Submitted to the Annals of Applied Statistics

A BAYESIAN DECISION FRAMEWORK FOR OPTIMIZING SEQUENTIAL COMBINATION ANTIRETROVIRAL THERAPY IN PEOPLE WITH HIV

BY WEI JIN[1,a], YANG NI[2,c], JANE O'HALLORAN[3,d], AMANDA B. SPENCE[4,e], LEAH H. RUBIN[5,f], AND YANXUN XU[1,b]

[1]*Department of Applied Mathematics and Statistics, Johns Hopkins University,* [a]*wjin@jhu.edu;* [b]*yanxun.xu@jhu.edu*

[2]*Department of Statistics, Texas A&M University,* [c]*yni@stat.tamu.edu*

[3]*Department of Internal Medicine, Washington University in St. Louis,* [d]*janeaohalloran@wustl.edu*

[4]*Department of Medicine, Georgetown University,* [e]*abs132@georgetown.edu*

[5]*Departments of Neurology and Psychiatry, Johns Hopkins University School of Medicine,* [f]*lrubin@jhu.edu*

Numerous adverse effects (e.g., depression) have been reported for combination antiretroviral therapy (cART) despite its remarkable success in viral suppression in people with HIV (PWH). To improve long-term health outcomes for PWH, there is an urgent need to design personalized optimal cART with the lowest risk of comorbidity in the emerging field of precision medicine for HIV. Large-scale HIV studies offer researchers unprecedented opportunities to optimize personalized cART in a data-driven manner. However, the large number of possible drug combinations for cART makes the estimation of cART effects a high-dimensional combinatorial problem, imposing challenges in both statistical inference and decision-making. We develop a two-step Bayesian decision framework for optimizing sequential cART assignments. In the first step, we propose a dynamic model for individuals' longitudinal observations using a multivariate Gaussian process. In the second step, we build a probabilistic generative model for cART assignments and design an uncertainty-penalized policy optimization using the uncertainty quantification from the first step. Applying the proposed method to a dataset from the Women's Interagency HIV Study, we demonstrate its clinical utility in assisting physicians to make effective treatment decisions, serving the purpose of both viral suppression and comorbidity risk reduction.

1. Introduction. The emergence of antiretroviral therapy (ART) has transformed HIV infection from a fatal to chronic disease by effectively reducing the viral load and decreasing HIV-related morbidity and mortality. Common ART drugs fall into six drug classes with different mechanisms, including nucleotide reverse transcriptase inhibitor (NRTI), non-nucleotide reverse transcriptase inhibitor (NNRTI), protease inhibitor (PI), integrase inhibitor (INSTI), entry inhibitor (EI), and pharmacokinetic enhancer (Booster). Despite the effectiveness of *combination* ART (cART) consisting of three or more ART drugs from different drug classes in viral suppression, numerous ART-related adverse effects have been reported, including mental health disorders, chronic kidney failure, and cardiovascular diseases (Checa et al., 2020; Dietrich et al., 2021). The U.S. Department of Health and Human Services provides a general guideline on initiating cART for treatment-naive people with HIV (PWH); however, the guideline mainly focuses on viral suppression but does not account for the reported adverse effects. Furthermore, ART-related adverse effects can vary greatly from person to person due to various individualized risk factors such as sociodemographic, clinical, and behavioral characteristics. Therefore, personalizing cART that not only suppresses viral

Keywords and phrases: Antiretroviral therapy, Multivariate Gaussian process, Offline reinforcement learning, Precision medicine, Uncertainty-penalized policy optimization.

load but also manages ART-related adverse effects is now one of the most pressing challenges in the field of HIV, especially considering that cART is recommended for PWH indefinitely. In this paper, we aim to develop a decision-making framework for optimizing personalized cART in PWH.

In many medical applications with chronic conditions (e.g., diabetes, HIV infections, and chronic kidney diseases), it is often important for treatments to be adaptive to individuals' disease progression and treatment responses over time. Such scenarios can be formalized as a dynamic treatment regime (DTR, Robins 1986; Murphy 2003), which is a sequence of treatment decision rules at multiple stages, each of which maps an individual's up-to-date information to a recommended treatment. For example, PWH are recommended to follow up with their physicians semiannually by the current HIV treatment guidelines. At each visit, their sociodemographics, medication use, and laboratory test results are collected. Then physicians prescribe their cARTs based on clinical observations such as staying on the previous cART or switching to a new cART until the next visit.

Many statistical methods have been developed to estimate the optimal sequential treatment assignments from observational data such as marginal structural model (Wang et al., 2012), G-computation formula (Robins, 2004), stochastic tree search (Sun and Wang, 2021), and likelihood-based approaches (Xu et al., 2016; Hua et al., 2021). Most of these methods only consider a small number of possible actions at each decision stage. However, the number of possible drug combinations in cART assignments can be enormous since there are over 30 U.S. Food and Drug Administration (FDA)-approved ART drugs. Another related research area is offline reinforcement learning (RL) (Lange, Gabel and Riedmiller, 2012), in which a *policy* (a sequence of actions) model is reinforced, by the feedback from the offline (previously collected) data including individuals' longitudinal observations (also called the *state*) and treatments (also called the *action*), to optimize sequential decisions that maximize a *reward*. While they have been proven useful in applications such as robotics (Yu et al., 2020a), traditional *model-free* offline RL methods can result in unstable policy learning when out-of-distribution actions are evaluated (Fujimoto, Meger and Precup, 2019). In contrast, *model-based* offline RL methods (Yu et al., 2020b) learn a probabilistic dynamic model from the observational data to evaluate out-of-distribution actions and are more sample efficient. Most of the offline RL methods depend on the Markov decision process assumption, however, the dynamics in our setting is not Markov since how clinical measurements evolve over time may depend on the full history of these measurements and cARTs.

Large-scale HIV studies, such as the Women's Interagency HIV Study (WIHS), provide us unprecedented opportunities to learn personalized optimal sequential cART assignments. The WIHS is a large prospective, observational, multicenter study designed to investigate the impact of HIV infection on multimorbidity in women with HIV or at risk for HIV in the United States (Adimora et al., 2018). The complexity of cART assignments, longitudinal observations, individual heterogeneity, and long-term sequential decisions present three major analytical and modeling challenges, which we explain in detail below.

- *Learn how individuals' longitudinal states (i.e., health outcomes) evolve over time conditional on their preceding states and cART histories.* This requires us to estimate the longitudinal cART effects from a high-dimensional and unbalanced space. With more than 30 FDA-approved ART drugs, there are a large number of possible drug combinations, making the estimation a high-dimensional problem. In addition, some cARTs are frequently used whereas others are rarely used. For example, 3TC+D4T+NFV (two NRTIs + one PI) was recorded 993 times in the WIHS, while a similar cART 3TC+D4T+ATV (two NRTIs + one PI) was only recorded 12 times. Most prior studies only used simplistic cART representations that do not account for drug-drug interactions, such as using a binary variable to indicate whether an individual is on cART (Lundgren et al., 2002; Bogojeska et al., 2010).

One recent work (Jin et al., 2022) proposed a Bayesian approach to estimate the *cross-sectional* cART effects incorporating drug-drug iterations. However, it cannot be used to learn the *longitudinal* cART effects.

- *Generate a realistic cART from a large discrete space.* Assume there are a total number N of ART drugs. A straightforward way of representing a cART is an N-dimensional binary vector with each element indicating whether the cART contains that corresponding drug. Then we can generate possible cARTs using a multivariate logistic model. However, this leads to 2^N possible drug combinations, in which most of them are unrealistic and would never be prescribed in clinical practice. To the best of our knowledge, there are no existing methods that build a generative model for cART assignments to effectively explore the high-dimensional cART space.
- *Mitigate the distribution shift issue.* The fundamental challenge of optimizing sequential treatments from observational data is *distributional shift*: the offline training data may be collected under different policies from the one we try to evaluate. In other words, the distribution of states visited by the learned policy inevitably deviates from the distribution of offline data. Therefore, without accounting for the distribution shift in policy optimization, the learned policy may not achieve the expected optimality.

To address the aforementioned challenges, we propose a two-step Bayesian decision framework for optimizing sequential cART assignments, which is illustrated in Figure 1. In the first step, we develop a probabilistic dynamic model for individuals' irregular longitudinal observations using a multivariate Gaussian process (MGP, Alvarez et al. 2012), where the irregularity is caused by missing values in measurements. The MGP learns the transition dynamics that describes how individuals' states evolve over time conditional on their historical states and treatment histories, which not only mitigates the Markov assumption required in traditional RL methods but also captures the longitudinal drug combination effects of cART on individuals' states. We use the subset-tree (ST) kernel (Jin et al., 2022) that converts the drug combination into a conceptually simple but mathematically powerful representation. The ST kernel induces an appropriate similarity measure among different cARTs by explicitly accounting for known clinical knowledge on ART drugs. This formulation enables us to efficiently borrow information across cARTs and thus reduces the dimension of the drug combination space to a manageable size. We fit the Bayesian dynamic model to the observed data and obtain the posterior estimates of the dynamics with uncertainty quantification.

In the second step, we build a probabilistic generative model for the cART assignment by representing the selection of a cART via a tree structure with three levels. The mathematically analytic formulation of the three-level decision process allows for direct policy optimization by applying the stochastic gradient descent (SGD, Robbins and Monro 1951) algorithm. Moreover, the uncertainty quantification of posterior inference from the first step allows us to penalize the reward of each state-action pair by its uncertainty when optimizing the sequential cART assignments. Such a procedure can help mitigate the distribution shift issue via a trade-off between the reward gain and risk for exploring new policies and new state-action pairs. Through both simulations and the WIHS application, we demonstrate the capability of the proposed method in making effective personalized treatment decisions to optimize individual-level health outcomes. As expected, individual-level improvement accumulates to population-level health betterment.

The rest of paper is organized as follows. In Section 2, we outline the proposed two-step Bayesian decision framework for optimizing personalized sequential cART assignments. In Section 3, we elaborate on the first step of developing an MGP model for individuals' longitudinal states. In Section 4, we elaborate on the second step of an uncertainty-penalized policy optimization procedure. We evaluate the performance of the proposed approach through simulation studies and compare it to alternative methods in Section 5, and apply it to the WIHS dataset in Section 6. Lastly, we conclude with a discussion in Section 7.

