Developing and Evaluating Pharmacokinetics-Driven Dynamic Personalized Medicine: A Framework and Case Study

A. Demetri Pananos, Daniel J. Lizotte

1 Introduction

Personalized medicine has been characterized by four goals: 1) to identify drugs for which between-subject variability in effectiveness or toxicity is a key issue for effective treatment, 2) to identify predictors which may explain this variability, 3) to decide on the right dose of the right drug by considering these factors, and 4) to prevent adverse reactions to drugs [1]. Progress in all four goals has accelerated within the last decade: For example, recent studies on DPYD genotype testing prior to starting fluoropyrimidine-based chemotherapy showed promise in preventing adverse events, making good arguments for integration of DPYD genotype testing into standard of care practices [2].

With regard to the third goal—personalized dosing—the intent of most efforts has been what we call *static* personalization. Such approaches inform dose at one point in time (usually induction) with the goal of eliminating the need for "trial-and-error" adjustments (titration) where the dose is adapted to the patient over time in response to its effects, both therapeutic and adverse [1]. Although significant progress has been made, for example in warfarin dosing [], the need for titration has been reduced but not eliminated. Thus, there is an opportunity to personalize not only the initial doses but also the titration process to achieve the best result—we call this *dynamic* personalization. This approach has been used in other contexts by applying techniques from disciplines such as control theory, operations research, machine learning, and biostatistics to define and apply models for optimal sequential decision-making for patient care **cite RL healthcare examples**.

Despite its potential to improve care, dynamic personalization imposes additional burden on the patient and provider, because it requires ongoing monitoring, for example by gathering lab results and returning for additional clinic visits. It is therefore natural to ask whether dynamic personalization is "worth it." Is the additional control over dose worth the additional burden? To help answer this question, we present a unified framework for the development and simulation-based evaluation of static and dynamic personalization based on PK modelling. The knowledge created by our framework can be integrated into a system-level decision-making framework like Know4Go, for example, which can be used to evaluate whether such a personalized medicine program should be implemented into a particular health care system [3]. Having established our

framework, we investigate the static and dynamic personalization of apixaban dosing as a case study.

The paper proceeds as follows. We begin in Section 2 with an overview of dynamic treatment regimes, which underpin our models for dynamic personalization, and we review Bayesian pharmacokinetic modelling, which allow us to predict drug concentrations and to generate simulated patient data. In Section 3, we present our framework, which describes how to estimate an optimal dynamic treatment regimes for personalization by combining Bayesian pharmacokinetic modelling with Q-learning, and describes a simulation-based approach for assessing the potential benefits of different levels of static and dynamic personalization. We then present our case study of personalized apixaban dosing in Section 4. Finally in Section 5 we discuss the results of the case study, and we identify broader issues relevant to the further development and implementation of PK-driven static and dynamic personalization.

2 Background

In the following two subsections, we present background material on dynamic treatment regimes, which are used to develop optimal decision-making models, and Bayesian pharmacokinetic models, which are used to capture relationships between patient characteristics, measurements, pharmacokinetics, and dose so that optimal dosing decisions can be derived using the dynamic treatment regime framework.

2.1 Dynamic Treatment Regimes

In this section, we review the theory of dynamic treatment regimes and how dynamic personalization using pharmacokinetic models can be formulated using a dynamic treatment regime.

2.1.1 Trajectories

Our work considers personalization of a dose or sequences of doses for a patient to keep their blood serum concentration of a drug within a desired range for as long as possible given the constraints: a) subject's blood serum concentrations cannot be measured very frequently, and b) we are limited to pre-dose clinical measurements to make our initial dosing decision. The theory of dynamic treatment regimes and statistical reinforcement learning offers a way of operationalizing dynamic personalization that can combine prior information with data to learn how best to select doses.

A dynamic treatment regime (DTR) is a sequence of decision rules for adapting a treatment plan to the time-varying state of an individual subject [4]. In DTRs, and their cousin topic in computer science *reinforcement learning*, an agent (often thought of as a robot in reinforcement learning, but within medicine sometimes thought of as a physician's computerized decision support system) interacts with a system for a number of stages. At each stage, the agent receives an *observation* of the system and then decides which *action* to take. This action will result in an observed *reward* which is followed by a new observation of the

system after it has been impacted by the action. This cycle of observation, action, reward then repeats, with the agent aiming to take actions which yield the largest total reward.

Key to our DTR is the concept of a *trajectory*. Define a stage to be a triple containing an observation, chosen action, and resulting reward. Let O_i denote an observation at the ith stage, A_i be the action at the i^{th} stage, and Y_i denote the reward at the i^{th} stage, denoted in capital letters when considering the observation, action, and reward as random variables. A trajectory is then the tuple $(O_1, A_1, O_2, A_2, \cdots, O_K, A_K, O_{K+1})$. Following notation by Chakraborty and Moodie [4], we will denote a system's history at stage j as $H_j = (O_1, A_1, O_2, A_2, \cdots, O_{j-1}, A_{j-1}, O_j)$. The reward at stage j is then a function of the system's history, the action taken, and the next observation $Y_j = Y_j(H_j, A_j, O_{j+1})$.

2.1.2 Policies, Value Functions, and Q-Learning

A deterministic policy $d = (d_1, \cdots, d_k)$ is a vector of decision rules each which take as input the system's history and output an action to take. Mathematically, $d_j : \mathcal{H}_j \to \mathcal{A}_j$ where $\mathcal{H}_j$ and $\mathcal{A}_j$ are the history and action spaces at stage j respectively. The stage j value function for a policy d is the expected reward the agent would receive starting from history h_j(here in lower case since it is an observed quantity) and then choose actions according to d for every action thereafter. The value function is written as

$$V_j^d(h_j) = E_d \left[\sum_{k=j}^{K} Y_k(H_k, A_k, O_{k+1}) \middle| H_j = h_j \right] . \tag{1}$$

Here, the expectation is over the distribution of trajectories. Importantly, the stage j value function can be decomposed into the expectation of reward at stage j plus the stage $j+1$ value function [4]

$$V_j^d(h_j) = E_d \left[Y_j(H_j, A_j, O_{j+1}) + V_{j+1}^d(H_{j+1}) | H_j = h_j \right] . \tag{2}$$

The optimal stage j value function is the value function under a policy which yields maximal value. Mathematically,

$$V_j^{opt}(h_j) = \max_{d \in \mathscr{D}} V_j^d(h_j) \tag{3}$$

A natural question is "what policy maximizes the value?". Estimating such a policy can be achieved by estimating the optimal Q function [4]. The optimal Q function at stage j is a function of the system's history h_j and a proposed action a_j,

$$Q_j^{opt} = E \left[Y_j(H_j, A_j, O_{j+1}) + V_{j+1}^{opt}(H_{j+1}) | H_j = h_j, A_j = a_j \right] \tag{4}$$

Note that the optimal Q function has similar form and interpretation to the optimal value function (namely, it is the expected reward starting at stage j but with the added condition that we take action a_j and then follow the optimal policy thereafter). Due to the decomposition of the reward function at stage

j, estimation of the optimal Q function can be performed by choosing the action which yields the largest reward at each stage assuming we act optimally in the future.

