see **Table 2**), whose predictions also presented a slightly higher degree of bias (ME of 3.81 vs. 1.50 IU/dL). The most accurate alternative[23] depicts similar performance to our model (RMSE = 15.4 IU/dL, $R^2 = 0.89$).

Finally, the accuracy of the generative model was evaluated in the two missing data scenarios (see **Table 3**). The Bayesian

Table 2 Accuracy of population PK models

Model	Training data	RMSE of typical predictions (IU/dL)	ME of typical predictions (IU/dL)	R^2
Bjorkman *et al.*[1]	Octocog alfa and plasma-derived FVIII	16.6	3.85	0.87
Nesterov *et al.*[16,21]	Octocog alfa	17.6	3.82	0.85
McEneny-King *et al.*[22]	Octocog alfa and other SHL	19.0	3.76	0.86
Allard *et al.*[23]	Octocog alfa and other SHL	15.4	1.13	0.89
Causal DCM (ours)	Lonoctocog alfa	14.6	1.50	0.90

DCM, deep compartment model; FVIII, factor VIII; ME, mean error; PK, pharmacokinetic; R^2, coefficient of determination; RMSE, root mean squared error; SHL, standard half-life.
Root mean squared error, mean error, and coefficient of determination for each of the models on the test set are shown.

Table 3 Accuracy of the generative model

Scenario	Approach	MAPE (%) ± SD			
		Height	Weight	FFM	VWF:Ag (assumed BG)
VWF:Ag (and blood group) missing	*a priori*	–	–	–	30.0 ± 25 (non-O) 32.1 ± 19 (O)
	Bayesian	–	–	–	**17.6 ± 14**
All PK model covariates missing	*a priori*	4.3 ± 3.4	25.5 ± 24	16.8 ± 16	30.0 ± 25 (non-O) 32.1 ± 19 (O)
	Bayesian	**3.9 ± 3.1**	**22.4 ± 22**	**14.7 ± 14**	**17.9 ± 15**

BG, blood group; FFM, fat-free mass; MAPE, mean absolute percentage error; PK, pharmacokinetic; SD, standard deviation; VWF:Ag, von Willebrand factor antigen.
The average mean absolute percentage error between the true and generated covariate values along with its standard deviation is shown. Bold text indicates the most accurate model in each of the two scenarios.

approach outperformed the *a priori* approach in terms of MAPE in all cases. When using the *a priori* approach to impute VWF levels, MAPE of predictions was 30.0% when assuming all individuals had blood group non-O and 32.1% when assuming blood group O. The MAPE of the median VWF:Ag levels obtained from the Bayesian approach was 17.6%. Overall, imputation of height was the most accurate (MAPE of 3.9–4.3%), with imputation of body weight having relatively high error (MAPE of 22.4–25.5%). Interestingly, the MAPE of imputed VWF:Ag levels was similar in both missing data scenarios (MAPE of 17.6% and 17.9%).

DISCUSSION

In this work, we aimed to develop a population PK model that follows techniques from causal inference. First, relationships of relevant variables and potential confounders were described using a DAG. The graph supports the selection of important covariates to include in the PK model while offering a natural way to interpret consequences of interventions on any of the variables. Next, a hybrid ML/PK model was fit to predict lonoctocog alfa levels measured using the chromogenic assay. Because part of the patients in the data set also received octocog alfa shortly before their lonoctocog alfa PK profile was taken, the model could be extended to correct for the difference in PK between these two concentrates. By estimating the difference with respect to the individual PK parameters estimates for lonoctocog alfa, we simulate the intervention of only changing the FVIII concentrate. The resulting predictions for octocog alfa were highly accurate based on a proportional change in the PK parameters. Only for a single patient were discordant results observed, potentially as a result of an unseen variable that specifically affects the PK of octocog alfa (e.g., rFVIII specific inhibitors).

We then determined the generalization capacity of the model by comparing the error to predictions from previous PK models on data of patients who had received octocog alfa and turoctocog alfa measured using the one-stage assay. Predictions from our model thus needed to be corrected for differences between FVIII concentrates as well as the measurement assay used. Nonetheless, our model presented lower RMSE compared with three of the previous models (with roughly similar performance to the most accurate alternative), even though an important covariate – VWF:Ag – was missing in more than half of the patients. Although it is difficult to determine the clinical impact with respect to prediction accuracy, it is encouraging that we obtained at worst similar accuracy to models specifically trained on data of a different rFVIII concentrate and measurement assay.