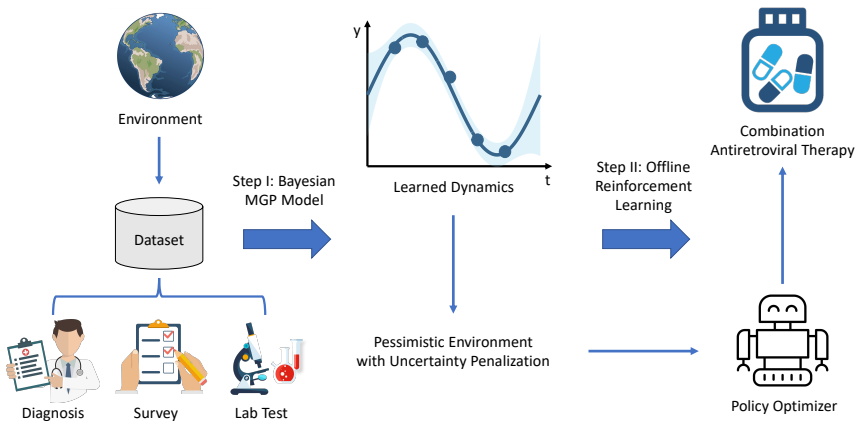


FIG 1. *Illustration of the proposed two-step Bayesian decision framework for optimizing sequential cART assignments with proper uncertainty propagation.*

2. Two-Step Bayesian Decision Framework Formulation. For each individual $i = 1, 2, \ldots, I$, assume that we have an S-dimensional vector of baseline covariates denoted by $\boldsymbol{X}_{i0}$. At times $\boldsymbol{t}_i = (t_{i1}, \ldots, t_{i,J_i})$, we have M time-varying variables that characterize the individual's health state such as depression score, denoted by $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{i,J_i})$ with $\boldsymbol{Y}_{ij} \in \mathbb{R}^M$ for each visit $j = 1, 2, \ldots, J_i$. Let $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{i,J_i})$ with Z_{ij} denoting the cART used by individual i during the time period $(t_{i,j-1}, t_{ij}]$, where $t_{i0} = 0$. Thus our data can be summarized as $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^I = \{\boldsymbol{X}_{i0}, \boldsymbol{t}_i, \boldsymbol{Y}_i, \boldsymbol{Z}_i\}_{i=1}^I$. Assume that the physician assigns a cART $Z_{i,j+1}$ at time t_{ij} for the individual i to take during the time period $(t_{ij}, t_{i,j+1}]$ based on her baseline covariates $\boldsymbol{X}_{i0}$, longitudinal state history $\overline{\boldsymbol{Y}_{ij}} = \{\boldsymbol{Y}_{ij'} : j' \leq j\}$, and treatment history $\overline{Z_{ij}} = \{Z_{ij'} : j' \leq j\}$. Then the individual takes the prescribed cART until the next visit at time $t_{i,j+1}$, and her state is updated to $\boldsymbol{Y}_{i,j+1}$ following a probabilistic dynamic model parameterized by $\phi : \boldsymbol{Y}_{i,j+1} = f(\overline{\boldsymbol{Y}_{ij}}, \overline{Z_{i,j+1}}; \phi)$. The objective is to optimize personalized sequential cART assignments to maximize the individual's long-term health outcomes, e.g., lowest cumulative depression scores in the next two years. We first define our problem in an optimization framework.

For any individual i with baseline covariates $\boldsymbol{X}_{i0}$, suppose that she already has J_i visits with recorded states history $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}, \ldots, \boldsymbol{Y}_{i,J_i})$ and treatment history $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{i,J_i})$. Let $\boldsymbol{Y}_i^{\text{new}} = \{\boldsymbol{Y}_{ij} : j > J_i\}$ and $\boldsymbol{Z}_i^{\text{new}} = \{Z_{ij} : j > J_i\}$ denote her future longitudinal states and cART assignments, respectively. Assume for any future visit $j > J_i$, the cART is assigned through a policy function parameterized by $\boldsymbol{\theta} : \pi(Z_{i,j+1} \mid \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta})$. We assign to each individual some stochastic reward function of future longitudinal states: $r_i(\boldsymbol{Y}_i^{\text{new}})$. For example, if our goal is to select sequential cARTs that result in the lowest cumulative depression scores (higher is worse) in the next two years (i.e., the next 4 visits if 2 visits per year), and let Y_{ij1} denote the predicted depression score at the future visit j, we will define $r_i(\boldsymbol{Y}_i^{\text{new}}) = -\sum_{j=J_i+1}^{J_i+4} Y_{ij1}$. Denote the expected reward for any individual i to be:

$$R_i(\boldsymbol{\theta}) = \int E_{(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}}) \sim p(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}} \mid \mathcal{D}, \boldsymbol{\phi}, \boldsymbol{\theta})}[r_i(\boldsymbol{Y}_i^{\text{new}})] p(\boldsymbol{\phi} \mid \mathcal{D}) d\boldsymbol{\phi}. \tag{2.1}$$

Note that even though the reward function $r_i(\boldsymbol{Y}_i^{\text{new}})$ only depends on $\boldsymbol{Y}_i^{\text{new}}$, the expectation in (2.1) is taken over all stochastic realizations of both $\boldsymbol{Y}_i^{\text{new}}$ and $\boldsymbol{Z}_i^{\text{new}}$ to highlight their coupled relationship, which is equivalent to taking the expectation over $\boldsymbol{Y}_i^{\text{new}}$ only with respect to its marginal distribution $p(\boldsymbol{Y}_i^{\text{new}} \mid \mathcal{D}, \boldsymbol{\phi}, \boldsymbol{\theta})$ with $\boldsymbol{Z}_i^{\text{new}}$ integrating out. We aim to find

the optimal personalized cART assignment policy $\pi(\cdot,\cdot;\boldsymbol{\theta}_i^\star)$ that maximizes the expected reward $R_i(\boldsymbol{\theta})$, $\boldsymbol{\theta}_i^\star = \arg\max_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta})$, while accounting for the uncertainty in the longitudinal dynamic model by integrating out its parameter $\boldsymbol{\phi}$ with respect to the posterior $p(\boldsymbol{\phi} \mid \mathcal{D})$.

To find $\boldsymbol{\theta}_i^\star$ and the optimal sequential cART assignments from (2.1), we will use stochastic gradient descent (SGD, Robbins and Monro 1951), i.e., $\boldsymbol{\theta}_{i,q+1} = \boldsymbol{\theta}_{i,q} + s_{i,q}\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{i,q}}$, which requires computing the gradient of the expected reward: $\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta})$. As the expectation is taken over realizations of the joint distribution $p(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}} \mid \mathcal{D}, \boldsymbol{\phi}, \boldsymbol{\theta})$, it is intractable to directly compute $\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta})$. Fortunately, we can indirectly compute this gradient by taking the expectation of the reward-weighted gradient of log-policy:

(2.2)
$$\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta}) = \int E_{(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}}) \sim p(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}} \mid \mathcal{D}, \boldsymbol{\phi}, \boldsymbol{\theta})} \left[r_i(\boldsymbol{Y}_i^{\text{new}}) \nabla_{\boldsymbol{\theta}} \log \left(\prod_{j \geq J_i} \pi(Z_{i,j+1} \mid \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}) \right) \right] p(\boldsymbol{\phi} \mid \mathcal{D}) d\boldsymbol{\phi},$$

where the policy $\pi(Z_{i,j+1} \mid \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta})$ maps the individual's up-to-date longitudinal states and treatment history to a recommended cART at each future visit j, $j > J_i$. We provide the proof of equation (2.2) in Supplementary Material Section B.

The form of (2.2) allows us to use Monte Carlo to approximate $\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta})$. Specifically, we need to 1) sample future longitudinal states $\boldsymbol{Y}_i^{\text{new}}$, which requires us to learn how the individual's states evolve over time conditional on her preceding states and treatment history from the data $\mathcal{D}$; and 2) parameterize the cART assignment policy π so that we can compute the gradient of log-policy $\nabla_{\boldsymbol{\theta}} \log \left(\prod_{j \geq J_i} \pi(Z_{i,j+1} \mid \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}) \right)$. To fulfill these two objectives, we propose a two-step approach. In the first step (Section 3), we propose to use a multivariate Gaussian process (MGP) to model the joint distribution of individual's longitudinal states. The transition dynamics $\boldsymbol{Y}_{i,j+1} = f(\overline{\boldsymbol{Y}_{ij}}, \overline{\boldsymbol{Z}_{ij}}; \boldsymbol{\phi})$ is then induced by the conditional distribution of the MGP model, and can be subsequently used for sampling future longitudinal states. The MGP is able to model multivariate longitudinal data observed at irregular time points with uncertainty quantification, which will be incorporated into the optimization. In the second step (Section 4), to mitigate the distribution shift issue arising from optimizing sequential cART assignments from observational data, we construct a pessimistic environment as a surrogate for the underlying true environment, by equipping the reward function with uncertainty penalization for safe exploration in the cART space. We conduct policy optimization with respect to the following uncertainty-penalized reward: $\widetilde{r}_i(\boldsymbol{Y}_i^{\text{new}}) = r_i(\boldsymbol{Y}_i^{\text{new}}) - \lambda u(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}})$, where the function $u(\cdot,\cdot)$ quantifies the uncertainty of the estimated dynamic model at future states $\boldsymbol{Y}_i^{\text{new}}$ with cART assignments $\boldsymbol{Z}_i^{\text{new}}$, and $\lambda \geq 0$ is a hyperparameter that controls the degree of uncertainty penalization to the reward function. To find the gradient of the log-policy, we develop a probabilistic generative model for the cART assignment $Z_{i,j+1} = \pi(\overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta})$ by representing the decision process of selecting cARTs via a tree structure with three levels. The functional form of the cART assignment π allows us to directly compute $\nabla_{\boldsymbol{\theta}} \log \left(\prod_{j \geq J_i} \pi(Z_{i,j+1} \mid \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}) \right)$, which can be then used for estimating $\boldsymbol{\theta}_i^\star$ from (2.1) though SGD.

3. First Step: Modeling Longitudinal States.

3.1. *Probability model.* In this section, we describe the proposed MGP model for individuals' longitudinal states. MGPs are a popular choice for modeling irregularly spaced multivariate longitudinal data with great flexibility and natural uncertainty quantification (Alvarez et al., 2012). Motivated by our application, we focus on continuous states but if desired, MGPs can be extended to handle non-normal states (e.g., binary) by introducing an appropriate link function (e.g., the probit link) between the non-normal states and the latent Gaussian processes (Albert and Chib, 1993).

Let $Y_{im}(t)$ denote the m-th variable for individual i at time t. Note that $Y_{im}(t)$ can be missing at any time t, and we assume they are missing at random. We construct a sampling model for individuals' longitudinal states $Y_{im}(t) = f_{im}(t) + \epsilon_{im}$, where $f_{im}(t)$ is a smooth function representing the mean of variable m for individual i at time t and $\epsilon_{im} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma_m^2)$. We place independent GP priors over functions $f_{im}(t)$'s with a shared time correlation kernel $C^t(t,t')$ for ease of computation. Because we do not expect $f_{im}(t)$ to be overly smooth, we consider the Ornstein-Uhlenbeck (OU) kernel $C^t(t,t') = \rho_t^{|t-t'|}$ whose realizations are only first-order continuous. Given $C^t(t,t')$, $(f_{i1}(t), \ldots, f_{iM}(t))$ are MGP-distributed with mean $(\mu_{i1}(t), \ldots, \mu_{iM}(t))$ and a separable covariance function $cov(f_{im}(t), f_{im'}(t')) = C^M_{mm'} C^t(t,t')$, where C^M is an $M \times M$ covariance matrix characterizing the dependence among the variables.