To use Q-learning for personalization in the context of optimal dosing and titration, we will define the actions to be possible doses or dose adjustments, and we define the reward to be a function of the resulting concentrations, for example a measurement of how well concentrations are kept in a specified range. The relationship between possible actions and rewards, which Q-learning can use to produce optimal policies, can be captured by Bayesian pharmacokinetic modelling, which we now review.

2.2 Bayesian Models of Pharmacokinetics

In order to estimate the optimal Q functions, we need to be able to predict how a patient's concentration is likely to evolve over time in response to a hypothetical dose (action.) Our approach is to build a Bayesian model of patient pharmacokinetics that can use baseline clinical information, as well as any available concentration measurements, to make tailored predictions of future concentrations that are as accurate as possible given the model structure and available data. The model is flexible in that it can condition on whatever information is available - for example, if previous dose and measurement information is not available for a specific patient, the model will rely on baseline information alone. If it is available, the model will use it to (hopefully) make improved predictions. This allows us to optimize both initial doses and later dose adjustments after additional information about concentration is acquired.

Bayesian models have another key property that we use in our framework. They are able to simulate the trajectories of patients drawn from a distribution that is similar to the distribution of the data that the models were trained on, but in the simulated data, *all* variables—including hidden PK parameters—are fully observed. This allows us to conduct a form of internal validation where we use the simulated patients to assess the relative benefits of different modes of static and dynamic personalization, because we can know for each simulated patient exactly what the effect of any dose would be. This process is described in detail in the next section, where we present our framework.

3 A Framework for Assessing Static and Dynamic Personalization

In this section, we present the components of our framework for assessing static and dynamic personalization, including details for fitting a heirarchical Bayesian pharmacokinetic model to concentration data from a cohort of patients, assessing the behaviour of markov chains via diagnostics, and using the Bayesian model to generate simulated data for evaluation. We then outline several modes of static and dynamic personalization ranging from no personalization (every patient gets the same dose) to a complex dynamic mode of personalization (estimation of the optimal policy for dosing from a dynamic treatment regime). Finally, we outline steps for assessing the benefits of each mode of personalization.

3.1 Bayesian Modelling

The first step in our framework is to fit a Bayesian model that relates patient covariates and dose to drug concentration as a function of time. For example, previous work [5] describes a hierarchical Bayesian model of apixaban pharmacokinetics, in which the clearance rate (L/hour), time to max concentration (hours), absorption time delay (hours), and ratio between the elimination and absorption rate constants (called alpha, a unitless parameter) are hierarchically modelled. In our case study, we extend that model by regressing the latent pharmacokinetic parameters on baseline clinical variables (age, sex, weight, and creatinine) to permit personalization. The model could equally well be extended with pharmacokinetic or biomarker information if the relevant theory and data were available for a particular use case.

Once the form of the model is specified, creating simulated patients or estimating the PK parameters of a real patient requires computation of or sampling from the posterior distribution of the relevant variables given the relevant data. However, exact computation of the posterior distribution is intractable for all but very simple models, so Markov chain Monte Carlo (MCMC) techniques are often used to approximate the expectations with respect to the posterior distribution. Presently, the gold standard for generating samples from the posterior is Hamiltonian Monte Carlo (HMC), which works by generating a sequence of samples that "explores" the space of possible sampled values in a way that reflects the posterior distribution. Many implementations of HMC come with diagnostics which monitor the behaviour of the Markov chains that are used to generate samples and help to ensure that they are representative of the posterior distribution. That these Markov chains behave well is crucial, as any inferences about or from the model are obtained from samples generated by the chains. To assess the quality of the markov chains, several diagnostics are commonly used including: number of divergences, the Gelman-Rubin convergence diagnostic, and effective sample size [6].

In practice, several Markov chains are used simultaneously to generate samples from the posterior. The chains are assessed with within-chain and between-chain diagnostics. First, individual chains may sometimes *diverge*. A divergence in a Markov chain indicates that the HMC Markov chain has encountered a region of high curvature in the posterior distribution which cannot be adequately explored. Consequently, Monte Carlo estimators of any expectations can be biased due to incomplete exploration of the posterior distribution. It is important that none of the Markov chains generated by HMC display a divergence, and that many chains (typically 4 or more) are initialized and are allowed to explore the posterior distribution.

Having ensured that no chains are diverging, a group-level diagnostic is used to assess whether all chains have converged to the same limiting distribution. The *Gelman-Rubin (sometimes called $\hat{R}$) convergence diagnostic* is designed to detect if the Markov chains have converged to the same distribution by measuring the within-chain variance to the between chain-variance. In practice, $1.05 < \hat{R}$ indicates that there is poor mixing of the Markov chains and inference from the samples should not be performed lest the Monte Carlo estimators are biased by this poor mixing.

Even if the chains do not exhibit divergences and arrive at the same limiting distribution, the Markov

chains could still exhibit high within-chain correlation, thereby increasing the uncertainty of estimation of key posterior quantities such as (means, variances, or quantiles) [7]. The *effective sample size* is a measure of how much the within chain autocorrelation increases uncertainty estimates. In some software packages, the effective sample size is reported as a fraction of the total number of samples drawn from the Markov chains. Presently, the guidance is that the effective sample size ratio should be no smaller than 1%.

Once the model is fit, important diagnostics indicate no pathelogical behaviour, and the model is deemed to fit the data sufficiently well, the model can then used to generate synthetic pharmacokinetic data for use in experiments to compare different forms of personalization. Each generated data point may be thought of as one synthetic patient, with observed covariates and observed pharmacokinetic parameters. These parameters, which are never observed in real data, allow us to compute the effects of any dosing decisions (which are made *without* direct knowledge of the parameters), and thus allow us to evaluate the performance of different modes of personalized dosing on the sampled population.

3.2 Modes of Personalization

The second step in our framework is to identify modes of personalization that we wish to evaluate. We classify these modes of personalization into two types: static and dynamic personalization.

Static modes of personalization seek to inform the dose at one point in time (usually induction) with the goal of eliminating the need for "trial-and-error" adjustments. We consider two modes of static personalization:

1. One size fits all. This mode of personalization is not very personal at all. All patients receive the same dose size at the onset of treatment.

2. Doses size selected on clinical variables. In this mode of personalization, the patient's covariates, for example age, sex, weight, creatinine measurements, and possibly genetic or biomarker information, are provided to the pharmacokinetic model. A dose size is then selected using the model to maximize the reward function conditional on these measurements.

Dynamic modes of personalization seek to personalize the initial doses but also the titration process. We consider four modes of dynamic personalization:

1. One size fits all dose *and* one blood sample. This mode of personalization provides patients the same dose to start, but requires a concentration measurement to be made sometime in the future. In our simulations, subjects take their initial dose once every 12 hours with perfect adherence. Our pharmacokinetic model conditions on this measurement, and the dose is adjusted in order to maximize the reward in the latter 5 days.

2. Dose size selected on clinical variables *and* one blood sample. Here, the initial dose provided to the patient is determined by the patient's clinical measurements. In the second half of the fifth day, a

concentration measurement is made at a random time. The model is conditioned on this concentration and the dose is adjusted to optimize the reward.

