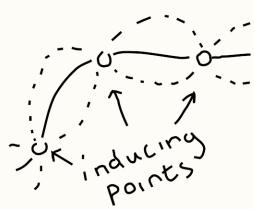


$$x(t) \rightarrow \text{NeuralODE} \rightarrow z(t)$$

"latent effect" Flows $\hookrightarrow z_0 \leftarrow z_1 \dots \leftarrow z_k$

$$\frac{1}{M} \sum_{i=1}^M \log p(y_i | n_i)$$



Alexander Janssen

two compartmental models

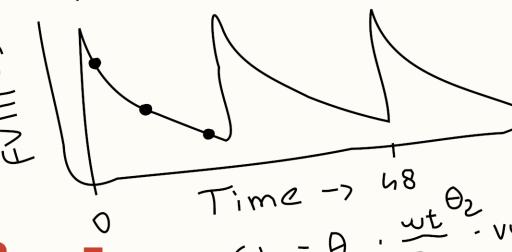
$$dA_1 = \frac{I}{V} + k_{21} A_2$$

$$dA_2 = k_{12} A_1$$

$$KL[q(u) || p(u|y)]$$

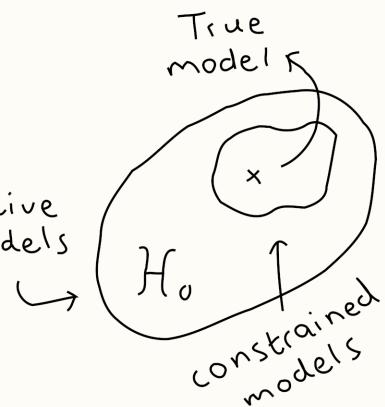
How to deal with overfitting?

3 samples?



Artificial Intelligence

for intelligent care



$$p(y) = \text{intractable}$$

$$p_K = f(x)$$

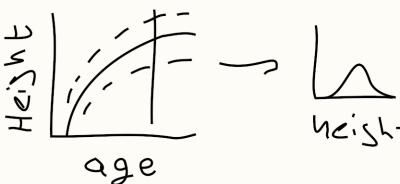
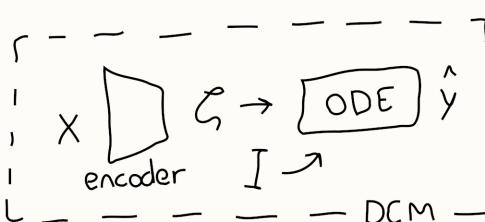
$$f(\underline{x}) \sim GP$$

$$\text{ELBO} = \mathbb{E}_{z \sim q} [\log p(y|z)] + \underbrace{\mathbb{H}[q_\phi(z)]}_{\text{entropy of variational distribution}}$$

$$x_1 \times \square \quad x_2 \circ - \zeta \quad x_3 \triangleright 0$$

How machine learning algorithms can enhance the personalised treatment of patients with haemophilia A.

$$\eta \sim \mathcal{N}(0, \Sigma)$$



$$(a) \rightarrow (h) \rightarrow (\omega)$$

$$-2 \log p(\theta, \omega | y_i) = -2 \log \int p(\theta, \omega | y_i) d\theta d\omega \approx \Phi(\omega) - \log S$$

ARTIFICIAL INTELLIGENCE FOR INTELLIGENT CARE

*How machine learning algorithms can enhance the personalised
treatment of patients with haemophilia A.*

ALEXANDER JANSSEN

COLOPHON

The studies performed in this thesis were financially supported by the Dutch Research Council (Nederlandse Organisatie voor Wetenschappelijk Onderzoek; NWO) under grant agreement NWA.1160. 18.038. The copyright of the articles that have been published or accepted for publication has been transferred to the respective journals. No part of this thesis may be reproduced without the permission of the author.

Financial support for the printing of this thesis was kindly provided by CSL Behring and Amsterdam UMC.

This basic structure of this document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*".

Cover design and formatting: Alexander Janssen,
Printing: Ridderprint, www.ridderprint.nl.

Final Version as of December 10, 2024.

Artificial Intelligence for intelligent care
How machine learning algorithms can enhance the personalised treatment of patients with
haemophilia A

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op donderdag 23 januari 2025, te 11.00 uur

door Alexander Janssen
geboren te Amsterdam

Promotiecommissie

<i>Promotores:</i>	prof. dr. R.A.A. Mathôt prof. dr. M.H. Cnossen	AMC-UvA Erasmus Universiteit Rotterdam
<i>Copromotores:</i>	dr. F.C. Bennis	AMC-UvA
<i>Overige leden:</i>	prof. dr. P. Liò prof. dr. P.H. van der Graaf dr. P.W.G. Elbers prof. dr. A. Abu-Hanna dr. S.C. Gouw prof. dr. A.H. Zwinderman	University of Cambridge Universiteit Leiden Vrije Universiteit Amsterdam AMC-UvA AMC-UvA AMC-UvA

Faculteit der Geneeskunde

CONTENTS

1	General introduction and thesis outline	1
I	Machine learning in pharmacometrics	
2	Adoption of machine learning in pharmacometrics: an overview of recent implementations and their considerations	27
3	Application of SHAP values for inferring the optimal functional form of covariates in pharmacokinetic modelling	73
II	Deep compartment models	
4	Deep compartment models: a deep learning approach for the reliable prediction of time-series data in pharmacokinetic modelling	99
5	On inductive biases for the robust and interpretable prediction of drug concentrations using deep compartment models	121
6	Mixed effect estimation in deep compartment models: Variational methods outperform first-order approximations	159
III	Applying machine learning to improve treatment of haemophilia A	
7	A generative and causal pharmacokinetic model for factor VIII in haemophilia A: a machine learning framework for continuous model refinement	201
8	Variable pharmacokinetics of coagulation factor VIII in the perioperative setting complicates personalisation of treatment in patients with haemophilia A	229
9	A repeated time-to-event model for personalised treatment of patients with haemophilia A based on individual bleeding risk	261
IV	The OPTI-CLOT web-portal	
10	Individualised treatment with a personal touch: introducing the OPTI-CLOT web-portal, an open web-application implementing PK-guided dosing in patients with rare bleeding disorders	289
V	Discussion	
11	General discussion and perspectives	305
VI	Appendix	
	Data Availability	327
	Summary / Samenvatting	329
	About the Author / PhD Portfolio / Publications	345
	Dankwoord / Acknowledgements	353

GENERAL INTRODUCTION AND THESIS OUTLINE

1.1 HAEMOPHILIA A

1.1.1 *Background*

Bleeding disorders are rare conditions that are caused by a deficiency or qualitative defect of platelets or coagulation factors. These disorders involve a disruption in the process of coagulation, known as *haemostasis*, and can be differentiated into primary or secondary haemostatic disorders, fibrinolytic disorders, and bleeding disorders of unknown cause (where the precise aetiology has not (yet) been deciphered). The most well-known bleeding disorders are haemophilia A (deficiency of factor VIII; FVIII), haemophilia B (deficiency in factor IX; FIX), and von Willebrand disease (deficiency of von Willebrand factor; VWF). Amongst these three, von Willebrand disease is the most common with a prevalence rate of 1 per 100-1,000 individuals, compared to roughly 13 and 3 per 100,000 males for haemophilia A and B, respectively [1, 2].

Patients with haemophilia A have impaired haemostasis, resulting in an elevated risk of (spontaneous) bleeding. Haemophilia A is *X-linked*, meaning that it almost exclusively affects males. The severity of the disorder is characterised in terms of the residual endogenous factor activity level, which is measured in international units (IU) using one stage and chromogenic clotting assays. Patients with mild haemophilia A have endogenous FVIII activity levels of around 5-40 IU/dL, moderate patients have levels between 1-5 IU/dL, and patients with less than 1 IU/dL are classified as having severe haemophilia A. Without adequate treatment, haemophilia A patients present with frequent (spontaneous) bleeding typically in joints and muscles leading to arthropathy, and have an elevated risk of life threatening bleeding events such as gastrointestinal and intracranial bleeding. In general, most moderate and mild haemophilia A patients have a milder bleeding phenotype with bleeding usually occurring following (minor) trauma or dental and other medical procedures.

In this thesis, we focus mainly on data collected from severe and moderate haemophilia A patients with a more severe bleeding phe-

Haemostasis involves the process of forming a clot at the site of blood vessel damage to stop bleeding.

Women have two copies of the X chromosome. Although a deficiency in both F8 genes is unlikely, haemophilia carriers can still present with low FVIII levels and elevated bleeding risk.

notype who require regular treatment in daily life or around medical procedures.

1.1.2 *Evolving treatment of haemophilia A*

For decades, replacement therapy with FVIII concentrates has been the cornerstone of the treatment of haemophilia A (see figure 1.1). Before the 1970s, excessive bleeding could only be treated using whole blood, plasma or cryoprecipitate infusions, which contain only small amounts of coagulation factor [3]. Shortages of blood products meant that most patients were left untreated, resulting in severe blood loss which was often lethal. Management of haemophilia improved in the 1970s, when lyophilized plasma concentrates with a higher concentration of FVIII became more widely available. Unfortunately, the widespread adoption of plasma-derived factor concentrates and lack of appropriate viral screening procedures also resulted in a high rate of infections with blood-borne viruses, such as hepatitis C and human immunodeficiency virus (HIV). By the early 1980s, the majority of haemophilia patients were infected with HIV in western countries [4]. Although viral inactivation techniques and screening procedures in later years greatly improved the safety of these plasma-derived concentrates, recombinant FVIII (rFVIII) concentrates introduced in the 1990s greatly improved the standard of care for haemophilia A patients. Home-based treatment became possible and patients saw great improvements in quality of life. To this day, rFVIII concentrates remain one of the main pillars of haemophilia A treatment. Recent advances have mainly focused on the development of specifically modified rFVIII molecules such as extended half-life (EHL) or VWF-decoupled concentrates (BIVVoo1) as well as non-factor replacement therapy options such as emicizumab [5]. Most of these treatment options promise longer drug half-lives, improving drug efficacy while reducing injection frequency and easing patient burden. Novel drugs such as emicizumab can also be administered subcutaneously, reducing the pain associated with injections.

1.1.3 *Prophylaxis*

Prophylactic treatment involves medical measures taken to prevent disease.

In the 1960s, after positive experiences in Sweden and the Netherlands showing significant reductions in bleeding rates [6, 7], most resource-rich countries adopted *prophylaxis* as the standard treatment approach

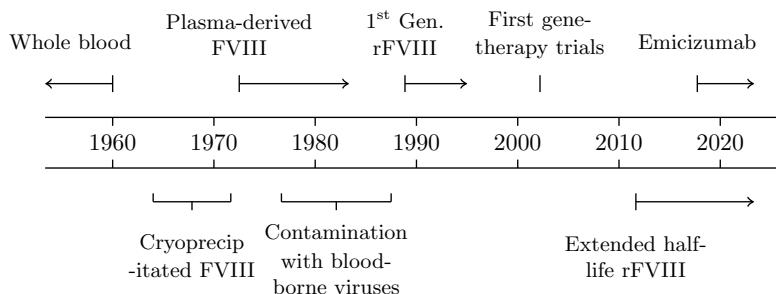


Figure 1.1: Timeline of important events related to the treatment of haemophilia A.

for haemophilia A patients with a severe bleeding phenotype. Early prophylactic treatment was based on the observation that patients with severe haemophilia could be converted to milder variants by keeping FVIII levels above 1 IU/dL [8]. To this end, patients on prophylaxis self-administer FVIII several times per week. Sweden and the Netherlands diverged in their specific approach to prophylaxis: in Sweden, high dose (25-40 IU/kg) and high frequency administration of FVIII (three times weekly) at very early age was recommended [8]. In contrast, Dutch physicians focused on the more individually oriented intermediate approach, where prophylaxis was started after the occurrence of the first bleed and where dosage and frequency were iteratively increased in response to breakthrough bleeding [9]. Comparisons of both approaches depicted only minor differences in bleeding rates and, perhaps more relevant, markedly reduced rFVIII consumption in intermediate prophylaxis regimens (average costs of prophylaxis are roughly €200,000/year per patient) [9–11]. In the present-day, most physicians in resource-rich countries strive for zero bleeds and opt for high-dose prophylaxis. However, the high costs of rFVIII concentrates and prospects of life-long treatment complicates more widespread adoption in resource-limited countries. Focus has thus been placed on a more efficient use of factor concentrates, for example by adapting target FVIII trough levels on an individual basis.

1.1.4 Role of pharmacokinetics in personalisation of prophylaxis

As the half-life of many rFVIII concentrates is relatively short (around 12 hours), maintaining trough levels >1 IU/dL requires relatively high dosage and frequent administration [12]. Unfortunately, the introduction of EHL concentrates (featuring a roughly 1.5-fold increase in half-life) have generally not allowed for a significant reduction in infusion frequency, but do result in higher trough levels [13]. The personalisation of prophylaxis is complicated by the significant inter-individual variability in FVIII exposure when patients are given the same dose per kilogram of body weight [14]. This has prompted the introduction of methods from the field of pharmacometrics to quantify pharmacological differences between patients. Population *pharmacokinetic* (PK) analysis can be applied to individualise drug dosage based on mathematical models [15, 16]. Patients first receive a test dose of a rFVIII concentrate followed by the collection of multiple plasma samples in order to determine individual estimates of so-called PK parameters (e.g. drug clearance and volume of distribution). These parameters can be used to simulate FVIII exposure in response to different prophylaxis schedules on an individualised basis. This way, one can select the optimal treatment regime that achieves pre-specified target levels with acceptable degree of patient burden and rFVIII consumption. Multiple clinical guidelines have since recommended the use of PK-guided dosing for the optimisation of prophylactic treatment regimens [17–19].

Pharmacokinetics describes the processes of drug absorption, distribution, metabolism, and elimination.

1.1.5 Population pharmacokinetics

Several population PK models, most even specific to certain rFVIII concentrates, have since been developed to optimise treatment of haemophilia A patients. Population PK models offer a mathematical representation of the processes of drug absorption, distribution, metabolism, and elimination based on (semi-)mechanistic models (see figure 1.2). Non-linear mixed effect (NLME) modelling is the dominant statistical method for developing population PK models. In mixed effects models, variability between subjects is divided into fixed and random effects. Fixed effects describe the relationship between covariates (e.g. basic patient characteristics such as body weight or age) and the PK parameters using explicit mathematical equations. Random effects describe the remaining inter-individual variability in the parameters, and are used to correct model predictions based on

the measurements. The random effect can for example be thought of as correcting for the effect of unobserved covariates. Finally, NLME models describe the residual error which could for example arise as a consequence of measurement errors, data collection (e.g. inaccurate reporting of dose administration or measurement times), and model misspecification. The resulting population PK models provide prior information regarding the expected FVIII response of typical patients (dashed line in figure 1.2). Accurate individual PK parameters can then be obtained using *maximum a posteriori* (MAP) estimation based on a limited number of plasma samples (solid line in figure 1.2). Three samples collected at 0.5-4, 24, and 48 hours after dose generally are sufficient for standard half-life (SHL) concentrates, significantly reducing patient burden compared to previous rich sampling schemes required during classical PK analyses (5-11 samples) [12, 17, 20, 21].

Maximum a posteriori estimation finds the most probable parameters given the data.

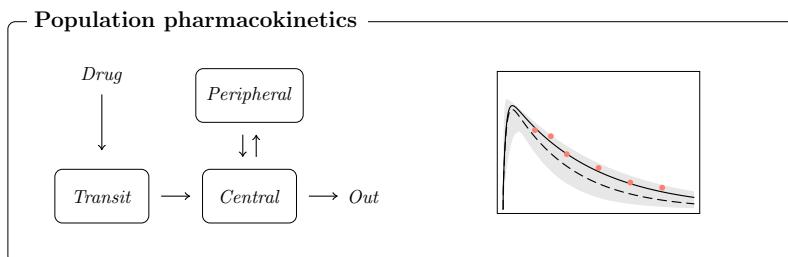


Figure 1.2: Population PK models .

1.1.6 Advances in the treatment of haemophilia A

In the context of haemophilia A, PK-guided dosing often focuses on maintaining FVIII levels above 1 IU/dL. However, many studies have demonstrated considerable differences in bleeding outcomes between patients with similar FVIII exposure [8, 22–25]. Genetic differences, physical activity levels, and risk-taking behaviour all contribute to differences in the individual bleeding phenotype [17]. Personalisation of treatment should thus ideally also focus on optimising *pharmacodynamic* (PD) endpoints. Current research focuses on clinical markers obtained from the thrombin generation assay (TGA) in order to estimate the adequacy of haemostasis. The endogenous thrombin potential (ETP) has been suggested as a promising marker predictive of bleeding outcomes. A recent study found that all patients with ma-

Pharmacodynamics involves the study of physiological effects of drugs on the body.

ajor bleeding had significantly lower ETP levels [26]. However, relevant variability in bleeding outcomes was still seen at the lower range of ETP levels [27, 28]. In addition, appropriate targets for TGA-derived parameters are currently unknown. Further research is thus required before the method is ready for clinical implementation. Since there is a strong need for a PD-based method for optimising treatment, research into alternative methods might be desirable.

As previously mentioned, recent innovations in the treatment of haemophilia A have led to the development of highly effective and more patient-friendly non-factor based therapies. One prominent example is emicizumab, for which prophylactic treatment once every 1-4 weeks was shown to have high efficacy in various patient populations during the HAVEN studies [29]. At the moment, factor concentrates still play a role in prophylaxis and are still widely used for the acute treatment of breakthrough bleeding or to maintain haemostatic balance during and after medical procedures. Since these novel therapies are often more expensive, factor concentrates will also likely remain mainstay of treatment in resource-limited countries. There is however no doubt that the various new therapeutic options for haemophilia will significantly change the clinical landscape in the near future. Since most patients on emicizumab attain impressive bleeding control, future approaches to optimise treatment might not be necessary be focused on improving bleeding outcomes but instead on reducing costs [30].

Gene therapy involves the manipulation of gene expression, for example by introducing genetic material in a patient's cells.

Finally, *gene therapy* has long been suggested as an option for the potential cure of haemophilia A, especially after FIX gene therapy for haemophilia B was successfully developed. Gene therapy in haemophilia A involves the injection of adeno-associated viruses to transfet hepatic cells with a functional version of the F8 gene in an attempt to restore endogenous production of FVIII. In 2022, the first FVIII gene therapy product, Roctavian®, was granted conditional approval for haemophilia A patients on prophylaxis in the European Union. Unfortunately, gene therapy cannot be applied in all patients as those with antibodies against the viral vector are excluded. Moreover, expressed FVIII levels can be seen to diminish in the years following treatment [31, 32]. Some patients also still present with breakthrough bleeding [33]. Repetitive treatment is however complicated by the development of neutralising antibodies, meaning different viral vectors need to be used at every iteration of treatment.

1.2 MACHINE LEARNING

1.2.1 *Background*

Researchers have long been fascinated with the prospects of constructing an artificial intelligence (AI): thinking machines that equal or surpass human capabilities. Recent developments in the field of machine learning had many sceptics reconsider their previous disbelief in the attainability of that goal. Most notably, the introduction of ChatGPT 3.0 to the public in November 2022 had many fear a future where human workers are replaced by machines. Reality is more nuanced; machine learning algorithms like ChatGPT appear highly capable but are frequently found to report incorrect information or cite non-existing sources (a phenomenon since known as *hallucinating*) [34–36]. Nonetheless, the prospects of machine learning in fields like healthcare are extremely promising. Several machine learning algorithms have already been successfully implemented in clinical practice: examples include methods that provide discharge decision support, identify regions-of-interest in magnetic resonance images, or that can detect atrial fibrillation based on data from wearables [37–39]. Most approaches utilise highly specialised algorithms designed to improve performance when working with image (e.g. convolutional methods) or time-series (e.g. recurrent methods) data.

Deep learning is a sub-field of machine learning focusing on neural network-based model architectures. Although most specialised techniques are based on deep learning, classical algorithms such as tree-based models (decision trees and random forests) can sometimes still outperform more complex models, especially when working with relatively small or tabular data sets (e.g. the majority of health record data is in tabular format) [40, 41]. In general, there is no "best" machine learning algorithm, and it is likely the case that the performance of different algorithms is highly dependent on the type of data available. There has also been a gradual shift away from Big Data to *right data*. Especially within the context of rare disease, where data is evidently limited, it is important to collect data of high quality. To successfully implement machine learning methods in this context, algorithms should be able to perform well on smaller data sets, while researchers should be aware of their different qualities and intricacies. In the following sections we discuss the algorithms relevant to this thesis.

The attribution of human-like qualities to AI algorithms is quite common, but frequently considered to be unjustified.

With right data, quality outweighs the importance of sheer volume.

1.2.2 Terminology for the uninitiated

We use the term *algorithm* to denote a model architecture that adheres to specific nomenclature.

Before we start our review of relevant algorithms, it might be useful for those less versed with terminology from the field of mathematics and machine learning to offer a short introduction. First, the words "model", "method", and "*algorithm*" are used somewhat interchangeably in this thesis. A model offers an abstract representation of observable phenomena using mathematical concepts. Generally, a model consists of mathematical equations (i.e. "functions") that have "parameters". The simplest model is the linear model: $y = ax + b$, where $\{a, b\}$ are the parameters. Here, y is our quantity of interest and x is a "covariate", e.g. the predictor that is used to make predictions. Often, we denote our prediction of y with \hat{y} (pronounced as "y-hat"), to indicate that we do not expect it to be exactly the same as the true observed value (for example due to noise). Machine learning algorithms have a specific model architecture that is "data-driven" rather than "hypothesis-driven". One way to think about this is these algorithms have a generic structure that can learn the relationships between the covariates and the outcome of interest from the data. The user thus does not need to specify an explicit model structure.

Whenever we speak of "optimisation" or "learning", the goal is find the value of the parameters that results in the lowest error of predictions. This error is represented by the "objective function". This function can be as simple as the residual error ($y - \hat{y}$) or more complicated, for example by using likelihood functions that also take into account model complexity and prediction uncertainty. Model development includes defining model architecture (or several alternatives based on multiple hypotheses), selection of covariates, parameter optimisation based on the objective function, and finally some sort of model evaluation (for example by comparing learned effects with our expectation of reality). This last step can be complicated for *black-box* methods (which most machine learning algorithms are), as it impossible to infer exactly what such models are doing.

A black-box refers to a system that produces output based on input, but whose inner workings are opaque.

1.2.3 Random forests

Random forests are an extension of the decision tree model (see figure 1.3). A decision tree learns to organise data into bins (e.g. patients with age < 40 or body weight > 70kg) in a hierarchical manner. At each level, the model divides the data into two groups based on a cut-off point in one of the covariates. Increasing the depth of the decision tree

generally improves predictive performance, but might result in overly complex trees that generalise poorly (i.e. have low accuracy on new data).

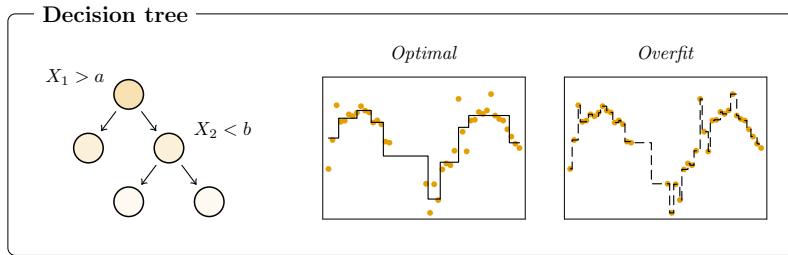


Figure 1.3: Decision trees.

In a random forest model (figure 1.4), predictions from a large number of decision trees are averaged in an attempt to improve generalisability. The training conditions of individual trees are highly randomised (e.g. by removing a percentage of rows and columns of the data set at random) to create a diverse model ensemble. Random forests feature a set of parameters (so-called hyper-parameters) that can be tuned in order to adjust the learning procedure and to create models better suited to the data. Manually choosing appropriate values for these parameters is difficult, and so cross validation procedures are often employed to search the hyperparameter space for the optimal setting. Aside from hyperparameter tuning, performance of random forest models is difficult to control. Their architecture is also relatively rigid, hampered by the fact that their learning procedure is non-differentiable. Random forests are thus often replaced by more specialised algorithms when data sets become more complex (for example when dealing with images). Tree-based models might nonetheless still have a role to play for tabular data sets where the data naturally organises into cohorts with different outcomes. The method for example has seen some success relative to other algorithms in classification problems.

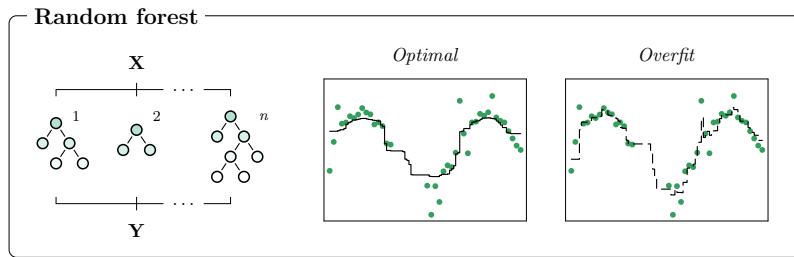


Figure 1.4: Random forests.

TO PUT IT SIMPLY...

Random forest models are made up of many decision trees (hence the name "forest"). Decision trees are prevalent in the medical domain as decision support tools, and most physicians will be familiar with them from flow diagrams (e.g. "*if [severe disease] and [blood marker elevated] then [give drug X]*"). Decision trees can be automatically constructed from data but can quickly become complex, reducing their performance in practice. Random forests improve performance by fitting many decision trees, randomising the learning process, and averaging predictions. They work well on small, tabular data sets but are challenging to apply to more complex data sets.

1.2.4 *Neural networks*

Artificial neural networks are a biologically-inspired approach to replicate the neuronal structure present in the human brain. A neural network (see figure 1.5) consists of a hierarchical set of nodes (representing neurons) that are connected to other nodes via edges (representing synapses). Each node processes the information received from previous nodes and passes on the processed signal further downstream. Models consisting of many such neuronal layers (known as "hidden layers") are called deep neural networks. Each node in the network performs a linear transformation of the data and passes the result through a non-linear function (so-called "activation functions"). Based on this relatively simple architecture, deep neural networks can perform very

complex tasks. The choice of activation function and depth of the model affect its capabilities and performance. For example, single layer neural networks based on the rectified linear unit (ReLU) activation function are already considered *universal function approximators* [42]. Deep neural networks with large numbers of neurons (e.g. GPT-3 has 175 billion parameters) can generate text, images, or video [43–45].

*Universal function
approximators can
represent any
continuous
function.*

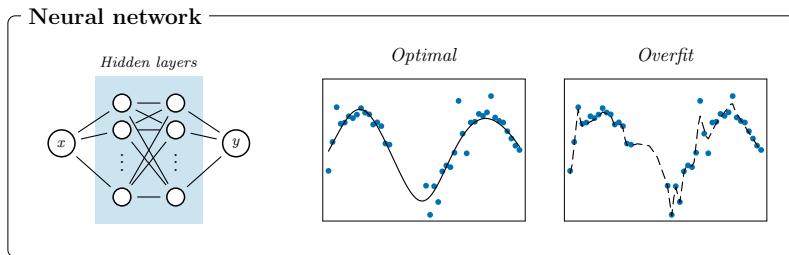


Figure 1.5: Neural networks.

The success of neural networks can in part be attributed to their extensibility: highly specialised architectures such as convolutional layers (scanning groups of pixels in images rather than single pixels), recurrent networks (tracking a latent state as a form of memory), and most recently transformers (efficiently learning historical dependencies in data) can be used to adapt models to specific data types [46–48]. These architectures are fully differentiable, meaning that complex combinations of structures can be implemented in order to create highly specialised model. In this thesis, we will focus mainly on the function approximation capabilities of relatively simple neural networks and their composability with other methods, such as differential equations.

TO PUT IT SIMPLY...

The structure of an artificial neural network is meant to mimic that of the human brain, where signals pass between neurons while requiring a certain intensity to activate. Although some of the similarities are lost in practice, neural networks still perform surprisingly well at various tasks. The reasons are however not well understood. Generally, combining bigger models with larger data sets yields more complex behaviour. Their ability to produce (potentially) useful models without the need for prior knowledge, combined with their extensibility to different types of data (e.g. audio or image data) make them very appealing.

1.2.5 Gaussian Processes

Gaussian Processes (GPs) are an extension of the multivariate normal distribution to infinite dimensions [49]. To aid the reader in grasping this concept, we can frame GPs as a collection of Gaussian variables indexed by time, such that a GP is described by a Gaussian (i.e. Normal) distribution with an unique mean and variance at each time point (see figure 1.6). These variables are described by a mean function and a kernel function. The kernel function describes the correlation between each variable. When correlation is high, distant variables (in time) have similar values and when correlation is low the random variables appear to be more randomly distributed. We can also think of a GP as representing a distribution over functions, with the kernel function describing the complexity of the functions. Similar to neural networks, GPs using specific kernels are universal function approximators [50]. Similarities do not end there: neural networks with infinite width (i.e. an infinite number of neurons) featuring Gaussian distributions over their parameters (also known as Bayesian neural networks) are equivalent to GPs [51].

GPs are attractive since they offer a probabilistic, data-driven approach to the modelling of data, providing predictions with uncertainty estimates. The choice of kernel function can be used to add prior knowledge to model structure, potentially improving performance in smaller data sets. One downside of GPs is their computational complexity, resulting in poor scaling to larger data sets. Standard GPs are

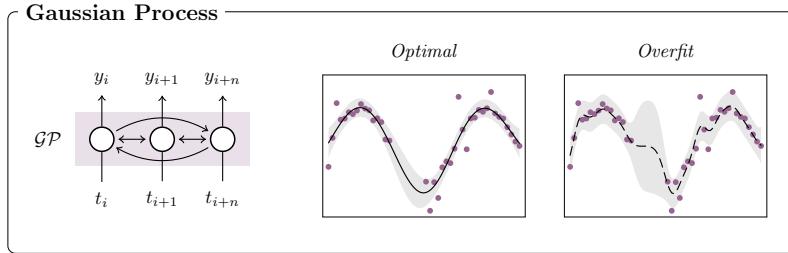


Figure 1.6: Gaussian Processes.

also not suitable in cases where the data does not follow a Gaussian distribution (e.g. when modelling categorical data). This has led to the development of sparse and approximate methods for fitting GPs [52, 53]. Although much of the machine learning field is dominated by neural network based models, GP-based algorithms might still be beneficial in a setting where model extrapolation and uncertainty is important.

TO PUT IT SIMPLY...

Gaussian Processes are a collection of infinitely many Normal distributions, each indexed by a variable such as time, with specific means and variances at each point. Since it is represented by distributions, its predictions have confidence intervals which provides a measure of uncertainty. A key aspect of learning a Gaussian Process is learning the covariance between points; high covariance indicates higher similarity, while low covariance suggests they are more distinct. The result is a smooth curve, with parameters controlling the curve's complexity. Gaussian Processes are a flexible and powerful tool for modelling data and learning underlying patterns with quantified uncertainty.

1.2.6 Overfitting

One of the strengths of machine learning algorithms – their ability to learn from data – also represents one of their most important caveats. Since these algorithms can learn almost any complex function,

they are prone to learning spurious effects. Indeed, methods like neural networks can make highly accurate predictions even when data or labels are completely randomised [54]. This concept is known as overfitting, a phenomenon where the model learns to (perfectly) reproduce the data instead of learning useful relationships between the variables. When training machine learning algorithms, one of the most important factors to test is whether the model has learned something "useful". In practice, one evaluates how well the model generalises to unseen data. Ideally, machine learning methods are first trained on one data set, and performance is then evaluated on additional (external) sets of data. However, this is often not possible as data is limited, which is especially the case in the context of rare diseases such as haemophilia A. Alternatively, methods explaining model output (known as explainable AI) might be useful to judge whether the model has learned "correct" information. These methods can be used to rank covariates based on their importance, or to visualise the approximate relationship with the dependent variable. However, such methods generally add an additional black-box layer, while interpretation of the resulting model explanations is not always straightforward. Frequently however, the available data is simply divided into a train and test set. The training data is then used to develop the model, while accuracy is evaluated on the test data. This process is then replicated many times in order to estimate generalisability of the model. Unfortunately, there is no single method that can definitively determine whether a model has overfit to the data, so model evaluation procedures generally require expertise with respect to the methods used.

1.2.7 Domain-specific challenges in pharmacometrics and rare disease

The adoption of machine learning in the context of pharmacometrics and rare disease holds great potential [55–57]. Examples include the ability to learn features from medical imaging data, or to find complex patterns in *-omics* data to improve the treatment of disease. Machine learning algorithms might reduce the complexity of model development as they can be used to learn covariate effects, disease characteristics, or treatment outcomes. This can also be useful in settings where disease physiology is not yet fully understood, as is often the case for rare disease. There are however several barriers that complicate the implementation of machine learning in these fields. First, data is often sparse, both in the number of subjects as in the number of relevant clinical measurements per subject. Second, measurements

*The suffix *-omics* references the study of large-scale data of biological molecules like genes, proteins, or metabolites.*

are often collected at irregular intervals, such that a large fraction of data is missing per subject. Third, inter-individual variability in drug exposure and response is often high, meaning that models should account for some form of prediction uncertainty. Fourth, pharmacometric models are frequently used to perform counterfactual analysis, for example by simulating how drug exposure is affected by changes in the dosing schedule. The algorithm should thus reliably extrapolate to unseen treatment settings. Finally, a large degree of trust in model predictions is required in the context of medical decision making. Model interpretation and explanation are thus likely important components of a successful system.

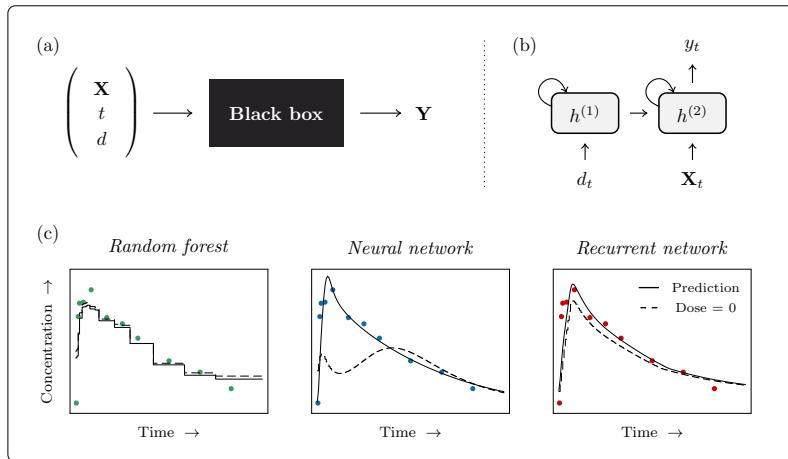


Figure 1.7: Model architectures that fail to extrapolate to unseen settings.

Typical machine learning algorithms are not necessarily well suited for tackling these problems. Strikingly, we can easily show that most standard algorithms (and even specialised architectures such as recurrent neural networks) are inadequate. Problems arise as a consequence of the use of treatment information as input to the model. Since we can never be certain that the model correctly interprets these inputs, extrapolation is inherently unreliable. In figure 1.7, we show examples of different models failing to extrapolate. We depict two generic model architectures that make incorrect predictions in counterfactual scenarios. In figure 1.7a, we depict models that provide the current time point or time after dose t and the dose amount d as direct inputs to a

black box model. Since the model has no explicit causal representation of the dose, the model attempts to correlate the dose with the observed drug measurements. We cannot guarantee that the model correctly interpolates between dose amounts given that it has observed only a fraction of the possible options. The same problem arises with respect to the direct use of time t : making predictions far outside of the time frame can result in unexpected behaviour. Since the covariates \mathbf{X} are provided together with the current time point and dose, any changes in \mathbf{X} can also affect how t and d are used by the model.

In figure 1.7b, a schematic representation of a recurrent network architecture is shown. These models are often used to improve performance when using time-series data. Here, hidden states h describe how the dose and concentration evolve over time (represented by $h^{(1)}$ and $h^{(2)}$, respectively). By using a recurrent architecture, time is handled explicitly as the time window is discretized and the change in h at each time step is evaluated. However, the model still has no causal representation of the effect of dose, and is still prone to making errors on unseen data.

When we compare the predictions from the three models when the patient is given the true dose (see figure 1.7c; solid lines) and the counterfactual case where no dose is given (dashed lines), we see that all models erroneously predict drug exposure when no drug is administered. For simple counterfactual cases as in the above example, we can explicitly train the model to reproduce the expected behaviour (no drug exposure). However, preparing the model for all counterfactual cases is fundamentally impossible. This means that specialised algorithms adhering to existing concepts of pharmacometrics will need to be developed. The development of a reliable and robust approach that performs well on small data sets is an open problem.

1.3 OBJECTIVE

The objective of this thesis is three-fold:

- i Identify opportunities for machine learning to enhance pharmacometric analyses.
- ii Develop a reliable and robust machine learning-based approach for the prediction of drug exposure and effects in the rare disease setting.
- iii Apply machine learning methods to improve the treatment of patients with haemophilia A.

We focus on describing novel machine learning-based approaches that facilitate more comprehensive analyses of data within the domain of pharmacometrics. These models have important additional requirements within the context of the medical domain. Specifically within the context of haemophilia A (and other rare diseases), we will be looking to describe algorithms that reliably handle sparse data.

1.4 OUTLINE OF THIS THESIS

This thesis is divided into four parts. Part i and ii have a technical focus, where we describe the adoption and development of novel machine learning based architectures in pharmacometrics. Part iii has a clinical focus, where we show how machine learning methods can be applied to improve the prediction of FVIII exposure and treatment outcomes in haemophilia A. In part iv we introduce the OPTI-CLOT web-portal, a web-application aiming to improve the accessibility of PK-guided dosing for treatment teams of patients with rare bleeding disorders.

1.4.1 Part one: Machine learning in pharmacometrics

In part i, we discuss recent applications of machine learning within the domain of pharmacometrics. In chapter 2, we performed a review of the recent literature and discuss several opportunities for the implementation of machine learning approaches in pharmacometrics. In chapter 3, we describe one such approach in the context of covariate selection, where machine learning algorithms are fit to data and explainability methods are used in order to identify covariate effects.

1.4.2 Part two: Deep compartment models

In section ii, we present the deep compartment model (DCM) framework as a reliable and robust approach for the prediction of drug exposure and effects. In chapter 4, we present the general architecture of the DCM and show how this method can reliably handle complex dosing schedules. We also show that the algorithm still performs reasonably well when data is sparse. In chapter 5, we continue building on this framework by introducing additional forms of constraints to model structure that further improve performance on sparse data sets. In addition, we show that models can be designed in such a

way that covariate effects can be visualised, making the model inherently interpretable. Finally, in chapter 6, we facilitate the estimation of mixed-effects, allowing model predictions to be adjusted based on observed measurements. This is essential for implementation in clinical practice, as individually adjusted predictions are used to simulate drug exposure and response to select individualised treatment regimens.

1.4.3 Part three: Machine learning for improving the treatment of patients with haemophilia A

In part iii, we discuss three examples where we use machine learning methods to tackle open issues in the treatment of haemophilia A patients. First, in chapter 7, we describe how the DCM can be used in a causal inference setting to create a model that can accurately extrapolate to new conditions. Specifically, we are interested in the counterfactual scenario where we estimate drug exposure if the patient were given a different drug. In addition, we augment the model with a generative component in order to impute missing covariate data. Next, in chapter 8, we explore the differences in FVIII PK between the prophylactic and perioperative setting. Here, the method from chapter 2 is used to identify covariates that explain observed differences in FVIII exposure. In addition, we use GPs to identify time-dependent changes in the postoperative clearance of FVIII to enable more accurate prediction of FVIII exposure over time. Finally, in chapter 9, we describe a novel approach for the personalisation of treatment using factor concentrates based on individual bleeding risk rather than just FVIII levels. Such an approach could significantly change the approach for personalised treatment of patients with haemophilia A.

1.4.4 Part four: The OPTI-CLOT web-portal

Finally, in part iv, we describe the OPTI-CLOT web-portal, an online web-application allowing physicians in the Netherlands to request dosing advice for their patients with bleeding disorders. We describe the architecture of the system, how patient privacy is protected as well as its adoption.

REFERENCES

- [1] Angela C Weyand and Veronica H Flood. "Von Willebrand disease: current status of diagnosis and management". In: *Hematology/Oncology Clinics* 35.6 (2021), pp. 1085–1101.
- [2] Erik Berntorp, Kathelijn Fischer, Daniel P Hart, Maria Elisa Mancuso, David Stephensen, Amy D Shapiro, and Victor Blanchette. "Haemophilia". In: *Nature reviews Disease primers* 7.1 (2021), p. 45.
- [3] Massimo Franchini and Pier Mannuccio Mannucci. "Past, present and future of hemophilia: a narrative review". In: *Orphanet journal of rare diseases* 7 (2012), pp. 1–8.
- [4] Pier Mannuccio Mannucci. "Hemophilia: treatment options in the twenty-first century". In: *Journal of Thrombosis and Haemostasis* 1.7 (2003), pp. 1349–1355.
- [5] Hussien Ahmed H Abdelgawad, Rachel Foster, and Mario Otto. "Nothing short of a revolution: Novel extended half-life factor VIII replacement products and non-replacement agents reshape the treatment landscape in hemophilia A". In: *Blood Reviews* (2023), p. 101164.
- [6] IM Nilsson. "Experience with prophylaxis in Sweden." In: *Seminars in hematolgy*. Vol. 30. 3 Suppl 2. 1993, pp. 16–19.
- [7] AHM Triemstra, C Smit, HM Van der Ploeg, E Briët, and FR Rosendaal. "Two decades of haemophilia treatment in the Netherlands, 1972–92". In: *Haemophilia* 1.3 (1995), pp. 165–171.
- [8] Erik Berntorp. "Prophylactic therapy for haemophilia: early experience". In: *Haemophilia* 9 (2003), pp. 5–9.
- [9] K Fischer, Jan Astermark, JG Van Der Bom, R Ljung, Erik Berntorp, DE Grobbee, and HM Van Den Berg. "Prophylactic treatment for severe haemophilia: comparison of an intermediate-dose to a high-dose regimen". In: *Haemophilia* 8.6 (2002), pp. 753–760.
- [10] Kathelijn Fischer, Katarina Steen Carlsson, Pia Petrini, Margareta Holmström, Rolf Ljung, H Marijke van den Berg, and Erik Berntorp. "Intermediate-dose versus high-dose prophylaxis for severe hemophilia: comparing outcome and costs since the 1970s". In: *Blood, The Journal of the American Society of Hematology* 122.7 (2013), pp. 1129–1136.
- [11] Jamie O'Hara, David Hughes, Charlotte Camp, Tom Burke, Liz Carroll, and Daniel-Anibal Garcia Diego. "The cost of severe haemophilia in Europe: the CHESS study". In: *Orphanet journal of rare diseases* 12 (2017), pp. 1–8.
- [12] Alfonso Iorio. "Using pharmacokinetics to individualize hemophilia therapy". In: *Hematology 2014, the American Society of Hematology Education Program Book* 2017.1 (2017), pp. 595–604.
- [13] J Oldenburg and T Albert. "Novel products for haemostasis—current status". In: *Haemophilia* 20 (2014), pp. 23–28.
- [14] Sven Björkman, Anna Folkesson, and Siv Jönsson. "Pharmacokinetics and dose requirements of factor VIII over the age range 3–74 years: a population analysis based on 50 patients with long-term prophylactic treatment for haemophilia A". In: *European journal of clinical pharmacology* 65 (2009), pp. 989–998.

- [15] Giovanni Longo, Marzia Matucci, Massimo Morfini, Silvia Vannini, and Andrea Messori. "A calculator program for individualizing factor VIII dosage". In: *Drug Intelligence & Clinical Pharmacy* 18.9 (1984), pp. 726–730.
- [16] Andrea Messori, Giovanni Longo, Marzia Matucci, Massimo Morfini, and Pier Luigi Rossi Ferrini. "Clinical pharmacokinetics of factor VIII in patients with classic haemophilia". In: *Clinical pharmacokinetics* 13 (1987), pp. 365–380.
- [17] Alok Srivastava, Elena Santagostino, Alison Dougall, Steve Kitchen, Megan Sutherland, Steven W Pipe, Manuel Carcao, Johnny Mahlangu, Margaret V Ragni, Jerzy Windyga, et al. "WFH guidelines for the management of hemophilia". In: *Haemophilia* 26 (2020), pp. 1–158.
- [18] MV Ragni, SE Croteau, Massimo Morfini, MH Cnossen, A Iorio, et al. "Pharmacokinetics and the transition to extended half-life factor concentrates: communication from the SSC of the ISTH". In: *Journal of Thrombosis and Haemostasis* 16.7 (2018), pp. 1437–1441.
- [19] Annamaria Iorio, V Blanchette, J Blatny, P Collins, K Fischer, E Neufeld, et al. "Estimating and interpreting the pharmacokinetic profiles of individual patients with hemophilia A or B using a population pharmacokinetic approach: communication from the SSC of the ISTH". In: *Journal of Thrombosis and Haemostasis* 15.12 (2017), pp. 2461–2465.
- [20] M Morfini, M Lee, A Messori, et al. "The design and analysis of half-life and recovery studies for factor VIII and factor IX". In: *Thrombosis and haemostasis* 66.09 (1991), pp. 384–386.
- [21] Sven Björkman. "Limited blood sampling for pharmacokinetic dose tailoring of FVIII in the prophylactic treatment of haemophilia A". In: *Haemophilia* 16.4 (2010), pp. 597–605.
- [22] K Van Dijk, K Fischer, JG Van Der Bom, DE Grobbee, and HM Van Den Berg. "Variability in clinical phenotype of severe haemophilia: the role of the first joint bleed". In: *Haemophilia* 11.5 (2005), pp. 438–443.
- [23] Josefina Ahnström, Erik Berntorp, K Lindvall, and S Björkman. "A 6-year follow-up of dosing, coagulation factor levels and bleedings in relation to joint status in the prophylactic treatment of haemophilia". In: *Haemophilia* 10.6 (2004), pp. 689–697.
- [24] Peter William Collins, VS Blanchette, K Fischer, Sven Björkman, M Oh, S Fritsch, P Schroth, G Spotts, Jan Astermark, B Ewenstein, et al. "Break-through bleeding in relation to predicted factor VIII levels in patients receiving prophylactic treatment for severe hemophilia A". In: *Journal of Thrombosis and Haemostasis* 7.3 (2009), pp. 413–420.
- [25] Rolf Ljung, Kathelijn Fischer, Manuel Carcao, Elena Santagostino, Marilyn J Manco-Johnson, and Prasad Mathew. "Practical considerations in choosing a factor VIII prophylaxis regimen: role of clinical phenotype and trough levels". In: *Thrombosis and haemostasis* 116.05 (2016), pp. 913–920.
- [26] M Van Geffen, M Menegatti, A Loof, P Lap, M Karimi, BAP Laros-van Gorkom, P Brons, and WL Van Heerde. "Retrospective evaluation of bleeding tendency and simultaneous thrombin and plasmin generation in patients with rare bleeding disorders". In: *Haemophilia* 18.4 (2012), pp. 630–638.

- [27] Y Dargaud, C Negrier, L Rusen, J Windyga, P Georgiev, J Bichler, C Solomon, S Knaub, T Lissitchkov, and R Klamroth. "Individual thrombin generation and spontaneous bleeding rate during personalized prophylaxis with Nuwiq®(human-cl rh FVIII) in previously treated patients with severe haemophilia A". In: *Haemophilia* 24.4 (2018), pp. 619–627.
- [28] Debnath Maji, Michael A Suster, Divyashwathi Citla Sridhar, Maria Alejandra Pereda, Janet Martin, Lalitha V Nayak, Pedram Mohseni, and Sanjay P Ahuja. "Assessment of Bleeding Phenotype in Hemophilia A By a Novel Point-of-Care Global Assay". In: *Blood* 134 (2019), p. 4662.
- [29] Michael U Callaghan, Claude Negrier, Ido Paz-Priel, Tiffany Chang, Sammy Chebon, Michaela Lehle, Johnny Mahlangu, Guy Young, Rebecca Kruse-Jarres, Maria Elisa Mancuso, et al. "Long-term outcomes with emicizumab prophylaxis for hemophilia A with or without FVIII inhibitors from the HAVEN 1-4 studies". In: *Blood, The Journal of the American Society of Hematology* 137.16 (2021), pp. 2231–2242.
- [30] Anouk AMT Donners, Carin MA Rademaker, Lisanne AH Bevers, Alwin DR Huitema, Roger EG Schutgens, Toine CG Egberts, and Kathelijn Fischer. "Pharmacokinetics and associated efficacy of emicizumab in humans: a systematic review". In: *Clinical Pharmacokinetics* 60.11 (2021), pp. 1395–1406.
- [31] Steven W Pipe, Francesca Ferrante, Muriel Reis, Sara Wiegmann, Claudia Lange, Manuela Braun, and Lisa A Michaels. "First-in-human gene therapy study of AAVhu37 capsid vector technology in severe hemophilia A-BAY 2599023 has broad patient eligibility and stable and sustained long-term expression of FVIII". In: *Blood* 136 (2020), pp. 44–45.
- [32] Johnny Mahlangu, Radoslaw Kaczmarek, Annette Von Drygalski, Susan Shapiro, Sheng-Chieh Chou, Margareth C Ozelo, Gili Kenet, Flora Peyvandi, Michael Wang, Bella Madan, et al. "Two-year outcomes of valoctocogene roxaparvovec therapy for hemophilia A". In: *New England Journal of Medicine* 388.8 (2023), pp. 694–705.
- [33] Margareth C Ozelo, Johnny Mahlangu, K John Pasi, Adam Giermasz, Andrew D Leavitt, Michael Laffan, Emily Symington, Doris V Quon, Jiaan-Der Wang, Kathelijne Peerlinck, et al. "Valoctocogene roxaparvovec gene therapy for hemophilia A". In: *New England Journal of Medicine* 386.11 (2022), pp. 1013–1025.
- [34] Hussam Alkaissi and Samy I McFarlane. "Artificial hallucinations in ChatGPT: implications in scientific writing". In: *Cureus* 15.2 (2023).
- [35] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions". In: *arXiv preprint arXiv:2311.05232* (2023).
- [36] Konstantinos C Sontis, Zachi I Attia, Samuel J Asirvatham, and Paul A Friedman. "ChatGPT hallucinating: can it get any more humanlike?" In: (2024).
- [37] Patrick J Thoral, Mattia Fornasa, Daan P de Bruin, Michele Tonutti, Hidde Hovenkamp, Ronald H Driessens, Armand RJ Girbes, Mark Hoogendoorn, and Paul WG Elbers. "Explainable machine learning on AmsterdamUMCdb for ICU discharge decision support: uniting intensivists and data scientists". In: *Critical care explorations* 3.9 (2021), e0529.

- [38] David Zopfs, Kai R Laukamp, Stefanie Paquet, Simon Lennartz, Daniel Pinto dos Santos, Christoph Kabbasch, Alexander Bunck, Marc Schlamann, and Jan Borggrefe. "Follow-up MRI in multiple sclerosis patients: automated co-registration and lesion color-coding improves diagnostic accuracy and reduces reading time". In: *European Radiology* 29 (2019), pp. 7047–7054.
- [39] Marco V Perez, Kenneth W Mahaffey, Haley Hedlin, John S Rumsfeld, Ariadna Garcia, Todd Ferris, Vidhya Balasubramanian, Andrea M Russo, Amol Rajmane, Lauren Cheung, et al. "Large-scale assessment of a smartwatch to identify atrial fibrillation". In: *New England Journal of Medicine* 381.20 (2019), pp. 1909–1917.
- [40] Ravid Shwartz-Ziv and Amitai Armon. "Tabular data: Deep learning is not all you need". In: *Information Fusion* 81 (2022), pp. 84–90.
- [41] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on typical tabular data?" In: *Advances in neural information processing systems* 35 (2022), pp. 507–520.
- [42] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks". In: *Neural networks* 3.5 (1990), pp. 551–560.
- [43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (2023).
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), pp. 10684–10695.
- [45] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. "Photorealistic video generation with diffusion models". In: *arXiv preprint arXiv:2312.06662* (2023).
- [46] Yann LeCun, Yoshua Bengio, et al. "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995.
- [47] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.
- [48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).
- [49] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. Vol. 2. 3. MIT press Cambridge, MA, 2006.
- [50] Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. "Universal Kernels". In: *J. Mach. Learn. Res.* 7 (2006), 2651–2667. ISSN: 1532-4435.
- [51] Radford M. Neal. "Priors for Infinite Networks". In: *Bayesian Learning for Neural Networks*. New York, NY: Springer New York, 1996, pp. 29–53. ISBN: 978-1-4612-0745-0.

- [52] Michalis Titsias. "Variational Learning of Inducing Variables in Sparse Gaussian Processes". In: *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*. Ed. by David van Dyk and Max Welling. Vol. 5. Proceedings of Machine Learning Research. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR, 2009, pp. 567–574.
- [53] Joaquin Quiñonero Candela and Carl Edward Rasmussen. "A Unifying View of Sparse Approximate Gaussian Process Regression". In: *J. Mach. Learn. Res.* 6 (2005), 1939–1959. ISSN: 1532-4435.
- [54] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization". In: (2017). arXiv: 1611.03530 [cs.LG].
- [55] Sergio Decherchi, Elena Pedrini, Marina Mordini, Andrea Cavalli, and Luca Sangiorgi. "Opportunities and challenges for machine learning in rare diseases". In: *Frontiers in medicine* 8 (2021), p. 747612.
- [56] Jineta Banerjee, Jaclyn N Taroni, Robert J Allaway, Deepashree Venkatesh Prasad, Justin Guinney, and Casey Greene. "Machine learning in rare disease". In: *Nature Methods* 20.6 (2023), pp. 803–814.
- [57] Julia Carrasco-Zanini, Maik Pietzner, Jonathan Davitte, Praveen Surendran, Damien C Croteau-Chonka, Chloe Robins, Ana Torralbo, Christopher Tomlinson, Florian Grünschläger, Natalie Fitzpatrick, et al. "Proteomic signatures improve risk prediction for common and rare diseases". In: *Nature Medicine* (2024), pp. 1–10.

Part I

MACHINE LEARNING IN
PHARMACOMETRICS

2

ADOPTION OF MACHINE LEARNING IN PHARMACOMETRICS: AN OVERVIEW OF RECENT IMPLEMENTATIONS AND THEIR CONSIDERATIONS

Alexander Janssen, Frank C. Bennis, Marjon H. Cnossen, and Ron A.A. Mathôt

Pharmaceutics 14(9) (2022): 1814

ABSTRACT

Pharmacometrics is a multidisciplinary field utilising mathematical models of physiology, pharmacology, and disease to describe and quantify the interactions between medication and patient. As these models become more and more advanced, the need for advanced data analysis tools grows. Recently, there has been much interest in the adoption of machine learning (ML) algorithms. These algorithms offer strong function approximation capabilities and might reduce the time spent on model development. However, ML tools are not yet an integral part of the pharmacometrics workflow. The goal of this work is to discuss how ML algorithms have been applied in four stages of the pharmacometrics pipeline: data preparation, hypothesis generation, predictive modelling, and model validation. We will also discuss considerations before the use of ML algorithms with respect to each topic. We conclude by summarising applications that hold potential for adoption by pharmacometrists.

2.1 INTRODUCTION

2.1.1 *Background*

Pharmacometrics is a multidisciplinary field utilising mathematical models of physiology, pharmacology, and disease to describe and quantify the interactions between medication and patient. This involves models of drug pharmacokinetics (PK), pharmacodynamics (PD), exposure-response (PK/PD), and disease progression. One of the main themes of interest is the explanation of variability in drug response between patients. Various statistical techniques have been adopted to quantify such inter-individual variation (IIV) [1].

Non-linear mixed effect (NLME) modelling has been embraced as a statistical method for describing treatment effect on a population and individual level [2, 3]. Population PK modelling makes efficient use of sparse data by pooling information of multiple individuals, and breaking down treatment response in shared and individual effects. Observations of the dependent variable (i.e., drug concentrations or treatment effect) can then be used to adapt the prediction to the individual patient, resulting in higher accuracy.

Recently, however, advances in hospital digitisation, data collection, and inclusion of increasingly extensive laboratory testing in standard clinical care have resulted in the availability of richer data sets. This increased accessibility of complex data sources such as genomic or gene expression data stresses current modelling approaches as they can lack the flexibility to handle these data. As a response, more attention is being paid to the opportunity of using machine learning (ML) algorithms as an innovative strategy for pharmacometric modelling [4, 5]. The field of ML has seen an explosive boost of promising applications for image analysis, text recognition, and other high-dimensional data. There are many examples of their successful application in the medical domain, for example for the diagnosis of breast cancer [6], identification of biomarkers from gene expression data [7], and survival analysis [8]. As ML methods offer strong predictive performance there is no denying that its adoption in pharmacometrics brings with it exciting new modelling opportunities.

As the relatively young ML research field is maturing at a rapid pace, more advanced model architectures are frequently being proposed in order to further improve predictive accuracy. Consequently, understanding the differences and intricacies of distinct learning methods is becoming increasingly more difficult for non-experts. A proper

understanding of the advantages and pitfalls of these methods is essential for their responsible and reliable use, especially for clinical applications. As most of the emphasis has been put on the supposed high predictive accuracy of ML methods, it is easy to become overconfident in their abilities. It is thus important to monitor and guide the adoption of ML in pharmacometrics.

In this review, we will discuss recent approaches for the use of ML algorithms in the context of pharmacometrics, while also providing important considerations for their use. For some examples, we will provide demonstrations based on simulation experiments. We also discuss the important concept of model validation and the importance of understanding what is actually learned by the algorithm.

In this work, we will be assuming a general understanding of ML and the most common algorithms. For those wanting to learn more about the basic concepts of ML, *Badillo et al.* offer an excellent tutorial on ML aimed at pharmacometrists [9].

2.1.2 *Structure of this review*

This review is structured as follows. First, we discuss applications of ML algorithms in three stages of the pharmacometrics pipeline: data preparation, hypothesis generation, and predictive modelling. We define these stages as follows: data preparation deals with the imputation of missing data and dimensionality reduction. Next, in the section on hypothesis generation we discuss methods for clustering data, and how ML can be used for the detection of influential covariates. In the predictive modelling section, we discuss ML-based alternatives to traditional modelling approaches. We will conclude our review of recent application with a discussion on model validation, focusing mainly on estimating model generalisability and the interpretation of ML models.

For each topic, we first discuss the current approach, its (possible) limitations, followed by what ML techniques have been proposed to address the issues. At the end of each topic, we will summarise the discussion with considerations for the use of ML for each issue.

2.1.3 *Literature search*

In order to support the initial framing of our discussion we performed a literature search. Our objective was to find recent articles discussing

ML in the context of pharmacometrics. The following search query for PubMed was constructed:

```
("machine learning" [tiab] OR "artificial intelligence"  
[tiab] OR "random forest" [tiab] OR "gradient  
boosting" [tiab] OR "XGBoost" [tiab] OR "support  
vector" [tiab] OR "neural network" [tiab] OR "deep  
learning" [tiab]) AND ("pharmacometric*" OR "  
pharmacokinetic*" OR "pharmacodynamic*" [tiab] OR "  
pharmacogen*" [tiab] OR "drug concentration" [tiab]  
OR "dose estimation" [tiab] OR "dose optimization" [  
tiab]) AND ("2016/01/01" [Date - Publication]:  
"3000" [Date - Publication]) NOT (review[Publication  
Type]).
```

The search identified a total of 586 articles (as of 30 May 2022), of which 198 were included based on abstract screening. Additional articles were obtained by means of scanning the reference lists of included articles, or by specifically searching in the arXiv database (<https://arxiv.org/>; accessed from 30 May 2022 until 30 June 2022). Some ML papers are only indexed in pre-print servers, and thus can not be found in PubMed.

2.2 DATA PREPARATION

2.2.1 *Data Imputation*

Missing data are a frequent occurrence in the clinical setting. When encountering missing data one can drop all data entries or covariates with missing data, impute missing data, or employ maximum likelihood estimation techniques. As many clinical data sets are relatively small, the latter two options are often preferred. Missing data are often categorised in one of three categories; they are either missing completely at random (MCAR), missing at random (MAR; missingness depends on observed data), or missing not at random (MNAR; missingness depends on unobserved data). The source of the missing data can affect the choice of imputation method. In addition, the type of data (i.e., continuous or categorical) can also be a reason for choosing different methods. In the below sections, we will focus on the problem of data imputation, which requires us to choose an appropriate model for the prediction of missing data. How do we select such a model?

2.2.1.1 Standard Methods for Data Imputation

Imputation can either be performed once (single imputation), or multiple times (multiple imputation). Commonly used methods for single imputation include imputation by mean or mode, grouping missing data in a separate category (in the case of categorical covariates), or regression-based imputation. In multiple imputation, multiple samples are taken from a predictive distribution allowing for the quantification of the resulting variance of model output. This provides a measure of uncertainty of the imputation. A Bayesian multiple imputation strategy has been proposed for NLME models, which presented lower bias of parameter estimates compared to mean value imputation for MCAR and MAR data [10]. A maximum likelihood procedure based on this strategy was also shown to lead to less biased PK parameter estimates compared to mode and logistic regression based approaches [11]. Prior studies have been mainly concerned with the imputation of categorical variables. Model-based (i.e., multiple imputation and likelihood-based) approaches seem to perform well for this kind of data, but do require one to make assumption about the distribution of the data. This can be more difficult for continuous variables. In these cases, it might be compelling to also evaluate regression-based techniques for imputation. Unfortunately, choosing an appropriate regression model when the assumed relationship is non-linear can be difficult. This is especially the case when covariates are correlated. For this reason, ML-based regression techniques have been suggested with the goal of improving the accuracy of regression-based imputation. An early study suggests that when covariates are simulated based on non-linear relationships, the bias of PK parameters after performing imputation can be reduced by using a random forest or neural network prediction model rather than mean imputation [12].

2.2.1.2 Machine Learning Methods for Data Imputation

A paper by Batista and Mondard compared the accuracy of the k-nearest neighbour (k-NN) method to mode, decision tree, and rule-based methods for MCAR data imputation [13]. They found that k-NN generally was the most accurate method. In k-NN, individuals are grouped in k clusters based on similarity (for example based on Euclidean distance). Next, missing values can be imputed based on mean/median values from their respective cluster. The method is simple to implement, but might be less effective in small or very

homogeneous data sets. Although Batista and Mondard found that the decision tree-based method was not as accurate, random forest-based approaches have been more successful [14–16]. In the popular implementation missForest [14], a random forest is combined with multiple imputation by chained equations (MICE; [17]). MICE is an iterative procedure where each missing covariate is imputed based on the remaining covariates. Initially, missing data are imputed using an arbitrary method (e.g., by their mode) and a model is fit to predict missing data for each covariate independently. This process is repeated with the assumption that each iteration, more accurate imputations of the covariates are used for the predictions. The overall process can be repeated for multiple initial data sets. This way, MICE allows for multiple imputation based on deterministic regression models. Using missForest outperformed k-NN and linear MICE for single imputation of MCAR data [14]. However, performing multiple imputation using linear MICE was more accurate than single imputation using missForest on MCAR and MAR data [15]. Performing multiple imputation using missForest led to the overall best accuracy. This implies that multiple imputation procedures might also generally be preferred for regression-based imputation.

Several probabilistic approaches have also been proposed for performing regression-based multiple imputation. Unsupervised deep latent variable models, such as generative adversarial networks (GANs) and variational auto-encoders (VAE), have recently been successfully applied to data imputation problems [18, 19]. A GAN is a combination of two neural networks, a generator and a discriminator, which compete against each other. The discriminator learns to discern true from generated data, such that the generator becomes increasingly effective at reproducing real data. The generator learns to represent the generative distribution of the data. The GAIN approach, a GAN specifically developed for the imputation of missing data, was found to more accurately impute MCAR and MAR compared to missForest and MICE [18].

A VAE is a special neural network architecture that learns to encode its input into a distribution of latent variables by means of variational inference. A decoder neural network is then learned to reproduce the original input from samples of this distribution. Mattei and Frellsen describe the combination of an importance-weighted auto-encoder with a maximum likelihood objective for imputation of MAR data. This method was found to be more accurate than k-NN and missForest [19].

Finally, Gaussian Processes (GPs) are stochastic processes which represent a prior distribution over latent functions. GPs are a non-parametric framework to fit models to data, while simultaneously providing a measure of variance. This allows one to sample from the distribution of latent functions and perform multiple imputation. A recent study has proposed a deep GP approach which suggested more accurate imputation of MCAR data compared to k-NN, MICE, and GAIN [20].

2.2.1.3 Considerations

We have described several ML techniques for data imputation. These techniques might offer improved imputation of covariates that have non-linear relationships with the non-missing covariates. Several findings point to improved performance of regression models when using multiple imputation compared to single imputation. Since most standard regression methods are deterministic, strategies such as MICE are advisable for performing multiple imputation. Using missForest in this context might improve the prediction of more complex covariates. MICE is flexible in that a different model can be used for imputation of each covariate. We can thus use ML methods for the prediction of non-linear covariates while using likelihood-based approaches for categorical covariates. This might be a promising method to explore in the context of NLME modelling. Recent probabilistic approaches, such as deep latent variable models and GPs, offer an interesting take on regression-based imputation. These methods use likelihood-based approach which might improve imputation accuracy [18–20]. However, it is not clear if more accurate methods of data imputation offer significant benefits in terms of reducing the bias of parameter estimates in NLME models. Studies will have to evaluate the benefit of using these more complex methods in the context of pharmacometric modelling.

2.2.2 Dimensionality Reduction

Dimensionality reduction is a technique for detecting patterns in data and reducing this to a lower number of principal components. This can be useful when analysing very high-dimensional data (e.g., gene expression data). Such data are difficult to include in pharmacometric models (and thus rarely are) as their effect might be dependent on a specific combination of patterns. One of the main linear techniques for dimensionality reduction is principal component analysis

(PCA). It uses a linear mapping to project each data point to a lower-dimensional representation. The so called principal components are independent and aim to preserve as much of the original variance in the data. Such decompositions can be used to facilitate data visualisation and can be used for hypothesis generation (see Section 2.3). This technique has for example been used to predict the impact of different factor VIII mutations on haemophilia A disease severity [21]. Here, 544 amino acid properties were collected, which they were able to reduce to 19 components using PCA. The researchers could thus drastically reduce data dimensionality while reportedly retaining 99% of the information in the data set.

Non-linear methods have also been proposed which allow a more flexible mapping to lower-dimensional space. This way, these methods might be able to represent more complex patterns and thus increase the explained variance. One example is the VAE, where the input data are condensed into a set of latent variables. Other prominent examples include uniform manifold approximation and projection (UMAP) [22] and t-distributed stochastic neighbour embedding (t-SNE) [23]. *Xiang et al.*, have recently performed a comparison of ten dimensionality reduction methods for predicting cell type from RNA sequencing data [24]. Although the study shows that no one-size-fits-all method exists, UMAP, t-SNE, and VAE were found to generally outperform other methods of dimensionality reduction. Accuracy of PCA was also high, but it suffered in terms of stability with respect to changes in the number of cells, cell types, and genes.

Another study by *Becht et al.*, compared t-SNE to UMAP to discern cell populations based on patterns in single-cell RNA sequencing data [25]. In their case, UMAP led to more reproducible results and more meaningful visualisations. This is in contrast to the results of *Xiang et al.*, which found t-SNE to be the best performing method [24].

2.2.2.1 Considerations

To our knowledge, the use of dimensionality reduction techniques in the context of pharmacometrics is still quite limited. One of its current principal uses might be as a pre-processing step for generating hypotheses. The lower dimensional representations are ideal for visualisation and can be used to detect patterns in otherwise complex data. However, one downside can be that the meaning of the resulting lower dimensional components might be difficult to interpret. In a two dimensional t-SNE visualisation for example, samples that are closer

together and thus more similar would also have been more similar in high dimensional space. The actual features underpinning this similarity can not be discerned from the analysis. In such cases, further inspection of the data would be required to identify the explaining covariates. Another downside is that most of these methods require re-analysis of the data when including new samples. In addition, for each new sample we need to collect the same data to produce the lower dimensional representation.

An important finding is that there might not necessarily be a best method for performing dimensionality reduction [24]. This suggests that different methods should be compared based on the predictive performance of the resulting principal components. However, from our literature search we found that it was more common to use ML to directly learn to predict the outcome of interest from high-dimensional data and select influential covariates for downstream analysis [26–28]. We will further discuss such approaches in the context of covariate selection in Section 2.3. It is unclear how such an approach compares to the above-mentioned methods for dimensionality reduction. More studies are needed to explore the benefit of using these different techniques.