We model the GP mean $\mu_{im}(t)$ with a mixed-effects model,

$$\mu_{im}(t) = \boldsymbol{X}_{i0}\boldsymbol{\beta}_m + \boldsymbol{V}(t)\boldsymbol{\alpha}_{im} + h_m(\overline{Z_i(t)}), \tag{3.1}$$

where $\overline{Z_i(t)}$ denotes the treatment history of individual i until time t, $\boldsymbol{\beta}_m$ is the baseline fixed effects including an intercept, $\boldsymbol{V}(t) = (1, t)$, and $\boldsymbol{\alpha}_{im} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}_m})$ represents the random effects. The last term $h_m(\overline{Z_i(t)})$ is the key component of the GP mean, which measures not only the *instantaneous* effect of the current cART $Z_i(t)$ at time t for variable m, but also the *accumulated* effects of preceding cARTs,

$$h_m(\overline{Z_i(t)}) = \underbrace{\frac{\sum_{d=1}^D \kappa(Z_i(t), z_d)\gamma_{md}}{\sum_{d=1}^D \kappa(Z_i(t), z_d)} + \sum_{s=1}^S \frac{\sum_{d=1}^D \kappa(Z_i(t), z_d) X_{i0s}\tilde{\gamma}_{mds}}{\sum_{d=1}^D \kappa(Z_i(t), z_d)}}_{\text{instantaneous drug effect}} + \underbrace{\sum_{n=1}^N \delta_{mn} \int_0^t \mathbb{I}(\mathcal{A}_n \in Z_i(t')) e^{-(t-t')} dt'}_{\text{accumulated drug effect}}, \tag{3.2}$$

where $\mathcal{A}_n$ represents the n-th individual ART drug recorded in the dataset, $n = 1, 2, \ldots, N$. In the WIHS dataset, $N = 31$. The *instantaneous drug effect* includes the cART main effect and cART-covariate interaction effect. Since the cART space is high dimensional due to the large number of possible drug combinations, we use the subset-tree (ST) kernel approach (Jin et al., 2022) to reduce the dimension to a manageable size and encourage similar effects for similar cARTs. Specifically, we first pick a number D of representative cARTs that are commonly prescribed in clinical practice, denoted by $z_1, \ldots, z_D$. Then we calculate the similarities between the cART $Z_i(t)$ and those representatives using a similarity score function $\kappa(Z_i(t), z_d)$ induced by the ST kernel, which will be described later. The *accumulated drug effect* models the long-term effect δ_{nm} of each ART drug n that has been used by the individual i before time t on variable m, denoted by an indicator function $\mathbb{I}(\mathcal{A}_n \in Z_i(t')), t < t'$. This accumulated effect decays with time and will eventually decline to zero after an ART drug is terminated for a long time. In practice, the instantaneous drug effect is usually beneficial (e.g., viral suppression), while the accumulated drug effect can be toxic. For example, the long-term use of EFV (efavirenz) is associated with worse neurocognitive functioning (Ma et al., 2016).

Here we give a brief description of the ST kernel. We first represent each cART as a rooted tree $\mathcal{T}$ with three levels: 1) the first level indicates which drug classes are used; 2) the second level indicates how many drugs are used within each drug class; 3) the third level indicates which specific individual ART drugs are used within each drug class. Figure 2 illustrates the representation using two cARTs as an example. The main idea of the ST kernel is to compute the number of common substructures between two trees (highlighted by the yellow and blue boxes in Figure 2). Let $R_{\mathcal{T}}$ denote the set of nodes for any tree $\mathcal{T}$ and let $ch(r)$ denote the set of children nodes of the node $r \in R_{\mathcal{T}}$. The similarity score, $\kappa(\mathcal{T}_a, \mathcal{T}_b)$ between two cART trees $\mathcal{T}_a$ and $\mathcal{T}_b$, is calculated by $\kappa(\mathcal{T}_a, \mathcal{T}_b) = \sum_{r_a \in R_{\mathcal{T}_a}} \sum_{r_b \in R_{\mathcal{T}_b}} \rho(r_a, r_b)$, where $\rho(r_a, r_b)$ is defined for each pair of nodes as follows. (i) If r_a and r_b are terminal nodes

$(ch(r_a) = ch(r_b) = \emptyset)$, then $\rho(r_a, r_b) = 0$. (ii) If r_a and r_b have different sets of children nodes $(ch(r_a) \neq ch(r_b))$, then $\rho(r_a, r_b) = 0$. (iii) If r_a and r_b have the same nonempty set of children nodes, then $\rho(r_a, r_b) = \eta \prod_{s=1}^{|ch(r_a)|} \{1 + \rho(ch^s_{r_a}, ch^s_{r_b})\}$, where $|\cdot|$ is the cardinality of a set and $ch^s_{r_a}$ is a child of r_a for $s = 1, \ldots, |ch(r_a)|$. The ST kernel is able to incorporate the known clinical knowledge on drug classes and capture the similarity between cARTs across all levels of the tree representation. The hyperparameter $\eta \in (0, 1]$ is a decay factor to control the relative influence from nodes near the root to alleviate the peakiness of the ST kernel when the tree depth is large (Beck et al., 2015).

(a) cART: FTC+TDF+EFV (b) cART: FTC+TDF+DTG

FIG 2. *Tree representations of cARTs.*

In summary, the likelihood for $\boldsymbol{Y}_i \in \mathbb{R}^{M \times J_i}$ can be represented as $\text{vec}(\boldsymbol{Y}_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu}_i$ vectorizes $\{\mu_{im}(t_{ij})\}_{m=1,j=1}^{M,J_i}$ and $\boldsymbol{\Sigma}_i = C^M \bigotimes C^{J_i} + \boldsymbol{D} \bigotimes \boldsymbol{I}_{J_i}$. Here C^{J_i} is a $J_i \times J_i$ time correlation matrix specified by $C^t(t, t')$ and $\boldsymbol{D}$ is a diagonal matrix of $\{\sigma_m^2\}_{m=1}^M$.

3.2. *Posterior inference.* There are two computational challenges in the estimation procedure for the MGP. First, the ST kernels $\kappa(\cdot, \cdot)$ in (3.2) are potentially high-dimensional if the number of selected knots D is large, and are highly correlated for similar cARTs. Therefore, we consider a principal component regression method (Kendall, 1957). Specifically, let $\boldsymbol{H}_{ij}$ be a D-dimensional vector whose d-th element is $\kappa(Z_{ij}, z_d) / \sum_{d=1}^D \kappa(Z_{ij}, z_d)$ for individual i at time t_{ij}, and let $\boldsymbol{H} = (\boldsymbol{H}_{11}^{\mathrm{T}}, \ldots, \boldsymbol{H}_{1J_1}^{\mathrm{T}}, \ldots, \boldsymbol{H}_{I1}^{\mathrm{T}}, \ldots, \boldsymbol{H}_{IJ_I}^{\mathrm{T}})^{\mathrm{T}}$ be the $\sum_{i=1}^I J_i \times D$-dimensional ST kernel design matrix in (3.2). We perform the principal component analysis on $\boldsymbol{H}$ and retain the first $D^\star$ principal components that explain at least 99.9% of the total variance, where the resulting $\sum_{i=1}^I J_i \times D^\star$ matrix is denoted by $\boldsymbol{H}^\star$. Then the instantaneous drug effect, i.e., the first two terms in (3.2), can be approximated by $\boldsymbol{\gamma}^\star \boldsymbol{H}^\star_{ij} + \tilde{\boldsymbol{\gamma}}^\star \boldsymbol{H}^\star_{ij} \bigotimes \boldsymbol{X}_{i0}$, where $\boldsymbol{\gamma}^\star$ and $\tilde{\boldsymbol{\gamma}}^\star$ are the $M \times D^\star$ main effect matrix and $M \times (S \times D^\star)$ interaction effect matrix that need to be estimated. Second, estimating the covariance matrix C^M requires fitting $M(M+1)/2$ parameters, which can be computationally inefficient if M is moderately large (e.g., $M \geq 4$). We speed up the computation by dimension reduction following the idea of intrinsic coregionalization model (ICM, Alvarez et al. (2012)). Specifically, in ICM, each state function $f_{im}(t)$ is assumed be a linear combination of L independent latent GPs $g_{i1}(t), \ldots, g_{iL}(t)$ with common correlation kernel $C^t(t, t')$ such that $\boldsymbol{f}_i(t) = \sum_{l=1}^L \boldsymbol{b}_l g_{il}(t)$, where $\boldsymbol{b}_l = (b_{1l}, \ldots, b_{Ml})^{\mathrm{T}}$ is the vector representing the collection of linear coefficients associated with the l-th latent function. It follows that the covariance function for $\boldsymbol{f}_i(t)$ is $cov(\boldsymbol{f}_i(t), \boldsymbol{f}_i(t')) = \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} \bigotimes C^t(t, t')$, where $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L)$ is an $M \times L$ matrix. Therefore, the number of parameters to be estimated in the covariance matrix decreases from $M(M+1)/2$ to $M \times L$ by using the ICM.

We complete the model by imposing the following priors. To encourage parsimony, we assign the horseshoe prior (Carvalho, Polson and Scott, 2010) on the coefficients of baseline and cART effects $\boldsymbol{\psi}_m = (\boldsymbol{\beta}_m, \boldsymbol{\gamma}^\star_m, \tilde{\boldsymbol{\gamma}}^\star_m)$. In particular, we assume that $\psi_{ms} \mid \tau, \nu_s \sim \mathcal{N}(0, \tau^2 \nu_s^2)$,

$\tau \sim \mathcal{C}^+(0,1)$, and $\nu_s \sim \mathcal{C}^+(0,1)$, where $s = 2, \ldots, S + D^\star + S \times D^\star$ and $\mathcal{C}^+(0,1)$ is the standard half-Cauchy distribution. Note that the intercept ψ_{m1} is not shrunk by the horseshoe prior; instead it is assigned a normal prior $\psi_{m1} \sim \mathcal{N}(0, \sigma^2_{\psi_1})$. The horseshoe prior is a global-local shrinkage procedure in which the global parameter τ tries to push all of the coefficients towards zero, while the heavy-tail property of the Cauchy distribution for the local parameters ν_s allows sufficiently large coefficients (i.e., signals) to escape from the global shrinkage effect. We assume the long-term effect $\delta_{mn} \sim \text{Unif}[0, +\infty)$ if a higher value of the variable m indicates a worse symptom (e.g., depression); $\delta_{mn} \sim \text{Unif}(-\infty, 0]$ if a lower value of the variable m indicates a worse symptom (e.g., cognition). In addition, we assign a conjugate Inverse-Wishart(a_0, A_0^{-1}) prior on $\mathbf{\Sigma}_{\boldsymbol{\alpha}_m}$ and a conjugate Inverse-Gamma(d_1, d_2) on σ^2_m for ease of computation. We assume flat priors for ρ_t in the time correlation kernel $C^t(t, t')$ and $\boldsymbol{B}$. We carry out posterior inference using the Markov chain Monte Carlo (MCMC) sampler, the details of which are included in the Supplementary Material Section A.

4. Second Step: Optimizing cART Assignments. In this section, we propose an uncertainty-penalized policy optimization procedure to optimize personalized sequential cART assignments that maximize individuals' long-term health outcomes. We first introduce an uncertainty-penalized reward function in the setting of a pessimistic environment to mitigate the distribution shift issue, and then build a generative model for cART assignments by developing a three-level decision process. Lastly, we describe the details of the SGD algorithm for policy optimization.

4.1. *An uncertainty-penalized reward.* In HIV clinical practice, many factors contribute to PWH's long-term health and quality of life. Virologic control is the primary goal of cART since failing to suppress viral load can significantly increase HIV-related mortality and morbidity (Ledergerber et al., 1999). Moreover, PWH are at increased risks for a range of comorbidities, such as kidney diseases (D'Souza, Golub and Gange, 2019). Other clinical factors such as depression and cognition are also important when physicians make treatment decisions for PWH (Langebeek et al., 2017). Here we define our reward based on viral load, kidney function, and depression to illustrate the proposed method, which can be easily extended to include other factors.