3. Optimal sampling. Similar to the previous mode of personalization, but the time at which the measurement is made is under our control and tuned to maximize reward.

4. Q-learning. Need a bit more here - maybe explain in terms of 3 previous ones? We can discuss.Estimation of the optimal policy to maximize the reward function via a dynamic treatment regime and Q-learning.

Here, we stress that these are just examples of some modes of personalization, and that we do not mean that these modes should always be candidate modes for personalization. These modes may not be appropriate for all drugs across all clinics, and were selected in order to illustrate natural extensions and combinations of static personalization with additional information collection. A strength of our approach is that many possible modes of personalization may be considered depending on what is appropriate for the use-case at hand.

3.3 Evaluation

To evaluate the potential for personalization, we compare all of the modes identified in step 2 based on their achieved reward as well as the difference between the achieved reward and theoretically largest reward the reward we would achieve if we knew each patient's pharmacokinetic parameters exactly. Because we know the true latent pharmacokinetic parameters of the simulated subjects, we can optimize the reward with the known pharmacokinetics of the subject, thereby yielding the largest reward possible.

In reality, at the time of dose adjustment, the actual reward is not known. However, because a Bayesian model is a generative model, samples generated from that model can be used to compute the expected reward. The steps for computing the expected reward for a given dose are as follows:

1. Condition the model on available information if the mode of personalization permits so. For some modes of personalization, this might be clinical measurements, blood concentration measurements, both or neither.

2. Draw several sets of of pharmacokinetic parameters from the posterior distribution. Each of this represents a possible underlying truth for the patient under consideration. Choosing any one of these sets of pharmacokinetic parameters as "the truth", in conjunction with the dose schedule considered, completely determines the concentration profile and allows us to make a prediction of the patient's concentration into the future. Each set of parameters thus corresponds to a concentration, resulting in a distribution over possible future trajectories of concentrations.

3. Compute the concentration profile on a grid of times. The grid must be sufficiently fine to capture changes in the concentration over time. We make a prediction at 15 minute intervals.

4. Compute the reward for the proposed dose based on the resulting concentration trajectory.

5. Update the proposed dose and recompute the reward.

The concentration profile scales linearly with dose size; double the dose, double the concentration at a given time. This relationship allows for off-the-shelf optimizers to be used in the selection of dose size in this processes.

4 Case Study

In this section, we present the results of applying our framework to static and dynamic personalization of apixaban dosing. **Insert apixaban data info/background.**

4.1 Bayesian Modelling

We extend a previously proposed one-compartment Bayesian pharmacokinetic model [5] to include fixed effects of covariates on pharmacokinetic parameters in order to incorporate baseline clinical information (age, sex, weight, and creatinine.) Full details of the model structure are provided in Appendix A. We fit the model to previously-collected data on apixaban concentration [8] and then use the fitted model to simulate patients with known "ground truth" pharmacokinetic parameters. We will then use this population of simulated patients in our experiments to explore different modes of dose personalization and their relative benefits.

f We fit M1 to real pharmacokinetic data using the Stan software in R[]. Stan monitors several markov chain diagnostics, none of which detected problematic markov chain behavior, which indicates that Stan's sampling algorithm was able to converge to the target distribution (0 divergences, all all Gelman-Rubin diagnostics$<$1.01, all effective sample size ratios $>$ 22%).

The inclusion of covariates in the model results in a better fit than excluding them. Shown in figure 1 are the estimated random effects for the clearance pharmacokinetic parameter of each subject as a function of weight. Subject sex is indicated by color, the overall trend is shown in the black dashed line. Failing to include subject sex and weight results in males having on average a larger random effect than females of the same weight, and heavier subjects having a larger random effect than lighter subjects. When covariates are added into the model, the variation in the random effects attenuates, resulting in closer alignment to model assumptions. A better fit to the data means data generate from the model may be closer aligned with the true data generating process.

Examining the posterior distributions of the regression coefficients provides further insights into the relationships between covariates and pharmacokinetics. Subject weight increases the expected value of alpha (which is used to compute the elimination and absorption rates in the first order one compartment PK model. The parameter α is the ratio of how fast the drug exists the central compartment to how fast the

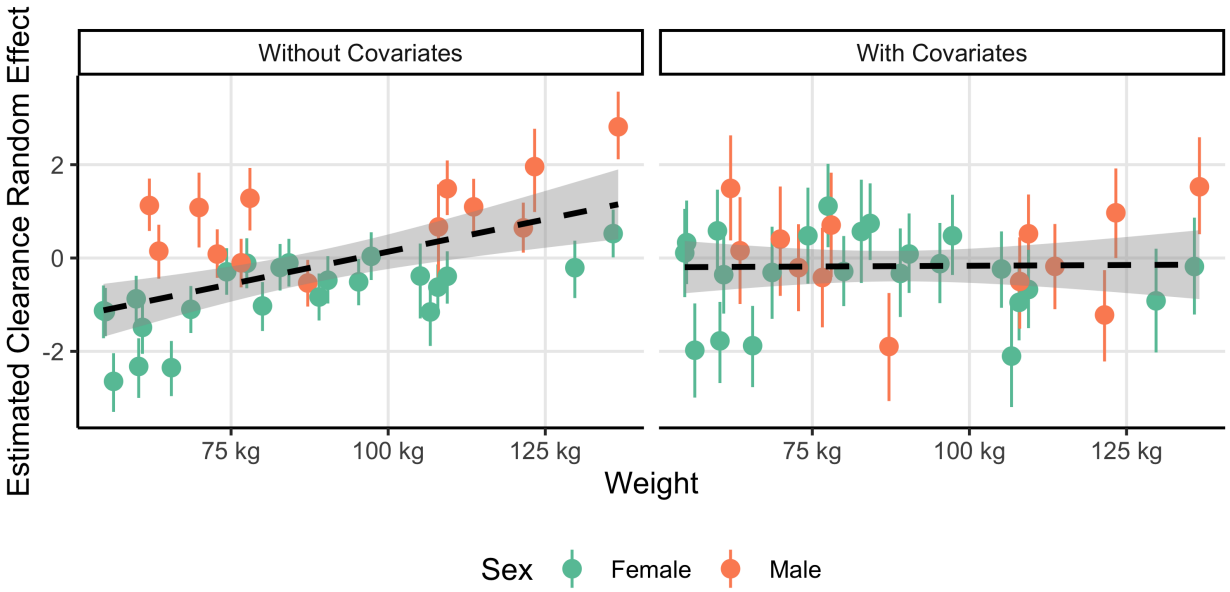

Figure 1: Random effects estimates for clearence rate Cl_i and 95% credible intervals (left). Random effects estimates are colored by patient sex. Prior to adjusting for covariates, a general trend in weight can be seen in the random effects. Subjects who are heavier tend to have larger random effect, and males tend to have larger random effects than females of the same weight. Patterns such as these indicate that weight and sex can be used to explain variation in the random effects. After adjusting for sex and weight (right), the random effects have no discernable pattern.