To support the model in settings involving missing data, we augmented the model with a generative model which reproduces the data based on the DAG. Evaluations of this model depicted good imputation performance, with < 18% error when imputing VWF:Ag levels in the lonoctocog alfa data set. This model even provided accurate (< 18% error) predictions of PK model covariates in a very limited setting when only patient age was known.

The above results indicate the benefit of viewing PK model development through a causal lens. The main applied tool of causal inference involved using a DAG to describe the relationships of relevant variables. In the graph, we assumed that any causal effect of age and blood group are largely mediated through VWF levels. Our results show that these covariates were largely uncorrelated to the PK parameters when VWF:Ag was already included in the model (see **Figure S6**). It has already been extensively reported that VWF:Ag levels are lower in individuals with blood group O.[9] Similarly, higher age correlates with an increase in VWF levels.[25] Interestingly, this relationship disappeared when correcting for the presence of specific comorbidities, which we included in the DAG.[26] We explicitly specify that VWF levels are partially observed, as these levels can vary over time related to factors such as stress. Relatively recent VWF levels might thus be necessary to correctly estimate the causal effect of interventions in the graph. The same applies to the individual estimates of the random effects.

In the PK model, we used an estimate of FFM to affect FVIII CL and V_1 rather than body weight. Although the use of body weight depicted similar predictive performance, the uncertainty of the learned functions was higher. Additionally, the functions seemed to indicate the model implicitly learning a measure of lean body mass as the function flattened at higher body weight (see **Figure S7**). These findings support the observation that body weight correlates poorly with the PK of rFVIII at higher body mass index (BMI).[27] A relevant assumption in the model was that Q and V_2 were not affected by any covariates. It is common in PK models to implement allometric scaling of these parameters. In our analysis, we did not find that adding the effect of FFM on Q and V_2 improved model accuracy. Additionally,

uncertainty in the learned functions was again large when their effects were added, discouraging its inclusion in the model. Alternatively, we included the effect of differences between rFVIII concentrates on all PK parameters (rather than on a single parameter). The model produced accurate predictions for turoctocog alfa after correcting for octocog alfa PK, suggesting that it might not be necessary to correct for each specific molecular formulation of FVIII.

The final component of the proposed DAG deals with variables that affect the measurement of FVIII levels. Corrections for discrepancies between assays are rarely described in detail by FVIII population PK models. There do exist models that incorporate such corrections,[6,28] or that correct for differences in measured FVIII levels between treatment centers (potentially related to the use of different reagents).[29] Although we do describe several sources of variability affecting FVIII measurements, we did not describe most of their potential effects in the current work due to limitations of the available data. Examples of additional sources of variability include different assay reagents, or bias arising from incompatibilities between specific assays and certain FVIII concentrates.[30] In order to correct for such biases, it might be necessary to develop models on multiple data sets which should be explored in future work.

A novel element of the current work is the addition of a generative model to support population PK models. Differences in covariate availability can complicate the implementation of PK models in clinical practice. Generative models can be used to impute missing values or to simulate realistic patients. Additionally, these models can be used to learn the joint distribution over the covariates with respect to a specific data set. When encountering new data, these joint distributions can be used to identify out-of-distribution samples for which the model might not be appropriate. Additionally, it allows models to continue training on new data, where new covariate effects are learned in regions where the model does not yet have sufficient support. Such an approach is an essential component of the Bayesian paradigm, where model priors are used in sequential studies to iteratively update the posterior. PK models can be trained locally, whereas model parameters can be shared, keeping actual patient data private. The use of automatic ML models greatly support such an approach, whereas the use of interpretable models proposed in the current work enable the identification of model bias and errors. Concrete examples of additional use cases of our approach include the sharing of synthetic data with outcomes to pool information on risk profiles for different mutations in rare cancers, or to continuously refine a PK model for vancomycin on specific patient populations,[31] utilizing information from previous studies.