For individual i at a future visit j, $j > J_i$, let Y_{ij1}, Y_{ij2}, and Y_{ij3} denote her future values of depression, viral load, and estimated glomerular filtration rate (eGFR, a kidney function indicator), respectively. We define a personalized reward function for the next two years (i.e., the next 4 visits for 2 visits per year),

(4.1)
$$r_i(\mathbf{Y}_i^{\text{new}}) = -\sum_{j=J_i+1}^{J_i+4} \left\{ \underbrace{w_{i1} Y_{ij1}}_{\text{depression}} + \underbrace{w_{i2} |Y_{ij2} - T_V| \mathbb{I}(Y_{ij2} > T_V)}_{\text{viral load}} + \underbrace{w_{i3} |Y_{ij3} - T_E| \mathbb{I}(Y_{ij3} < T_E)}_{\text{eGFR}} \right\}.$$

Here T_V and T_E denote the known clinical thresholds for viral load and eGFR, i.e., if $Y_{ij2} > T_V$, or $Y_{ij3} < T_E$, then individual i's viral load or eGFR at visit j is in the abnormal range and immediate medical care is needed. If an individual's viral load is in its normal range (i.e., $Y_{ij2} \leq T_V$), it is not necessary to adjust the cART assignment to further reduce the viral load; accordingly the proposed reward function does not warrant additional rewards due to the term $\mathbb{I}(Y_{ij2} > T_V)$. Same can be said for eGFR. In contrast, since lower depression is always better, the proposed reward function always encourages lower depression scores. The personalized weight $\boldsymbol{w}_i = (w_{i1}, w_{i2}, w_{i3})$ determines the relative contribution of depression, viral load, and eGFR to the reward, which should be chosen by practitioners. For example, for one individual whose viral load and eGFR are within the normal range but whose depression

score is high, a large weight can be assigned to the depression term so that the optimization procedure can find the cART that reduces the depression score the most.

Inspired by Yu et al. (2020b) who developed a model-based uncertainty-penalized policy optimization method to mitigate the distribution shift issue in offline RL, we build a pessimistic environment based on the uncertainty quantified from the learned probabilistic dynamic model in the first step. Specifically, we define an uncertainty-penalized reward $\widetilde{r}_i(\boldsymbol{Y}_i^{\text{new}})$ that penalizes $r_i(\boldsymbol{Y}_i^{\text{new}})$ in (4.1) for each pair of state (i.e., individual's longitudinal states $\boldsymbol{Y}_i^{\text{new}}$) and action (i.e., cART assignments $\boldsymbol{Z}_i^{\text{new}}$) by its estimated uncertainty in the learned dynamics: $\widetilde{r}_i(\boldsymbol{Y}_i^{\text{new}}) = r_i(\boldsymbol{Y}_i^{\text{new}}) - \lambda u(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}})$, where $u(\cdot,\cdot)$ quantifies the uncertainty of the state-action pair $(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}})$. In this paper, we use $u(\boldsymbol{Y}_i^{\text{new}}, \boldsymbol{Z}_i^{\text{new}}) = \sum_{j=J_i+1}^{J_i+4} \sum_{m=1}^{M} \sqrt{\text{Var}(Y_{ijm} \mid Z_{ij}, \mathcal{D})}$, where Y_{ijm} is the predicted value of the m-th variable for individual i at future visit j, and the variance is calculated by its posterior predictive distribution conditional on the cART assignment Z_{ij} and the observed data $\mathcal{D}$. Our formulation is motivated by the theoretical guarantee established in Yu et al. (2020b) who showed that the learned policy from the pessimistic environment performed at least as well as the behavior policy that generated the observational data.

4.2. *Decision process for assigning cART.* In order to find $\boldsymbol{\theta}^\star$ that maximizes the expected reward $R_i(\boldsymbol{\theta})$ defined in (2.1) for individual i via the policy optimization procedure, we need to compute the gradient of the expected reward: $\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta})$. After discussing with clinicians, we construct a clinically meaningful policy π (i.e., the probabilistic generative model for the cART assignment $Z_{i,j+1} = \pi(\overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta})$) by representing the decision process of selecting a cART conditional on individuals' preceding longitudinal states and treatment histories via a tree structure with three levels, illustrated in Figure 3.

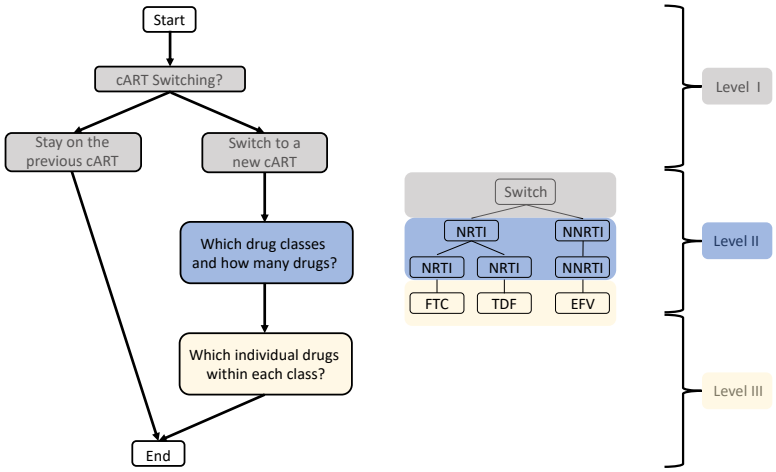


FIG 3. *Illustration of the three-level decision process for selecting a cART conditional on individuals' preceding longitudinal states and treatment histories.*

In the first level of the decision process, we determine whether individual i needs to switch to a new cART or stay on her previous cART. In HIV clinical practice, physicians make switch or no-switch decisions based on patients' health outcomes. Specifically, at any future visit j, $j > J_i$, individual i needs to switch to a new cART if any one of the outcomes is abnormal, i.e., $Y_{ij1} > T_D$, $Y_{ij2} > T_V$, or $Y_{ij3} < T_E$, where T_D is the clinical threshold for depression; otherwise, she will stay on her previous cART.

If switching is necessary, we need to determine the optimal new cART assignment. As there are $N = 31$ individual ART drugs recorded in the WIHS, a straightforward approach is to represent $Z_{i,j+1}$ as an N-dimensional binary vector and apply logistic regression model to each individual drug regressed on $\overline{\boldsymbol{Y}_{ij}}$ and $\overline{Z_{ij}}$. Although simple to use, this method ignores the structure of cART such as drug class information and yields a large number of parameters to be estimated. Also, considering each drug independently leads to 2^N possible drug combinations, in which most of them would be unrealistic and never be prescribed in clinical practice. For example, practical cARTs typically take two NRTI drugs as backbone then add drugs from other drug classes.

To efficiently explore the high-dimensional cART space when treatment switching is necessary, we represent the new cART as a tree rooted at node "Switch", as shown in Figure 3. Then we determine which drug classes to use and how many drugs used in each drug class in the second level of the decision process. Lastly, we determine what specific individual drugs to be selected in the third level. For example, the new cART in Figure 3 is a combination of two NRTI drugs FTC and TDF and an NNRTI drug EFV. During this decision process, known clinical knowledge can be incorporated to reduce the dimension. For instance, except for the NRTI drug class that is frequently used as the cART backbone, prescribing more than one individual drug from other drug classes is uncommon in practice.

We represent this three-level decision process as a hierarchical model,

$$\pi(Z_{i,j+1} \mid \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}) = \underbrace{p(a_{ij}^{(1)} \mid \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}^{(1)})}_{\text{first level}} \underbrace{p(\boldsymbol{a}_{ij}^{(2)} \mid a_{ij}^{(1)}, \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}^{(2)})}_{\text{second level}} \underbrace{p(\boldsymbol{a}_{ij}^{(3)} \mid a_{ij}^{(1)}, \boldsymbol{a}_{ij}^{(2)}, \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}^{(3)})}_{\text{third level}}, \tag{4.2}$$

where $a_{ij}^{(1)}$ is a binary indicator for cART switching, $\boldsymbol{a}_{ij}^{(2)}$ represents the number of drugs used in each drug class, and $\boldsymbol{a}_{ij}^{(3)}$ contains the selected individual ART drugs in each drug class, the details of which are given below.

First-level decision. We model the first level decision by applying hard thresholding on depression, viral load, and eGFR at known clinically abnormal thresholds,

$$p(a_{ij}^{(1)} = 1 \mid \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}^{(1)}) = \begin{cases} 1, & \text{if } Y_{ij1} > T_D \text{ or } Y_{ij2} > T_V \text{ or } Y_{ij3} < T_E, \\ 0, & \text{otherwise,} \end{cases} \tag{4.3}$$

where $a_{ij}^{(1)} = 1$ indicates cART switching, and $a_{ij}^{(1)} = 0$ represents staying on the previous cART Z_{ij} until the next visit $j + 1$.

Second-level decision. Assume that there are K different drug classes, and that the maximum number of drugs used in the k-th drug class is C_k, $k = 1, 2, \ldots, K$. Conditional on the cART switching, we model the number of drugs used in each drug class for the new cART independently using the following multi-class logistic regression model,

$$p(\boldsymbol{a}_{ij}^{(2)} \mid a_{ij}^{(1)} = 1, \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}^{(2)}) = \prod_{k=1}^{K} p(a_{ijk}^{(2)} \mid a_{ij}^{(1)} = 1, \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}_k^{(2)}), \tag{4.4}$$

where

$$p(a_{ijk}^{(2)} = c_k \mid a_{ij}^{(1)} = 1, \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}_k^{(2)}) = \begin{cases} \dfrac{\exp\left(\boldsymbol{Y}_{ij}^{\mathrm{T}} \boldsymbol{\theta}_{kc_k}^{(2)}\right)}{1 + \sum_{c'_k=1}^{C_k} \exp\left(\boldsymbol{Y}_{ij}^{\mathrm{T}} \boldsymbol{\theta}_{kc'_k}^{(2)}\right)}, & c_k = 1, 2, \ldots, C_k, \\ \dfrac{1}{1 + \sum_{c'_k=1}^{C_k} \exp\left(\boldsymbol{Y}_{ij}^{\mathrm{T}} \boldsymbol{\theta}_{kc'_k}^{(2)}\right)}, & c_k = 0. \end{cases} \tag{4.5}$$

Note that the second-level decision only depends on the individual's most recent state $\boldsymbol{Y}_{ij}$, which can be easily extended to her entire history if necessary.