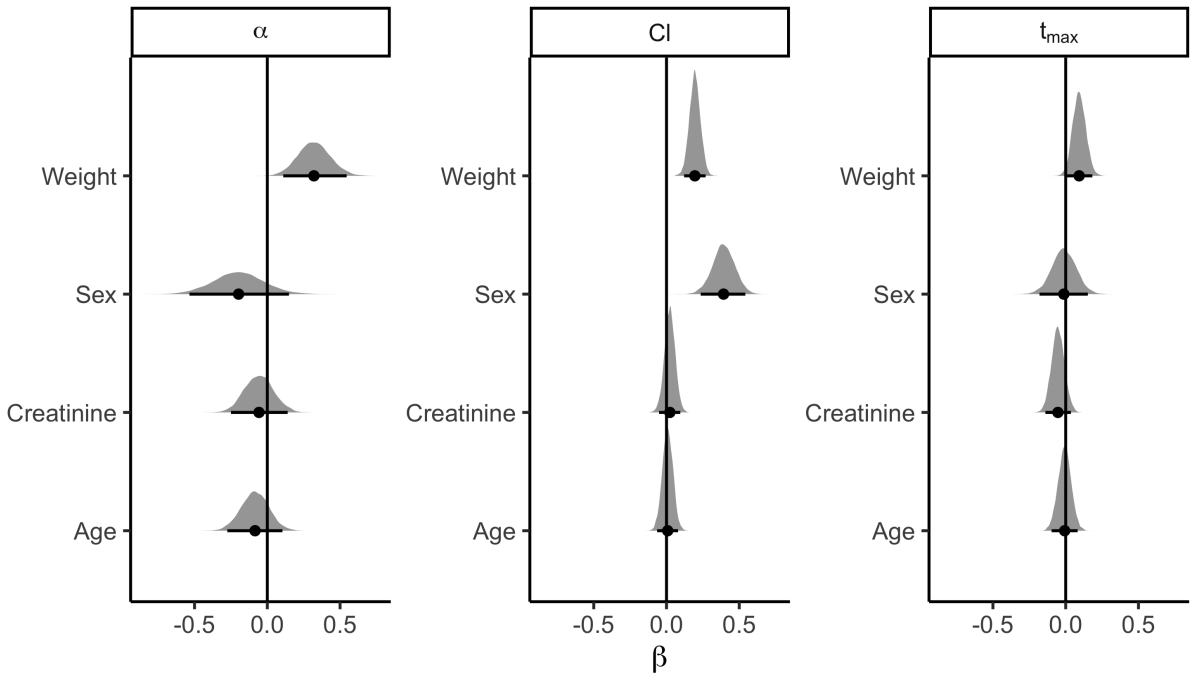

Figure 2: Posterior distributions of regression coefficients. Expectations are shown as black dots, 95% credible intervals are shown as horizontal black lines. Solid black vertical line is $\beta = 0$ for reference. Note, regression coefficients for Cl and t_{max} act multiplicatively (a one unit increase in weight leads to a change in Cl of $\exp(\beta)$), while regression coefficients for α are interpreted on the log odds scale.

drug enters the central compartment) as well as the time to max concentration. There is an estimated effect of sex on α (males have smaller alpha than females, meaning the drug leaves their central compartment slower or enters the central compartment quicker), however the uncertainty is large (estimated effect -0.2 on the logit scale, 95% credible interval -0.54 to 0.15). See Table X in the Appendix for a full summary of the regression coefficients.

Model training error sees a very small improvement. Including subject covariates decreases model training error from 6.89 ng/ml to 6.84 ng/ml. Estimates of concentration uncertainty remain similar between the two models as well. We conclude the inclusion of covariates in the model improves model inferences but does not improve the fit of the model to the data in any substantial way. Either model would require additional validation prior to using in a predictive capacity.

4.2 Modes of Personalization

Implementing static and dynamic personalization as a dynamic treatment regime [4], we propose six policies, each increasing in complexity and clinic/patient burden, for personalizing doses of apixaban with the goal of keeping blood serum concentrations within a desired range for as long as possible.

Thus far, we have motivated personalization of dose sizes through Q learning. Q learning is the most complex solution that could be implemented at this time, and its implementation in practice would be burdensome due to this complexity. This makes implementation of Q learning a tall order, especially considering alternative DTRs exist which are not as costly to implement. The cost of implementing Q learning may be worth paying if the benefit of Q learning over these other DTRs is substantial, but we need a framework in which to estimate the size of this benefit.

Our study considers the following modes of personalization. Each one has different requirements in terms of computation, data needs, clinical overhead, and patient burden. Our study aims to understand, in this particular setting (apixaban with potential PK monitoring) what the relative benefits of these different modes might be in practice. It also presents a general framework for evaluating these different modes of personalization in other settings. The six modes of personalization we consider are:

4.2.1 Experimental Design In Terms of Stages of a DTR

In our experiments, we develop a DTR for selecting the best dose for keeping a patient's blood plasma concentration within a desired range. Here, we present details of the experimental design in the DTR framework, leaving simulation details (including how the data were simulated) for our methods section.

Our experiment consists of 1000 simulated subjects taking a dose of apixaban once every 12 hours with perfect adherence for a total of 10 days. Sometime in the second 12 hour period on the fourth day (between 108 and 120 hours after the initial dose), we have the opportunity to measure the simulated subject's blood concentration, should our policy allow for it. At the start of the fifth day, the dose is adjusted based on all the pre-dose clinical measurements plus the observed concentration. The dose will be adjusted so as to attempt to maximize the time spent between 0.1 mg/L and 0.3 mg/L. Thus, our DTR consists of two stages (the first five days, and the latter five days), however the size of the range may be adapted for different scenarios. We choose this range as it is not so narrow that even optimal doses perform poorly, but not so wide that any dose can achieve high reward.

In terms of the DTR, the system is the patient for whom a dose is selected, the actions correspond to selection of dose sizes, and the reward is the proportion of time spent within the desired concentration range. The trajectories we will use to estimate the optimal Q functions are of the form

$$O_1, A_1, Y_1, O_2, A_2, Y_2, O_3 \tag{5}$$

The interpretation of a given trajectory is:

- O_1 is any pre-dose clinical measurements of the subject. In our experiments, we consider age in years, renal function (as measured by serum creatinine in mMol/L), weight in kilograms, and dichotmous biological sex (dummy coded so that male=1 and female=0). We choose these variables as they are known to affect the pharmacokinetics of apixaban [9].

- A_1 is dual action of initial dose to provide the subject plus a time in the future at which to measure the subject's blood serum concentration.
- Y_1 is the proportion of time spent within the concentration range in the first five days.
- O_2 is the pre clinical measurements of the subject plus the observed concentration made on the fourth day.
- A_2 is the dose adjustment
- Y_2 is the proportion of time spent within the concentration range in the last five days after the dose adjustment.
- O_3 would be pre-dose clinical measurements, the observed concentration made on the fourth day, and the next concentration measurement, were it to be made. As we examine just the two actions A_1 and A_2, we do not make use of O_3 but include it here to adhere with our definition of trajectories above.