There were also some limitations of the current study. The proposed PK model was mainly trained on a population of adult patients, and thus might not be appropriate for pediatric patients. Next, the models (including the previous population PK models) depicted an underestimation of octocog alfa peak levels in the OPTI-CLOT data set. This effect was not seen when making predictions for the subset of patients who received octocog alfa in the training data set. It is possible that differences between the used assay or patient population (e.g., higher BMI in the OPTI-CLOT data set) influenced the results. It is important that generative models are developed on large, representative data sets to reduce model bias when imputing missing values. The availability of sufficiently large data sets can be an issue, also for the development of data set specific generative models. Next, although not necessarily specified in the DAG, we chose to represent the effect of VWF levels using VWF:Ag, because public data on VWF:act levels was scarce. It is unknown whether the relative amount of VWF or its FVIII binding activity is more relevant for FVIII clearance. A combination of both quantities might be a more accurate representation of the effect of VWF. Finally, description of a comprehensive causal DAG might be complicated for some drugs, potentially making the proposed approach difficult to implement. In some cases, the DAG might contain several variables that are either rarely measured or difficult to determine even in an experimental setting. Although there might then not seem to be much benefit to the creation of a DAG, it can nonetheless be of use to identify confounders or to quantify a degree of uncertainty in the downstream effect prediction when data are scarce.

In conclusion, we present a hybrid ML/PK model utilizing causal inference techniques to predict FVIII levels in patients with hemophilia A. The model accurately extrapolated to a different FVIII concentrate and measurement assay in an external data set. By using probabilistic models to learn the data generating process, the proposed approach can also be used to generate missing data and simulate realistic virtual patients. Additionally, by sharing these generative models, information on otherwise sensitive data can still be made publicly available. The approach introduces an interesting new paradigm for the continuous refinement of population PK models.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

ACKNOWLEDGMENTS

SYMPHONY consortium: The SYMPHONY consortium which aims to orchestrate personalized treatment in patients with bleeding disorders, is a unique collaboration between patients, healthcare professionals, and translational and fundamental researchers specialized in inherited bleeding disorders, as well as experts from multiple disciplines. It aims to identify best treatment choice for each individual based on bleeding phenotype. To achieve this goal, workpackages have been organized according to 3 themes, for example, Diagnostics (workpackages 3 and 4), Treatment (workpackages 5-9), and Fundamental Research (workpackages 10–12). This research receives funding from the Netherlands Organization for Scientific Research (NWO) in the framework of the NWA-ORC Call grant agreement NWA.1160.18.038. Principal investigator: Dr. M.H. Cnossen. Project coordinator: Dr. S.H. Reitsma.

Beneficiaries of the SYMPHONY consortium: Erasmus University Medical Center-Sophia Children's Hospital, project leadership and coordination; Sanquin Diagnostics; Sanquin Research; Amsterdam University Medical Centers; University Medical Center Groningen; University Medical Center Utrecht; Leiden University Medical Center; Radboud University Medical Center; Netherlands Society of Hemophilia Patients (NVHP); Netherlands Society for Thrombosis and Hemostasis (NVTH); Bayer B.V., CSL Behring B.V., Swedish Orphan Biovitrum (Belgium) BVBA/SPRL. Additional beneficiaries, not included in the SYMPHONY consortium, currently funding parallel projects are: Novonordisk (OPTI-CLOT TARGET), Roche (Partitura), Stichting Hemophilia (patient-reported outcomes project).

OPTI-CLOT study group: OPTI-CLOT/To WiN study group aims to implement personalized treatment by pharmacometric-guided dosing of factor concentrates, desmopressin and nonfactor therapies in patients with bleeding disorders.

OPTI-CLOT/To WiN Steering Committee, the Netherlands: M.H. Cnossen (principal Investigator & chair OPTI-CLOT-To WiN) and R.A.A. Mathôt (coinvestigator). F.W.G. Leebeek, Rotterdam; M. Coppens, K. Fijnvandraat, Amsterdam; K. Meijer, Groningen, S.E.M. Schols, Nijmegen; H.C.J. Eikenboom, Leiden; R.E.G. Schutgens, Utrecht; F. Heubel-Moenen, Maastricht; L. Nieuwenhuizen, Veldhoven; P. Ypma, The Hague; M.H.E. Driessens, Nijkerk.

Trial bureau: I. van Vliet, Rotterdam.

Local collaborators the Netherlands: M.J.H.A. Kruip, S. Polinder, Rotterdam; P. Brons, Nijmegen; F.J.M. van der Meer, Leiden; K. Fischer, K. van Galen, Utrecht.

Principal investigators and local collaborators in the UK: P. W. Collins, Cardiff; M. Mathias, P. Chowdary, London; D. Keeling, Oxford.