Third-level decision. Assume that there are a number of N_k possible individual ART drugs for each drug class k. Note that $\sum_{k=1}^{K} N_k \leq N$ since some ART drugs are no longer available due to their sub-optimal antiviral potency or unacceptable toxicities. Given c_k drugs have been selected for each drug class k in the second level, we select individual ART drugs using the Wallenius' noncentral hypergeometric distribution (WNH, Wallenius 1963),

$$p(\boldsymbol{a}_{ij}^{(3)} \mid a_{ij}^{(1)}, \boldsymbol{a}_{ij}^{(2)}, \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}^{(3)}) = \prod_{k=1}^{K} \int_0^1 \prod_{n_k=1}^{N_k} (1 - x^{\xi_{n_k}})^{a_{ijkn_k}^{(3)}} dx, \tag{4.6}$$

where $\xi_{n_k} = \omega_{n_k} / \{\sum_{n_k'=1}^{N_k} \omega_{n_k'} (1 - a_{ijkn_k'}^{(3)})\}$, $\omega_{n_k} = \exp(\boldsymbol{Y}_{ij}^{\mathrm{T}} \boldsymbol{\theta}_{kn_k}^{(3)}) / \{\sum_{n_k'=1}^{N_k} \exp(\boldsymbol{Y}_{ij}^{\mathrm{T}} \boldsymbol{\theta}_{kn_k'}^{(3)})\}$, for $n_k = 1, 2, \ldots, N_k$. For each drug class k, $a_{ijkn_k}^{(3)}$ is a binary variable indicating whether the n_k-th ART drug is included in the cART. The binary vectors $\boldsymbol{a}_{ijk}^{(3)} = (a_{ijk1}^{(3)}, \ldots, a_{ijkN_k}^{(3)})$ satisfy the natural constraint $\sum_{n_k=1}^{N_K} a_{ijkn_k}^{(3)} = a_{ijk}^{(2)}$ by the assumption of the WNH distribution. Specifically, we need to select c_k out of N_k drugs for the drug class k, each of which has weight ω_{n_k} modeled by a logistic regression with covariates $\boldsymbol{Y}_{ij}$. The odds ratio ξ_{n_k} measures the relative probability of selecting the n_k-th individual drug compared to other drugs that have not been selected in the k-th drug class.

4.3. *Policy optimization.* The parametric form of the proposed three-level decision process allows for computing the gradient of the expected reward $\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta})$, a key quantity for applying SGD to obtain $\boldsymbol{\theta}_i^{\star}$ that optimizes sequential cART assignments. The details of the SGD algorithm are described in Algorithm 1. We first draw a number E samples of $\boldsymbol{\phi}$ from its posterior distribution $p(\boldsymbol{\phi} \mid \mathcal{D})$. At each iteration of the SGD, given the current value of $\boldsymbol{\theta}$, we simulate the individual's future states $\boldsymbol{Y}_{ij}$'s from its posterior predictive distribution and sample the future cARTs Z_{ij}'s from the three-level decision model for each sample of $\boldsymbol{\phi}$, $j = J_i + 1, \ldots, J_i + 4$ (see details in Supplementary Material Section C). Next we compute the gradient of the log-policy $\nabla_{\boldsymbol{\theta}} \log \left(\prod_{j \geq J_i} \pi(Z_{i,j+1} \mid \overline{\boldsymbol{Y}_{ij}}, \overline{Z_{ij}}; \boldsymbol{\theta}) \right)$ (see Supplementary Material Section D). Then we approximate $\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta})$ in (2.2) using Monte Carlo. Note that in the Step 11 of Algorithm 1, we subtract the average reward. This "baseline subtraction" trick significantly reduces the variance while still yielding an unbiased estimate of the gradient (Greensmith, Bartlett and Baxter, 2004). Lastly, we select the optimal policy parameter $\boldsymbol{\theta}^{\star}$ to be the one yielding the highest expected reward across all SGD iterations.

5. Simulation Study. We conducted simulation studies to evaluate performance of the proposed Bayesian decision framework and compared it to alternative methods in terms of the expected rewards under the estimated optimal cARTs. To illustrate the clinical utility of the proposed method, we demonstrated how the estimated personalized optimal sequential cARTs can improve PWH's health outcomes at both individual- and population-level.

5.1. *Simulation setup.* We simulated a dataset mimicking the WIHS dataset composed of longitudinal measurements with missing data. Assume that there were $I = 200$ individuals with $M = 3$ *state* variables including individuals' depression scores, viral load (in log-scale), and eGFR, and $S = 3$ baseline covariates with one intercept, one binary covariate, and one continuous covariate, i.e., $\boldsymbol{X}_{i0} = (1, x_{i1}, x_{i2})$, where x_{i1}'s and x_{i2}'s were respectively generated from Bernoulli(0.6) and a standard normal distribution, $i = 1, 2, \ldots, I$. Individuals' treatment histories $\boldsymbol{Z}_i$ were randomly sampled from the WIHS dataset without replacement, resulting in the number of visits per individual to range from 2 to 46. Conditional on the number of visits J_i for individual i, the number of observed measurements for each variable m was independently sampled from Poisson(25) truncated by 1 and J_i, for $m = 1, 2, \ldots, M$,

Algorithm 1 Stochastic Gradient Descent for optimizing $\boldsymbol{\theta}$ for any individual i

Input: initial value $\boldsymbol{\theta}_0$, step size $s_{i,q}(q = 1, 2, \ldots, Q)$, posterior samples $\boldsymbol{\phi}_{(e)}(e = 1, 2, \ldots, E)$, data $\mathcal{D}$, reward weight $\boldsymbol{w}_i$, known clinical thresholds T_D, T_V, T_E, and hyperparameter λ.
Initialize: $\boldsymbol{\theta}_{i,1} \leftarrow \boldsymbol{\theta}_0$
for $q = 1, \ldots, Q$ **do**
  for $e = 1, \ldots, E$ **do**
    for $j = J_i + 1, \ldots, J_i + 4$ **do**
      Sample $\boldsymbol{Y}_{ij(e)}$ and $Z_{ij(e)}$ conditional on $\mathcal{D}$, $\boldsymbol{\phi}_{(e)}$, and $\boldsymbol{\theta}_{i,q}$
    end for
    Compute the uncertainty-penalized reward $\widetilde{r}_i\left(\boldsymbol{Y}_{i(e)}^{\text{new}}\right) = r_i\left(\boldsymbol{Y}_{i(e)}^{\text{new}}\right) - \lambda u\left(\boldsymbol{Y}_{i(e)}^{\text{new}}, \boldsymbol{Z}_{i(e)}^{\text{new}}\right)$
  end for
  $\overline{R_i(\boldsymbol{\theta}_{i,q})} \leftarrow \frac{\sum_{e=1}^{E} \widetilde{r}_i\left(\boldsymbol{Y}_{i(e)}^{\text{new}}\right)}{E}$
  $\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta}_{i,q}) \leftarrow \frac{\sum_{e=1}^{E}\left(\widetilde{r}_i\left(\boldsymbol{Y}_{i(e)}^{\text{new}}\right) - \overline{R_i(\boldsymbol{\theta}_{i,q})}\right)\nabla_{\boldsymbol{\theta}} \log\left(\prod_{j \geq J_i} \pi(Z_{i,j+1(e)} | \overline{\boldsymbol{Y}_{ij(e)}}, \overline{Z_{ij(e)}}; \boldsymbol{\theta}_{i,q})\right)}{E}$
  $\boldsymbol{\theta}_{i,q+1} \leftarrow \boldsymbol{\theta}_{i,q} + s_{i,q} \nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\theta}_{i,q})$
end for
$q^\star \leftarrow \arg\max_q \overline{R_i(\boldsymbol{\theta}_{i,q})}$
Output: $\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}_{i,q^\star}$

yielding an overall 20% missing rate. There were $N = 30$ individual ART drugs in treatment histories of this simulated dataset. We selected representative cARTs $z_1, \ldots, z_D$ if a cART z_d has been used in more than 10 visits among all 200 individuals, yielding $D = 67$. We performed the principal component analysis on the kernel design matrix $\boldsymbol{H}$ based on these 67 representatives, and selected the first $D^\star = 41$ principal components that explained 99.9% variation of the original matrix. We set the decay factor $\eta = 0.5$ and then computed the similarity scores between different cARTs using the ST kernel.

We assumed that the simulated true fixed effect parameters were $\boldsymbol{\beta}_1 = (25, 1, 2)$, $\boldsymbol{\beta}_2 = (4.5, -0.5, 1)$, and $\boldsymbol{\beta}_3 = (75, -4, 2)$, corresponding to depression scores, viral load, and eGFR, respectively. We randomly generated the cART coefficients $\boldsymbol{\gamma}_m^\star$ and $\tilde{\boldsymbol{\gamma}}_m^\star$ from standard multivariate normal distributions. We set the drug toxicity coefficients to be $\boldsymbol{\delta}_n = (1, 0.5, -2)$ for NRTI drugs, and $\boldsymbol{\delta}_n = (0, 0, 0)$ for non-NRTI drugs. We assumed the random effect covariance matrices to be $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_1} = \boldsymbol{I}_Q$, $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_2} = 0.5\boldsymbol{I}_Q$, and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}_3} = 2\boldsymbol{I}_Q$. We set C^M to be a covariance matrix with diagonal elements $(c_{11}, c_{22}, c_{33}) = (5, 1, 10)$ and off-diagonal elements $(c_{12}, c_{13}, c_{23}) = (1.67, -3.53, -0.79)$, $\rho_t = 0.5$, and $\boldsymbol{\sigma}^2 = (10, 1, 20)$. Based on the proposed MGP model in Section 3, we generated the individuals' longitudinal states $\boldsymbol{Y}_i$'s.

In the decision process, we incorporated known clinical knowledge to make the cARTs from the generative model π clinically meaningful. Specifically, we first divided the NRTI drug class into two sub-classes (NRTI1 and NRTI2) so that drugs within the same sub-class share similar profiles, and hence at most one drug from each sub-class would be used in clinical practice. Furthermore, we removed individual ART drugs that were no longer recommended by FDA or discontinued, yielding $K = 6$ drug classes with $\sum_{k=1}^{K} N_k = 16$ individual ART drugs. They included NRTI1 drugs 3TC and FTC, NRTI2 drugs ABC, TAF, and TDF, NNRTI drugs EFV, ETR, NVP, and RPV, PI drugs ATV, DRV, and LPV, INSTI drugs DTG, EVG, and RAL, and EI drug MVC. In addition, the Booster RTV will be included in the cART if any of the PI drugs is selected, and the Booster COBI will be included if the INSTI drug EVG is selected. Such a setup makes the maximum number of drugs used in each defined drug class for a clinically meaningful cART to be $C_k = 1$ for $k = 1, 2, \ldots, K$. The thresholds for depression, viral load, and eGFR were set to be $T_D = 16$ (Zich, Attkisson and Greenfield, 1990), $T_V = \log(20)$ (Raboud et al., 1998), and $T_E = 60$ (Ma et al., 2017).