The reward function we use depends on the subject's true latent concentration. Let $c_j \quad j = 1...K$ be the j^{th} latent concentration value at time t_j. The reward function is

$$Y(c_1, c_2, \cdots, c_k) = \frac{1}{k} \sum_{j=1}^{K} \mathbb{I}(0.1 < c_j < 0.3) \tag{6}$$

Here, $\mathbb{I}$ is an indicator function returning 1 if the argument is true, and 0 else. We suppress the dependence on the history in the definition of the reward as the reliance on the history is implicit. The reward depends on the latent concentrations which depend on previous doses (actions) and potentially on the previous dose measurements (observations of the system). We approximate this reward function with a continuously differentialble function to facilitate optimization. See the appendix for details.

Our stage 2 optimal Q function is then

$$Q_2^{opt}(H_2, A_2) = E\left[Y(c_{j+1}, c_{j+2}, \cdots, c_{j+n}) \middle| H_2, A_2 \right], \tag{7}$$

and our stage 1 optimal Q function is

$$Q_1^{opt}(H_1, A_1) = E\left[Y(c_1, c_2, \cdots, c_j) + \max_{a_2 \in \mathscr{A}} Q_2^{opt}(H_2, a_2) \middle| H_1, A_1 \right] \tag{8}$$

We seek to maximize the stage 1 optimal Q function to learn the optimal policy for dosing subjects under the constraint we can measure them at most once and are limited to the aforementioned pre-dose clinical variables. The interpretation of stage 1 optimal Q function is as follows: *Given the pre-dose clinical variables of the subject and a proposed initial dose and measurement time, the stage 1 optimal Q function gives the proportion of time the subject's blood serum concentration is between 0.1mg/L and 0.3mg/L assuming that we provide the subject with the best dose possible at the start of the* 5^{th} *day.* The actions A_1 and A_2 which maximize these functions constitute the optimal policy.

The concentration values c_j in the optimal Q functions are latent, meaning we have no direct access to them in practice. Furthermore, obtaining measurements with high enough frequency so that the reward is faithfully estimated would be too burdensome on the patient.

We consider 6 modes of personalization which range in the amount of information used in the decision process as well as burden placed on the patient and clinic, and burden of implementation. We present the results of our simulation in figure X below in terms of difference between theoretically largest reward and reward achieved by the mode of personalization. The results are ordered from least amount of information and burden (top) to most amount of information and burden (bottom).

1) Dose selection using a hierarchical Bayesian model which does not incorporate subject covariates. This model was presented in Pananos & Lizotte [5]. We refer to this mode as the "No Covariate Model".

2) 1) and conditioning the model on a single sample from the subject taken sometime in the final 12 hours before the half way point. At the start of the fifth day, a new dose is selected and used for the remaining time. We refer to this mode as "No Covariate + 1 Sample".

3) Dose selection from M2. A single dose is selected at the start of the regiment and is used throughout the 10 simulated days. We refer to this mode as "Covariate Model".

4) 3) and conditioning the model on a single sample from the subject taken sometime in the final 12 hours before the half way point. At the start of the fifth day, a new dose is selected and used for the remaining time. We refer to this mode as "Covariate model + 1 Sample".

5) A two stage DTR, however the initial dose is the result of the procedure in 3). The best time to sample the patient is then determined via Q learning. We refer to this mode as "Optimal Sampling Time".

6) The two stage DTR we describe in the previous sections, estimated via Q learning. We refer to this mode as "Q Learning".

4.3 Evaluation of Benefits

Modes of personalization which use less information have a larger difference (i.e. yield smaller reward on average than what is theoretically possible). The no covariate model (which uses no information about the subject) performs worst with a median difference of 0.19. The distribution of differences for this mode is right skewed with some differences exceeding 0.95, meaning the subject could have been in range for nearly the entire time but the mode selected a dose which failed to put the subject in range.

The use of covariates in the model nearly cuts the difference in half, achieving an average difference of 0.1 with smaller right skew. There is a diminishing in the difference in rewards as additional burden is undertaken. Modes which use observed concentration information (Covariate Model + 1 Sample, Optimal Sampling Time, and Q Learning) lead to marginally more reward on average.

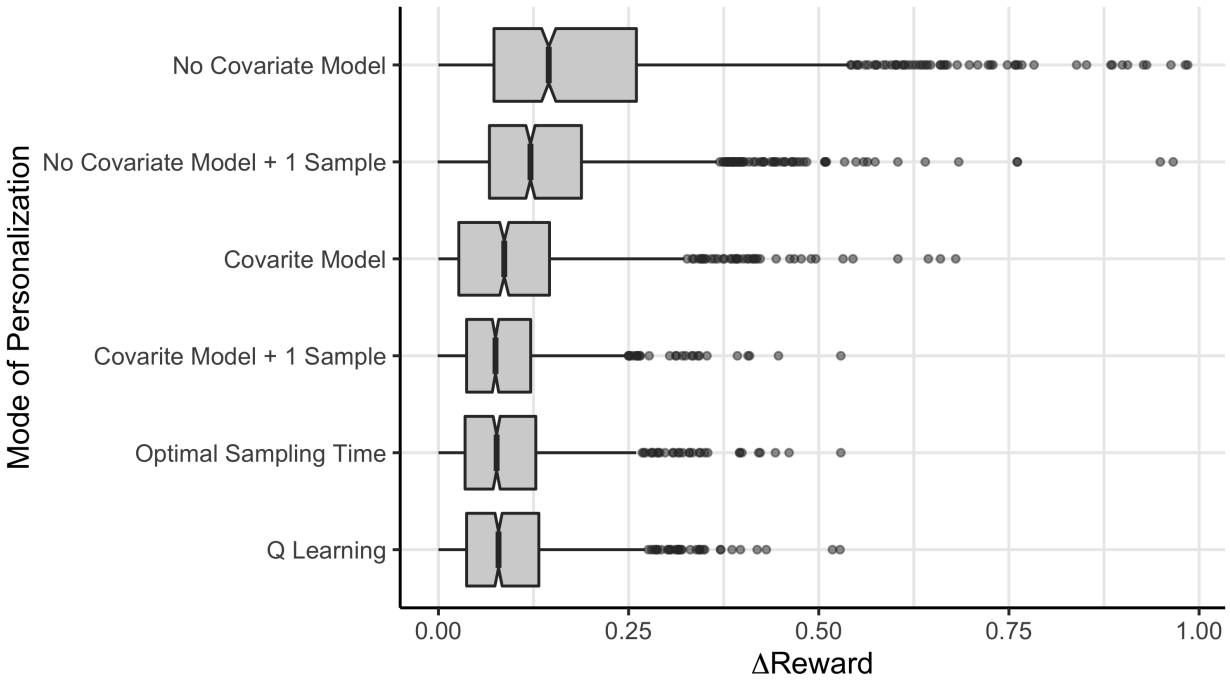

Figure 3: Boxplots of the difference between theoretically largest reward and achieved reward for each of the 1000 simulated subjects. Subjects who achieve a reward close to their maximium reward have a difference on 0, subjects who achieve a reward less than their maximum have larger differences, with the largest difference being 1.

5 Discussion

Including age, sex, weight, and creatinine in our Bayesian model improved the inferences from that model. Though the predicted concentrations and estimates of uncertainty changed negligibly, the covariates explained residual confounding in the random effects. This results in a model which better explains the observed variation in the data and hence should generate more plausible data for simulation.