2.3 HYPOTHESIS GENERATION

2.3.1 *Discovery of Patient Sub-Populations*

Detecting patient subgroups can help to understand why groups of individuals respond differently to treatment. We found several studies that describe the use of clustering techniques in the context of pharmacometrics. Most focus on grouping patients based on the similarity in terms of treatment response. Kapralos and Dokoumetzidis describe the use of k-means clustering for the detection of two patient sub-populations presenting distinctly different absorption patterns of Octreotide LAR [29]. Here, they used the Fréchet distance to define similarity between patients, which can be used to calculate the similarity of longitudinal data in terms of shape. These kinds of measures however do require that all patients are sampled at similar time points. This can be difficult to achieve in practice.

Another study describes the use of mixture models for dividing patients into different classes based on treatment response [30]. Next, the study attempted to predict subgroup based on patient characteristics. This allowed for the identification of clinical indicators that were

associated with the different subgroups. The resulting seven treatment classes were validated in an external data set. This study offers a nice representation how clustering can be used for hypothesis generation and subsequent analysis.

Clustering assumptions can also be implemented within predictive models. In one study, an expectation maximisation (EM) approach is used to group individuals based on drug concentration data and a predefined compartment model [31]. Each cluster has its own distinct reparameterisation and estimate of (residual) variance. This allows us to categorise new patients to a cluster and treat them as were similar patients. Another study has taken an interesting approach where a mixture model-based algorithm is described for grouping individuals into different PK models [32]. These PK models are automatically constructed during the clustering procedure. This approach can thus also be used to generate hypotheses about the appropriate PK models to use for different patient groups. Requirement of a predefined model can also be avoided by combining clustering and supervised ML algorithms. *Chapfuwa et al.*, describe a model for clustering patients in the context of time-to-event analysis [33]. A neural network is used to represent the covariates into a latent variable space, which is made to behave as a mixture of distributions. Each individual is assigned to a cluster in the latent space which contains a corresponding event-time distribution. This allows for the identification of heterogeneous treatment response groups based on covariate data.

2.3.1.1 Considerations

We have discussed examples of studies that have used k-means and mixture models to cluster patients in subgroups. Mixture models allow for probabilistic inference, and have been used in more complex model architectures [31–33]. These approaches are experimental, but may be of interest to apply to different problems for the purpose of hypothesis generation. Both mixture models and k-means require the user to specify the number of clusters beforehand. This can be difficult when there is no prior information to choose the number of subgroups or when the data cannot be visualised due to high-dimensionality. In those cases, we can either reduce data dimensionality (see Section 2.2.2), or use some criterion to select the optimal number of clusters [30, 34]. An additional downside of mixture models is that they are sensitive to local minima. One study found that prior initialisation

based on k-means++ (an adaptation of k-means) allows for a simple procedure to improve model convergence [35].

Clustering patients based on the dependent variable can pose issues in pharmacometric modelling. For example, difficulties arise when clustering patients based on drug concentration measurements when these are collected at different time points. Aside from increasing the speed of processing many samples, the benefit of clustering based solely on the dependent variable might be unclear when differences in drug exposure can be easily discerned from the concentration–time curve. Alternatively, clustering patients based on (individual) PK parameters, summary variables such as area under the concentration–time curve (AUC), or independent variables might be more informative in practice.

2.3.2 Covariate Selection

Considering the black-box nature of most ML algorithms, why would one consider using ML for covariate selection? Step-wise covariate modelling (SCM), which is perhaps the most commonly used covariate selection method in pharmacometrics, also has its limitations [36]. Step-wise approaches can lead to difficulties when the data contains many covariates, when there is high collinearity, or when covariate effects are highly non-linear and difficult to determine a priori. The potential of ML-algorithms in this context is that they can be used to learn the optimal implementation of covariates. By performing post-hoc analyses of the model, it can then be possible to determine influential covariates. In the below sections we first discuss limitations of step-wise methods in order to suggest a set of requirements for a successful covariate selection method. Next, we describe ML methods that have been used for this purpose and evaluate if they fit the derived requirements.

2.3.2.1 Limitations of Step-wise Covariate Selection Methods

In SCM, covariates are included one by one (forward inclusion), and each time the covariate leading to the largest significant decrease in objective function value is included. After all covariates have been tested, the included covariates are removed from the full model one by one (backward elimination). The covariates that do not result in a significant increase in objective function value are removed. This approach leads to some issues. First, due to the potentially large

number of statistical tests there is a risk of multiplicity. Second, for an honest implementation of step-wise methods, all hypotheses need to be defined beforehand. This includes all covariates to consider and their functional form. The latter can be quite difficult to determine without first extensively inspecting the data. Finally, the statistical tests are not independent, since the significance of the tests might depend on how and if other covariates have been included. This is especially a problem when there is high collinearity between covariates. Studies have indeed indicated that SCM has a relatively low power when covariates are correlated, have weak effects, or when the number of observations in the data set is limited [37–39].

To reduce the effect of multiplicity and to ensure tests are independent, full model methods are preferred [36]. However, this does not resolve the issue of choosing a suitable functional form to implement each covariate *a priori*. We suggest the following definition of ideal covariate selection method: it (1) should perform a full model fit (i.e., test multiple hypotheses simultaneously); (2) should be able to learn covariate relationships from data; while (3) penalising complex solutions (e.g., by regularisation); and (4) should allow for the interpretation of resulting relationships. If the method is unable to learn optimal implementations of covariates, we risk making type II errors. If the method does not constrain model complexity it risks inflating the importance of covariates (by fitting arbitrarily complex relationships) resulting in higher type I error. Finally, if the method is not interpretable, we run into problems when actually implementing the selected covariates. If sub-optimal functions are used to implement the selected covariates, they might still result in insignificant effects.

2.3.2.2 *Linear Machine Learning Methods*

The least absolute shrinkage and selection operator, or LASSO, is a regression-based method that performs covariate selection by regularisation. The LASSO employs the ℓ^1 -norm, which penalises the absolute size of the regression coefficients β . This causes the coefficients of unimportant covariates to be shrunk to exactly zero. All covariates of interest are tested simultaneously in the form of linear equations using a full model fit. Next, a hyperparameter s , which controls the size of β such that $\sum_{j=1}^{N_{cov}} |\beta_j| \leq s$, can be selected using cross-validation procedures. The use of s is a substitution for statistical testing as only the most important covariates will have coefficients greater than zero. The LASSO has seen applications for population PK and Cox

hazard models where it outperformed step-wise methods in terms of speed and predictive accuracy [38, 40]. Owing to its simplicity, direct integration into the non-linear mixed effects procedure is possible [38].

The LASSO performs a full model fit, penalises complex solutions, and is interpretable. However, due to the assumption of linear relationships the LASSO fails to meet our second requirement. Since this assumption might not hold for all covariates, there is a risk of type I errors in covariate selection. Although the predictive performance of the LASSO holds up relatively well [41], its performance suffers when the relationship of some of the covariates are non-linear [42].

Multivariate adaptive regression splines (MARS) is a ML algorithm for the approximation of non-linear functions based on piece-wise linear basis functions [43]. The method automatically learns the optimal number of splines and their location for single covariates and their combinations. Its classic implementation uses a step-wise approach to prune the number of basis functions to reduce model complexity. Alternatively, a LASSO-based implementation of MARS has been described which presented favourable performance compared to the classic approach [44]. This method has the potential of matching our requirements, but has not yet seen frequent use for the purpose of covariate selection. We have found one abstract mentioning its use, but it did not explore its benefit for approximating non-linear functions [45].

2.3.2.3 Tree-Based Methods

Tree-based ML algorithms, such as the random forest and gradient boosting trees, have seen recent applications for the purpose of covariate selection. These methods offer a flexible approach to learning non-linear functions, while offering a large number of hyperparameters that can be tuned for regularisation. Maximum tree depth, the change in minimum objective function change required for a split, or the minimal number of samples in each node can be empirically set (or automatically using cross validation) to reduce model complexity. The method fits our first three requirements, although the effects of regularisation are more difficult to interpret compared to the ℓ^1 -norm. In order to use tree-based methods for covariate selection, covariate importance scores based on “impurity” (also known as Gini importance) or permutation are often calculated. The covariates can be ranked based on these scores. Covariates can be included based on biological

plausibility or if they meet a certain threshold [46]. Permutation-based methods are preferred over impurity based methods as the latter can be biased for differently scaled or high cardinal covariates [47]. Simulation studies seem to suggest relative accurate identification of true covariates [42, 48].

It is important to note that there is no underlying theory that supports the use of these scores as selection criteria. In addition, another problem is that these scores do not provide information on what functional form to use in order to implement the covariate. It is possible that the relationship underlying the importance has a complicated functional form, and is less important when approximated using basic functions in the final model. As-is, this approach does not meet our requirement of interpretability. Novel approaches such as explainable gradient boosting [49, 50], might improve the interpretability of tree-based models.

2.3.2.4 *Genetic Algorithms*

Genetic algorithms are a special form of search space optimization techniques that rely on evolutionary concepts such as natural selection, cross-over, and random mutation for selecting the most optimal model. They have long been suggested as an alternative approach to model selection for pharmacometric applications [51]. Genetic algorithms allow for testing many opposing hypotheses with respect to model structure simultaneously. In this way, it matches our first requirement. Its direct output is an optimal model (according to the survival function) and matches our fourth requirement. The general procedure is as follows: first, the full search space is defined, containing all model features to be considered. Next, an objective function is chosen that describes model fitness. Usually this is a combination of the log likelihood of the model and additional penalties for model complexity. Then an initial population is formed containing random combinations of the selected features. For each model, the fitness function is evaluated and the ‘fittest’ models are selected to produce the next generation. This process is repeated for several iterations or when a stopping criteria is met. Since many models have to be fit and evaluated the computational cost of fitting genetic algorithms can be relatively high.

A recent study describes the development of a software-based resource for automating model selection using genetic algorithms, improving their accessibility [52]. This application was compared to step-wise methods and seemed to more accurately recover the true

model based on simulated data. Such comparisons are however difficult to make, since the penalty for model complexity was more conservative in the case of the step-wise methods versus the genetic algorithm. The reverse was found in another study, where a stricter fitness function resulted in overly simplified models [53]. Choosing an appropriate fitness function by balancing model accuracy and complexity is not straightforward. It is possible to use heuristic methods such as the Akaike or Bayesian information criterion, but it is likely that there is no one-size-fits-all solution. In addition, the method cannot be used to learn more complex representations of the covariates than were originally included in the search space. Genetic algorithms thus do not meet our second requirement.

2.3.2.5 Considerations

We have discussed several ML algorithms that can be used for covariate selection. We also proposed four requirements that underlie an ideal covariate selection tool. All discussed methods test all hypotheses simultaneously and match our first requirement. The LASSO offers the most comprehensible approach to regularisation but might risk higher type I error due to its assumption of linear relationships. More complex ML algorithms, such as tree-based methods, are more flexible with respect to the representation of non-linear relationships. Perhaps not surprisingly, these methods also suffer the greatest in terms of interpretability (with the exclusion of decision trees). This makes it difficult to translate the results of covariate importance to an appropriate model. The current principal use of tree-based methods might thus be for selecting covariates for subsequent analysis.

The MARS and explainable gradient boosting algorithms come closest to meeting all four requirements. By using piece-wise linear functions, MARS approximates non-linear functions and is interpretable. In explainable gradient boosting, a large number of simple models (e.g., small depth decision trees) are fit to each covariate, and relationships can be visualised by summarising over these models. The visualisations obtained from both methods could be useful in providing an initial intuition about the appropriate functional form to use when implementing covariates. Alternatively, model interpretation methods might be of interest to infer covariate relationships from ML models. We have previously performed an investigation into how one such explanation method can be used to visualise the relationships between covariates and estimated PK parameters [54]. We found that

these relationships matched implementations in previous PK models and biological concepts. It might be of interest to further investigate the application of such tools in the context of pharmacometrics. Model explanation methods will be further discussed in Section 2.6.

We have performed a simple simulation study (see Appendix 2.A for implementation details) to showcase the use of some of the previously mentioned methods for covariate selection. Each method was fit to predict individual clearance estimates based on covariate data containing two true covariates and 48 noise covariates. In figure 2.1, we depict the measures of covariate importance as determined by means of LASSO, MARS, random forest, or explainable gradient boosting. Each method has correctly identified the two true covariates as important. In addition, we have depicted the approximation of the covariate effect by MARS and explainable gradient boosting (see figure 2.1E,F).

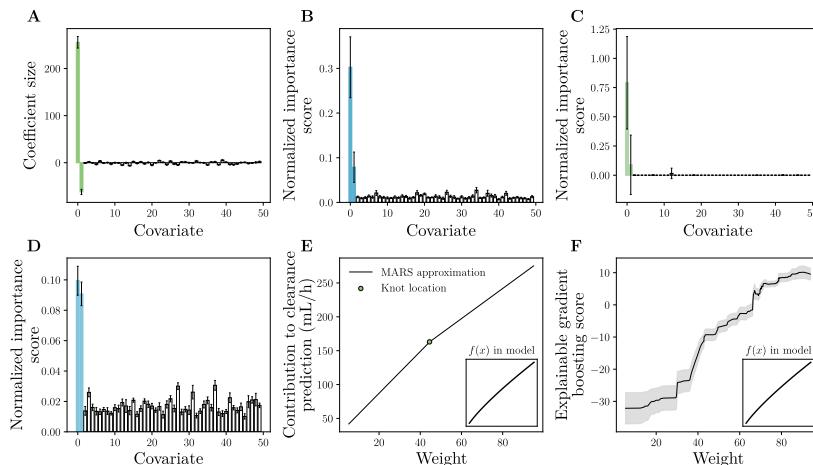


Figure 2.1: Examples of machine learning-based covariate importance scores. LASSO coefficients (A), random forest importance scores (B), MARS covariate importance (C), explainable gradient boosting scores (D), MARS (E) and explainable gradient boosting (F) approximation of the effect of covariate 1 are shown. Coloured bars indicate true covariates, whereas white bars represent noise covariates. Bar height represents the importance of each covariate. Importance should be larger for true covariates than for noise covariates. The resulting scores can for example be used to select covariates eligible for inclusion in a NLME model. Error bars indicate standard deviation of each score following a ten-fold cross validation. In (E), the point indicates the piece-wise split location (i.e., a knot). In (F), shaded area represents the standard deviation of model predictions in the explainable gradient boosting model. Figure inset represents the function used for covariate 1 in the simulations.

We have also discussed the use of genetic algorithms for automation of model selection. Compared to local search or step-wise methods, genetic algorithms offer an intuitive procedure based on evolutionary concepts for simultaneously testing multiple hypotheses with respect to model selection. Software-based resources such as presented by *Ismail et al.*, could help improve accessibility for performing experiments based on genetic algorithms [52]. Although they might be an improvement compared to step-wise methods, genetic algorithms do not meet the suggested requirements for a comprehensive covariate selection method. The main issue lies with selecting an appropriate fitness function. There is no consensus on a generally applicable fitness function. This is worrisome, as choosing an inappropriate fitness function can negatively affect the result.

In summary, none of the presented approaches can meet all our requirements for an ideal covariate selection method. Their purpose might thus mainly be for providing a more informed set of covariates to test. We have mentioned some methods such as MARS and explainable gradient boosting which might also provide intuition about appropriate functional forms to use. Next, genetic algorithms can be used as a full model based approach to testing the hypotheses. More research is however required to optimise this procedure.

2.4 PREDICTIVE MODELS

2.4.1 *Machine Learning for Pharmacokinetic Modelling*

The development of NLME models is a time-consuming process and requires extensive domain knowledge. Recently, ML algorithms have seen applications as efficient alternatives to NLME modelling [55–58]. Aside from reducing the time spent on the model building process, ML algorithms can be used as a flexible approach to handle complex and high-dimensional data sources. For example, ML algorithms have been used to directly estimate PK parameters from dynamic contrast enhanced MRI images [59], or to screen almost 2000 genetic markers to find variants affecting tacrolimus exposure [60].

Our search identified many different ML algorithms used for pharmacokinetic modelling. However, we observe that most of these models suffer from problems impeding their reliable use. For example, most models take the current time and drug dose as direct inputs. Aside from leading to issues when multiple drug doses are given, it is uncertain how these inputs will be interpreted by the model. In

addition, some models are trained to predict drug concentrations at specific time points, making them unreliable when extrapolating to unseen time points. Finally, since the translation from covariates to drug concentrations can be quite non-linear, these models are prone to overfitting and might require larger data sets in order to generalise well. Based on these issues, we again suggest a set of requirements: (1) the model should be able to produce a continuous solution (i.e., extrapolate to unseen time points); (2) it should be able to adapt to complex treatment schedules (e.g., frequent dosing, mixing different types of administration); (3) it should be able to handle differences in the timing and the number of measurements per patient; and (4) should be reasonably interpretable. Below, we discuss several of the ML algorithms obtained from our literature search and identify two reliable methods for predicting drug concentrations.

2.4.1.1 *Evaluation of Different Approaches*

A basic strategy has been to directly predict the concentration-time response based on patient characteristics (e.g., covariates), the dose, and the current time point of interest [55, 57]. By predicting a single concentration, we can make independent predictions at each time point per patient. This way, we can meet requirement one and three. In order to satisfy requirement two, we must treat each dosing event as independent and add the remaining concentration from the previous dosing event to the current prediction. A problem with this approach is that the prediction does not represent the total concentration of drug in the body (usually only blood levels) so we lose information about drug accumulation in peripheral tissue. In addition, we assume that the model will learn to predict drug exposure based on the covariates, make adjustments based on the dose, and use the supplied time point to obtain the concentration along the time dimension. It is impossible to completely validate that the model uses these quantities as assumed.

We can however easily show that this approach is unreliable. We performed a simulation study using a neural network to predict real-life warfarin concentrations based on patient age, sex, the dose given at to, and the time point for which to evaluate (see Appendix 2.B for implementation details). The neural network can provide a continuous solution (figure 2.2A), and is reasonably able to represent the kinetics of warfarin (e.g., it seems to recognise its absorption and early distribution behaviour). However, when we extend the time

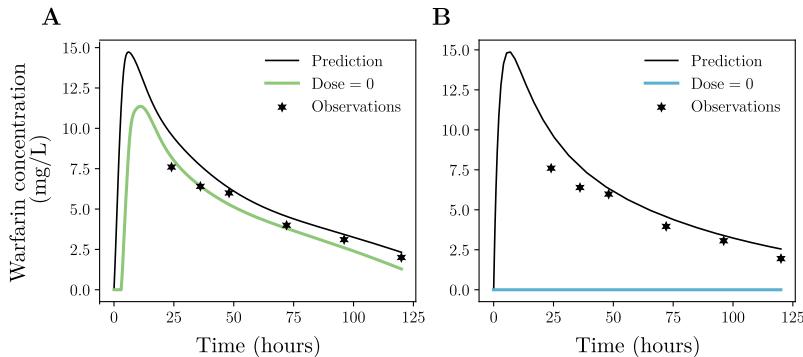


Figure 2.2: Examples of predicting drug concentrations using neural networks. Concentration-time curves for a single test set patient are shown as predicted using naive (A) and ODE-based (B) neural networks. Model prediction when artificially setting the dose to zero is depicted by the coloured lines. Stars represent the measured warfarin concentrations for the patient.

frame beyond what was seen during training, we found that the model incorrectly predicts an increase in the exposure (data not shown). In addition, when artificially setting the dose to zero, the model still predicts a response. Admittedly, we can use data augmentation to learn the neural network to predict no exposure when the dose is zero, or when the time point is long after dose administration. We cannot however augment the data with counterfactual cases (specifically with respect to the given dose) and thus the method is inherently unreliable.

Other approaches have used ML to learn optimal dose or AUC instead of a full concentration-time response [61–63]. These approaches have their own issues. They will likely be more accurate when measurements are provided as input, resulting in problems relating to requirement three. In addition, it is more difficult to interpret the credibility of the current prediction. In our previous example we could identify problems with the concentration-time curve, but in the case of direct AUC or optimal dose predictions it is more difficult to for example validate the prediction based on visualisations. Interpretation of the model using covariate importance scores can also be difficult [63]. Determination of the AUC or optimal dose based on a prediction of a full concentration-time curve, which can be verified by the observed measurements, will likely still be a more reliable approach.

The next strategy might thus be to use more complex ML algorithms better suited to time-series predictions, such as recurrent neural net-

works [56, 58, 64, 65]. These studies suggest that these methods can indeed accurately predict the changing drug concentration over time in the context of multiple dosing events [56, 58, 65]. However, *Lu et al.*, found that such methods did not extrapolate well to unseen dosing schedules [58]. Alternatively, the authors suggest a neural-ODE based approach. Here, a neural network is used to encode the covariates into a latent variable space z , which serves as the input to an ordinary differential equation (ODE) solver. This solver is a neural-ODE, a special form of recurrent neural network that learns to represent continuous dynamics similar to an ODE [66]. The neural-ODE is used to explicitly integrate the dosing and timing information. The resulting latent variables adjusted for the current time point are then fed into a decoder network, which produces concentration predictions. They show that this approach does correctly extrapolate to unseen dosing schedules, while also identifying no exposure when the dose is artificially set to zero [58]. This model fits our first three requirements. However, the model architecture is quite complex, and can be difficult to interpret.

We have also recently proposed a similar approach where we directly combine neural networks and ODEs [67]. Here, we also use an encoder network to transform the covariates into latent space variables. In contrast to the neural-ODE approach, we explicitly formulate an ODE system based on compartment models. Dosing events are then used to perturb this ODE system, which directly outputs concentration predictions at the desired time points. This offers several benefits over a neural-ODE: first, by explicitly defining drug kinetics we might reduce the data required to fit the model by imposing explicit constraints. Second, the latent variables now represent PK parameters (e.g., clearance or volume of distribution), which can be compared to previous results. Finally, since we are using a known ODE system, model predictions are credible and interpretable. This way, the method is a better match to requirement four compared to neural-ODE based methods.

2.4.1.2 Considerations

The last two approaches indicate that ODE-based ML methods are more reliable for the prediction of drug concentrations. In figure 2.2, we present a comparison between a naive and ODE-based architecture. We see that only the latter correctly identifies the absence of concentration response when the dose is set to zero (figure 2.2B). A neural-ODE can be used to learn the kinetics underlying drug exposure, whereas

an explicit ODE system can be used when prior knowledge is available. It is of interest to compare the performance of both these methods.

One remaining opportunity lies with the characterisation of prediction uncertainty. In NLME models, the estimation of residual IIV allows for MAP estimation of the PK parameters to correct the prediction based on concentration measurements. Adding this functionality to the above ODE-based models might be of interest for encouraging their adoption in general practice [68]. One approach for obtaining an estimate of predictive and parameter uncertainty are deep ensembles [69]. Here, the predictions of multiple randomly initialised neural networks are combined, and the mean and variance of predictions is presented. This approach is simple to implement and might outperform methods that explicitly estimate parameter uncertainty, such as Bayesian neural networks [70].

Finally, one important aspect of the pharmacometrics pipeline is simulation. As we have shown by removing the dosing event for the neural networks in our example, actively searching for errors in our model is essential for its evaluation. Learning if new patients are different from the data that the model was trained on might help to provide intuition about cases where we trust model output and cases we do not. Simulation is an important approach for facilitating such analyses. We further discuss the topic of model validation in Section 2.5.

2.4.2 Machine Learning for Predicting Treatment Effects

In the wake of -omics research, the interest to personalise patient treatment based on gene or protein expression profiles has increased greatly. High-dimensional data sources stress classical statistical modelling approaches, and many have turned to ML-based approaches [7, 26, 27]. In addition, some conventional methods such as the cox proportional hazard model, assume linear relationships with covariates. This has prompted the development of tree-based survival models [8], which might be better suited to problems where non-linear interactions can be expected [48]. In the following sections we will focus on the application of ML algorithms for exposure–response modelling (PK/PD models) and survival analysis (time-to-event models).

2.4.2.1 *Exposure-Response Modelling*

Exposure-response modelling involves the prediction of treatment effect in relation to the current dose or concentration of an administered drug. It is similar to PK modelling in that it often involves the use of differential equations for describing the dynamics of drug action (e.g., target-site distribution or target binding). Likewise, it is possible to use ODE-based neural network architectures to learn the effects of covariates on model parameters. However, the assumptions underlying the chosen ODE system are often weaker than in the case of PK models [71].

One can discern two types of model components: drug-specific and biological system-specific properties [72]. Drug-specific properties, such as receptor affinity, can be estimated from in vitro data. Biological system-specific properties, such as protein or receptor expression, can only be measured in vivo and can be highly variable between individuals. The latter properties are thus especially sensitive to errors in modelling assumptions. In addition, these effects are often governed by highly non-linear relationships [72]. One approach can be to use neural-ODE based models to learn the relationship between exposure and response from data [73]. Novel interpretable methods have also been described to infer such physical relationships from data [74]. However, in many cases such a direct relationship offers an overly simplistic representation of the biological situation. Alternatively, we can explicitly define part of the ODE system and use neural-ODE to estimate unknown components [75]. This allows the user to explicitly include reasonably certain model components (e.g., drug-specific properties), while neural-ODEs estimates more the more variable and complicated biological system-specific properties from data. It might be of interest to compare such approaches to classical PK/PD approaches.

Novel approaches have also aimed at improving the estimation of biological system-based effects, for example by extrapolating from cell-line data or animal models [76–78]. These approaches allow for more frequent measurement of treatment endpoints, and can be used to estimate otherwise difficult to obtain quantities (e.g., spatial distribution of drug in the target tissue [77]). As an example, patient-derived cancer xenograft models can be used to characterise the concentration-dependent effect of drugs on their target based on tumour growth data [76]. Obtaining such results from in vivo patient data would not

only be complicated but also undesirable due to the high frequency by with which tumour biopsies would have to be performed.

Adoption of ML algorithms in the context of exposure-response modelling hold exciting opportunities. For example, a recent study describes a method for predicting the drug response of thousands of cancer cell lines based on mutations and expression profiles [79]. Another study describes a method for quantifying the individual variability in tumour dose-response while also identifying important biomarkers [80]. ML techniques can also be used to learn PK or PD (e.g., drug absorption or receptor binding) parameters based on quantitative structure-activity relationships [81]. In short, there have already been many diverse applications of ML in this field, and we expect them to further increase in the future.

2.4.2.2 *Survival Analysis*

Time-to-event analysis refers to a set of methods that aim to describe the probability of a specified outcome occurring over time. In the case where only a single event is possible per individual (i.e., survival analysis), non-parametric methods such as the Kaplan-Meier estimator are used to estimate the distribution describing the proportion of individuals who have 'survived' over time. These methods allow for the statistical comparison of the efficacy of two competing treatment modalities. Often, we are also interested how covariates affect this efficacy. The standard method for estimating of such effects is the Cox proportional hazard model. Here, the covariates are assumed to affect the hazard in a proportional manner. However, this assumption might be too limiting for a flexible analysis of high-dimensional data, complex time-dependent effects, or multi-state survival models. One might instead turn to ML for learning the effect of covariates or the underlying model structure. As we have mentioned, the random survival forest model has been proposed for performing non-linear analysis of covariates. The random survival forest was suggested to obtain either similar or lower error compared to Cox proportional hazard models [8]. Recently, deep learning approaches have also been proposed for survival analysis [82, 83]. These were generally suggested to obtained higher accuracy (in terms of concordance index) when compared to Cox models and survival forests on several clinical data sets. In addition, these approaches allow for the calculation of the individual risk of prescribing a certain treatment [82]. In Cox models, this risk is constant unless treatment interaction effects are explicitly included,

which can be complicated. The neural network-based approach produced treatment recommendations that led to a higher rate of survival compared to random survival forests [82].

Recurrent neural networks have been suggested as a method for estimation of the effect of time-dependent covariates. These methods were again found to outperform previous methods (including neural networks) in terms of concordance index [84, 85]. By predicting the individual risk based on current and previous information at discrete time intervals, these methods might improve learning of time-dependent effects.

Multi-state models step away from the usual alive-death dichotomy and instead specify disease progression into intermediary (non-fatal) or competing states [86]. In oncology for example, this allows for the categorisation into induction, relapse, remission, and deceased states. This allows for prediction of the risk of relapse following complete remission or survival in the case of relapse [87]. Specification of the dynamics between different states and the influence of covariates might require strong assumptions. A generalise approach used neural-ODE to learn the likelihood of being in each state over time [88]. This approach obtained improved performance over multi-state Cox models in a competing risk setting.

2.4.2.3 Considerations

There are many interesting avenues exploring novel applications of ML in exposure-response modelling. The onset of Big Data has resulted in many opportunities for using more advanced and computationally efficient methods for analysing these data. However, some of these tools might still remain at the fringe due to their complexity. Domain-specific reviews providing an overview of recently developed algorithms might help to provide guidelines for optimal strategies for analysis and validation [81]. Without the availability of model code or comprehensive tutorials on the use of complex ML models, adoption of these methods will likely remain limited.

An important consideration for the use of ML for survival analysis is whether the current data set supports such analyses. Small data sets or those without frequent measurements of the covariates over time might lack the power to correctly describe non-linear effects. As a result, we would recommend evaluating multiple different models for the task at hand. For example, for some data sets, Cox models either performed equal to or better than neural network-based approaches

[85]. This could be the case in smaller data sets, or when the data does not support more complex models. In such scenarios, Cox models might be preferred due to their improved interpretability. One might also prefer Cox models when model interpretation is of the highest importance.

2.5 MODEL VALIDATION

2.5.1 Choosing a Validation Strategy

An essential component of any analysis using ML is a model validation strategy. Arguably, performing model validation is also more generally advisable in the context of pharmacometrics. In contrast to conventional statistical methods however, ML algorithms such as neural networks are extremely flexible. Even neural networks with a single hidden layer are considered to be universal function approximators, meaning that they can fit any data arbitrarily well [89, 90]. This flexibility results in a high risk of “overfitting”, a phenomenon where the resulting model is completely tailored to the current data set such that it generalises poorly. It is thus important to validate the generalisability of a ML model before it can be used in practice. Arguably the best validation method is to determine the predictive accuracy on independent data sets. Unfortunately, data are often limited. In this section, we report on alternatives for performing model validation.

2.5.1.1 Options for Estimating Model Generalisability

In the most simple case the data set is divided in a “train” and “test” set. The train set is used to fit the model, whereas the test set is used to estimate the accuracy of the model. In ML, usually a split using roughly 70–80% of data for training and 20–30% as test data is advised. This is however largely dependent on the size of the test set as it should contain a representative number of samples. Some ML models have additional parameters (i.e., hyperparameters) that can be tuned in order to affect performance. When performing such optimization, the data set should be split in three parts: a train set (for fitting the model), a “validation” set (for determination of the performance of the current hyperparameters), and a test set (for determination of the accuracy of the final model). A similar approach is be advisable when performing covariate selection.

Performing a single random split of the data set can be a poor estimate of model generalisability. For this reason, the accuracy is often evaluated on multiple train/test splits and their results are pooled. We will discuss three such techniques for estimating the generalisation error: random sub-sampling without replacement, bootstrapping (sub-sampling with replacement), and k -fold cross validation. A schematic overview of the three methods is provided in figure 2.3.

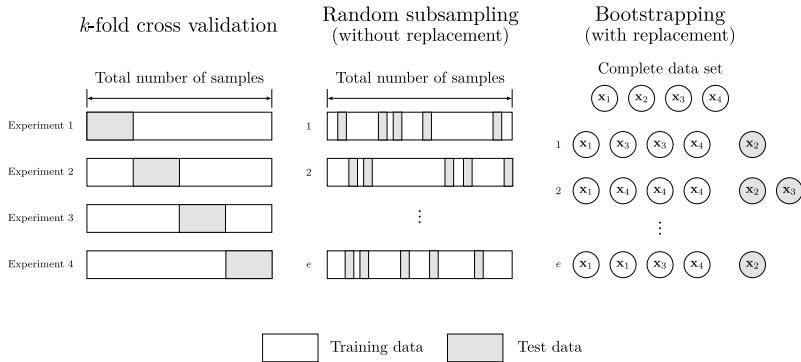


Figure 2.3: Examples of methods for estimation of model generalisation accuracy. Schematic overview of three common validation strategies: k -fold cross validation, random sub-sampling, and bootstrapping (with replacement). The white shapes denote the training data, whereas grey shapes denote testing data. Here, e represents the total number of experiments to run.

In random sub-sampling without replacement (also known as Monte Carlo cross validation), the model is fit to a random split of the data set multiple times, model accuracy is evaluated on the corresponding test sets, and the results are pooled. Since we are sampling without replacement, each sample occurs only once in either the training or test set. Crucially, one should understand that this leads to biased estimates of the true population mean and its standard error. This is because the samples in each split are not independent, and thus violate the Central Limit Theorem. Optionally, one can use the finite population correction factor to improve estimates. However, since the choice of validation strategy is independent of experimental design, possible problems can simply be avoided by performing a bootstrap instead.

In bootstrapping, samples are taken with replacement, resulting in independent samples. Usually n (size of the original data set) samples are taken, and the samples that are not in the training set are used

as the test set. Again the results are pooled for a large number of replicates. This estimate can be reliable when all the models converge, which is often not a problem in ML.

Finally, in k -fold cross validation, the data are partitioned into k *folds*, or subsets, and the model is trained on $k - 1$ folds. The remaining fold is used to estimate test accuracy. Models are fit iteratively so that each fold is used for testing once. k -fold cross validation is also frequently performed in the context of hyperparameter optimization.

2.5.1.2 Considerations

In general, drawing a consensus on the best method for estimating model generalisability is difficult. A downside of random subsampling and bootstrapping is the large computational cost associated with fitting a large number of models. In addition, when sampling with replacement, the number of unique samples in the training set is reduced, which may lead to higher bias of predictions in smaller data sets [91, 92]. A similar issue can occur when performing k -fold cross validation with a low number of folds [92]. The other extreme, known as leave-one-out cross validation (LOOCV; where $k = n$), has been suggested to have the best bias-variance trade-off when compared to the other methods [91, 92]. Performing ten-fold cross validation leads to similar results compared to leave-one-out cross validation at a lower computational cost [91, 92]. The latter is especially relevant as data set size increases in size (as models have to be fit). Papers have also reported on the inconsistency of LOOCV, specifically that selection of the true data generating model does not actually improve as data set size increases [93].

Another important consideration when estimating model generalisability is to prevent data leakage. When multiple observations are available per patient, a simple random split might result in different observations of a single individual appearing in both the train and test set. These observations should be grouped to prevent information leakage. Care should also be taken when optimising hyperparameters. The data that is used to test the current set of hyperparameters should not include samples from the final test set. This means that the data set should first be divided in a train and test set. The hyperparameters can then be optimised by performing k -fold cross validation on data from the train set only. The accuracy of the best model from the cross validation (containing the optimal set of hyperparameters) is then evaluated on the test set. This entire process can also be repeated for multiple

test sets, essentially performing an additional (outer) cross-validation. This approach also estimates sensitivity of the hyperparameters to random sub-sets of the data.

Another point to consider is that creating random subsets of the data can exacerbate class imbalances. For example, an algorithm trained to diagnose disease (classifying samples in "no disease" and "disease") can present an inflated accuracy if the model always predicts no disease while the test set does not contain many samples from the disease group. This can often be the case for rare diseases. Alternatively, in case-control studies, train sets can be saturated with control patients, resulting in a model that is unable to make accurate predictions for the case group. In such situations, the data can be stratified so that class proportions are roughly similar in each fold. Care should again be taken to prevent data leakage; data should not be stratified based on the independent variables as this results in more similar train and test sets.

There likely is no single best method of estimating model generalisability. In many cases, k -fold cross validation might be preferred when bootstrapping encompasses a too high computational cost. When choosing cross validation, the most important aspect is to choose a suitable value of k . Arlot and Celisse provide an excellent survey on model selection using cross validation [94]. Their findings may aid in choosing the appropriate cross validation procedure on a per-problem basis.

2.6 MODEL INTERPRETATION

Explainable artificial intelligence has emerged as an important sub-field of ML research. Especially with respect to the adoption of ML for medical applications, understanding why a certain prediction is made is crucial for instilling trust. As an example, model interpretation methods can be used to indicate regions-of-interest underlying predictions for medical image classification [95, 96]. There are two types of explanation methods: model-specific and model-agnostic. Model-specific explanation methods for example involve using the regression coefficients from linear models to explain the proportional relationship between covariates and the dependent variable. Although these coefficients have a straightforward meaning, they are not necessarily true; correlated covariates can complicate the estimation of true covariate effects. In more complex models, such as neural networks, the meaning of the model parameters is not immediately obvious.

Although neural network specific explanation methods exist [97], generally model-agnostic explanation methods are used. These methods themselves can be considered black-box as they aim to replace the complex model by a more simple, interpretable model.

There have been numerous suggestions for interpretation frameworks aimed at explaining ML model output, including Local Interpretable Model agnostic Explanations (LIME), Deep Learning Important FeaTures (DeepLIFT), and SHapley Additive exPlanations (SHAP) [97–99]. An overview of popular methods is provided by *Holzinger et al.* [100]. This reference also provides short discussions of each method which may provide intuition on what method to use for what goal. It can be difficult to understand how the function of each of these methods affects model interpretation. It is possible that using different frameworks on the same model results in different explanations. It has thus been of interest to find theoretical support for these methods. One method that aims to offer theoretical guarantees is SHAP [99].

SHAP has already seen use in pharmacokinetic modelling. One study used SHAP for the identification of important covariates when using neural networks to predict cyclosporin A clearance [101]. Aside from covariate importance, which only provides a limited interpretation of the model, SHAP can also be used to visualise covariate relationships [54]. To present an example, we performed a SHAP analysis on the prediction of warfarin absorption rate (k_a) by the previous discussed ODE-based neural network (implementation details in Appendix 2.B.3). In figure 2.4, we depict the relationship between age and k_a , stratified by sex, as represented by SHAP values. Since we only have a single continuous and categorical variable as input to our neural network we can also obtain their exact functional relationships. In cases when more covariates are included, this is generally not possible. The model predicts a different effect of age on k_a for males and females (see figure 2.4). The SHAP values allow for the evaluation if the relationships adhere to biological expectations of covariate effects. However, since there are only a few female patients, we should take caution when performing such evaluations. Although the SHAP values seem to be able to represent the effect of the covariates well, extrapolating to unseen samples might be unreliable. Other approaches have been developed in order to estimate the uncertainty of out-of-distribution samples [102]. The use of only a single explanation method might thus not be enough for the complete evaluation of ML models.

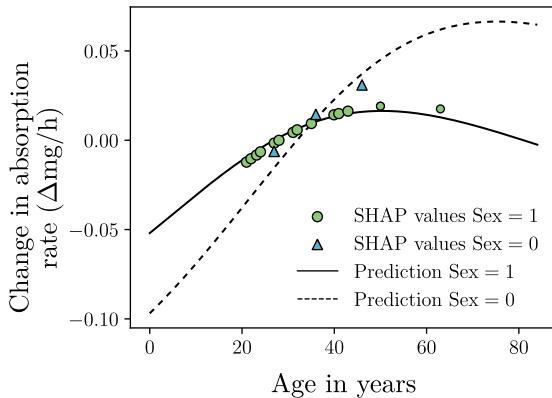


Figure 2.4: Example of using SHAP for model interpretation. Change in warfarin absorption rate ($\Delta\text{mg}/\text{h}$) prediction by the neural network as estimated by SHAP values. Here, circles represent the SHAP values calculated for male patients, whereas triangles represent SHAP values calculated for female patients. Lines represent the neural network predicting when fixing patient sex (solid for male, and dashed for female) and predicting absorption rate based on different values for age.

2.6.1 Considerations

There are many available explanation frameworks for ML models. Here, we have chosen to discuss only one technique to illustrate how these methods can be used. In the case of ODE-based methods, model-agnostic methods are useful to visualise the effect of covariates. However, it is possible that such methods alone are not sufficient for use in clinical practice. It might also be of interest to know if each individual prediction can be trusted, especially when predicting for samples that are different from training data.

Due to the large number of available explanation methods and difficulties with representing their accuracy, choosing the correct method can be a daunting task. In most cases, we can only lean on some theoretical guarantees or expected behaviour on simple examples [103]. Molnar *et al.*, present an excellent overview of pitfalls of these methods and how to resolve them [104]. Since most of the model-agnostic interpretation methods operate in a similar fashion (i.e., they perturb data and evaluate the effect on predictions), they share similar pitfalls. This study also makes the important suggestion that there again is no one-size-fits-all method.

Another great reference is the work by Yang *et al.* [105]. Here, the authors first outline frequently used concepts in model interpretation, after which they provide two showcases demonstrating how these concepts can be applied for explaining ML model output. This study shows how these techniques can provide insight into the strengths and weaknesses of ML models.

2.7 MAIN POINTS

We have discussed several recent applications of ML algorithms in the context of pharmacometrics. More specifically, we have presented how ML techniques can and have been used for data imputation, dimensionality reduction, unsupervised clustering, covariate selection, drug concentration prediction, and treatment response prediction. In general, tree-based models and neural networks were the most frequently used algorithms for these purposes. Most papers report an improvement in performance when comparing the use of these methods to classical approaches. In addition, more complex architectures, which were most frequently based on neural networks, were suggested to be the most accurate. More research is however needed to compare these methods to classical approaches, such as NLME models.

We have started our discussion with the application of ML methods for data preparation. With respect to missing data imputation, the literature suggests lower bias for estimated model parameters when using multiple imputation compared to single imputation. Several ML-based methods have been suggested for imputation of missing data based on regression of observed covariates. It is still unclear however if the added complexity of these methods actually leads to improved estimation of missing data. Evaluation of such methods in the context of the smaller data sets often seen in pharmacometrics is thus required. We have briefly discussed methods for dimensionality reduction. Such approaches might be interesting for facilitating the analysis of complex genetic or proteomics data. However, their benefit compared to using ML to detect influential covariates is not obvious. Although the latter approach can result in the loss of information on covariate dependencies, its results are more easily applicable.

Next, we discussed the application of ML techniques for hypothesis generation. Studies have used unsupervised clustering techniques for the detection of patient subgroups from data. These subgroups can for example be used to generate hypotheses regarding differences in treatment response between patients. ML methods have also been

used for the detection of influential covariates. A study has suggested that covariate importance scores produced by the random forest can be used to obtain a better selection of covariates compared to step-wise methods [42]. This and other methods do not yet offer a complete alternative to step-wise methods, but are useful for producing an initial set of hypotheses regarding important covariates to consider for inclusion. Methods for search space optimization such as genetic algorithms are a promising approach for improving the selection of model components that lead to the best performance. This approach requires the selection of a fitness function to control model complexity, which can be difficult to choose. More research is needed for an empirical method for selecting an appropriate fitness function.

ML models have also been used as predictive models in the context of pharmacometrics. We make the point that ODE-based methods outperform other methods in reliability regarding the prediction of drug concentrations. We have showcased this point by using a simple example of how naive methods can misinterpret drug doses when they are passed directly as input. Next, we discuss several ML-based methods for predicting treatment response and efficacy. Again, ODE-based methods show potential for improving prediction reliability, especially in the case of PK/PD modelling. There have also been ML-based approaches for survival analysis. It is the question however if these are appropriate for every analysis, as more complex models might not always result in improved performance.

Finally, we have discussed model validation. Due to the flexibility of many of the discussed methods, deciding on a suitable model validation strategy should be an integral part of the modelling process. The generalisation performance of the model is an important metric for judging its appropriateness. Validation of accuracy on a high quality external data-set is often regarded as one of the best options. It is however not clear what is the next best alternative when such data are not available. We would like to urge pharmacologists that are interested in using ML to first consider whether their use case supports the use of these tools. In our experience, we found that imposing constraints on these models (for example based on prior knowledge using ODEs) can help in improving performance when data are sparse. We also want to stress the importance of evaluating what the ML model has learned. Examples include the analysis of the most important covariates, or performing sensitivity analysis (e.g., using methods such as SHAP) with respect to the model parameters. Understanding how the model makes its predictions allows for the

removal of any biases, and adapting model regularisation to prevent it from making "mistakes". Examination of under-sampled regions of the input space can provide insight into the extend to which model predictions can be trusted. Specifically training the model on new data from under-sampled patients will help improve generalisability in the long run.

In the coming years, our expectation is that the number of studies exploring the use of ML in pharmacometrics will keep increasing. Perhaps some of the methods mentioned in this review have already become a standard part of the pharmacometrician's tool kit in the near future. This could be the time for researchers interested in ML to educate themselves in ML concepts and, perhaps, to develop new model architectures better suited to problems in the field of pharmacometrics.

REFERENCES

- [1] Stuart L Beal and Lewis B Sheiner. "Estimating population kinetics." In: *Critical reviews in biomedical engineering* 8.3 (1982), pp. 195–222.
- [2] Mary J Lindstrom and Douglas M Bates. "Nonlinear mixed effects models for repeated measures data". In: *Biometrics* (1990), pp. 673–687.
- [3] Amy Racine-Poon and Adrian FM Smith. "Population models". In: *Statistical methodology in the pharmaceutical sciences* (1990), pp. 139–162.
- [4] Ayyappa Chaturvedula, Stacie Calad-Thomson, Chao Liu, Mark Sale, Nandu Gattu, and Navin Goyal. "Artificial intelligence and pharmacometrics: time to embrace, capitalize, and advance?" In: *CPT: pharmacometrics & systems pharmacology* 8.7 (2019), p. 440.
- [5] Mason McComb, Robert Bies, and Murali Ramanathan. "Machine learning in pharmacometrics: Opportunities and challenges". In: *British Journal of Clinical Pharmacology* 88.4 (2022), pp. 1482–1499.
- [6] Alireza Osareh and Bita Shadgar. "Machine learning techniques to diagnose breast cancer". In: *2010 5th international symposium on health informatics and bioinformatics*. IEEE. 2010, pp. 114–120.
- [7] David GP van IJzendoorn, Karoly Szuhai, Inge H Brinaire-de Bruijn, Marie Kostine, Marieke L Kuijjer, and Judith VMG Bovée. "Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas". In: *PLoS computational biology* 15.2 (2019), e1006826.
- [8] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. "Random survival forests". In: *Ann. Appl. Stat.* 2 (2008), 841–860. doi: 10.1214/08-AOAS169..
- [9] Solveig Badillo, Balazs Banfa, Fabian Birzele, Iakov I Davydov, Lucy Hutchinson, Tony Kam-Thong, Juliane Siebourg-Polster, Bernhard Steiert, and Jitao David Zhang. "An introduction to machine learning". In: *Clinical pharmacology & therapeutics* 107.4 (2020), pp. 871–885.

- [10] Hulin Wu and Lang Wu. "A multiple imputation method for missing covariates in non-linear mixed-effects models with application to HIV dynamics". In: *Statistics in medicine* 20.12 (2001), pp. 1755–1769.
- [11] Åsa M Johansson and Mats O Karlsson. "Comparison of methods for handling missing covariate data". In: *The AAPS journal* 15 (2013), pp. 1232–1241.
- [12] Bräm D.S., Nahum U., Atkinson A., Koch G., and Pfister M. "Opportunities of Covariate Data Imputation with Machine Learning for Pharmacometrists in R". In: *Proceedings of the 30th Annual Meeting of the Population Approach Group in Europe. Abstract 9982*. 2022. URL: www.page-meeting.org/?abstract=9982.
- [13] Gustavo EAPA Batista and Maria Carolina Monard. "An analysis of four missing data treatment methods for supervised learning". In: *Applied artificial intelligence* 17.5-6 (2003), pp. 519–533.
- [14] Daniel J Stekhoven and Peter Bühlmann. "MissForest—non-parametric missing value imputation for mixed-type data". In: *Bioinformatics* 28.1 (2012), pp. 112–118.
- [15] Anoop D Shah, Jonathan W Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. "Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study". In: *American journal of epidemiology* 179.6 (2014), pp. 764–774.
- [16] Liang Jin, Yingtao Bi, Chenqi Hu, Jun Qu, Shichen Shen, Xue Wang, and Yu Tian. "A comparative study of evaluating missing value imputation methods in label-free proteomics". In: *Scientific reports* 11.1 (2021), p. 1760.
- [17] Buuren S. and Oudshoorn K. "Flexible Multivariate Imputation by MICE". In: (1999).
- [18] Jinsung Yoon, James Jordon, and Mihaela Schaar. "Gain: Missing data imputation using generative adversarial nets". In: PMLR. 2018, pp. 5689–5698.
- [19] Pierre-Alexandre Mattei and Jes Frellsen. "MIWAE: Deep generative modelling and imputation of incomplete data sets". In: *International conference on machine learning*. PMLR. 2019, pp. 4413–4423.
- [20] Bahram Jafrasteh, Daniel Hernández-Lobato, Simón Pedro Lubián-López, and Isabel Benavente-Fernández. "Gaussian processes for missing value imputation". In: *Knowledge-Based Systems* 273 (2023), p. 110603.
- [21] Tiago JS Lopes, Ricardo Rios, Tatiane Nogueira, and Rodrigo F Mello. "Prediction of hemophilia A severity using a small-input machine-learning framework". In: *NPJ systems biology and applications* 7.1 (2021), p. 22.
- [22] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: (2020). arXiv: 1802.03426 [stat.ML]. URL: <https://arxiv.org/abs/1802.03426>.
- [23] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).
- [24] Ruizhi Xiang, Wencan Wang, Lei Yang, Shiyuan Wang, Chaohan Xu, and Xiaowen Chen. "A comparison for dimensionality reduction methods of single-cell RNA-seq data". In: *Frontiers in genetics* 12 (2021), p. 646936.

- [25] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. "Dimensionality reduction for visualizing single-cell data using UMAP". In: *Nature biotechnology* 37.1 (2019), pp. 38–44.
- [26] Diana-Maria Ciuculete, Marcus Bandstein, Christian Benedict, Gérard Waeber, Peter Vollenweider, Lars Lind, Helgi B Schiöth, and Jessica Mwinyi. "A genetic risk score is significantly associated with statin therapy response in the elderly population". In: *Clinical genetics* 91.3 (2017), pp. 379–385.
- [27] Sofia H Kanders, Claudia Pisanu, Marcus Bandstein, Jörgen Jonsson, Enrique Castelao, Giorgio Pistis, Mehdi Gholam-Rezaee, Chin B Eap, Martin Preisig, Helgi B Schiöth, et al. "A pharmacogenetic risk score for the evaluation of major depression severity under treatment with antidepressants". In: *Drug development research* 81.1 (2020), pp. 102–113.
- [28] Laura B Zwep, Kevin LW Duisters, Martijn Jansen, Tingjie Guo, Jacqueline J Meulman, Parth J Upadhyay, and JG Coen van Hasselt. "Identification of high-dimensional omics-derived predictors for tumor growth dynamics using machine learning and pharmacometric modeling". In: *CPT: pharmacometrics & systems pharmacology* 10.4 (2021), pp. 350–361.
- [29] Iasonas Kapralos and Aristides Dokoumetzidis. "Population Pharmacokinetic Modelling of the Complex Release Kinetics of Octreotide LAR: Defining Sub-Populations by Cluster Analysis". In: *Pharmaceutics* 13.10 (2021), p. 1578.
- [30] Riya Paul, Till FM Andlauer, Darina Czamara, David Hoehn, Susanne Lucae, Benno Pütz, Cathryn M Lewis, Rudolf Uher, Bertram Müller-Myhsok, Marcus Ising, et al. "Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models". In: *Translational Psychiatry* 9.1 (2019), p. 187.
- [31] Elson Tomás, Susana Vinga, and Alexandra M Carvalho. "Unsupervised learning of pharmacokinetic responses". In: *Computational Statistics* 32 (2017), pp. 409–428.
- [32] Kerstin Bunte, David J Smith, Michael J Chappell, Zaki K Hassan-Smith, Jeremy W Tomlinson, Wiebke Arlt, and Peter Tiňo. "Learning pharmacokinetic models for in vivo glucocorticoid activation". In: *Journal of Theoretical Biology* 455 (2018), pp. 222–231.
- [33] Paidamoyo Chapfuwa, Chunyuan Li, Nikhil Mehta, Lawrence Carin, and Ricardo Henao. "Survival cluster analysis". In: *Proceedings of the ACM Conference on Health, Inference, and Learning*. 2020, pp. 60–68.
- [34] Rui P Guerra, Alexandra M Carvalho, and Paulo Mateus. "Model selection for clustering of pharmacokinetic responses". In: *Computer Methods and Programs in Biomedicine* 162 (2018), pp. 11–18.
- [35] Johannes Blömer and Kathrin Bujna. "Adaptive seeding for Gaussian mixture models". In: *Pacific-asia conference on knowledge discovery and data mining*. Springer. 2016, pp. 296–308.
- [36] Frank E Harrell et al. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Vol. 608. Springer, 2001.
- [37] Jakob Ribbing and E Niclas Jonsson. "Power, selection bias and predictive performance of the population pharmacokinetic covariate model". In: *Journal of pharmacokinetics and pharmacodynamics* 31 (2004), pp. 109–134.

- [38] Jakob Ribbing, Joakim Nyberg, Ola Caster, and E Niclas Jonsson. "The lasso—a novel method for predictive covariate model building in nonlinear mixed effects models". In: *Journal of pharmacokinetics and pharmacodynamics* 34 (2007), pp. 485–517.
- [39] Malidi Ahamadi, Anna Largajolli, Paul M Diderichsen, Rik de Greef, Thomas Kerbusch, Han Witjes, Akshita Chawla, Casey B Davis, and Ferdous Gheysen. "Operating characteristics of stepwise covariate selection in pharmacometric modeling". In: *Journal of Pharmacokinetics and Pharmacodynamics* 46 (2019), pp. 273–285.
- [40] Robert Tibshirani. "The lasso method for variable selection in the Cox model". In: *Statistics in medicine* 16.4 (1997), pp. 385–395.
- [41] Phyllis Chan, Xiaofei Zhou, Nina Wang, Qi Liu, René Bruno, and Jin Y Jin. "Application of machine learning for tumor growth inhibition—overall survival modeling platform". In: *CPT: pharmacometrics & systems pharmacology* 10.1 (2021), pp. 59–66.
- [42] Emeric Sibieude, Akash Khandelwal, Jan S Hesthaven, Pascal Girard, and Nadia Terranova. "Fast screening of covariates in population models empowered by machine learning". In: *Journal of pharmacokinetics and pharmacodynamics* 48.4 (2021), pp. 597–609.
- [43] Jerome H Friedman. "Multivariate adaptive regression splines". In: *The annals of statistics* 19.1 (1991), pp. 1–67.
- [44] Dohyeong Ki, Billy Fang, and Adityanand Guntuboyina. "MARS via LASSO". In: (2024). arXiv: 2111.11694 [math.ST]. URL: <https://arxiv.org/abs/2111.11694>.
- [45] Mitov V., Kuemmel A., Gobeau N., Cherkaoui M., and Bouillon T. "Dose selection by covariate assessment on the optimal dose for efficacy—Application of machine learning in the context of PKPD". In: 2022. URL: www.page-meeting.org/?abstract=10066.
- [46] Rui Wang, Xiao Shao, Junying Zheng, Abdel Saci, Xiaozhong Qian, Irene Pak, Amit Roy, Akintunde Bello, Jasmine I Rizzo, Fareeda Hosein, et al. "A machine-learning approach to identify a prognostic cytokine signature that is associated with nivolumab clearance in patients with advanced melanoma". In: *Clinical Pharmacology & Therapeutics* 107.4 (2020), pp. 978–987.
- [47] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8 (2007), pp. 1–21.
- [48] Xiajing Gong, Meng Hu, and Liang Zhao. "Big data toolsets to pharmacometrics: application of machine learning for time-to-event analysis". In: *Clinical and translational science* 11.3 (2018), pp. 305–311.
- [49] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. "Accurate intelligible models with pairwise interactions". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013, pp. 623–631.
- [50] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. "InterpretML: A Unified Framework for Machine Learning Interpretability". In: (2019). arXiv: 1909.09223 [cs.LG]. URL: <https://arxiv.org/abs/1909.09223>.

- [51] Bies R.R., Muldoon M.F., Pollock B.G., Manuck S., Smith G., and Sale M.E. "A genetic algorithm-based, hybrid machine learning approach to model selection". In: *J. Pharmacokinet. Pharmacodyn.* 33 (2006), 195–221. doi: 10.1007/s10928-006-9004-6..
- [52] Robert R Bies, Matthew F Muldoon, Bruce G Pollock, Steven Manuck, Gwenn Smith, and Mark E Sale. "A genetic algorithm-based, hybrid machine learning approach to model selection". In: *Journal of pharmacokinetics and pharmacodynamics* 33.2 (2006), p. 195.
- [53] Emeric Sibieude, Akash Khandelwal, Pascal Girard, Jan S Hesthaven, and Nadia Terranova. "Population pharmacokinetic model selection assisted by machine learning". In: *Journal of pharmacokinetics and pharmacodynamics* 49.2 (2022), pp. 257–270.
- [54] Alexander Janssen et al. "Application of SHAP values for inferring the optimal functional form of covariates in pharmacokinetic modeling". In: *CPT: Pharmacometrics & Systems Pharmacology* 11.8 (2022), pp. 1100–1110.
- [55] Yichao Xu, Honggang Lou, Jinliang Chen, Bo Jiang, Dandan Yang, Yin Hu, and Zourong Ruan. "Application of a backpropagation artificial neural network in predicting plasma concentration and pharmacokinetic parameters of oral single-dose rosuvastatin in healthy subjects". In: *Clinical pharmacology in drug development* 9.7 (2020), pp. 867–875.
- [56] Oscar J Pellicer-Valero, Isabella Cattinelli, Luca Neri, Flavio Mari, José D Martín-Guerrero, and Carlo Barbieri. "Enhanced prediction of hemoglobin concentration in a very large cohort of hemodialysis patients by means of deep recurrent neural networks". In: *Artificial Intelligence in Medicine* 107 (2020), p. 101898.
- [57] Xiaohui Huang, Ze Yu, Shuhong Bu, Zhiyan Lin, Xin Hao, Wenjun He, Peng Yu, Zeyuan Wang, Fei Gao, Jian Zhang, et al. "An ensemble model for prediction of vancomycin trough concentrations in pediatric patients". In: *Drug design, development and therapy* (2021), pp. 1549–1559.
- [58] James Lu, Kaiwen Deng, Xinyuan Zhang, Gengbo Liu, and Yuanfang Guan. "Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens". In: *Iscience* 24.7 (2021).
- [59] Cagdas Ulas, Dhritiman Das, Michael J Thripplenton, Maria del C Valdés Hernández, Paul A Armitage, Stephen D Makin, Joanna M Wardlaw, and Bjoern H Menze. "Convolutional neural networks for direct inference of pharmacokinetic parameters: application to stroke dynamic contrast-enhanced MRI". In: *Frontiers in neurology* 9 (2019), p. 1147.
- [60] Jeong-An Gim, Yonghan Kwon, Hyun A Lee, Kyeong-Ryoon Lee, Soohyun Kim, Yoonjung Choi, Yu Kyong Kim, and Howard Lee. "A machine learning-based identification of genes affecting the pharmacokinetics of tacrolimus using the DMETTM plus platform". In: *International journal of molecular sciences* 21.7 (2020), p. 2517.
- [61] Yanyun Tao, Yenming J Chen, Ling Xue, Cheng Xie, Bin Jiang, and Yuzhen Zhang. "An ensemble model with clustering assumption for warfarin dose prediction in Chinese patients". In: *IEEE journal of biomedical and health informatics* 23.6 (2019), pp. 2642–2654.

- [62] Jean-Baptiste Woillard, Marc Labriffe, Jean Debord, and Pierre Marquet. "Tacrolimus exposure prediction using machine learning". In: *Clinical Pharmacology & Therapeutics* 110.2 (2021), pp. 361–369.
- [63] Xiaohui Huang, Ze Yu, Xin Wei, Junfeng Shi, Yu Wang, Zeyuan Wang, Jihui Chen, Shuhong Bu, Lixia Li, Fei Gao, et al. "Prediction of vancomycin dose on high-dimensional data using machine learning techniques". In: *Expert Review of Clinical Pharmacology* 14.6 (2021), pp. 761–771.
- [64] Xiangyu Liu, Chao Liu, Ruihao Huang, Hao Zhu, Qi Liu, Sunanda Mitra, and Yaning Wang. "Long short-term memory recurrent neural network for pharmacokinetic-pharmacodynamic modeling". In: *International journal of clinical pharmacology and therapeutics* 59.2 (2021), p. 138.
- [65] Dominic Stefan Bräm, Neil Parrott, Lucy Hutchinson, and Bernhard Steiert. "Introduction of an artificial neural network-based method for concentration-time predictions". In: *CPT: Pharmacometrics & Systems Pharmacology* 11.6 (2022), pp. 745–754.
- [66] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. "Neural ordinary differential equations". In: vol. 31. 2018.
- [67] Alexander Janssen et al. "Deep compartment models: a deep learning approach for the reliable prediction of time-series data in pharmacokinetic modeling". In: *CPT: Pharmacometrics & Systems Pharmacology* 11.7 (2022), pp. 934–945.
- [68] Janssen A., Leebeek F.W.G., Cnossen M.H., and Mathôt R.A.A. "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. URL: www.page-meeting.org/?abstract=9826.*
- [69] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in neural information processing systems* 30 (2017).
- [70] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. "Deep Ensembles: A Loss Landscape Perspective". In: (2020). arXiv: 1912.02757 [stat.ML]. URL: <https://arxiv.org/abs/1912.02757>.
- [71] Huixi Zou, Parikshit Banerjee, Sharon Shui Yee Leung, and Xiaoyu Yan. "Application of pharmacokinetic-pharmacodynamic modeling in drug delivery: development and challenges". In: *Frontiers in pharmacology* 11 (2020), p. 997.
- [72] Meindert Danhof, Elizabeth CM De Lange, Oscar E Della Pasqua, Bart A Ploeger, and Rob A Voskuyl. "Mechanism-based pharmacokinetic-pharmacodynamic (PK-PD) modeling in translational drug research". In: *Trends in pharmacological sciences* 29.4 (2008), pp. 186–191.
- [73] James Lu, Brendan Bender, Jin Y Jin, and Yuanfang Guan. "Deep learning prediction of patient response time course from early data via neural-pharmacokinetic/pharmacodynamic modelling". In: *Nature machine intelligence* 3.8 (2021), pp. 696–704.
- [74] Daria Kurz, Carlos Salort Sánchez, and Cristian Axenie. "Data-driven discovery of mathematical and physical relations in oncology data using human-understandable machine learning". In: *Frontiers in Artificial Intelligence* 4 (2021), p. 713690.

- [75] Zhaozhi Qian, William Zame, Lucas Fleuren, Paul Elbers, and Mihaela van der Schaar. "Integrating expert ODEs into neural ODEs: pharmacology and disease progression". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 11364–11383.
- [76] Harvey Wong, Bruno Aliche, Kristina A West, Patricia Pacheco, Hank La, Tom Januario, Robert L Yauch, Frederic J de Sauvage, and Stephen E Gould. "Pharmacokinetic–pharmacodynamic analysis of vismodegib in preclinical models of mutational and ligand-dependent Hedgehog pathway activation". In: *Clinical Cancer Research* 17.14 (2011), pp. 4682–4692.
- [77] Elizabeth C Randall, Kristina B Emdal, Janice K Laramy, Minjee Kim, Alison Roos, David Calligaris, Michael S Regan, Shiv K Gupta, Ann C Mladek, Brett L Carlson, et al. "Integrated mapping of pharmacokinetics and pharmacodynamics in a patient-derived xenograft model of glioblastoma". In: *Nature communications* 9.1 (2018), p. 4904.
- [78] JungHo Kong, Heetak Lee, Donghyo Kim, Seong Kyu Han, Doyeon Ha, Kunyoo Shin, and Sanguk Kim. "Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients". In: *Nature communications* 11.1 (2020), p. 5485.
- [79] Yu-Chiao Chiu, Hung-I Harry Chen, Tinghe Zhang, Songyao Zhang, Aparna Gorthi, Li-Ju Wang, Yufei Huang, and Yidong Chen. "Predicting drug response of tumors from integrated genomic profiles by deep neural networks". In: *BMC medical genomics* 12 (2019), pp. 143–155.
- [80] Dennis Wang, James Hensman, Ginte Kutkaite, Tzen S Toh, Ana Galhoz, GDSC Screening Team Lightfoot Howard Yang Wanjuan Soleimani Maryam Barthorpe Syd Mironenko Tatiana Beck Alexandra Richardson Laura Lleshi Ermira Hall James Tolley Charlotte Barendt William, Jonathan R Dry, Julio Saez-Rodriguez, Mathew J Garnett, Michael P Menden, et al. "A statistical framework for assessing pharmacological responses and biomarkers using uncertainty estimates". In: *Elife* 9 (2020), e60352.
- [81] Mohammad R Keyvanpour and Mehrnoush Barani Shirzad. "An analysis of QSAR research based on machine learning concepts". In: *Current Drug Discovery Technologies* 18.1 (2021), pp. 17–30.
- [82] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network". In: *BMC medical research methodology* 18 (2018), pp. 1–12.
- [83] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. "Deephit: A deep learning approach to survival analysis with competing risks". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.
- [84] Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. "Deep recurrent survival analysis". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 4798–4805.
- [85] Eleonora Giunchiglia, Anton Nemchenko, and Mihaela van der Schaar. "Rnn-surv: A deep recurrent model for survival analysis". In: *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4–7, 2018, Proceedings, Part III* 27. Springer. 2018, pp. 23–32.

- [86] Luís Meira-Machado, Jacobo de Uña-Álvarez, Carmen Cadarso-Suárez, and Per K Andersen. "Multi-state models for the analysis of time-to-event data". In: *Statistical methods in medical research* 18.2 (2009), pp. 195–222.
- [87] Moritz Gerstung, Elli Papaemmanuil, Inigo Martincorena, Lars Bullinger, Verena I Gaidzik, Peter Paschka, Michael Heuser, Felicitas Thol, Niccolo Bolli, Peter Ganly, et al. "Precision oncology for acute myeloid leukemia using a knowledge bank approach". In: *Nature genetics* 49.3 (2017), pp. 332–340.
- [88] Stefan Groha, Sebastian M Schmon, and Alexander Gusev. "A General Framework for Survival Analysis and Multi-State Modelling". In: (2021). arXiv: 2006.04893 [stat.ML]. URL: <https://arxiv.org/abs/2006.04893>.
- [89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5 (1989), pp. 359–366.
- [90] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization". In: (2017). arXiv: 1611.03530 [cs.LG]. URL: <https://arxiv.org/abs/1611.03530>.
- [91] Ron Kohavi et al. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Proceedings of the IJCAI*. Vol. 14. 2. Montreal, Canada. 1995, pp. 1137–1145.
- [92] Annette M Molinaro, Richard Simon, and Ruth M Pfeiffer. "Prediction error estimation: a comparison of resampling methods". In: *Bioinformatics* 21.15 (2005), pp. 3301–3307.
- [93] Quentin F Gronau and Eric-Jan Wagenmakers. "Limitations of Bayesian leave-one-out cross-validation for model selection". In: *Computational brain & behavior* 2.1 (2019), pp. 1–11.
- [94] Sylvain Arlot and Alain Celisse. "A survey of cross-validation procedures for model selection". In: *Statistics Surveys* 4.none (2010). ISSN: 1935-7516. doi: 10.1214/09-ss054. URL: <http://dx.doi.org/10.1214/09-SS054>.
- [95] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. "Mdnet: A semantically and visually interpretable medical image diagnosis network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6428–6436.
- [96] Amitojdeep Singh, Abdul Rasheed Mohammed, John Zelek, and Vasudevan Lakshminarayanan. "Interpretation of deep learning using attributions: application to ophthalmic diagnosis". In: 11511 (2020), pp. 39–49.
- [97] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences". In: *International conference on machine learning*. PMIR. 2017, pp. 3145–3153.
- [98] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.
- [99] Scott M. Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: (2017), 4768–4777.

- [100] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. "Explainable AI methods-a brief overview". In: *International workshop on extending explainable AI beyond deep models and classifiers*. Springer. 2022, pp. 13–38.
- [101] Chika Ogami, Yasuhiro Tsuji, Hiroto Seki, Hideaki Kawano, Hideto To, Yoshiaki Matsumoto, and Hiroyuki Hosono. "An artificial neural network- pharmacokinetic model and its interpretation using Shapley additive explanations". In: *CPT: pharmacometrics & systems pharmacology* 10.7 (2021), pp. 760–768.
- [102] Danijar Hafner, Dustin Tran, Timothy Lillicrap, Alex Irpan, and James Davidson. "Noise contrastive priors for functional uncertainty". In: *Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 905–914.
- [103] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.
- [104] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. "General pitfalls of model-agnostic interpretation methods for machine learning models". In: *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer. 2020, pp. 39–68.
- [105] Guang Yang, Qinghao Ye, and Jun Xia. "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond". In: *Information Fusion* 77 (2022), pp. 29–52.
- [106] Sven Björkman, MyungShin Oh, Gerald Spotts, Phillip Schroth, Sandor Fritsch, Bruce M Ewenstein, Kathleen Casey, Kathelijn Fischer, Victor S Blanchette, and Peter W Collins. "Population pharmacokinetics of recombinant factor VIII: the relationships of pharmacokinetics to age and body weight". In: *Blood, The Journal of the American Society of Hematology* 119.2 (2012), pp. 612–618.
- [107] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

APPENDIX

2.A MACHINE LEARNING FOR COVARIATE SELECTION

2.A.1 *Data*

A data set of severe haemophilia A patients receiving a single dose of 50 IUkg^{-1} blood clotting factor VIII (FVIII) concentrate was simulated based on a previous pharmacokinetic (PK) model [106]. Pharmacokinetic time profiles were based on a two compartment model with inter-individual variation on the clearance and central volume parameters. Typical clearance was estimated based on patient weight (power function) and age (linear function), whereas central volume was based on weight only (power function). Next, we sampled values to produce individual estimates of the clearance parameter. Patient age was sampled from a Uniform(1.1, 66.0) distribution. We fit a Gaussian Process to predict patient weight based on age. The simulated age values were subsequently used to sample corresponding weight estimates from the Gaussian Process. Finally, we augmented the data set with 48 noise covariates from a standard normal distribution.