5.2. *Simulation results.* We first applied the proposed MGP model to the simulated dataset. The hyperparameters were set to be $\sigma_{\psi_1}^2 = 100$, $a_0 = 3$, $A_0^{-1} = \boldsymbol{I}_2$, $d_1 = 1$, $d_2 = 1$,

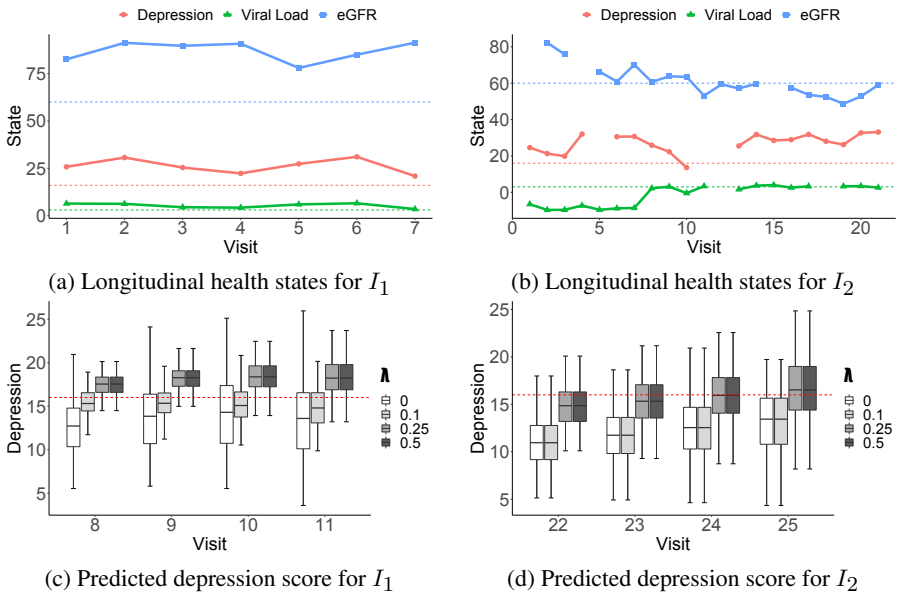
$\eta = 0.5$, and $L = 1$. We ran 10,000 MCMC iterations with an initial burn-in of 5,000 iterations and a thinning factor of 10. The convergence diagnostic was assessed using R package **coda**, including trace plots of the post-burn-in MCMC samples for some randomly selected parameters (Supplementary Figure S1), showing no issues of non-convergence. Supplementary Figure S2 plots the 95% estimated credible intervals (CI) for selected parameters, showing that all 95% CIs are centered around the simulated true values. Supplementary Table S3 summarizes the mean squared error (MSE) of the post-burn-in MCMC posterior samples, indicating that the proposed method performs well in terms of parameter estimation.

To demonstrate that the MGP can handle discrete state variables, we conducted an additional simulation study in Supplementary Section F1, where individuals' longitudinal states consist of both binary and continuous outcomes. The proposed MGP was able to recover the simulated true parameter values. In Supplementary Section F2, we compared the performance of the proposed MGP model to the random forest model and found that the MGP outperformed the random forest in terms of prediction accuracy. To demonstrate the robustness of the MGP model with respect to model misspecification, we conducted an additional simulation study with a sequential data generation scheme in Supplementary Section F3. The MGP yielded satisfactory prediction performance under this model misspecification scenario.

We then applied the proposed uncertainty-penalized policy optimization in Section 4 to the simulated dataset to estimate the personalized optimal sequential cART assignments under different choices of $\lambda = 0, 0.1, 0.25, 0.5$. We implemented the SGD algorithm with 1000 steps and used a fixed step size, i.e., $s_{i,q} = 0.1$, $q = 1, 2, \ldots, Q$. The starting parameter values $\boldsymbol{\theta}_0$ in Algorithm 1 were set to be all zeros, so that all possible drug combinations can be generated with equal probabilities.

At the individual level, we considered two randomly-selected subjects (denoted by I_1 and I_2) to demonstrate that the proposed method can recommend personalized optimal sequential cART. We assigned equal weights to depression, viral load, and eGFR in the reward function for both individuals, i.e., $\boldsymbol{w}_i = (1/3, 1/3, 1/3)$, $i = 1, 2$. Individual I_1 had $J_1 = 7$ visits and stayed on the same cART 3TC+AZT+NFV (two NRTIs + one PI) all the time, while individual I_2 had $J_2 = 21$ visits, and used the cART 3TC+AZT+NVP (two NRTIs + one NNRTI) in her first seven visits, then switched to FTC+TDF+ATV+RTV (two NRTIs + one PI + one Booster) for the rest of her visits. As shown in Figure 4(a, b), the depression scores of both I_1 and I_2 at their last visits were beyond the normal range, indicating that switching to a new cART would be desired according to the first-level decision in (4.3). Note that the discontinuity of the curve in Figure 4(b) is due to missing data.

For individual I_1, switching to the cART FTC+ABC+NVP+MVC (two NRTIs + one NNRTI + one EI) for the next three visits (i.e., visits 8-10) and then replacing the NNRTI drug NVP with another NNRTI drug ETR at the last visit was her personalized optimal sequential cART when $\lambda = 0$. The sequence of cARTs 3TC+ABC+NVP (two NRTIs + one NNRTI) for her next four visits was optimal under $\lambda = 0.1$, while the sequence of cARTs 3TC+ABC+NVP+ATV+RTV (two NRTIs + one NNRTI + one PI + one Booster) was optimal under $\lambda = 0.25$ and 0.5. For individual I_2, prescribing FTC+ABC+RPV (two NRTIs + one NNRTI) for the next two years (i.e., visits 22 to 25) was optimal when $\lambda = 0$ and 0.1, and the sequence of cARTs FTC+TDF+EFV (two NRTIs + one NNRTI) was optimal under $\lambda = 0.25$ and 0.5. As shown in Figure 4(c,d), all the recommended cARTs under different choices of λ would alleviate depressive symptoms for both individuals, while the optimal cARTs under $\lambda = 0$ and $\lambda = 0.1$ can reduce their depression scores to the normal range. Furthermore, Supplementary Figure S4 plots the predicted viral loads and eGFRs for individual I_1 and I_2 under their personalized optimal sequential cART with respect to different choices of λ, indicating that all the recommended cARTs can successfully suppress their viral load and control their eGFR within the normal range.



(a) Longitudinal health states for I_1

(b) Longitudinal health states for I_2

(c) Predicted depression score for I_1

(d) Predicted depression score for I_2

FIG 4. *Panels (a, b) plot the longitudinal health states for two randomly selected individuals I_1 and I_2 in the simulation study. The dashed lines represent the thresholds for depression score, viral load, and eGFR. Panels (c, d) plot the predicted depression scores for I_1 and I_2 under their personalized optimal sequential cART assignments with respect to $\lambda = 0, 0.1, 0.25, 0.5$. The dashed red lines represent the thresholds for depression score.*

Moreover, Figure 4(c,d) show that when λ increased, both individuals' predicted depression scores also increased, while the predicted uncertainties decreased, demonstrating the trade-off between the gain and risk introduced by the uncertainty-penalized policy optimization. Specifically, high uncertainties when λ was small arose from the fact that these cARTs were rarely used in the observed data, which was captured by our MGP model in the first step. For example, in the simulated dataset, the optimal cART for individual I_2 under $\lambda = 0.25$ and 0.5 (i.e., FTC+TDF+EFV) was observed for 265 times, while the optimal cART under $\lambda = 0$ and 0.1 (i.e., FTC+ABC+RPV) was never used. Therefore, there is a good practical reason why FTC+TDF+EFV may be preferred over FTC+ABC+RPV for individual I_2 even though the latter has a lower predicted depression score.

At the (sub)population level, we demonstrated how the estimated optimal sequential cART can improve PWH's health outcomes. Specifically, we selected individuals who were depressed but with normal viral load and kidney function at their last visits (i.e., $Y_{i,J_i,1} > T_D$, $Y_{i,J_i,2} \leq T_V$, and $Y_{i,J_i,3} \geq T_E$), resulting in a subpopulation of 27 individuals. We set their reward weights to be $\boldsymbol{w}_i = (0.8, 0.1, 0.1)$ since reducing their depression scores is the priority for their long-term health. Under the personalized optimal sequential cART found by the SGD algorithm when $\lambda = 0$, the average reward (4.1) for this subpopulation increased from -88.5 during their last four visits to -25.1 during their future four visits. In addition, the average depression scores of these 27 individuals reduced from 27.4 to 7.5, with 24 of them being not depressed in the future two years (recall that $T_D = 16$).

Lastly, to demonstrate the advantage of the proposed two-step Bayesian decision framework, we compared its performance in estimating personalized optimal sequential cARTs with three alternative strategies. The first alternative prescribes the same cART as what the individual has been using. The second alternative optimizes the individual's cARTs one step at a time, each of which maximizes the expected reward of the next visit. The third alternative is the neural fitted-Q (NFQ) learning (Riedmiller, 2005), which is a model-free RL algorithm to estimate the Q-function using neural networks. The proposed method achieved the high-

est expected reward compared to the three alternatives. The details of the three alternative methods and comparison are described in Supplementary Material Section F4.

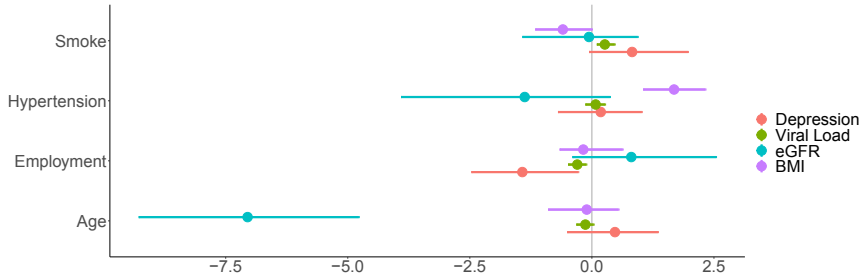
6. Application: WIHS Data Analysis. We applied the proposed method to the WIHS dataset to demonstrate its clinical utility. We included all women from the Washington, D.C. site with at least two visits, yielding a total of $I = 339$ individuals. Depression scores assessed by the Center for Epidemiological Studies Depression Scale (CES-D, Radloff (1977)), viral load (in log-scale), eGFR, and BMI were collected as individuals' longitudinal *state* variables ($M = 4$) at each follow-up visit, resulting in 8% missing data rate. We extracted the following sociodemographic, behavioral, and clinical risk factors as baseline covariates: age, smoking status, substance use (e.g., heroin), employment status, hypertension, and diabetes. A total of $N = 31$ ART drugs in $K = 6$ drug classes were recorded in this dataset. We selected $D = 105$ representative cARTs using the same criterion as in the simulation study.

6.1. *Results: MGP model fitting.* We first applied the proposed MGP to the WIHS dataset using the same hyperparameters as in the simulation study. We retained the first $D^\star = 51$ principal components that explained 99.9% variation of the original ST kernel matrix. We ran 5,000 MCMC iterations after a burn-in of 20,000 iterations, and a thinning factor of 10.

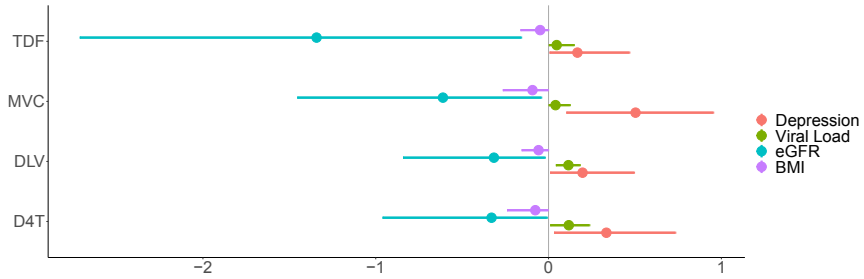
We summarized the parameter estimation for some selected covariates effects on *state* variables in Figure 5. Figure 5(a) plots the posterior means with 95% CIs for the estimated coefficients with respect to baseline covariates age, employment, hypertension, and smoking status. All these results are consistent with the findings in medical literature. Age was negatively associated with eGFR, indicating that older people had an increased risk of renal disease since kidney function declined over time due to aging (Islam et al., 2012). Unemployment status was associated with higher levels of depressive symptoms and viral load in PWH (Zeng et al., 2019). Furthermore, there was a positive relationship between hypertension and BMI since obesity is a major risk factor for hypertension (Bloomfield et al., 2011). Figure 5(a) also indicates that smokers with HIV had a higher HIV viral load (Pollack et al., 2017). Figure 5(b) plots the posterior means and 95% CIs for selected ART drug toxicity coefficients. The NRTI drug D4T was positively associated with depressive symptoms. Arenas-Pinto et al. (2016) reported that D4T can cause a variety of systemic discomforts including peripheral neuropathy. Figure 5(b) also shows that the NNRTI drug DLV was associated with a higher level of viral load in the long term, which is also supported by existing studies (Yazdanpanah et al., 2004). In fact, both D4T and DLV are no longer recommended by the U.S. Department of Health and Human Services in the general guidelines. Furthermore, a positive relationship was observed between the EI drug MVC and depression (Williams et al., 2020), and a negative relationship was observed between NRTI drug TDF and eGFR (Surial et al., 2020). Lastly, we analyzed the cART effects in Supplementary Material Section G.