That modes of personalization which use more information result in larger reward is ultimately unsurprising. From our perspective, the more pertinent result is the diminishing return on investment observed when using additional information (and consequently, taking on additional implementation burden to effectively use that information). Were doses to be personalized for the simulated population, we would recommend the covariate model be used, as it is easier to implement, puts smaller burden on the clinic, and results in approximately mean/median rewards as compared to other methods. Taking an additional sample doesn't seem to improve the expected mean/median reward appreciably to be worth the burden of having the patient take time to come in, drawing their blood, measuring their concentration in the lab, and reporting results to the decision maker, in which a decision to not adjust the dose might be made anyway. Similar arguments can be made for Q learning, which has even higher implementation burden.

But mean/median reward does not tell the whole story. As noted, the distribution of differences in reward is right skewed. Some subjects have a very large difference, and the possibility of these differences might not be acceptable in different contexts with different drugs. There is a tradeoff between less extreme differences and taking on additional clinic and implementation burden, and that tradeoff should be examined on a case by case basis.

Context is crucial, and how we adapt to that context is perhaps a question in need of closer examination. Traditional methods of personalization include conditioning only on a subject's covariates (not unlike the Covariate model we present here). But of course patients are not their age, sex, weight, and creatinine. Additionally, safety information and best available practices might change in the future as more research on drugs is performed. Were new safety information to be published, one might imagine the reward function might be affected, which may result in a new mode of personalization being more/less preferable or more/less feasible. Any number of factors in flux can change the context in which personalization occurs, and that change in context may prompt for a re-evaluation in how personalization is done.

Thus, our results are not about apixaban per se. We don't offer recommendations on how personalization for apixaban should be done because we can't anticipate the context. What we offer is a framework for developing strategies of personalization and evaluating their performance against their implementation and clinic burden. Context can be changed where needed, either through the reward function, or by adjusting when the clinic is able to take measurements, or by including additional information such as genotype in the Bayesian model. Using this framework, clinics have flexibility to personalize the personalization.

Moved down from introduction.

Personalized medicine still faces several barriers to widespread adoption, including economic burden,

patient burden, and expertise burden required for new methods of personalization. Personalized medicine can increase safety and reduce costs to the healthcare system by identifying patients who are at greater risk for adverse events or dose adjustments. For example, if personalization enables a patient to avoid an adverse event, then this avoids associated costs to the healthcare system, example from a hospital stay [10]. More ambitiously, personalized medicine has the potential to save the healthcare system costs by more effectively using resources [11].

The cost of instruments, technicians, and leadership required to operate a personalized medicine clinic are high burden, and it is not yet clear if personalized medicine is sufficiently cost effective to offset operating costs in all circumstances [12]. In their 2019 scoping review of personalized medicine cost effectiveness, Kasztura et. al [12] found that willingness-to-pay thresholds vary wildy from country to country (citing that cost per quality adjusted life year for some modes of personalized medicine range from \$20, 000 USD per quality adjusted life year in for studies in Europe and the United Kingdom to \$200,000 USD per quality adjusted life year for studies in the United States). This high variability in cost effectiveness means the burden required for start up may result in a positive return on investment in some areas but not others. This variability should prompt would be adopters to more closely examine if taking on the initial burden is worth the result.

The dominant perspective on personalized medicine focuses on the use of clinical and physiological information (including biomarkers, genotyping, and diagnostic tests) as a means of optimizing treatments, but largely ignore needs, constraints, and utilities of the patient [13, 14]. Patients can be burdened by frequent followup for clinical measurement (as in the case with Warfarin), be burdened by costly expenses related to obtaining care, or may be more risk adverse/tolerant than the "typical" patient. As an example, transportation has been found to be a large financial burden for patients recieving cancer treatment [15], and continues to burden patients, with a 2020 study finding that the cost of parking alone can climb as high as \$1600 over the course of treatment in the United States [16]. Additional visits to a clinic have the potential to further burden patients by requiring them to miss a day of work, and find means of childcare during their absence (if neccesary). Incorporating patient preferences and reducing the burden of personalization on the patient can result in sustained adherence [17], thereby increasing effectiveness and further preventing adverse events.

An additional expertise burden is added as machine learning (used interchangeably with the term "artificial intelligence") is adopted into personalized medicine initiatives. Cutting edge machine learning models for prediction or decision making can be prohibitively burdensome to implement effectively. Failure to carefully implement a prediction model may result in pernicious bias inadvertently affecting subpopulations, as was found to be the case in algorithms for credit scoring [18], crime prediction [19], and hiring [20]. A 2019 study found an instance of this bias in a widely used risk scoring algorithm in healthcare [21], demonstrating that despite the best intentions of those involved, the use of a model can lead to worse rather than better care if investigators are not careful in considering what sorts of bias may be present in the data used to train these models. Implementation of new approaches and methods requires the close collaboration of experts

in data science with physicians, domain experts, and other stakeholders. Close collaboration should allow for domain experts to identify what kinds of biases the data might have, and for data science experts to implement methods to help ammeliorate that bias (or to admit the data are not fit for purpose). The result of iterating on this collaborative process (wherein domain experts help inform the appraoches methodologists take, and the methodologists provide model checks which help domain experts decide if decisions from the model are reasonable or suspicious) is a model which more closely aligns with domain expertise, a model which is sufficiently flexible to capture the true data generating mechanism, an effective use of data, a more transparent modelling process, and calibrated expectations surrounding algorithms and their abilities [22]. Presently, this form of collaboration between methodologists and domain experts is not the norm, with developement of machine learning solutions in healthcare bdeing developed in silos [23].

These burdens may be surmountable for some, but the question then turns to if the result is worth the expense. Answering that question is difficult without an idea how the additional burden of collecting data, or implementing new algorithms, will benefit the clinic or the patient subject to inherent constraints.

5.1 Limitations

We've examined six modes for making decisions. The next mode improves on a deficiency of the previous mode in a natural manner, and so our experiment constitutes a kind of ablation study. We believe the decision making aspect of our study extracts information in a responsible way and uses the best decision making methodology available. That being said, the experiment is not without limitations.

The bayesian model of the pharmacokinetics is integral to the methodology we present. Any shortcomings in the model affect the quality of the decision and decision process. Bayesian models are not as ubiquitous as other models in pharmacology, and so particular expertise is required for model development and evaluation. That expertise increases the implementation burden of any decision process involving Bayesian models. However, we demonstrate how one such model can be constructed in a past stdy [CITE] and include open sourced code and data for practitioners to replicate our model fitting.

Additionally, the data required to construct a high quality Bayesian model of pharmacokinetics require multiple observations of a single patient over an extended time, preferably over multiple well timed doses with near perfect adherence. Obtaining such data requires well organized efforts and is high burden for both investigators and participating subjects. This makes acquiring a robust Bayesian model for use in dose personalization difficult.