2.A.2 *Models*

We fit a LASSO (sklearn python package, v1.0.2 [107]), MARS (py-earth python package, v0.1.0 [107]), random forest model (sklearn python package, v1.0.2 [107]), and explainable gradient boosting model (interpret python package, v0.2.7 [50]) to predict the individual clearance estimates based on the augmented set of covariates. A ten-fold cross validation was performed. The test sets were used to calculate the accuracy of model predictions using the root mean squared error (RMSE). MARS was the most accurate model with a mean RMSE of $65.5 \pm 10.3 \text{ mL/h}$, compared to 67.3 ± 9.74 for LASSO, 68.3 ± 8.45 for random forest, and 75.4 ± 16.3 for the explainable gradient boosting model. Next, we visualised LASSO coefficients and normalised importance scores for the random forest, MARS and explainable gradient boosting models. All four methods correctly identified the true covariates as most important, while noise covariates were less influential

(see figure 2.1A–D). For the MARS and explainable gradient boosting model, the learned function which approximates the effect of weight on clearance was also visualised (see figure 2.1E,F).

2.B NEURAL NETWORK FOR DRUG CONCENTRATION PREDICTION

2.B.1 *Data*

A publicly available data set of 32 patients who received warfarin was used to depict how neural networks can be used to predict drug concentrations. The data set contained a total of 251 warfarin concentration measurements, with a median of six measurements per patient. Every patient was given a single dose of warfarin at $t = 0$ and measurements were performed at $t \in \{0.25, 0.5, 1, 2, 4, 6, 12, 24, 48, 72, 96, 120\}$. Available covariates were patient weight, age, and sex.

2.B.2 *Prediction of Warfarin Concentrations*

The patients from the real-world warfarin data set were split into a training ($n = 22$) and testing ($n = 10$) set. A neural network with two hidden layers (with, 16, and 4 neurons, respectively) was fit to predict single warfarin concentrations based on patient age, sex, dose, and time point. The swish activation function was used for the hidden layers. Final layer used the softplus activation function in order to constrain output to be positive. Next, we trained an ODE-based neural network [67], which predicts the parameters of a set of differential equations representing a compartment model. The same neural network architecture was used (hidden layer of 16 neurons and output layer of 4 neurons). Real-world warfarin concentrations were predicted based on patient age and sex. This model was trained on the same train set and accuracy was evaluated on the test set. Naive neural network achieved slightly lower RMSE 1.41 IUmL^{-1} compared to the ODE-based neural network (1.60 IUmL^{-1}).

The results for the same test set patient was plotted for the naive neural network model and ODE-based model (figure 2.2). We also artificially set the dose to zero for both models and plotted the results. Only the ODE-based model correctly recognised that no concentration response should be expected in this case (figure 2.2B).

2.B.3 SHAP Analysis

We performed a SHAP analysis (ShapML Julia package, v0.3.2, Nick Redell (Chicago, IL, USA)) on the ODE-based model to visualise covariate relationships with absorption rate predictions. Since our neural network has only a single continuous covariate we could directly visualise the predicted absorption rate for a range of age values for male and female patients. These are visualised alongside the SHAP values (figure 2.4). We can see that the SHAP values match the prediction by the neural network. However, visualising the complete functional relationship of patient age allows us to infer the prediction for out-of-distribution data. In this case, the effect of age on the absorption rate can be quite different from the observed data. This is especially true for female patients, of whom the number of samples is low. Here we see a sharp decrease in absorption rate for female patients aged below 25 years, which might lead to poor generalisation. However, as we do not have any samples in this age range, we are unable to reliably represent this effect using SHAP values.

3

APPLICATION OF SHAP VALUES FOR INFERRING THE OPTIMAL FUNCTIONAL FORM OF COVARIATES IN PHARMACOKINETIC MODELLING

Alexander Janssen, Mark Hoogendoorn, Marjon H. Cnossen, and Ron A.A. Mathôt

CPT: *Pharmacometrics and Systems Pharmacology* 11(8) (2022): 1100-1110.

ABSTRACT

Pharmacometrics is a multidisciplinary field utilising mathematical models of physiology, pharmacology, and disease to describe and quantify the interactions between medication and patient. As these models become more and more advanced, the need for advanced data analysis tools grows. Recently, there has been much interest in the adoption of machine learning (ML) algorithms. These algorithms offer strong function approximation capabilities and might reduce the time spent on model development. However, ML tools are not yet an integral part of the pharmacometrics workflow. The goal of this work is to discuss how ML algorithms have been applied in four stages of the pharmacometrics pipeline: data preparation, hypothesis generation, predictive modelling, and model validation. We will also discuss considerations before the use of ML algorithms with respect to each topic. We conclude by summarising applications that hold potential for adoption by pharmacometrists.

3.1 INTRODUCTION

In population pharmacokinetic (PK) modelling, identification of the relationship between PK parameters and covariates is important for the explanation of inter-individual variation (IIV). The classic step-wise method is among the most popular methods but is not without flaws. In step-wise methods, covariate selection is determined by a significant change in the objective function value following inclusion or exclusion of each covariate. Due to the ordered nature of this process, the method may suffer from bias and multiplicity issues [1-3].

The full fixed effects model (FFEM), which is based on a full model fit, was introduced to reduce selection bias [4]. In this method, all covariates of interest are tested simultaneously and included if they result in a clinically relevant change of the typical PK parameters. Although an improvement over the step-wise method, the FFEM is not able to solve all prior issues. In both methods, an assumption must be made about the functional form describing the relationship between the covariate and the PK parameters. This encourages data dredging because various functional forms can be tested until one satisfies the criteria for inclusion. Furthermore, true covariates may be excluded when sub-optimal functional forms are used. In summary, we identify a need for a covariate selection method which performs both a full model fit, while simultaneously estimating the optimal functional form of each covariate.

A recent study describes the use of machine learning (ML) for performing covariate selection for PK models [5]. Here, the authors discuss how combining ML algorithms with covariate importance scores can be used to obtain a similar or better selection of covariates compared to step-wise methods. Other studies further discuss using such an approach on real-life data to obtain a set of predictive covariates [6, 7]. ML algorithms might be suitable for this task as they can learn covariate relationships directly from data. These methods might thus reduce the issue of selecting sub-optimal functional forms when testing covariates for inclusion. Many ML software packages provide measures of covariate importance. For tree-based methods (e.g., random forests[8] or gradient boosting trees [9]), examples include counting the number of uses of each covariate, or more sophisticated measures, such as Gini or permutation importance. Although often found to be relatively accurate, there are situations where these measures may be biased [10]. In addition, they only provide a single score

of importance without information about the relationship between each covariate and model output. After obtaining a set of important covariates, how do we now select the functional form to implement these covariates without again resorting to step-wise methods?

SHapley Additive exPlanations (SHAP) is a promising model explanation technique due to its strong theoretical base [11]. In addition to a more robust benchmark performance compared to other approaches[12], SHAP allows for identification of the influence of specific covariates and their effect on each individual prediction. The use of SHAP might improve upon importance scores by also allowing for the analysis of the relationship between covariates and model output. Its use for covariate selection has, however, not yet been explored.

In this study, we will focus on tree-based ML algorithms, as there exists an exact method for the computation of SHAP values for these types of models [12]. Specifically, we will use the random forest and XGBoost algorithms [13]. Both methods create an ensemble model of decision trees. A decision tree is an algorithm that groups observations into bins (appropriately called leaves), which share a similar value for the response variable. Each tree is composed of multiple layers, where the observation is split into two leaves based on the value of one of the covariates. In a random forest, the model prediction is averaged over multiple independently fit trees. Each tree is fit using a subset of the data adding stochasticity to the learning process aiming to reduce overfitting. In gradient boosting trees (e.g., XGBoost), the trees are built sequentially, so that additional decision trees are added if they improve the prediction of the previous model ensemble. Each tree is thus fit to improve the mistakes of the previous tree. The objective function also contains a regularisation term which penalises the addition of complex models. In contrast to the classic random forest implementation, XGBoost supports missing values [13].

Our goal is to evaluate the value of combining ML and SHAP for enriching ML-based covariate analysis in the context of PK models. To this end, we will fit a random forest and XGBoost model to predict empirical Bayes estimates of PK parameters and perform a SHAP analysis on the most accurate model. As a case study, we use a retrospective data set of patients with haemophilia A receiving clotting factor VIII (FVIII) while undergoing surgery [14]. We explore the output of the SHAP analysis and present how it can be used for understanding the relationship between covariates and PK parameters.

3.2 METHODS

3.2.1 Data set

We used retrospective data of 119 individuals with haemophilia A undergoing surgery in five different haemophilia treatment centres in the Netherlands [14]. Patients received clotting factor FVIII concentrate (via bolus or continuous doses) to reach target FVIII levels as set by the Dutch National Haemophilia Consensus. This guideline recommends the following FVIII peak levels during the perioperative window: $0.80\text{-}1.00 \text{ IUml}^{-1}$ at $0\text{-}24\text{h}$, $0.50\text{-}0.80 \text{ IUml}^{-1}$ at $24\text{-}120\text{h}$, and $0.30\text{-}0.50 \text{ IUml}^{-1}$ beyond 120h post-surgery. A total of 3350 FVIII levels were measured during 197 surgical procedures. All FVIII levels were measured using the one-stage clotting assay. Timing and dosage of measurements was determined at the discretion of the treating physician. For most patients, this resulted in more frequent measurements early in the perioperative window, and occasional measurements post-surgery to validate if the patient still met target levels. The following 13 covariates were chosen for analysis: treatment centre (1-5), pre-assessed surgical risk (low vs. high [15]), use of β -domain deleted recombinant FVIII (BDD-FVIII, moroctocog alfa/Refacto AF), haemophilia severity (moderate vs. severe), FVIII baseline levels, blood group (O vs. non-O), blood loss during surgery, occurrence of a bleeding complication, body weight, body mass index (BMI), age in years, and von Willebrand factor antigen (VWF:Ag) and activity (VWF:act) levels. Five covariates contained missing values. Missing values were either imputed by mean (for continuous variables) or addition of a separate category (for categorical variables).

3.2.2 Prediction of PK parameters using machine learning

Empirical Bayes estimates of the PK parameters were obtained by fitting a base two-compartment model to the data using NONMEM (ICON Development Solutions, Ellicott City, MD [16]). Random effects were only estimated for the clearance and central volume parameters in order to improve model stability. A combined additive and proportional error model was used. We fixed the residual error estimates to $\sigma_1 = 0.08$ (additive error) and $\sigma_2 = 0.17$ (proportional error) to improve model stability and shrinkage while matching earlier findings [17, 18]. Random forest (Python scikit-learn package, version 0.23.2)

and XGBoost (Python `xgboost` package, version 1.4.2) models were fit to predict the empirical Bayes estimated clearance and central volume distribution parameters independently. We fit the XGBoost models to both the original (containing missing values) and imputed data set. We performed a 10-fold cross-validation for the estimation of test set error and for SHAP value calculation. Default model hyperparameters were used (see Material S1 for details). Model accuracy was represented as the average mean absolute error (MAE) \pm one SD of PK parameter predictions on the 10 test sets. We also calculated the root mean squared error (RMSE) of predicted FVIII levels by solving a two-compartment model using the test set predicted PK parameters. The empirical Bayes estimated inter-compartmental clearance and peripheral volume parameters were directly used for all patients. FVIII level predictions were performed in the Julia programming language (Julia Computing, Inc., version 1.6.0) using the `DifferentialEquations` package (version 6.17.1) [19]. The RMSE was presented as the mean and SD of the RMSE calculated for each individual patient.

3.2.3 SHAP analysis

A SHAP analysis (Python `shap` package, version 0.36.0) was performed to explain model output. This method decomposes a model $f(x)$ into a simpler additive model:

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_{x_i} \quad (3.1)$$

Here, the SHAP value ϕ_{x_i} of covariate $i \in M$ represents its direct effect on the model prediction, whereas ϕ_0 represents the typical prediction. By accumulating the SHAP values for each individual, we can visualise their relationships with each of the covariates. For each of the 10 cross-validations, we calculated SHAP values on the corresponding test set. The SHAP values were pooled and a smoothed representation of the effect was obtained by means of locally estimated scatterplot smoothing (LOESS; Python `statsmodels` package, version 0.12.2). SHAP values for missing continuous covariates were removed from visualisations.

3.2.4 Model code

All model code, including implementation instructions, will be made available at <https://github.com/Janssena/pkSHAP> at the time of publication.

3.3 RESULTS

3.3.1 Patient characteristics and model accuracy

An overview of the patient characteristics, missing data, and the base model parameter estimates is shown in table 3.3.1. RMSE of FVIII level predictions by the base nonlinear mixed effects (NLME) model was $0.23 \text{ IUml}^{-1} \pm 0.27$ (SD). Accuracy of the ML models is depicted in table 3.3.2. The MAE of PK parameter predictions by both ML algorithms fit to the imputed data set was similar. The XGBoost model fit to the original data set resulted in higher MAE of both clearance (43.8 vs. 40.4 ml/h), as well as central volume predictions (893 vs. 807 ml) compared to the random forest model. In addition, the RMSE of the resulting FVIII level predictions was higher for the XGBoost model (0.36 vs. 0.32 IUml^{-1}). The MAE of PK parameter predictions was indicative of the presence of residual IIV unexplained by the current set of covariates.

3.3.2 SHAP analysis

We present an overview of the SHAP values for the random forest models in figure 3.3.1. This visualisation can, for example, be used for the identification of influential covariates, as indicated by the horizontal span of SHAP values. Alternatively, we can use feature importance scores or the mean absolute SHAP value to rank the covariates based on influence. We have provided a comparison of these two scores in figure 3.A.1. Both scores seem to lead to relatively similar results.

	NO. OF PROCEDURES (%) OR MEDIAN [MIN / MAX]	NO. OF MISSING DATA (%)
Weight, kg	75.0 [5-111]	0 (0)
Age, years	39.8 [0.24-77.7]	0 (0)
BMI	24.1 [13.6-32.8]	21 (10.7)
Treatment centre		0 (0)
- One	40 (20.3)	
- Two	45 (22.8)	
- Three	76 (38.6)	
- Four	16 (8.1)	
- Five	20 (10.2)	
Blood group		26 (13.2)
- Non-O	82 (41.6)	
- O	80 (40.6)	
FVIII concentrate		3 (1.5)
- BDD-rFVIII	28 (14.2)	
- Non BDD-rFVIII	166 (84.3)	
High pre-assessed surgical risk	97 (49.2)	0 (0)
Has severe haemophilia	147 (74.6)	0 (0)
Blood loss, ml	0 [0-6700]	0 (0)
Had bleeding complication	30 (15.2)	0 (0)
FVIII baseline level, IUml ⁻¹	0.0 [0.0-0.05]	0 (0)
VWF:Ag, %	120 [25-250]	79 (40.1)
VWF:Act, %	130 [24-270]	99 (50.3)
<hr/>		
NLME model parameters		
CL, ml/h	163 [29.5-387]	
V ₁ , ml	3030 [260-9710]	
Q, ml/h	56.9	
V ₂ , ml	1270	
CL (%CV)	65.2	
V ₁ (%CV)	83.5	

Abbreviations: %CV = percent coefficient of variation, BDD-rFVIII = β -domain deleted recombinant clotting factor FVIII, BMI = body mass index, CL = clearance, NLME = nonlinear mixed effects, Q = inter-compartmental clearance, V₁ = central volume, V₂ = peripheral volume, VWF:Act = von Willebrand factor activity, VWF:Ag = von Willebrand factor antigen.

Table 3.3.1: Patient characteristics.

	RANDOM FOREST	XGBOOST	XGBOOST IMPUTE
MAE of CL predictions, ml/h	40.4 ± 10.5 SD ($R^2 = 0.56$)	43.8 ± 10.8 SD ($R^2 = 0.48$)	42.4 ± 11.0 SD ($R^2 = 0.50$)
MAE of V_1 predictions, ml	807 ± 320 SD ($R^2 = 0.49$)	893 ± 356 SD ($R^2 = 0.37$)	817 ± 308 SD ($R^2 = 0.47$)
RMSE of concentration predictions, IUml ⁻¹	0.32 ± 0.20 SD	0.36 ± 0.26 SD	0.33 ± 0.22 SD

Abbreviations: CL = clearance, MAE = mean absolute error, PK = pharmacokinetic, RMSE = root mean squared error, SD = standard deviation, V_1 = central volume.

Table 3.3.2: Accuracy of PK parameter and concentration predictions.

For both PK parameters, patient weight was the most influential covariate. For clearance (figure 3.3.1a), treatment centre, blood group, age, and VWF:Ag appeared to be relatively influential. For central volume (figure 3.3.1b), BMI and use of BDD-rFVIII concentrate seem to be the most important covariates aside from patient weight. The remaining covariates seem to be less influential for explaining the prediction. We can also take a look at the SHAP values for a single individual (figure 3.3.2). Here, we can see the exact change in clearance and central volume resulting from the inclusion of each covariate.

Our main motivation for performing the SHAP analysis was the ability to visualise the relationship between the calculated SHAP values and each covariate of interest. In figure 3.3.3, we present the resulting relationships for six covariates from the clearance model and three covariates from the central volume model. We observed a positive relationship between body weight and clearance, which flattened for weights above 65 kg (figure 3.3.3a). For age, we saw a negative relationship with clearance, similar to earlier findings [17]. We noticed that individuals with VWF:Ag levels below 100% had higher clearance than those with higher levels (figure 3.3.3c). In addition, we observed that patients with blood group O displayed an increased clearance compared to non-O individuals (figure 3.3.3d). Both these findings were in line with physiological concepts of haemostasis. Next, we saw that the model predicts a decrease in clearance for individuals in centre one, possibly as result of a confounder (figure 3.3.3e). Finally, individuals who received a BDD-rFVIII concentrate displayed slightly increased clearance compared to those who did not (figure 3.3.3f).

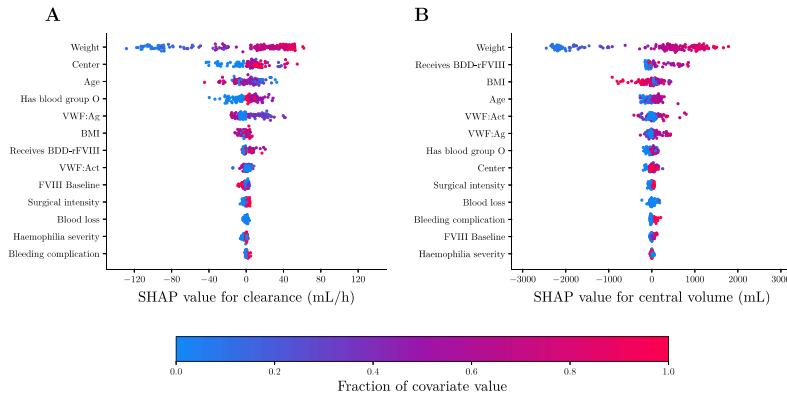


Figure 3.3.1: Overview of SHAP values for random forest model. SHAP values of the clearance (a) and central volume (b) are shown as calculated for the random forest model. The covariate value is indicated by colour. The horizontal span of the SHAP values indicate the change in the parameters value. The larger the span, the larger the changes in PK parameter and thus the more important the covariate. Covariates are ranked from most (top) to least (bottom) influential by means of their mean absolute SHAP value. Abbreviations: BDD-rFVIII: β -domain deleted recombinant clotting factor FVIII; BMI: body mass index; PK: pharmacokinetic; SHAP: SHapley Additive exPlanations; VWF: von Willebrand factor.

For central volume, we also noted a positive relationship with body weight, which flattened slightly with increasing body weight (figure 3.3.3g). We saw a sharp decrease in the SHAP values for central volume for individuals with a BMI 25 (i.e., those classified as overweight; figure 3.3.3h). Finally, we saw an increase in the SHAP values for individuals who received BDD-rFVIII concentrate (figure 3.3.3i).

We could further push the analysis by examining the combined effects of multiple covariates (figure 3.3.4). Because body weight, BMI, and age were correlated, the true effect of either covariate might have been obscured by the others. We combined their respective SHAP values to determine if there was a unique effect of including the separate covariates. After this intervention, there were only small differences between the SHAP values of weight alone versus those of weight and BMI combined for clearance. The same was true for the combined SHAP values of weight and age for central volume. However, combining the SHAP values of weight and age for clearance showed that part of its variance could be well explained by age for

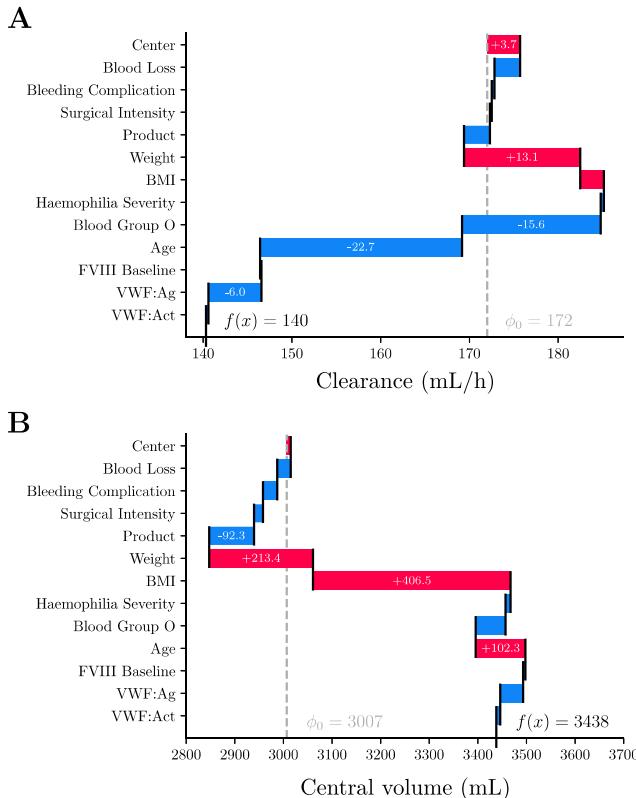


Figure 3.3.2: SHAP values for a typical patient. SHAP values are shown for the clearance (a) and central volume (b) predictions by the random forest. Data is shown for a 70 kg, 63 year-old individual with blood group non-O. SHAP value for each covariate is shown in the corresponding bar. Vertical dashed line indicates expected SHAP value. The SHAP values sum up to the final model prediction. Abbreviations: BDD-FVIII: β -domain deleted clotting factor FVIII; BMI: body mass index; SHAP: SHapley Additive exPlanations; VWF: von Willebrand factor.

individuals with a body weight above 65 kg (figure 3.3.4a). Combining the SHAP values of weight and BMI for central volume resulted in a more pronounced flattening of SHAP values for individuals with a body weight above 65 kg, although considerable variance remained (comparing figures 3.3.3g and 3.3.4b). Earlier, we identified a difference in the SHAP values of clearance for patients receiving treatment in centre one. The SHAP analysis suggests that individuals without blood group O had SHAP values closer to zero compared to individuals

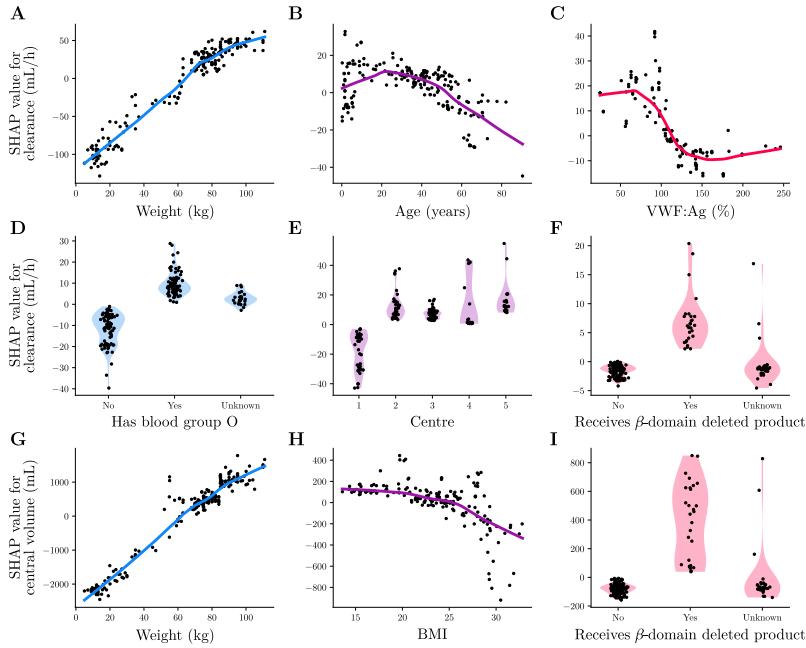


Figure 3.3.3: Relationship between covariates and PK parameters based on SHAP values. Here we visualise the relationship between PK parameter and covariate by plotting SHAP value against covariate value. Points represent the SHAP values, while lines indicate the LOESS fitted smooth representation of the relationship. For the categorical covariates the SHAP value density is also shown by means of a violin plot. We have shown the results for the most important covariates for clearance (a-f) and central volume (g-i). Abbreviations: BMI: body mass index; LOESS: locally estimated scatterplot smoothing; PK: pharmacokinetic; SHAP: SHapley Additive exPlanations; VWF: von Willebrand factor.

with blood group O (figure 3.3.4c). No such effect is seen for the other centres. For the SHAP values of blood group for clearance, we see a similar result. Here, individuals with lower body weight (65 kg) seem to have SHAP values closer to zero than those with higher body weight (figure 3.3.4d).

A classical approach to obtain intuition on what functional forms to use would be to plot the empirical Bayes estimates of the PK parameters against each of the covariates. This visualisation is shown in figure 3.A.2. Here, we see that for highly correlated covariates (i.e., weight), it is possible to derive some intuition on the functional form

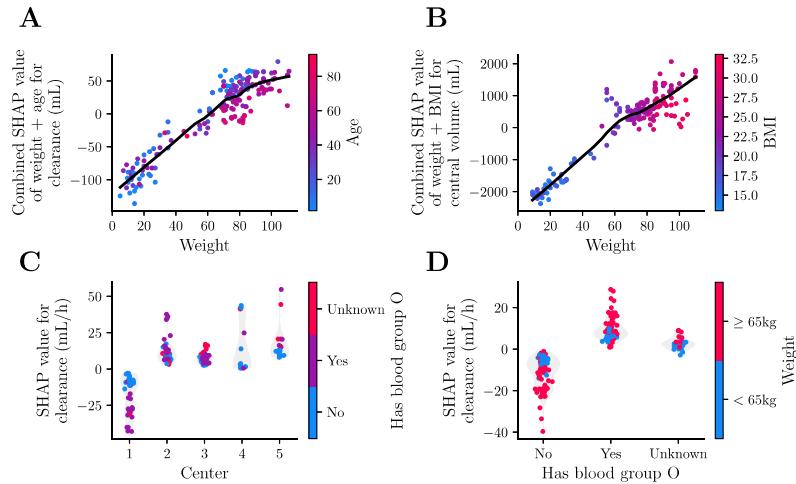


Figure 3.3.4: Interaction between SHAP values of the covariates. SHAP values of interactions between covariates are shown for the clearance (a, c, and d) and central volume (b) models. Points represent the SHAP values, while lines indicate the LOESS fitted smooth representation of the relationship. The value of the interacting covariate is indicated by colour. For the categorical covariates the SHAP value density is also shown by means of a violin plot. Abbreviations: BMI: body mass index; LOESS: locally estimated scatterplot smoothing; SHAP: SHapley Additive exPlanations.

to use, but for most covariates their effect is difficult to discern. This is because we are unable to visualise the contribution of each covariate in isolation. Because unexplained residual variance is also present in the PK parameters, choosing a function to use can be more difficult due to large variation. This can mean that we have to iteratively select functions to implement covariates, reproduce the visualisations, and re-evaluate, thus again resorting to a step-wise approach. With SHAP, we can decide on appropriate functions based on a single full model fit.

Although not shown, the functional forms of the covariates as described by the SHAP values of the two XGBoost models were very similar to those from the random forest. This suggested that the choice between a random forest and XGBoost had only minor effects on the subsequent SHAP analysis.

3.4 DISCUSSION

In this study, we aimed to enrich ML-based covariate selection methods using SHAP in order to infer the optimal function form to use when including covariates in PK models. We fit both a random forest and XGBoost model to predict empirical Bayes estimated PK parameters originating from a base NLME model. The random forest resulted in slightly more accurate PK parameter predictions compared to the XGBoost models. Next, influential covariates can, for example, be selected using importance scores [5]. Finally, after performing a SHAP analysis, we are able to examine the relationship between each covariate and the PK parameters in greater detail. The SHAP analysis also allowed us to explore more complex interaction effects of covariates resulting from the sequential binning in tree-based methods. Because SHAP values depict the absolute change in output value, the user can intuitively determine clinical relevance. These features display the benefit of SHAP values compared to using importance measures in isolation, where often only a single score of importance is obtained.

The SHAP analysis identified covariates that have previously been associated with the PK of FVIII concentrates. In addition, the suggested relationships of the covariates are similar to their implementation in previous PK models [17, 20, 21]. First, we found that patient weight was the most important covariate to explain IIV for both clearance and central volume. The concept of allometric scaling is often applied to the relationship between weight and FVIII clearance. This is mirrored in the flattening of the SHAP values as weight increases (figure 3.3.3a,g). As the central volume compartment represents the blood plasma, a relationship resembling a linear interaction with weight might be expected. An obvious exception exists for overweight individuals, which is represented by the SHAP values in the sharp decline in SHAP values seen for individuals with a BMI greater than 25 (figure 3.3.3h). Measures of fat-free mass have been suggested to better predict central volume, which could remove the need to model the effect of BMI [22].

Next, we saw a negative interaction between age and clearance. This effect has been demonstrated before, 16 and there might be multiple possible explanations for this effect. One such explanation is the finding that several blood coagulation factors, including VWF, increase with age [23, 24]. It is well known that VWF binds to FVIII to protect it from degradation in the blood circulation. Similar to this effect, SHAP values for patients with blood group O depicted increased

FVIII clearance, an effect likely linked to lower VWF:Ag levels seen in patients with blood group O [25]. Looking at the interaction between blood group and weight (figure 3.3.4d), we see that individuals below 65 kg (i.e., usually younger individuals) with blood group non-O have relatively higher clearance than heavier individuals. This might also be linked to the previously observed increase in VWF:Ag levels with age [23, 24]. It is possible that weight was used by the random forest as a proxy for age. Higher VWF:Ag levels were also directly associated with a decrease in FVIII clearance by the model (figure 3.3.3c). However, considering the large fraction of missing data (40.1%), a low number of patients at the extremes of VWF levels, and the fact that the measurements were outdated (i.e., not measured during the surgical procedure) there remains uncertainty about the observed relationship between VWF:Ag and clearance. Interpreting the effects of covariates with large fractions of missing data should be handled with care.

The SHAP values indicate that individuals from centre one had lower clearance compared to other centres. One possible explanation is the use of different assay reagents in this centre. The results, however, also indicate that this effect is correlated with the patient blood group (figure 3.3.4c). There could thus be some other factor influencing this effect. Because we worked with retrospective data, it is difficult to underpin the origin of this effect.

Finally, we notice an increase in clearance and central volume associated with patients who received BDD-rFVIII concentrate. It is well known that use of BDD-rFVIII leads to an underestimation of FVIII activity levels when using the one-stage assay versus the chromogenic assay [26, 27]. By changing the phospholipid source in the one-stage assay, similar FVIII activity levels compared to the chromogenic assay are measured. This suggests that this effect is not due to increased clearance or distribution volume of BDD-rFVIII [27]. It is possible that this effect leaked into the PK parameter estimates (instead of being part of the estimated error) by the base NLME model. Most of its effect was on increasing the central volume estimate. This can be expected as it would lead to a decrease in predicted FVIII levels.

From the previous discussions, we see the possibility of identifying many subtle effects captured by the random forest model using SHAP. However, the method also has limitations. First, the quality of the empirical Bayes estimated PK parameters is an important factor affecting the accuracy of the ML model and quality of the SHAP analysis. In our case, this required fixing the residual error parameters and only

including random effects on clearance and central volume. It might not be clear in advance what measures need to be taken to obtain reliable results. Inspecting the distribution of the resulting PK parameters and comparing these to prior results can be a way to decide on an effective strategy in obtain good quality PK parameter estimates.

Next, we used LOESS to obtain an average representation of the relationship between the covariates and PK parameters. Although this may be helpful for the identification of effects, it might also bias the user to find relationships that do not exist. The method might falsely represent the true effect when SHAP values have high variance or when data are sparse.

Another possible issue lies in the inclusion of covariates that displayed substantial fractions of missing values. For example, roughly 40% of VWF:Ag levels were missing. Although its relationship with clearance suggested by the SHAP values matches previous biological understanding, we might not want to include the covariate based on the current analysis alone. Previous studies have, however, included this covariate using a function matching the SHAP values [20, 21].

A more general issue with the application of SHAP values in the context of PK models is that it results in an additive breakdown of the model. Often, covariate effects in PK models are instead implemented as a product of functions. This makes it difficult to compare the outcomes of SHAP analyses with classic methods of covariate analysis, such as forest plots obtained from an FFEM. In addition, by using products, we can prevent the PK parameters from becoming negative. However, because the relationships of the covariates suggested by the SHAP values match those used in previous PK studies, we assume that the functional forms might hold (up to a difference in parameters) [14, 17, 18]. Such an assumption will have to be validated.

Finally, although SHAP might be able to explain the covariate relationships in the ML model, this does not mean that the results are biologically interpretable. ML algorithms remain black box models, simply deconstructing the model in components does not guarantee that the results are humanly interpretable. For example, we found an effect of centre one on clearance, which was correlated with patient blood group. With the current data we are unable to provide an explanation of this effect. Consequently, not every effect found by the SHAP analysis should necessarily be included in PK models.

In summary, we show that combining ML and SHAP allows for an in-depth review of the relationships between covariates and PK parameters. We have mainly focused on using SHAP values for visualising

covariate relationships in ML models. SHAP values can also be used to perform covariate selection. Its benefit over importance scores will have to be evaluated. Covariate selection is a difficult issue, and our method is one of the first to allow one to infer the optimal function form to include covariates based on ML algorithms. The method can prove useful for covariate analysis and hypothesis generation.

REFERENCES

- [1] Shelley Derksen and Harvey J Keselman. "Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables". In: *British Journal of Mathematical and Statistical Psychology* 45.2 (1992), pp. 265–282.
- [2] Jakob Ribbing and E Niclas Jonsson. "Power, selection bias and predictive performance of the population pharmacokinetic covariate model". In: *Journal of pharmacokinetics and pharmacodynamics* 31 (2004), pp. 109–134.
- [3] Virginia F Flack and Potter C Chang. "Frequency of selecting noise variables in subset regression analysis: a simulation study". In: *The American Statistician* 41.1 (1987), pp. 84–86.
- [4] M. Gastonguay. "Full covariate models as an alternative to methods relying on statistical significance for inferences about covariate effects: A review of methodology and 42 case studies". In: *Annual meeting of the population approach group in Europe*. Athens, Greece, 2011.
- [5] Emeric Sibieude, Akash Khandelwal, Jan S Hesthaven, Pascal Girard, and Nadia Terranova. "Fast screening of covariates in population models empowered by machine learning". In: *Journal of pharmacokinetics and pharmacodynamics* 48.4 (2021), pp. 597–609.
- [6] Rui Wang, Xiao Shao, Junying Zheng, Abdel Saci, Xiaozhong Qian, Irene Pak, Amit Roy, Akintunde Bello, Jasmine I Rizzo, Fareeda Hosein, et al. "A machine-learning approach to identify a prognostic cytokine signature that is associated with nivolumab clearance in patients with advanced melanoma". In: *Clinical Pharmacology & Therapeutics* 107.4 (2020), pp. 978–987.
- [7] Diana-Maria Ciuculete, Marcus Bandstein, Christian Benedict, Gérard Waeber, Peter Vollenweider, Lars Lind, Helgi B Schiöth, and Jessica Mwinyi. "A genetic risk score is significantly associated with statin therapy response in the elderly population". In: *Clinical genetics* 91.3 (2017), pp. 379–385.
- [8] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.
- [9] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.
- [10] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8 (2007), pp. 1–21.
- [11] Scott M. Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: (2017), 4768–4777.

- [12] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.
- [13] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 785–794.
- [14] Hendrika Hazendonk, Karin Fijnvandraat, Janske Lock, Mariëtte Driessens, Felix Van Der Meer, Karina Meijer, Marieke Kruip, Britta Laros-van Gorkom, Marjolein Peters, Saskia de Wildt, et al. "A population pharmacokinetic model for perioperative dosing of factor VIII in hemophilia A patients". In: *haematologica* 101.10 (2016), p. 1159.
- [15] M Koshy, SJ Weiner, ST Miller, LA Sleeper, E Vichinsky, AK Brown, Y Khakoo, and TR Kinney. "Surgery and anesthesia in sickle cell disease. Cooperative Study of Sickle Cell Diseases". In: *Blood* 86.10 (Nov. 1995), pp. 3676–3684. ISSN: 0006-4971. DOI: 10.1182/blood.V86.10.3676.bloodjournal86103676.
- [16] S.L. Beal, L.B. Sheiner, and A.J. Boeckmann. "NONMEM 7.4 users guides". In: 1989. URL: <https://nonmem.iconplc.com/nonmem743/guides>.
- [17] Sven Björkman, MyungShin Oh, Gerald Spotts, Phillip Schroth, Sandor Fritsch, Bruce M Ewenstein, Kathleen Casey, Kathelijn Fischer, Victor S Blanchette, and Peter W Collins. "Population pharmacokinetics of recombinant factor VIII: the relationships of pharmacokinetics to age and body weight". In: *Blood, The Journal of the American Society of Hematology* 119.2 (2012), pp. 612–618.
- [18] Alanna McEneny-King, Pierre Chelle, Gary Foster, Arun Keepanasseril, Alfonso Iorio, and Andrea N Edginton. "Development and evaluation of a generic population pharmacokinetic model for standard half-life factor VIII for use in dose individualization". In: *Journal of Pharmacokinetics and Pharmacodynamics* 46 (2019), pp. 411–426.
- [19] Christopher Rackauckas and Qing Nie. "Differentialequations. jl-a performant and feature-rich ecosystem for solving differential equations in julia". In: *Journal of open research software* 5.1 (2017), pp. 15–15.
- [20] Ivan Nestorov, Srividya Neelakantan, Thomas M Ludden, Shuanglian Li, Haiyan Jiang, and Mark Rogge. "Population pharmacokinetics of recombinant factor VIII Fc fusion protein". In: *Clinical pharmacology in drug development* 4.3 (2015), pp. 163–174.
- [21] Y Zhang, J Roberts, M Tortorici, A Veldman, K St Ledger, A Feussner, and J Sidhu. "Population pharmacokinetics of recombinant coagulation factor VIII-SingleChain in patients with severe hemophilia A". In: *Journal of Thrombosis and Haemostasis* 15.6 (2017), pp. 1106–1114.
- [22] Iris van Moort, Tim Preijers, Hendrika CAM Hazendonk, Roger EG Schutgens, Britta AP Laros-van Gorkom, Laurens Nieuwenhuizen, Felix JM van der Meer, Karin Fijnvandraat, Frank WG Leebeek, Karina Meijer, et al. "Dosing of factor VIII concentrate by ideal body weight is more accurate in overweight and obese haemophilia A patients". In: *British journal of clinical pharmacology* 87.6 (2021), pp. 2602–2613.

- [23] JC Gill, J Endres-Brooks, PJ Bauer, WJ Jr Marks, and RR Montgomery. "The effect of ABO blood group on the diagnosis of von Willebrand disease". In: *Blood* 69.6 (June 1987), pp. 1691–1695. ISSN: 0006-4971. doi: 10.1182/blood.V69.6.1691.1691.
- [24] Massimo Franchini. "Hemostasis and aging". In: *Critical reviews in oncology/hematology* 60.2 (2006), pp. 144–151.
- [25] Dieter Klarmann, Christine Eggert, Christof Geisen, Sabine Becker, Erhard Seifried, Thomas Klingebiel, and Wolfhart Kreuz. "Association of ABO (H) and I blood group system development with von Willebrand factor and Factor VIII plasma levels in children and adolescents". In: *Transfusion* 50.7 (2010), pp. 1571–1580.
- [26] M Mikaelsson, U Oswaldsson, and MA Jankowski. "Measurement of factor VIII activity of B-domain deleted recombinant factor VIII". In: 38 (2001), pp. 13–23.
- [27] CA Lee, CM Kessler, D Varon, U Martinowitz, M Heim, M MIKAELOSSON, U OSWALDSSON, and H SANDBERG. "Influence of phospholipids on the assessment of factor VIII activity". In: *Haemophilia: State of the Art* 4.4 (1998), pp. 646–650.

APPENDIX

3.A SUPPLEMENTARY FIGURES

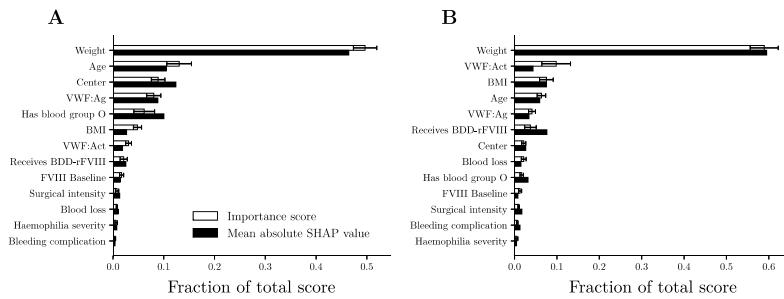


Figure 3.A.1: Covariate importance scores for the random forest model.

Permutation importance (white) and mean absolute SHAP value (black) scores for the clearance (A) and central volume (B) random forest models. The fraction of the total score is shown. Error bars for the permutation importance scores indicate the standard deviation of the importance scores from the ten models fit during cross validation.

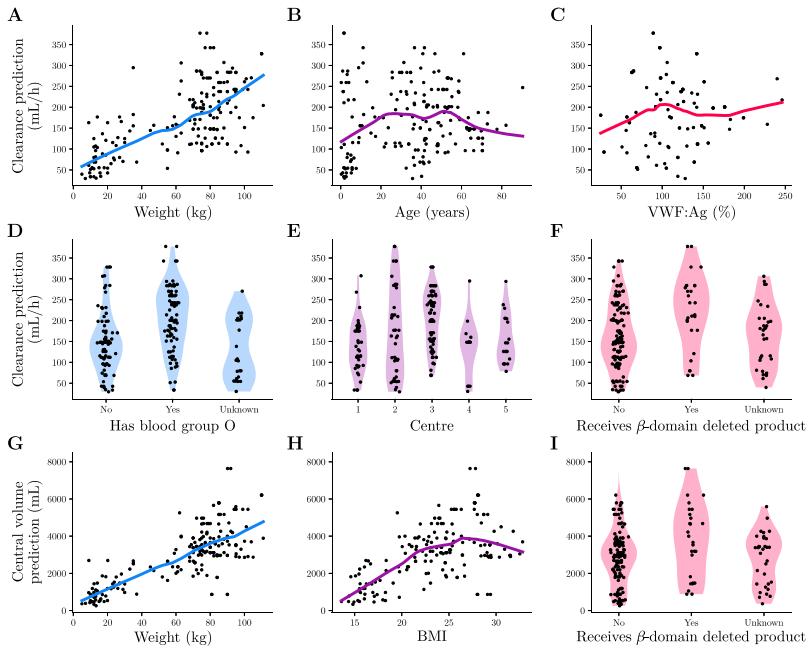


Figure 3.A.2: Correlation between covariates and PK parameter estimates. Here the correlation between empirical Bayes estimates of the PK parameter and covariate values are shown. Points represent the PK parameter predictions, while lines indicate the LOESS fitted smooth representation of the relationship. For the categorical covariates the density is also shown by means of a violin plot. We again show the results for the most important covariates for clearance (A-F) and central volume (G-I).

3.B DEFAULT HYPER-PARAMETERS FOR RANDOM FOREST AND XGBOOST MODELS

Below we list the most important hyper-parameters (e.g. excluding those not influencing model performance such as logging etc.) of the random forest and XGBoost model. Hyper-parameter descriptions were directly obtained from the sci-kit learn documentation (<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>) and the XGBoost documentation (<https://xgboost.readthedocs.io/en/stable/parameter.html>).

3.B.1 Random forest model

`n_estimators = 100`: The number of trees in the forest.

`max_depth = None`: The maximum depth of the tree. If `None`, then nodes are expanded until all leaves are pure or until all leaves contain less than `min_samples_split` samples.

`min_samples_split = 2`: The minimum number of samples required to split an internal node.

`min_samples_leaf = 1`: The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least `min_samples_leaf` training samples in each of the left and right branches.

`min_weight_fraction_leaf = 0.0`: The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Samples have equal weight when `sample_weight` is not provided.

`max_features = "auto"`: The number of features to consider when looking for the best split. If `"auto"`, then `max_features = n_features`.

`max_leaf_nodes = None`: Grow trees with `max_leaf_nodes` in best-first fashion. Best nodes are defined as relative reduction in impurity. If `None` then unlimited number of leaf nodes.

`min_impurity_decrease = 0.0`: A node will be split if this split induces a decrease of the impurity greater than or equal to this value.

`bootstrap = true`: Whether bootstrap samples are used when building trees. If `False`, the whole data set is used to build each tree.

`max_samples = None`: If `bootstrap` is `True`, the number of samples to draw from `X` to train each base estimator. If `None` (default), then draw `X.shape[0]` samples.

`ccp_alpha = 0.0`: Complexity parameter used for Minimal Cost-Complexity Pruning. The sub-tree with the largest cost complexity that is smaller than `ccp_alpha` will be chosen. By default, no pruning is performed.

3.B.2 XGBoost model

`booster = "gbtree"`: Which booster to use. Can be `gbtree`, `gblinear` or `dart`; `gbtree` and `dart` use tree based models while `gblinear` uses linear functions.

`eta = 0.3`: Step size shrinkage used in update to prevents overfitting. After each boosting step, we can directly get the weights of new features, and `eta` shrinks the feature weights to make the boosting process more conservative.

`gamma = 0`: Minimum loss reduction required to make a further partition on a leaf node of the tree. The larger `gamma` is, the more conservative the algorithm will be.

`max_depth = 6`: Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit. `0` is only accepted in `lossguide` growing policy when `tree_method` is set as `hist` or `gpu_hist` and it indicates no limit on depth. Beware that XGBoost aggressively consumes memory when training a deep tree.

`min_child_weight = 1`: Minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than `min_child_weight`, then the building process will give up further partitioning. In linear regression task, this simply corresponds to minimum number of instances needed to be in each node. The larger `min_child_weight` is, the more conservative the algorithm will be.

`max_delta_step = 0`: Maximum delta step we allow each leaf output to be. If the value is set to `0`, it means there is no constraint. If it is set to a positive value, it can help making the update step more conservative. Usually this parameter is not needed,

but it might help in logistic regression when class is extremely imbalanced. Set it to value of 1-10 might help control the update.

`subsample = 1`: Subsample ratio of the training instances. Setting it to 0.5 means that XGBoost would randomly sample half of the training data prior to growing trees. and this will prevent overfitting. Sub-sampling will occur once in every boosting iteration.

`colsample_bytree`, `colsample_bylevel`, `colsample_bynode = 1`: This is a family of parameters for sub-sampling of columns. All `colsample_by*` parameters have a range of (0, 1], the default value of 1, and specify the fraction of columns to be sub sampled.

`colsample_bytree` is the subsample ratio of columns when constructing each tree. Sub-sampling occurs once for every tree constructed.

`colsample_bylevel` is the subsample ratio of columns for each level. Sub-sampling occurs once for every new depth level reached in a tree. Columns are subsample from the set of columns chosen for the current tree.

`colsample_bynode` is the subsample ratio of columns for each node (split). Sub-sampling occurs once every time a new split is evaluated. Columns are subsample from the set of columns chosen for the current level.

`lambda = 1`: L₂ regularisation term on weights. Increasing this value will make model more conservative.

`alpha = 0`: L₁ regularisation term on weights. Increasing this value will make model more conservative.

`tree_method = "auto"`: The tree construction algorithm used in XGBoost. Choices: auto, exact, approx, hist, gpu_hist, this is a combination of commonly used updaters. For other updaters like refresh, set the parameter updater directly.

"auto": Use heuristic to choose the fastest method. For small data set, exact greedy (exact) will be used. For larger data set, approximate algorithm (approx) will be chosen. It's recommended to try hist and gpu_hist for higher performance with large data set. (gpu_hist) has support for external memory.

scale_pos_weight = 1: Control the balance of positive and negative weights, useful for unbalanced classes. A typical value to consider: sum(negative instances) / sum(positive instances).

num_parallel_tree = 1: Number of parallel trees constructed during each iteration. This option is used to support boosted random forest.

objective = "reg:squarederror": reg:squarederror: regression with squared loss.

base_score = 0.5: The initial prediction score of all instances, global bias

Part II

DEEP COMPARTMENT MODELS

DEEP COMPARTMENT MODELS: A DEEP LEARNING APPROACH FOR THE RELIABLE PREDICTION OF TIME-SERIES DATA IN PHARMACOKINETIC MODELING

Alexander Janssen, Frank W.G. Leebeek, Marjon H. Cnossen, and Ron A.A. Mathôt

CPT: *Pharmacometrics and Systems Pharmacology* 11(7) (2022): 934-945.

ABSTRACT

Nonlinear mixed effect (NLME) models are the gold standard for the analysis of patient response following drug exposure. However, these types of models are complex and time-consuming to develop. There is great interest in the adoption of machine-learning methods, but most implementations cannot be reliably extrapolated to treatment strategies outside of the training data. In order to solve this problem, we propose the deep compartment model (DCM), a combination of neural networks and ordinary differential equations. Using simulated data sets of different sizes, we show that our model remains accurate when training on small data sets. Furthermore, using a real-world data set of patients with haemophilia A receiving factor VIII concentrate while undergoing surgery, we show that our model more accurately predicts a priori drug concentrations compared to a previous NLME model. In addition, we show that our model correctly describes the changing drug concentration over time. By adopting pharmacokinetic principles, the DCM allows for simulation of different treatment strategies and enables therapeutic drug monitoring.

4.1 INTRODUCTION

There is much interest in the adoption of machine learning (ML) in the field of pharmacometrics. Implementation of covariates in population pharmacokinetic (PK) models can be very complex, and might benefit from the automatic learning capabilities of ML algorithms. Previous studies have examined the accuracy of such models for predicting drug concentrations [1–3]. Although these studies report similar or improved accuracy compared to nonlinear mixed effect (NLME) models, which are widely considered to be the gold standard in the field, none of these models allow for practical use. For example, most of the proposed ML models have only been trained to predict drug concentrations at specific timepoints. Extrapolating from these timepoints can lead to highly inaccurate results. In addition, dosing and timing information is often a direct input to the model, even though we are uncertain that they will be interpreted as such. As a result, trust in the ML algorithm is low because we do not understand the translation from covariates to drug concentrations. A simple way to overcome these issues is to constrain the solution space to satisfy knowledge about drug dynamics. This involves using an ML model to predict the latent parameters z of another function, such as the one compartment model:

$$C(t, D) = \frac{D}{V_d} \cdot \exp(-k_e t), z \in \{V_d, k_e\} \quad (4.1)$$

Here, the elimination rate constant (k_e) and the distribution volume (V_d) of the drug are estimated by an ML model, whereas dose D and time since dose t can be supplied directly to $C(t, D)$. If the drug is eliminated at a constant concentration-dependent rate, we can thus reliably extrapolate to different timepoints or doses. Unfortunately, for most drugs, this assumption does not hold, and as soon as the complexity of the compartment model or dosing schedule increases, no simple closed form solution exists.

A recent paper by *Chen et al.* reports on an automatic differentiation method for calculating the gradient of an ordinary differential equation (ODE) solution with respect to its inputs [4]. This means that methods relying on automatic differentiation for gradient calculations, such as neural networks, can be constrained based on ODEs. Because we can represent any compartment model using a system of ODEs, this opens the door for a reliable use of ML algorithms in the field of pharmacometrics. In addition, interventions (such as drug doses) can

be defined to perturb the ODE system at specific timepoints, allowing for the differentiation of the solution with respect to individual treatment schedules.

In this study, we present the deep compartment model (DCM). In a DCM, a neural network is used to predict the latent parameters of a system of ODEs representing a compartment model. This technique allows for a full model-based approach which automatically implements covariates in PK models. We will test the accuracy of this model for predicting drug concentrations using simulated data sets of different sizes. In addition, we will compare its accuracy to an NLME model on real-world data of patients with haemophilia A receiving standard half-life (SHL) factor VIII (FVIII) concentrate while undergoing surgery. Both models will be fit on a retrospective data set, and will be validated on data collected during the OPTI-CLOT randomised controlled trial [5, 6].

4.1.1 Related work

Brier et al. discussed a comparison of steady-state peak and trough gentamicin concentrations predictions made by a neural network and NLME model [1]. The neural network predicted peak gentamicin concentrations between 2.5 and 6.0 $\mu\text{g}/\text{ml}$ with lower bias compared to the NLME model. However, when extrapolating to samples which were outside of this range (and not in the training set) the NLME model was more accurate. This indicated that using ML algorithms as-is likely results in problems with respect to extrapolating to unseen data.

Lai et al. introduce an implementation of neural networks (and regression splines) in the likelihood function for a data-driven estimation of covariate effects in population PK models [7]. The neural network was used to directly learn the relationship between covariates and the PK parameters of a one-compartment model. They show how the neural network is able to accurately represent nonlinear effect of covariates. However, the approach focuses on the use of compartmental models with a closed-form solution and is difficult to extend to more complex models.

Finally, *Lu et al.* reported on the deep-learning-based approach which utilises a NeuralODE to handle time and dose irregularities [8]. A recurrent neural network encoder is used to learn the initial state for an ODE solver. The solver translates this state based on the current time interval between doses into a latent variable space z . Finally,

a decoder is used to translate samples from z to the concentration predictions. The authors show how this approach can be used to correctly extrapolate to treatment schedules not seen during training, in contrast to other ML-based methods. However, a possible issue is its inherent reliance on black box methods for estimation. It is difficult to understand what the latent variables z represent, how the NeuralODE produces them, and finally how the decoder relates them to the observations.

Results from the above papers indicate how using time and dose as direct inputs to ML models will likely lead to poor extrapolation to samples outside of the training data. This is eloquently shown by *Lu et al.*, where such models still predict drug exposure even when the given dose is set to zero [8]. In this work, neural networks are used to predict parameters for an ODE (similar to NLME models), which makes it easier to implement complex compartment models and dosing schedules. The proposed architecture is relatively simple compared to the NeuralODE [8]. The latent variables z predicted by the neural network now represent PK parameters, which are more interpretable and can be compared to previous results.

4.2 METHODS

4.2.1 Problem definition

We consider a data set of n patients with d observed covariates $\mathbf{x}_i \in \mathbb{X}^{n \times d}$, $i \in \{1 \dots n\}$; and corresponding drug concentration measurements $\mathbf{y}_i \in \mathbb{R}_+^k$ for k measurements in time window $t \in [0, T]$. The number of measurements may differ between patients. For each patient i , we can define a set of clinical interventions \mathbf{I}_i , which, for example, contains information of drug doses given at specific timepoints. In classical PK modelling, we can represent the dynamics of this drug using a system of ODEs $A(t, z, \mathbf{I})$ with p latent parameters $z \in \mathbb{R}_+^p$ (aptly named the PK parameters). We often assume that the information in \mathbf{X} is insufficient to completely describe the inter-individual variation (IIV) in the concentration measurements, so our goal is to predict the typical or population predicted concentrations $\mathbb{E}[y_i]$. The corresponding typical PK parameters ζ_i for each patient are predicted directly from the covariates using a set of functions f_θ so that:

$$\zeta_i = f_\theta(\mathbf{x}_i). \quad (4.2)$$

The algebraic form of f_θ has to be specified but its parameters θ can be estimated from data. In many cases, prior knowledge is present for choosing an appropriate compartment model, but not f_θ . As a result, implementations of f_θ can be sub-optimal, resulting in lower accuracy of $\mathbb{E}[\mathbf{y}_i]$. To combat this issue, NLME models introduce two random variables: one describing the IIV: $\eta \sim \mathcal{N}(\mathbf{0}, \Omega)$, and one describing the residual variability: $\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma)$. η is used to transform ζ to obtain a distribution of z which describes the residual IIV in the population:

$$z = \zeta \cdot \exp(\eta) \quad (4.3)$$

Here, we have depicted a commonly used transformation of ζ which results in a log normally distributed random variable z . NLME models predict a set of parameters $\Theta = \{\theta, \Omega, \Sigma\}$ and produces a *maximum a posteriori* estimate of η which maximises $p(\eta | \mathbf{y}_i, \Theta)$. A downside of this approach is the requirement of sufficient measurements in \mathbf{y}_i , especially when T is large. Because the a priori predicted $\mathbb{E}[\mathbf{y}_i]$ can be inaccurate, we often need to generate a PK profile for new patients. This can be perceived as an additional burden for the patient, especially when measurements need to be taken over the span of multiple days.

4.2.2 Deep compartment model

In order to improve the prediction of ζ we developed the DCM. Here, a neural network ϕ_w with weights w is used to predict the latent parameters of a compartment model based on \mathbf{I}_i . Because ϕ_w directly predicts ζ instead of \mathbf{y}_i , we can better interpret its output. The neural network learns to represent ζ from a latent z in a data-driven manner. When we assume that each concentration measurement y_{ij} is drawn i.i.d. from a Gaussian distribution with mean μ_{ij} and variance σ^2 so that $y_{ij} = \mu_{ij} + \epsilon_{ij}, \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$; we can find the optimal weights w^* by minimising the mean squared error (MSE) objective function:

$$w^* = \min_w \mathcal{L}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n (\mathbf{y}_i - A(t_i, \phi_w(\mathbf{x}i), \mathbf{I}_i))^2 \quad (4.4)$$

The DCM model was developed in the Julia programming language (Julia Computing, Inc., version 1.6.0). Dosing events in \mathbf{I}_i were implemented as time-based callbacks to the ODE solver. These callbacks affected the rate of drug flowing into the central compartment. Consequently, bolus doses were converted to short duration infusions with

a fixed duration of 1 minute and rate $D \cdot 60$ IU/h. Model covariates were normalised between zero and one using minimum-maximum normalisation. Two variants of the DCM were developed. The first directly outputs ζ in the final layer, using a softplus activation function to ensure $\zeta \geq 0$. The second can be passed a set of initialisation parameters ζ_0 . In the latter case, the final layer of ϕ_w has the following form:

$$l_n = \zeta_0 \odot (\pi(l_{L-1}) + \mathbf{1}) \quad (4.5)$$

Here, L denotes the number of layers l in ϕ_w , \odot indicates the Hadamard product, $\pi(\cdot)$ is the CELU activation function with $\alpha < 1$ [9], and $\mathbf{1}$ is a vector of ones of length p . In this case, the model learns the deviation from ζ_0 based on x_i . The CELU activation function acts as an implicit constraint to penalise the gradient of l_{L-1} as it reaches $1 - \alpha$, preventing ζ to be zero. The "standard" DCM can be used in cases where measurement data is rich, whereas the DCM with initialisation can help to improve parameter predictions when data are sparse.

In this paper, we have used a basic neural network encoder structure in order to reduce the number of parameters in the model. The model contained two fully connected hidden layers: the first had 64 neurons, and the second had 16 neurons. The swish activation function was used for the hidden layers [10]. The output layer contained four neurons representing the PK parameters. No optimisation of model architecture was performed. The ADAM optimiser was used for updating neural network weights with a learning rate of 1e-3 [11].

All relevant code and results will be made available for public access at <https://github.com/Janssena/DeepCompartmentModels.jl> at the time of publication.

4.2.3 *Simulation experiment*

We simulated a data set of 500 patients based on a previously published NLME model [5]. This model was developed using retrospective data from 119 patients with haemophilia A treated with an SHL FVIII concentrate perioperatively. This model predicted ζ based on patient weight, age, blood group, and surgical risk score. A two-compartment model with clearance (CL), central volume of distribution (V_1), inter-compartmental clearance (Q), and peripheral volume (V_2) parameters was used.

The goal of our simulation was to evaluate the accuracy of the DCM in sparse and dense data scenarios. For each patient, we simulated a single intravenous dose of 25–50 IUkg⁻¹ (rounded to nearest multiple of 250) of SHL FVIII concentrate at $t = 0$. Typical PK parameters were calculated based on samples from covariate distributions fit to the original data set. FVIII levels were simulated based on these PK parameters and collected at $t = 0.5h$ and every hour until $t = 48h$. Average simulated FVIII peak level was 0.89 IUml⁻¹ (0.43–1.31), and average trough level at $t = 48h$ was 0.09 IUml⁻¹ (0.01–0.21). Gaussian noise ($\sigma = 0.05$) was added to produce training measurements. Any resulting negative concentrations were fixed to zero. Multiple sets of measurements were collected to evaluate an extremely limited ($t = 24$), limited ($t = 8, 30$), routine ($t = 4, 24, 48$), and extensive ($t = 0.5, 4, 12, 24, 36$, and 48) sampling strategy [12]. The DCM was trained on 20, 60, or 120 patients representing data sets of low, medium, and large size, respectively. Corresponding test sets contained the remaining 480, 440, or 380 patients. Models were trained until MSE stopped improving. Both a standard DCM and DCM with initialisation were fit for all scenarios. A reasonable set of initialisation parameters $\zeta_0 = [150, 2500, 150, 2000]$ was used for CL (ml/h), V_1 (ml), Q (ml/h), and V_2 (ml), respectively. Training procedure was replicated five times to account for the influence of the random initialisation of w on the accuracy.

Accuracy of FVIII level predictions was defined as the percentage of predictions within a range of the "true" simulated FVIII level (without noise) evaluated at all simulated timepoints. This target range was set at 0.05 IUml⁻¹ for $\mu_{true} \geq 0.15$ IUml⁻¹, and at 0.02 IUml⁻¹ for $\mu_{true} < 0.15$. These values represent clinically relevant differences in the FVIII level. Because patients with levels above 0.15 IUml⁻¹ hardly suffer from joint bleeding, we chose this as the lower limit [13]. The 0.05 IUml⁻¹ range represents an estimate of assay accuracy. This range was decreased to 0.02 IUml⁻¹ to emphasise the importance of making accurate predictions of FVIII trough levels (e.g., < 0.15 IUml⁻¹). A large difference in accuracy between the train and test set was indicative of model over-fitting. The mean accuracy \pm one standard deviation (SD) was presented for each model.

Finally, speed of the algorithm was evaluated by determining the calculation time per epoch. We calculated the gradient and updated the parameter for 100 epochs, recorded the total duration, and presented the average time spend per epoch. We used a 16 GB, Intel Core i7-

9750H CPU computer for our tests. Models were trained on the CPU only.

4.2.4 Validation using real-world data sets

Following the simulation experiment, we compared the accuracy of a priori predicted perioperative FVIII levels of a DCM and NLME model using real-world data. Both models were developed on the retrospective data set from Hazendonk *et al.* [5]. Data from the OPTI-CLOT trial was used as an independent validation data set [6]. In this study, perioperative FVIII consumption was compared between PK-guided and standard dosing regimens. FVIII levels were actively monitored and dosing was adjusted following daily measurements if required.

The one-stage assay used in both data sets was known to significantly under-report FVIII levels from a β -domain deleted recombinant FVIII product (BDD-rFVIII; moroctocog alfa/ReFacto AF) [14]. The proposed DCM architecture did not support estimation of the effect of covariates that influence the drug concentration directly. We removed all patients who received this product (9 and 4 patients in the train and validation set, respectively). The final retrospective data set contained 110 patients with a total of 1380 perioperative FVIII measurements, and the validation set contained 62 patients with 526 measurements. Re-estimating the NLME model parameters on the retrospective data without these patients did not lead to meaningful differences so the final model was used as-is.

We fit a DCM based on patient weight, age, and having blood group O using a two-compartment model as these covariates have generally accepted biological significance with respect to FVIII drug dynamics. We used the same ζ_0 as in the simulation study. Additional covariates shared between the two data sets were von Willebrand factor antigen (VWF:Ag) and activity (VWF:Act) levels, haemophilia severity, and pre-assessed surgical risk score. They were added to the base set of covariates if inclusion improved objective function value on the training data. This was somewhat similar to a step-wise procedure, although we could not use p values as there were no explicit parametric assumptions. Accuracy of the resulting models was evaluated on the validation set. Models were trained for 100 epochs and the set of parameters w from the epoch resulting in the highest accuracy on the retrospective data set were selected. We again performed five replications of the training procedure, resulting in five

independently fit models. For the NLME model, the final model from Hazendonk *et al.* was implemented in NONMEM (ICON Development Solutions, version 7.4.2) [5]. Covariates used in the NLME model were patient weight, age, blood group, and surgical risk score. Accuracy was again represented as the percentage of predictions within 0.05 IUml^{-1} of measured FVIII levels greater than or equal to 0.15 IUml^{-1} , and 0.02 IUml^{-1} for levels < 0.15 .

4.3 RESULTS

4.3.1 DCM accuracy on simulated data

The accuracy of FVIII predictions by the DCM for the different scenarios is shown in table 4.3.1. In general, a higher number of measurements or training samples resulted in improved accuracy. However, accuracy was higher for the standard DCM trained on limited measurements compared to the routine set. Slight model over-fitting was seen when training on 20 samples but not for the other sample sizes. In all cases, we saw that initialisation using ζ_0 increased both train and test accuracy. When using initialisation, there was no large improvement in accuracy when increasing the number of measurements from three (routine) to six (extensive). Furthermore, using initialisation greatly improved model accuracy when only one measurement was available (from roughly 29% to 65–75%).

In figure 4.3.1, we have depicted the mean residuals including SD for the different sampling strategies at $n = 120$ or 20. For the standard DCM, we can appreciate that decreasing the number of training samples increases variance of the residuals, whereas decreasing the number of measurements increases bias. We also see that for all but the extended measurements set high bias can be seen for peak concentration predictions. For some scenarios, using initialisation is able to reduce this bias.

In figure 4.3.2, we have shown predictions for a random patient for each of the sampling strategies. Here, we can notice that an insufficient number of measurements can allow the standard DCM to predict unrealistic FVIII responses (figure 4.3.2d). Using initialisation, we guide the DCM to find a solution that follows an initial belief about the value of each of the PK parameters.

With respect to algorithm speed, we found that time spent per epoch increased proportional to the number of samples in the train

SAMPLING STRATEGY	N	STANDARD DCM		DCM WITH INITIALISATION	
		TRAIN	TEST	TRAIN	TEST
$t = 0.5,$ $4, 12, 24,$ $36, 48$	120	99.0 ± 0.21	99.1 ± 0.25	99.6 ± 0.12	99.4 ± 0.16
	60	93.3 ± 13.0	93.0 ± 12.5	98.9 ± 0.42	97.9 ± 0.18
	20	89.5 ± 1.09	84.4 ± 1.79	92.8 ± 1.76	88.7 ± 3.27
$t = 4,$ $24, 48$	120	65.2 ± 8.68	65.3 ± 8.86	97.8 ± 0.33	97.8 ± 0.41
	60	60.7 ± 0.61	59.5 ± 0.62	96.0 ± 0.85	94.8 ± 0.97
	20	58.2 ± 0.99	59.1 ± 0.71	96.3 ± 1.18	90.1 ± 2.00
$t = 8,$ 30	120	75.9 ± 0.65	76.1 ± 1.08	90.8 ± 6.63	90.3 ± 6.19
	60	72.4 ± 1.33	73.6 ± 1.19	81.4 ± 3.29	83.0 ± 3.08
	20	66.8 ± 1.78	61.2 ± 1.41	77.7 ± 4.82	76.5 ± 2.19
$t = 24$	120	28.6 ± 3.69	28.9 ± 5.31	76.2 ± 2.74	76.0 ± 2.41
	60	29.2 ± 1.21	29.4 ± 1.02	66.8 ± 2.23	65.2 ± 2.14
	20	29.6 ± 2.68	32.2 ± 1.92	73.7 ± 1.83	72.9 ± 1.80

Abbreviations: DCM = deep compartment model.

