A Framework For Considering Costs and Benefits of Additional Complexities in Personalized Medicine

Working Title Need a New One

Demetri Pananos, Dan Lizotte

1 Introduction

Personalized medicine has 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 aforementioned factors, 4) and to aid in the prevention of adverse reactions of said drugs [1]. Progress in all four goals has accelerated within the last decade. As an 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]. Despite this progress, personalized medicine still faces barriers to widespread adoption.

The cost of instruments, technicians, and leadership required to operate a personalized medicine clinic are non-negligible, and it is not yet clear if personalized medicine is sufficiently cost effective to offset operating costs [3]. Additionally, some perspectives on personalized medicine focus on the use of demographic and clinical/biological information (including biomarkers, genotyping, and diagnostic tests) as a means of optimizing treatments, but largely ignore needs, constraints, and utilities of the patient [4] (for example their availability or willingness to be subject to multiple blood draws to measure blood serum concentrations). Ignoring these constraints may result in a method of personalization which, although effective, may not be realistic to implement at scale because the cost to the clinic, or the burden to the patient, is just too large.

An additional expertise cost 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 complex to implement correctly and at scale, requiring the partnership of experts in data science, computer science, statistics, engineering, etc. Collaboration between experts in these disciplines and physicians (among other stakeholders) is crucial to make effective use of data, rigorously internally validate models, and temper expectations which may be skewed from hype surrounding these algorithms [5]. Thus, new approaches to personalization may be out of reach without financial means to hire experts should they not be available for collaborative work (vis à vis large research grants, etc).

These costs may be payable 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 cost of collecting data, or implementing new algorithms, will benefit the clinic or the patient subject to inherent constraints. In this study, we present a new framework for helping practitioners interested in implementing personalized medicine to answer these questions while considering practical limitations, and we present a case study using the framework applied to apixaban dosing.

For our case study, we fit a Bayesian model to existing data on the pharmacokinetics of apixaban. The resulting Bayesian model is used to generate synthetic pharmacokinetic data for use in experiments to compare different forms of personalization. Treating personalization as a dynamic treatment regime, 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. Under the assumption that the fitted Bayesian model can produce similar data to what might be observed in the future from new patients, we can make inferences as to how different policies for personalizing doses may improve upon one another, and compare if the additional burden of implementing a more complex or costly form of personalization can generate a more desirable outcome for the patient or healthcare provider.

We begin with an overview of dynamic treatment regimes and how personalizing doses can be framed as such. Afterwards, we describe how estimating an optimal policy for a dynamic treatment regime can be done using Q learning (the most complex method we entertain here). We discuss the details of the Bayesian model we use to fit the real pharmacokinetic data, and present model fit diagnostics to argue that our model is satisfactory for generating synthetic data for use in our simulations. We then present and discuss the results of our simulation in light of the costs presented to a clinic to implement personalized medicine, and how this framework can be integrated to make decisions about what form of personalization to implement.

2 Dynamic Treatment Regimes

In this section, we discuss the theory of dynamic treatment regimes and how personalization can be thought of as a dynamic treatment regime. We describe our experiments in the context of dynamic treatment regimes and introduce our reward function.

2.1 Trajectories

Our goal is to find the dose or sequences of doses for a subject to keep their blood serum concentration 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 framework through which to understand our problem and construct one possible solution.

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 [6]. 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 [6], 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 $Y_j = Y_j(H_j, A_j, O_{j+1})$.

2.2 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. 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 computed 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 [6]

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

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 [6]. 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. Below, we describe how we choose optimal actions using a posterior distribution of a subject's pharmacokinetics.

2.3 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 [7].
- 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 $c_j j = 1 ... K$ 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. To leverage off-the-shelf optimization tools, we approximate this reward function with a continuously differentiable function, namely

$$Y(c_1, c_2, \cdots, c_k) = \frac{1}{k} \sum_{j=1}^{K} \exp\left(- \left[\frac{c_j - 0.15}{0.05} \right]^{2\beta} \right) \tag{7}$$

Here, $\beta$is a positive integer. For sufficiently large beta, our approximation becomes arbitrarily close to our intended reward function. In practice we set beta=5 to balance between good approximation of our intended reward and vanishing gradients impeding our optimization. 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).

Our stage 2 optimal Q function is then

$$Q_2^{opt}\left(H_2, A_2\right) = E\left[Y\left(c_{j+1}, c_{j+2}, \cdots, c_{j+n}\right) \middle| H_2, A_2\right] \ , \tag{8}$$

and our stage 1 optimal Q function is

$$Q_1^{opt}\left(H_1, A_1\right) = E\left[Y\left(c_1, c_2, \cdots, c_j\right) + \max_{a_2} Q_2^{opt}\left(H_2, a_2\right) \middle| H_1, A_1\right] \tag{9}$$

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. What is left to explain is how these concentrations are computed. In the next section, we describe how we use a Bayesian model to obtain latent concentration predictions and compute the required expectations.

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] 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.

[4] 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.

[5] 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.

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

[7] 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.