5.2 Future Work

Because the data required to build reliable Bayesian pharmacokinetic models are difficult to collect in practice, research into developing these models from observational data may prove fruitful in extending this work. If clinics record data on measured blood concentrations, they may have dozens or hundreds of subjects with only one or two measurements per subject. Moreover, the subjects in question may be on multiple drugs

or have comorbidities which may affect the pharmacokinetics of the drug under study. Additional research into constructing Bayesian models which can adjust for polypharmacy and comorbidities while learning an individual's pharmacokinetics from a large but sparse sample would drive this work towards being easier to implement in practice.

References

[1] Bridget L Morse and Richard B Kim. Is personalized medicine a dream or a reality? *Critical reviews in clinical laboratory sciences*, 52(1):1–11, 2015.

[2] Theodore John Wigle, Brandi Povitz, Wendy Teft, Robin Legan, John Gordon Lenehan, Markus Gulilat, Stephanie Nevison, Justin Kritzinger, Veera Punaganty, Denise Keller, et al. Prospective cohort study of the impact of hospital-wide dihydropyrimidine dehydrogenase (dpyd) genotype testing for fluoropyrimidine-based chemotherapy on adverse events and hospital costs. *American Society of Clinical Oncology*, 2019.

[3] Janet Martin, Avtar Lal, Jessica Moodie, Fang Zhu, and Davy Cheng. *Hospital-Based HTA and Know4Go at MEDICI in London, Ontario, Canada*, pages 127–152. Springer International Publishing, Cham, 2016. ISBN 978-3-319-39205-9.

[4] Bibhas Chakraborty. *Statistical methods for dynamic treatment regimes*. Springer, 2013.

[5] A Demetri Pananos and Daniel J Lizotte. Comparisons between hamiltonian monte carlo and maximum a posteriori for a bayesian model for apixaban induction dose & dose personalization. In *Machine Learning for Healthcare Conference*, pages 397–417. PMLR, 2020.

[6] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo, 2018.

[7] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

[8] Rommel G Tirona, Zahra Kassam, Ruth Strapp, Mala Ramu, Catherine Zhu, Melissa Liu, Ute I Schwarz, Richard B Kim, Bandar Al-Judaibi, and Melanie D Beaton. Apixaban and rosuvastatin pharmacokinetics in nonalcoholic fatty liver disease. *Drug Metabolism and Disposition*, 46(5):485–492, 2018.

[9] Wonkyung Byon, Samira Garonzik, Rebecca A Boyd, and Charles E Frost. Apixaban: a clinical pharmacokinetic and pharmacodynamic review. *Clinical pharmacokinetics*, 58(10):1265–1279, 2019.

[10] Margot de Looff, Bob Wilffert, Cornelis Boersma, Lieven Annemans, Stefan Vegter, Job FM van Boven, and Maarten J Postma. Economic evaluations of pharmacogenetic and pharmacogenomic screening tests: a systematic review. second update of the literature. *PloS one*, 11(1):e0146262, 2016.

[11] Fatiha H Shabaruddin, Nigel D Fleeman, and Katherine Payne. Economic evaluations of personalized medicine: existing challenges and current developments. *Pharmacogenomics and personalized medicine*, 8:115, 2015.

[12] Miriam Kasztura, Aude Richard, Nefti-Eboni Bempong, Dejan Loncar, and Antoine Flahault. Cost-effectiveness of precision medicine: a scoping review. *International journal of public health*, 64(9): 1261–1271, 2019.

[13] Wolf Rogowski, Katherine Payne, Petra Schnell-Inderst, Andrea Manca, Ursula Rochau, Beate Jahn, Oguzhan Alagoz, Reiner Leidl, and Uwe Siebert. Concepts of 'personalization'in personalized medicine: implications for economic evaluation. *Pharmacoeconomics*, 33(1):49–59, 2015.

[14] Antonello Di Paolo, François Sarkozy, Bettina Ryll, and Uwe Siebert. Personalized medicine in europe: not yet personal enough? *BMC health services research*, 17(1):1–9, 2017.

[15] Peter S Houts, Allan Lipton, Harold A Harvey, Barbara Martin, Mary A Simmonds, Richard H Dixon, Santo Longo, Thomas Andrews, Robert A Gordon, John Meloy, et al. Nonmedical costs to patients and their families associated with outpatient chemotherapy. *Cancer*, 53(11):2388–2392, 1984.

[16] Anna Lee, Kanan Shah, and Fumiko Chino. Assessment of parking fees at national cancer institute–designated cancer treatment centers. *JAMA oncology*, 6(8):1295–1297, 2020.

[17] Rachel A Elliott, Judith A Shinogle, Pamela Peele, Monali Bhosle, and Dyfrig A Hughes. Understanding medication compliance and persistence from an economics perspective. *Value in health*, 11(4):600–610, 2008.

[18] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[19] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

[20] Ifeoma Ajunwa. The paradox of automation as anti-bias intervention, 41 cardozo, l, 2020.

[21] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[22] Holger Fröhlich, Rudi Balling, Niko Beerenwinkel, Oliver Kohlbacher, Santosh Kumar, Thomas Lengauer, Marloes H Maathuis, Yves Moreau, Susan A Murphy, Teresa M Przytycka, et al. From hype to reality: data science enabling personalized medicine. *BMC medicine*, 16(1):1–15, 2018.

[23] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

A Appendix

The Bayesian model used to predict personalized concentration in response to dose, which we refer to as $\mathcal{M}_1$, is

$$y_{i,j} \sim \text{Lognormal}\left(C_i(t_j), \sigma_y^2\right) \tag{9}$$

$$\sigma^2 \sim \text{Lognormal}\left(0.1, 0.2\right) \tag{10}$$

$$C_i(t_j) = \begin{cases} \dfrac{D_i \cdot F}{Cl_i} \cdot \dfrac{k_{e,i} \cdot k_{a,i}}{k_{e,i} - k_{a,i}} \left(e^{-k_{a,i}(t_j - \delta_i)} - e^{-k_{e,i}(t_j - \delta_i)}\right) & t_j > \delta_i \\ 0 & \text{else} \end{cases} \tag{11}$$

$$k_{e,i} = \alpha_i \cdot k_{a,i} \tag{12}$$

$$k_{a,i} = \frac{\log(\alpha_i)}{t_{max,i} \cdot (\alpha_i - 1)} \tag{13}$$

$$\delta_i \sim \text{Beta}(\phi, \kappa) \tag{14}$$

$$\text{logit}(\alpha_i) | \beta_\alpha, \sigma_\alpha^2 \sim \text{Normal}(\mu_\alpha + \mathbf{x}_i^T \beta_\alpha, \sigma_\alpha^2) \tag{15}$$

$$\log(t_{max,i}) | \beta_{t_{max}}, \sigma_{t_{max}} \sim \text{Normal}(\mu_{t_{max}} + \mathbf{x}_i^T \beta_{t_{max}}, \sigma_{t_{max}}^2) \tag{16}$$

$$\log(Cl_i) | \beta_{Cl}, \sigma_{Cl} \sim \text{Normal}(\mu_{Cl} + \mathbf{x}_i^T \beta_{Cl}, \sigma_{Cl}^2) \tag{17}$$