Table 4.3.1: Accuracy of predicted FVIII levels in the simulation experiment. Note: Train and test accuracy is represented as the percentage of predictions within 0.05 IU mL^{-1} of true simulated factor VIII levels ≥ 0.15 and within 0.02 IU mL^{-1} of levels < 0.15 . Time points are in hours. n is the number of patients in the train set. Test set size is the remainder of $500 - n$. Values are represented as the mean \pm one SD of five replicates.

set (table 4.A.1). The type of DCM or the number of available measurements did not affect computational time.

4.3.2 Comparison with NLME model using real-world data

In table 4.3.2, we show the accuracy of a priori predictions of the DCM and NLME model using real-world data. Only adding VWF:Ag to the base set of covariates resulted in an improvement of the objective function value. The DCM + VWF:Ag model showed improved accuracy on the validation set compared to the NLME model (23.1% vs. 21.6%). The base DCM had similar accuracy to the NLME model (22.0%). Time spent on training a single replicate for 100 epochs took $\sim 25\text{s}$.

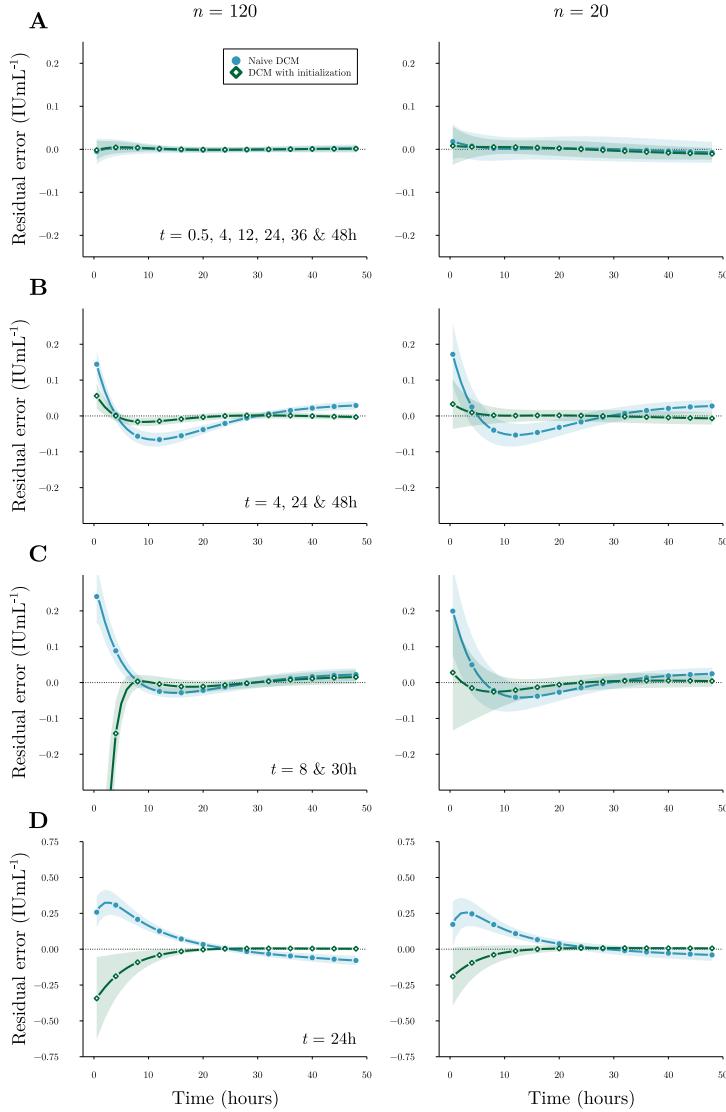


Figure 4.3.1: Bias and variance of residuals. Mean residuals on the test set of a single replicate of the standard DCM (circles), DCM with initialisation (diamonds), and corresponding SD (shaded areas) are shown for the extensive (a), routine (b), limited (c), and extremely limited (d) sampling strategies. Points were added for the purpose of comparison. Dotted line indicates zero residual error. Images on the left were trained on 120 patients, and images on the right on 20. Positive residuals indicate underestimation of FVIII levels while negative residuals indicate overestimation. Abbreviations: DCM = deep compartment model, FVIII = factor VIII.

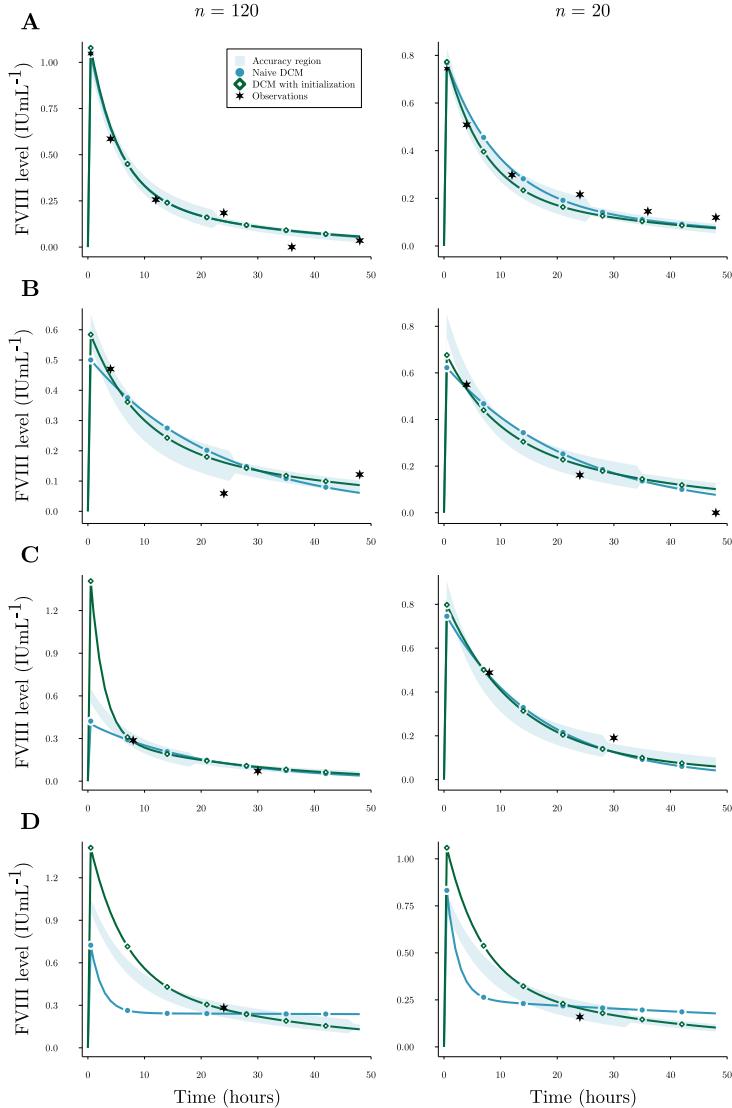


Figure 4.3.2: Examples of FVIII level predictions in the simulation experiment. Here, predicted FVIII levels by a single replicate of the standard DCM (circles) and DCM with initialisation (diamonds) are compared. The accuracy threshold (shaded area) is also shown. Points were added for the purpose of comparison. Results are shown for a single patient for the extensive (a), routine (b), limited (c), and extremely limited (d) sampling strategies. Stars represent the observed FVIII levels. Images on the left were trained on 120 patients, and images on the right on 20. Abbreviations: DCM = deep compartment model, FVIII = factor VIII.

MODEL	ACCURACY
NLME	21.9%
DCM	$22.0 \pm 0.417\%$
DCM + VWF:Ag	$23.1 \pm 1.12\%$

Abbreviations: DCM = deep compartment model, FVIII = factor VIII, NLME = nonlinear mixed effect, VWF:Ag = von Willebrand factor antigen.

Table 4.3.2: Accuracy of a priori predicted FVIII levels for the independent OPTI-CLOT data set. Note: Here we show the accuracy of the models as the percentage of predictions within 0.05 IU mL^{-1} of observed FVIII levels ≥ 0.15 . For observations < 0.15 this threshold was set at 0.02 IU mL^{-1} . DCM accuracy is shown as the mean accuracy of five independent runs \pm SD. The DCM + VWF:Ag model included VWF:Ag as an additional covariate. Bold text indicates the most accurate model.

In figure 4.3.3, the residuals of the NLME model and DCM are compared per 24h from the day of surgery. The residual error of DCM predictions suggest lower bias, as judged by the median residual error being generally within the accuracy threshold. In contrast, the NLME model more often underestimated FVIII levels. For all models, variance of the residual error started decreasing after $t = 72$.

In figure 4.3.4, we have shown the prediction by the DCM + VWF:Ag compared to the NLME model for six patients. Here, we see that the DCM can accurately represent the changing FVIII levels over time when subjected to complex dosing schemes. For some patients, the DCM and NLME model predicted concentrations are very similar.

4.4 DISCUSSION

In this study, we present a technique for improving the performance of ML models for predicting drug concentrations by constraining the solution space. Here, we have used a neural network to predict the latent parameters of a system of ODEs and determined its accuracy in different scenarios during a simulation experiment. We show that when using initialisation parameters, the accuracy of such an approach is high ($>80\%$) when training on medium-sized data sets with at least two measurements. Next, we compared the accuracy of the DCM to an NLME model using real-world data. The DCM displayed increased accuracy of FVIII level predictions on an independent validation set ($23.1\% \pm 1.12$ SD compared to 21.9% for the NLME model). Even though many measurements were available, achieved model accuracy

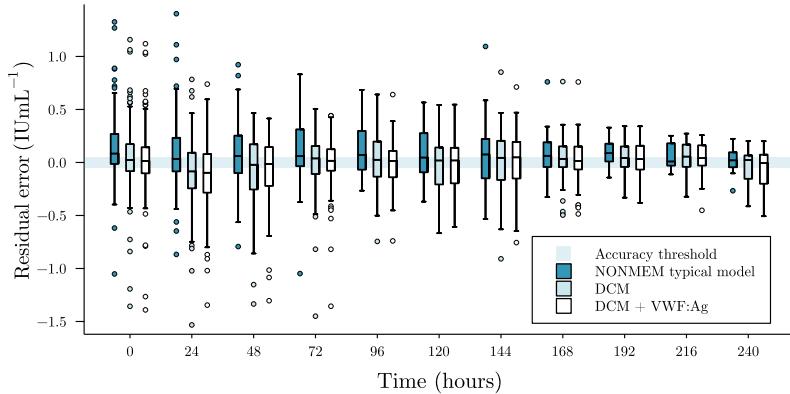


Figure 4.3.3: Box-plots of residual error of predicted perioperative FVIII levels. Here, we show the residual error of a priori predictions grouped per 24h for the NLME model (dark boxes), DCM (lightly shaded boxes), and DCM with VWF:Ag (white boxes). The shaded area indicates the 0.05 IUml^{-1} accuracy threshold. $t = 0$ corresponds to the day of surgery. Mean prediction from the five independent DCM runs was taken to calculate residual error. Positive residuals indicate underestimation of FVIII levels, whereas negative residuals indicate overestimation. Abbreviations: DCM = deep compartment model, FVIII = factor VIII, NONMEM = nonlinear mixed-effect modelling, VWF:Ag = von Willebrand factor antigen.

was lower compared to the simulation experiment. This is indicative of the complexity of predicting perioperative FVIII levels, where other (unknown) factors seem to contribute to the IIV.

In the simulation experiment, we found that the accuracy of the standard DCM was higher for the limited sampling strategy compared to the routine sampling strategy. This suggests that it is not only the number of measurements but also their timing that can affects model bias. This is reflected in figure 4.3.2b,c, where we can see that the routine sampling strategy leads to higher bias between $t = 4$ and $t = 24$ compared to the limited sampling strategy. For all scenarios, we found that using initialisation parameters improved prediction accuracy. Especially when training on smaller data sets ($n = 20$), bias of residual error greatly reduced compared to a standard DCM. In small data sets, there is likely not enough data to correctly characterise the relationship between the covariates and the PK parameters. When measurements were extremely limited, a standard DCM was completely free to choose how to fit the single FVIII level and often

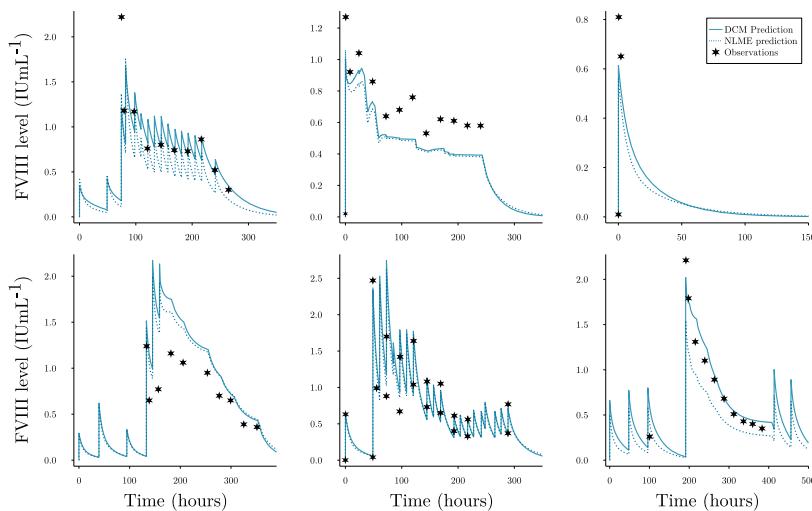


Figure 4.3.4: Examples of a priori perioperative predicted FVIII levels. DCM predictions represent the predicted FVIII levels by a single replicate of the DCM + VWF:Ag model. Stars represent observed FVIII levels. Both the prediction by the DCM (solid line) and the typical prediction from the NLME model (dotted line) are shown. For some patients, pre-surgery prophylactic doses are also shown. Abbreviations: DCM = deep compartment model, FVIII = factor VIII, NLME = nonlinear mixed-effect modelling, VWF:Ag = von Willebrand factor antigen.

degenerated to a flattened concentration curve (i.e., very low clearance; see figure 4.3.2d). By using initialisation, we can drive the model to follow an initial guess of compartment dynamics. However, we found that the current ζ_0 could still lead to a biased estimation of peak concentration predictions. Similar to choosing an informative prior in the Bayesian setting, choosing the "correct" ζ_0 can be difficult. In our case, we noticed that the DCM could maintain accurate predictions of the measurements while excessively adjusting V1. As no measurements were present at early timepoints for many of the scenarios, the model was not penalised for over or underestimating peak FVIII levels. It is thus important to choose ζ_0 carefully by, for example, monitoring the distribution of residual errors during training and adjusting initial estimates accordingly.

The results suggest, however, that a more rigid constraint against extreme predictions is required. One such approach would be to include a prior belief over the PK parameters and performing maximum a posteriori estimation. By setting a prior distribution over our param-

eters we can penalise more extreme estimates. However, in the case of a neural network, this prior has to be set over the weights of each layer. Choosing a correct weight distribution that matches our prior belief over the PK parameters is very complex, and is an area of active research [15, 16]. Another related improvement is the use of a Bayesian neural network [17]. Again, using a prior over the neural network weights, we can obtain a credible interval for our parameter estimates, similar to the standard error estimates NLME produces. This allows us to contribute a measure of certainty to the PK parameters, and identify patients for which the prediction is inaccurate. It might be difficult to implement such methods relating to prior selection so other approaches might have to be evaluated.

In the real-world experiment, the DCM trained using patient weight, age, having blood group O, and VWF:Ag achieved higher accuracy than the NLME model. Although this improvement was not extremely large, fitting and adjusting a DCM is far less time-consuming. Training the model required only roughly 25s, whereas development of NLME models can take far longer. A downside, however, can be that the DCM was programmed in the Julia programming language, which is unfamiliar to many pharmacometrists. We have made our model code publicly available and include a tutorial on how to fit a DCM model to any NLME compatible data set using only a few lines of code. This way, we hope to reduce the complexity of using this new technique. New covariates can simply be added to a base set of covariates and accuracy can be monitored during training. The method also allows for the user to simulate new treatment strategies by adjusting I_i . As seen in figure 4.3.4, the model accurately represents the changing concentration over time.

We have shown examples where we use a DCM to estimate the effect of all covariates, but it is also possible to add a layer where the relationship between a covariate and the PK parameters is explicitly stated. An example would be to use allometric scaling to represent the effect of weight on the PK parameters, while having the neural network learn the effect of the other covariates using standard layers. The practical use of this concept will have to be evaluated.

From the above experiments some limitations of the DCM have come to light. First, it is sometimes the case that no prior knowledge exists for choosing an appropriate compartment model to describe the drug concentrations. In these cases, we suggest fitting multiple DCM models with different model structures and inspect the solution in order to resolve model misspecification. Next, the proposed

architecture of the DCM does not support covariates that affect the predicted concentration directly. This has resulted in the removal of all patients in the data sets who received BDD-rFVIII. In the NLME model, this effect can be directly estimated in the model, whereas for the DCM estimating this quantity next to w can be difficult. The DCM also does not quantify any form of residual variability. Use of the MSE implicitly assumes simple additive error, where in many cases a combined additive and proportional error model is more appropriate. In addition, the model does also not quantify residual IIV, making the model potentially more susceptible to over-fitting. We have performed some prior work on combining the DCM with the extended least squares objective function as a possible solution to these problems [18]. We, however, found that the implementation is unstable and requires careful tuning of training parameters. More work is required to improve the random effect estimation when using neural networks. Finally, although the relationships between PK parameters and covariates can be visualised after fitting the DCM, understanding the relationships between covariates and PK parameters can be difficult. ML explanation methods, such as SHAP [19], can be performed in order to help visualise these relationships. Fact remains that neural networks are black box models, and the discussion of trust in ML method in the field of pharmacometrics is still in its infancy.

In conclusion, the DCM is a reliable tool for introducing ML models in population PK analysis. The DCM can automatically learn covariate relationships from data reducing the need for tedious covariate analysis. In contrast to other ML models, the DCM is based on compartment models allowing for the implementation of prior knowledge of drug dynamics. In addition, the DCM can be used with any dosing scheme, and allows for reliable extrapolation to different timepoints.

REFERENCES

- [1] Michael E Brier, Jacek M Zurada, and George R Aronoff. "Neural network predicted peak and trough gentamicin concentrations". In: *Pharmaceutical research* 12 (1995), pp. 406–412.
- [2] Hsiao-Hui Chow, Kristin M Tolle, Denise J Roe, Victor Elsberry, and Hsinchun Chen. "Application of neural networks to population pharmacokinetic data analysis". In: *Journal of pharmaceutical sciences* 86.7 (1997), pp. 840–845.
- [3] Rong Liu, Xi Li, Wei Zhang, and Hong-Hao Zhou. "Comparison of nine statistical model based warfarin pharmacogenetic dosing algorithms using the racially diverse international warfarin pharmacogenetic consortium cohort database". In: *PloS one* 10.8 (2015), e0135784.

- [4] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. "Neural ordinary differential equations". In: *Advances in neural information processing systems* 31 (2018).
- [5] Hendrika Hazendonk, Karin Fijnvandraat, Janske Lock, Mariëtte Driessens, Felix Van Der Meer, Karina Meijer, Marieke Kruip, Britta Laros-van Gorkom, Marjolein Peters, Saskia de Wildt, et al. "A population pharmacokinetic model for perioperative dosing of factor VIII in hemophilia A patients". In: *haematologica* 101.10 (2016), p. 1159.
- [6] Iris van Moort, Tim Preijers, Laura H Bikkemps, Hendrika CAM Hazendonk, Johanna G van der Bom, Britta AP Laros-van Gorkom, Erik AM Beckers, Laurens Nieuwenhuizen, Felix JM van der Meer, Paula Ypma, et al. "Perioperative pharmacokinetic-guided factor VIII concentrate dosing in haemophilia (OPTI-CLOT trial): an open-label, multicentre, randomised, controlled trial". In: *The Lancet Haematology* 8.7 (2021), e492–e502.
- [7] Tze Leung Lai, Mei-Chiung Shih, and Samuel P Wong. "A new approach to modeling covariate effects and individualization in population pharmacokinetics-pharmacodynamics". In: *Journal of Pharmacokinetics and Pharmacodynamics* 33 (2006), pp. 49–74.
- [8] James Lu, Kaiwen Deng, Xinyuan Zhang, Gengbo Liu, and Yuanfang Guan. "Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens". In: *Iscience* 24.7 (2021).
- [9] Jonathan T. Barron. "Continuously Differentiable Exponential Linear Units". In: (2017). arXiv: 1704.07483 [cs.LG]. URL: <https://arxiv.org/abs/1704.07483>.
- [10] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. "Searching for Activation Functions". In: (2017). arXiv: 1710.05941 [cs.NE]. URL: <https://arxiv.org/abs/1710.05941>.
- [11] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: (2017). arXiv: 1412.6980 [cs.LG]. URL: <https://arxiv.org/abs/1412.6980>.
- [12] S Björkman, P Collins, et al. "Measurement of factor VIII pharmacokinetics in routine clinical practice". In: *Journal of Thrombosis and Haemostasis* 11.1 (2013), pp. 180–182.
- [13] IEM Den Uijl, K Fischer, JG Van Der Bom, DE Grobbee, FR Rosendaal, and I Plug. "Analysis of low frequency bleeding data: the association of joint bleeds according to baseline FVIII activity levels". In: *Haemophilia* 17.1 (2011), pp. 41–44.
- [14] Anthony R Hubbard, Lynne J Weller, and Sally A Bevan. "A survey of one-stage and chromogenic potencies in therapeutic factor VIII concentrates". In: *British journal of haematology* 117.1 (2002), pp. 247–247.
- [15] Mariia Vladimirova, Jakob Verbeek, Pablo Mesejo, and Julian Arbel. "Understanding priors in Bayesian neural networks at the unit level". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6458–6467.
- [16] Wanqian Yang, Lars Lorch, Moritz A. Graule, Srivatsan Srinivasan, Anirudh Suresh, Jiayu Yao, Melanie F. Pradier, and Finale Doshi-Velez. "Output-Constrained Bayesian Neural Networks". In: (2019). arXiv: 1905.06287 [cs.LG]. URL: <https://arxiv.org/abs/1905.06287>.

- [17] Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. "Hands-on Bayesian neural networks—A tutorial for deep learning users". In: *IEEE Computational Intelligence Magazine* 17.2 (2022), pp. 29–48.
- [18] A. Janssen, F.W.G. Leebeek, M.H. Cnossen, and R.A.A. Mathôt. "The neural mixed effects algorithm: leveraging machine learning for pharmacokinetic modelling". In: *29th Annual Meeting of the Population Approach Group in Europe, Abstr 9826*. 2021. URL: <https://www.page.org>.
- [19] Scott M. Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2017, 4768–4777.

APPENDIX

4.A SUPPLEMENTARY TABLES

SAMPLING STRATEGY	N	TIME PER EPOCH STANDARD DCM (SECONDS)	TIME PER EPOCH DCM WITH INITIALISATION (SECONDS)
$t = 0.5, 4, 12, 24, 36, 48$	120	0.083	0.083
	60	0.046	0.042
	20	0.015	0.015
$t = 24$	120	0.091	0.084
	60	0.043	0.042
	20	0.016	0.014

Abbreviations: DCM = deep compartment model.

Table 4.A.1: Time spend per epoch in the simulation experiment. Time per epoch (seconds) represents the average time spend on gradient calculation and parameter update when training for 100 epochs.

ON INDUCTIVE BIASES FOR THE ROBUST AND INTERPRETABLE PREDICTION OF DRUG CONCENTRATIONS USING DEEP COMPARTMENT MODELS

Alexander Janssen, Frank C. Bennis, Marjon H. Cnossen, and Ron A.A.

Mathôt

Journal of Pharmacokinetics and Pharmacodynamics (2024): 1-12.

ABSTRACT

Conventional pharmacokinetic (PK) models contain several useful inductive biases guiding model convergence to more realistic predictions of drug concentrations. Implementing similar biases in standard neural networks can be challenging, but might be fundamental for model robustness and predictive performance. In this study, we build on the deep compartment model (DCM) architecture by introducing constraints that guide the model to explore more physiologically realistic solutions. Using a simulation study, we show that constraints improve robustness in sparse data settings. Additionally, predicted concentration–time curves took on more realistic shapes compared to unconstrained models. Next, we propose the use of multi-branch networks, where each covariate can be connected to specific PK parameters, to reduce the propensity of models to learn spurious effects. Another benefit of this architecture is that covariate effects are isolated, enabling model interpretability through the visualisation of learned functions. We show that all models were sensitive to learning false effects when trained in the presence of unimportant covariates, indicating the importance of selecting an appropriate set of covariates to link to the PK parameters. Finally, we compared the predictive performance of the constrained models to previous relevant population PK models on a real-world data set of 69 haemophilia A patients. Here, constrained models obtained higher accuracy compared to the standard DCM, with the multi-branch network outperforming previous PK models. We conclude that physiological-based constraints can improve model robustness. We describe an interpretable architecture which aids model trust, which will be key for the adoption of machine learning-based models in clinical practice.

5.1 INTRODUCTION

Selection of appropriate drug dosage is an important aspect underlying the efficacy of treatment and the prevention of drug-induced toxicity. However, selecting optimal doses on an individual basis can be challenging, which has historically led most to follow weight-based dosing regimens. It has frequently been reported that these conventional regimens can result in considerable inter-individual variability of achieved drug concentrations [1–3]. For example, in a study of haemophilia A patients receiving factor VIII (FVIII) concentrate, weight-based dosing (50 IU/kg) was observed to result in as high as a tenfold variation in peak FVIII levels [1]. Such large discrepancies could be especially concerning during surgical procedures, where maintaining appropriate FVIII levels is thought to be important for reducing the risk of (severe) bleeding [4]. Previous studies have shown that personalisation of treatment based on the individual pharmacokinetic (PK) profile of the patient resulted in improved achievement of target FVIII levels during the perioperative setting compared to weight-based dosing [4].

Population PK involves the study of inter-individual differences in drug absorption, distribution, metabolism, and elimination [5]. PK models leverage mathematical representations of these processes to predict *in vivo* drug concentrations. PK models estimate a set of latent variables (PK parameters; which for example represent drug clearance or volume of distribution) based on covariate data using hand-picked closed-form expressions describing covariate effects. These estimates are then fed into a system of differential equations—so-called compartment models—which encode prior knowledge of drug distribution [6]. Considerable time and expertise is required for the development of population PK models, partly due to the manual selection of covariates and fine-tuning of the functions describing their effect. Another downside of the classical approach is that more complex and unconventional functions are rarely considered in favour of linear or power functions. This might hurt the predictive performance of such methods. Finally, development of population PK models rarely involves internal or external validation procedures, and the use of simple covariate effects and significance testing of model components might mask risks of overfitting and poor generalisability.

Recently there has been increased interest in the use of machine learning (ML) based approaches for performing PK analysis [7]. Several methods have been suggested to screen covariates based on feature

importance [8, 9] or to inform function selection [10, 12]. Population PK models can also directly leverage ML methods which has the potential to improve model accuracy while reducing time spent on model development by for example directly learning drug kinetics [13] or covariate implementation [14] from data. However, the design of a reliable approach in the context of pharmacometrics is non-trivial: drug concentration data is often sparsely and irregularly sampled, while treatment interventions (e.g. drug administration) can be notably different between individuals. Additionally, we wish to use these models to evaluate counterfactual scenarios (evaluating different treatment strategies) meaning that these models should reliably extrapolate to unseen data. It might therefore be necessary to include prior knowledge into model structure to allow for more data-efficient learning. Most ML methods are also prone to overfitting, so it might be difficult for physicians to place their trust in these methods without some form of interpretability or prediction uncertainty [15, 16]. The European Commission's proposed Regulation on Artificial Intelligence also explicitly places such requirements on ML models before they can be used for healthcare applications (AI Act recital 47, <https://www.euaiact.com/recital/47>, accessed 19 December 2023). A potential positive consequence of these requirements might be that ML-based algorithms will be more extensively validated compared to classical methods.

5.1.1 *Inductive biases*

In population PK models, three sources of inductive biases help to improve model convergence: the structure of the compartment model, the equations chosen to represent covariate effects, and the use of informed initial estimates of model parameters. In contrast, naive neural networks encode weak inductive biases for dealing with tabular or time-series data. It can be shown that naive neural networks incorrectly handle important variables such as dose leading to incorrect extrapolation [7]. This problem is inherent to the inclusion of dose as a model input and is likely equally problematic in other standard ML methods (e.g. random forests and gradient boosting) [13]. Importantly, *Lu et al.* have even shown how neural network architectures specialised for time series predictions such as recurrent neural networks (RNNs) and long short-term memory models (LSTMs) fail to reliably extrapolate to unseen dosing schedules [13]. These limitations cannot be overcome without a causal use of variables such as dose,

which potentially necessitates the use of ordinary differential equation (ODE) based methods [11]. Neural-ODE-based approaches, where treatment directly affects the latent state of the model at discrete time points, indeed do correctly respond to new and complex dosing regimens. These models are fully data-driven, greatly simplifying model development. Multiple Neural-ODE-based approaches have been suggested which can be used to learn unknown parts of the dynamical system [17, 18], or augment expert models by learning latent effects [19]. The deep compartment model (DCM) approach by [14] uses neural networks to predict the PK parameters for a compartment model, implementing doses as time-based events directly affecting drug concentrations in specific compartments. This approach has the benefit of predicting the same variables used in PK models allowing for the comparison of results. Additionally, prior knowledge on drug kinetics can be included through the compartment model, potentially improving data efficiency.

Explicitly learning the dynamical system underlying observations likely serves as a useful inductive bias to improve the reliability of predictions. However, as drug concentration measurements are often sparse, the solution space given the data of potential models for these ODE-based methods can still be considerably large (see Fig. 5.1.1). As a result, unconstrained models might place similar likelihood on many different model parametrizations (Fig. 5.1.1a). Alternatively, well-specified models with physiological-based constraints result in more concentrated posterior distributions. If these biases are well-adjusted and informative, the resulting posterior might be more similar to the true model. In Fig. 5.1.1b, we depict an example of models within the solution space of unconstrained models. Some of the models might learn potentially physiologically implausible or unlikely concentration-time curves (dashed lines).

In this work, we introduce simple inductive biases within the deep compartment model framework by placing domain-specific constraints on model architecture to improve robustness. We define models as not robust if they have a high propensity of learning spurious effects. We investigated effects on model accuracy and stability of providing bounds for the value of the PK parameters, estimating global values for difficult to identify parameters, and connecting covariates to specific PK parameters. Expanding on the latter constraint, fully-connected neural networks encode an implicit assumption that part of the signal potentially originates from complex interactions between the covariates. Model generalisability and robustness can potentially

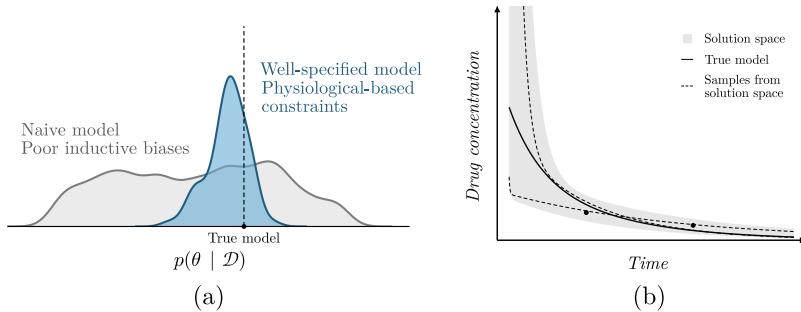


Figure 5.1.1: Schematic representation of the solution space of naive and well-specified models. In (a), we show the solution space of naive and well-specified models. In (b), the solution space of a model with poor inductive biases is shown. Samples from the solution space (dashed lines) can be physiologically unrealistic when data is sparse, and can differ greatly with the true solution (solid line).

be improved by only linking covariates with causal effects to specific PK parameters in sub-models. An additional benefit of this approach is that the learned function from each sub-model can be visualised, enabling model interpretation.

5.2 METHODS

5.2.1 Problem definition

Our focus is on haemophilia A, a blood clotting disorder where a deficiency of FVIII results in elevated (spontaneous) bleeding risk. Haemophilia A patients are treated by intravenous injection of FVIII at regular intervals. The PK of FVIII is often described using a two-compartmental structure, where the first compartment represents the distribution of FVIII into the blood and the second is often thought to represent the initial rapid clearance of FVIII or its binding to intra or extra-vascular space [20–22]. The two-compartmental model can be represented by the following system of partial differential equations:

$$\begin{aligned} \frac{dA_1}{dt} &= \frac{I}{V_1} + A_2 \cdot k_{21} - A_1(k_{10} + k_{12}) \\ \frac{dA_2}{dt} &= A_1 \cdot k_{12} - A_2 \cdot k_{21} \end{aligned} \quad (5.1)$$

Here, the rate constants k describe the flow between the compartments specified in its subscript, A_1 represents the concentration in the

1st compartment (and so on), and I represents the rate of drug entering the first compartment after drug administration. The rate constants are functions of the PK parameters: $k_{10} = \frac{CL}{V_1}$, $k_{12} = \frac{Q}{V_1}$, $k_{21} = \frac{Q}{V_2}$, with $\mathbf{z} = \{CL, V_1, Q, V_2\}$ referring to clearance, inter-compartmental clearance, central distribution volume, and peripheral distribution volume, respectively.

Consider a population of n individuals with $\mathcal{D} = (x^{(i)}, t^{(i)}, y^{(i)})_{i \in [1..n]}$, each with irregular drug concentration measurements $y^{(i)} \in \mathbb{R}_+^K$ sampled over time horizon $t^{(i)} \in [0, T_i]$ with T_i indicating the follow-up time for individual i . Drug concentration predictions are produced based on the compartment model and a matrix of interventions $\mathbf{I}^{(i)}$ containing information on time of dose, dosage, and infusion rates that affect the integrator at the specified time points:

$$\hat{y}^{(i)}(t) = A(t; \mathbf{z}^{(i)}, \mathbf{I}^{(i)}) \quad (5.2)$$

In non-linear mixed effects models, individual estimates of each of the PK parameters $\zeta^{(i)} \in \mathbb{R}_+^M$ are obtained based on covariates $\mathbf{x}^{(i)} \in \mathbb{R}^D$ and subject-specific random effects $\eta^{(i)} \sim \mathcal{N}(\mathbf{0}, \Omega)$, where Ω is a $M \times M$ covariance matrix when random effects are included on all PK parameters. The following implementation is frequently observed within the pharmacometrics literature:

$$z_m^{(i)} = \theta_m \cdot \exp(\eta_m^{(i)}) \cdot \prod_s^{S_m} f_s(x_s; \theta_s) \quad (5.3)$$

Here, θ represent model fixed effect parameters and $S_m \subset [1..D]$ indicates the subset of covariates used to predict z_m . After specifying a model for the residual error $\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma)$ on $y^{(i)}$ (e.g. additive, proportional, or a combination of both), model parameters $\Theta = \{\theta, \Omega, \Sigma\}$ can be optimised by maximising:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}} \mathcal{L}(\Theta) = \prod_{i=1}^N p(y^{(i)} | \Theta, \eta^{(i)}) p(\eta^{(i)}) \quad (5.4)$$

As previously mentioned, development of non-linear mixed effects models requires considerable time and expertise, partly due to the manual selection of covariates and the functions f to represent their effect on \mathbf{z} . In DCMs, the fixed effect model is learned by a neural network ϕ with parameters w , and the covariates are used to predict typical PK parameters $\zeta^{(i)}$:

$$\zeta^{(i)} = \phi(\mathbf{x}^{(i)}; w) \quad (5.5)$$

And the model minimises the squared error:

$$\hat{w} = \underset{w}{\operatorname{argmin}} \mathcal{L}(w) = \sum_{i=1}^N \sum_{k=1}^{K_i} \left(y_k^{(i)} - A(t_k^{(i)}; \zeta^{(i)}, \mathbf{I}^{(i)}) \right)^2 \quad (5.6)$$

These models are relatively unconstrained in their prediction of $\zeta^{(i)}$, as long as it results in low error with respect to the observations. It can thus be the case that the model is not penalised for making extreme predictions outside of the observed data.

5.2.2 Model constraints

We propose three simple approaches for constraining the solution space of DCMs (Fig. 5.2.1). First, boundary conditions were imposed on the PK parameters by using a transformed sigmoidal function following the output layer of the neural network (referenced as boundary constraint; Fig. 5.2.1b). The boundaries can be set empirically based on prior knowledge. For example, bounds for the volume of distribution of drugs tightly bound to plasma proteins can be based on the expectation that the plasma volume of a typical male is roughly around 46-52 mL/kg [23]. Lower bounds of [0, 0.3, 0.05, 0] and upper bounds of [0.5, 7, 0.5, 2] for respectively CL (L/h), V_1 (L), Q (L/h), and V_2 (L) were used.

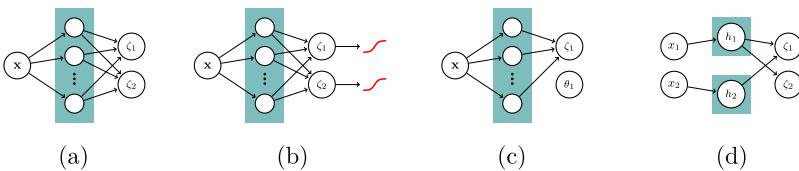


Figure 5.2.1: Graphical models representing the model structure of the proposed architectures. Naive (a), boundary constraint (b), global parameter (c), and multi-branch network (d) architectures are depicted. Nodes represent neurons, with the coloured box representing the hidden layer of the neural network.

Next, global parameters θ for a subset of the PK parameters were estimated in parallel to w (referenced as *global parameters constraint*; Fig. 5.2.1c). We chose to estimate $\theta = \{Q, V_2\}$ since these parameters

affect the early distribution of FVIII, and drug concentration measurements at early time points are usually too sparse to identify covariate effects on these parameters.

Finally, we describe a neural network architecture where each covariate (or specific combinations thereof) are connected to specific PK parameters via independent sub-models, whose predictions are combined using a product (referenced as the multi-branch network; Fig. 5.2.1d). This architecture is similar to a generalised additive model, using product accumulation rather than the sum of covariate effects. The use of a product matches the standard implementation of covariates in population PK models (Eq. 3), and facilitates the interpretation of the clinical relevance of each covariate. For example, covariates resulting in a maximal net change 20% of the corresponding PK parameter are often deemed clinically insignificant in the pharmacometrics literature [24]. An additional benefit of the approach is that the output of each sub-model can be visualised, allowing for the interpretation of the learned covariate effects. A schematic overview of the multi-branch network is provided in Supplementary Fig. 5.A.1.

More details about the specific implementation of the model constraints can be found in Supplementary Data 5.B.

5.2.3 *Synthetic experiments*

5.2.3.1 *Data generation*

We simulated a data set of haemophilia A patients based on population data from the American National Health and Nutrition Examination Survey (NHANES) [25]. The weight, height, and age of 756 male individuals without missing data were collected from the data set and used as covariates. FVIII levels were simulated based on an existing population PK model of an extended half-life FVIII concentrate [26]. This model was chosen as it used an estimate of the fat-free mass (FFM) from [27] to predict FVIII clearance (CL) and central volume of distribution (V_1):

$$FFM = \left(0.88 + \frac{1 - 0.88}{1 + \frac{Age}{13.4}} \right) \cdot \left(\frac{9270 \cdot Weight}{6680 + (216 \cdot BMI)} \right) \quad (5.7)$$

This allowed for the comparison of model accuracy when training models using FFM directly as well as using its components weight, height (as part of BMI), and age. This could give an indication of

the accuracy at which non-linear interactions of covariates could be learned.

The previous PK model was based on a two-compartment model, with inter-individual variability on the CL and V_1 parameters. Using the structural equations reported by [26], typical estimates of the PK parameters were produced. Next, samples of the random effects were drawn to produce individual estimates of the PK parameters. Each individual received a dose of 50 IU/kg, rounded to the nearest 250 IU. To allow for stochasticity of measurement times, samples were taken from a multivariate normal distribution $t^{(i)} \sim \mathcal{N}([4, 24, 48], \sigma = [2, 5, 5])$. Sampling times were truncated at $t = 0.25$ (i.e. 15 min after dose) to prevent samples at negative time points or too close to the time of dose administration. Finally, the ODE was solved based on the individual PK parameters to simulate FVIII levels for each individual and additive error ($\sigma = 5.0$ IU/dL) was added to create the training data.

5.2.3.2 *Evaluation of model constraints*

Prediction accuracy of the proposed constraints was compared to a naive neural network as well as the initialisation approach suggested in [14]. In all experiments, covariates were scaled between 0 and 1 using min–max scaling. Models were trained using patient weight, height and age, or FFM and age. A two compartment model was used. Each neural network was trained using a single hidden layer of either 8, 32, or 128 neurons followed by the swish activation function [28]. A softplus activation function was used in the output layer of the naive neural network as well as in the model estimating global parameters for Q and V_2 to constrain latent variables to \mathbb{R}_+ . All models were trained for 500 epochs using the ADAM optimiser with a learning rate of 1e-2 [29]. We found that these settings were sufficient for each model to converge before the end of optimisation. First, prediction accuracy and robustness of the naive, initialisation, boundary, and global parameter models were compared. The multi-branch network was not tested in this context due to similarities to the global parameter model. Each model was fit to a random subset of the simulated data of size 20, 60, or 120 to represent data sets of small, medium, and large size, respectively. A Monte Carlo cross validation of 20 different train and test sets was performed in order to estimate the stability of model predictions. In addition, model training was replicated five times on each train-test split. This resulted in a total number of 100 replicates of

each model, which was deemed sufficient to estimate model variability given our computational budget. Model accuracy was represented by the root mean squared error (RMSE) of predicted FVIII levels compared to the true, simulated concentration–time curves on the test set. To this end, true and predicted FVIII levels were collected at five minute intervals until $t = 72h$ as a means to approximate the error compared to the full concentration–time curve. Models were compared in terms of their median RMSE over the 20 data sets and five model replicates. Model robustness for each of the architectures was represented by the percentage of models with RMSE greater than 150% of the median RMSE (references as divergent models).

In order to evaluate differences between using fully-connected versus multi-branch neural networks, the data set was augmented with two continuous and one categorical covariate without correlations to the other covariates. For the continuous covariates, random samples were drawn from Uniform(0,1) distributions, while for the categorical covariate samples were randomly assigned to one of five categories with equal probability. Next, three models with global Q and V_2 parameters were fit to FFM, age, and the noise covariates: (1) a fully-connected model, (2) a multi-branch network with all covariates independently connected to CL and V_1 , and (3) a multi-branch network with the ground truth covariate connections as used in the simulation (referenced as the causal model). Fully-connected models were trained using a single hidden layer of 32 neurons. The number of neurons in the hidden layer of each sub-model was set to 16 to ensure that models had roughly similar number of parameters. Accuracy was again compared using the RMSE. Results were compared with the global parameter models trained on FFM and age from the first experiment (models trained using 32 neurons). Covariate effects from the multi-branch network were visualised to facilitate model interpretation (See Supplementary Data 5.D for implementation details).

All model code and synthetic data will be made available at <https://github.com/Janssena/dcm-constrained>.

5.2.4 Real-world experiments

We compared the predictive performance of two previously published population PK models [30, 31] to the DCM with or without the proposed constraints and a Neural-ODE based model [13]. Data consisted of 69 severe haemophilia A patients who received a single dose of 25–50 IU/kg standard half-life FVIII. For each patient, three measure-

ments were available roughly 4, 24, and 48 h after dose. Available covariates without missing data were patient weight, height, age, and blood group.

The population PK model by [22] included the effect of weight on all PK parameters, as well as the effect of age on CL . The model implemented allometric scaling, which is very common in PK models. We also evaluated the performance of a more recent model by [31]. Instead of using weight, this model implements the effect of FFM on clearance and volume of distribution. An effect of patient age was also included on clearance. Since it is well documented that patients with blood group O have higher FVIII CL compared to non-O patients, we also fitted models including a proportional effect of having blood group O on CL [32].

DCMs were fit using neural networks with a single layer containing 8, 32 or 128 neurons (halved for each sub-model for the multi-branch network). For the fully-connected network, model input was patient weight, height, age, and BGO. Models were fit without constraints, using boundary constraints (same as used during simulation experiments), and using global parameters for Q and V_2 . In the multi-branch network clearance was predicted based on patient a combination of weight and height, age, and BGO, while estimating volume of distribution based on a combination of weight and height. Global parameters were estimated for Q and V_2 in all models.

For the NeuralODE based model we followed the general architecture by *Lu et al.* [13]. Hyper-parameters were the number of neurons in the encoder, NeuralODE, and decoder (8, or 32), the number of hidden layers (1 or 2) in the NeuralODE, and the number of the latent variables (2 or 6). Encoder and decoder consisted of a single hidden layer. Tanh activation functions were used in the NeuralODE to improve model stability. Model input was patient weight, height, age, and BGO and values were normalised between -1 and 1. This marks an important difference to the model by *Lu et al.*, where part of the drug concentration measurements were used as input to the encoder and decoder.

5.2.5 Model training and evaluation

A ten-fold cross-validation was performed for both the non-linear mixed effects models and DCMs. Both PK models were implemented in the NONMEM software (ICON Development Solutions, Ellicott City, MD) and model parameters were re-estimated on each full train

fold. Exponents of the effects of weight on the PK parameters were not re-estimated in the model by [22] since they follow the concept of allometric scaling. The accuracy of typical predictions were reported. The ML models were trained for 4000 epochs which was more than sufficient for model convergence, and neural network weights resulting in the lowest validation error (20% of training fold) were saved. Hyper-parameter selection was performed by comparison of the RMSE on the validation sets. Results for the models with lowest average validation error were presented. The average RMSE of predictions with respect to the test fold was reported.

5.3 RESULTS

5.3.1 *Constraints improve model robustness*

In the first experiment, highest model accuracy was generally obtained when using a hidden layer size of 8 neurons (see table 5.3.1). Results for models trained with larger hidden layer sizes can be found in Supplementary Tables 5.A.1 and 5.A.2. All models seemed to perform similarly well when sufficient data was available. When training on smaller data-sets, the median RMSE and its variance increases for all models. However, when training naive models, a relevant proportion of models (18%) presented with a highly divergent error on the small data set (mean RMSE 44.4 IU/dL).

In contrast, model accuracy was more stable when using model constraints, with only two divergent models (0.33%) over all models with global parameters (including those with larger hidden layer sizes). Setting boundary constraints reduced the number of divergent models compared to the previously suggested approach of initialisation, but was less effective compared to the global parameter model (nine divergent models overall). Looking at the naive models fit with 128 neurons, both model accuracy and robustness was negatively affected when training at lower sample sizes (Supplementary Table 5.A.1). Median RMSE for the naive model trained on 20 samples increased from 14.7 to 16.5 IU/dL when changing hidden layer size from 8 to 128 neurons. In contrast, models fit using global parameters were almost unaffected by hidden layer size in the same context (RMSE 13.9 to 14.1 IU/dL). Models trained using FFM and age resulted in slightly more accurate predictions when trained on 20 samples, with almost no differences in medium to large data sets. A more extensive investigation of the effect of the constraints on model training can

MODEL	MEDIAN RMSE ± ONE SD (%-AGE DIVERGENT)		
	N=20	N=60	N=120
WEIGHT, HEIGHT, AGE			
None	14.6 ± 14 (18)	13.1 ± 1.2 (0)	12.3 ± 0.34 (0)
Initialisation	15.3 ± 21 (6)	12.9 ± 2.0 (2)	12.0 ± 0.45 (0)
Boundary	14.9 ± 2.5 (3)	12.6 ± 0.55 (0)	12.0 ± 0.45 (0)
Global parameters	13.9 ± 0.94 (0)	12.9 ± 0.44 (0)	12.3 ± 6.0 (1)
FFM, AGE			
None	14.1 ± 10 (12)	12.8 ± 0.71 (0)	12.2 ± 0.39 (0)
Initialisation	14.2 ± 16 (6)	12.5 ± 1.2 (2)	11.9 ± 0.3 (0)
Boundary	13.8 ± 1.2 (0)	12.4 ± 0.33 (0)	11.9 ± 0.3 (0)
Global parameters	13.5 ± 0.75 (0)	12.6 ± 0.35 (0)	12.2 ± 0.38 (0)

Abbreviations: RMSE = root mean squared error, SD = standard deviation.

Table 5.3.1: Test set accuracy and divergence rate for models with a hidden layer size of 8. Median RMSE over all replicates of model training (5×20 data sets) during experiment 1 is reported along with its standard deviation.

be found in Supplementary Data 5.C. Here, we found that divergent behaviour was specific to certain data folds, and was related to the estimate of V2. Adding constraints to this specific parameter was sometimes sufficient to improve models.

Next, we inspected the predicted concentration–time curves from each model. In Fig. 5.3.1, we show the predictions for a single, representative patient for naive (a), boundary (b), and global parameter (c) models. Here, we see that all models accurately predict the three observed FVIII levels. However, the naive model seems biased to predict unrealistically high FVIII peak levels (with predictions for some patients as high as 1340 IU/dL). In contrast, the constrained models resulted in less extreme and more similar solutions.

5.3.2 False covariates degrade model performance

We then compared the fully-connected and multi-branch networks on the augmented data set (see Table 5.3.2). The addition of false covariates degraded the accuracy of the fully-connected network with global parameters when using small data sets compared to the first experiment (RMSE of 18.1 vs. 13.3 IU/dL). We found that, initially, the multi-branch network using all covariates depicted high error

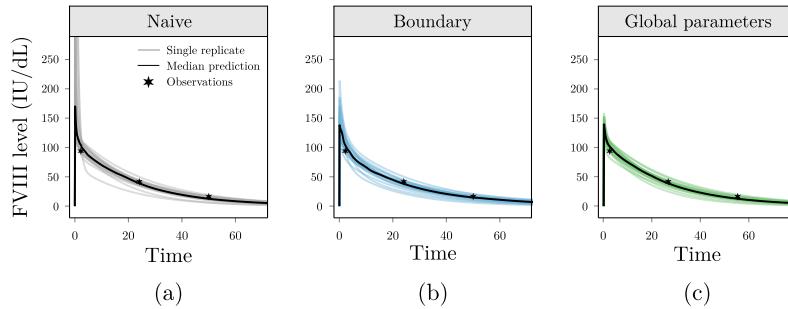


Figure 5.3.1: Predicted concentration–time curves from the proposed constraints are more realistic compared to naive models. Results are shown for the naive (a), boundary constraint (b), and global parameter (c) model. The median prediction (black line) over the 20 data set replicates (lightly coloured lines) along with the observations (stars) are shown for the same patient.

in several replicates ($\text{RMSE} > 40 \text{ IU/dL}$). In these replicates, poor initialisation resulted in initial V_1 estimates close to zero, resulting in high peak predictions as seen in Fig. 5.3.1a. To solve this issue, we initialised the bias of the neuron connecting to V_1 in the final layer of each sub-model to 0.5, increasing initial estimates close to 1 L. The resulting model performs slightly better at low sample sizes compared to the fully-connected network ($\text{RMSE } 15.6 \text{ vs. } 18.1 \text{ IU/dL}$). This suggests that part of the decrease in accuracy of the fully-connected network might be related to the model learning spurious interactions between the covariates.

MODEL	MEDIAN RMSE \pm ONE SD (%-AGE DIVERGENT)		
	N=20	N=60	N=120
Fully-connected	18.1 ± 2.9 (2)	13.6 ± 0.52 (o)	12.7 ± 0.35 (o)
Multi-branch	15.6 ± 2.9 (1)	12.8 ± 0.48 (o)	12.1 ± 0.19 (o)
Causal	13.3 ± 1.0 (o)	12.5 ± 0.34 (o)	12.1 ± 0.25 (o)

Abbreviations: RMSE = root mean squared error, SD = standard deviation.

Table 5.3.2: Introduction of noise covariates deteriorates accuracy of fully-connected networks. Median RMSE of the test set over all replicates of model training (5×20 data sets) is reported along with its standard deviation.

By including only true effects, the causal model achieved very similar accuracy to the global parameter model from the first experiment

at all data set sizes. In Fig. 5.3.2 we depict the learned covariate effects for the network containing noise covariates. As the number of training samples decreases, the variance of learned functions across replicates seemed to increase (i.e. the functions became more diverse). When trained on $n = 20$, the effect of the noise covariates on clearance was quite substantial in some replicates. It is still possible to identify these covariates as unimportant overall, as their mean effect over replicates is close to 1.

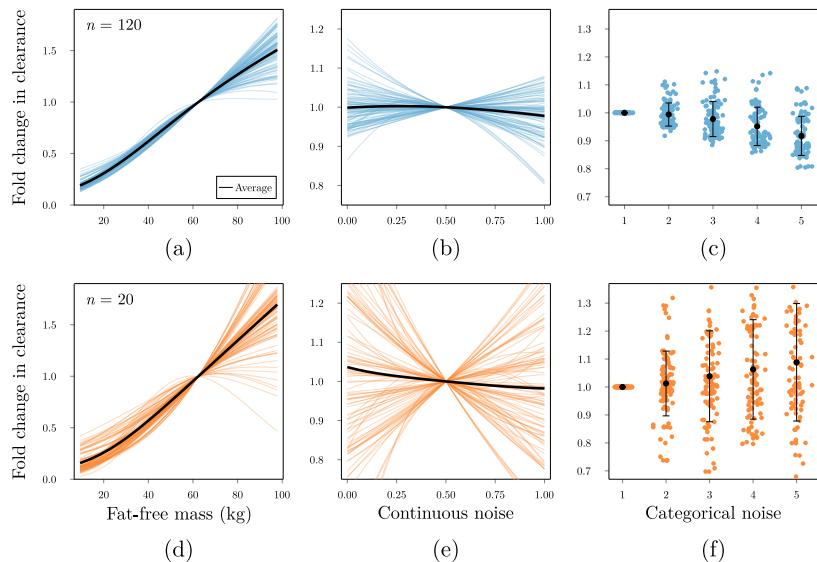


Figure 5.3.2: Visualisation of learned functions from the multi-branch network enable model interpretability. The top panel (a, b, and c) depict the learned functions for the model trained on 120 samples, with the bottom panel (d, e, and f) showing results on $n = 20$. Black curves depict the average effect over the 100 model replicates (coloured lines)

5.3.3 Constrained models perform better on real-world data

In Table 5.3.3, we summarise the results from the models fit to real data. Here, we see that the addition of the effect of blood group O on CL improved accuracy of the expert models. Addition of the covariate resulted in a statistically significant decrease in objective function value by more than 20 points in both models ($p < 0.01$; $\chi^2 = 6.635$).

The model by [31] was more accurate on our data set than the model from [22] (RMSE 13.7 vs 14.8 IU/dL). The addition of model constraints improved accuracy compared to the naive DCM, and the use of boundary constraints and global parameters resulted in models with relatively similar performance to the best performing NLME model. The multi-branch network achieved the lowest RMSE overall (13.0 IU/dL) while the Neural-ODE based model achieved the highest RMSE (19.5 IU/dL). Similar to Fig. 5.3.1, the naive model predicted higher FVIII peak levels, followed by an initial rapid re-distribution of FVIII (see Supplementary Fig. 5.A.2a). The addition of model constraints resulted in more smooth concentration time curves. Concentration time-curves produced by the Neural-ODE were unrealistic and suggestive of model overfitting (Supplementary Figs. 5.A.2e and 5.A.3). Again, learned covariate effects in the multi-branch network were visualised, enabling model interpretation (see Supplementary Fig. 5.A.4).

MODEL	MEAN RMSE (IU / dL) \pm ONE SD
EXPERT MODELS	
<i>Björkman et al.</i> [22]	15.8 ± 3.3
<i>Björkman et al.</i> [22] + BGO	14.8 ± 3.1
<i>McEneny-King et al.</i> [31]	14.7 ± 2.9
<i>McEneny-King et al.</i> [31] + BGO	13.7 ± 3.0
MACHINE LEARNING MODELS	
Neural-ODE (32 neurons, 1 hidden layer, 2 latent variables) [13]	19.5 ± 3.5
Fully-connected DCM (32 neurons)	14.8 ± 2.8
DCM + boundary (32 neurons)	14.1 ± 2.4
DCM + global parameters (32 neurons)	13.9 ± 1.8
Multi-branch network (16 neurons)	13.0 ± 2.1

Abbreviations: RMSE = root mean squared error, SD = standard deviation, BGO = blood group O, DCM = deep compartment model.

Table 5.3.3: Comparison of model accuracy on real-world data.

5.4 DISCUSSION

In this work, we investigated how model constraints affected the predictive performance of deep compartment models [14]. Without any constraints, models potentially learn unrealistic concentration-time

curves when data is sparse. Although these models accurately predicted observed concentration measurements (i.e. had low training loss), they were not penalised for making extreme predictions at time points outside the training data. The results from the first experiment indicated that roughly one-fifth of fitted models resulted in divergent results when data was sparse ($n = 20$). This limits the clinical implementation of such algorithms. Our results indicate that the introduction of simple constraints improved model robustness as represented by the number of divergent models. As a consequence, the constraints could improve model accuracy when trained on smaller data sets and resulted in more realistic concentration–time curves. The constraints can be based on prior knowledge, making them easier to implement in practice. Finally, the proposed multi-branch network architecture is an interpretable alternative to fully-connected networks, trading ease-of-implementation for increased model trust.

In the second synthetic experiment we found that the presence of false covariates affected model accuracy. Visualisations of the learned functions in the multi-branch network indicated that models were sensitive to learning false effects irrespective of the size of the training set. However, models trained on sparse data were more likely to inflate the importance of the false covariates, resulting in higher error on new data. This is indicative of the importance of the careful selection of (causal) covariates to include in these models. One approach for covariate selection can for example be the use of cross-validation based procedures to identify covariates that can be removed based on the uncertainty/absence of their effect across replicates. Similarly, this approach can be used to perform an initial screening of the covariates for downstream model analysis. This approach can both identify covariate importance as well as their relationship to the PK parameters. Comparing learned functions from multiple replicates also allows for the identification of regions of covariate space that have higher data uncertainty. For example, <2% of patients had a FFM in the data set, which is reflected by higher uncertainty of the effect of FFM on CL (Fig. 5.3.2a). For patients in these regions one can decide to first collect more data before making predictions. Knowing when to trust model predictions is important, especially in the context of medical decision-making.

The results of the real data experiment support the findings of the synthetic experiments. The addition of model constraints improved model performance in terms of test set accuracy. Importantly, the shape of the resulting concentration–time curves were again more

realistic compared to those from unconstrained models. By understanding how inductive biases are encoded in conventional methods used for PK analysis, we show that hybrid architectures can be a promising approach for improving model performance in settings with limited data. Fully ML-based architectures, such as the Neural-ODE, greatly simplify model development but suffer when data is sparse. In addition, diagnosing and resolving overfitting issues in these models is more complicated. We show that hybrid architectures can alleviate these issues, and can be designed in such a way that the model is inherently interpretable. This eliminates the need for (post-hoc) ML explainability methods such as SHAP, which do not necessarily offer a true representation of model predictions [33]. Using Neural-ODEs for learning parts of the mechanistic model can also be an interesting hybrid approach [18]. In the multi-branch network, covariates are organised into sub-models, allowing for the visualisation of learned functions. Such an approach can improve model trust while also aiding with the ability to critique the model during development. Compared to classical population PK modelling, this method holds great potential for reducing the complexity of model development especially when paired with the ability to detect and manage overfitting.

There were also some limitations to this study. First of all, in the PK model used to generate the synthetic data, Q and V_2 parameters were fixed for all individuals. This might partly explain the higher accuracy of the models estimating global parameters for these variables. However, due to data sparsity at early time-points and the addition of noise, it is not necessarily clear in what degree this affects the results. Regardless of potential biases during the synthetic experiments, the estimation of global parameters also resulted in more accurate predictions in the experiment using real world data. Next, we found that the estimation of global parameters resulted in higher accuracy compared to the use of boundaries. Inspections of model predictions showed that estimates of Q and V_2 were often stuck in flat regions of the sigmoid during early training (Supplementary Data 5.C). Resulting gradients shrink to zero, making it more difficult for the model to correct for early misspecification. This approach could thus potentially be improved by only placing boundaries on a subset of the PK parameters, by combining it with the estimation of global parameters, or by using less aggressive functions to constrain the parameters (e.g. softsign or cdf of a Cauchy(0, 2), see Supplementary Data 5.C).

A limitation of the proposed multi-branch network is that poor initialisation could still be prone to fitting unrealistic models. Unfortunately, placing additional constraints on this architecture is difficult as it changes model interpretation. For example, setting boundaries on the predicted values of the PK parameters in the final layer of the network breaks the interpretation of the learned functions. Another downside is that learned effects can only be visualised when the number of covariates used in each sub-model facilitates 2 or 3-dimensional visualisation. Next, we evaluated only a relatively small number of hyper-parameters settings, i.e. only a single hidden layer with three options for the number of neurons. Extensive searches over appropriate hyper-parameters can be problematic, especially when data is sparse. In the real-world experiment for example, only 12 patients were used to find the optimal weights during training as well as the optimal hyper-parameters. When evaluating a large set of hyper-parameters, we risk overfitting the hyper-parameters to the validation set. A promising alternative is to perform hyper-parameter selection based on the desired complexity of the learned functions in the multi-branch network.

In the real world experiment, we compared model performance based on prediction accuracy represented by the RMSE, similar to previous studies [13, 19]. This metric might not be sufficient to fully compare the models. However, common tools for comparing population PK models, such as the Akaike and Bayesian information criterion, are not suitable for use with neural networks as they generally over-estimate model complexity when penalising the number of parameters. Although the current results suggest improvement of models when adding constraints, more research on multiple data sets might be needed to draw conclusions.

Finally, we only evaluated the use of constraints in the context of a drug with relatively simple kinetics. How performance is affected in more complex settings was not within the scope of the current work. It is possible that the selection of appropriate constraints can be difficult in models with an extremely large number of PK parameters. Similarly, setting constraints on parameters with a more complicated interpretation can also be difficult.

Future work could investigate the implementation of more sophisticated inductive biases. It might be of interest to selectively tighten boundaries based on patient covariates. We would for example expect lower distribution volumes for children compared to adults. Other approaches could focus on placing constraints on the learned functions in

the multi-branch network, for example by encouraging monotonicity at unseen values of the covariates. Maximum a posteriori estimation of the neural network weights can also be performed using prior distributions that favour less extreme functions. Alternatively, Gaussian Processes are an interesting alternative to neural networks, as they provide a more practical approach for placing priors over the functional form of the relationships. Additionally, Gaussian Processes allow for a practical method for estimating uncertainty over learned functions. Finally, a method for performing covariate selection using the multi-branch network would be of interest to aid model development.

5.5 CONCLUSION

This work has focused on improving the robustness of the deep compartment model framework. The suggested model constraints can be used to improve the performance of this model class when data is sparse, which is frequently the case in the pharmacometric literature. The proposed hybrid model has many of the benefits of current ML methods used in the pharmacometrics literature, and addresses some of their main limitations. The suggested improvements further demonstrate the method as a viable alternative to classical population PK modelling.

REFERENCES

- [1] Sven Björkman, MyungShin Oh, Gerald Spotts, Phillip Schroth, Sandor Fritsch, Bruce M Ewenstein, Kathleen Casey, Kathelijn Fischer, Victor S Blanchette, and Peter W Collins. "Population pharmacokinetics of recombinant factor VIII: the relationships of pharmacokinetics to age and body weight". In: *Blood, The Journal of the American Society of Hematology* 119.2 (2012), pp. 612–618.
- [2] Nienke AG Lankheet, Lotte M Knapen, Jan HM Schellens, Jos H Beijnen, Neeltje Steeghs, and Alwin DR Huitema. "Plasma concentrations of tyrosine kinase inhibitors imatinib, erlotinib, and sunitinib in routine clinical outpatient cancer care". In: *Therapeutic drug monitoring* 36.3 (2014), pp. 326–334.
- [3] Jason A Roberts, Sanjoy K Paul, Murat Akova, Matteo Bassetti, Jan J De Waele, George Dimopoulos, Kirsi-Maija Kaukonen, Despoina Koulenti, Claude Martin, Philippe Montravers, et al. "DALI: defining antibiotic levels in intensive care unit patients: are current β -lactam antibiotic doses sufficient for critically ill patients?" In: *Clinical infectious diseases* 58.8 (2014), pp. 1072–1083.
- [4] Iris van Moort, Tim Preijers, Laura H Bukkems, Hendrika CAM Hazendonk, Johanna G van der Bom, Britta AP Laros-van Gorkom, Erik AM Beckers, Laurens Nieuwenhuizen, Felix JM van der Meer, Paula Ypma, et al. "Perioperative

- pharmacokinetic-guided factor VIII concentrate dosing in haemophilia (OPTI-CLOT trial): an open-label, multicentre, randomised, controlled trial". In: *The Lancet Haematology* 8.7 (2021), e492–e502.
- [5] LB Sheiner and TM Ludden. "Population pharmacokinetics/dynamics". In: *Annual review of pharmacology and toxicology* 32.1 (1992), pp. 185–209.
 - [6] Martin Holz and Alfred Fahr. "Compartment modeling". In: *Advanced Drug Delivery Reviews* 48.2-3 (2001), pp. 249–264.
 - [7] Alexander Janssen, Frank C Bennis, and Ron AA Mathôt. "Adoption of machine learning in pharmacometrics: an overview of recent implementations and their considerations". In: *Pharmaceutics* 14.9 (2022), p. 1814.
 - [8] Lina Keutzer, Hufang You, Ali Farnoud, Joakim Nyberg, Sebastian G Wicha, Gareth Maher-Edwards, Georgios Vlasakakis, Gita Khalili Moghaddam, Elin M Svensson, Michael P Menden, et al. "Machine learning and pharmacometrics for prediction of pharmacokinetic data: differences, similarities and challenges illustrated with rifampicin". In: *Pharmaceutics* 14.8 (2022), p. 1530.
 - [9] Emeric Sibieude, Akash Khandelwal, Jan S Hesthaven, Pascal Girard, and Nadia Terranova. "Fast screening of covariates in population models empowered by machine learning". In: *Journal of pharmacokinetics and pharmacodynamics* 48.4 (2021), pp. 597–609.
 - [10] Ylva Wahlquist, Jesper Sundell, and Kristian Soltesz. "Learning pharmacometric covariate model structures with symbolic regression networks". In: *Journal of Pharmacokinetics and Pharmacodynamics* 51.2 (2024), pp. 155–167.
 - [11] Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. "Universal Differential Equations for Scientific Machine Learning". In: (2021). arXiv: 2001.04385 [cs.LG]. URL: <https://arxiv.org/abs/2001.04385>.
 - [12] Alexander Janssen et al. "Application of SHAP values for inferring the optimal functional form of covariates in pharmacokinetic modeling". In: *CPT: Pharmacometrics & Systems Pharmacology* 11.8 (2022), pp. 1100–1110.
 - [13] James Lu, Kaiwen Deng, Xinyuan Zhang, Gengbo Liu, and Yuanfang Guan. "Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens". In: *Iscience* 24.7 (2021).
 - [14] Alexander Janssen et al. "Deep compartment models: a deep learning approach for the reliable prediction of time-series data in pharmacokinetic modeling". In: *CPT: Pharmacometrics & Systems Pharmacology* 11.7 (2022), pp. 934–945.
 - [15] Zhe He, Rui Zhang, Gayo Diallo, Zhengxing Huang, and Benjamin S Glicksberg. "Explainable artificial intelligence for critical healthcare applications". In: *Frontiers in artificial intelligence* 6 (2023), p. 1282800.
 - [16] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective". In: *BMC medical informatics and decision making* 20 (2020), pp. 1–9.
 - [17] D.S. Bräm, U. Nahum, and Schropp J. "Neural ODEs in pharmacokinetics: concepts and applications". In: *J Pharmacokinet Pharmacodyn* (2023). DOI: 10.21203/rs.3.rs-2428689/v1.