OPTI-CLOT-To WiN, DAVID and SYMPHONY PhDs: PhDs: J. Lock, H.C.A.M. Hazendonk, T. Preijers, N.C.B. de Jager, L. Schutte, L.H. Bukkems.

PhDs ongoing: M.C.H.J. Goedhart, J.M. Heijdra, L. Romano, W. Al Arashi, M.E. Cloesmeijer, A. Janssen, S.F. Koopman, C. Mussert.

FUNDING

This study is supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO; Dutch Research Council) under grant agreement NWA.1160.18.038.

CONFLICT OF INTEREST

M.H.C.'s institution has received investigator-initiated research and travel grants as well as speaker fees over the years from the Netherlands Organization for Scientific Research (NWO) and Netherlands National research Agenda (NWA), the Netherlands Organization for Health Research and Development (ZonMw), the Dutch Innovatiefonds Zorgverzekeraars, Baxter/Baxalta/Shire/Takeda, Pfizer, Bayer Schering Pharma, CSL Behring, Sobi Biogen, Novo Nordisk, Novartis and Nordic Pharma and for serving as a steering board member for Roche, Bayer and Novartis for which fees go to the Erasmus MC as an institution. R.A.A.M. has received grants from governmental and societal research institutes such as NWO, ZonMW, Dutch Kidney Foundation and Innovation Fund and unrestricted investigator research grants from Baxter/Baxalta/Shire/Takeda, Bayer, CSL Behring, Sobi, and CelltrionHC. He has served as advisor for Bayer, CSL Behring, Merck Sharp & Dohme, and Baxter/Baxalta/Shire/Takeda. All grants and fees paid to the institution. All other authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

A.J., L.S., F.C.B., M.H., and R.A.A.M wrote the manuscript. A.J., L.S., F.C.B., and R.A.A.M. designed the research. A.J. performed the research. A.J. analyzed the data.

1. Björkman, S. *et al*. Population pharmacokinetics of recombinant factor VIII: the relationships of pharmacokinetics to age and body weight. *Blood* **119**, 612–618 (2012).
2. Turecek, P.L., Johnsen, J.M., Pipe, S.W. & O'Donnell, J.S. iPATH biological mechanisms underlying inter-individual variation in factor VIII clearance in haemophilia. *Haemophilia* **26**, 575–583 (2020).
3. Goedhart, T.M. *et al*. Population pharmacokinetic modeling of factor concentrates in hemophilia: an overview and evaluation of best practice. *Blood Adv.* **5**, 4314–4325 (2021).
4. Zhu, J. *et al*. Pharmacokinetics of perioperative FVIII in adult patients with haemophilia a: an external validation and development of an alternative population pharmacokinetic model. *Haemophilia* **27**, 974–983 (2021).
5. Preijers, T. *et al*. Validation of a perioperative population factor VIII pharmacokinetic model with a large cohort of pediatric hemophilia a patients. *Br. J. Clin. Pharmacol.* **87**, 4408–4420 (2021).
6. Bukkems, L.H. *et al*. A novel, enriched population pharmacokinetic model for recombinant factor VIII-fc fusion protein concentrate in hemophilia a patients. *Thromb. Haemost.* **120**, 747–757 (2020).
7. Rogers, J.A., Maas, H. & Pitarch, A.P. An introduction to causal inference for pharmacometricians. *CPT Pharmacometrics Syst. Pharmacol.* **12**, 27–40 (2023).
8. Biguzzi, E. *et al*. Rise of levels of von willebrand factor and factor VIII with age: role of genetic and acquired risk factors. *Thromb. Res.* **197**, 172–178 (2021).
9. Ward, S.E., O'Sullivan, J.M. & O'Donnell, J.S. The relationship between ABO blood group, von willebrand factor, and primary hemostasis. *Blood* **136**, 2864–2874 (2020).
10. Kahlon, A. *et al*. Quantification of perioperative changes in von willebrand factor and factor VIII during elective orthopaedic surgery in normal individuals. *Haemophilia* **19**, 758–764 (2013).
11. van Moort, I. *et al*. Von willebrand factor and factor VIII clearance in perioperative hemophilia a patients. *Thromb. Haemost.* **120**, 1056–1065 (2020).
12. Al-Sallami, H.S. *et al*. Prediction of fat-free mass in children. *Clin. Pharmacokinet.* **54**, 1169–1178 (2015).
13. St. Ledger, K. *et al*. International comparative field study evaluating the assay performance of AFSTYLA in plasma samples at clinical hemostasis laboratories. *J. Thromb. Haemost.* **16**, 555–564 (2018).
14. Janssen, A. *et al*. Deep compartment models: a deep learning approach for the reliable prediction of time-series data in pharmacokinetic modeling. *CPT: Pharm. Syst. Pharmacol.* **11**, 934–945 (2022).
15. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Proces. Syst.* **30** (2017).
16. Janssen, A., Leebeek, F., Cnossen, M. & Mathôt, R. The neural mixed effects algorithm: Leveraging machine learning for pharmacokinetic modelling. In *Proceedings of the 29th annual meeting of the population approach group in europe. abstract* **9826** (2021).
17. Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS). National health and nutrition examination survey data (2013) https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2013. Accessed November 11, 2022.
18. Davies, J.A., Collins, P.W., Hathaway, L.S. & Bowen, D.J. Effect of von willebrand factor y/C1584 on in vivo protein level and function and interaction with ABO blood group. *Blood* **109**, 2840–2846 (2007).
19. van Moort, I. *et al*. Perioperative pharmacokinetic-guided factor VIII concentrate dosing in haemophilia (OPTI-CLOT trial): an open-label, multicentre, randomised, controlled trial. *The Lancet Haematology* **8**, e492–e502 (2021).
20. Viuff, D., Barrowcliffe, T., Saugstrup, T., Ezban, M. & Lillicrap, D. International comparative field study of N8 evaluating factor VIII assay performance. *Haemophilia* **17**, 695–702 (2011).
21. Nestorov, I., Neelakantan, S., Ludden, T.M., Li, S., Jiang, H. & Rogge, M. Population pharmacokinetics of recombinant factor VIII fc fusion protein. *Clin. Pharmacol. Drug Dev.* **4**, 163–174 (2015).
22. McEneny-King, A., Chelle, P., Foster, G., Keepanasseril, A., Iorio, A. & Edginton, A.N. Development and evaluation of a generic population pharmacokinetic model for standard half-life factor VIII for use in dose individualization. *J. Pharmacokinet. Pharmacodyn.* **46**, 411–426 (2019).