$$p(\phi) \sim \text{Beta}(20, 20) \tag{18}$$

$$p(\kappa) \sim \text{Beta}(20, 20) \tag{19}$$

$$p(\mu_{Cl}) \sim \text{Normal}(\log(3.3), 0.15^2) \tag{20}$$

$$p(\mu_{t_{max}}) \sim \text{Normal}(\log(3.3), 0.1^2) \tag{21}$$

$$p(\mu_\alpha) \sim \text{Normal}(-0.25, 0.5^2) \tag{22}$$

$$p(\sigma_y) \sim \text{Lognormal}(\log(0.1), 0.2^2) \tag{23}$$

$$p(\sigma_{CL}) \sim \text{Gamma}(15, 100) \tag{24}$$

$$p(\sigma_{t_{max}}) \sim \text{Gamma}(5, 100) \tag{25}$$

$$p(\sigma_\alpha) \sim \text{Gamma}(10, 100) \tag{26}$$

$$p(\beta_{Cl,k}) \sim \text{Normal}(0, 0.25^2) \quad k = 1...4 \tag{27}$$

$$p(\beta_{t_{max},k}) \sim \text{Normal}(0, 0.25^2) \quad k = 1...4 \tag{28}$$

$$p(\beta_{\alpha,k}) \sim \text{Normal}(0, 0.25^2) \quad k = 1...4 \tag{29}$$

Here, normal distributions are parameterized by their mean and variance, lognormal distributions are parameterized by the mean and variance of the random variable on the log scale, and gamma distributions are parameterized by their shape and rate. The μ in the model above represent population means on either

the log or logit scale, the β are regression coefficients for the indicated pharmacokinetic parameter, the sigmas are the population level standard deviations on the log or logit scale, δ is aparameter which relaxes the assumption that the dose is absorbed into the blood immeditately upon ingestion, F is the bioavailability of apixiban (which we fix to 0.5 [9]) and D si the size of the dose in milligrams. All continuous variables were standardized using the sample mean and standard deviation prior to being passed to the model.

Once fit, $\mathcal{M}_1$ can be used to predict the pharmacokinetics of new patients, using the patient's covariates as predictors. To do so, the marginal posterior distributions for μ_{Cl}, $\mu_{t_{max}}$, μ_α, β_{Cl}, $\beta_{t_{max}}$, β_α, σ_{Cl}, $\sigma_{t_{max}}$, σ_α, and σ_y must be summarized. We use maximum likelihood on the posterior samples to summarize the marginal posterior distributions. We model the population means and regression coefficients as normal, and the standard deviations as gamma. The maximum likelihood estimates are used to construct priors for a new model, which we call $\mathcal{M}_2$. We construct $\mathcal{M}_2$ so as to be able to predict plasma concentration after multiple doses (of potentially different sizes) administered over time, and remove the time delay (δ) to simplify our simulations. Model priors for $\mathcal{M}_2$ are then

$$p(\mu_{Cl}) \sim \text{Normal}(0.5, 0.04) \tag{30}$$

$$p(\mu_{t_{max}}) \sim \text{Normal}(0.93, 0.05) \tag{31}$$

$$p(\mu_{\alpha}) \sim \text{Normal}(-1.35, 0.13) \tag{32}$$

$$p(\sigma_{Cl}) \sim \text{Gamma}(69.15, 338.31) \tag{33}$$

$$p(\sigma_{t_{max}}) \sim \text{Gamma}(74.96, 349.56) \tag{34}$$

$$p(\sigma_{\alpha}) \sim \text{Gamma}(10.1, 102.07) \tag{35}$$

$$p(\beta_{Cl,1}) \sim \text{Normal}(0.39, 0.08^2) \tag{36}$$

$$p(\beta_{Cl,2}) \sim \text{Normal}(0.19, 0.04^2) \tag{37}$$

$$p(\beta_{Cl,3}) \sim \text{Normal}(0.02, 0.04^2) \tag{38}$$

$$p(\beta_{Cl,4}) \sim \text{Normal}(0.01, 0.04^2) \tag{39}$$

$$p(\beta_{t_{max},1}) \sim \text{Normal}(-0.01, 0.08^2) \tag{40}$$

$$p(\beta_{t_{max},2}) \sim \text{Normal}(0.09, 0.05^2) \tag{41}$$

$$p(\beta_{t_{max},3}) \sim \text{Normal}(-0.05, 0.04^2) \tag{42}$$

$$p(\beta_{t_{max},4}) \sim \text{Normal}(-0.01, 0.04^2) \tag{43}$$

$$p(\beta_{\alpha,1}) \sim \text{Normal}(-0.19, 0.17^2) \tag{44}$$

$$p(\beta_{\alpha,2}) \sim \text{Normal}(0.33, 0.11^2) \tag{45}$$

$$p(\beta_{\alpha,3}) \sim \text{Normal}(-0.06, 0.1^2) \tag{46}$$

$$p(\beta_{\alpha,4}) \sim \text{Normal}(-0.09, 0.1^2) \tag{47}$$

$$\tag{48}$$

For our experiments, we generate the pharmacokinetic parameters of 1000 simulated patients from the prior predictive model of $\mathcal{M}_2$. Bayesian models are generative models, meaning they can generate pseudo-data by drawing random variables according to the model specification going from top (model priors) to bottom (model likelihood). To do so, we begin by resampling 1000 tuples of age, sex, weight, and creatinine from the dataset used to fit $\mathcal{M}_{\infty}$. We sample one draw of r μ_{Cl}, $\mu_{t_{max}}$, μ_{α}, β_{Cl}, $\beta_{t_{max}}$, and β_{α} from their respective prior distributions in $\mathcal{M}_2$. The values of these parameters remained fixed for all 1000 patients. Conditioned on the values of these mus and betas, we compute the expectation of the population distribution for each pharmacokinetic parameter by computing $\mu_{Cl} + \mathbf{x}^T\beta_{Cl}$, $\mu_{t_{\max}} + \mathbf{x}^T\beta_{t_{max}}$, $\mu_{\alpha} + \mathbf{x}^T\beta_{\alpha}$, where $\mathbf{x}^T$ is

the resampled tuple. From the prior distribution of M2, we sample one draw ofσ_{Cl}, $\sigma_{t_{max}}$, σ_α, and σ_y. These remained fixed for all 1000 patients. Using the previously computed expectations and σ, we sample 1000 tuples of pharmacokinetic parameters, one for each of the simulated patients. The clearance rate and time to max concentration were sampled assuming a lognormal distribution. Alpha was sampled using a logitnormal distribution. The pharmacokinetics can then be determined conditional on the pharmacokinetic parameters. Each of simulated patients' pharmacokinetic parameters remained fixed through the experiments. We simulate the latent concentration using $C(t)$ as written in $\mathcal{M}_2$, and can simulate observed concentrations by drawing a sample from a lognormal distribution with mean $\ln(C(t))$ and standard deviation σ_y

We use Stan, an open source probabilistic programming language, for fitting our Bayesian models via Hamiltonian Monte Carlo (a Markov Chain Monte Carlo technique) and computing markov chain diagnostics. Twelve chains are initialized and run for 2000 iterations each (1000 for warmup allowing the Markov chain the opportunity to find the correct target distribution and 1000 to use as samples from the posterior).