- [18] Dominic Stefan Bräm, Uri Nahum, Johannes Schropp, Marc Pfister, and Gilbert Koch. "Low-dimensional neural ODEs and their application in pharmacokinetics". In: *Journal of Pharmacokinetics and Pharmacodynamics* 51.2 (2024), pp. 123–140.
- [19] Zhaozhi Qian, William Zame, Lucas Fleuren, Paul Elbers, and Mihaela van der Schaar. "Integrating expert ODEs into neural ODEs: pharmacology and disease progression". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 11364–11383.
- [20] Jan Over, JJ Sixma, MH Bruine, MC Trieschnigg, RA Vlooswijk, NH Beeser Visser, BN Bouma, et al. "Survival of ¹²⁵Iodine-labeled Factor VIII in normals and patients with classic hemophilia. Observations on the heterogeneity of human Factor VIII." In: *The Journal of clinical investigation* 62.2 (1978), pp. 223–234.
- [21] Dennis A Noe. "A mathematical model of coagulation factor VIII kinetics". In: *Haemostasis* 26.6 (1996), pp. 289–303.
- [22] Sven Björkman, Maj Carlsson, Erik Berntorp, and Pål Stenberg. "Pharmacokinetics of factor VIII in humans: obtaining clinically relevant data from comparative studies". In: *Clinical pharmacokinetics* 22 (1992), pp. 385–395.
- [23] Marvin J Yiengst and Nathan W Shock. "Blood and plasma volume in adult males". In: *Journal of applied physiology* 17.2 (1962), pp. 195–198.
- [24] Xu Steven Xu, Min Yuan, Hao Zhu, Yaning Yang, Hui Wang, Honghui Zhou, Jinfeng Xu, Liping Zhang, and Jose Pinheiro. "Full covariate modelling approach in population pharmacokinetics: understanding the underlying hypothesis tests and implications of multiplicity". In: *British journal of clinical pharmacology* 84.7 (2018), pp. 1525–1534.
- [25] Centers for Disease Control, Prevention (CDC), and National Center for Health Statistics (NCHS). *National Health and Nutrition Examination Survey Data*. Hyattsville, MD, 2009. doi: <https://www.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2009>.
- [26] Pierre Chelle, Cindy HT Yeung, Stacy E Croteau, Jennifer Lissick, Vinod Balasa, Christina Ashburner, Young Shil Park, Santiago Bonanad, Juan Eduardo Megías-Vericat, Azusa Nagao, et al. "Development and validation of a population-pharmacokinetic model for rurioctacog alfa pegol (Adynovate®): a report on behalf of the WAPPS-Hemo Investigators Ad Hoc Subgroup". In: *Clinical pharmacokinetics* 59 (2020), pp. 245–256.
- [27] Hesham Saleh Al-Sallami, Ailsa Goulding, Andrea Grant, Rachael Taylor, Nicholas Holford, and Stephen Brent Duffull. "Prediction of fat-free mass in children". In: *Clinical pharmacokinetics* 54 (2015), pp. 1169–1178.
- [28] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. "Searching for Activation Functions". In: (2017). arXiv: 1710.05941 [cs.NE]. URL: <https://arxiv.org/abs/1710.05941>.
- [29] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: (2017). arXiv: 1412.6980 [cs.LG]. URL: <https://arxiv.org/abs/1412.6980>.

- [30] Sven Björkman, Anna Folkesson, and Siv Jönsson. "Pharmacokinetics and dose requirements of factor VIII over the age range 3–74 years: a population analysis based on 50 patients with long-term prophylactic treatment for haemophilia A". In: *European journal of clinical pharmacology* 65 (2009), pp. 989–998.
- [31] Alanna McEneny-King, Pierre Chelle, Gary Foster, Arun Keepanasseril, Alfonso Iorio, and Andrea N Edginton. "Development and evaluation of a generic population pharmacokinetic model for standard half-life factor VIII for use in dose individualization". In: *Journal of Pharmacokinetics and Pharmacodynamics* 46 (2019), pp. 411–426.
- [32] J O'donnell and MA Laffan. "The relationship between ABO histo-blood group, factor VIII and von Willebrand factor". In: *Transfusion Medicine* 11.4 (2001), pp. 343–351.
- [33] Indra Kumar, Carlos Scheidegger, Suresh Venkatasubramanian, and Sorelle Friedler. "Shapley Residuals: Quantifying the limits of the Shapley value for explanations". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 26598–26608.

APPENDIX

5.A SUPPLEMENTARY TABLES AND FIGURES

MODEL	MEAN RMSE ± ONE SD (%-AGE DIVERGENT)		
	N = 20	N = 60	N = 120
WEIGHT, HEIGHT, AGE			
None	15.7 ± 32 (17)	12.9 ± 4.5 (4)	12.1 ± 0.39 (o)
Initialisation	16.7 ± 19 (8)	13.0 ± 4.6 (4)	12.2 ± 0.69 (o)
Boundary	15.8 ± 3.0 (2)	12.7 ± 0.81 (o)	12.1 ± 0.78 (o)
Global parameters	14.1 ± 5.7 (1)	12.8 ± 0.43 (o)	12.1 ± 0.27 (o)
FFM, AGE			
None	14.3 ± 49 (18)	12.6 ± 1.6 (3)	12.1 ± 0.32 (o)
Initialisation	15.1 ± 26 (9)	12.6 ± 2.8 (5)	11.9 ± 0.37 (o)
Boundary	14.2 ± 2.1 (2)	12.5 ± 0.34 (o)	11.9 ± 0.3 (o)
Global parameters	13.3 ± 0.87 (o)	12.6 ± 0.33 (o)	12.0 ± 0.28 (o)

Abbreviations: RMSE = root mean squared error, SD = standard deviation.

Table 5.A.1: Results for the models trained using a hidden layer size of 32.

MODEL	MEAN RMSE \pm ONE SD (%-AGE DIVERGENT)		
	N = 20	N = 60	N = 120
WEIGHT, HEIGHT, AGE			
None	16.8 ± 36 (21)	13.0 ± 8.1 (5)	12.1 ± 0.58 (o)
Initialisation	16.5 ± 31 (6)	13.2 ± 9.7 (5)	12.4 ± 0.87 (o)
Boundary	15.8 ± 2.2 (2)	12.8 ± 0.63 (o)	12.2 ± 0.93 (o)
Global parameters	14.1 ± 1.3 (o)	13.0 ± 0.57 (o)	12.2 ± 0.39 (o)
FFM, AGE			
None	15.5 ± 41 (24)	12.6 ± 4.6 (5)	12.1 ± 0.43 (o)
Initialisation	15.4 ± 16 (13)	12.6 ± 9.1 (5)	12.0 ± 0.4 (o)
Boundary	14.7 ± 1.4 (o)	12.6 ± 0.42 (o)	12.0 ± 0.37 (o)
Global parameters	13.5 ± 0.85 (o)	12.5 ± 0.43 (o)	12.0 ± 0.27 (o)

Abbreviations: RMSE = root mean squared error, SD = standard deviation.

Table 5.A.2: Results for the models trained using a hidden layer size of 128.

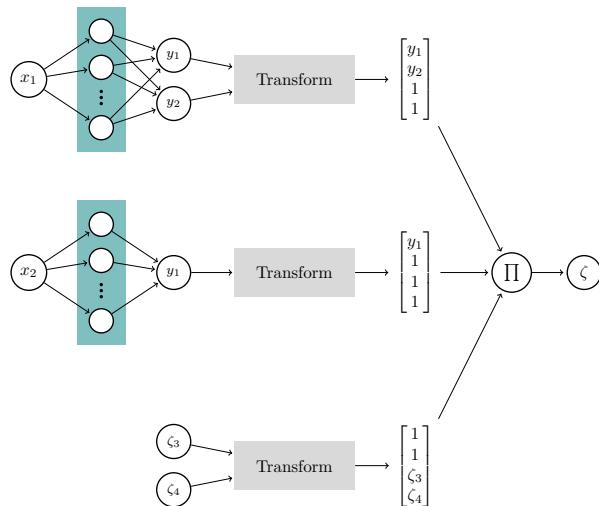


Figure 5.A.1: Schematic representation of the multi-branch network.

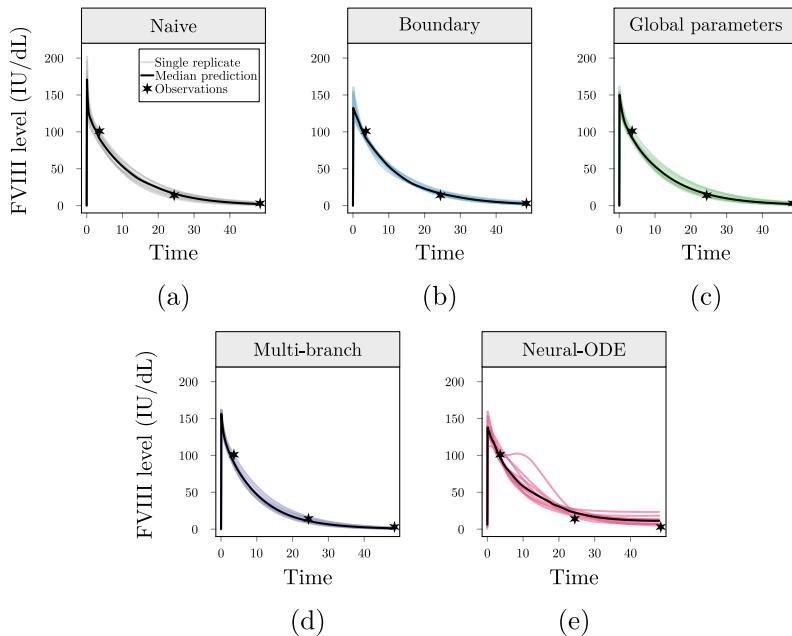


Figure 5.A.2: Comparison of the predicted concentration-time curves in the real-world data experiment. Results are shown for the naive (a), boundary constraint (b), global parameter (c), multi-branch network (d), and Neural-ODE (e) models. The median prediction (black line) over the 10 data set folds (lightly coloured lines) along with the observations (stars) are shown for the same patient.

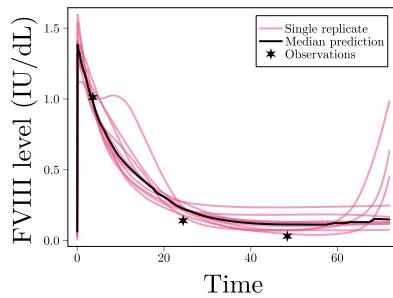


Figure 5.A.3: Extrapolation of predictions in the Neural-ODE quickly degenerate. Showing the prediction for the same patient as in Figure s2, but with an expanded time window. Predictions by the Neural-ODE can behave unexpectedly when data is insufficient to fully describe drug kinetics. The median prediction (black line) over the 10 data set folds (lightly coloured lines) along with the observations (stars) are shown.

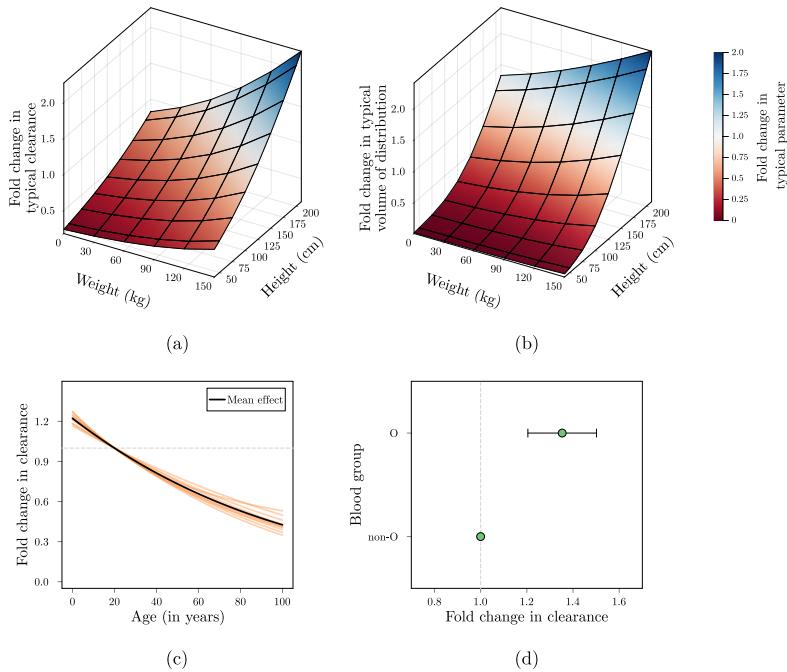


Figure 5.A.4: Learned effects of the multi-branch network in the real-world data experiment. In the top panel, the combined effect of weight and height on clearance and volume of distribution are shown. Horizontal and vertical lines depict the marginal effect of respectively weight or height at a fixed value of the other covariate. In a, we can see that weight and height are similarly important for predicting changes in clearance. However, in b we see that the importance of height is stronger than that of weight. In the bottom panel, the effect of age on clearance (c) and blood group on clearance (d) are shown.

5.B DETAILED DESCRIPTION OF MODEL CONSTRAINTS

We propose three simple approaches for constraining the solution space of DCMs (figure 5.2.1). First, boundary conditions were imposed on the PK parameters by using a transformed sigmoidal function following the output layer of the neural network (referenced as boundary constraint):

$$\zeta^{(i)} = \pi(\phi(\mathbf{x}^{(i)})) \cdot (\mathbf{u} - \mathbf{l}) + \mathbf{l} \quad (5.8)$$

Here, $\pi(\cdot)$ corresponds to the sigmoid function and $\mathbf{l} \leq \zeta^{(i)} \leq \mathbf{u}$ is the constrained PK parameter vector. The boundaries can be set empirically based on prior knowledge. For example, bounds for the volume of distribution of drugs tightly bound to plasma proteins can be based on the expectation that the plasma volume of a typical male is roughly around 46 - 52 mL/kg. We focus on setting a single boundary for all individuals, although it is possible to use a function to adapt \mathbf{l} or \mathbf{u} based on the covariates. Lower bounds of [0, 0.3, 0.05, 0] and upper bounds of [0.5, 7, 0.5, 2] for respectively CL (L/h), V₁ (L), Q (L/h), and V₂ (L) were used.

Next, global parameters θ for a subset of the PK parameters were estimated in parallel to w (referenced as global parameters constraint). We chose to estimate $\theta = \{Q, V_2\}$ since these parameters affect the early distribution of FVIII, and drug concentration measurements at early time points are usually too sparse to identify covariate effects on these parameters. PK parameter vectors were reconstructed in the correct order using design matrices constructed using indicator functions $\mathbf{1}_A$. This function is specified in algorithm 1. An example of $A = \{1, 3\}$ using one-based indexing results in:

$$\mathbf{1}_{\{1,3\}} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \quad (5.9)$$

The PK parameter vector can then be reconstructed using the following equation (continuing the example of $A = \{1, 3\}\}:$

$$\zeta^{(i)} = \mathbf{1}_A \cdot \phi(x^{(i)}; w) + \mathbf{1}_{-\{A\}} \cdot \theta = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \zeta_1 & \zeta_3 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \zeta_2 & \zeta_4 \end{bmatrix} \quad (5.10)$$

Where $-\{A\}$ corresponds to the indexes from $[1..M] \notin A$. Finally, we describe a neural network architecture where each covariate (or specific combinations thereof) are connected to independent sub-models ψ , whose predictions are combined using a product (referenced as the multi-branch network). This architecture is similar to a generalised additive model, using product accumulation rather than the sum of covariate effects. The use of a product ensures that $\zeta^{(i)}$ remains positive, regardless of the prediction from each sub-model as long as these are constrained to be positive. The use of a product also matches the standard implementation of covariates in population PK models (equation 3 in the main manuscript), and facilitates the interpretation of the clinical relevance of each covariate. For example, covariates resulting in a maximal net change 20% of the corresponding PK parameter are often deemed clinically insignificant in the pharmacometrics literature [16]. Again, indicator functions are used to determine the position of each prediction in the PK parameter vector:

$$\zeta^{(i)} = \prod_{d=1}^D \mathbf{1}_0[\mathbf{1}_{A_d} \cdot \psi_d(x_d^{(i)})] \quad (5.11)$$

Here, $\mathbf{1}_0$ is an indicator function mapping zeros to ones. Softplus activation functions were used in the output layer of each sub-model. Single covariates can also be linked to multiple PK parameters. Likewise, multiple covariates can be passed to a sub-model when an interaction between covariates is expected. An added benefit of this approach is that the output of each sub-model can be visualised, allowing for the interpretation of the learned covariate effects. A schematic overview of the multi-branch network is provided in supplementary figure 5.

5.C DETAILED INVESTIGATION OF EFFECTS OF CONSTRAINTS ON MODEL TRAINING

In this section, we further investigate the effect of each constraint on model training. First, we found that there was roughly 4.5x greater

Algorithm 1 Pseudo-code describing the indicator function.

```

1:  $M \leftarrow \text{length}(\zeta)$ 
2:  $A \leftarrow a \subset [1..M]$ ; subset of indexes to set to 1
3:  $N \leftarrow \text{length}(A)$ 
4:  $Z \leftarrow \text{zeros}(M, N)$ 
5: for  $i$  in  $\text{eachindex}(A)$  do
6:    $Z[i, A[i]] = 1$ 
7: end
8: return  $Z$ 

```

variability in median RMSE over replicates over data folds compared to within fold replicates. This suggest that weight initialisation have a relatively small effect on final model accuracy. Looking at the number of divergent models when fitting the naive model, the frequency of divergent models seemed to be related to specific data folds:

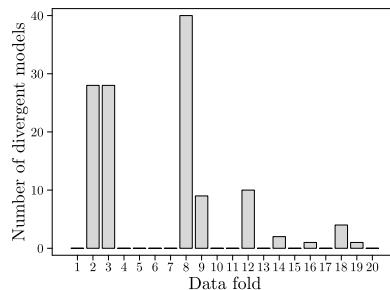


Figure 5.C.1: Divergent models per training fold.

In 21/36 of examples where divergent models occurred, 4 or 5 of the within fold replicates resulted in divergent models. Most divergent models were trained on data from fold 2, 3, and 8. We therefore look specifically into the distribution of the covariates as well as the initial PK parameter estimates for the models trained on these folds. We did not observe very large differences between covariate distributions, with possibly a slightly higher fraction of younger patients in data folds 2, 3 and 8. This might cause the model to lean slightly more towards lower volume of distributions (as children have lower volume of distribution), resulting in sharper peaks. There do not seem to be very distinct differences between the drug levels between ‘bad’ and ‘good’ data folds:

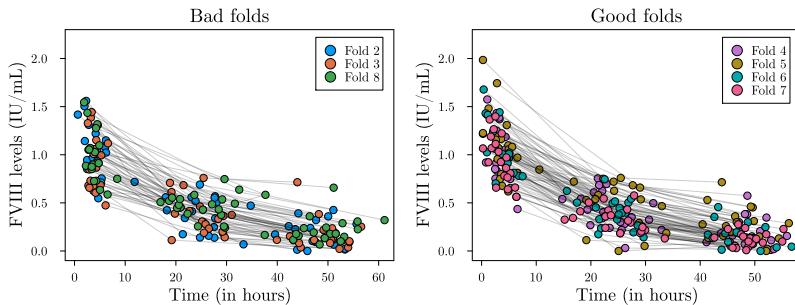


Figure 5.C.2: FVIII measurements per fold.

Random initialisation of the naive neural network weights results in initial estimates of the PK parameters around 0.7 (L or L/h):

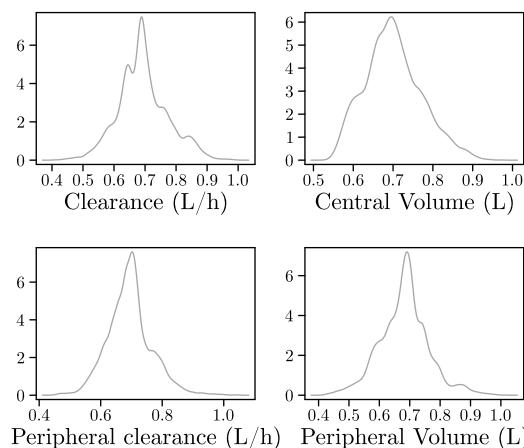


Figure 5.C.3: Distribution of PK parameters after initialisation.

This image shows 100 random initialisation of the network weights and the distribution of resulting initial PK parameter estimates for subjects in training fold 2 with $n = 120$. The distributions are data agnostic, they initialise around 0.7 no matter the input data. Due to the relatively low initial estimate for V_1 and high estimate for CL , concentration time curves start out with high peak concentrations and short half-life. Next we look at how the PK parameters change during optimisation when training on $n = 120$ and $n = 20$ on one of the problematic folds (and compare to $n = 20$ on a good fold):

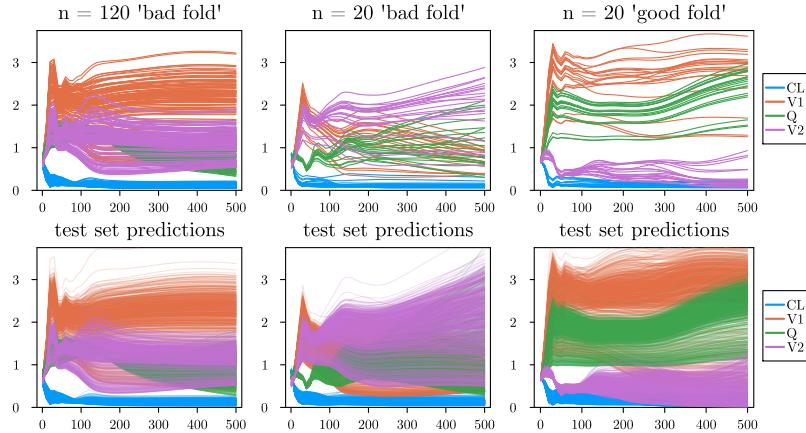


Figure 5.C.4: PK parameters during training.

We see that for one of the problematic folds (fold 2) the estimate of volume of distribution (V_1) drops down towards zero after an initial increase for all subjects at $n = 20$, while V_2 increases rapidly. This results in high peak concentration predictions due to a rapid distribution from a small central volume into a larger peripheral volume. This is not seen when training on larger patient data sets, or on ‘good’ data folds. Removing the younger patients from the data fold is not sufficient to improve optimisation:

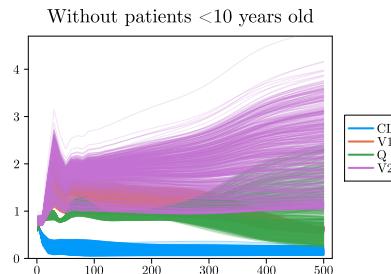


Figure 5.C.5: PK parameters during training without children.

We then look at the effect of adding constraints to the model:

Above we show the PK parameter predictions after training on $n = 20$ subjects for a ‘bad’ fold, and show predictions in the test set. Similar to initialisation, the addition of bounds to the value of the PK parameters changes the initial estimates of the PK parameters. The

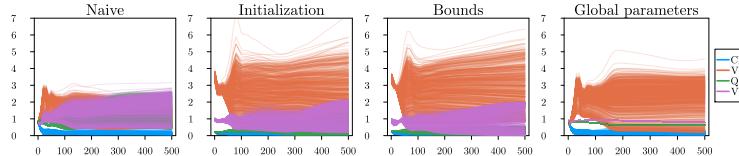


Figure 5.C.6: PK parameters when adding constraints.

effect of initialisation and providing bounds seem to be somewhat similar, with a notable exception that placing bounds causes the estimates of V_2 to be mainly located at the extremes of the bound (i.e. 0 and $2 L$). Additionally, the variability in V_1 is larger compared to the other models. It is possible that placing bounds seems to have a similar effect as initialisation. In the global parameter model, PK estimates during the first 50 epochs of training look somewhat similar to those obtained using the naive model:

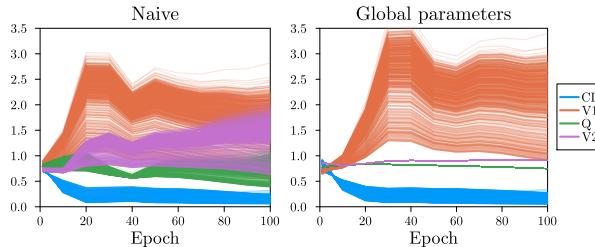


Figure 5.C.7: Comparison of first 100 epochs.

An important distinction is that in the naive model, the estimate for V_2 rises concurrently with the value of V_1 , whereas in the global parameter model, V_2 remains somewhat stable during optimisation. It seems that setting global variables regularises the optimisation procedure in such a way that gradients of these parameters are potentially smaller compared to the other parameters. We can show that specifically using a global variable for V_2 has a similar effect as using global parameters for both parameters. Using a global parameter for Q still results in unrealistic solutions.

If we set the initial estimates of V_2 to be very high ($3 L$) or low ($0.1 L$) the model still results in reasonable models when using global parameters for V_2 :

Optimisation can also be improved by only setting bounds on the value of V_2 :

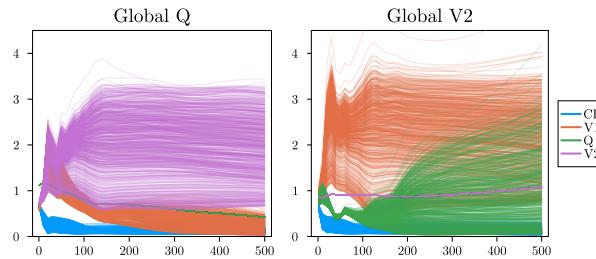


Figure 5.C.8: Using global variables for Q or V₂.

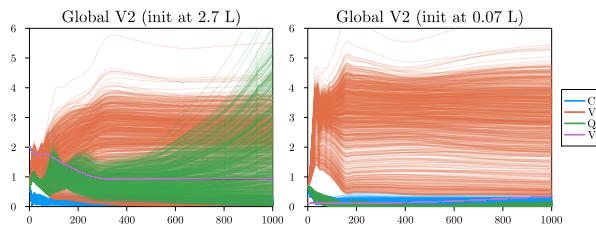


Figure 5.C.9: Effect of parameter initialisation.

We find here that placing softer bounds (for example by using a softsign instead of a sigmoid) might be a reasonable approach, since we observe a propensity of V₂ estimates ‘getting stuck’ at the extremes of the sigmoid during optimisation. We ran the synthetic experiment again for these two V₂ specific constraints on the data sets with $n = 20$ subjects, using 8 neurons in the hidden layer of the neural network:

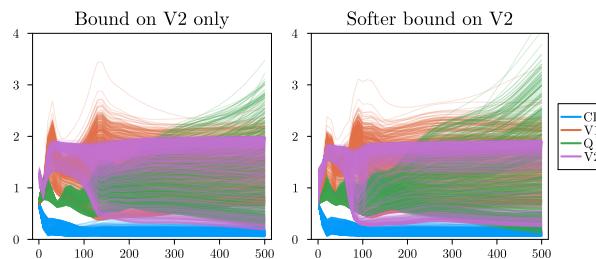


Figure 5.C.10: Setting bounds on V₂.

MODEL	MEDIAN RMSE ± ONE SD (%-AGE DIVERGENT)	
	WEIGHT + HEIGHT + AGE	FFM + AGE
Naive (from original experiment)	14.7 ± 0.42 (18%)	14.1 ± 0.48 (12%)
Softsign bound on V2	17.4 ± 3.5 (5%)	17.5 ± 21.8 (5%)
Global parameter for V2	16.8 ± 2.1 (2%)	16.4 ± 41.1 (1%)

Abbreviations: RMSE = root mean squared error, SD = standard deviation, FFM = fat-free mass.

Table 5.C.1: Model accuracy when using specific constraints.

5.C.1 Conclusion

The results hint at the importance of the data used for training (especially when the number of samples is sparse), since it can bias the optimisation procedure to converge to unrealistic solutions. Unfortunately, we could not find specific data that causes the behaviour. Although initialisation and the use of bounds seem to improve optimisation, there is still a large variability in the PK parameters after convergence. Global parameters on the other hand seem to result in distinct solutions, which we found (in this case) to be more similar to models trained on larger data sets. We can further identify the behaviour to be specifically related to the estimate of V2. Setting this PK parameter to be global, or setting bounds on it specifically can improve solutions. It could be the case that the models have identifiability issues between V1 and V2 during optimisation, such that additional constraints can potentially be useful to improve the model.

5.D VISUALISATION OF COVARIATE EFFECTS

Visualisation of the learned effects for each of the covariates were obtained from the multi-branch networks by taking each sub-model and entering dummy input within the domain of the training data. In order to compare learned effects across model replicates, the output of each sub-model was normalised with respect to its prediction at the location of the median covariate value:

$$\psi_s^*(x_s) = \frac{\psi_s(x_s)}{\psi_s(\text{Med}[x_s])} \quad (5.12)$$

The prediction of $\zeta^{(i)}$ now decomposes into:

$$\zeta_m = \theta_{TV} \cdot \prod_s^{S_m} \psi_s(x_s) \quad (5.13)$$

Where $\theta_{TV} = \prod_s^{S_m} \psi_s(\text{Med}[x_s])$ is the typical value for PK parameter ζ_m over all individuals. Note the similarity of this equation to equation 3 in the main manuscript. This way the prediction from each neural network is anchored to 1 at the median value of each of the covariates, similar to how covariates are implemented in NONMEM. Since covariate effects are combined using a product, the full model is unconstrained with respect to the scale of the predictions from each sub-model and variance of the unnormalised learned effects is high. For example, if the prediction of ψ_1 is very low after random initialisation of the network, the model can still produce the same PK parameter predictions to other replicates by increasing the scale of predictions from ψ_2 . The normalisation corrects for these differences between replicates.

Now that we have $\psi_s^*(x_s)$, we can query this model to obtain predictions at any value of the covariate. Visualisation of these predictions results in the figures as reported in the manuscript.

6

MIXED EFFECT ESTIMATION IN DEEP COMPARTMENT MODELS: VARIATIONAL METHODS OUTPERFORM FIRST-ORDER APPROXIMATIONS

Alexander Janssen, Frank C. Bennis, Marjon H. Cnossen, and Ron A.A. Mathôt

Journal of Pharmacokinetics and Pharmacodynamics (2024): 1-12.

ABSTRACT

This work focuses on extending the deep compartment model (DCM) framework to the estimation of mixed-effects. By introducing random effects, model predictions can be personalised based on drug measurements, enabling the testing of different treatment schedules on an individual basis. The performance of classical first-order (FO and FOCE) and machine learning based variational inference (VI) algorithms were compared in a simulation study. In VI, posterior distributions of the random variables are approximated using variational distributions whose parameters can be directly optimised. We found that variational approximations estimated using the path derivative gradient estimator version of VI were highly accurate. Models fit on the simulated data set using the FO and VI objective functions gave similar results, with accurate predictions of both the population parameters and covariate effects. Contrastingly, models fit using FOCE depicted erratic behaviour during optimisation, and resulting parameter estimates were inaccurate. Finally, we compared the performance of the methods on two real-world data sets of haemophilia A patients who received standard half-life factor VIII concentrates during prophylactic and perioperative settings. Again, models fit using FO and VI depicted similar results, although some models fit using FO presented divergent results. Again, models fit using FOCE were unstable. In conclusion, we show that mixed-effects estimation using the DCM is feasible. VI performs conditional estimation, which might lead to more accurate results in more complex models compared to the FO method.

6.1 INTRODUCTION

Non-linear mixed effect (NLME) models serve as the established methodology for the analysis of time-series data within the domain of pharmacometrics. These models allow for the simultaneous estimation of population and individual level effects using (semi-)mechanistic models, and are particularly useful for disentangling different sources of variability from data. The inclusion of random variables η imposes a distribution over the model parameters and can be thought of as representing the effect of unseen covariates. At prediction-time, an individual estimate of the parameters can be obtained based on the observations. Aside from improving prediction accuracy, these individual estimates can also be used to simulate drug exposure or effects based on unseen treatment strategies, facilitating the selection of optimal treatment on a personalised basis.

Recently, the field of pharmacometrics has seen an influx of interest in the use of machine learning (ML) methods [1–3]. Most ML techniques favour data-driven learning of relationships between covariates and observations based on large amounts of data. However, the availability of large data sets is often a limiting factor within the context of pharmacometrics, rendering most standard ML methods ineffective. Moreover, algorithms such as neural networks and tree-based methods require the utilisation of drug dose as model input, which has been shown to be problematic for reliable extrapolation to unseen data [1, 4]. Combining prior knowledge with machine learning methods in so-called hybrid model architectures poses a promising alternative, potentially improving both data efficiency and predictive performance.

One such architecture is the deep compartment model (DCM), which uses neural networks to learn the relationship between covariates and the parameters of a system of differential equations representing the (semi-)mechanistic model [5]. This architecture is highly flexible: it supports all problems involving ODEs, can learn the effects of specific covariates only (using explicit equations for others), or can be used to learn the partial differential equations describing drug kinetics/dynamics or parts thereof using Neural-ODEs [4, 6, 7]. In its current form, the framework focuses on the estimation of fixed effects. As these models use highly flexible neural networks, failing to assign part of the variability to random effects can potentially result in the model internalising noise. Another downside is that model

predictions cannot be individualised, limiting its potential for use in clinical practice.

In the work by *Lu et al.*, a variational auto-encoder (VAE) [8] is used to produce individual prior distributions over the Neural-ODE parameters, enabling the personalisation of predictions [4]. In VAEs, neural networks are used to estimate parameters (e.g. mean and variances) for a set of random variables describing the Neural-ODE parameters. Optimisation is simplified by amortisation of the learning procedure [8, 9], often optimising the mean squared error of predictions combined with a regularising term restricting complexity of the latent variables (e.g. using hyper-priors such as a standard Normal). However, this approach breaks the typical assumption that random effects are independent of the covariates, and in practice often results in the variance of (part of) the latent variables shrinking to zero to benefit prediction accuracy [10, 11]. To circumvent these issues, estimation of random effects should be decoupled from the estimation of fixed-effects as is the case in classical NLME models.

The aim of this work is to formulate a robust approach to jointly estimate fixed and random effects within the DCM framework. We investigate the performance of classical first-order approximation methods used in NLME models as well as machine learning based variational methods [12]. The accuracy and stability of these different algorithms are tested on a simulated data set using a population pharmacokinetic (PK) approach. Finally, we showcase the use of the mixed-effect DCM on two real world data sets of haemophilia A patients receiving standard half-life (SHL) factor VIII (FVIII) concentrates during prophylaxis and surgery.

6.1.1 *Estimation of random variables*

Given a data set of covariates \mathbf{X} , interventions \mathbf{I} (e.g. drug administration), and measurements \mathbf{y} for each subject $i \in \{1, \dots, n\}$, we typically use an ODE-based model $A(t)$ to represent the evolution of y_i over time:

$$y_i(t) = A(t; \zeta_i, \mathbf{I}_i) + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \Sigma) \quad (6.1)$$

Here, matrix \mathbf{I}_i contains individual treatment information with corresponding time points and $\zeta_i = f(x_i; \theta)$ are typical ODE parameters (e.g. PK parameters) whose relationship to the covariates \mathbf{X} are described by a set of functions f with fixed effect parameters θ . Mixed

effects models introduce a subject-specific random variable $\eta_i \in \mathbb{R}^K$ on (part of) the parameters of the ODE in order to account for additional heterogeneity between subjects:

$$z_i = g(\zeta_i, \eta_i), \text{ where } \eta_i \sim \mathcal{N}(0, \Omega) \quad (6.2)$$

Here, z_i represents the individual estimate of the ODE parameters and Ω is a $K \times K$ covariance matrix. We drop the subscript i in subsequent equations to reduce cluttering. Following from the Bayes rule $p(\eta | y) = p(y | \eta)p(\eta)/p(y)$, we can obtain maximum a posteriori (MAP) estimates of η based on the measurements y by maximising the joint likelihood $p(y, \eta) = p(y | \eta)p(\eta)$. However, obtaining maximum likelihood estimates of the fixed effect parameters is more complicated. One way is to marginalise out the random variables, which results in a complex integral often lacking a closed-form solution:

$$p(y; \Theta) = \int p(y, \eta; \Theta) d\eta, \text{ where } \Theta = \{\theta, \Omega, \Sigma\} \quad (6.3)$$

Classical methods approximate this integral using a Laplace approximation around the mode of the random effects and linearise the model by performing a first-order Taylor expansion. This results in a Gaussian approximation of the random effect posterior, and is known as the First-Order Conditional Estimation (FOCE) extended least squares objective function (see appendix 6.B for derivation) [13, 14]. When using the FOCE objective, the model iterates through producing MAP estimates of η followed by optimisation of Θ based on the linearised model. Further approximation of the FOCE objective results in the FO objective function, where the mode of η is fixed at the population mean (i.e. zero), removing the need for the calculation of MAP estimates (see appendix 6.B)[15]. However, individual random effects are rarely located at zero (unless shrinkage is high) and the resulting objective function is less accurate. In practice, the FO method is only appropriate when the inter-individual variances are small [16].

6.1.2 Variational Inference

Model performance likely depends on the accuracy of the approximation. The Laplace approximation (and the FO and FOCE by extension) suffers especially when η posteriors are non-Gaussian, or have multiple modes. Alternatively, we can apply Markov Chain Monte Carlo (MCMC) methods to obtain samples of model parameters that converge to their true posterior distributions. Unfortunately, MCMC

quickly becomes computationally prohibitive when the number of subjects and dimension of the random variables increases. This is especially the case when the fixed effects model is a neural network with ill-defined posterior distributions over its weights [17]. Fortunately, several approximate methods for Bayesian inference have been developed to reduce computational complexity.

A notable example is Variational Inference (VI), where the true posterior is approximated by a (simpler) variational distribution q [12]. The variational approximation is optimised by minimising its Kullback–Leibler (KL) divergence with respect to the true posterior. Since the true posterior is unknown, the evidence lower bound (ELBO) is maximised instead, which places a lower bound on the marginal likelihood $p(y)$ (see appendix 6.B):

$$\log p(y) = \underbrace{\mathbb{E}_{q_\phi(\eta)} [\log p(y, \eta) - \log q_\phi(\eta)]}_{\text{ELBO}} + \underbrace{\text{KL}(q_\phi(\eta) \| p(\eta | y))}_{\text{divergence}} \quad (6.4)$$

Here, q_ϕ is a tractable distribution parameterised by ϕ (e.g. $\phi = \{\mu, \sigma\}$ in the case of a Normal distribution). Since $p(y)$ is a constant, maximising the ELBO implicitly minimises the KL divergence. An unbiased estimate of the expectation in Eq. 6.4 can be obtained using Monte Carlo methods, but the resulting gradients have high variance. Roeder *et al.* describe the path-derivative gradient estimator of the ELBO, which has the property that the gradient variance shrinks to zero as $q_\phi(\eta)$ approaches $p(\eta | y)$ [18]. This means that a potentially very close approximation of the true posterior can be obtained based on the chosen complexity of q_ϕ . Choosing a Gaussian approximation will result in a similar approximation of the integral in Eq. 6.3 as with FOCE, albeit a stochastic one due to the Monte Carlo approximation in Eq. 6.4.

It is of interest to compare VI to the classical first-order approximations when using the DCM framework to see if there are differences in performance. Since VI performs conditional estimation, we expect improved performance over the FO method in more complex models. A potential benefit of VI over FOCE might be reduced computational time as MAP optimisation over η is not required. It is also unknown how well these models will behave when simultaneously learning fixed and random effect parameters when covariate effects are learned during the optimisation, as is the case in the DCM.

6.2 METHODS

6.2.1 Synthetic data generation

A total of 500 samples of patient age, height, weight, blood group, and von Willebrand factor antigen (VWF:Ag) levels were simulated from a recently proposed generative model for haemophilia A patients [19]. This generative model implements non-linear relationships to represent the joint distribution over these covariates. Covariate relationships were based on a directed acyclic graph (DAG) representing the causal effects of the covariates. The resulting samples are more realistic than samples from multivariate normal or marginal distributions. After generating synthetic covariate data, factor VIII levels were simulated based on a hypothetical population PK model implementing the following covariate effects:

$$CL = 0.1 \cdot \frac{\text{weight}^{0.75}}{70} \cdot \left(\frac{\text{leaky_softplus}(-\text{VWF} + 100)}{55} + 0.9 \right) \cdot \exp(\eta_1)$$

$$V_1 = 2.0 \cdot \frac{\text{weight}}{70} \cdot \exp(\eta_2)$$

$$Q = 0.15$$

$$V_2 = 0.75$$

$$\text{where leaky_softplus}(x, \alpha = \frac{1}{20}, \beta = \frac{1}{10}) = \alpha \cdot x + (1 - \alpha) \cdot \frac{\log(\exp(x \cdot \beta) + 1)}{\beta}$$

Each virtual patient was given a single dose of 25 IU/kg rounded to the nearest 250 IU. Random samples $\eta \sim \mathcal{N}(0, \Omega)$ with $\Omega = \begin{bmatrix} 0.037 & 0.0113 \\ 0.0113 & 0.017 \end{bmatrix}$ were drawn to produce individual estimates of the PK parameters. Next, simulated FVIII concentration–time curves were generated based on a two compartment model. FVIII measurements were collected at 4, 24, and 48 h after dose.

6.2.2 Evaluating the accuracy of variational approximations

The accuracy of variational posterior approximations was determined by comparing learned random effect posteriors obtained from VI to those obtained from MCMC sampling when using the true model from the simulation. Posteriors were compared in two settings: (1) using the true typical PK and population parameters (i.e. Ω and Σ), and (2) when only using the true typical PK parameters (also approximating the posterior over Ω and Σ). Covariance matrices M were decomposed

in terms of marginal standard deviations S and correlation matrix C such that $M = S \cdot C \cdot S'$. More information on prior and hyper-prior selection for the MCMC model can be found in appendix 6.C.5.

For the MCMC model in scenario 1, a single chain was run to generate 10000 posterior samples using the NUTS algorithm. In scenario 2, 5000 samples were taken. Models were fit to the first data fold of the simulated data set, and 20 replicates of the VI algorithm were fit to compare to results from MCMC. The same prior distributions were used in the VI model. Posterior similarity was determined based on visualisations and quantified using the Wasserstein distance. The ADAM optimiser using a learning rate of 0.1 was used.

6.2.3 Comparison of methods for estimating random variables

Given our computational budget, we decided on fitting 100 models for each of the methods. The complete data set was divided into 20 random subsets of 60 subjects drawn with replacement for model training with the remaining samples for determining model accuracy. Previous results indicated that data from 60 subjects was sufficient to fit accurate models [5, 20]. On each data fold, five replicates of model training were performed which we deemed to be a minimal requirement to represent variability induced by random initialisation of model parameters. We chose to run a larger number of training replicates over data folds rather than within a single data fold (i.e. 20 vs. 5) as we assumed that the specific training data had a larger effect on parameter variability compared to random initialisation following previous findings [20].

A multi-branch network based architecture of the DCM [20] was fit to each training fold of the simulated data set. In a multi-branch network, covariates are linked to specific ODE parameters such that each covariate effect is learnt in isolation. This contrasts standard fully-connected networks where all covariates are linked to all ODE parameters, potentially making the model susceptible to learning spurious covariate effects. In addition, the approach enables the direct visualisation of learned functions for each of the covariates, making the model inherently interpretable without the need for post-hoc ML explanation methods. Subject weight and VWF:Ag were used as covariates. Global parameters were estimated for Q and V2. In the multi-branch network, weight was connected to CL and V1, and VWF:Ag was connected to CL. The same model was optimised using each of the objective functions. For each training replicate, random

initial parameters were drawn from initial distributions. More information on model architecture and initial parameter settings can be found in appendix 6.C.

Again, covariance matrices M were decomposed in marginal standard deviations and correlation matrices. All variance estimates were constrained to be positive using the softplus function. Models were compared based on the root mean squared error (RMSE) of typical predictions, accuracy of the estimated population parameters (represented by the KL divergence of Ω and mean absolute error (MAE) of σ), and the similarity of the learned functions with respect to the true covariate effects. Models were fit based on the MSE (no estimation of population parameters), FO, FOCE, and VI objective functions. When using the VI objective, random effect posteriors were approximated using full-rank multivariate normal distributions. The expectation in the ELBO was approximated using Monte Carlo simulation, taking three random samples and using the reparameterisation trick [8] to generate samples from q . For the models trained using FOCE, MAP estimates of the random effects were obtained by minimisation of the negative joint likelihood for each subject using the BFGS method at the start of each epoch of training. Estimates were constrained between [-3, 3] to improve stability during optimisation.

Models were trained for 2000 epochs and parameters were saved every 25 epochs to determine model convergence and stability during training. Most models converged within 250 – 500 epochs, so additional training iterations allowed insights into parameter stability after convergence and risks of overfitting when overextending training time. The ADAM optimiser using a learning rate of 0.1 or 0.01 was used depending on training stability. Results at the end of optimisation were compared based on the mean of saved parameter estimates from the last 500 epochs of training. Uncertainty estimates over model parameters were obtained by taking the standard deviation of final parameter estimates for each of the training replicates. An overview of the approach is shown in Fig. 6.2.1.

6.2.4 *Evaluation on real world data*

The performance of the algorithms was also evaluated on two real world data sets of haemophilia A patients receiving SHL FVIII concentrates during prophylaxis (data set one) and following surgery (data set two). The data originates from the OPTI-CLOT clinical trial [21], where FVIII consumption was compared between standard weight-

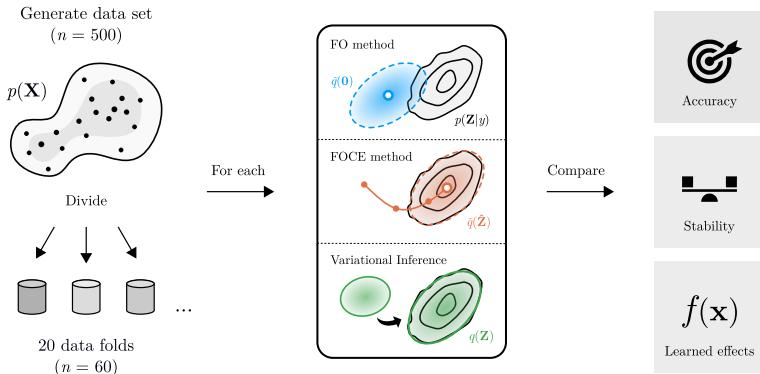


Figure 6.2.1: Comparison of the different methods in the simulation study. First, a data set was simulated containing 500 virtual subjects based on a previously published generative model $p(\mathbf{X})$. The data set was divided in 20 random data subsets with replacement to create the training ($n = 60$) and testing ($n \approx 440$) data sets. On each data fold, models were fit using based on the different methods (FO, FOCE, and VI). In the FOCE method, a Gaussian approximation \tilde{q} of the random effect posterior $p(\mathbf{Z} | \mathbf{y})$ centred at its maximum a posteriori estimate (white circle) is obtained. In the FO method, the mode is fixed at zero, resulting in lower accuracy due to a potential mismatch with the true posterior. In VI, the divergence between a variational approximation $q(\mathbf{Z})$ and the true posterior is minimised. After fitting the models, the methods were compared based on the accuracy of parameter estimates, their stability during training, and the similarity of learned covariate effects to true effects.

based dosing regimens and PK-guided dosing in moderate and severe haemophilia A patients undergoing surgery. The first data set contains a total of 69 subjects who received a PK profile following a 25–50 IU/kg test dose of one of five SHL FVIII concentrates. Three FVIII measurements were collected roughly 4, 24, and 48 h after administration. Available covariates were haemophilia severity, body weight, height, age, and VWF:Ag levels. A large proportion of VWF:Ag levels were missing (65.2%), with some subjects missing body weight or height data (1.4% and 4.3%, respectively). Missing values were imputed based on the mode of prior distributions produced by the generative model (i.e. the same model used for generation of the synthetic data) [19].

The second data set contained data on 66 subjects from data set one who underwent a minor or moderate risk surgical procedure within 12 months after their PK assessment. FVIII levels were measured before and after surgery and FVIII peak and trough levels were collected during follow-up. Compared to the first data set, follow-up time was longer (median of 144 vs 44 h) and subjects received a more complex combination of bolus doses and continuous infusions. Available covariates were haemophilia severity, body weight, height, VWF:Ag and VWF activity (VWF:act) levels, pre-assessed surgical risk scores, blood loss, and NaCL administration during surgery. In this data set, most subjects had multiple VWF measurements. Missing VWF:Ag values were imputed based on the mode of the prior distributions from the generative model multiplied by a factor of 1.3 (VWF:Ag levels are higher following surgery [22]). This factor was calculated from the mean difference between imputed VWF levels in data set one and average VWF levels per subject in data set two. The mean VWF:Ag value was used for each individual.

We fitted a multi-branch DCM with either an additive or combined residual error model to both data sets. Subject CL and V1 was predicted based on fat-free mass (FFM) calculated from body weight, BMI, and age using Al Sallami's equation [23], with an additional effect of VWF:Ag on CL. Random effects were estimated for CL and V1 and global parameters were estimated for Q and V2. These choices match the results from a recent study on the PK of FVIII [19]. The goal of our analysis was to compare results from the different algorithms rather than to produce optimal models for these two data sets. For this reason, no additional covariate selection was performed. Models were trained until convergence (roughly 1000 epochs for MSE, FO, and VI; 2000 for FOCE) and parameters were saved every 25 epochs. Mean parameters from the last 250 epochs were presented. The ADAM optimiser with a learning rate of 0.1 was used. A larger number of epochs (2000 instead of 1000) were required for the FOCE model to converge when using a lower learning rate (0.01 instead of 0.1). Models were again compared based on the accuracy of typical predictions, final parameter estimates and their stability during training, and the learned functions.

6.2.5 *Model code*

Model code and the simulated data set are available at <https://github.com/Janssena/ME-DCM.jl>.

6.3 RESULTS

6.3.1 Accuracy of variational approximations compared to MCMC

First, we compared the accuracy of the variational posterior approximations obtained using VI to those obtained from MCMC. In Fig. 6.3.1, we can see that applying the path derivative gradient estimator results in accurate posteriors approximations and low variability across replicates compared to the standard estimator. Results for the two scenarios (with and without estimation of Ω and Σ posteriors) are summarised in supplementary Table 6.A.1. Approximate posteriors were most similar (represented by the Wasserstein distance) to the MCMC posteriors when using the path derivative gradient estimator. In both scenarios, variational posteriors of the individual random effects were highly accurate (see supplementary Fig. 6.A.1). Contrastingly, posteriors for the population parameters were less accurate as variational posteriors tended to underestimate the variance of the MCMC posteriors. We focus the remainder of the manuscript on results obtained using the path derivative estimator.

6.3.2 Comparison of VI to first-order objectives

Next, we compare the performance of the different objective functions on the simulated data. We found that models fit using the FOCE objective function behaved erratically during optimisation. Several models failed optimisation (non-positive definite Ω) which seemed to be related to the specific formulation of the objective function used (supplementary Fig. 6.A.2). A reduction of the learning rate (from 0.1 to 0.01) also improved stability of models fit using FOCE (data not shown). In the remainder of the manuscript we thus show results from the FOCE formulation based on Eq. s10 using a learning rate of 0.01 (see appendix 6.B).

In Fig. 6.3.2, we display the objective function value, log KL divergence of Ω , and residual error estimate during training for the FO, FOCE (Eq. s10 + reduced learning rate), and VI objectives. We notice that the FO and VI objectives quickly converge to accurate estimates of the population parameters. These models were not affected by an over-extension of training time, as judged by the stability of parameter estimates during the final 1500 epochs. In contrast, large fluctuations in the KL divergence of Ω are observed when using the

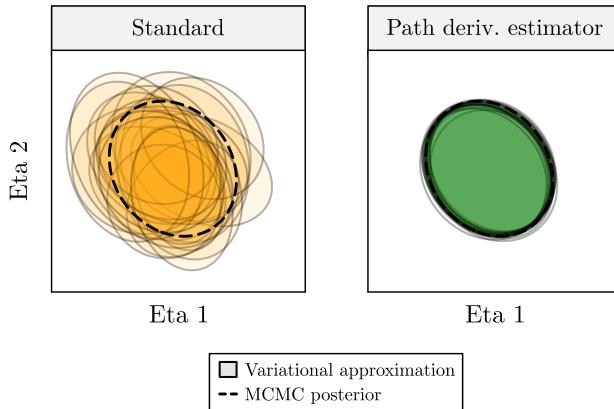


Figure 6.3.1: Accuracy of variational approximations of the random effect posterior obtained through MCMC. 95% confidence regions of the posterior produced by MCMC (dashed lines) and VI (coloured ellipses) are shown for a single subject across 20 replicates of model training. Variational approximations when using the standard VI algorithm (left figure) and the path-derivative estimator (right figure) are shown. The path-derivative estimator results in highly accurate posterior approximations compared to the standard VI objective.

FOCE objective. These fluctuations are not always reflected by the objective function value, making it difficult to determine actual model convergence. Looking at the individual elements of the Ω matrix (i.e. marginal standard deviations S and correlation matrix C), we notice that estimates obtained using FOCE generally underestimated the variances (supplementary Fig. 6.A.2).

The results at the end of optimisation for the MSE, FO, FOCE, and VI objectives are summarised in Table 6.3.1. All methods resulted in similar median root mean squared error of typical predictions. Results for the FO and VI objectives were highly similar, with low error of population parameter predictions. Models fit using the FOCE objective displayed biased parameter estimates as well as high variability between replicates. We can see that models fit using VI completed training slightly faster than models fit using FO (median run time of 14.7 vs. 16.2 min), with FOCE models taking significantly longer (37.7 min). The computational burden of VI can potentially be further reduced close to the training time of MSE-based models by decreasing

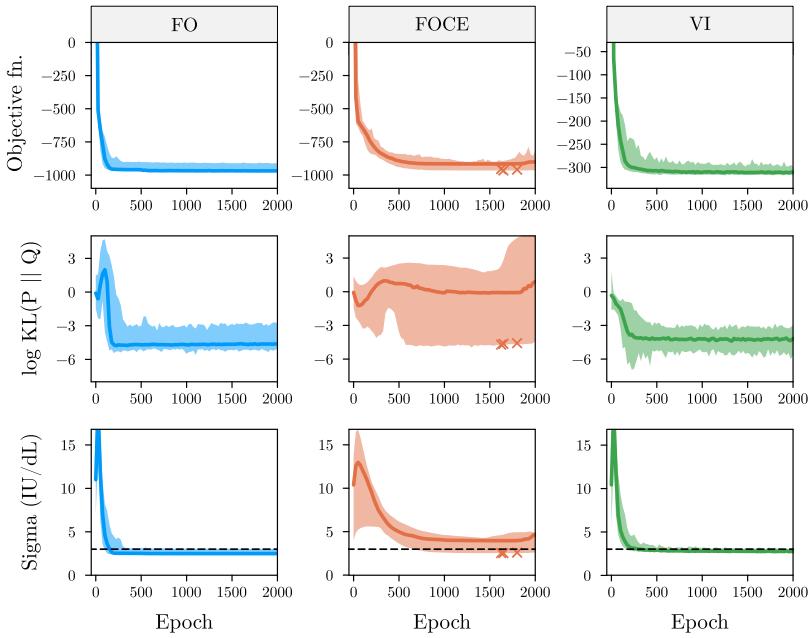


Figure 6.3.2: Objective function value and parameter accuracy during training on the simulated data. Objective function value (top row), log KL divergence of Ω (middle row), and the residual error estimate (bottom row) are shown for the models fit using the FO, FOCE, and VI method. Solid lines indicate median value across replicates along with 95% confidence intervals. Dashed line indicates the true value of the additive error (sigma). Crosses indicate models that failed optimisation. Models fit using the FOCE objective present higher bias of estimated and lower stability during training.

the number of Monte Carlo samples to 1 (median run time of 5.2 min) without loss of parameter accuracy (see supplementary Table 6.A.2).

Finally, we investigate the learned functions at the end of optimisation for each of the models (supplementary Fig. 6.A.4). For all objectives, median covariate effects were very similar to the ground truth functions used in the simulation. Interestingly, we notice a low degree of bias of the learned covariate effects when using the FOCE objective, even though the population parameters were inaccurate. Compared to the mixed-effects models, use of the MSE objective

METHOD	RUN TIME (MIN- UTES)	RMSE (IU / DL)	KL DIVER- GENCE OF Ω	MAE OF ω_1	MAE OF ω_2	MAE OF ADDI- TIVE ERROR (IU / DL)
MSE	3.2 ± 0.73	6.34 ± 0.37	-	-	-	-
FO	16.2 ± 6.5	5.86 ± 0.25	0.009 ± 0.01	0.011 ± 0.01	0.0087 ± 0.001	0.47 ± 0.12
FOCE (Eq. s10)	37.7 ± 7.5	5.75 ± 0.36	1.0 ± 313	0.11 ± 0.05	0.046 ± 0.03	0.92 ± 0.60
VI	14.7 ± 2.6	5.80 ± 0.59	0.011 ± 0.005	0.013 ± 0.008	0.0086 ± 0.002	0.23 ± 0.03

Abbreviations: SD = standard deviation, RMSE = root mean squared error, KL = Kullback–Leibler, MAE = mean absolute error.

Table 6.3.1: Accuracy of model parameters after convergence for the simulated data set. Median values \pm SD for the models at the end of convergence are shown. Parameter estimates obtained from the FOCE objective function presented higher error and variability between training replicates.

seemed to potentially result in a higher degree of variance in the learned effects between model replicates.

6.3.3 Comparison on real world data

Next, we evaluated the performance of the different algorithms on two real-world data sets. Patient characteristics for both data sets are shown in Table 6.3.2. Models fit using a combined error model depicted at least a 20 point decrease in objective function value for all methods. In Table 6.3.3, we show the final parameter estimates for the models with combined error. Models fit using FO or VI resulted in similar median parameters estimates after convergence. However, parameter estimates in some of the replicates of the FO model were less stable, most notably with respect to ω_1 and the proportional error estimate (see supplementary Fig. 6.A.5). Parameter estimates obtained from the FOCE method were again different from the other algorithms. Both the ω_2 and additive error estimates were notably higher in both data sets. Again, the FOCE objective function value was a poor indicator of model convergence, with parameters still changing after apparent convergence (see supplementary Fig. 6.A.5). In contrast,

models fit using VI quickly converged and parameter estimates were stable.

Visualisation of covariate effects can help to provide insights in the covariate effects learned by the models, as well as regions of higher uncertainty due to data sparsity in parts of the covariate space (see Fig. 6.3.3). Learned functions in the perioperative setting (data set two) were similar to those learned based on the PK profiles (see Fig. 6.3.3 and supplementary Fig. 6.A.6). Lower uncertainty over the learned functions was observed when using FOCE, but this result could be replicated for the other objectives by lowering the learning rate (see supplementary Fig. 6.A.7).

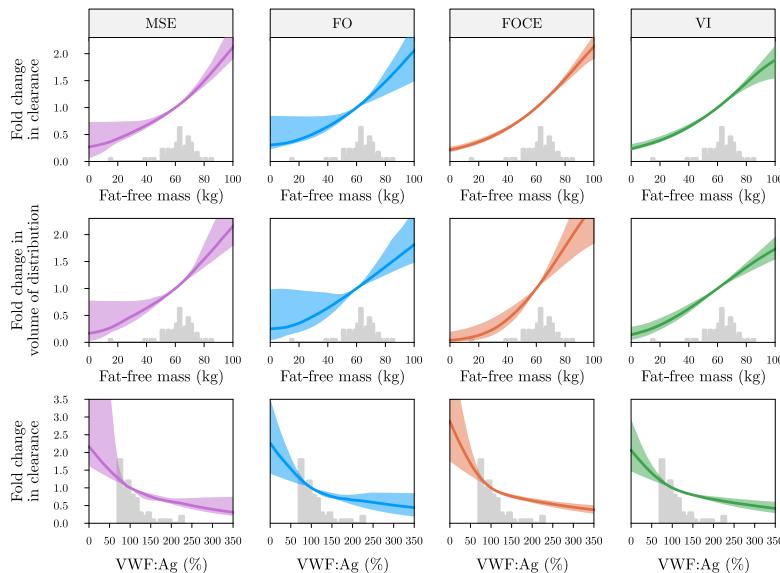


Figure 6.3.3: Learned covariate effects from models fit on real-world data set one. Covariate effects for models fit using the MSE (left column), FO (centre left column), FOCE (centre right column), and VI (right column) are shown. Learned functions are shown for the effect of fat-free mass on clearance (top row), fat-free mass on volume of distribution (middle row) and von Willebrand factor antigen levels on clearance (bottom row) at the end of training on data set one. Median covariate effect (solid line) along with 95% confidence intervals are shown. Grey histograms represent the corresponding covariate distributions.

COVARIATE	DATA SET ONE:		DATA SET TWO	
	PK PROFILES (N = 69)		FOLLOWING SURGERY (N = 66)	
	NUMBER (%-AGE) OR MEAN [RANGE]	NUMBER OF ENTRIES WITH MISSING VALUES (%)	NUMBER (%-AGE) OR MEAN [RANGE]	NUMBER OF ENTRIES WITH MISSING VALUES (%)
Body weight (kg)	86.0 [50.4—134]	1 (1.4%)	85.7 [50.4—134]	0 (0%)
Height (cm)	179 [148—198]	3 (4.3%)	178 [148—198]	0 (0%)
Age (years)	47.6 [12.1—76.9]	0 (0%)	47.6 [12.4—76.9]	0 (0%)
Blood group		0 (0%)		0 (0%)
- A	19 (28%)		18 (27%)	
- B	3 (4.3%)		3 (4.5%)	
- AB	5 (7.2%)		5 (7.5%)	
- O	42 (61%)		40 (61%)	
Pre-assessed surgical risk		NA		0 (0%)
- Low	NA		35 (53%)	
- Medium	NA		31 (47%)	
Haemophilia severity		0 (0%)		0 (0%)
- Moderate	22 (32%)		22 (33%)	
- Severe	47 (68%)		44 (67%)	
Expected blood loss		NA		0 (0%)
- Mild	NA		42 (64%)	
- Moderate	NA		24 (36%)	
Blood loss during surgery (mL)	NA	NA	227 [0—1200]	21 (32%)
Brand of FVIII concentrate		0 (0%)		0 (0%)
Octocog alfa (Kogenate [®])	18 (26%)		18 (27%)	
Octocog alfa (Advate [®])	22 (32%)		21 (34%)	
Moroctocog alfa (ReFacto AF [®])	4 (5.8%)		4 (6.1%)	
Plasma-derived FVIII Concentrate (Aafact [®])	3 (4.3%)		3 (4.5%)	
Turoctocog alfa (NovoEight [®])	22 (32%)		20 (30%)	
VWF:Ag (%)	113 [61—225]	45 (65.2%)	131 [0.43—384]	9 (13.6%)
VWF:act (%)	106 [58—185]	45 (65.2%)	127 [32—396]	9 (13.6%)
FVIII measurements per patient	3.26 [3—10]	-	8.61 [2—21]	-

Abbreviations: kg = kilogram, cm = centimetre, FVIII = blood clotting factor VIII, aPTT = activated partial thromboplastin time, s = seconds, PT = Prothrombin time, VWF = von Willebrand factor, NA = not applicable.

Table 6.3.2: Patient characteristics for the two real-world data sets. Patient characteristics and missing data are shown for data set one and two. A point to note are the differences in the amount of missing data between the two clinical settings. Most prominently, VWF:Ag levels were missing for most (65%) subjects in data set one.

METHOD	RUN TIME ± SD (MINUTES)	RMSE ± SD (IU / DL)	ω_1 (%CV) ± SD	ω_2 (%CV) ± SD	ADDITIONAL ERROR (IU / DL) ± SD	PROPORTIONAL ERROR ± SD
DATA SET ONE (PROPHYLACTIC SETTING)						
MSE	2.1 ± 0.16	14.1 ± 0.24	-	-	-	-
FO	9.1 ± 2.2	14.3 ± 0.77	0.289 (29.5) ± 0.044	0.127 (12.8) ± 0.020	3.09 ± 0.43	0.105 ± 0.013
FOCE (Eq. s10)	54.2 ± 14 ^a	19.0 ± 4.3	0.240 (24.4) ± 0.019	0.465 (49.1) ± 0.052	3.70 ± 0.05	0.108 ± 0.004
VI	8.0 ± 0.51	14.3 ± 0.69	0.282 (28.8) ± 0.012	0.160 (16.1) ± 0.004	2.89 ± 0.077	0.094 ± 0.017
DATA SET TWO (PERIOPERATIVE SETTING)						
MSE	2.3 ± 0.17	27.6 ± 1.13	-	-	-	-
FO	19.5 ± 3.8	32.0 ± 1.66	0.300 (30.7) ± 0.012	0.211 (21.3) ± 0.018	2.89 ± 1.63	0.151 ± 0.012
FOCE (Eq. s10)	113 ± 20 ^a	31.5 ± 1.66	0.321 (32.9) ± 0.014	0.326 (33.5) ± 0.020	4.53 ± 0.37	0.152 ± 0.005
VI	14.6 ± 1.2 ^b	30.0 ± 1.17	0.316 (32.4) ± 0.005	0.179 (18.0) ± 0.001	2.46 ± 0.024	0.165 ± 0.001

^a = convergence after 2000 epochs, ^b = convergence after 1250 epochs. Abbreviations:
SD = standard deviation, RMSE = root mean squared error, CV = coefficient of variation.

Table 6.3.3: Accuracy of model parameters on real world data sets
Median values ± SD are shown. Coefficient of variation was calculated using the following formula: $CV(\%) = \sqrt{\exp(\omega^2) - 1} \cdot 100\%$. Compared to the other methods, the FOCE objective results in divergent parameter estimates. Higher RMSE in data set two is indicative of the higher inter-individual variability in FVIII levels observed during surgical procedures.

6.4 DISCUSSION

In this work, we investigated the performance of classical first-order approximations as well as ML-based variational methods for estimating mixed-effects in DCMs. Results from our simulation experiment suggest that both the FO and VI objectives reliably converged to accurate solutions, whereas the FOCE objective function resulted in biased estimates and high variability amongst training replicates. These results were replicated in two real-world data sets, where we again observed divergent results when using the FOCE objective. Here, VI resulted in the most reliable results as some models fit using FO depicted lower parameter stability during training. Learned covariate effects for all models could be visualised by using the multi-branch architecture of the DCM. This enables model interpretation and is useful for critiquing the model during development.

Even though the FOCE objective function is widely regarded to be more accurate than the FO method, our results indicate that this is not always the case. When the underlying model is highly flexible and is trained using gradient descent, as is the case when using neural networks, the FOCE algorithm seemed to result in poor convergence behaviour. Although a different formulation of the objective function and lowering of the learning rate slightly improved results, optimisation still was not reliable. Population parameter estimates were highly variable during training, even after apparent convergence based on the stabilisation of the objective function value. We hypothesise that frequent changes to the loss landscape affect the stability of optimisation when using gradient descent. Since the fixed effects model initially has low accuracy, early η estimates shrink to the prior mean with relatively high posterior variance. As a result, the prior variances (Ω) might have a tendency to shrink to zero. After a few iterations, the accuracy of typical PK parameter improves, resulting in jumps in the estimates of η away from zero and potentially large changes to the loss landscape. Methods such as gradient descent might perform poorly in such settings, getting stuck in poor local optima and frequently changing the direction of gradients in response to changes to the loss landscape. For both the FO objective and VI such changes do not occur, since the random effects are either fixed during training (as in FO) or part of the parameter space (as in VI). Additional research is needed to investigate why the FOCE objective fails in this setting.

As an alternative to the FOCE objective, we suggest VI for the concurrent optimisation of fixed effect parameters and subject-specific

random effect posteriors. We show that variational posteriors were very accurate when using the path derivative gradient estimator, which is simple to implement. Most probabilistic programming languages such as Turing.jl or Pyro provide functionality for fast implementation of VI [24, 25]. Results from our experiments indicate fast and stable convergence to an accurate set of parameter estimates. Additional benefits of VI are improved computational speed compared to FOCE (even outperforming FO for one of our data sets) as well as it being part of an active field of research, potentially bringing more improvements in terms of speed and accuracy [26]. Furthermore, the complexity of the variational approximation can be controlled, making the method suitable for problems where the random effect posterior is multi-modal or better described by a more complex distribution by for example using Gaussian mixture models or normalising flows based variational posteriors, respectively [18, 27].

VI is conceptually very similar to (stochastic) expectation maximisation (EM) procedures [28, 29]. In Stochastic approximation EM (SAEM), samples from the random effect posterior are taken (for example using MCMC) and a stochastic averaging procedure with adaptive step sizes is performed to approximate the integral in Eq. 6.3 [29]. This is followed by maximisation of the fixed-effects parameters based on the obtained approximation. In VI, samples are instead taken from a Variational distribution whose parameters are directly optimised along with the fixed-effects parameters. A benefit of the latter approach is that we obtain a closed-form expression for the random effect posterior and that no adaptive step size procedures are required. It might be of interest to compare the performance of these two approaches to see if there are notable differences.

Even though the FO method resulted in reasonable median parameter estimates in our experiments, the use of VI might be preferred. In more complex models, FO is likely to result in less accurate parameter estimates. We already found that some training replicates on the second real-world data set showed signs of lower stability and poor accuracy. It has been shown that the FO method can often produce biased parameter estimates with incorrect uncertainty estimates in certain settings [30]. Furthermore, it is well known that the FO method is not suited for problems with high levels of inter-individual variability [16]. Especially in the context of pharmacodynamic (PD) models, this variability is expected to be relatively large (often $>100\%$ coefficient of variation) and so the FO method might be unsuited in most cases. In contrast, accuracy of VI depends on the chosen variational approxima-

tion (Gaussian approximations are often sufficient) and the number of Monte Carlo samples, both of which can be adapted based on the complexity of the problem at hand.