23. Allard, Q. *et al*. Real life population pharmacokinetics modelling of eight factors VIII in patients with severe haemophilia a: is it always relevant to switch to an extended half-life? *Pharmaceutics* **12**, 380 (2020).
24. Rackauckas, C. & Nie, Q. Differentialequations. Jl–a performant and feature-rich ecosystem for solving differential equations in julia. *J. Open Res. Softw.* **5**, 1 (2017).
25. Favaloro, E.J., Franchini, M. & Lippi, G. Aging hemostasis: changes to laboratory markers of hemostasis as we age—a narrative review. *Semin Thromb Hemost* **40**, 621–633 (2014).
26. Atiq, F. *et al*. Comorbidities associated with higher von willebrand factor (VWF) levels may explain the age-related increase of VWF in von willebrand disease. *Br. J. Haematol.* **182**, 93–105 (2018).
27. van Moort, I. *et al*. Dosing of factor VIII concentrate by ideal body weight is more accurate in overweight and obese haemophilia a patients. *Br. J. Clin. Pharmacol.* **87**, 2602–2613 (2021).
28. Abrantes, J.A., Nielsen, E.I., Korth-Bradley, J., Harnisch, L. & Jönsson, S. Elucidation of factor VIII activity pharmacokinetics: a pooled population analysis in patients with hemophilia a treated with moroctocog alfa. *Clin. Pharmacol. Therap.* **102**, 977–988 (2017).
29. Hazendonk, H. *et al*. A population pharmacokinetic model for perioperative dosing of factor VIII in hemophilia a patients. *Haematologica* **101**, 1159–1169 (2016).
30. Pouplard, C. *et al*. The use of the new ReFacto AF laboratory standard allows reliable measurement of FVIII: C levels in ReFacto AF mock plasma samples by a one-stage clotting assay. *Haemophilia* **17**, e958–e962 (2011).
31. Monteiro, J.F., Hahn, S.R., Gonçalves, J. & Fresco, P. Vancomycin therapeutic drug monitoring and population pharmacokinetic models in special patient subpopulations. *Pharmacol. Res. Perspect.* **6**, e00420 (2018).