6.2. *Results: personalized optimal sequential cART assignments.* We applied the proposed policy optimization to estimate the personalized optimal sequential cART assignments initially with $\lambda = 0$ (i.e., no uncertainty penalty) and later with different choices of λ. The settings were the same as in the simulation study.

We demonstrate how the estimated optimal sequential cART improves individuals' health outcomes by randomly selecting two individuals: S_1 and S_2. Individual S_1 had $J_1 = 13$ visits with a depression score of 5, viral load of 4.4, and eGFR of 121.3 at her last visit. In her reward function, the weights on depression score, viral load, and eGFR were set to be $\boldsymbol{w}_1 = (0.1, 0.8, 0.1)$ since her viral load was in the abnormal range (recall that $T_D = 16$, $T_V = \log(20)$, and $T_E = 60$). Individual S_2 had $J_2 = 31$ visits with a depression score of 16, viral load of 5, and eGFR of 102.8 at her last visit. We set the weights in her reward



(a) Baseline coefficients



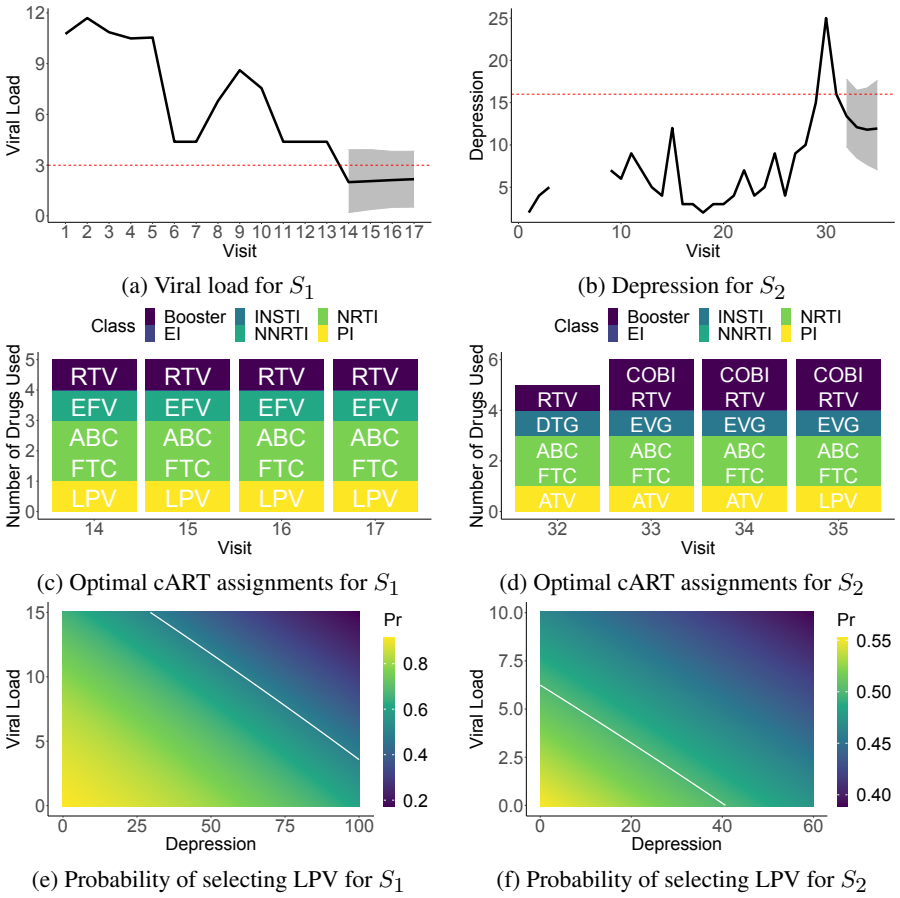
(b) ART drug toxicity coefficients

FIG 5. *Posterior means and* 95% *CIs for the estimated effects of selected baseline covariates and ART drug toxicities in the WIHS data analysis. The dots represent the posterior means.*

function to be $\boldsymbol{w}_2 = (1/3, 1/3, 1/3)$ since her health states were relatively normal. Figure 6(a) shows that S_1's personalized optimal sequential cART assignments (shown in Figure 6(c)) successfully suppressed her viral load from 4.4 at her last visit to 2.1 in the next four visits, a 52% improvement. Figure 6(b) plots the predicted depression scores for individual S_2 under her optimal sequential cART assignments (shown in Figure 6(d)), which decreased the depression score from 16 at her last visit to 12.3 in the next four visits, a 23% improvement.

To further interpret the estimated optimal parameters $\boldsymbol{\theta}_i^\star$ and their corresponding decision rules, we plot the probabilities of including the PI drug LPV for both individuals in their optimal sequential cART assignments versus their predicted viral loads and depression scores in Figure 6(e, f). The white lines represent decision boundaries. As shown in Figure 6(e, f), a lower level of viral load and depression score led to a higher probability for selecting LPV compared to the other two PI drugs ATV and DRV in both S_1 and S_2's optimal cARTs. For example, conditional on the viral load to be 4.4 and the depression score to be 5 at individual S_1's visit 12, the probability of including LPV in her optimal cART at her visit 13 was greater than 0.5. Since the selected cART FTC+ABC+EFV+LPV+RTV was able to control her viral load and depression score in the normal range, the optimal cARTs always included LPV for her next four visits. For individual S_2, the probability of including LPV at her visit 32 was less than 0.5, conditional on a viral load of 5 and a depression score of 16 at her visit 31. However, the selected cARTs for her visits 32-34 decreased S_2's depression score and viral load over time, which increased the probability of selecting LPV in her optimal cART to be greater than 0.5 at her visit 35.

In addition, note that our decision rule was built using a probabilistic generative model. Its stochastic nature informally accounts for exploration versus exploitation. If desired, one can always turn an optimized stochastic policy into a deterministic policy by e.g., taking the mode. On the other hand, compared to the deterministic policy, the stochastic policy assigns positive probabilities to several optimal or nearly-optimal deterministic policies, which provides physicians more flexibility in clinical practice.



(a) Viral load for S_1

(b) Depression for S_2

(c) Optimal cART assignments for S_1

(d) Optimal cART assignments for S_2

(e) Probability of selecting LPV for S_1

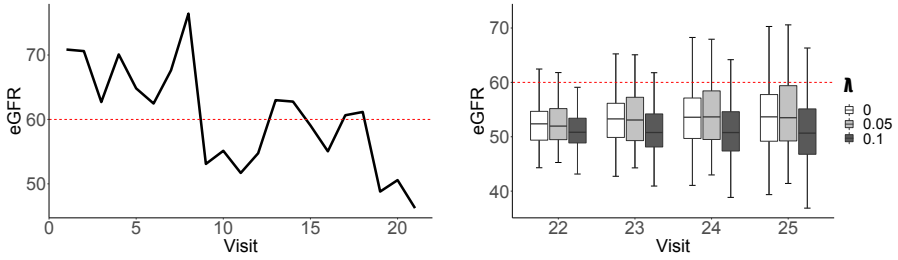
(f) Probability of selecting LPV for S_2

FIG 6. *Panels (a, b) plot the observed and predicted values of viral load and depression for two randomly selected individuals S_1 and S_2 in the WIHS data analysis. The predictions are made under the scenarios where both of them select their personalized optimal sequential cART assignments. The shaded areas represent the 95% predictive credible bands, and the dashed red lines represent the thresholds for viral load and depression. Panels (c, d) plot the personalized optimal sequential cART assignments for individual S_1 and S_2. Panels (e, f) plot the probabilities of selecting LPV in the personalized optimal sequential cART assignments for S_1 and S_2 conditional on their viral loads and depression scores. The white lines represent the contours for probability equals to 0.5.*

Similarly to the simulation study, we selected all individuals who were depressed but with normal viral load and kidney function at their last visits to illustrate the clinical utility of the proposed approach at the subpopulation level, resulting in 29 individuals. We set the reward weights $\boldsymbol{w}_i = (0.8, 0.1, 0.1)$ to stress the importance of reducing their depression scores. Under their personalized optimal sequential cART assignments, the average rewards (4.1) for this subpopulation increased from -65.7 during their last four visits to -51.3 during their future four visits. In addition, their average depression scores were reduced from 20.5 to 15.9, with 14 individuals becoming not depressed in the future two years.

To demonstrate the advantage of using uncertainty-penalized policy optimization, we investigated several different choices of hyperparameter $\lambda = 0, 0.05, 0.1$ for one randomly selected individual P_1 in the WIHS dataset who had a number of $J_1 = 21$ visits. Individual P_1 initially received a cART with triple NRTI therapies 3TC+D4T+TDF, which was then switched to a cART of two NRTIs FTC+TDF with a PI drug ATV boosted by RTV at her fifth visit, and finally replaced the two NRTIs with two new NRTIs 3TC+ABC since her eighth visit. We assigned a large reward weight on eGFR for individual P_1, i.e., $\boldsymbol{w}_1 = (0.1, 0.1, 0.8)$, as her renal function was abnormal for a long time, as shown in Figure 7(a). Figure 7(b) plots

the posterior predictive distributions of eGFR under the three estimated optimal sequences of cARTs (corresponding to three values of λ) for P_1 at her next four visits (i.e., from visit 22 to 25). Specifically, the sequence of cARTs FTC+ABC+EFV (two NRTIs + one NNRTI) was optimal under $\lambda = 0$, 3TC+ABC+EFV (two NRTIs + one NNRTI) was optimal under $\lambda = 0.05$, and 3TC+ABC+ATV+RTV (two NRTIs + one PI + one Booster) was optimal under $\lambda = 0.1$. Although none of these sequential cART assignments was able to improve P_1's eGFR to the normal range, all of them stabilized her renal function around the threshold. As shown in Figure 7(b), when λ increased, the predicted mean values of eGFR decreased but the corresponding predicted uncertainties also decreased. In the WIHS dataset, the cARTs 3TC+ABC+ATV+RTV and 3TC+ABC+EFV were recorded for 96 and 71 times, respectively, while the cART FTC+ABC+EFV was not recorded. The combination of 3TC and ABC is sold under the brand name Epzicom as one pill, making it more commonly prescribed than FTC+ABC in clinical practice since one pill usually lead to better adherence in PWH than two pills (Weisser et al., 2020). The proposed method can easily take into account the pill burden by adding a penalty term on the number of pills in the reward. In summary, there is a trade-off between exploring cARTs that are rarely or never used in the data with higher expected rewards and selecting commonly-prescribed cARTs with lower risks. We leave the final decision to HIV physicians.