There were also some limitations to this work. First, our results indicated that variational approximations estimated over population parameters depicted an underestimation of posterior variance compared to MCMC. Unfortunately, estimation of the population parameter posteriors using MCMC is computationally intensive as it still requires iteration over all subjects in the data set. This might only be feasible in small data sets (e.g. ≤ 30 subjects) and when using relatively simple models (simple ODEs, small neural network, and small number of random effect parameters). To estimate uncertainty over model parameters we might need to resort to deterministic methods to estimate standard errors. Similar to the approach used by NLME models, reasonable estimates can be obtained based on post-hoc Gaussian approximations based on the Fisher information matrix. Second, we use deterministic methods to optimise neural network weights. Since models could be prone to overfitting, we might want to marginalise over predictions from many model replicates to reduce spurious effects and to obtain estimates of functional uncertainty. Ideally, uncertainty over covariate effects can be estimated in a single model replicate. Alternatively, the use of priors over the desired function space in this context can be of interest in order to regularise function complexity. It would be of interest to investigate how these improvements can be implemented in practice. Finally, we did not perform an exhaustive evaluation of the performance of the objective functions in many different data sets, different degrees model complexity, or for very different initial parameter and prior distributions settings. More research might be desirable to evaluate the performance of VI in multiple practical settings.

6.5 CONCLUSION

In summary, our work introduces mixed-effects estimation in the DCM framework. Highly accurate posterior approximations for the random effects could be obtained using VI, and estimated population parameters were accurate and stable during training. We found that the FOCE method did not provide reliable results and might not be suited for this purpose. In our experiments, VI was the most reliable approach for the estimation of mixed effects and might perform better in more complex models compared to FO. Mixed-effects models enable

the individualisation of predictions based on clinical measurements, enhancing the likelihood of the clinical adoption of these algorithms. This extension to the DCM framework further promotes the use of ML-based methods as a viable alternative to classical NLME models.

REFERENCES

- [1] Alexander Janssen, Frank C Bennis, and Ron AA Mathôt. "Adoption of machine learning in pharmacometrics: an overview of recent implementations and their considerations". In: *Pharmaceutics* 14.9 (2022), p. 1814.
- [2] Kamilė Stankevičiūtė, Jean-Baptiste Woillard, Richard W Peck, Pierre Marquet, and Mihaela van der Schaar. "Bridging the worlds of pharmacometrics and machine learning". In: *Clinical Pharmacokinetics* 62.11 (2023), pp. 1551–1565.
- [3] Mason McComb, Robert Bies, and Murali Ramanathan. "Machine learning in pharmacometrics: Opportunities and challenges". In: *British Journal of Clinical Pharmacology* 88.4 (2022), pp. 1482–1499.
- [4] James Lu, Kaiwen Deng, Xinyuan Zhang, Gengbo Liu, and Yuanfang Guan. "Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens". In: *Iscience* 24.7 (2021).
- [5] Alexander Janssen et al. "Deep compartment models: a deep learning approach for the reliable prediction of time-series data in pharmacokinetic modeling". In: *CPT: Pharmacometrics & Systems Pharmacology* 11.7 (2022), pp. 934–945.
- [6] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. "Neural ordinary differential equations". In: *Advances in neural information processing systems* 31 (2018).
- [7] Dominic Stefan Bräm, Uri Nahum, Johannes Schropp, Marc Pfister, and Gilbert Koch. "Low-dimensional neural ODEs and their application in pharmacokinetics". In: *Journal of Pharmacokinetics and Pharmacodynamics* 51.2 (2024), pp. 123–140.
- [8] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: (2022). arXiv: 1312.6114 [stat.ML]. URL: <https://arxiv.org/abs/1312.6114>.
- [9] Ankush Ganguly, Sanjana Jain, and Ukrit Watchareeruetai. "Amortized Variational Inference: A Systematic Review". In: *Journal of Artificial Intelligence Research* 78 (2023), pp. 167–215.
- [10] Andrea Aspertì and Matteo Trentin. "Balancing reconstruction error and kullback-leibler divergence in variational autoencoders". In: *Ieee Access* 8 (2020), pp. 199440–199448.
- [11] Bin Dai and David Wipf. "Diagnosing and Enhancing VAE Models". In: (2019). arXiv: 1903.05789 [cs.LG]. URL: <https://arxiv.org/abs/1903.05789>.
- [12] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

- [13] Lewis B Sheiner, Barr Rosenberg, and Kenneth L Melmon. "Modelling of individual pharmacokinetics for computer-aided drug dosage". In: *Computers and Biomedical Research* 5.5 (1972), pp. 441–459.
- [14] Mary J Lindstrom and Douglas M Bates. "Nonlinear mixed effects models for repeated measures data". In: *Biometrics* (1990), pp. 673–687.
- [15] Lewis B Sheiner and Stuart L Beal. "Evaluation of methods for estimating population pharmacokinetic parameters. I. Michaelis-Menten model: routine clinical pharmacokinetic data". In: *Journal of pharmacokinetics and biopharmaceutics* 8.6 (1980), pp. 553–571.
- [16] B Jones and J Wang. "Constructing optimal designs for fitting pharmacokinetic models". In: *Statistics and Computing* 9 (1999), pp. 209–218.
- [17] Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. "What are Bayesian neural network posteriors really like?" In: *International conference on machine learning*. 2021, pp. 4629–4640.
- [18] Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. "Sticking the landing: Simple, lower-variance gradient estimators for variational inference". In: *Advances in Neural Information Processing Systems* 30 (2017).
- [19] Alexander Janssen et al. "A Generative and Causal Pharmacokinetic Model for Factor VIII in Hemophilia A: A Machine Learning Framework for Continuous Model Refinement". In: *Clinical Pharmacology & Therapeutics* 115.4 (2024), pp. 881–889.
- [20] Alexander Janssen, Frank C Bennis, Marjon H Cnossen, Ron AA Mathôt, OPTI-CLOT Study Group, and SYMPHONY Consortium. "On inductive biases for the robust and interpretable prediction of drug concentrations using deep compartment models". In: *Journal of Pharmacokinetics and Pharmacodynamics* (2024), pp. 1–12.
- [21] Iris van Moort, Tim Preijers, Laura H Bukkems, Hendrika CAM Hazendonk, Johanna G van der Bom, Britta AP Laros-van Gorkom, Erik AM Beckers, Laurens Nieuwenhuizen, Felix JM van der Meer, Paula Ypma, et al. "Perioperative pharmacokinetic-guided factor VIII concentrate dosing in haemophilia (OPTI-CLOT trial): an open-label, multicentre, randomised, controlled trial". In: *The Lancet Haematology* 8.7 (2021), e492–e502.
- [22] Iris van Moort, Laura H Bukkems, Jessica M Heijdra, Roger EG Schutgens, Britta AP Laros-van Gorkom, Laurens Nieuwenhuizen, Felix JM van der Meer, Karin Fijnvandraat, Paula Ypma, Moniek PM de Maat, et al. "von Willebrand factor and factor VIII clearance in perioperative hemophilia A patients". In: *Thrombosis and haemostasis* 120.07 (2020), pp. 1056–1065.
- [23] Hesham Saleh Al-Sallami, Ailsa Goulding, Andrea Grant, Rachael Taylor, Nicholas Holford, and Stephen Brent Duffull. "Prediction of fat-free mass in children". In: *Clinical pharmacokinetics* 54 (2015), pp. 1169–1178.
- [24] Hong Ge, Kai Xu, and Zoubin Ghahramani. "Turing: a language for flexible probabilistic inference". In: *International conference on artificial intelligence and statistics*. 2018, pp. 1682–1690.
- [25] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. "Pyro: Deep universal probabilistic programming". In: *Journal of machine learning research* 20.28 (2019), pp. 1–6.

- [26] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. "Advances in variational inference". In: *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018), pp. 2008–2026.
- [27] Danilo Rezende and Shakir Mohamed. "Variational inference with normalizing flows". In: *International conference on machine learning*. 2015, pp. 1530–1538.
- [28] Dimitris G Tzikas, Aristidis C Likas, and Nikolaos P Galatsanos. "The variational approximation for Bayesian inference". In: *IEEE Signal Processing Magazine* 25.6 (2008), pp. 131–146.
- [29] Bernard Delyon, Marc Lavielle, and Eric Moulines. "Convergence of a stochastic approximation version of the EM algorithm". In: *Annals of statistics* (1999), pp. 94–128.
- [30] Céline Dartois, Annabelle Lemenuel-Diot, Christian Laveille, Brigitte Tranchand, Michel Tod, and Pascal Girard. "Evaluation of uncertainty parameters estimated by different population PK software and methods". In: *Journal of pharmacokinetics and pharmacodynamics* 34 (2007), pp. 289–311.

APPENDIX

6.A SUPPLEMENTARY TABLES AND FIGURES

WASSERSTEIN DISTANCE WITH RESPECT TO MCMC POSTERIOR $\times 10^{-3}$					
VI RITHM	η ($W_2 \pm \text{SD}$)	σ ($W_1 \pm \text{SD}$)	ω_1 ($W_1 \pm \text{SD}$)	ω_2 ($W_1 \pm \text{SD}$)	ρ ($W_1 \pm \text{SD}$)
TYPICAL PK AND POPULATION PARAMETERS KNOWN					
Standard estimator	9.0 ± 3.5	-	-	-	-
Path derivative estimator	5.5 ± 2.9	-	-	-	-
TYPICAL PK PARAMETERS KNOWN					
Standard estimator	8.8 ± 3.3	5.0 ± 0.6	7.2 ± 3.1	10.9 ± 3.2	46.9 ± 23.9
Path derivative estimator	4.5 ± 2.3	0.6 ± 0.3	6.7 ± 3.5	7.2 ± 4.0	43.4 ± 14.3

Abbreviations: VI = Variational Inference, W_2 = z -Wasserstein distance, W_1 = 1 -Wasserstein distance, SD = standard deviation.

Table 6.A.1: Accuracy of the variational posteriors compared to MCMC. Wasserstein distances were calculated with respect to multivariate normal distributions fit to the samples obtained through MCMC. For σ , ω_1 , and ω_2 , MCMC posteriors were represented by fitting a LogNormal distributions to the samples, while a SkewNormal distribution was fit to represent the posterior for ρ . Bold text represents the lowest Wasserstein distances.

METHOD	RUN TIME ± SD (MINUTES)	FINAL OBJECTIVE VALUE ± SD		KL DIVERGENCE OF Ω ± SD		MAE OF ω_1 ± SD	MAE OF ω_2 ± SD	MAE OF σ ± SD (IU/DL)
		TIVE FUNC- TION	RMSE ± SD (IU/DL)	DIVER- GENCE	MAE	MAE	MAE	MAE OF σ ± SD (IU/DL)
One sample	5.20 ± 0.41	298 ± 3.6 ^a	5.83 ± 0.68	0.0089 ± 0.022	0.014 ± 0.001	0.0029 ± 0.003	0.050 ± 0.042	
Three samples	14.7 ± 2.6	309 ± 2.4 ^a	5.80 ± 0.59	0.011 ± 0.005	0.013 ± 0.008	0.0086 ± 0.002	0.23 ± 0.03	

^a = Based on stochastic estimates of the ELBO. Higher is better. SD = standard deviation, RMSE = root mean squared error, KL = Kullback-Leibler, MAE = mean absolute error.

Table 6.A.2: Comparison of parameter estimates of VI objective based on the number of Monte Carlo samples. Median value ± SD are shown for the VI models fit to the synthetic data experiment. Decreasing the number of Monte Carlo samples from three to one did not seem to affect parameter accuracy.

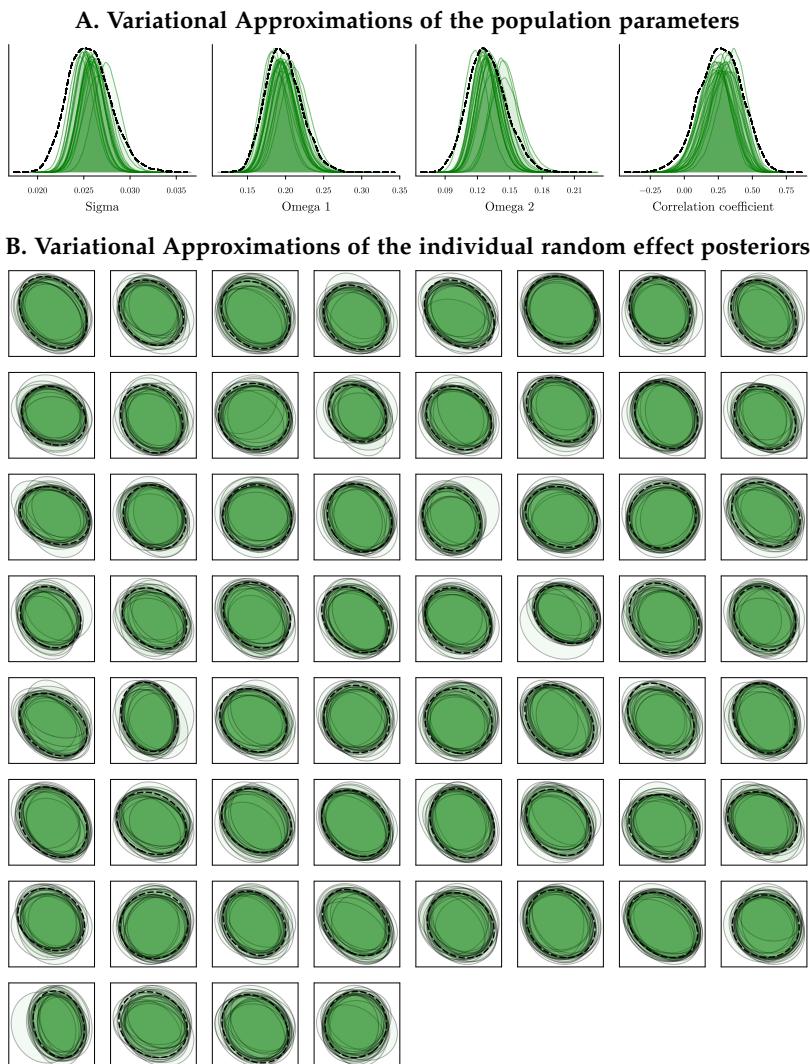


Figure 6.A.1: Posterior approximations using variational inference. Variational posteriors for the population parameters (A) as well as the individual random effect parameters (B) are shown. Black dashed lines represent posteriors obtained through MCMC. Results are shown for 20 replicates of the model fit using the path derivative gradient estimator.

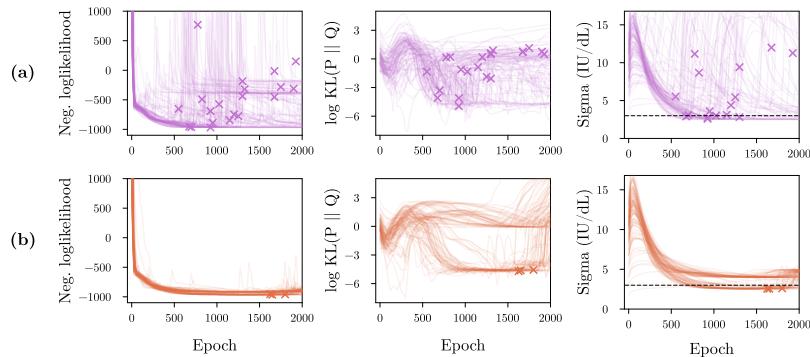


Figure 6.A.2: Objective function value and parameter accuracy for the FOCE objectives during training. The objective function value (left column), $\log KL$ divergence of the estimated random effect prior (centre column), and residual error estimate (right column) during training are shown for the FOCE based objectives. Results are shown for the FOCE objective according to equation s9 (a), equation s10 (b) using the reduced learning rate. Each line represents a single replicate fit to one of the data folds. Dashed line indicates the true value for sigma. Crosses indicate models that failed convergence. The formulation of the FOCE objective based on equation s9 (a) depicts lower stability during training and a higher fraction of models failing optimisation.

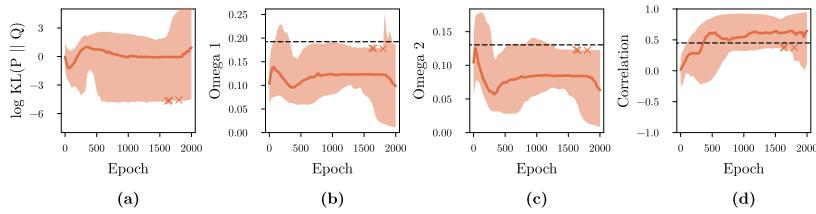


Figure 6.A.3: Population parameter estimates during optimisation using the FOCE objective. The log KL divergence of the estimated random effect prior (a), marginal standard deviation of ω_1 (b) and ω_2 (c), and their correlation coefficient (d) during training are shown. Results are shown for the FOCE objective according to equation s10 with reduced learning rate. Each line represents a single replicate along the data folds. Dashed lines indicate the true parameter value. Crosses represent early end of optimisation due to errors. The model generally seems to underestimate the marginal variances of the true prior distribution.

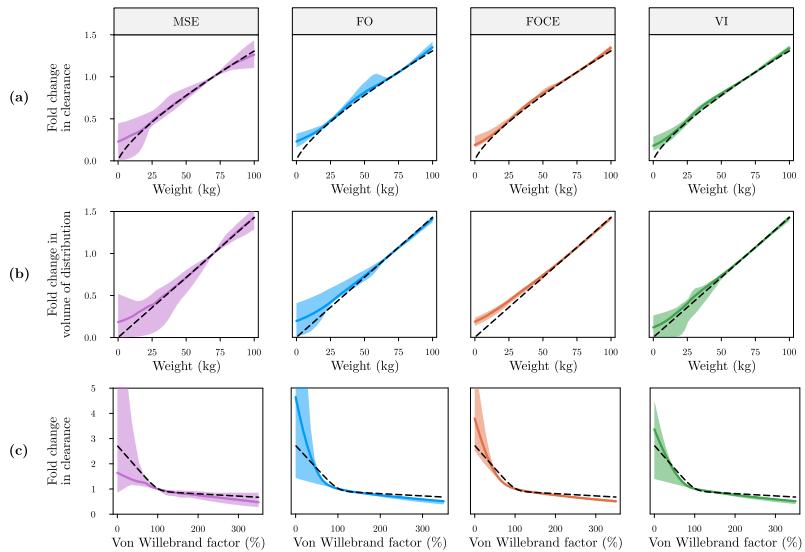


Figure 6.A.4: Learned covariate effects after training on the synthetic data set. Covariate effects for models fit using the FO (left column), FOCE (Eq. s10; centre left column), VI (centre right column), and mean squared error (right column) are shown. Learned functions are shown for the effect of weight on clearance (a), weight on volume of distribution (b) and von Willebrand factor on clearance (c). Median covariate effect (solid line) along with 95% confidence intervals are shown. Dashed black lines indicates the ground truth functions used in the simulation.

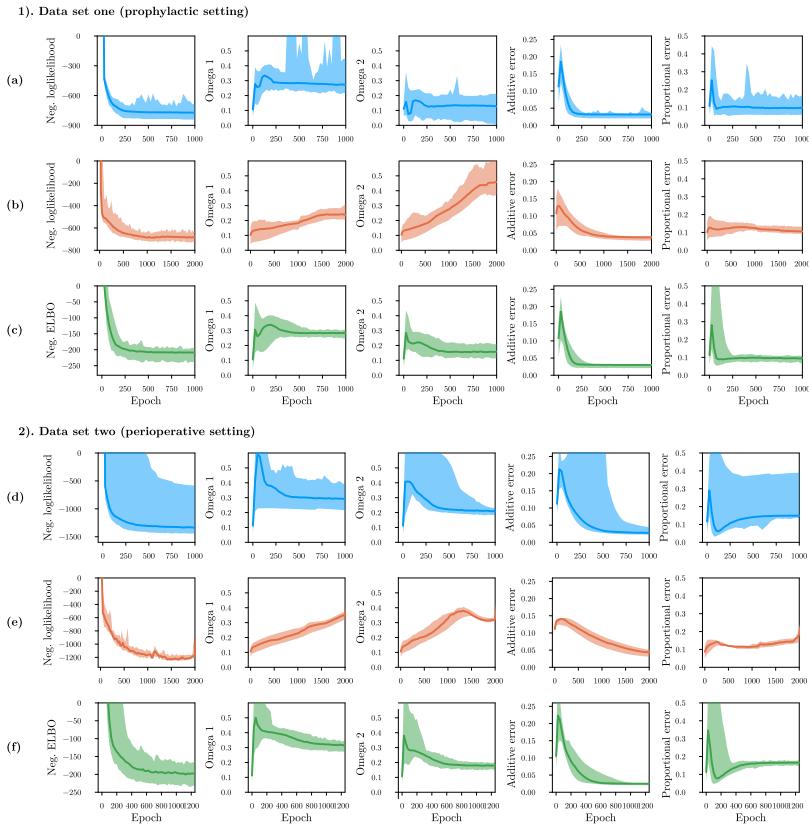


Figure 6.A.5: Parameter estimates during training on the real-world data sets. Parameter estimates during training are shown for the FO (a & d), FOCE (b & e), and VI (c & f) based objectives. Results are shown for data set one (a-c) and data set two (d-f). Median estimate (solid line) along with 95% confidence interval across replicates are shown.

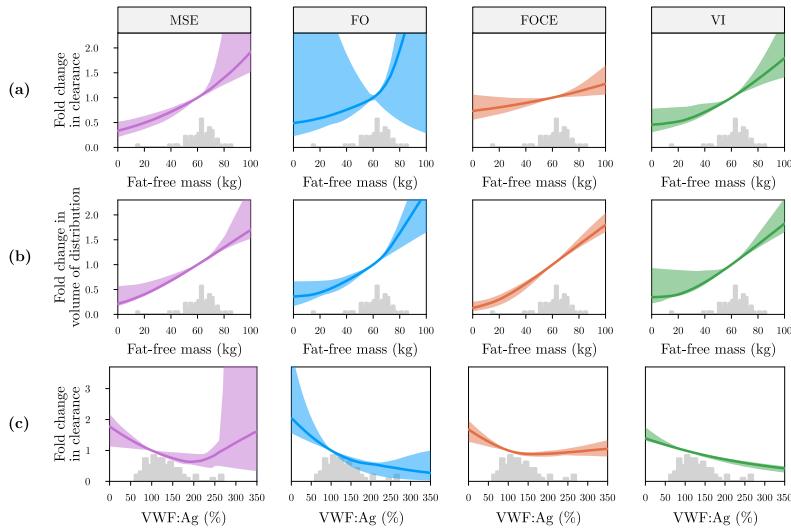


Figure 6.A.6: Learned functions after training on real-world data set two. Covariate effects for models fit using the MSE (left column), FO (centre left column), FOCE (centre right column), and VI (right column) are shown. Learned functions are shown for the effect of fat-free mass on clearance (a), fat-free mass on volume of distribution (b) and von Willebrand factor antigen (VWF:Ag) levels on clearance (c) at the end of training for data set two. Median covariate effect (solid line) along with 95% confidence intervals are shown. Grey histograms represent the corresponding covariate distributions.

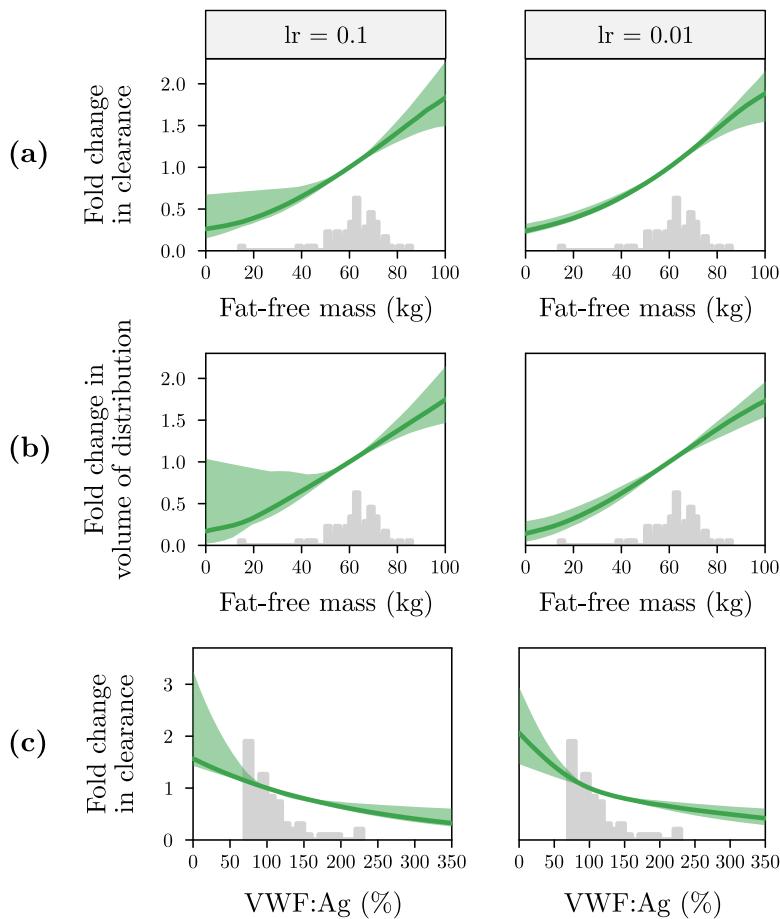


Figure 6.A.7: Decreasing the learning rate lowers uncertainty over learned effects for VI. Results are shown for the VI models trained using the regular learning rate (left column) and reduced learning rate (right column). Learned functions are shown for the effect of fat-free mass on clearance (a), fat-free mass on volume of distribution (b) and von Willebrand factor antigen levels on clearance (c) at the end of training for data set one. Median covariate effect (solid line) along with 95% confidence intervals are shown. Grey histograms represent the corresponding covariate distributions. lr = learning rate, VWF:Ag = von Willebrand factor antigen.

6.B DERIVATION OF OBJECTIVE FUNCTIONS

6.B.1 Laplace approximation

Given a complex integral of the form $\int f(x)dx$, where $f(\cdot)$ is a twice-differentiable function, $f(x)$ can be re-expressed as $g(x) = \log f(x)$, such that $\int f(x)dx = \int \exp g(x)dx$. Consider the second-order Taylor series expansion of $g(x)$ around a point x_0 :

$$g(x) \approx g(x_0) + g'(x_0)(x - x_0) + \frac{1}{2}g''(x_0)(x - x_0)^2 \quad (\text{s1})$$

If we set x_0 to be the mode of $g(x)$ the second term becomes zero (since $g'(x_0) = 0$). We thus obtain the following approximation of the integral:

$$\int f(x)dx \approx \int \exp(g(x_0) + \frac{1}{2}g''(x_0)(x - x_0)^2)dx \quad (\text{s2})$$

Since $\exp g(x_0) = f(x_0)$ is a constant, we can move it out of the integral to obtain:

$$\begin{aligned} \int \exp(g(x_0) + \frac{1}{2}g''(x_0)(x - x_0)^2)dx &= f(x_0) \cdot \int \exp(\frac{1}{2}g''(x_0)(x - x_0)^2)dx \\ &= f(x_0) \cdot \sqrt{\frac{2\pi}{-g''(x_0)}} \end{aligned} \quad (\text{s3})$$

The second term in the last equation originates from integration of the probability density function of a normal distribution:

$$\int p(X)dx = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) dx = 1 \quad (\text{s4})$$

From Eq. s4 we can see that $\sigma \cdot \sqrt{2\pi} = \int \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) dx$. If we set $\sigma = -g''(x_0)^{-1}$ we recover the second term in Eq. s3. The Laplace approximation thus results in a Gaussian approximation around the mode of the random effects:

$$\int f(x)dx \approx f(x_0) \cdot \sqrt{\frac{2\pi}{-g''(x_0)}} \quad (\text{s5})$$

In the context of non-linear mixed effects models this results in the following objective function after simplification:

$$\mathcal{L}(\Theta, \hat{\eta}) = p(y | \hat{\eta}; \Theta) + \log |\Omega| + \hat{\eta} \cdot \Omega^{-1} \cdot \hat{\eta}^T + \left| \Omega^{-1} + \frac{H(\hat{\eta})}{2} \right| \quad (\text{s6})$$

Where H is the hessian of the likelihood with respect to η : $H(\eta) = \frac{\partial^2 p(y|\eta)}{\partial \eta}$.

6.B.2 First-order conditional estimation (FOCE)

To avoid the computation of the second order derivatives, the Hessian matrix can be approximated as a function of the Jacobian vector of $\hat{\eta}$:

$$\mathbb{E}[H(\eta)] \approx \frac{1}{2} \mathbb{E}[J(\eta) \cdot J(\eta)^T] \quad (\text{s7})$$

This additional approximation results in the first-order conditional estimation objective function (also see [1]):

$$\mathcal{L}(\Theta, \hat{\eta}) = p(y | \hat{\eta}; \Theta) + \log |\Omega| + \hat{\eta} \cdot \Omega^{-1} \cdot \hat{\eta}^T + \left| \Omega^{-1} + \frac{\mathbb{E}[J(\eta) \cdot J(\eta)^T]}{4} \right| \quad (\text{s8})$$

We can further simplify this equation to obtain the FOCE objective function that is used in NONMEM:

$$-2\mathcal{L}(\Theta, \hat{\eta}) = \log |C| + \frac{(y - A(t; \hat{z}, I) + J(\hat{\eta}) \cdot \hat{\eta})^2}{C} \quad (\text{s9})$$

Where $C = J(\hat{\eta}) \cdot \Omega \cdot J(\hat{\eta})^T + \Sigma$ and \hat{z} is the individual estimate of the ODE parameters based on $\hat{\eta}$. The Jacobian is calculated with respect to the output of the ODE. An equivalent expression exists:

$$-2\mathcal{L}(\Theta, \hat{\eta}) = \log |C| + \frac{(y - A(t; \hat{z}, I))^2}{\Sigma} + \hat{\eta} \cdot \Omega^{-1} \cdot \hat{\eta}^T \quad (\text{s10})$$

We use this objective function (Eq. s10) in the manuscript.

6.B.3 The FO objective

In the FO objective, the mode of η is assumed to be located at the population mean (i.e. zero). This results in the following objective function following from Eq. s9:

$$-2\mathcal{L}(\Theta, \hat{\eta}) = \log |C_0| + \frac{(y - A(t; z_0, I))^2}{C_0} \quad (\text{s11})$$

where $C_0 = J(0) \cdot \Omega \cdot J(0)^T + \Sigma$.

6.B.4 Derivation of the ELBO

In Bayesian inference, given a set of observations $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ we are often interested in obtaining the posterior distribution over a set of latent variables $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$. We are however often unable to compute the model evidence $p(\mathbf{X})$ as this requires integration over all possible values of \mathbf{Z} . The goal of Variational Inference (VI) is to instead minimise the differences between the true posterior and a (simpler) variational approximation $q(\mathbf{Z})$. One way to represent the differences between two distributions is via their KL-divergence:

$$\begin{aligned} \text{KL}(q(\mathbf{Z}) \| p(\mathbf{Z} | \mathbf{X})) &= \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z} | \mathbf{X})} dz \\ &= \mathbb{E}_{q(\mathbf{Z})} \left[\log \frac{q(\mathbf{Z})}{p(\mathbf{Z} | \mathbf{X})} \right] \\ &= \mathbb{E}_{q(\mathbf{Z})} [\log q(\mathbf{Z})] - \mathbb{E}_{q(\mathbf{Z})} [\log p(\mathbf{X}, \mathbf{Z})] + \log p(\mathbf{X}) \end{aligned} \quad (\text{s12})$$

We can rewrite this expression to obtain:

$$\log p(\mathbf{X}) = \underbrace{\mathbb{E}_{q(\mathbf{Z})} [\log p(\mathbf{X}, \mathbf{Z}) - \log q(\mathbf{Z})]}_{\text{ELBO}} + \underbrace{\text{KL}(q(\mathbf{Z}) \| p(\mathbf{Z} | \mathbf{X}))}_{\text{divergence}} \quad (\text{s13})$$

Note that the KL divergence is an asymmetric measure, i.e. $\text{KL}(q(\mathbf{Z}) \| p(\mathbf{Z} | \mathbf{X})) \neq \text{KL}(p(\mathbf{Z} | \mathbf{X}) \| q(\mathbf{Z}))$. Swapping terms in the KL divergence results in a different objective function with different behaviour.

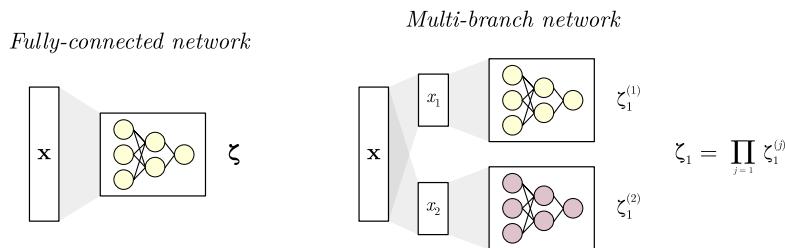
REFERENCES

- [1] Wang, Y. (2007). Derivation of various NONMEM estimation methods. Journal of Pharmacokinetics and pharmacodynamics, 34, 575-593.

6.C MODEL ARCHITECTURE

6.C.1 Multi-branch network

In a multi-branch neural network architecture, the covariates are connected to independent sub-networks, such that the model learns the effect of each covariate in isolation. The independent covariates are combined using a product, similar to the common implementation of covariates in non-linear mixed effects models. In this sense, the architecture is similar to a generalised additive model, using product accumulation rather than the sum of covariate effects. Typical fully-connected neural networks can learn complex interactions between the covariates. By removing the possibility of learning such potentially spurious correlations, model performance and generalisability can potentially be improved. Similarly, we can specifically link covariates with known causal effects on one of the parameters, preventing the model from learning any spurious effects with respect to the other parameters. An additional benefit of the approach is that the output of each sub-model can be visualised, allowing for the interpretation of the learned covariate effects. A schematic representation of fully-connected and multi-branch networks is provided below.



6.C.2 Model architecture

A multi-branch architecture was used to learn the effect of weight (or fat-free mass in the real-world experiment) on clearance and volume of distribution, and the effect of von Willebrand factor antigen (VWF:Ag) levels on clearance. The first model consisted of a single hidden layer containing 12 neurons feeding into a transformed softplus activation

function: $\pi(x) = \frac{1}{10} \cdot \log(\exp(10 \cdot x) + 1)$. Inputs were normalised between roughly 0 and 1 by dividing model input by 150 kg. Output from this hidden layer was fed into two independent hidden layers, each again consisting of 12 neurons connected to a single output neuron. The two independent output neurons represent the effect of the covariate on clearance or volume of distribution. This way, the two effects share a similar base relationship based on the first hidden layer which and individual differences between their effects on the different PK parameters can be learned based on the second set of hidden layers.

The second model (VWF:Ag on clearance) consisted of a single hidden layer of 12 neurons feeding into a single output neuron. Again the transformed softplus activation function was used. Inputs were normalised between roughly 0 and 1 by dividing model input by 350%. Global parameters were estimated for Q and V_2 . All parameters were constrained to be positive using a softplus activation function. We chose 12 neurons in all hidden layers as this allowed a sufficient level of complexity of the learned functions, while not being so large as to result in excessive overfitting (which could be likely when using 128 neurons for example). The number of neurons can potentially be optimised (by means of hyperparameter tuning), but we found the risks of overfitting to be already sufficiently managed when using 12 neurons. Bias parameters in the output layers were initialised to ones to initialise the model at reasonable estimates at the start of training.

6.c.3 Visualisation of learned effects

Visualisations of learned functions were obtained by entering dummy input to each of the sub-networks. First, typical estimates for each of the PK parameters were obtained by dividing the prediction of each neural network to its prediction for the median covariate value (using typical clearance, CL_{TV} , as an example):

$$CL_{TV} = \frac{f_1(x_1)}{f_1(\text{Med}[x_1])} \cdot \frac{f_2(x_2)}{f_2(\text{Med}[x_2])} \quad (\text{s14})$$

Here f_1 represents the sub-network for the effect of weight (or fat-free mass), while f_2 represents the effect of VWF:Ag. We chose to use a value of 60 kg for fat-free mass, and 100% for VWF:Ag. Each model in the deep ensemble produces estimates of the typical value for the PK parameters. This way the prediction from each neural network are anchored to 1 at the median values of the covariates, similar to how

covariates are implemented in non-linear mixed effects models. After calculation of the typical PK parameter estimates we can investigate the variance of these values over replicates to determine their uncertainty.

Predictions from each sub-network divided by their prediction at the median covariate value can then be evaluated at any value of the covariate. We can thus visualise model predictions along the entire covariate space in order to obtain the visualisations.

6.c.4 Parameter initialisation

Model parameters were randomly drawn from initial guess distributions at the start of optimisation for each training replicate. Since the three optimisation algorithms share the same parameters ($\Theta = \{w, \Omega, \Sigma\}$), the same initial guess distributions were used.

- Neural network parameters w were initialised using Xavier initialisation: $w \sim \text{Uniform}(-x, x)$ where $x = \sqrt{(6/(in + out))}$ where in and out reference the number of input neurons and output neurons for that layer, respectively.
- Covariance matrix Ω is used in the prior distribution over the random effects $\eta \sim \mathcal{N}(\mathbf{0}, \Omega)$, and was decomposed into marginal standard deviations ω and correlation coefficient ρ . The following distributions were used for these parameters: $\omega \sim \text{Normal}(0.1, 0.025)$ truncated at [0, Inf] and $\rho \sim \text{Normal}(0, 0.1)$.
- Covariance matrix Σ represents the estimates of residual error. The initial distribution for additive error was sampled from $\sigma \sim \text{Normal}(0.1, 0.025)$ truncated at [0, Inf]. The same distribution was used to sample the initial proportional error estimate whenever applicable.

The same initial guess distributions were used for the simulation and real-world experiments. Only additive error was used in the simulation experiment.

6.c.5 MCMC model

We first compared the accuracy of posterior distributions over the random effects η obtained through MCMC and VI. We evaluated the

different approaches in two settings: (1) the ground truth parameters used in the simulation were known (i.e. we only estimate posterior distributions over the random effects) and (2) only the typical PK parameters were known (i.e. also estimate posterior distributions over the population parameters Ω and Σ). For the MCMC model, we fit a single chain to the data. Since this problem was relatively simple, the model converged well and multiple chains were not required. Pseudo-code representing the probabilistic models are shown in listing 6.1.

The following hyper-priors were used in setting 2:

$S \sim \text{LogNormal}(-1.5, 1)$: marginal standard deviations of Ω .
 $\rho \sim \text{Beta}(2, 2)$: correlation coefficient in Ω .
 $\sigma \sim \text{LogNormal}(-3, 1)$: additive error.

Listing 6.1: Pseudo-code for MCMC models.

```

using Turing

@model function model(zeta, omega, sigma, y) # setting (1).
    eta ~ MultivariateNormal(zeros(2), omega)
    z = zeta .* exp(eta) # individual estimates of PK
    parameters
    yhat = solve_ode(z)
    y ~ MultivariateNormal(yhat, sigma)
end

@model function model(zeta, y) # setting (2).
    sigma ~ LogNormal(-3, 1)
    S ~ LogNormal(-1.5, 1)
    rho ~ Beta(2, 2)
    C = [1 rho; rho 1]
    omega = S * C * transpose(S)
    eta ~ MultivariateNormal(zeros(2), omega)
    z = zeta .* exp(eta)
    yhat = solve_ode(z)
    y ~ MultivariateNormal(yhat, sigma)
end

```

Part III

APPLYING MACHINE LEARNING TO IMPROVE TREATMENT OF HAEMOPHILIA A

A GENERATIVE AND CAUSAL PHARMACOKINETIC MODEL FOR FACTOR VIII IN HAEMOPHILIA A: A MACHINE LEARNING FRAMEWORK FOR CONTINUOUS MODEL REFINEMENT

Alexander Janssen, Louk Smabil, Frank C. Bennis, Marjon H. Cnossen, and Ron A.A. Mathôt

Clinical Pharmacology and Therapeutics 115(4) (2024): 881-889.

ABSTRACT

In rare diseases, such as haemophilia 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 haemophilia treatment centres are combined but this is generally difficult as patient data are kept private. In this work, we utilise 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.

7.1 INTRODUCTION

Haemophilia A is an X-linked recessive bleeding disorder caused by a deficiency or dysfunction of the blood clotting factor VIII (FVIII). Severe haemophilia A (endogenous FVIII activity level <1% or <1 IU/dL) are at increased risk of prolonged bleeding, significant morbidity, and reduced quality of life. Personalised prophylaxis involving the administration of exogenous FVIII is the cornerstone of haemophilia 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 inter-individual variability in the PK of FVIII makes accurately predicting FVIII concentration-time profiles challenging [1, 2].

Population PK modelling has emerged as a valuable tool for characterising 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 centre 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 sub-populations, 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 haemostatic 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:

$$X \rightarrow Z \rightarrow Y \quad (7.1)$$

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 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 haemophilia A, but the proposed framework of combining causal inference, generative models, and ML-based population PK modelling can of course be applied to other problems.

7.2 METHODS

7.2.1 *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 (paediatric) haematologists (see Figure 7.2.1). 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 (V1) acts through unobserved factors U , which could, for example, represent plasma volume. We hypothesised 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 inter-compartmental 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, Ionoctocog alfa levels are known to be underestimated by roughly twofold when using the one-stage assay [13]. We first fit

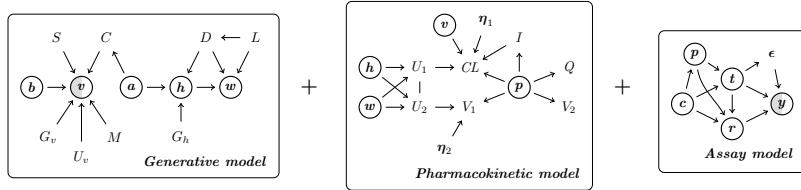


Figure 7.2.1: Directed acyclic graphs describing covariate relationships.

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 centre; CL , clearance; V_1 , volume of distribution; Q , inter-compartmental clearance; V_2 , peripheral volume of distribution; 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.

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.

7.2.2 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 optimising 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 haemophilia 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.

7.2.3 *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 ageing 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 and are not suited for data imputation.

7.2.4 *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.

7.2.5 Data set specific generative model

A generative model was developed for the data from the Ionoctocog 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.

7.2.6 Model evaluation

Accuracy of the generative model with covariate-focus was evaluated using the Ionoctocog 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 Ionoctocog 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 data set of FVIII PK profiles collected for patients with moderate and severe haemophilia 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).

7.2.7 *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>.

7.3 RESULTS

An overview of the patient characteristics for the Ionoctocog alfa data set and the OPTI-CLOT data set are shown in Table 7.3.1. 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 Ionoctocog 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 7.2.1. 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%.

	TRAINING DATA		OVERALL	TEST DATA	
	LONOCTOCOG ALFA	OCTOCOG ALFA		OCTOCOG ALFA	TUROCTOCOG ALFA
<i>n</i>	103	27	40	19	21
Age in years	26 [1–60]	32 [19–60]	49 [18–77]	48 [18–77]	49 [21–77]
Height in cm	172 [84–194]	178 [163–190]	182 [148–198]	183 [143–195]	179 [170–198]
Weight in kg	68 [12–112]	77 [59–100]	89 [61–134]	88 [61–133]	95 [63–134]
BMI	21 [13–37]	25 [19–30]	27 [19–43]	27 [19–36]	27 [21–43]
Fat-free mass in kg	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 (% 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 centre specific (unspecified)		

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

Table 7.3.1: Patient characteristics. Medians and [ranges] are shown.

Learned functions could be visualised and matched expectations about the causal effect of the covariates (see Figure 7.3.1). Investigations on alternative hypotheses supported the proposed final model (see Table S1).

Next, the conversion model was created to adjust individual lonoc tocog alfa PK parameters to octocog alfa PK parameters. Estimated *CL* of octocog alfa was increased by 15% (95% credible interval (Crl): 13–17), *V*₁ was decreased by 19% (95% Crl: 16–23), *Q* was decreased by 74% (95% Crl: 49–83), and *V*₂ was 223% higher (95% Crl: 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:

$$\text{osa} = \max \left(0, \frac{-3.06 + 4.76 \cdot \text{csa}^{0.66}}{2.10^{\text{Lonoctocog alfa}}} \right) \quad (7.2)$$

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

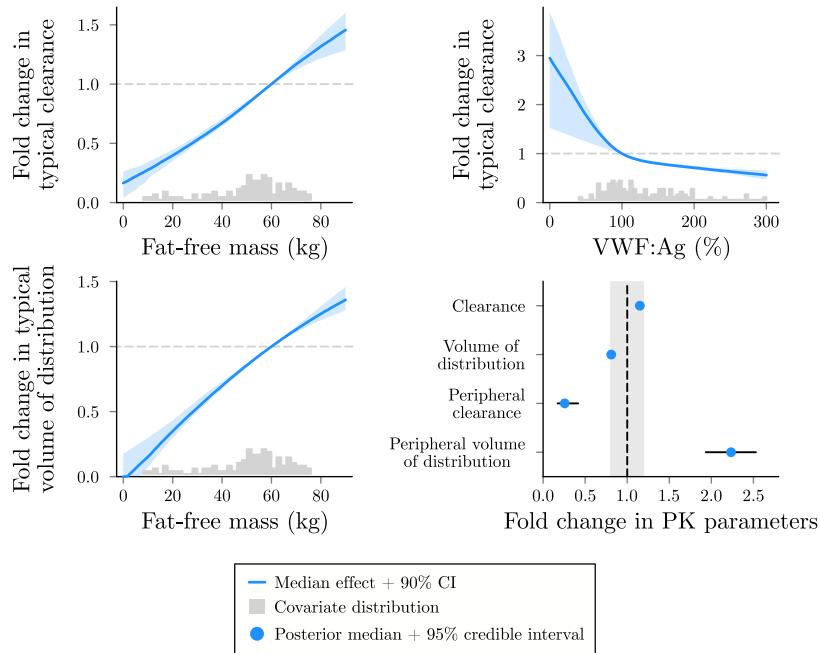


Figure 7.3.1: Visualisations of learned covariate effects. Each line depicts the median effect over the predictions from the deep ensemble, along with its 90% CI. Histograms represent the distribution of the observed covariates. In the bottom right, the median and its 95% credible interval from the posterior distributions of the difference in PK parameters between lonootocog 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.

Table 7.3.2), 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 7.3.3). The Bayesian 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

MODEL	TRAINING DATA	RMSE OF TYPICAL PREDICTIONS (IU / DL)	ME OF TYPICAL PREDICTIONS (IU / DL)	R^2
Björkman <i>et al.</i> [1]	Octocog alfa + PD-FVIII	16.6	3.85	0.87
Nesterov <i>et al.</i> [21]	Octocog alfa	17.6	3.82	0.85
McEneny-King <i>et al.</i> [22]	Octocog alfa + other SHL	19	3.76	0.86
Allard <i>et al.</i> [23]	Octocog alfa + other SHL	15.4	1.13	0.89
Causal DCM (ours)	Lonootcog alfa	14.6	1.5	0.9

Abbreviations: DCM = deep compartment model, FVIII = factor VIII, PD = plasma-derived, ME = mean error, PK = pharmacokinetic, R^2 = coefficient of determination, RMSE = root mean squared error, SHL = standard half-life.

Table 7.3.2: Accuracy of population PK models. Root mean squared error, mean error, and coefficient of determination for each of the models on the test set are shown.

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%).

7.4 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 lonootcog alfa levels measured using the chromogenic assay. Because part of the patients in the data set also received octocog alfa shortly before their lonootcog alfa PK profile was taken, the

SCENARIO	APPROACH	MAPE (%) \pm SD				VWF:AG (ASSUMED BG)
		HEIGHT	WEIGHT	FFM		
VWF:Ag (and blood group) missing	a priori	-	-	-		30.0 \pm 25 (non-O) 32.1 \pm 19 (O)
	Bayesian	-	-	-		17.6 \pm 14
All PK model covariates missing	a priori	4.3 \pm 3.4	25.5 \pm 24	16.8 \pm 16		30.0 \pm 25 (non-O) 32.1 \pm 19 (O)
	Bayesian	3.9 \pm 3.1	22.4 \pm 22	14.7 \pm 14		17.9 \pm 15

Abbreviations: BG = blood group, FFM = fat-free mass, MAPE = mean absolute percentage error, PK = pharmacokinetic, SD = standard deviation, VWF:Ag = von Willebrand factor antigen.

Table 7.3.3: Accuracy of the generative model. 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.

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 Ionoctocog 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 generalisation 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 centres (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], utilising 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 paediatric 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 utilising causal inference techniques to predict FVIII levels in patients with haemophilia 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.

REFERENCES

- [1] Sven Björkman, MyungShin Oh, Gerald Spotts, Phillip Schroth, Sandor Fritsch, Bruce M Ewenstein, Kathleen Casey, Kathelijn Fischer, Victor S Blanchette, and Peter W Collins. "Population pharmacokinetics of recombinant factor VIII: the relationships of pharmacokinetics to age and body weight". In: *Blood, The Journal of the American Society of Hematology* 119.2 (2012), pp. 612–618.
- [2] Peter L Turecek, Jill M Johnsen, Steven W Pipe, James S O'Donnell, and iPATH Study Group. "Biological mechanisms underlying inter-individual variation in factor VIII clearance in haemophilia". In: *Haemophilia* 26.4 (2020), pp. 575–583.
- [3] Tine MHJ Goedhart, Laura H Bukkems, C Michel Zwaan, Ron AA Mathôt, Marjon H Cnossen, OPTI-CLOT study group, and SYMPHONY consortium. "Population pharmacokinetic modeling of factor concentrates in hemophilia: an overview and evaluation of best practice". In: *Blood advances* 5.20 (2021), pp. 4314–4325.
- [4] Jing Zhu, Yi Shuan Wu, Ryan J Beechinor, Ryan Kemper, Laura H Bukkems, Ron AA Mathôt, Marjon H Cnossen, Daniel Gonzalez, Sheh-Li Chen, Nigel S Key, et al. "Pharmacokinetics of perioperative FVIII in adult patients with haemophilia A: An external validation and development of an alternative population pharmacokinetic model". In: *Haemophilia* 27.6 (2021), pp. 974–983.
- [5] Tim Preijers, Ri Liesner, Hendrika CAM Hazendonk, Pratima Chowdary, Marïëtte HE Driessens, Dan P Hart, Britta AP Laros-van Gorkom, Felix JM van der Meer, Karina Meijer, Karin Fijnvandraat, et al. "Validation of a perioperative population factor VIII pharmacokinetic model with a large cohort of pediatric hemophilia a patients". In: *British Journal of Clinical Pharmacology* 87.11 (2021), pp. 4408–4420.
- [6] Laura H Bukkems, Jessica M Heijdra, Mary Mathias, Peter W Collins, Charles RM Hay, Robert C Tait, Sarah Mangles, Bethan Myers, G Evans, Benjamin Bailiff, et al. "A novel, enriched population pharmacokinetic model for recombinant factor VIII-Fc fusion protein concentrate in hemophilia A patients". In: *Thrombosis and haemostasis* 120.05 (2020), pp. 747–757.
- [7] James A Rogers, Hugo Maas, and Alejandro Pérez Pitarch. "An introduction to causal inference for pharmacometrists". In: *CPT: Pharmacometrics & Systems Pharmacology* 12.1 (2023), pp. 27–40.
- [8] Eugenia Biguzzi, Filippo Castelli, Willem M Lijfering, Suzanne C Cannegieter, Jeroen Eikenboom, Frits R Rosendaal, and Astrid van Hylckama Vlieg. "Rise of levels of von Willebrand factor and factor VIII with age: role of genetic and acquired risk factors". In: *Thrombosis Research* 197 (2021), pp. 172–178.
- [9] Soracha E Ward, Jamie M O'Sullivan, and James S O'Donnell. "The relationship between ABO blood group, von Willebrand factor, and primary hemostasis". In: *Blood, The Journal of the American Society of Hematology* 136.25 (2020), pp. 2864–2874.
- [10] A Kahlon, J Grabell, A Tuttle, D Engen, W Hopman, D Lillicrap, and P James. "Quantification of perioperative changes in von Willebrand factor and factor VIII during elective orthopaedic surgery in normal individuals". In: *Haemophilia* 19.5 (2013), pp. 758–764.

- [11] Iris van Moort, Laura H Bikkens, Jessica M Heijdra, Roger EG Schutgens, Britta AP Laros-van Gorkom, Laurens Nieuwenhuizen, Felix JM van der Meer, Karin Fijnvandraat, Paula Ypma, Moniek PM de Maat, et al. "Von willebrand factor and factor VIII clearance in perioperative hemophilia A patients". In: *Thrombosis and haemostasis* 120.07 (2020), pp. 1056–1065.
- [12] Hesham Saleh Al-Sallami, Ailsa Goulding, Andrea Grant, Rachael Taylor, Nicholas Holford, and Stephen Brent Duffull. "Prediction of fat-free mass in children". In: *Clinical pharmacokinetics* 54 (2015), pp. 1169–1178.
- [13] K St Ledger, A Feussner, U Kalina, C Horn, HJ Metzner, D Bensen-Kennedy, N Blackman, A Veldman, A Stowers, and KD Friedman. "International comparative field study evaluating the assay performance of AFSTYLA in plasma samples at clinical hemostasis laboratories". In: *Journal of Thrombosis and Haemostasis* 16.3 (2018), pp. 555–564.
- [14] Alexander Janssen et al. "Deep compartment models: a deep learning approach for the reliable prediction of time-series data in pharmacokinetic modeling". In: *CPT: Pharmacometrics & Systems Pharmacology* 11.7 (2022), pp. 934–945.
- [15] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *Advances in neural information processing systems* 30 (2017).
- [16] A Janssen, FWG Leebeek, MH Cnossen, and RAA Mathôt. "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.* Vol. 9826. 2021.
- [17] Centers for Disease Control, Prevention (CDC), and National Center for Health Statistics (NCHS). *National Health and Nutrition Examination Survey Data*. Hyattsville, MD: U.S. Department of Health, Human Services, Centers for Disease Control, and Prevention, 2009. doi: <https://www.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2009>.
- [18] James Anthony Davies, Peter William Collins, Lee Sarah Hathaway, and Derrick John Bowen. "Effect of von Willebrand factor Y/C1584 on in vivo protein level and function and interaction with ABO blood group". In: *Blood* 109.7 (2007), pp. 2840–2846.
- [19] Iris van Moort, Tim Preijers, Laura H Bikkens, Hendrika CAM Hazendonk, Johanna G van der Bom, Britta AP Laros-van Gorkom, Erik AM Beckers, Laurens Nieuwenhuizen, Felix JM van der Meer, Paula Ypma, et al. "Perioperative pharmacokinetic-guided factor VIII concentrate dosing in haemophilia (OPTI-CLOT trial): an open-label, multicentre, randomised, controlled trial". In: *The Lancet Haematology* 8.7 (2021), e492–e502.
- [20] D Viuff, TW Barrowcliffe, T Saugstrup, M Ezban, and D Lillicrap. "International comparative field study of N8 evaluating factor VIII assay performance". In: *Haemophilia* 17.4 (2011), pp. 695–702.
- [21] Ivan Nestorov, Srividya Neelakantan, Thomas M Ludden, Shuanglian Li, Haiyan Jiang, and Mark Rogge. "Population pharmacokinetics of recombinant factor VIII Fc fusion protein". In: *Clinical pharmacology in drug development* 4.3 (2015), pp. 163–174.

- [22] Alanna McEneny-King, Pierre Chelle, Gary Foster, Arun Keepanasseril, Alfonso Iorio, and Andrea N Edginton. "Development and evaluation of a generic population pharmacokinetic model for standard half-life factor VIII for use in dose individualization". In: *Journal of Pharmacokinetics and Pharmacodynamics* 46 (2019), pp. 411–426.
- [23] Quentin Allard, Zoubir Djerada, Claire Pouplard, Yohann Repessé, Dominique Desprez, Hubert Galinat, Birgit Frotscher, Claire Berger, Annie Harroche, Anne Ryman, et al. "Real life population pharmacokinetics modelling of eight factors VIII in patients with severe hemophilia A: is it always relevant to switch to an extended half-life?" In: *Pharmaceutics* 12.4 (2020), p. 380.
- [24] Christopher Rackauckas and Qing Nie. "Differentialequations. jl—a performant and feature-rich ecosystem for solving differential equations in julia". In: *Journal of open research software* 5.1 (2017), pp. 15–15.
- [25] Emmanuel J Favaloro, Massimo Franchini, and Giuseppe Lippi. "Aging hemostasis: changes to laboratory markers of hemostasis as we age—a narrative review". In: *Seminars in thrombosis and hemostasis*. Vol. 40. 06. Thieme Medical Publishers. 2014, pp. 621–633.
- [26] Ferdows Atiq, Karina Meijer, Jeroen Eikenboom, Karin Fijnvandraat, Eveline P Mauser-Bunschoten, Karin PM van Galen, Marten R Nijziel, Paula F Ypma, Joke de Meris, Britta AP Laros-van Gorkom, et al. "Comorbidities associated with higher von Willebrand factor (VWF) levels may explain the age-related increase of VWF in von Willebrand disease". In: *British journal of haematology* 182.1 (2018), pp. 93–105.
- [27] Iris van Moort, Tim Preijers, Hendrika CAM Hazendonk, Roger EG Schutgens, Britta AP Laros-van Gorkom, Laurens Nieuwenhuizen, Felix JM van der Meer, Karin Fijnvandraat, Frank WG Leebeek, Karina Meijer, et al. "Dosing of factor VIII concentrate by ideal body weight is more accurate in overweight and obese hemophilia A patients". In: *British journal of clinical pharmacology* 87.6 (2021), pp. 2602–2613.
- [28] João A Abrantes, Elisabet I Nielsen, J Korth-Bradley, L Harnisch, and Siv Jönsson. "Elucidation of factor VIII activity pharmacokinetics: a pooled population analysis in patients with hemophilia A treated with moroctocog alfa". In: *Clinical Pharmacology & Therapeutics* 102.6 (2017), pp. 977–988.
- [29] Hendrika Hazendonk, Karin Fijnvandraat, Janske Lock, Mariëtte Driessens, Felix Van Der Meer, Karina Meijer, Marieke Kruip, Britta Laros-van Gorkom, Marjolein Peters, Saskia de Wildt, et al. "A population pharmacokinetic model for perioperative dosing of factor VIII in hemophilia A patients". In: *haematologica* 101.10 (2016), p. 1159.
- [30] C Pouplard, C Caron, MF Aillaud, C Ternisien, C Desconclois, A Dubanchet, and F Sobas. "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". In: *Haemophilia* 17.5 (2011), e958–e962.
- [31] Joaquim F Monteiro, Siomara R Hahn, Jorge Gonçalves, and Paula Fresco. "Vancomycin therapeutic drug monitoring and population pharmacokinetic models in special patient subpopulations". In: *Pharmacology research & perspectives* 6.4 (2018), e00420.

APPENDIX

7.A SUPPLEMENTARY TABLES AND FIGURES

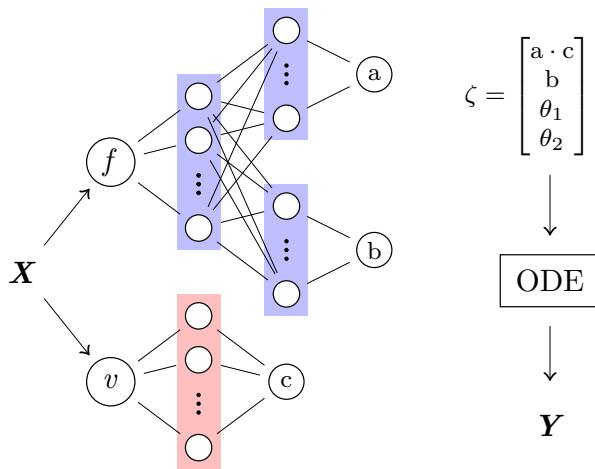


Figure 7.A.1: Schematic overview of hybrid ML/PK model architecture. Hidden layers are indicated by filled rectangles. Output of the independent neural networks is combined using a product to produce the typical PK parameter estimates. X = covariate matrix, f = fat-free mass, v = VWF:Ag, a, b, c = output of neural networks, ζ = typical PK parameters [CL, V₁, Q, V₂], ODE = system of ordinary differential equations, Y = dependent variable.

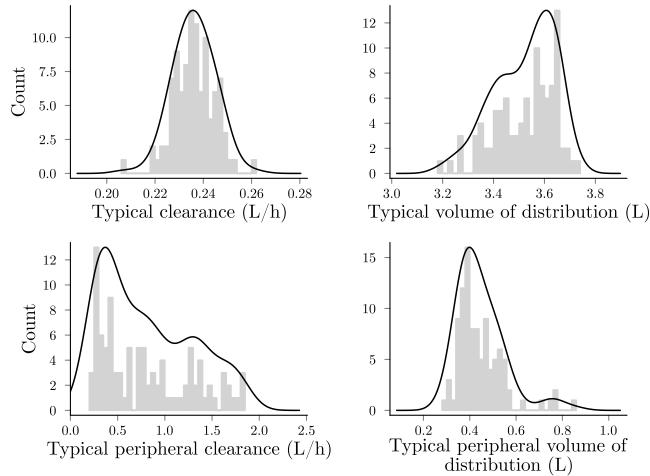


Figure 7.A.2: Uncertainty over typical PK parameter estimates in the deep ensemble. Black line represents the kernel density estimate of the histogram.

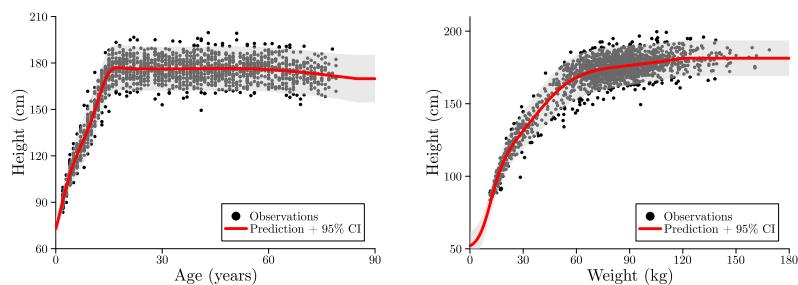


Figure 7.A.3: Generative models based on neural networks. Left: generative model for height based on patient age. Right: model for generating height based on body weight.

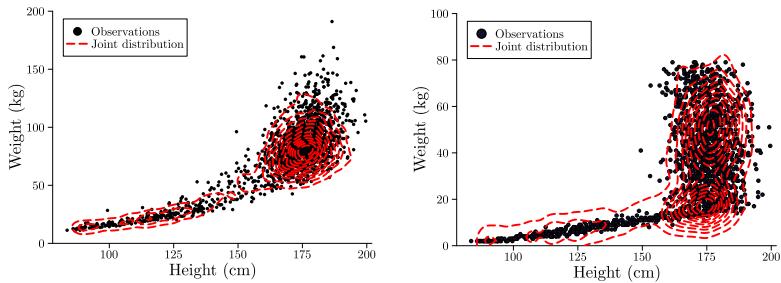


Figure 7.A.4: Generative models based on neural spline flows. Joint distributions for body weight and height (left) and age and height (right).

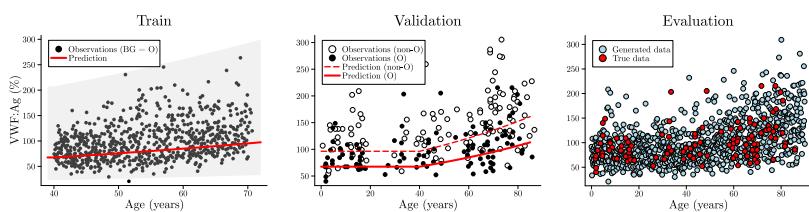


Figure 7.A.5: Generative models for VWF. From left to right: model fit on the training data, adjustment of the model based on the validation data, and evaluation of the model by comparing simulated values to true VWF:Ag..

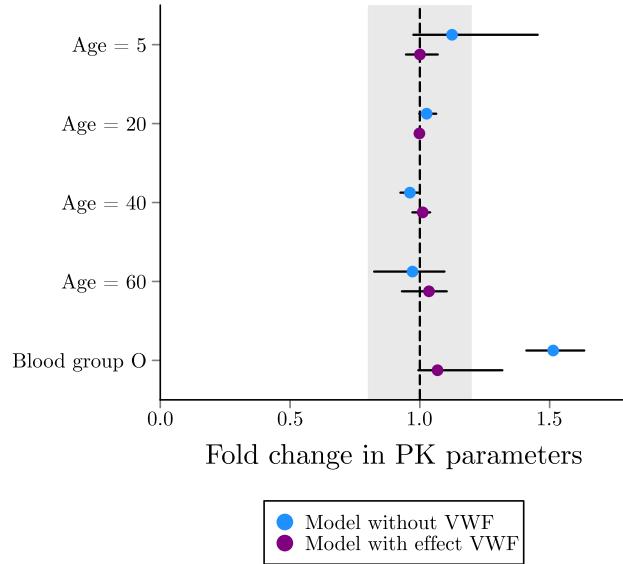


Figure 7.A.6: Assessment of the effect of age and blood group in two alternative hypotheses. Horizontal bars indicate 95% confidence interval. Shaded region covers a 20% change (in both directions) of the typical PK parameter. Covariate effects with effects outside of this range are often associated with clinically relevant effects.

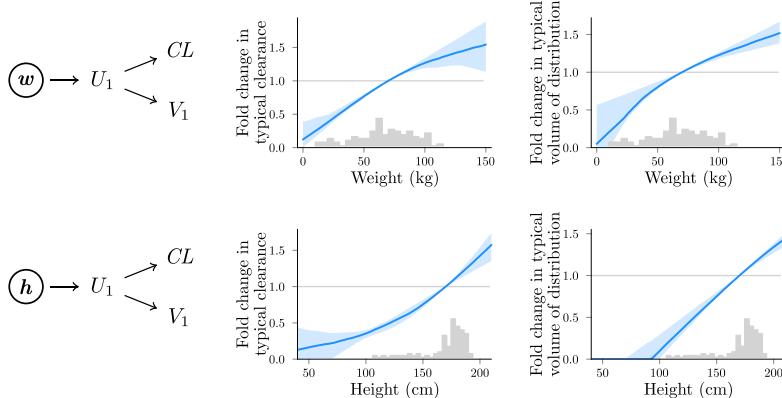


Figure 7.A.7: Comparison of the learned effects in alternative hypotheses. CL = clearance, V_1 = volume of distribution, w = body weight, h = height, U = latent variable.

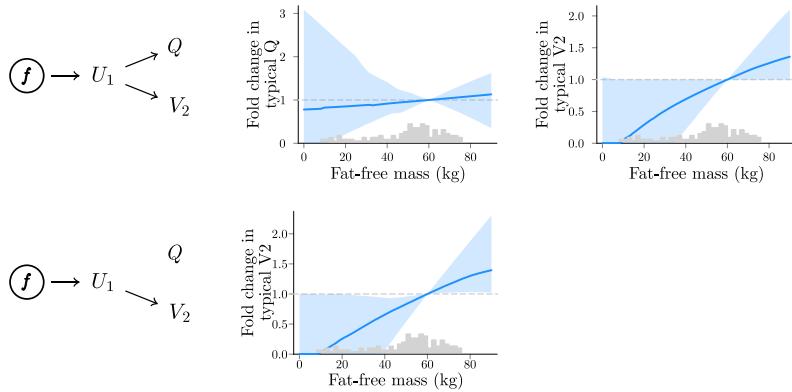


Figure 7.A.8: Alternative hypotheses with respect to the inclusion of fat-free mass (*f*) on inter-compartmental clearance (Q) and peripheral volume of distribution (V₂).

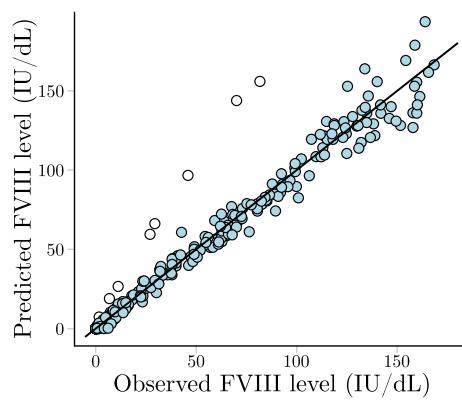


Figure 7.A.9: Goodness-of-fit plot for rFVIII predictions after product conversion correction. White filled dots represent the predictions for the individual for whom the learned conversion from Ionoctocog alfa to Octocog alfa was not accurate, whereas coloured dots represent the predictions for the other patients.

7.B MODEL DEVELOPMENT

7.B.1 Deep compartment model

A deep compartment model was fit to predict lonoctocog alfa levels based on patient fat-free mass and VWF:Ag levels. We use a specific architecture where each covariate was linked to specific PK parameters and effects are combined using a proportional model. In this architecture, independent models are fit to learn the effect of each of the covariates (or combinations thereof) allowing for the interpretation of learned effect by visualising the output of each sub-model. As a result, the model is fully interpretable. Two independent neural networks were fit, the first learning the effect of fat-free mass on clearance (CL) and volume of distribution (V_1), and the second learning the effect of VWF on clearance. Predictions from these models were then lower bounded at 0.01 to prevent predictions too close to zero and combined using a product to produce the PK parameter estimates. Adjusted softplus activation functions were used as a smooth alternative to the relu in all layers except output layers:

$$\pi(x) = \frac{1}{\beta} \cdot \text{softplus}(\beta \cdot x) \quad (7.3)$$

Using $\beta = 10$.

A schematic overview of the model is provided in figure S1. In the first network, a shared hidden layer with 12 neurons followed by two separated heads were used to learn the effect of fat-free mass on CL and V_1 . For the second neural network a single hidden layer with 6 neurons was used, feeding into a single output neuron. Model inputs were normalised by dividing fat-free mass inputs by 90 and VWF:Ag inputs by 300. Model predictions (of CL and V_1) were concatenated with a vector of two learned parameters (Q and V_2), shared between all individuals, to produce the PK parameter vector.

A deep ensemble model was fit by stochastic gradient descent using the ADAM optimiser using a learning rate of 1e-2. A Monte Carlo cross validation procedure was performed dividing the full data set in an 80% train and 20% validation set for each replicate of the deep ensemble. A total of 100 models were trained to build the deep ensemble. At each epoch, a random sample of 60 patients without replacement was taken from the train set and gradients were calculated. Models were optimised for 3000 epochs. Optimal weights during the 3000 epochs were selected based on the lowest validation set error. After

models were fit, PK parameter predictions were produced for each of the 100 model replicate in the deep ensemble and the median prediction was taken along with the 90% confidence interval for final predictions.

Median typical PK parameter predictions were then used to optimise the population parameters for the random effect variances and the residual error. For this purpose, PK parameter estimates were directly entered into a two compartment model in NONMEM and population parameters were optimised using the FOCEI objective function.

7.B.2 Visualising learned functions

Visualisations of learned functions were obtained by entering dummy input to each of the neural networks for each of the replicates of the deep ensemble. First, typical estimates for each of the PK parameters were obtained by dividing the prediction of each neural network to its prediction for the median covariate value (using CL_{TV} as an example):

$$CL_{TV} = \frac{f_1(x_1)}{f_1(\text{Med}[x_1])} \cdot \frac{f_2(x_2)}{f_2(\text{Med}[x_2])} \quad (7.4)$$

We chose to use a value of 60 kg for fat-free mass, and 100% for VWF:Ag. Each model in the deep ensemble produces estimates of the typical value for the PK parameters.

This way the prediction from each neural network are anchored to 1 at the median values of the covariates, similar to how covariates are implemented in NONMEM. This removes the issue of complex interplay's between the neural networks potentially inflating the variance of the learned functions. For example, if the prediction of f_1 is very low after random initialisation of the network, the model can still produce similar PK parameter predictions to other replicates by increasing the magnitude of predictions from f_2 . This can greatly inflate the variance of the learned functions when naively calculating quantiles.

After calculation of the typical PK parameter estimates we can investigate the variance of these values to determine uncertainty in the PK parameter estimates. Results for the final model used in the manuscript (with fat-free mass on CL and V_1 and VWF:Ag on CL) is shown in figure S2. Here we see that the typical predictions of CL , V_1 , and peripheral volume of distribution (V_2) depict relatively low variance between replicates. Predictions for inter-compartmental

clearance (Q) show higher variance, suggesting poorer identifiability of this parameter.

Predictions from each neural network divided by their prediction at the median covariate value can then be evaluated at any value of the covariate. We can thus visualise model predictions along the entire covariate space in order to obtain the visualisations. The results are shown in figure 2 of the main manuscript.

7.C REPRESENTING RELATIONSHIPS IN THE DAG USING GENERATIVE MODELS

Neural networks with two hidden layers (16 or 24 neurons) were used to predict the mean and the variance of height conditional on either body weight or age. Neural networks were fit for 5000 epochs using the ADAM optimiser. Model and hyper-parameter (neurons etc.) selection was based on visual inspection of the learned functions. After fitting the models, model output was constrained in order to improve extrapolation to unseen data. This was achieved by fixing the predicted distribution over height for individuals with weights above 120 kg or those above the age of 85. Model output was similarly constrained between 40 and 210cm by using the following activation function in the final layer of the model:

$$y = \text{sigmoid}(x) \cdot (210 - 40) + 40 \quad (7.5)$$

Figures showing the predictions of the final models are shown in figure S3.

For the inverse relationships, i.e. predicting either body weight or age from height, normalising flows (NF) models were fit. NF is a technique to learn complex distributions by performing invertible transformations of a random variable with a known density (usually a standard normal). By using an invertible mapping, the probability density function of the transformed density can be computed exactly. This approach was chosen as the shape of the distribution of weight and age was dependent on the value of height. For example, at lower height (i.e. for children) the weight distribution more closely resembled a normal distribution, whereas at higher height (e.g. adults) the weight distribution was more similar to a log normal distribution. The chosen approach involved the use of a neural network to predict the parameters for a neural spline flows model based on height. The resulting joint distributions are shown in figure S4.

Finally, generative models for describing the joint distribution over age, blood group, and VWF:Ag were produced as follows. First, a linear model was fit to predict 645 log transformed VWF:Ag levels collected from figures presented in *Biguzzi et al.* based on age. The mean effect of having blood group O on the mean VWF:Ag level was inferred from a previous study. The subsequent model was then validated on the remaining 225 levels. Based on the validation data, the coefficient for age in the linear model was reduced by a factor of 0.8 and was fixed for individuals below the age of 40. Finally, a log normal distribution was fit to all VWF:Ag levels normalised by age and blood group (e.g. using equation 4), assuming homoscedastic noise with respect to age and blood group (see figure S5 for results). The following equation adequately described the VWF:Ag levels:

$$VWF = \epsilon \cdot \exp \left(4.11 + 0.515 \cdot \max \left(\frac{\text{age}}{45}, \frac{40}{45} \right) \cdot 0.7^{\text{bg}=0} \right) \quad (7.6)$$

Where $\epsilon \sim \text{LogNormal}(\mu = 0.158, \sigma = 0.348)$. Final evaluation of the model involved comparing synthetic data to observed VWF:Ag levels per blood group, which closely matched (see figure S5).

7.D EVALUATION OF ALTERNATIVE HYPOTHESES IN THE DAG

The correctness of the DAG given the data was evaluated by comparing model performance using different versions of the graph.

The absence of an independent causal effect of age and blood group was evaluated by first fitting a model using fat-free mass, age and blood group and visualising the learned functions. The resulting functions indicated low importance of patient age, but an almost 50% increase in clearance for individuals with blood group O compared to those who did not. Next, we took PK parameters from the final model (containing the effect of fat-free mass and VWF:Ag) and attempted to add the effect of age and blood group. Compared to the model without VWF:Ag, the learned effects for age and blood group were greatly diminished. In figure S6 we show a forest plot depicting the covariate effects in both models. The median difference in CL for individuals with blood group O was now only 6%, including 0% in its 95% confidence interval. For all ages, the previous effect was completely eliminated. Although this does not provide conclusive evidence, the results do point in the direction of both covariates acting as mediators of VWF levels rather than having significant independent causal effects.

Next we investigated whether patient weight, height, or fat-free mass were better predictors of clearance. The learned functions for these models are shown in figure S7 and corresponding accuracy in supplementary table 1. We see that the model using fat-free mass performs similarly well as the model using weight as covariate, but that the variance of the learned functions for weight are higher. Additionally, at higher weight we can see the learned function flatten, roughly starting at weights above 90 kg. Since most individuals above this weight are likely at the higher spectrum of the BMI range, this could indicate that the model is implicitly learning the effect of lean body weight. Adding that the learned function for the effect of weight depict higher uncertainty, it might be more appropriate to use fat-free mass.

We can also have the model learn the appropriate combination of weight and height, which was tested in the last hypothesis. This model very slightly improves validation set error compared to the other models. However, combining the two covariates in a single model complicates its interpretation. In this case, the slight increase in model accuracy does not weight up against the decrease in interpretability.

Finally, we would expect that the height of a patient correlates reasonably well to variables such as lean body mass. However, the high variance of the learned functions as well as the high validation set error for the model that uses height seem to indicate that the models cannot learn its effect well. Additionally, the absence of patients in the 50 – 100 cm range of height will likely result in the model having poor extrapolation capabilities, especially seeing as the learned function for V_1 essentially degrades towards zero for this range.

Additional hypotheses that were tested are related to including the effect of fat-free mass on Q and V_2 . The accuracy of the different models are also shown in supplementary table 1, and the learned functions are shown in figure S8. Again, although the addition of the effect of fat-free mass on Q and V_2 suggest a slight improvement in accuracy, the uncertainty over the learned effects is very high.

Part IV

THE OPTI-CLOT WEB-PORTAL

INDIVIDUALISED TREATMENT WITH A PERSONAL TOUCH: INTRODUCING THE OPTI-CLOT WEB-PORTAL, AN OPEN WEB-APPLICATION IMPLEMENTING PK-GUIDED DOSING IN PATIENTS WITH RARE BLEEDING DISORDERS

Alexander Janssen, Frank C. Bennis, Ron A.A. Mathôt, and Marjon H. Cnossen

In preparation.

ABSTRACT

We introduce the OPTI-CLOT portal, a web-application aiming to facilitate the adoption of pharmacokinetic (PK)-guided dosing to personalise treatment for people living with rare bleeding disorders. One of the most important barriers for the adoption of PK-guided dosing is its accessibility. Most centres do not have access to the expertise required to perform and analyse PK profiles. The OPTI-CLOT portal (<https://opticlot.com/>) aims to resolve this issue by giving all hospitals in the Netherlands access to pharmacometric experts specialised in giving dosing advice to caregivers of people with rare bleeding disorders. We discuss the design of the portal, its adoption, and future aims.

10.1 BACKGROUND

Haemophilia and von Willebrand disease are a rare bleeding disorder characterised by a deficiency or qualitative defect in coagulation factor VIII (FVIII; haemophilia A), IX (FIX; haemophilia B), or von Willebrand factor (VWF, von Willebrand disease). Patients with these bleeding disorders have an increased (spontaneous) bleeding risk, which can lead to debilitating arthropathy or life-threatening haemorrhages if not treated adequately. Hallmark of treatment is replacement of the deficient coagulation factor using factor concentrates, non-factor replacement therapy or desmopressin if applicable. Treatment is administered preventively (prophylaxis) or acutely (on demand) whenever bleeding has occurred, or when there is a high risk of bleeding, for example during and after medical procedures. Prophylactic treatment is based on the observation that patients with moderate haemophilia (i.e. those with residual endogenous factor activity levels > 1 IU/dL) have a less severe bleeding phenotype [1]. The overall aim is thus often to maintain minimal factor activity levels above 1 IU/dL. However, this is complicated by a relatively high degree of inter-individual variability in drug exposure when performing body weight-based dosing [2].

Several clinical guidelines and expert groups recommend the use of pharmacokinetic (PK)-guided dosing to personalise the treatment of patients with haemophilia [3–6]. In PK-guided dosing, a PK profile is constructed which provides individual estimates of parameters such as drug *in vivo* recovery and half-life. These parameters are subsequently used to simulate drug exposure following different treatment regimens. The optimal treatment regimen that achieves pre-specified target levels can then be selected on an individual basis in consultation with patient and treatment team. In spite of its potential benefits, more widespread adoption of PK-guided dosing in routine clinical practice might be desirable [7]. Two recent surveys have reviewed the use of PK-guided dosing when switching from standard half-life (SHL) to extended half-life (EHL) factor concentrates. They found that full PK analysis was performed by less than 10% of 70 respondents within the Subcommittee on FVIII and FIX of the Scientific and Standardisation Committee of the International Society on Thrombosis and Haemostasis and by 51% of 37 physicians from European haemophilia treatment centres [6, 8]. This is indicative of a discrepancy between recommendations by clinical guidelines and the adoption of PK-guided dosing in clinical practice.

One important barrier to tackle is accessibility to pharmacometric expertise [9, 10]. Most clinicians and haemophilia treatment centres lack the experience and expertise to perform and evaluate PK analyses. Ideally, dedicated pharmacometrists with an expertise in haematology are desirable. To facilitate the implementation PK-guided dosing within the Netherlands, work-package 6 within the SYMPHONY consortium [11] introduced the OPTI-CLOT portal (<https://opticlot.com/>), an online web application where caregivers are able to request dosing advice for patients with rare bleeding disorders. The current version of the portal is aimed at providing dosing recommendations for patients with haemophilia A, haemophilia B, or von Willebrand disease. Users can request dosing advice for prophylaxis, perioperative dosing, and dosing around bleeding events for various factor and non-factor based therapies. In contrast to other web-applications such as WAPPS-Hemo and MyPKFit, each dosing advice is personally curated by an expert pharmacometrist in consultation with the requesting caregiver [12, 13]. In addition, the portal is transparent in its use of (published) population PK models and publishes all newly developed models. The main aim of the OPTI-CLOT portal is to provide a user-friendly approach to support PK analyses. We will discuss the current design of the portal, its adoption, challenges, as well as future perspectives.

10.2 DESIGN

10.2.1 *Security and privacy*

Since the OPTI-CLOT portal works with sensitive patient information, it is crucial to ensure that proper security and privacy measures are in place. To prevent unauthorised access, the portal makes use of SURFconext, a service linking (academic) institutions to web-services. This enables users to authenticate themselves using their institutional account (i.e. not requiring new log-in information), while the portal relies on information from the institutions to identify and authorise users. Next, a two-factor authentication system is required to authenticate users during each session to prevent unauthorised access. Each session expires after a relatively short time interval when the user is inactive, subsequently requiring re-authentication to access the application. Design of the portal follows security recommendations made for software in the medical domain [14, 15], and has gone through rigorous quality control by the IT department at the Amsterdam UMC.

The privacy policy of the OPTI-CLOT web-portal is to store as little information as possible about each patient as needed for producing dosing recommendations. Each patient is given an OPTI-CLOT pseudo-ID rather than directly using actual medical health record numbers. These pseudo-IDs can be stored in the health record system of the associated institution. If the patient linked to a specific pseudo-ID is unknown, the portal allows users to view the patient's birth date for identification. Before patient information is sent to the portal, the requesting clinician must confirm that the patient has given their consent. Patients can also request the removal of their data at any point in time, and previous dosing recommendations can be downloaded from the portal and stored locally at each institution so that no information is lost.

10.2.2 Workflow

A schematic overview of the portal workflow is shown in figure 10.2.1. Several screenshots of the application are available in supplementary figures 10.A.

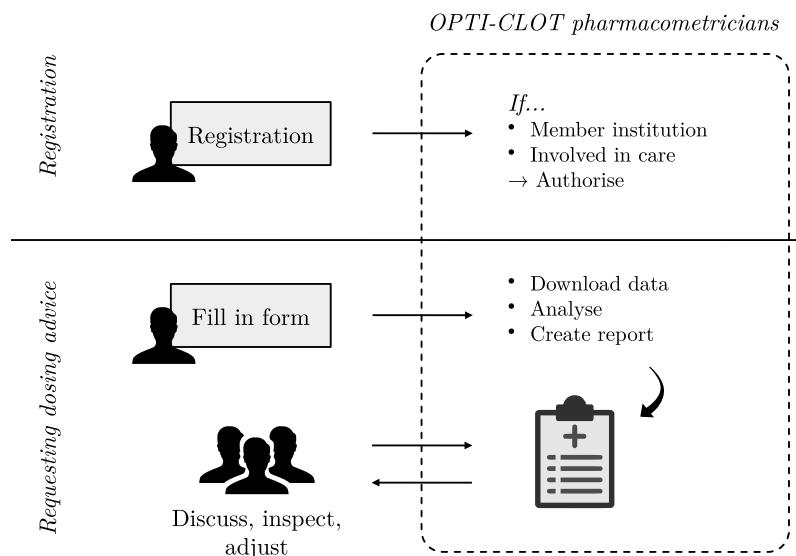


Figure 10.2.1: Schematic overview of portal workflow.

Before a new user is able to access the OPTI-CLOT portal, they must first register. After registering via the online form of the web-application, the institutional account provided by the prospective user is verified. If the user is involved in the care of patients with bleeding disorders, the account is authorised, and the user sets up a two-factor authentication method.

After completing the registration procedure, the new user has access to the previous dosing requests solicited by their institution. At this point, the new user is able to request dosing recommendation for both new and existing patients. The workflow for a new patient is as follows. The OPTI-CLOT portal contains input forms that are dynamically changed based on the entered information. For example, different information is required for patients with haemophilia A compared to those with von Willebrand disease. This reduces the burden of the user by only requiring them to fill in information relevant to the case at hand. The input form is organised into four sections: Basic patient characteristics (e.g. body weight, height and blood group), information on the current treatment (e.g. current factor concentrate, prophylaxis regimen), relevant clinical measurements (e.g. factor levels), and the desired information with regard to the dosing recommendation(s) (e.g. target levels, what to do in case of bleeding). When moving through each section, all inputs are checked for errors in order to notify the user of any problems. The procedure for existing patients is similar, with the exception that most of the information is pre-filled based on previous requests, and the user is requested to verify the information to determine if all data is still up-to-date.

After completing the request, the pharmacometrists from the OPTI-CLOT group are notified and the request is assigned to one of the analysts. The data can then be downloaded in a format specifically prepared for analysis. A secondary goal of the OPTI-CLOT portal is to streamline and standardise report generation. The final result of the analysis is a HTML file containing information on the PK profile along with several proposed treatment regimens that achieve the desired target levels. The report is interactive, and users can inspect drug levels at any time point during the week or zoom into specific areas of the produced figures. The initial report is uploaded to the portal and any member from that institution can download and inspect the advice. The requesting user is notified and can suggest any changes to the report if so desired.

Finally, the OPTI-CLOT portal has a specific environment for clinical trials that involve dose adjustments based on the PK of each

subject. One example of a clinical trial that makes use of the portal is the DosEmi trial [16]. This crossover study investigates whether a reduction in drug concentrations of emicizumab, a non-factor replacement therapy option for patients with haemophilia A, is equally effective compared to standard treatment regimens. Tailored dosing recommendations are given for each subject, and treating physicians are able to discuss patient preferences regarding the minimal dose and dosing frequency with the patient in a shared decision making process. The portal offers a dedicated input form for the study that matches the information gathered during specific visits during the trial.

10.3 ADOPTION

In figure 10.3.1, we show the number of requests for dosing recommendations for patients with a bleeding disorder made to our team. In December of 2022, we introduced the OPTI-CLOT web-portal to the members of the Dutch society of haemophilia (Nederlandse Vereniging van Hemofilie Behandelaren; NVHB), and since then, the number of requests has doubled with respect to earlier periods. We have observed a general increase in the rate of incoming dosing requests after COVID, likely due to centres attempting to clear backlogs. However, this rate has been constant even in 2024, partly due to a large volume of requests for dosing recommendations for subjects in the DosEmi trial [16]. The portal currently has 24 users from 9 institutions and is looking to expand to international partners.

10.4 PERSPECTIVES

The main benefit of using the OPTI-CLOT portal is that it streamlines both the process of requesting dosing advice as well as analysis of the received data. The produced reports are interactive, making them useful as an educative tool for both treatment teams and patients. There are several ways these reports can be further improved. For example, it might be interesting to allow users to directly make adjustments to the provided dosing recommendations within the reports. This facilitates shared decision making, especially when patients are able to directly see the effect of changes in dose timing or altering doses on drug exposure during consultation. Another way of helping patients understand the effects of their treatment is combine the PK

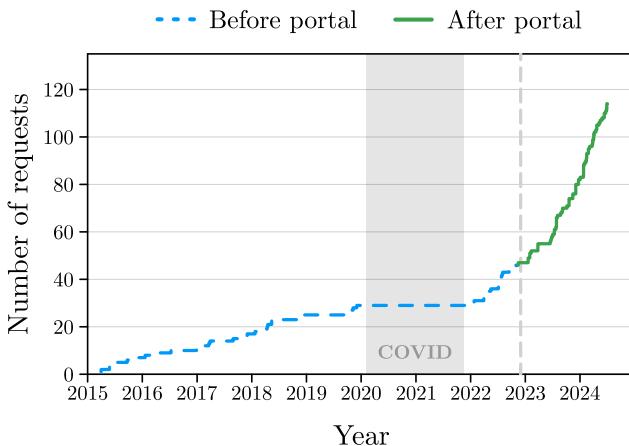


Figure 10.3.1: Number of requests for dosing recommendations over the years. Dashed line indicates the date of the introduction of the OPTI-CLOT web-portal (December 2022). Coloured area indicates time period where no requests were received, likely as a consequence of the COVID pandemic.

profile with applications that record dose administration (e.g. patient logging applications such as VastePrik©). Such applications can use the PK and dosing information to visualise real-time drug exposure. An additional benefit of this approach would be that the recording of treatment information might be improved as patients would want real-time drug levels to be accurate. This might also improve patient adherence as they can directly observe the consequences of missing treatment.

In the near future, it may also be of interest to base the selection of the optimal treatment regimen on the patients' pharmacodynamic (PD) profile. The PD profile describes the effect of the drug on the body, and in the context of bleeding disorders would describe the innate bleeding phenotype and risk of patients. One way this can be achieved is by using repeated time-to-event models [17, 18]. These models can be used to estimate the individual bleeding risk for each patient, which can be used to simulate the projected number of bleeds given a specific treatment regimen. However, PD-based methods are relatively new and clinical studies will have to be performed in order to evaluate their effectiveness and accuracy.

Finally, users of the OPTI-CLOT portal are currently required to manually enter patient information. This information often originates

from the health record system, and the action of copying this information over to another application is potentially redundant and error-prone. It is of interest to look at opportunities to link the portal to electronic health record systems to automatically extract relevant data. This lowers the burden for the requesting user, and improves the accessibility of the portal. One example of such an initiative is the Digizorg app (<https://www.digizorg.app/>), which collects the data from the electronic health record in a specific format and displays it to the user. Adding the OPTI-CLOT portal as a specific sub-module in this app for patients with a bleeding disorder allows treatment teams to automatically send relevant information to the portal such that patient data can be exchanged automatically. This application also has a patient facing side, which can for example also interface with the portal to present real-time drug exposure.

10.5 CONCLUSION

In conclusion, we present the OPTI-CLOT portal, a web-application aiming to facilitate the adoption of PK-guided dosing for patients with bleeding disorders. The application offers a simple workflow and brings physicians and pharmacometrists together to improve the treatment of this group patient population.

REFERENCES

- [1] Erik Berntorp. "Prophylactic therapy for haemophilia: early experience". In: *Haemophilia* 9 (2003), pp. 5–9.
- [2] Sven Björkman, Anna Folkesson, and Siv Jönsson. "Pharmacokinetics and dose requirements of factor VIII over the age range 3–74 years: a population analysis based on 50 patients with long-term prophylactic treatment for haemophilia A". In: *European journal of clinical pharmacology* 65 (2009), pp. 989–998.
- [3] Alok Srivastava, Elena Santagostino, Alison Dougall, Steve Kitchen, Megan Sutherland, Steven W Pipe, Manuel Carcao, Johnny Mahlangu, Margaret V Ragni, Jerzy Windyga, et al. "WFH guidelines for the management of hemophilia". In: *Haemophilia* 26 (2020), pp. 1–158.
- [4] Daniel P Hart, Davide Matino, Jan Astermark, Gerard Dolan, Roseline d'Oiron, Cédric Hermans, Victor Jiménez-Yuste, Adriana Linares, Tadashi Matsushita, Simon McRae, et al. "International consensus recommendations on the management of people with haemophilia B". In: *Therapeutic advances in hematolgy* 13 (2022), p. 20406207221085202.
- [5] Annamaria Iorio, V Blanchette, J Blatny, P Collins, K Fischer, E Neufeld, et al. "Estimating and interpreting the pharmacokinetic profiles of individual patients with hemophilia A or B using a population pharmacokinetic approach: com-

- munication from the SSC of the ISTH". In: *Journal of Thrombosis and Haemostasis* 15.12 (2017), pp. 2461–2465.
- [6] MV Ragni, SE Croteau, Massimo Morfini, MH Cnossen, A Iorio, et al. "Pharmacokinetics and the transition to extended half-life factor concentrates: communication from the SSC of the ISTH". In: *Journal of Thrombosis and Haemostasis* 16.7 (2018), pp. 1437–1441.
- [7] Stacy E Croteau, Michael U Callaghan, Joanna Davis, Amy L Dunn, Michael Guerrera, Osman Khan, Ellis J Neufeld, Leslie J Raffini, Michael Recht, Michael Wang, et al. "Focusing in on use of pharmacokinetic profiles in routine hemophilia care". In: *Research and practice in thrombosis and haemostasis* 2.3 (2018), pp. 607–614.
- [8] Marijn van der Sluijs, Nicole Huyghe, Caroline Wood, and Sally Tawil. "A survey of physicians' treatment switching practice in people on long-term prophylaxis for hemophilia in five European countries". In: *Current Medical Research and Opinion* 38.1 (2022), pp. 65–73.
- [9] Alfonso Iorio, Arun Keepanasseril, Gary Foster, Tamara Navarro-Ruan, Alanna McEneny-King, Andrea N Edginton, Lehana Thabane, et al. "Development of a web-accessible population pharmacokinetic service—Hemophilia (WAPPS-Hemo): study protocol". In: *JMIR research protocols* 5.4 (2016), e6558.
- [10] Tine MHJ Goedhart, A Janssen, Ron AA Mathôt, Marjon H Cnossen, OPTI-CLOT Study Group, SYMPHONY Consortium, et al. "The road to implementation of pharmacokinetic-guided dosing of factor replacement therapy in hemophilia and allied bleeding disorders. Identifying knowledge gaps by mapping barriers and facilitators". In: *Blood Reviews* 61 (2023), p. 101098.
- [11] Marjon H Cnossen, Iris van Moert, Simone H Reitsma, Moniek PM de Maat, Roger EG Schutgens, Rolf T Urbanus, Hester F Lingsma, Ron AA Mathot, Samantha C Gouw, Karina Meijer, et al. "SYMPHONY consortium: Orchestrating personalized treatment for patients with bleeding disorders". In: *Journal of Thrombosis and Haemostasis* 20.9 (2022), pp. 2001–2011.
- [12] Alexandros Arvanitakis, Erik Berntorp, and Jan Astermark. "A comparison of MyPKFiT and WAPPS-Hemo as dosing tools for optimizing prophylaxis in patients with severe haemophilia A treated with Octocog alfa". In: *Haemophilia* 27.3 (2021), pp. 417–424.
- [13] T Preijers, Iris van Moert, K Fijnvandraat, FWG Leebeek, MH Cnossen, RAA Mathôt, et al. "Cross-evaluation of pharmacokinetic-guided dosing tools for factor VIII". In: *Thrombosis and haemostasis* 118.03 (2018), pp. 514–525.
- [14] Majid A Al-Taee, Waleed Al-Nuaimy, Zahra J Muhsin, Ali Al-Ataby, and Ahmad M Al-Taee. "Mapping security requirements of mobile health systems into software development lifecycle". In: *2016 9th International Conference on Developments in eSystems Engineering (DeSE)*. IEEE. 2016, pp. 87–93.
- [15] Mikael Lindvall, Madeline Diep, Michele Klein, Paul Jones, Yi Zhang, and Eugene Vasserman. "Safety-focused security requirements elicitation for medical device software". In: *2017 IEEE 25th International Requirements Engineering Conference (RE)*. IEEE. 2017, pp. 134–143.

- [16] Anouk Donners, Konrad van der Zwet, Antoine CG Egberts, Karin Fijnvandraat, Ron Mathôt, Ilmar Kruis, Marjon H Cnossen, Roger Schutgens, Rolf T Urbanus, and Kathelijn Fischer. "DosEmi study protocol: a phase IV, multicentre, open-label, crossover study to evaluate non-inferiority of pharmacokinetic-guided reduced dosing compared with conventional dosing of emicizumab in people with haemophilia A". In: *BMJ open* 13.6 (2023), e072363.
- [17] Laura H Bukkems, Siv Jönsson, Marjon H Cnossen, Mats O Karlsson, Ron AA Mathôt, OPTI-CLOT studies, and the SYMPHONY consortium. "Relationship between factor VIII levels and bleeding for rFVIII-SingleChain in severe hemophilia A: A repeated time-to-event analysis". In: *CPT: Pharmacometrics & Systems Pharmacology* 12.5 (2023), pp. 706–718.
- [18] João A Abrantes, Alexander Solms, Dirk Garmann, Elisabet I Nielsen, Siv Jönsson, and Mats O Karlsson. "Relationship between factor VIII activity, bleeds and individual characteristics in severe hemophilia A patients". In: *haematologica* 105.5 (2020), p. 1443.

APPENDIX

10.A SUPPLEMENTARY FIGURES

The screenshot shows a registration form titled "OPTI-CLOT webportal". The title features a stylized logo consisting of three curved lines above the word "webportal". Below the title, the text "Register new user." is displayed. The form consists of five input fields with placeholder text: "First name:" (with placeholder "John"), "Last name:" (with placeholder "Doe"), "Business email:" (with placeholder "john.doe@example.com"), "Role:" (with placeholder "Treatment centre"), and "Treatment centre:" (with placeholder "Treatment centre"). Below the input fields is a checkbox labeled "I agree with the [user agreement](#) of the OPTI-CLOT portal." At the bottom is a "Confirm" button.

OPTI-CLOT webportal

Register new user.

First name:

Last name:

Business email:

Role:

Treatment centre:

I agree with the [user agreement](#) of the OPTI-CLOT portal.

Confirm

Figure 10.A.1: Register page.

Welcome, Alexander

Continue with saved request ▶

New request ▶

Below is an overview of all the request made for your institution.

You can use filters to find specific requests and see information on which requests are completed.

Q Filter...

Request number	OPTI-CLOT pseudo-id	Status	Date
#89364519	99021	COMPLETED	31-1-2023
#13719326	99024	COMPLETED	16-6-2023
#72531646	99027	COMPLETED	28-7-2023
#51023572	99032	COMPLETED	24-1-2024

Figure 10.A.2: Request overview.

← New request

In order for us to give an accurate dosing advice, we need some information about your patient. Completely fill in the below form and add remarks at the end of each page if you want to add additional information. For some variables it is also possible to add the date of the measurement. This is optional.

Note: do not add more than 3 digits.

Patient details	Current treatment	Measurements	Advice
TYPE OF PATIENT	<input type="checkbox"/> Haemophilia A <input type="checkbox"/> Haemophilia B <input type="checkbox"/> von Willebrand disease		
BIRTHDATE	<input type="text"/> dd / mm / yyyy		
WEIGHT	<input type="text"/> kg <input type="text"/> dd / mm / yyyy		
HEIGHT	<input type="text"/> cm <input type="text"/> dd / mm / yyyy		
HAS BLOODGROUP O	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Unknown <input type="radio"/> ?		
REMARKS	<input type="text"/>		
Next			

Figure 10.A.3: Request form.

Elocta

Request number: ——, OPTI-CLOT pseudo-id: ——, age: -- years, body weight: — kg,
Date created: 01-01-1990, by: A. Janssen

PK profile Elocta

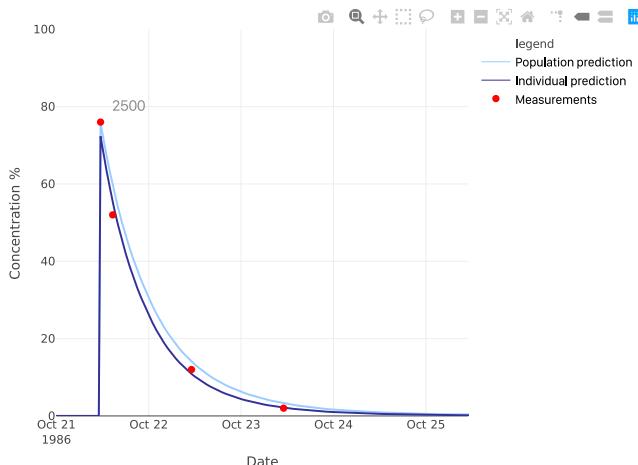


Figure 1: Elocta PK profile. The dark blue line indicates the simulated plasma concentration for the individual prediction and the light blue line represents the population-based prediction.. The red points depict the measured FVIII levels. The administered dose is shown at the time of administration. The time after dose and the day of the week are visible when moving the cursor over the lines. The measurement at $t = 72\text{h}$ was below the limit of detection.

Table 1: PK Parameters

CL	VSS	Terminal half-life
2.545285	37.46097	13.54563

Clearance (CL) is depicted in dL/h, steady state distribution volume (VSS) in dL and the terminal half-life in ours.

Advice #1: 2500 IU three times weekly



Figure 10.A.4: Example of dosing advice.

Part V
DISCUSSION

GENERAL DISCUSSION AND PERSPECTIVES

The aim of this thesis was to identify opportunities for machine learning methods to improve the treatment of patients with haemophilia A as well as to describe ways of adapting existing algorithms to requirements relevant to the field of pharmacometrics. In chapter 2, we performed a literature review to discuss the recent adoption of machine learning algorithms within pharmacometrics and offer considerations for their use. We have described one such method in chapter 3, which can be used to support covariate selection during classical pharmacometric analyses. Next, in chapters 4–6, we have introduced the deep compartment model (DCM) framework, a reliable and robust machine learning approach for pharmacometric analysis. We then developed several machine learning based models to tackle three open issues related to the clinical treatment of haemophilia A using factor concentrates (see chapter 7–9). We concluded our experimental section with the introduction of the OPTI-CLOT web-portal, an online application that offers dosing advice to treatment teams of patients with bleeding disorders (see chapter 10). In the future, we hope to offer the proposed machine learning methods to the haemophilia community through this web-portal.

In the following sections, we discuss perspectives for the use of machine learning in pharmacometrics and haemophilia A.

11.1 MACHINE LEARNING IN PHARMACOMETRICS

11.1.1 *Bringing machine learning to pharmacometrics and rare disease*

The field of machine learning is often linked to the concept of Big Data. Extracting (useful) information from large, complex data sets using traditional methods can be burdensome, whereas machine learning algorithms excel at extracting complex and potentially meaningful patterns from high-dimensional data. It is therefore not surprising that most successful implementations involve the use of massive data sets. In contrast, applying machine learning techniques in a field

such as pharmacometrics is much less straightforward: data is often sparse and requirements for model interpretability and robustness are strict. Recent interest in the adoption of machine learning in this field is nonetheless well reflected by a significant rise in the number of publications mentioning both topics (see figure 11.1.1).

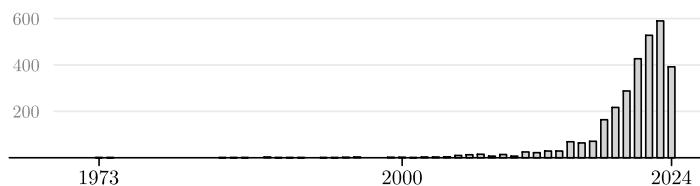


Figure 11.1.1: Number of publications indexed by PubMed mentioning "pharmacometrics" and "machine learning". Last updated: 6th of June 2024.

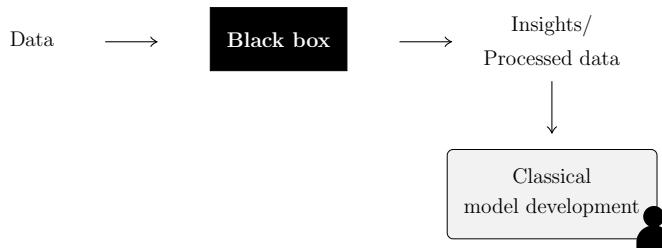
Only three papers are tagged with "haemophilia", excluding papers from this thesis.

Several literature reviews already describe a wide range of potential use-cases within pharmacometrics and offer future perspectives [1–4]. Most experimental papers focus on implementations in oncology ($n = 234$) or COVID-19 ($n = 195$), coincidentally two fields where larger data sets are more prevalent. In contrast, significantly fewer articles discuss applications within rare diseases. In this context, additional restrictions regarding the amount of data combined with an often incomplete understanding of disease physiology complicate the implementation of machine learning algorithms. Potential benefits can however be transformative: recent examples include using machine learning to support drug discovery, improve disease diagnosis, or to personalise treatment with often expensive medication [5–7]. Aside from a need to specifically adjust methods to meet requirements relevant to rare diseases and pharmacometrics, there is also a gap between the interest and actual adoption of machine learning. This is in part due to a lack of flexible and user-friendly software implementations of machine learning methods specific to pharmacometrics.

In chapter 2, we have identified three generic approaches for implementing machine learning techniques in pharmacometrics. First, machine learning can be used as a tool to support classical analyses, for example by performing (initial) covariate screening (see chapter 3). Second, methods have been suggested to fully learn models from data, greatly simplifying model development [9]. Finally, hybrid models combine machine learning with prior knowledge from pharmacomet-

rics to improve data efficiency (for example by explicitly specifying drug kinetics in a dynamical system; see chapter 4 and 5). In the following sections we discuss future perspectives regarding these three approaches in more detail. In addition, we will evaluate their suitability to applications within rare diseases.

11.1.2 Machine learning as a tool to support pharmacometric analyses



Many of the early adopters of machine learning in pharmacometrics will likely focus on methods supporting classical analyses. This way, the typical pharmacometrist can still rely on their own expertise in pharmacology while benefitting from the advantages machine learning algorithms can offer. There are two key use-cases for machine learning in this setting: performing data pre-processing (for example by imputing missing values, see chapter 7), or to screen data for potentially relevant covariates (see chapter 3) [8, 10].

First, a large fraction of data is often missing in the healthcare setting, which is especially the case when working with real-world data collected from patient health records. In the low-dimensional setting or when the covariates are strongly correlated (as is for example the case with age, height, and body weight) it is possible to manually describe models to impute missing data. However, the development of such models for high-dimensional or time-series data quickly becomes complicated. In these cases, machine learning methods such as multiple imputation by chained equations (MICE), variational autoencoders (VAE), generative adversarial networks (GANs), or diffusion models can potentially improve the accuracy of imputed data [11–14]. An initial evaluation of the performance of machine learning methods for data imputation showed positive results, so further research is of interest [10].

Synthetic data is artificially generated and meant to closely resemble real-world data.

In federated learning, models are shared rather than data, ensuring that sensitive information stays within local institutions.

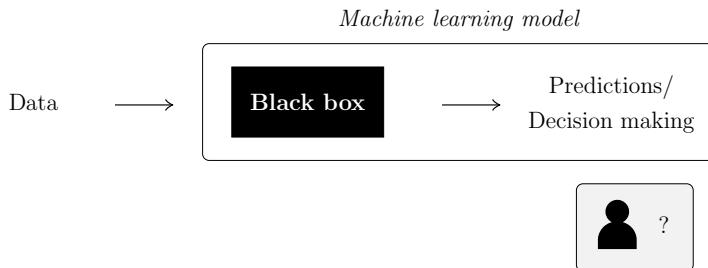
Machine learning methods can also be used to generate *synthetic data*, which is of great interest in the context of highly sensitive patient data. Privacy-preserving synthetic data can be shared instead of real data, offering a potential solution to data limitations. Smaller data sets can be augmented with synthetic data to create richer data sets, which is of course especially relevant within the context of rare disease (see chapter 7). However, it is still to be explored whether generative machine learning methods perform well on smaller data sets. Privacy requirements will also be more strict when utilising data from patients with rare diseases, as they are more easily traced based on certain characteristics despite anonymisation [15].

Our efforts with regard to these challenges will be pursued as part of the PHEMS consortium (<https://phems.eu/>), a research initiative launched by the European Children's Hospitals Organisation (ECHO) and funded by the European Union. The consortium aims to facilitate the development of machine learning models on data obtained from electronic healthcare records of several European paediatric hospitals using *federated learning*. Participating centres will be able to train machine learning models on the data from other centres without the need for direct access. A secondary goal of the PHEMS project is to develop generative models that produce a synthetic version of the data from each centre which can be shared with external partners. To this end, we will investigate whether accurate and privacy-preserving generative models can still be developed when data is sparse.

Second, machine learning methods can be used to speed up covariate screening. Here, classical linear methods such as LASSO can be used to identify important covariates by shrinking regression coefficients of unimportant covariates to zero [16]. More advanced machine learning algorithms, such as random forests, relax the assumption of linear effects and can potentially improve the identification of covariates with more complex relationships. After fitting the machine learning model, AI explainability tools can be used to rank the covariates based on model-specific importance metrics. More advanced analyses can also be performed by visualising learned covariate effects, which can help provide an indication of the functional forms to use when implementing each covariate (see chapter 3). However, interpretation of the covariate effects as represented by the explanation model is not necessarily straightforward, especially in high-dimensional settings. Issues arising from collinearity, confounding, or model overfitting also complicate the use of these methods as standard pre-screening tools. It is thus probable that users will still have to critically evaluate the

results of these analyses, which at minimal requires experience with limitations and pitfalls of the specific machine learning algorithm and explainability tool used.

11.1.3 Should pharmacometrists be replaced by machines?



The ultimate goal of machine learning research is to create autonomous systems that continuously learn from data. Current focus is on large language models (LLMs), chatbot-like systems that appear to offer human-like responses to queries. Recently, researchers have delved into the use of LLMs in the context of pharmacometrics. These systems can be used to support users by writing code and finding errors, summarising scientific papers, and by suggesting model components to add [17–19]. However, some initiatives take it one step further and use specifically trained LLM models to guide pharmacometrists through the entire model development process [20]. Specific LLMs are constructed to read input documents, create data sets, find relevant information online, generate model code, iteratively evaluate & improve models, and finally to simulate data for decision-making.

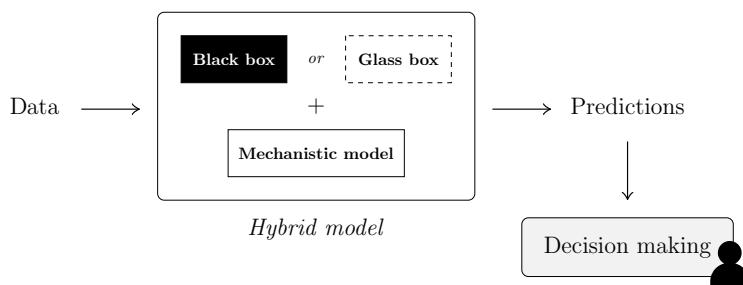
Initial evaluations of such a system showed large increases in the productivity of its users, but the grim reality is that these systems reduce the role of pharmacometrists to *prompt engineering* and verification of decisions made by the AI. As time goes on, the actual model development experience of pharmacometrists will diminish, hampering their ability to detect errors made by the AI system. There are many points along the data analysis pipeline at which the introduction of errors can lead to incorrect conclusions. Since LLMs are prone to hallucinate false information [21], the widespread implementation of these AI systems without adequate supervision can result in real-world harm to patients. It might also take pharmacometrists

Prompt engineering involves finding the best wording to use when querying a LLM to get optimal responses.

considerable time to validate all the actions taken by the AI system (especially so when data is automatically extracted from free-text documents), mitigating any increases in productivity.

An alternative future involves expanding the toolset of pharmacometrists with more advanced methods to promote the rate of model development. Designing such algorithms is not straightforward as was pointed out in chapter 1. Recent developments with respect to NeuralODEs are nonetheless promising [22]. These differential equation based models may be able to improve extrapolation capabilities of machine learning models by learning the dynamical system governing the observations [23, 24]. When sufficient data is available, these models can potentially elucidate complex disease dynamics [9, 25]. A recent extension to the method incorporates inter-individual variability in responses, providing a measure of uncertainty to predictions and allowing for the selection of the most likely response based on observed data [26]. However, an important open question is whether these methods will behave well on smaller data sets (see chapter 5). Since managing model complexity is not straightforward, NeuralODEs can still learn unexpected behaviour outside of the training data. It will thus probably be necessary to include some form of model regularisation to improve performance, but it is not necessarily clear what sort of an approach will be effective. In addition, when disease physiology is sufficiently complex, it is possible that the data simply does not support the learning of the true dynamical system. More research is therefore required to produce more reliable algorithms.

11.1.4 Combining the two approaches in hybrid models



Finally, hybrid models have the potential to combine the best of both worlds: it enables researchers to utilise their previous domain

knowledge while using machine learning methods to simplify model development. By incorporating prior domain knowledge, these models can improve data efficiency as the scope of the learning problem is reduced. One example of such a hybrid model architecture is the DCM, a model combining neural networks with compartment models to automatically learn covariate effects from data (see chapter 4). In contrast to NeuralODEs, we have demonstrated that the DCM can still obtain relatively high accuracy when trained on sparse data sets (see chapter 4 and 5). In addition, these models use the same interpretable parameters (e.g. drug clearance and volume of distribution) as classical pharmacometric models, allowing for comparisons to prior results. Data efficiency can be further improved by adding physiologically-based constraints on the model parameters (see chapter 5). Results obtained from explainable AI methods might also be more readily interpretable as the relationship between covariates and pharmacokinetic (PK) or pharmacodynamic (PD) parameters are often relatively simple. For example, the relationship between body weight and clearance closely resembles a linear correlation, whereas the association between body weight and drug concentrations follows a more complex, non-linear relationship. Explanations for models that directly predict the dependent variable are also time-dependent, further complicating model interpretation. The use of specialised architectures that are inherently interpretable (so-called *glass-box models*) can even be used, which as we have shown do not necessarily incur a loss of accuracy (see chapter 5).

Hybrid models can also be used to learn unknown parts of the dynamical system [24, 27]. A NeuralODE can for example be used to learn absorption kinetics while explicit partial differential equations are used to represent drug distribution and clearance as in typical compartment models. By giving explicit meaning to the dynamics learned by the NeuralODE, it becomes feasible to regularise the solution and to set rules so that the model produces expected behaviour. For example, total drug concentration should always go to zero, cannot be negative, and cannot exceed the administered dose amount. Although these rules seem very straightforward, without constraints there is no guarantee that the model will not learn to produce incorrect behaviour. It will be of great interest to further develop hybrid model applications that combine prior knowledge and machine learning methods in innovative ways. We have developed a software package (DeepCompartmentModels.jl) that combines the DCM framework

In contrast to a "black box", predictions from glass box models can be directly explained based on model structure.

with NeuralODEs in order to facilitate further research and adoption of these techniques.

11.1.5 Future directions

There are several interesting future directions for machine learning in pharmacometrics. First, given the advancements in NeuralODEs and hybrid models, it is possible that practitioners will forego intermediate methods (e.g. those supporting classical analyses) and instead directly use machine learning based models. For example, by adding mixed-effects estimation to the DCM framework (see chapter 6), we provide a simple to implement alternative to classical non-linear mixed effects models. We also suggest constructing these models in such a way that they are inherently interpretable, facilitating the identification of important covariates while providing the ability to critique models based on learned effects (see chapter 5). Covariate analysis can be performed in a similar way as in classical methods, but the use of neural networks means that the selection of specific functions is not necessary. Furthermore, a *full covariate model* based approach combined with deep ensembles can potentially be used to identify covariate importance from a single model run (see chapter 5). These methods could significantly reduce the time spend on model development.

Full model estimation involves directly including all candidate covariates and assessing their importance based on learned effects.

Second, the more widespread implementation of techniques from (Bayesian) causal inference would be of interest to improve the development of pharmacometric models [1, 28, 39]. These techniques can for example be used to compare potential outcomes of two different treatments. This is a complex problem, seeing as generally only a single outcome is actually observed for each patient. By representing (known) causal relationships between variables in a directed acyclic graph (DAG), one can perform specific interventions (e.g. give drug X and measure response) that enable the estimation of causal effects [37]. In chapter 7, we show that extrapolation performance of machine learning models can potentially be improved by fitting models that follow a causal diagram. For example, it is well known that von Willebrand factor (VWF) binds to FVIII to protect it from proteolytic degradation [38]. In a causal model we might therefore include a relationship between VWF and FVIII clearance but not with other PK parameters. When explicitly not allowing such an effect to be learned (for example by only linking specific covariates to specific PK parameters, see chapter 5), we might prevent machine learning models from learning spurious effects.

Machine learning methods can also support causal discovery, for example by learning causal relationships based on observational data and augmenting existing causal graphs [40–42]. In this context, it is likely beneficial to consider Bayesian methods in order to quantify the uncertainty of learned effects. By supplying prior information about covariate importance or even function complexity, these models can potentially identify important covariates. When data is spare, it is sensible to discourage the model from learning overly complex relationships to reduce overfitting. Gaussian Processes (GPs) offer an intuitive way of setting such priors (see chapter 1).

Third, the introduction of GPs in non-linear mixed effects models might be of a more general interest to pharmacometrists: it is more intuitive to specify the desired complexity of a function versus manually composing and tuning explicit algebraic functions. By using Bayesian *non-parametric* methods like GPs, we can perform data-driven learning of covariate effects while obtaining posterior distributions representing the uncertainty over their effects. These posteriors can be used as priors in subsequent studies to enable continuous model development. We can further facilitate such a process by supporting models with a generative component, for example one that simulates synthetic patients similar to the training data (see chapter 7). These generative models can be used to augment data sets in subsequent studies, or to identify a population PK model that was trained on a patient population similar to a new patient to improve predictions. The sharing of data and models in such a way would be of great interest to improve models in the long run, especially so in the context of rare diseases.

Non-parametric methods can make predictions without assuming a fixed parameter structure.

Finally, probabilistic methods such as GPs and neural stochastic differential equations (NeuralSDEs) can potentially be used to learn uncertainty over the structure of the dynamical model [26, 29]. By using these methods, we can obtain uncertainty estimates over the inclusion of specific model components, for example when adding additional compartments. Interpretation of the obtained uncertainty estimates (which are probabilities) is more straightforward than comparing p -values, which are typically obtained during such comparisons. Priors can again be used to control model complexity and to specify expected behaviour outside of the observed data to improve extrapolation. The use of Bayesian methods in this manner gives greater control to pharmacometrists to make decisions during model development.

11.2 BRINGING MACHINE LEARNING TO HAEMOPHILIA A

11.2.1 Opportunities related to drug discovery and bleeding phenotype

Unlike numerous other rare conditions where therapeutic options remain limited, the treatment of haemophilia A has seen considerable advancements. Factor replacement therapy (and non-factor based alternatives) feature high efficacy and have resulted in large improvements in quality of life. In addition, treatment is highly personalised, either by applying PK-guided dosing or by making iterative adjustments to dosing in response to breakthrough bleeding. Consequently, many patients with haemophilia A are able to lead relatively normal lives. Despite these advancements, research into haemophilia A remains highly active, with ongoing efforts to refine current therapies, explore novel treatment modalities (including gene therapy), and search for biomarkers representative of the individual bleeding phenotype. Although the focus of this thesis has mostly been on the use of machine learning methods to personalise haemophilia A treatment, we will also briefly highlight its opportunities with respect to these other research areas.

First, machine learning algorithms can significantly accelerate drug discovery processes by predicting how different compounds interact with biological targets, identifying potential therapeutic candidates, and optimising lead compounds [30]. For example, one pioneering study in haemophilia A has utilised a machine learning framework to predict disease severity based on the molecular structure of FVIII [31]. This model takes FVIII amino acid properties, structural relationships, and clinical characteristics as inputs and returns the probability of observing a severe haemophilia A phenotype. These probabilities can be used to identify 'hotspots' that are especially sensitive to mutations. Similar models can be constructed that could for example learn to predict how changes in the protein sequence of FVIII affect drug half-lives. Such information can in turn be used to guide the design of novel therapeutic compounds. The introduction of machine learning in drug discovery and repurposing pipelines has high potential in the context of rare disease. Not only would these approaches enhance the efficiency of drug discovery, but they could also potentially reduce the time and cost associated with bringing new therapies to market. It will be of interest for pharmaceutical companies to invest in the introduction of these algorithms to accelerate drug development.

Second, one of the key capabilities of machine learning is its potential to identify patterns from large-scale data sets that might be difficult to detect using traditional statistical methods. Deep learning methods can for example be used to analyse high-dimensional -omics data to uncover biomarkers and predict risk of disease [33, 34]. In the context of haemophilia A, it is well known that there is considerable variability in bleeding outcomes between patients, even after correcting for FVIII exposure [60]. Proteomics data can potentially be used to identify biomarkers that explain these differences [35, 36]. Work-package 10 of the SYMPHONY consortium focuses on the detection of proteolytic signatures that correlate with differences in bleeding phenotypes between patients with haemophilia A or von Willebrand disease. Due to the complexity and high-dimensionality of this data, it might be of interest to evaluate potential benefits of using machine learning algorithms to aid such analyses.

11.2.2 *Open issues for personalised treatment in haemophilia A*

Population PK models have long been used as a quantitative method to personalise the prophylactic treatment of haemophilia A patients. In chapters 7–9 of this thesis, we have described three examples where machine learning algorithms can be applied to improve treatment. Specifically, we show how model extrapolation can be improved by adopting techniques from causal inference, how to identify time-dependent changes in FVIII clearance to improve predictions in the perioperative setting, and finally how to predict bleeding outcomes in response to treatment based on FVIII PK and estimated bleeding risk. Aside from these contributions, there still are some clinical settings where personalised treatment is complicated or where there is room for improvement.

First, residual FVIII levels from a previous dose can affect the measurements taken for a PK profile. Without correcting for these residual levels, the estimated PK parameters will be biased. Although estimates can be easily corrected by including previous doses in the model, this information is often missing. This is especially the case when working with (retrospective) data from health record systems: prior doses are not available in health record systems since these were most likely administered by the patient at home. It might therefore be necessary to augment PK models with a probabilistic model predicting the most likely dosage as well as its time of administration. Methods

from the field of Bayesian Inference such as Markov Chain Monte Carlo (MCMC) might be well suited for this problem.

Second, current population PK model are generally not suited to predict FVIII exposure with respect to multiple different factor concentrates based on only a single PK profile [57]. Typically, we would not only want to know the FVIII consumption necessary to maintain certain target levels for one drug, but also whether switching to an alternative drug with more favourable PK is cost-effective. Theoretically, when we correctly adjust for sources of variability in the PK of different FVIII concentrates, we should be able to obtain reasonable estimates of FVIII exposure for alternative drugs. To develop such a system, we can build upon advancements from the field of machine learning, such as the Bayesian causal inference framework suggested in section 11.1.5. By using probabilistic machine learning methods, we can iteratively improve models on many different data sets of patients receiving different FVIII concentrates. Researchers can obtain the latest model version and continue training on local data sets. After fitting the model, updated model parameters can be shared instead of patient data, keeping sensitive data private. By training on a large number of such data sets, the model can learn to correct for more diverse sources of variability, improving prediction accuracy.

Third, model development may be complicated by the limited availability of data. Since haemophilia A is a rare disease, most single centres will not be able to collect large-scale data sets required for complex machine learning analyses. It might therefore be of interest to create generative models that accurately simulate synthetic patients that are similar to those from local data sets. By sharing these generative models alongside PK models, users can identify whether their patient is similar to those used to develop the PK model (see chapter 7). Meanwhile, the augmentation of existing data sets with synthetic data offers an opportunity to build larger data sets to train PK models.

Coincidentally, the problems encountered within use-case three of the PHEMS consortium (which is focused on improving a population PK model for haemophilia A) align well with the previously mentioned issues. The data from each participating centre is collected during routine clinical practice, meaning that a large fraction of dosing information at the time of FVIII measurements is expected to be missing. Furthermore, there are inherent differences between the patient populations from each centre as well as FVIII concentrate and measurement assay use. Finally, the implementation of federated learning might be able to address the issue of data availability by allowing

models to be trained on data from multiple centres. It will be of interest to determine whether generalisable solutions to these problems can be found as they will benefit the broader research community.

A final hurdle is the selection of appropriate FVIII target levels to reduce bleeding rates on an individualised basis. In chapter 9 we address this issue by using repeated time-to-event (RTTE) models to predict individual bleeding risk. In the current clinical landscape, physicians frequently base treatment decisions on expert opinion regarding the patient's bleeding phenotype rather than generic FVIII target levels. This phenotype is difficult to define concisely, and is often a subjective assessment of joint state, bleeding frequency, the severity and types of bleeding, physical activity levels, risk-aversion of the patient, and the current treatment regimen. Rather than directly attempting to quantify the bleeding phenotype, RTTE models can estimate the bleeding risk based on reported bleeding and treatment information. Not only can these models be used to compare patients based on differences in bleeding risk, they can also be used to simulate individual bleeding outcomes in response to different treatment regimens. When combined with the aforementioned causal population PK models, it might also be possible to estimate the effect of switching treatment on bleeding outcomes, for example when switching from standard to extended half-life concentrates or non-factor based therapies. Since we can directly compare treatment options based on FVIII consumption, these tools can be used to perform more comprehensive cost-effectiveness analyses and might provide stronger rationale for changing treatment. Additionally, such an approach supports shared decision making, as selection of the optimal treatment regimen based on bleeding outcomes is more illustrative to patients and healthcare professionals. It would be of great interest to clinically validate the performance of the use of RTTE-based methods for the optimisation of treatment.

11.2.3 Improving the access to personalised treatment in clinical practice

Although several clinical guidelines advise the use of PK-guided dosing to personalise treatment of patients with haemophilia A, its actual clinical adoption is somewhat lacking [49]. There are several factors that contribute to this finding. Importantly, ambiguities regarding measured FVIII levels or model predictions mean that expert pharma-

cologists are frequently required to aid in interpreting and evaluating PK profiles. The resulting process can be time-intensive, and requires specific expertise not available in all treatment centres. In chapter 10, we have introduced the OPTI-CLOT web-portal, a web-application that offers free access to dosing advice for treatment teams of patients with rare bleeding disorders. Here, patient data is shared with the external platform and tailored dosing advice is provided by a team of expert pharmacometrists. This web-portal currently provides dosing advice for patients with haemophilia A or B and von Willebrand disease, and adoption in the Netherlands has been promising. The uptake of the web-portal might be further enhanced by reducing the burden of entering relevant patient information, for example by directly interfacing with electronic health record systems to automatically extract data. By developing the web-portal, we hope to bring novel techniques closer to actual clinical adoption. In the future, we aim to offer dosing advise based on the patient's individual bleeding phenotype, for example through the use of RTTE-based models.

11.3 CONCLUSIONS

The work undertaken as part of this thesis supports the adoption of machine learning algorithms in the field of pharmacometrics. Principally, we have introduced the DCM framework, which is a reliable and robust machine learning based approach to predict drug exposure and effects. Next, we have shown that machine learning methods can achieve clinical benefits by tackling issues with respect to the personalisation of treatment for patients with haemophilia A. Perhaps one of the most important contributions is the development of a RTTE model to predict the individual bleeding risk in response to treatment. This method has the potential to revolutionise the approach taken to personalised treatment in haemophilia A by focusing on bleeding outcomes rather than just FVIII levels. To conclude, this thesis can hopefully serve as an example for the implementation of machine learning in the context of rare diseases in a more general sense. Our focus has been on improving model performance in sparse data settings, and is likely of interest to assist the treatment of patients with other rare conditions.

REFERENCES

- [1] Alicia Curth, Richard W Peck, Eoin McKinney, James Weatherall, and Mihaela van Der Schaar. "Using Machine Learning to Individualize Treatment Effect Estimation: Challenges and Opportunities". In: *Clinical Pharmacology & Therapeutics* (2024).
- [2] Kamilė Stankevičiūtė, Jean-Baptiste Woillard, Richard W Peck, Pierre Marquet, and Mihaela van der Schaar. "Bridging the worlds of pharmacometrics and machine learning". In: *Clinical Pharmacokinetics* 62.11 (2023), pp. 1551–1565.
- [3] Ethan A Powelet, Alexander A Vinks, and Tomoyuki Mizuno. "Artificial intelligence and machine learning approaches to facilitate therapeutic drug management and model-informed precision dosing". In: *Therapeutic drug monitoring* 45.2 (2023), pp. 143–150.
- [4] Mason McComb, Robert Bies, and Murali Ramanathan. "Machine learning in pharmacometrics: Opportunities and challenges". In: *British Journal of Clinical Pharmacology* 88.4 (2022), pp. 1482–1499.
- [5] Magda Wojtara, Emaan Rana, Taibia Rahman, Palak Khanna, and Heshwin Singh. "Artificial intelligence in rare disease diagnosis and treatment". In: *Clinical and Translational Science* 16.11 (2023), pp. 2106–2111.
- [6] Sunil Mathur and Joseph Sutton. "Personalized medicine could transform healthcare". In: *Biomedical reports* 7.1 (2017), pp. 3–5.
- [7] Marisa P Dolled-Filhart, Amanda Lordemann, William Dahl, Rajini Rani Haraksingh, Chih-Wen Ou-Yang, and Jimmy Cheng-Ho Lin. "Personalizing rare disease research: how genomics is revolutionizing the diagnosis and treatment of rare disease". In: *Personalized medicine* 9.8 (2012), pp. 805–819.
- [8] Emeric Sibiede, Akash Khandelwal, Jan S Hesthaven, Pascal Girard, and Nadia Terranova. "Fast screening of covariates in population models empowered by machine learning". In: *Journal of Pharmacokinetics and Pharmacodynamics* 48.4 (2021), pp. 597–609.
- [9] James Lu, Brendan Bender, Jin Y Jin, and Yuanfang Guan. "Deep learning prediction of patient response time course from early data via neural-pharmacokinetic/pharmacodynamic modelling". In: *Nature machine intelligence* 3.8 (2021), pp. 696–704.
- [10] Dominic Stefan Bräm, Uri Nahum, Andrew Atkinson, Gilbert Koch, and Marc Pfister. "Evaluation of machine learning methods for covariate data imputation in pharmacometrics". In: *CPT: Pharmacometrics & Systems Pharmacology* 11.12 (2022), pp. 1638–1648.
- [11] Stef Van Buuren, Jaap PL Brand, Catharina GM Groothuis-Oudshoorn, and Donald B Rubin. "Fully conditional specification in multivariate imputation". In: *Journal of statistical computation and simulation* 76.12 (2006), pp. 1049–1064.
- [12] Jinsung Yoon, James Jordon, and Mihaela Schaar. "Gain: Missing data imputation using generative adversarial nets". In: *International conference on machine learning*. PMLR. 2018, pp. 5689–5698.
- [13] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.
- [15] Khaled El Emam, Fida K Dankar, Régis Vaillancourt, Tyson Roffey, and Mary Lysyk. "Evaluating the risk of re-identification of patients from hospital prescription records". In: *The Canadian journal of hospital pharmacy* 62.4 (2009), p. 307.
- [16] Jakob Ribbing, Joakim Nyberg, Ola Caster, and E Niclas Jonsson. "The lasso—a novel method for predictive covariate model building in nonlinear mixed effects models". In: *Journal of pharmacokinetics and pharmacodynamics* 34 (2007), pp. 485–517.
- [17] Michael E Cloesmeijer, Alexander Janssen, Sjoerd F Koopman, Marjon H Cnossen, Ron AA Mathôt, and SYMPHONY consortium. "ChatGPT in pharmacometrics? Potential opportunities and limitations". In: *British journal of clinical pharmacology* 90.1 (2024), pp. 360–365.
- [18] Euibeom Shin, Yifan Yu, Robert R Bies, and Murali Ramanathan. "Evaluation of ChatGPT and Gemini large language models for pharmacometrics with NONMEM". In: *Journal of Pharmacokinetics and Pharmacodynamics* (2024), pp. 1–11.
- [19] Euibeom Shin and Murali Ramanathan. "Evaluation of prompt engineering strategies for pharmacokinetic data analysis with the ChatGPT large language model". In: *Journal of Pharmacokinetics and Pharmacodynamics* 51.2 (2024), pp. 101–108.
- [20] Ercan Suekuer. "PMx-AI Bot: Changing the way of traditional Pharmacometrics work with AI Bots". In: *Proceedings of the 32nd Annual Meeting of the Population Approach Group in Europe. Abstract 11257.* 2024. URL: www.page-meeting.org/?abstract=11257.
- [21] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions". In: *arXiv preprint arXiv:2311.05232* (2023).
- [22] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. "Neural ordinary differential equations". In: *Advances in neural information processing systems* 31 (2018).
- [23] James Lu, Kaiwen Deng, Xinyuan Zhang, Gengbo Liu, and Yuanfang Guan. "Neural-ODE for pharmacokinetics modeling and its advantage to alternative machine learning models in predicting new dosing regimens". In: *Iscience* 24.7 (2021).
- [24] Dominic Stefan Bräm, Uri Nahum, Johannes Schropp, Marc Pfister, and Gilbert Koch. "Low-dimensional neural ODEs and their application in pharmacokinetics". In: *Journal of Pharmacokinetics and Pharmacodynamics* (2023), pp. 1–18.
- [25] Dominic Stefan Bräm, Gilbert Koch, Karel Allegaert, John van den Anker, and Marc Pfister. "Applying Neural ODEs to Derive a Mechanism-Based Model for Characterizing Maturation-Related Serum Creatinine Dynamics in Preterm Newborns". In: *The Journal of Clinical Pharmacology* (2024).

- [26] Samira Pakravan, Nikolaos Evangelou, Maxime Usdin, Logan Brooks, and James Lu. "From Noise to Signal: Unveiling Treatment Effects from Digital Health Data through Pharmacology-Informed Neural-SDE". In: *arXiv preprint arXiv:2403.03274* (2024).
- [27] Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. "Universal differential equations for scientific machine learning". In: *arXiv preprint arXiv:2001.04385* (2020).
- [28] Fan Li, Peng Ding, and Fabrizia Mealli. "Bayesian causal inference: a critical review". In: *Philosophical Transactions of the Royal Society A* 381.2247 (2023), p. 20220153.
- [29] Markus Heinonen, Cagatay Yıldız, Henrik Mannerström, Jukka Intosalmi, and Harri Lähdesmäki. "Learning unknown ODE models with Gaussian processes". In: *International conference on machine learning*. PMLR. 2018, pp. 1959–1968.
- [30] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. "Applications of machine learning in drug discovery and development". In: *Nature reviews Drug discovery* 18.6 (2019), pp. 463–477.
- [31] Tiago JS Lopes, Ricardo Rios, Tatiane Nogueira, and Rodrigo F Mello. "Prediction of hemophilia A severity using a small-input machine-learning framework". In: *NPJ systems biology and applications* 7.1 (2021), p. 22.
- [32] Sergio Decherchi, Elena Pedrini, Marina Mordini, Andrea Cavalli, and Luca Sangiorgi. "Opportunities and challenges for machine learning in rare diseases". In: *Frontiers in medicine* 8 (2021), p. 747612.
- [33] Polina Mamoshina, Marina Volosnikova, Ivan V Ozerov, Evgeny Putin, Ekaterina Skibina, Franco Cortese, and Alex Zhavoronkov. "Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification". In: *Frontiers in genetics* 9 (2018), p. 242.
- [34] Julia Carrasco-Zanini, Maik Pietzner, Jonathan Davitte, Praveen Surendran, Damien C Croteau-Chonka, Chloe Robins, Ana Torralbo, Christopher Tomlinson, Florian Grünschläger, Natalie Fitzpatrick, et al. "Proteomic signatures improve risk prediction for common and rare diseases". In: *Nature Medicine* (2024), pp. 1–10.
- [35] Yassene Mohammed, Bart J van Vlijmen, Juncong Yang, Andrew J Percy, Magnus Palmblad, Christoph H Borchers, and Frits R Rosendaal. "Multiplexed targeted proteomic assay to assess coagulation factor concentrations and thrombosis-associated cancer". In: *Blood Advances* 1.15 (2017), pp. 1080–1087.
- [36] Yassene Mohammed, Carolina E Touw, Banne Nemeth, Raymond A van Adrichem, Christoph H Borchers, Frits R Rosendaal, Bart J van Vlijmen, and Suzanne C Cannegieter. "Targeted proteomics for evaluating risk of venous thrombosis following traumatic lower-leg injury or knee arthroscopy". In: *Journal of Thrombosis and Haemostasis* 20.3 (2022), pp. 684–699.
- [37] Judea Pearl. "Causal diagrams for empirical research". In: *Biometrika* 82.4 (1995), pp. 669–688.
- [38] V Terraube, JS O'donnell, and PV Jenkins. "Factor VIII and von Willebrand factor interaction: biological, clinical and therapeutic importance". In: *Haemophilia* 16.1 (2010), pp. 3–13.