(a) Observed eGFR for P_1 (b) Predicted eGFR for P_1

FIG 7. *Observed and predicted eGFR data for one randomly selected individual P_1 in the WIHS data analysis. The predictions are made under her personalized optimal sequential cART assignments with respect to different choices of $\lambda = 0, 0.05, 0.1$ at her next four visits. The dashed red lines represent the threshold for eGFR.*

7. Conclusion. To maximize long-term health outcomes for PWH, we developed a two-step Bayesian decision framework for optimizing personalized sequential cART assignments with proper uncertainty propagation. In the first step, we used an MGP model to characterize how individuals' longitudinal states evolve over time conditional on their historical states and treatment histories. In the second step, we designed an uncertainty-penalized policy optimization procedure to find the optimal sequential cART assignments. The uncertainty quantification in the first step was embedded in the decision framework by adding a penalty term to the reward function to help mitigate the distribution shift issue via a trade-off between the reward gain and risk for exploring new policies. Through simulation studies and the analysis of the WIHS dataset, we demonstrated that the proposed method has the potential to assist physicians' decisions on precision cART in PWH.

There are several potential extensions. First, for illustration purpose, we considered a personalized reward function depending on depression, viral load, and kidney function; other clinical factors such as cognition, BMI, and pill burden can also be incorporated into the decision framework. Second, the theoretical guarantee of the uncertainty-penalized policy optimization method was proved by Yu et al. (2020b) in a frequentist setup; it will be interesting to extend the theory to our Bayesian setup. Lastly, combination therapies are needed for

many complex diseases beyond HIV. The proposed method can be applied to such electronic health records datasets to learn the optimal treatment policies, potentially yielding better therapy management and improving the quality of life for people with chronic health conditions. For example, polypharmacy, the use of multiple drugs to treat different diseases and chronic health conditions at the same time, is a growing concern for older adults (Masnoon et al., 2017). The proposed approach can be used to optimize the combination of drugs for elders with multiple diseases in order to optimize their long-term health outcomes. Suppose that one or more drugs can be chosen from multiple possible drugs for treating each disease, then we can use the proposed three-level decision process as the decision model. Specifically, at the first level, we will determine whether we need to switch the individual's combination therapy to a new one or stay on the previous one. At the second level, we will determine how many drugs we need for treating each disease. At the third level, we will determine which specific drugs we will select for treating each disease.

Acknowledgment. This work was supported by NSF 1940107 to Dr. Xu, NSF 1918854 and NIH R01MH128085 to Drs. Xu and Rubin, and NSF 1918851 to Dr. Ni.

SUPPLEMENTARY MATERIAL

We provide all computational details, supplementary figures and tables, and additional simulation studies and WIHS data analysis in the supplementary material.

REFERENCES

ADIMORA, A. A., RAMIREZ, C., BENNING, L., GREENBLATT, R. M., KEMPF, M.-C. et al. (2018). Cohort profile: The women's interagency HIV study (WIHS). *International Journal of Epidemiology* **47** 393–394i.

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88** 669–679.

ALVAREZ, M. A., ROSASCO, L., LAWRENCE, N. D. et al. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning* **4** 195–266.

ARENAS-PINTO, A., THOMPSON, J., MUSORO, G., MUSANA, H., LUGEMWA, A. et al. (2016). Peripheral neuropathy in HIV patients in sub-Saharan Africa failing first-line therapy and the response to second-line ART in the EARNEST trial. *Journal of Neurovirology* **22** 104–113.

BECK, D., COHN, T., HARDMEIER, C. and SPECIA, L. (2015). Learning structural kernels for natural language processing. *Transactions of the Association for Computational Linguistics* **3** 461–473.

BLOOMFIELD, G. S., HOGAN, J. W., KETER, A., SANG, E., CARTER, E. J. et al. (2011). Hypertension and obesity as cardiovascular risk factors among HIV seropositive patients in Western Kenya. *PloS One* **6** e22288.

BOGOJESKA, J., BICKEL, S., ALTMANN, A. and LENGAUER, T. (2010). Dealing with sparse data in predicting outcomes of HIV combination therapies. *Bioinformatics* **26** 2085–2092.

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480.

CHECA, A., CASTILLO, A., CAMACHO, M., TAPIA, W., HERNANDEZ, I. and TERAN, E. (2020). Depression is associated with efavirenz-containing treatments in newly antiretroviral therapy initiated HIV patients in Ecuador. *AIDS Research and Therapy* **17** 1–5.

DIETRICH, L. G., THORBALL, C. W., RYOM, L., BURKHALTER, F., HASSE, B. et al. (2021). Rapid Progression of Kidney Dysfunction in People Living With HIV: Use of Polygenic and Data Collection on Adverse Events of Anti-HIV Drugs (D: A: D) Risk Scores. *The Journal of Infectious Diseases* **223** 2145–2153.

D'SOUZA, G., GOLUB, E. T. and GANGE, S. J. (2019). The changing science of HIV epidemiology in the United States. *American Journal of Epidemiology* **188** 2061–2068.

FUJIMOTO, S., MEGER, D. and PRECUP, D. (2019). Off-Policy Deep Reinforcement Learning without Exploration. In *Proceedings of the 36th International Conference on Machine Learning* (K. CHAUDHURI and R. SALAKHUTDINOV, eds.). *Proceedings of Machine Learning Research* **97** 2052–2062. PMLR.

GREENSMITH, E., BARTLETT, P. L. and BAXTER, J. (2004). Variance Reduction Techniques for Gradient Estimates in Reinforcement Learning. *Journal of Machine Learning Research* **5**.

HUA, W., MEI, H., ZOHAR, S., GIRAL, M. and XU, Y. (2021). Personalized Dynamic Treatment Regimes in Continuous Time: A Bayesian Approach for Optimizing Clinical Decisions with Timing. *Bayesian Analysis* **1** 1–30.

ISLAM, F. M., WU, J., JANSSON, J. and WILSON, D. P. (2012). Relative risk of renal disease among people living with HIV: a systematic review and meta-analysis. *BMC Public Health* **12** 1–15.

JIN, W., NI, Y., RUBIN, L. H., SPENCE, A. B. and XU, Y. (2022). A Bayesian nonparametric approach for inferring drug combination effects on mental health in people with HIV. *Biometrics* **78** 988–1000.

KENDALL, M. G. (1957). A course in multivariate analysis: London. *Charles Griffin & Co*.

LANGE, S., GABEL, T. and RIEDMILLER, M. (2012). Batch reinforcement learning. In *Reinforcement Learning* 45–73. Springer.

LANGEBEEK, N., KOOIJ, K. W., WIT, F. W., STOLTE, I. G., SPRANGERS, M. A. et al. (2017). Impact of comorbidity and ageing on health-related quality of life in HIV-positive and HIV-negative individuals. *AIDS* **31** 1471–1481.

LEDERGERBER, B., EGGER, M., OPRAVIL, M., TELENTI, A., HIRSCHEL, B. et al. (1999). Clinical progression and virological failure on highly active antiretroviral therapy in HIV-1 patients: a prospective cohort study. *The Lancet* **353** 863–868.

LUNDGREN, J. D., MOCROFT, A., GATELL, J. M., LEDERGERBER, B., MONFORTE, A. D., HERMANS, P. et al. (2002). A clinically prognostic scoring system for patients receiving highly active antiretroviral therapy: results from the EuroSIDA study. *The Journal of Infectious Diseases* **185** 178–187.

MA, Q., VAIDA, F., WONG, J., SANDERS, C. A., KAO, Y.-T. et al. (2016). Long-term efavirenz use is associated with worse neurocognitive functioning in HIV-infected patients. *Journal of Neurovirology* **22** 170–178.

MA, J., YANG, Q., HWANG, S.-J., FOX, C. S. and CHU, A. Y. (2017). Genetic risk score and risk of stage 3 chronic kidney disease. *BMC Nephrology* **18** 1–6.

MASNOON, N., SHAKIB, S., KALISCH-ELLETT, L. and CAUGHEY, G. E. (2017). What is polypharmacy? A systematic review of definitions. *BMC Geriatrics* **17** 1–10.

MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65** 331–355.

POLLACK, T. M., DUONG, H. T., PHAM, T. T., DO, C. D. and COLBY, D. (2017). Cigarette smoking is associated with high HIV viral load among adults presenting for antiretroviral therapy in Vietnam. *PLoS One* **12** e0173534.

RABOUD, J. M., MONTANER, J. S., CONWAY, B., RAE, S., REISS, P. et al. (1998). Suppression of plasma viral load below 20 copies/ml is required to achieve a long-term response to therapy. *AIDS* **12** 1619–1624.

RADLOFF, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement* **1** 385–401.

RIEDMILLER, M. (2005). Neural fitted Q iteration - first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning* 317–328. Springer.

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics* 400–407.

ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* **7** 1393–1512.

ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics* 189–326. Springer.

SUN, Y. and WANG, L. (2021). Stochastic Tree Search for Estimating Optimal Dynamic Treatment Regimes. *Journal of the American Statistical Association* **116** 421–432.

SURIAL, B., LEDERGERBER, B., CALMY, A., CAVASSINI, M., GÜNTHARD, H. F. et al. (2020). Changes in renal function after switching from TDF to TAF in HIV-infected individuals: a prospective cohort study. *The Journal of Infectious Diseases* **222** 637–645.

WALLENIUS, K. T. (1963). Biased sampling; the noncentral hypergeometric probability distribution Technical Report, Stanford University Applied Mathematics And Statistics Labs.

WANG, L., ROTNITZKY, A., LIN, X., MILLIKAN, R. E. and THALL, P. F. (2012). Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of the American Statistical Association* **107** 493–508.

WEISSER, B., PREDEL, H.-G., GILLESSEN, A., HACKE, C., VOR DEM ESCHE, J. et al. (2020). Single pill regimen leads to better adherence and clinical outcome in daily practice in patients suffering from hypertension and/or dyslipidemia: results of a meta-analysis. *High Blood Pressure & Cardiovascular Prevention* **27** 157.

WILLIAMS, D. W., LI, Y., DASTGHEYB, R., FITZGERALD, K. C., MAKI, P. M. et al. (2020). Associations between antiretroviral drugs on depressive symptomatology in homogenous subgroups of women with HIV. *Journal of Neuroimmune Pharmacology* 1–14.

XU, Y., MÜLLER, P., WAHED, A. S. and THALL, P. F. (2016). Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *Journal of the American Statistical Association* **111** 921–950.

YAZDANPANAH, Y., SISSOKO, D., EGGER, M., MOUTON, Y., ZWAHLEN, M. et al. (2004). Clinical efficacy of antiretroviral combination therapy based on protease inhibitors or non-nucleoside analogue reverse transcriptase inhibitors: indirect comparison of controlled trials. *BMJ* **328** 249.

YU, T., QUILLEN, D., HE, Z., JULIAN, R. et al. (2020a). Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning* 1094–1100. PMLR.

YU, T., THOMAS, G., YU, L., ERMON, S., ZOU, J. Y., LEVINE, S., FINN, C. and MA, T. (2020b). MOPO: Model-based offline policy optimization. *Advances in Neural Information Processing Systems* **33** 14129–14142.

ZENG, C., GUO, Y., HONG, Y. A., GENTZ, S., ZHANG, J. et al. (2019). Differential effects of unemployment on depression in people living with HIV/AIDS: a quantile regression approach. *AIDS Care*.

ZICH, J. M., ATTKISSON, C. C. and GREENFIELD, T. K. (1990). Screening for depression in primary care clinics: the CES-D and the BDI. *The International Journal of Psychiatry in Medicine* **20** 259–277.