- [39] Ioana Bica, Ahmed M Alaa, Craig Lambert, and Mihaela Van Der Schaar. "From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges". In: *Clinical Pharmacology & Therapeutics* 109.1 (2021), pp. 87–100.
- [40] Yunan Luo, Jian Peng, and Jianzhu Ma. "When causal inference meets deep learning". In: *Nature Machine Intelligence* 2.8 (2020), pp. 426–427.
- [41] Christian Toth, Lars Lorch, Christian Knoll, Andreas Krause, Franz Pernkopf, Robert Peherz, and Julius Von Kügelgen. "Active bayesian causal inference". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 16261–16275.
- [42] Kai Lagemann, Christian Lagemann, Bernd Taschler, and Sach Mukherjee. "Deep learning of causal structures in high dimensions under data limitations". In: *Nature Machine Intelligence* 5.11 (2023), pp. 1306–1316.
- [43] Alok Srivastava, Elena Santagostino, Alison Dougall, Steve Kitchen, Megan Sutherland, Steven W Pipe, Manuel Carcao, Johnny Mahlangu, Margaret V Ragni, Jerzy Windyga, et al. "WFH guidelines for the management of hemophilia". In: *Haemophilia* 26 (2020), pp. 1–158.
- [44] Annamaria Iorio, V Blanchette, J Blatny, P Collins, K Fischer, E Neufeld, et al. "Estimating and interpreting the pharmacokinetic profiles of individual patients with hemophilia A or B using a population pharmacokinetic approach: communication from the SSC of the ISTH". In: *Journal of Thrombosis and Haemostasis* 15.12 (2017), pp. 2461–2465.
- [45] Rachel Rayment, Elizabeth Chalmers, Katherine Forsyth, Richard Gooding, Anne M Kelly, Susan Shapiro, Kate Talks, Oliver Tunstall, Tina Biss, and British Society for Haematology. "Guidelines on the use of prophylactic factor replacement for children and adults with Haemophilia A and B". In: *British journal of haematology* 190.5 (2020), pp. 684–695.
- [46] K Fischer, Jan Asterman, JG Van Der Bom, R Ljung, Erik Berntorp, DE Grobbee, and HM Van Den Berg. "Prophylactic treatment for severe haemophilia: comparison of an intermediate-dose to a high-dose regimen". In: *Haemophilia* 8.6 (2002), pp. 753–760.
- [47] MT Álvarez-Román, I Fernandez-Bello, H De La Corte-Rodríguez, AL Hernández-Moreno, M Martín-Salces, N Butta-Coll, MI Rivas-Pollmar, S Rivas-Muñoz, and V Jiménez-Yuste. "Experience of tailoring prophylaxis using factor VIII pharmacokinetic parameters estimated with myPKFiT® in patients with severe haemophilia A without inhibitors". In: *Haemophilia: the official journal of the World Federation of Hemophilia* 23.1 (2017), e50–e54.
- [48] Erik Berntorp, Daniel Hart, Maria Elisa Mancuso, Roseline d'Oiron, David Perry, Brian O'Mahony, Radoslaw Kaczmarek, Miguel Crato, John Pasi, Alec Miners, et al. "The first Team Haemophilia Education meeting, 2015, Amsterdam, The Netherlands". In: *European journal of haematology* 97 (2016), pp. 3–18.
- [49] Tine MHJ Goedhart, A Janssen, Ron AA Mathôt, Marjon H Cnossen, OPTI-CLOT Study Group, SYMPHONY Consortium, et al. "The road to implementation of pharmacokinetic-guided dosing of factor replacement therapy in hemophilia and allied bleeding disorders. Identifying knowledge gaps by mapping barriers and facilitators". In: *Blood Reviews* (2023), p. 101098.
- [50] Joachim J Potgieter, Michael Damgaard, and Andreas Hillarp. "One-stage vs chromogenic assays in haemophilia A". In: *European journal of haematology* 94 (2015), pp. 38–44.

- [51] J Ingerslev, MA Jankowski, SB Weston, and LA Charles. "Collaborative field study on the utility of a BDD factor VIII concentrate standard in the estimation of BDDr Factor VIII: C activity in hemophilic plasma using one-stage clotting assays". In: *Journal of Thrombosis and Haemostasis* 2.4 (2004), pp. 623–628.
- [52] Tine MHJ Goedhart, Laura H Bukkems, C Michel Zwaan, Ron AA Mathôt, Marjon H Cnossen, OPTI-CLOT study group, and SYMPHONY consortium. "Population pharmacokinetic modeling of factor concentrates in hemophilia: an overview and evaluation of best practice". In: *Blood advances* 5.20 (2021), pp. 4314–4325.
- [53] Cedric Hermans, C Altisent, A Batorova, H Chambost, P De Moerloose, A Karafoulidou, R Klamroth, M Richards, B White, G Dolan, et al. "Replacement therapy for invasive procedures in patients with haemophilia: literature review, European survey and recommendations". In: *Haemophilia* 15.3 (2009), pp. 639–658.
- [54] Nederlandse Vereniging van Hemofiliebehandelaars. "Richtlijn: diagnostiek en behandeling van hemofylie en aanverwante hemostasestoornissen". In: *Alphen aan den Rijn: Van Zuiden* (2009).
- [55] G Longo, A Messori, M Morfini, F Baudo, N Ciavarella, S Cinotti, E Filimberti, G Giustarini, AC Molinari, and Pierluigi Rossi Ferrini. "Evaluation of factor VIII pharmacokinetics in hemophilia-A subjects undergoing surgery and description of a nomogram for dosing calculations". In: *American journal of hematology* 30.3 (1989), pp. 140–149.
- [56] Angelika Batorova and Uri Martinowitz. "Intermittent injections vs. continuous infusion of factor VIII in haemophilia patients undergoing major surgery". In: *British journal of haematology* 110.3 (2000), pp. 715–720.
- [57] K Yu Jacky, Alfonso Iorio, Pierre Chelle, and Andrea N Edginton. "A comparison of methods for prediction of pharmacokinetics across factor concentrate switching in hemophilia patients". In: *Thrombosis Research* 184 (2019), pp. 31–37.
- [58] Peter William Collins, VS Blanchette, K Fischer, Sven Björkman, M Oh, S Fritsch, P Schroth, G Spotts, Jan Asterman, B Ewenstein, et al. "Break-through bleeding in relation to predicted factor VIII levels in patients receiving prophylactic treatment for severe hemophilia A". In: *Journal of Thrombosis and Haemostasis* 7.3 (2009), pp. 413–420.
- [59] Laura H Bukkems, Siv Jönsson, Marjon H Cnossen, Mats O Karlsson, Ron AA Mathôt, OPTI-CLOT studies, and the SYMPHONY consortium. "Relationship between factor VIII levels and bleeding for rFVIII-SingleChain in severe hemophilia A: A repeated time-to-event analysis". In: *CPT: Pharmacometrics & Systems Pharmacology* 12.5 (2023), pp. 706–718.
- [60] IEM Den Uijl, K Fischer, JG Van Der Bom, DE Grobbee, FR Rosendaal, and I Plug. "Analysis of low frequency bleeding data: the association of joint bleeds according to baseline FVIII activity levels". In: *Haemophilia* 17.1 (2011), pp. 41–44.
- [61] Laura H Bukkems, Olav Versloot, Marjon H Cnossen, Siv Jönsson, Mats O Karlsson, Ron AA Mathôt, and Kathelijn Fischer. "Association between sports participation, factor VIII levels and bleeding in hemophilia A". In: *Thrombosis and haemostasis* 123.03 (2023), pp. 317–325.

- [62] Anouk AMT Donners, Carin MA Rademaker, Lisanne AH Bevers, Alwin DR Huitema, Roger EG Schutgens, Toine CG Egberts, and Kathelijn Fischer. "Pharmacokinetics and associated efficacy of emicizumab in humans: a systematic review". In: *Clinical Pharmacokinetics* 60.11 (2021), pp. 1395–1406.
- [63] Michael U Callaghan, Claude Negrier, Ido Paz-Priel, Tiffany Chang, Sammy Chebon, Michaela Lehle, Johnny Mahlangu, Guy Young, Rebecca Kruse-Jarres, Maria Elisa Mancuso, et al. "Long-term outcomes with emicizumab prophylaxis for hemophilia A with or without FVIII inhibitors from the HAVEN 1-4 studies". In: *Blood, The Journal of the American Society of Hematology* 137.16 (2021), pp. 2231–2242.

Part VI
APPENDIX

DATA AVAILABILITY

Personal data of patients collected as part of clinical trials, retrospective analyses, or other sources of data that are used in experiments or are otherwise mentioned as part of this thesis are private, in line with each patient's legal rights concerning their privacy. All simulated data and model code is made available as part of online repositories hosted on GitHub. An overview of the corresponding links are replicated below:

- Chapter two:
<https://github.com/Janssena/SI-AIEP-paper>
- Chapter three:
<https://github.com/Janssena/pkSHAP>
- Chapter four:
<https://github.com/Janssena/DeepCompartmentModels.jl/tree/old>
- Chapter five:
<https://github.com/Janssena/dcm-constrained>
- Chapter six:
<https://github.com/Janssena/ME-DCM.jl>
- Chapter seven:
<https://github.com/Janssena/DeepFVIII.jl>
- Chapter eight:
<https://github.com/Janssena/PerioperativeFVIII>
- Chapter nine:
<https://github.com/Janssena/RTTE-FVIII>

SUMMARY

GENERAL INTRODUCTION

In chapter 1, we provide background on haemophilia A, a rare X-linked bleeding disorder. Individuals with haemophilia A have an elevated risk of (spontaneous) bleeding and those with severe bleeding phenotypes require life-long prophylactic treatment in order to counteract debilitating joint damage and life-threatening bleeding. We discuss the history of treatment and the role of pharmacokinetics (PK) in the personalisation of prophylactic treatment. We end with presenting future perspectives for the treatment of haemophilia A, including developments in non-factor replacement therapies.

In the second part of the introduction we discuss notable developments within the field of machine learning, and elaborate on its potential for pharmacometric applications. We introduce the main three machine learning algorithms relevant to this thesis: random forests, neural networks, and Gaussian Processes. We end the chapter with a discussion of the concept of overfitting and the domain specific challenges hindering the adoption of machine learning in pharmacometrics. The main challenges are related to the sparsity and irregularity of observed data as well as the strict reliability and interpretability requirements relevant to medical applications. These barriers underline the need for algorithms specifically adapted to the pharmacometric setting.

PART I: MACHINE LEARNING IN PHARMACOMETRICS

In chapter 2, we discuss recent applications of machine learning algorithms within the context of pharmacometrics. We review the recent literature regarding the use of machine learning to support data preparation, hypothesis generation, and predictive modelling. At the end of each discussion, we provide important considerations before applying machine learning for the specified purpose. With respect to data preparation, we mainly identify methods supporting missing data imputation in high-dimensional or otherwise complex data sets. For hypothesis generation, most published research discusses the use of machine learning algorithms for performing covariate screening. This

is especially useful in high-dimensional settings, as is for example the case with genomics data. We then discuss the use of machine learning for predictive modelling. Here, we highlight the potential of differential equation based methods, such as NeuralODEs and hybrid models. We end with a discussion on model validation, which should play a principal role in any project involving machine learning. Since most algorithms are prone to overfitting, model biases and generalisation should be evaluated thoroughly. This is especially important in the context of medical applications, where it is important that the model can be trusted to not make incorrect predictions that can hurt patients.

In chapter 3, we propose a more advanced method for performing covariate screening. Instead of only ranking the covariates based on a measure of feature importance, we used explainable AI tools to visualise the relationships between covariates and PK parameters as implicitly learned by a machine learning model. To showcase the use of such a method, we fit a random forest model to predict individual estimates of PK parameters obtained from a retrospective data set of 119 haemophilia A patients who have undergone surgery. We then used SHapley Additive exPlanations (SHAP) to visualise the learned covariate effects from the random forest model. These visualisations can provide the user with an initial set of hypotheses for the implementation of covariates and helps to detect those with potentially unimportant effects (e.g. due to data artefacts). We found that the resulting visualisations adhered to our expectations regarding the effect of each covariate given prior knowledge on FVIII PK.

PART II: DEEP COMPARTMENT MODELS

In chapter 4, we introduce deep compartment models (DCM), a modelling framework that combines deep learning with universal differential equations. In this paper, we propose using neural networks to predict the parameters of a differential equation (e.g. a compartment model). This framework simplifies model development by automating the implementation of covariates, enables reliable extrapolation to new dosing regimens, and offers interpretable predictions (e.g. PK parameters). We perform a simulation experiment to show that this model can still perform well in sparse data settings, especially when initial guesses for the PK parameters are provided. We then compare its performance to a previous population PK model which was developed on a retrospective data set of 119 haemophilia A patients undergoing surgery. The accuracy of predictions from both models

was compared on an external perioperative data set of 62 haemophilia A patients collected during the OPTI-CLOT clinical trial. Here, we found that the performance of the DCM matched that of the previous model, even though it was developed in only a fraction of the time (population PK models often require weeks of development).

In chapter 5, we build upon the DCM framework by proposing the use of model constraints to improve performance on very sparse data sets. We show that without constraints, these models are capable of predicting unrealistic concentration-time curves in small data sets. Setting bounds on the PK parameters or using global parameters for hard-to-identify parameters improves accuracy and prevents the learning of unrealistic models. These results make the method a more promising alternative to classical non-linear mixed effects models, while subverting the typical assumption that machine learning methods require large data sets to achieve good performance. We also show that standard neural network architectures risk learning false effects in the presence of unimportant covariates. To tackle this issue, we suggest linking covariates to specific PK parameters (for example based on causal graphs). An added benefit is that the model becomes fully interpretable, since covariate effects can be isolated and visualised. This greatly improves model trust, while covariate effects can be critiqued to improve model performance.

In chapter 6, we propose an approach for mixed-effects estimation using the DCM. In this context, we compare the performance of classical first-order approximations of the marginal likelihood to a Variational Inference (VI) based method. We perform a simulation experiment to show that the principal method for estimating mixed effects in non-linear mixed effects model, the first-order conditional estimation (FOCE) method, results in erratic behaviour during optimisation when using the DCM. In contrast, VI results in a fast and stable convergence to accurate estimates of the population parameters. We repeat our experiment on real-world data from the OPTI-CLOT clinical trial, where results from the synthetic experiments are replicated. The resulting modelling framework represents a promising alternative to classical non-linear mixed effects models that can directly learn covariate effects from data.

PART III: MACHINE LEARNING FOR IMPROVING THE TREATMENT OF HAEMOPHILIA A PATIENTS

In chapter 7, we apply the DCM to predict FVIII PK in haemophilia A patients in the prophylactic setting. Our aim was to adopt techniques from causal inference in order to improve extrapolation performance, specifically when switching recombinant FVIII (rFVIII) concentrates. We first defined the causal graph representing the relationships between variables relevant to FVIII PK. Based on this graph, we fit a DCM to data of 103 severe haemophilia A patients treated using lonoctocog alfa. Next, we used an external data set of 40 patients to evaluate model performance. Importantly, patients in this data set received different rFVIII concentrates (octocog alfa and turoctocog alfa), measured using a different assay (one-stage instead of chromogenic assay). In addition, this data set had a high rate of missing data for von Willebrand factor antigen (VWF:Ag) levels. The model was thus augmented with components that corrected for differences between rFVIII concentrates and measurement assays. The model was also augmented with a generative component to impute missing values. The resulting model achieved higher accuracy on the external data set compared to previous models specifically trained on similar data. These results indicate that the adoption of techniques from causal inference might be beneficial for the development of population PK models. Specifically, we show that resulting models can potentially be used to estimate drug exposure of different rFVIII concentrates based on a single PK profile.

In chapter 8, we describe a population PK model that can be used to predict FVIII exposure during and after medical procedures. To this end, we use the method described in chapter 3 to find covariates that were predictive of differences in FVIII PK between the prophylactic and perioperative setting. The perioperative model corrects individual PK parameter estimates from the prophylactic setting to better match FVIII PK after surgery. We found that perioperative FVIII clearance was generally lower compared to the prophylactic setting, with covariates related to the complexity of the procedure indicating larger decreases in clearance. Next, we found that there were time-related discrepancies between model predictions and observed FVIII levels. We thus also fitted subject-specific Gaussian Processes to capture potential time-dependent changes in FVIII clearance. We found that roughly half of patients presented with potentially relevant changes ($>15\%$) in FVIII clearance. Effects were highly individual, although

larger changes in clearance were generally observed when subjects underwent more complex medical procedures. The resulting model depicted markedly improved prediction accuracy compared to the typical approach of using PK parameters from the prophylactic setting (mean absolute percentage error of 10.3% versus 26.3%). This study shows that the a priori selection of optimal postoperative treatment regimens is complicated by the presence of inter-individual variability in the response to surgery (in terms of changes in PK) as well as time-dependent effects on FVIII clearance. The proposed approach can be used to optimise treatment in real-time, although frequent measurement of FVIII levels is likely still required.

In chapter 9, we describe the development of a repeated time-to-event (RTTE) model that can be used to personalise treatment of haemophilia A patient based on bleeding risk. Typically, personalisation of treatment is achieved through PK-guided dosing, where the optimal dosing regimen is determined based on drug exposure and pre-specified FVIII target levels. In this work, we suggest using RTTE models to estimate the individual bleeding risk and to use this estimate to simulate projected annual bleeding rates based on a specific dosing regimen. This enables the selection of optimal treatment based on bleeding outcomes rather than just FVIII (target) levels. To improve upon previous models, we fit multiple RTTE models to predict risk for specific bleeding categories (damage-causing and nuisance bleeds), and fit the random effects using a Gaussian mixture model to learn the variability in the bleeding risk for patients with low, medium, and high bleeding frequencies. We fit the model to a data set of 264 severe haemophilia A patients with a median follow-up time of 881 days and 3106 total observed bleeds. Model evaluation showed that a large proportion (>70%) of median projected bleeding rates were within one bleed of the true observed bleeding rate. We then showcase how the method can be used to compare the bleeding outcomes of three alternative dosing regimens for a single patient. Here we see that different dosing regimens with similar total weekly rFVIII consumption can nonetheless result in different outcomes based on the timing of specific doses. In conclusion, the ability to select the optimal dosing regimen based on bleeding outcomes introduces an exciting new paradigm for the personalised treatment of haemophilia A.

PART IV: THE OPTI-CLOT WEB-PORTAL

In chapter 10, we introduce the OPTI-CLOT web-portal, a free web application that enables treatment teams of patients with rare bleeding disorders to request dosing advice. The web-portal enables the secure transfer of patient information after which a team of OPTI-CLOT pharmacometrists create an initial report containing the individually tailored dosing advice. The requesting team can discuss the options with their patient and can inquire about alternative dosing regimens if so desired. The final report is stored in the web-portal, so that the complete history of previous advice is always available to all members of the treatment team. In the future, we aim to also provide dosing advice based on individual bleeding phenotypes. In addition, we hope to reduce the burden of entering data by automatically extract necessary patient information from electronic health record systems. This way, we hope to increase the adoption of PK or PD-guided dosing for patients with rare bleeding disorders.

GENERAL DISCUSSION

Finally, in chapter 11, we review the most important findings discussed in this thesis and offer perspectives. We provide an overview of the approaches for implementing machine learning algorithms in pharmacometrics and discuss benefits and limitations. We highlight the potential of hybrid methods, where machine learning algorithms are combined with prior knowledge of pharmacometrics to improve prediction accuracy and data efficiency. The deep compartment model is one such method, and work performed in this thesis support the method as a reliable and robust alternative to classical non-linear mixed-effect models.

Next, we discuss current open issues for the personalised treatment of haemophilia A and offer insights into machine learning based approaches that could offer solutions. Several of these issues will be tackled as part of the PHEMS consortium, a research initiative launched by the European Children's Hospitals Organisation (ECHO). We hope to bring these and other novel methods for the personalisation of treatment of patients with haemophilia A to the OPTI-CLOT web-portal in order to facilitate their adoption in clinical practice.

To conclude, we have shown how the use of machine learning algorithms can result in clinical benefits, and hopefully inspire others to implement similar algorithms in the context of other rare diseases.

SAMENVATTING

ALGEMENE INTRODUCTIE EN DOEL VAN HET PROEFSCHRIFT

In hoofdstuk 1 geven we achtergrondinformatie over hemofilie A, een zeldzame bloedstollingsstoornis waarbij patiënten een tekort hebben aan stollingsfactor VIII. De ziekte is *X-chromosoom gebonden*, wat inhoudt dat de ziekte vooral mannen treft. Mensen met hemofilie A hebben een verhoogd risico op (spontane) bloedingen en hebben levenslange behandeling nodig om gewrichtsschade en levensbedreigende bloedingen te voorkomen. De meeste patiënten worden behandeld met kunstmatige vormen van factor VIII (zogeheten 'recombinant' factor VIII concentraten) om te zorgen dat het bloed weer kan stollen om eventuele bloedingen te stoppen (*on demand* behandeling). Patiënten die frequent bloeden worden 'profylactisch' (preventief) behandeld. Dit houdt in dat de patiënt enkele keren per week factor VIII 'intraveneus' (in de aderen) moet toedienen om het risico op bloedingen te verlagen. Dit wordt vooral bij jonge kinderen als moeilijk of vervelend ervaren. In Nederland valt het overgrote deel van de medicatie die bij hemofilie A wordt toegediend onder de categorie geneesmiddelen.

Naast ongemak bij het toedienen van medicatie en de hoge kosten van de behandeling zijn ook grote verschillen in het effect van de toegediende factor VIII concentraten tussen patiënten een probleem. Patiënten die dezelfde hoeveelheid medicijn per kilogram lichaamsgewicht krijgen kunnen verschillen in de concentratie van factor VIII in het bloed, wat vervolgens leidt tot verschillen in de effectiviteit. Daarnaast kunnen er ook verschillen zijn in het bloedingsrisico van patiënten, zelfs wanneer de concentratie van factor VIII in het bloed vergelijkbaar is: de ene patiënt bloedt wel bij een lage concentratie en een ander niet. Dit laat zien dat het belangrijk is om voor iedere patiënt een behandeling op maat op te stellen waarbij een op de persoon-gerichte dosering en doseerfrequentie wordt vastgesteld. Bij het vaststellen van deze persoonlijke behandel schema's wordt gebruik gemaakt van de 'farmacokinetiek' (PK; Engels *pharmacokinetics*). De PK beschrijft de processen van absorptie, verdeling en klaring (afbraak/uitscheiding) van geneesmiddelen in het lichaam en wordt gebruikt om met behulp van wiskundige modellen de concentratie

Het gen voor factor VIII bevindt zich op het X-chromosoom, en mannen (XY) hebben maar een enkele kopie van het gen en dus een grotere kans op fouten.

van medicatie in het bloed te voorspellen. Deze methode wordt in Nederland bij de behandeling van hemofilie A met factor VIII concentraten veel toegepast.

Machine learning is een vorm van kunstmatige intelligentie waarbij wiskundige modellen worden gebruikt die zelfstandig van gegevens kunnen leren om zo beter te presteren op een bepaalde taak.

Bij big data zijn grote hoeveelheden, vaak complexe gegevens beschikbaar die moeilijker te analyseren zijn met klassieke statistische methodes.

In dit proefschrift wordt onderzocht of de toepassing van *machine learning* technieken de persoonsgerichte behandeling van patiënten met hemofilie A kan verbeteren. Onze focus is hierbij specifiek op het gebruik van zulke technieken voor 'farmacometrische' toepassingen. De farmacometrische wetenschap houdt zich bezig met de relatie tussen de doseringen van een geneesmiddel, de concentratie in het bloed (de PK, zoals hierboven beschreven) en het effect op de ziekte (farmacodynamiek, PD; Engels *pharmacodynamics*). Het is belangrijk om rekening te houden met de individuele PK en PD gezien het de keuze voor de juiste dosering voor de juiste patiënt met wetenschappelijk bewijs kan ondersteunen. Het gebruik van *machine learning* technieken in deze context maakt het mogelijk om complexere relaties bloot te leggen en zo de behandeling te verbeteren.

In het eerste en tweede deel van het proefschrift (hoofdstuk 2–6) richten we ons op de ontwikkeling van methodes die betrouwbare voorspellingen kunnen geven wanneer data schaars is. *Machine learning* methodes worden vaak in verband gebracht met het concept van *big data*, waarbij het succes van deze methodes bepaald wordt door de hoeveelheid data die beschikbaar is. Bij zeldzame aandoeningen zoals hemofilie is er per definitie weinig data beschikbaar omdat de prevalentie van de ziekte laag is. Om *machine learning* in deze context te kunnen gebruiken zijn er dus specifiek ontwikkelde methodes nodig. Daarnaast is het belangrijk dat deze methodes betrouwbaar en interpreteerbaar zijn, vooral wanneer deze gebruikt worden voor medische toepassingen. De meeste *machine learning* modellen zijn een zogeheten *black-box*, wat betekent dat de innerlijke werking van deze modellen onbekend is. Het is dus belangrijk om vast te stellen of het model nuttige informatie geleerd heeft en correcte voorspellingen geeft.

In deel drie van het proefschrift (hoofdstuk 7–9) focussen we op drie aspecten van de behandeling van hemofilie A waar de toepassing van *machine learning* technieken kan leiden tot verbeteringen.

Ten slotte introduceren we in deel vier (hoofdstuk 10) het OPTI-CLOT webportaal, een website waar we behandelaren van patiënten met zeldzame bloedstollingsstoornissen persoonlijke behandeladviezen aanbieden.

DEEL I: MACHINE LEARNING IN PHARMACOMETRICS

In hoofdstuk 2 worden recente toepassingen van *machine learning* methodes binnen de farmacometrie besproken. We behandelen de recente literatuur over het gebruik van *machine learning* met betrekking tot de voorbereiding van datasets vóór analyse, het genereren van hypotheses, en de ontwikkeling van modellen. Met betrekking tot de voorbereiding van datasets identificeren we voornamelijk methodes die ontbrekende gegevens kunnen invullen om zo een dataset compleet te krijgen. Bij het genereren van hypotheses zien we dat *machine learning* technieken gebruikt kunnen worden om belangrijke 'covariaten' (voorspellers zoals lichaamsgewicht of leeftijd) die gerelateerd zijn aan de uitkomst van interesse, zoals factor VIII concentraties, te vinden. Dit is vooral handig wanneer er een groot aantal covariaten gescreend moeten worden: *machine learning* methodes kunnen bijvoorbeeld snel de tien belangrijkste covariaten identificeren die door de onderzoeker verder getoetst kunnen worden. Vervolgens bespreken we het gebruik van *machine learning* voor het ontwikkelen van farmacometrische modellen. Hier benadrukken we de potentie van methodes gebaseerd op differentiaalvergelijkingen, zoals NeuralODEs en hybride modellen, die mogelijk efficiënter met de beschikbare gegevens om kunnen gaan.

Ten slotte eindigen we dit hoofdstuk met het bespreken van manieren om *machine learning* modellen te valideren. Dit is van groot belang, aangezien deze modellen complexe verbanden kunnen leren/beschrijven die niet per se de realiteit weerspiegelen. Zo kan een *machine learning* model heel goed werken op een specifieke dataset door simpelweg te onthouden wat de uitkomst is *dat bij elk voorbeeld hoort*. Wanneer het model in de praktijk gebruikt wordt heeft het deze nieuwe gegevens vaak nog niet eerder gezien en zullen voorspellingen waarschijnlijk incorrect zijn. Een uitgebreide evaluatie van *machine learning* modellen is met name belangrijk in de context van medische toepassingen, waar het cruciaal is dat het model betrouwbare voorspellingen maakt die geen schade aan patiënten toebrengen.

In hoofdstuk 3 presenteren we een geavanceerdere methode om de relatie tussen covariaten en PK parameters te visualiseren. Deze methode kan worden gebruikt om de effecten van covariaten beroep te leggen en zo verschillen tussen patiënten te verklaren. In plaats van alleen de covariaten te rangschikken op basis van een maat voor hoe belangrijk deze zijn in het model (zoals bij andere methodes), gebruiken wij zogeheten *explainable AI* methodes om de covariaat effecten te visualiseren zoals deze impliciet door een *machine learning*

Dit concept staat bekend als 'overfitting', waarbij het model compleet is toegespitst op de specifieke dataset.

Explainable AI modellen proberen voorspellingen van een ander *machine learning* model uit te leggen.

model zijn geleerd. Om het gebruik van een dergelijke methode te illustreren, hebben we een *random forest* model (een *machine learning* methode) ontwikkeld die individuele schattingen van PK parameters voorspeld op basis van een retrospectieve dataset van 119 hemofilie A patiënten die een operatie hebben ondergaan. Vervolgens gebruikten we '*SHapley Additive exPlanations*' om de geleerde covariaat effecten van het *random forest* model te visualiseren. Deze visualisaties kunnen onderzoekers helpen om mogelijk belangrijke covariaten te detecteren. We constateerden dat de resulterende visualisaties overeenkwamen met onze verwachtingen van de effecten van de geteste covariaten op basis van eerdere kennis over de PK van factor VIII.

DEEL II: DEEP COMPARTMENT MODELS

In hoofdstuk 4 introduceren we het 'diepe compartimentenmodel' (DCM; Engels *deep compartment model*), een techniek voor het ontwikkelen van farmacometrische modellen waarbij neurale netwerken gecombineerd worden met universele differentiaalvergelijkingen. In dit hoofdstuk onderzoeken we de toepassing van deze *machine learning* modellen om bijvoorbeeld de PK parameters van een compartimentenmodel te schatten. Deze aanpak vereenvoudigt de ontwikkeling van modellen door de relatie tussen covariaten en PK parameters automatisch te leren op basis van patiëntgegevens, maakt betrouwbare extrapolatie naar nieuwe doseringsregimes mogelijk en biedt bovendien interpreteerbare voorspellingen (de voorspelde PK parameters van het geneesmiddel kunnen bijvoorbeeld vergeleken worden met eerder onderzoek). We voeren een simulatie-experiment uit om te laten zien dat dit model nog steeds goed kan presteren in situaties waarbij slechts een beperkte hoeveelheid patiëntgegevens beschikbaar zijn. Vervolgens vergelijken we de prestaties van onze techniek met een eerder gepubliceerd populatie PK model. Beide modellen zijn ontwikkeld op basis van een retrospectieve dataset van 119 hemofilie A patiënten die een operatie hebben ondergaan. De nauwkeurigheid van voorspelde factor VIII concentratie voorspellingen van beide modellen wordt bepaald op basis van gegevens van een externe *perioperatieve* dataset van 62 hemofilie A patiënten, verzameld tijdens het prospectieve OPTI-CLOT onderzoek. Hier vonden we dat de voorspellingen van het DCM minstens even accuraat waren als die van het eerdere model. Een voordeel van het DCM was dat het echter in een veel kortere tijd ontwikkeld kon worden vergeleken met

Een compartimentenmodel versimpelt het lichaam in compartimenten (zoals de maag en het bloed) om zo de spreiding van een geneesmiddel door het lichaam schematisch weer te geven.

Perioperatief verwijst naar de tijd rond een operatie.

de gemiddelde tijd dat wordt besteed aan de ontwikkeling van een (klassiek) populatie PK model.

In hoofdstuk 5 bouwen we voort op het DCM. Binnen het model maken we het mogelijk om grenzen aan te geven voor de PK parameters zodat deze binnen realistische waarden blijven. Zo is het gemiddelde bloedvolume van een volwassene tussen de vier tot zes liter, en kan het dus nuttig zijn om het model tegen te houden wanneer deze voorspellingen maakt die lager zijn dan bijvoorbeeld één liter. Daarnaast maken we het mogelijk om een enkele waarde voor de PK parameters te schatten voor alle patiënten wanneer een parameter moeilijk te identificeren is op basis van de gebruikte dataset. We laten zien dat deze aanpak nodig kan zijn om onrealistische voorspellingen te voorkomen wanneer er beperkte gegevens beschikbaar zijn. De resultaten van deze studie laten zien dat het gebruik van zulke beperkingen het DCM een veelbelovend alternatief maakt voor klassieke populatie PK modellen. Daarnaast is het duidelijk dat het niet nodig is om de beschikking te hebben over een grote dataset (zoals bij big data) om het DCM goed te laten presteren.

In dit hoofdstuk laten we ook zien dat standaard neurale netwerken het risico lopen om valse effecten te leren wanneer onbelangrijke covariaten in het model gebruikt worden. Om dit probleem aan te pakken stellen we voor om de covariaten te koppelen aan specifieke PK parameters. Deze connecties kunnen gebaseerd worden op eerdere kennis. Zo is het bekend dat een ander eiwit in het bloed, von Willebrand factor, kan binden aan factor VIII om het vervolgens te beschermen tegen afbraak. Von Willebrand factor heeft dus een effect op de klaring van factor VIII, en dit soort effecten kunnen specifiek worden toegevoegd aan het model. Een bijkomend voordeel van deze aanpak is dat het model volledig interpreteerbaar wordt, aangezien de relatie tussen elk covariaat en de PK parameters geïsoleerd en gevisualiseerd kan worden. Dit vergroot het vertrouwen in het model, en bevordert de adoptie voor medische toepassingen gezien gebruikers kunnen verklaren waarom het model specifieke voorspellingen maakt.

In hoofdstuk 6 beschrijven we een verdere verbetering van het DCM door het mogelijk te maken om de verschillen tussen patiënten te beschrijven door middel van '*mixed-effects*'. In '*fixed-effect*' modellen (zoals het originele DCM) is de voorspelling voor twee patiënten hetzelfde als de waarde van de covariaten gelijk zijn. Vaak kunnen deze covariaten niet alle verschillen tussen patiënten verklaren en wordt er een *random-effect* aan het model toegevoegd om de overgebleven variatie te bepalen. In de praktijk kan deze informatie

Bij een mixed-effect model worden verschillen in PK parameters beschreven als een gemiddelde waarde plus het effect van covariaten (het fixed-effect) en een patiënt-specifiek effect (het random-effect)

vervolgens gebruikt worden om voorspellingen te corrigeren op basis van gemeten medicijn concentraties. Wanneer factor VIII spiegels bij voorbeeld sneller dalen dan voorspeld, kan het *random-effect* gebruikt worden om de voorspelde factor VIII klaring te verhogen om zo een individuele schatting te krijgen die beter bij de metingen past. De resulterende individuele PK parameters kunnen vervolgens gebruikt worden om nauwkeurigere simulaties uit te voeren van de gevolgen van verschillende behandelingen. In dit hoofdstuk vergelijken we de prestaties van klassieke statische methodes voor het schatten van random effects met een techniek uit machine learning: '*Variational Inference*'. We voeren een simulatie-experiment uit om te laten zien dat de meestgebruikte klassieke methode, de '*first-order conditional estimation*' (FOCE) methode, onvoorspelbaar gedrag vertoont wanneer het gebruikt wordt in combinatie met het DCM. Daarentegen is het gebruik van *Variational Inference* sneller, stabieler en nauwkeuriger. We herhalen ons experiment op klinische gegevens verzameld tijdens het prospectieve OPTI-CLOT onderzoek, waar we de resultaten van de simulatie-experimenten konden repliceren. Deze uitbreiding van het DCM maakt het een veelbelovender alternatief voor klassieke populatie PK modellen.

DEEL III: MACHINE LEARNING VOOR HET VERBETEREN VAN DE BEHANDELING VAN PATIËNTEN MET HEMOFILIE A

In hoofdstuk 7 passen we het DCM toe om de PK van factor VIII te voorspellen bij hemofylie A patiënten tijdens profylactische behandeling. Het doel van dit onderzoek was om te evalueren of het mogelijk is om een model te ontwikkelen dat kan corrigeren voor verschillen tussen recombinant factor VIII concentraties. Dit zou het mogelijk maken om het verloop van de bloedconcentratie in de tijd van een nieuw middel te voorspellen op basis van de gegevens van een middel dat eerder gebruikt is. Wanneer de bloedconcentraties van het huidige middel onvoldoende zijn, zou dit model gebruikt kunnen worden om een ander middel te selecteren dat wel de gewenste blootstelling heeft. Om dit model te ontwikkelen hebben we eerst een causaal diagram opgesteld die alle (inter)relaties tussen relevante covariaten weergeeft. Op basis van de relaties in dit diagram hebben we een DCM ontwikkeld met behulp van gegevens van 103 ernstige hemofylie A patiënten die behandeld werden met lonoctocog alfa (een specifiek factor VIII concentraat). Vervolgens gebruikten we een externe dataset van 40 patiënten om de nauwkeurigheid van voorspel-

lingen te evalueren. Een belangrijk gegeven in deze vergelijking was dat patiënten in deze externe dataset andere factor VIII concentratessen kregen toegevoegd (octocog alfa en turoctocog alfa). Bovendien werd de factor VIII concentratie bepaald meteen andere laboratorium test, de zogeheten 'one-stage' test in plaats van de chromogene test. Het model werd daarom uitgebreid met componenten die corrigeerden voor verschillen tussen de verschillende factor VIII concentratessen en meetmethodes. Ten slotte ontbraken in deze dataset voor een groot deel van de patiënten informatie over de von Willebrand factor bloedconcentraties. Om dit probleem op te lossen hebben we een *generatief* component aan ons model toegevoegd dat ontbrekende waarden voor alle covariaten in het model in zou kunnen vullen. Het resulterende model was in staat de factor VIII concentraties in de externe dataset met een hoge nauwkeurigheid te voorspellen en liet een verbetering zien in vergelijking met eerdere modellen die specifiek op soortgelijke data waren ontwikkeld. Deze resultaten laten zien dat het inderdaad mogelijk is om modellen te ontwikkelen die gebruikt kunnen worden om de PK van meerdere verschillende factor concentratessen te voorspellen.

In hoofdstuk 8 beschrijven we een populatie PK model dat kan worden gebruikt om factor VIII concentraties tijdens en na medische ingrepen te voorspellen. Daartoe gebruiken we de methode beschreven in hoofdstuk 3 om covariaten te vinden die voorspellend waren voor verschillen in factor VIII PK parameters tijdens profylaxe en na een operatie. Het perioperatieve model corrigeert individuele PK parameters geschat op basis van profylactische gegevens zodat voorspellingen beter overeenkomen met factor VIII concentraties na een medische ingreep. Met behulp van dit model kunnen patiënten na een operatie beter ingesteld worden om *gewenste factor VIII concentraties* te behalen. Onze resultaten gaven aan dat de perioperatieve factor VIII klaring over het algemeen lager was na een medische ingreep, waarbij covariaten gerelateerd aan de complexiteit van de procedure grotere dalingen in de klaring lieten zien. Vervolgens ontdekten we dat er verschillen waren tussen voorspellingen van het model en de waargenomen factor VIII concentraties op specifieke tijdstippen na de operatie. Daarom hebben we ook per patiënt een *Gaussian Process* model (een *machine learning* methode) gebruikt om mogelijke veranderingen in factor VIII klaring over de tijd te identificeren. We vonden dat ongeveer de helft van de patiënten potentieel relevante veranderingen ($>15\%$) in factor VIII klaring vertoonden na de operatie. Dit effect kon sterk variëren tussen patiënten. Over het algemeen wer-

Generatieve modellen kunnen kunstmatige data creëren dat lijkt op echte data. ChatGPT is een voorbeeld van een generatief model.

Er zijn in Nederland protocollen opgesteld die aangeven dat hogere factor VIII concentraties nodig zijn om tijdens en na een operatie te beschermen tegen bloedingen.

den er grotere veranderingen in de klaring waargenomen wanneer patiënten complexere medische ingrepen ondergingen. Ons model resulteerde in een aanzienlijke verbetering van de nauwkeurigheid van factor VIII voorspellingen in vergelijking met de typische aanpak waarbij profylactische PK parameters worden gebruikt (gemiddelde absolute percentuele fout van 10,3% versus 26,3%). Deze studie toont aan dat de selectie van de optimale perioperatieve behandeling vóór aanvang van de ingreep wordt bemoeilijkt door de aanwezigheid van inter-individuele variabiliteit in de PK parameters en veranderingen in factor VIII klaring over de tijd. De voorgestelde aanpak kan worden gebruikt om de behandeling met factor VIII na een medische ingreep te optimaliseren, hoewel frequente meting van factor VIII concentraties in het bloed waarschijnlijk nog steeds vereist is.

Een time-to-event model voorspeld hoe het risico op een bepaalde gebeurtenis zich ontwikkelt over de tijd. In een repeated time-to-event model kan de gebeurtenis meerdere keren plaatsvinden.

In hoofdstuk 9 beschrijven we de ontwikkeling van een zogeheten *repeated time-to-event* (RTTE-)model dat kan worden gebruikt om de behandeling van hemofilie A patiënten te personaliseren op basis van het individuele bloedingsrisico. Doorgaans wordt personalisatie van de behandeling bereikt door te doseren op basis van vooraf bepaalde factor VIII concentraties (PK-gestuurd doseren). In dit hoofdstuk stellen we voor om met behulp van een RTTE-model het individuele bloedingsrisico te schatten, wat vervolgens gebruikt kan worden om de verwachte jaarlijkse bloedingsfrequentie te voorspellen op basis van een specifiek en individueel aangepast doseringsregime. Dit stelt ons in staat om de optimale behandeling te selecteren op basis van de hoeveelheid voorspelde bloedingen, in plaats van enkel op basis van factor VIII concentraties. Om eerdere RTTE-modellen te verbeteren, hebben we meerdere modellen ontwikkeld om bloedingen behorende tot specifieke categorieën te voorspellen. Hierbij maken we onderscheid tussen beschadigende bloedingen (bijv. in gewrichten) en kleinere bloedingen (zoals neusbloedingen). Een patiënt met een grotere hoeveelheid gewrichtsbloedingen heeft doorgaans een ernstiger bloedingsfenotype dan patiënten die frequent blauwe plekken of neusbloedingen hebben. Om een correcte schatting van het bloedingsrisico te krijgen is het dus belangrijk om onderscheid te maken in de categorieën van doorgevoerde bloedingen. Daarnaast maken we gebruik van een '*Gaussian mixture model*' om de variatie in het bloedingsrisico te beschrijven voor verschillende subgroepen van patiënten. Het model deelt patiënten in groepen met lage, gemiddelde, en hoge frequenties aan bloedingen om zo de verschillen tussen patiënten in deze subgroepen te verlagen. Het model is ontwikkeld op basis van een dataset van 264 ernstige

hemofilie A patiënten met een mediane follow-up van 881 dagen en in totaal 3106 bloedingen.

We hebben vervolgens het model geëvalueerd door voorspellingen van de jaarlijkse bloedingsfrequentie te vergelijken met de geobserveerde bloedingsfrequentie voor de verschillende categorieën van bloedingen. Hier zagen we dat het grootste deel van de voorspellingen (>70%) binnen één bloeding van de daadwerkelijke jaarlijkse bloedingsfrequentie lag. Deze methode kan worden gebruikt om de bloedingsuitkomsten van verschillende doseringsregimes te vergelijken voor een individuele patiënt. We laten zien hoe het model onderscheid kan maken tussen de bloedingsuitkomsten van drie verschillende doseringsregimes die vergelijkbare hoeveelheden factor VIII concentraat gebruiken per week. Hieruit blijkt dat het tijdstip waarop de specifieke doseringen worden toegediend een belangrijk effect kan hebben op het bloedingsrisico. Geconcludeerd kan worden dat de mogelijkheid om het optimale doseringsregime te selecteren op basis van bloedingsuitkomsten een veelbelovend nieuwe aanpak vormt voor de gepersonaliseerde behandeling van hemofilie A.

DEEL IV: HET OPTI-CLOT WEBPORTAAL

In hoofdstuk 10 introduceren we het OPTI-CLOT webportaal, een gratis website waar behandelteams van patiënten met een zeldzame bloedstollingsstoornis zoals hemofilie of von Willebrand ziekte een doseringsadvies voor de behandeling met stollingsfactorconcentraten kunnen aanvragen. Het webportaal ondersteunt de veilige overdracht van patiëntgegevens, waarna een team van OPTI-CLOT farmacometristen een initiële rapportage opstelt met een individueel afgestemd doseringsadvies. Het behandelteam kan de opties vervolgens bespreken met de patiënt. Indien gewenst kunnen in overleg alternatieve doseringsregimes worden opgesteld. Als de keuze is bepaald wordt een eindrapport opgesteld en opgeslagen in het webportaal, zodat de volledige geschiedenis van eerdere adviezen altijd beschikbaar blijft voor alle leden van het behandelteam. Dit webportaal kan in de toekomst ook gebruikt worden om doseeradviezen op basis van het individuele bloedingsrisico te geven, zoals beschreven in hoofdstuk 9. Daarnaast willen we de tijd die het kost om gegevens in te voeren verlagen door automatisch de benodigde patiëntinformatie uit elektronische patiëntendossiers te halen. Op deze manier hopen we de toepassing van PK en PD-gestuurd doseren in de klinische praktijk.

tijk te stimuleren voor de behandeling van patiënten met zeldzame bloedstollingsstoornissen.

ALGEMENE DISCUSSIE

In hoofdstuk 11 bespreken we de belangrijkste bevindingen van dit proefschrift en bieden we toekomstperspectieven. We geven onze visie voor de implementatie van *machine learning* methodes in de farmacometrie en bespreken voordelen en beperkingen. In het bijzonder bespreken we de potentie van hybride modellen, die *machine learning* methodes combineren met eerdere kennis van de farmacometrie om zo de nauwkeurigheid van voorspellingen te verbeteren en efficiënter gebruik te maken van de beschikbare data. Het DCM is een voorbeeld van een hybride methode, en hoofdstukken in dit proefschrift ondersteunen het gebruik van deze methode als een betrouwbaar en robuust alternatief voor klassieke populatie PK modellen.

Vervolgens bespreken we hoe de gepersonaliseerde behandeling van hemofilie A nog verder verbeterd kan worden en bieden we inzichten in de oplossingen die *machine learning* technieken hierbij zouden kunnen bieden. De toepassing van *machine learning* ter verbetering van de behandeling van hemofilie A wordt onder andere verder onderzocht in het PHEMS consortium, een onderzoekinitiatief gestart door de Europese Kinderziekenhuis organisatie (ECHO: European Children's Hospitals Organisation). In toekomstig onderzoek zullen deze en andere nieuwe methodes voor het personaliseren van de behandeling van patiënten met hemofilie A in het OPTI-CLOT webportaal ondergebracht worden om zo de toepassing in de klinische praktijk te bevorderen.

Tot slot, hopen we dat ons werk als voorbeeld zal dienen voor hoe *machine learning*-gebaseerde methodes op een betrouwbare en robuuste manier kunnen worden toegepast op problemen in de farmacometrie. We hebben laten zien hoe het gebruik van deze methodes klinische voordelen kan opleveren, en hopen anderen te inspireren om vergelijkbare methodes toe te passen in de context van andere zeldzame aandoeningen.

ABOUT THE AUTHOR

Alexander Janssen was born on the 26th of March 1994 in Amsterdam, The Netherlands. He attended secondary school at the Hervormd Lyceum Zuid and received his Atheneum diploma in 2012. Given his interest in Biology, he chose to study Biomedical sciences at the University of Amsterdam. He performed his master's thesis at the University of Edinburgh, where he discovered his passion for research under the supervision of dr. Sowmya Sekizar and prof. dr. Anna Williams. In the last few months of his studies, he developed an interest in computational biology and machine learning, and rather than apply for a PhD position he chose to follow the data science traineeship program at an IT consultancy.



© Alexandra Verkerk

Pressed by his desire for greater freedom and creativity, he chose to leave the corporate sector and return to the University. In 2020, at the onset of the COVID-19 pandemic, he started a PhD position as part of the SYMPHONY consortium under supervision of prof. dr. R.A.A. Mathôt and prof. dr. M.H. Cnossen. Here, he developed skills in Mathematics, Computer Science, and Pharmacometrics. The combination of these skills mixed with his medical background resulted in an unique set of competencies, which was honoured by him receiving the Lewis Sheiner Student Award in 2023 for his work bringing machine learning techniques to haemophilia A. As part of a European Joint Programme on Rare Diseases fellowship, he spent several months further growing his technical skills and his academic network as a visiting scholar at the Cambridge Centre for Artificial Intelligence in Medicine at the University of Cambridge. After his PhD, Alexander aims to continue his academic career by applying for grants to develop physiologically-based machine learning methods in the critical care setting.

PUBLICATIONS

PUBLICATIONS INCLUDED IN THIS THESIS

- [1] Alexander Janssen et al. "Deep compartment models: a deep learning approach for the reliable prediction of time-series data in pharmacokinetic modeling". In: *CPT: Pharmacometrics & Systems Pharmacology* 11.7 (2022), pp. 934–945.
- [2] Alexander Janssen et al. "Application of SHAP values for inferring the optimal functional form of covariates in pharmacokinetic modeling". In: *CPT: Pharmacometrics & Systems Pharmacology* 11.8 (2022), pp. 1100–1110.
- [3] Alexander Janssen, Frank C Bennis, and Ron AA Mathôt. "Adoption of machine learning in pharmacometrics: an overview of recent implementations and their considerations". In: *Pharmaceutics* 14.9 (2022), p. 1814.
- [4] Alexander Janssen et al. "A Generative and Causal Pharmacokinetic Model for Factor VIII in Hemophilia A: A Machine Learning Framework for Continuous Model Refinement". In: *Clinical Pharmacology & Therapeutics* (2024).
- [5] Alexander Janssen, Frank C Bennis, Marjon H Cnossen, and Ron AA Mathôt. "On inductive biases for the robust and interpretable prediction of drug concentrations using deep compartment models". In: *Journal of Pharmacokinetics and Pharmacodynamics* (2024), pp. 1–12.
- [6] Alexander Janssen, Frank C Bennis, Marjon H Cnossen, and Ron AA Mathôt. "Mixed effect estimation in deep compartment models: Variational methods outperform first-order approximations". In: *Journal of Pharmacokinetics and Pharmacodynamics* (2024), pp. 1–12.

OTHER PUBLICATIONS BY THE AUTHOR

- [1] MD Slooter, A Janssen, WA Bemelman, PJ Tanis, and R Hompes. "Currently available and experimental dyes for intraoperative near-infrared fluorescence imaging of the ureters: A systematic review". In: *Techniques in coloproctology* 23 (2019), pp. 305–313.
- [2] Alexander Janssen, Jan J De Waele, and Paul WG Elbers. "Towards adequate and automated antibiotic dosing". In: *Intensive care medicine* 49.7 (2023), pp. 853–856.
- [3] Tine MHJ Goedhart, A Janssen, Ron AA Mathôt, Marjon H Cnossen, OPTI-CLOT Study Group, SYMPHONY Consortium, et al. "The road to implementation of pharmacokinetic-guided dosing of factor replacement therapy in hemophilia and allied bleeding disorders. Identifying knowledge gaps by mapping barriers and facilitators". In: *Blood Reviews* (2023), p. 101098.
- [4] Michael E Cloesmeijer, Alexander Janssen, Sjoerd F Koopman, Marjon H Cnossen, Ron AA Mathôt, and SYMPHONY consortium. "ChatGPT in pharmacometrics? Potential opportunities and limitations". In: *British journal of clinical pharmacology* 90.1 (2024), pp. 360–365.

PHD PORTFOLIO

A. Janssen

PhD period

01 April 2020 – 01 April 2024

PhD supervisors

prof.dr. R.A.A. Mathôt

prof.dr. M.H. Cnossen

dr. F.C. Bennis

PHD TRAINING	YEAR	WORKLOAD (ECTS)
--------------	------	-----------------

Graduate school courses

E-science	2020	1.0
Scientific writing	2022	1.5

Other courses

AI and Machine Learning in Health-care Summer School (University of Cambridge)	2023	1.0
--	------	-----

Teaching

Supervision of bachelor thesis (Medical Informatics)	2020	1.0
OOR Lecture "PK/PD van stollingsfactoren bij hemofiliebehandelingen" (Amsterdam UMC)	2022	0.5

(Inter)national conferences

European Association of Haemophilia and Allied disorders congress	2021	0.8
World Conference of Pharmacometrics	2022	0.5
Pharmacometrics Benelux Network meeting (4x)	2022–2023	1.0

PHD TRAINING	YEAR	WORKLOAD (ECTS)
<i>(Inter)national conferences (continued)</i>		
International Society on Thrombosis and Haemostasis congress	2022	1.5
Population Approach Group Europe congress (4x)	2021–2024	3.2
Ledendag Nederlandse Vereniging van Hemofilie Patiënten	2023	0.5
<i>Poster presentations</i>		
Deep compartment models: combining machine learning and differential equations for reliable drug concentration predictions (WCoP)	2022	0.5
A Bayesian optimization procedure for the automated determination of optimal limited sampling strategies (PAGE)	2022	0.5
Time-dependent random effects (PAGE)	2024	0.5
<i>Oral presentations</i>		
The neural mixed effects algorithm: leveraging machine learning for pharmacokinetic modelling (PAGE)	2021	0.5
SHAP values for inferring the optimal functional form of covariates in pharmacokinetic modelling	2022	0.5
Deep compartment models: combining machine learning and differential equations for reliable drug concentration predictions (WCoP)	2022	0.5
A generative and causal pharmacokinetic model for haemophilia A: towards an unified model for all factor VIII concentrates (PAGE)	2023	0.5
<i>Invited talks</i>		
Regression versus machine learning (Erasmus MC Medische Besliskunde meeting)	2021	0.5

PHD TRAINING	YEAR	WORKLOAD (ECTS)
<i>Invited talks (continued)</i>		
So you made a machine learning model, now what? (Amsterdam public health spring meeting)	2021	0.5
Machine learning in pharmacometrics: how novel algorithms can facilitate pharmacometric analyses (Erasmus MC NONMEM meeting)	2022	0.5
Robust and reliable machine learning for predicting drug concentrations (Erasmus MC clinical pharmacology meeting)	2022	0.5
Potential of machine learning algorithms within the context of pharmacometrics (LAP&P Science Day)	2023	0.5
Potential of machine learning algorithms within the context of pharmacometrics (Utrecht MC NURD meeting)	2024	0.5
Treatment personalization in haemophilia A: application of AI in rare disease (Emma Center for Personalized Medicine meeting)	2024	0.5
<i>Other</i>		
SYMPHONY meetings	2020–2024	2.0
OPTI-CLOT meetings, monthly	2020–2024	2.0
NONMEM lab meetings, monthly	2020–2024	4.0
Pharmacy journal club, monthly	2022	1.0
Internship at machine learning group University of Cambridge, 4 months	2023	5.0
<i>Grants and awards</i>		
PAGE Stuart Beal methodology session	2021	—
European joint program rare disease fellowship	2023	—
PAGE Lewis Sheiner student session award	2023	—

DANKWOORD

Vanzelfsprekend heb ik al dit werk niet in mijn eentje verricht vanuit het comfort van mijn woon- of studeerkamer (zoals dat ging in jaar 1 en 2). Dit werk is uiteindelijk de culminatie van de inzet van talloze collega's en patiënten die hun tijd beschikbaar hebben gesteld voor de wetenschap. Dit dankwoord is een poging om zo veel mogelijk van deze mensen te benoemen en te bedanken.

Allereerst mijn promotieteam, mijn promotoren prof. dr. Ron Matthôt & prof. dr. Marjon Cnossen en copromotor, dr. Frank Bennis. Toen ik na een korte fling met de corporate wereld weer terugkeerde naar de universiteit op zoek naar vrijheid en creativiteit hebben jullie er voor gezorgd dat die overstap zo succesvol is geweest als dat ik het heb ervaren.

Beste **Ron**, ik moet je heel erg bedanken voor al het vertrouwen en de vrijheid die je mij hebt gegeven tijdens mijn onderzoek. Bijna vijf jaar geleden ben ik bij je gekomen als een ex-bio medisch student en IT-consultant die nog nooit van farmacologie gehoord had en graag iets met 'machine learning' wilde gaan doen. Ik had een paar kleine projecten gedaan, maar er was weinig om van uit te gaan dat ik een beetje kon integreren in de wereld van de farmacometrie. Van tevoren had ik ook zeker niet verwacht hoe diep ik in de materie zou gaan en hoe goed het mij allemaal ging bevallen. Het spijt me dat ik je op dit traject allerlei enigszins onsaamhangende drafts van manuscripten met stoffige wiskundige vergelijkingen heb gestuurd. Ik hoop dat ik op dat vlak wat heb geleerd van jouw advies om een concreet en simpel verhaal te vertellen. Ik denk dan ook dat in dit proefschrift jouw sturing tussen de wiskunde duidelijk doorschemert.

Beste **Marjon**, ik ben trots dat ik je in mijn proefschrift en dankwoord kan adresseren als professor! Deze titel is zeer verdienstelijk voor al het werk dat jij hebt verricht voor kinderen met zeldzame bloedstollingsstoornissen en natuurlijk sikkcelziekte. Het was dan ook jouw enthousiasme op de eerste dag dat ik je ontmoette die mij meteen liet weten dat een promotie onderzoek als onderdeel van het SYMPHONY consortium een prettige ervaring ging worden. Jouw klinische en patiëntgerichte blik is in veel van de manuscripten in

dit proefschrift duidelijk terug te vinden, en je hebt me geleerd om de kliniek op de eerste plaats te zetten bij al deze projecten (okee soms kropen technische details ook naar boven). Ik weet zeker dat de machine learning projecten waar we mee begonnen zijn via jou hun weg naar de patiënt gaan vinden. Wij farmacometristen houden heel erg van onze modellen in onze achterkamertjes, en het vergt heel wat om dit allemaal in het daglicht te brengen. In dat opzicht ben jij zeker de juiste persoon op de juiste plek.

Beste **Frank**, toen ik op zoek was naar een copromotor om ondersteuning te bieden op het gebied van machine learning ben ik terugblikkend heel blij dat ik bij jou beland ben. Ondanks dat de farmacometrie ook nog een onbekend veld voor jou was (alles wat wij hier gedaan hebben is een niche in een niche), heb je direct mee kunnen denken bij mijn projecten. Het was jouw kunde om door alle obscure machine learning technieken waar ik mee kwam heen te kijken en een kritische blik te houden op de structuur van de experimenten. Je hield mij bescheiden door altijd de vraag te stellen of alle toeters en bellen wel nodig waren. Ik heb altijd heel erg genoten van onze discussies in de cafetaria van het AMC. Ik kwam vaak naar je toe met een project dat alle kanten op ging, maar liep gelukkig altijd weer weg met een concreet plan. Dit heeft uiteindelijk geleid tot een pittig dik boek (als ik het zelf mag zeggen), maar gelukkig heeft jouw interventie ervoor gezorgd dat er vele pagina's onzin aan bespaart zijn gebleven.

Dear **members of the Doctorate committee**, I thank you all for your time and effort spent in reading and critiquing this thesis. I hope you have found the contents newsworthy, which might have served as some relief for having to go through the many pages I have presented you. Mijn dank gaat ook uit naar **prof. Schut**, die als gast-opponent wil optreden om zo weer een commissie van zes te vormen tijdens de plechtigheid.

Mijn dank gaat ook uit naar alle co-auteurs en samenwerkingen die ik heb ondergaan tijdens dit onderzoek. Beste **prof. Hoogendoorn**, bedankt voor de hulp bij mijn eerste manuscript en de introductie tot Frank binnen jouw onderzoeksgroep. Dit project heeft veel tijd gekost, maar heeft me veel geleerd over hoe een goede voorbereiding tijd kan besparen. Beste **prof. Leebeek**, bedankt dat je naar een van mijn vroege manuscripten hebt willen kijken. Hoewel het wellicht lastig was om in dit stadium de klinische interpretatie te vinden, heb jij

me hier toch op punten kunnen helpen. Beste **Louk**, bedankt voor de fijne samenwerking en de behulpzame blik vanuit de hoek van causal inference. Mijn kennis van dat veld reikte op dat moment tot het lezen van *The Book of Why* van Judea Pearl, dus jouw inbreng kwam op het goede moment. Beste **Tine**, bedankt voor de fijne samenwerking bij het schrijven van onze scoping review over de klinische implementatie van hemofilie. Dit bleek ongelofelijk veel werk en ondanks dat ik je geduld soms testte hebben we een mooi stuk geleverd! Dear **Jessica & Michael**, I was glad to be able to contribute to your work on the RISE data, and I am sure it will lead to a great paper. Beste **Paul**, naast jouw verdiensten als onderdeel van de promotiecommissie heb ik al een flinke tijd terug een idee bij jou gepitcht voor de toepassing van machine learning op de ICU. Hoewel ik je nog steeds laat wachten op een mooie uitkomst, hoop ik de komende tijd grote stappen te zetten zodat we dit onderzoek door kunnen zetten. Ik bedank je ook voor de introductie bij prof. dr. Michaela van der Schaar en dr. Ari Ercole bij de Universiteit van Cambridge. Het bezoek aan Cambridge heb ik als heel inspirerend ervaren en resulteert hopelijk in verdere samenwerkingen. That brings me to **Ari and Pietro**, I want to thank you both for receiving me with open arms in Cambridge. Although our time together there has been relatively short, I consider my time in Cambridge as very inspiring and hope that we will be able to continue our collaboration. Ook dank aan de medewerkers van CSL Behring Nederland die ons toegang hebben verleend tot gegevens van patiënten die mee hebben gedaan aan klinische studies geïnitieerd door het bedrijf. Met deze gegevens hebben we twee mooie projecten weten te volbrengen. Ten slotte natuurlijk **Konrad**, bedankt voor de prettige samenwerking als onderdeel van de DosEmi studie. Ik ben natuurlijk maar betrokken geweest bij een klein deel van de studie, dus respecteer alle moeite die je in deze studie hebt gestoken. Ik weet zeker dat deze studie een mooi resultaat gaat brengen!

Daarnaast wil ik ook graag mijn collega's, de Ronderzoekers (misschien *ooit* Testosteron), **Laura, Amadou, Michael, Matteo, Medhat, Rafael, en Jelien** bedanken voor de fijne samenwerking, leuke discussies en gezellige koffierondjes. Hoewel een deel van ons in het begin thuis hebben gezeten en elkaar weinig écht zagen, hebben we het toch erg gezellig gehad als groep. Laura, Amadou, en Michael hebben onderhand al grote mensen banen en denken waarschijnlijk vol emotie en gevuld met nostalgie terug aan ons simpele bestaan naast de Albert Heijn. Steffie, Medhat, Matteo, Rafael en Jelien wens ik

veel succes (en sterke) bij de afronding van hun PhD. Ik wens Snoerd en Jelly natuurlijk ook veel voorspoed met het uitkopen van de Deli met alle opgespaarde koffie punten.

A special thank you also to my international colleagues and the new friends I have made over the years. The pharmacometrics research community, friends made during midnight snacks at the pizza vending machine during the PAGE meetings in Ljubljana, A Coruña, and Rome, and of course during WCOP in Cape Town. I had never before seen a professor order multiple rounds of tequila shots for everyone present at a conference social event. *Je suis Paolo*. I have always regarded my time in Cape Town and South Africa as truly special, especially after spending a year and a half indoors. Finally being released into the wild after COVID to meet such a friendly and open group to go on some adventures was amazing. Thanks **Paolo, Roeland, Marie, Eduardo, Victor, Allen**, and others. Wishing you the best. Part of becoming independent as an academic is unfortunately that going to PAGE means paying the big bucks for admission, so I hope I will still be able to meet my international friends at future events.

Ook veel van mijn dank gaat uit naar de klinische tak van het SYMPHONY consortium. **Tine, Iris, Wala, Caroline**, en eerdere collega's, bedankt voor alle moeite die jullie in ons gezamenlijk onderzoek hebben gestopt. Veel van jullie hebben talloze uren gespendeerd met het verzamelen van de patiënt gegevens waar onderzoekers zoals ik gretig gebruik van maken. **Lieke en Martijn**, ik heb veel lol met jullie gehad tijdens borrels. I wish **Bas, Minka, Caroline, Diaz, Snoerd, Lieke, Lorenzo, Huang, Jessica, Rianne, Martijn, and Wala** a lot of success with the finalization of their PhD. Ook alle andere leden van het SYMPHONY consortium en de OPTI-CLOT studie groep bedankt voor alle inzet in het succesvol maken van deze samenwerkingen! Ook al mijn andere collega's van de apotheek bedank ik voor de gezellige lunches in het hok ondanks de matriarchale sfeer die ik daar als man heb moeten ervaren (grapje).

Beste **Julian**, je bent een van mijn oudste vrienden en daar ben ik erg blij om. Bedankt voor de avonden waar ik de hele nacht je de oren van het lijf gepraat heb met mijn onderzoek, de technieken die ik interessant vond, en de vraagstukken over het leven. Ik hoop dat we lang goede vrienden blijven.

As an AI language model, I may not have the full context or details to perfectly address your question, but here is an attempt:

"Beste mannen en dame van *Zwaar weer*, **Bauk, Bram, Daan, Jan, Laurens, Marley, Nick, dr. Steef, Sven en Tankie**. Ik laat jullie natuurlijk niet ongenoemd. Ik had niet durven dromen dat ik nog zulke goede vrienden over zou houden aan een eerstejaarsploeg. We hebben een hechte band ontwikkeld tijdens het roeien die heel belangrijk voor mij geworden is. Dit alles culmineerde natuurlijk in het zilveren blik voor Jan in Parijs. Het was ongelofelijk om hier met z'n allen bij te zijn en met zweet onder de okseltjes langs de kant Jan voorbij te zien knallen. We wachten natuurlijk allemaal nog op de credits voor ons aandeel in dit succes, any time Jan. De doktoren gaan ons om de oren vliegen in de toekomst, dus speciaal dan ook nog erg veel succes gewenst bij de afronding van jullie onderzoek **Martijn, Jasmijn en Daan**, wordt zeker fantastisch! Ten slotte, Nick, ik hoop dat je me binnenkort kan uitleggen wat ik nou precies gedaan heb met mijn PhD nadat je de tijd hebt kunnen vinden om even snel de samenvatting te scannen."

Natuurlijk ook grote dank aan **Yoël**, toch wel een van mijn beste vrienden overgehouden aan Skøll. Ik vind het mooi wat voor goede match wij zijn (#emotioneel). Jij hebt altijd door gehad dat ik gek was om een PhD te gaan doen, positieve ervaringen zoals jij met onderzoek hebt gehad zijn immers schaars. Ik wens je dan toch ook nog zeker een eigen PhD ervaring toe, misschien iets van een multi-centre trial over het hulpbehoevende gedrag van expats in het buitenland.

Bedankt vrienden en familie voor alle etentjes, verjaardagen, borrels, en gezelligheid van de afgelopen jaren, ik hoop dat nu mijn PhD voorbij is ik weer wat vaker mijn kop laat zien!

Lieve **pap** en **mam**, jullie natuurlijk ook heel erg bedankt voor al jullie steun en liefde in mijn leven. Wie had gedacht dat ik na de basisschool zou studeren aan de universiteit, laat staan een poging zou doen om een doctoraat te behalen? Ik denk dat jullie invloed doorschemert in al het werk in dit proefschrift. Aan de ene kant heb ik een hele boel af geprogrammeerd, aan de andere kant heb ik de cover zelf ontworpen, en altijd geprobeerd om al mijn figuren visueel aantrekkelijk te maken tegenover alle stoffige zwart-wit figuren. Bedankt voor alles!

Jij ook bedankt broertje, **Jur**, hoewel het leek dat we in de middelbare school twee kanten op gingen, blijkt uit de laatste jaren toch dat we op elkaar lijken. Ik Bio-medische wetenschappen, jij Bio-farmaceutische wetenschappen. Ik toch richting machine learning, jij richting machine learning. Ik een PhD, jij een PhD. Ik een rijbewijs, jij een rijbewijs! Technisch gezien ben jij er alleen steeds iets eerder bij. Ik weet zeker dat jij ook een fantastisch proefschrift gaat afleveren!

Als laatste mag ik natuurlijk mijn lieve vriendin **Laura** bedanken. Je hebt veel met mij moeten doorstaan, vooral ook tijdens de jaren dat ik elk uur in de dag aan mijn promotie leek te wijden. Ik mag van geluk spreken dat ik zo'n goede band heb ontwikkeld met zo'n fantastisch iemand, en met zo'n leuke **schoonfamilie** (jullie natuurlijk ook bedankt voor alles in deze periode!). Bedankt voor al je geduld (en misschien ook soms de totale afwezigheid daarvan, denk aan dit proefschrift). Binnenkort doen we dit nog een keer voor jouw verdediging!

ACKNOWLEDGEMENTS

The SYMPHONY consortium [1] which aims to orchestrate personalised treatment in patients with bleeding disorders, is a unique collaboration between patients, health care professionals and translational & fundamental researchers specialised 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. In order to achieve this goal, workpackages (WP) have been organised according to three themes e.g. Diagnostics (WPs 3 & 4); Treatment (WPs 5-9) and Fundamental Research (WPs 10-12). This research received funding from the Netherlands Organisation 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 manager: Dr. S.H. Reitsma. More information: www.symphonyconsortium.nl.

Beneficiaries of the SYMPHONY consortium: Erasmus MC and Erasmus MC Sophia Children's Hospital, University Medical Center Rotterdam, 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 (The Netherlands) B.V..

REFERENCES

- [1] Cnossen, M.H., van Moert, I., Reitsma, S.H., de Maat, M.P., Schutgens, R.E., Urbanus, R.T., Lingsma, H.F., Mathot, R.A., Gouw, S.C., Meijer, K. and Bredenoord, A.L., 2022. SYMPHONY consortium: Orchestrating personalized treatment for patients with bleeding disorders. *Journal of Thrombosis and Haemostasis*, 20(9), pp.2001-2011.