Submitted to the Annals of Applied Statistics

BAGEL: A BAYESIAN GRAPHICAL MODEL FOR INFERRING DRUG EFFECT LONGITUDINALLY ON DEPRESSION IN PEOPLE WITH HIV

BY YULIANG LI[1], YANG NI[2], LEAH H. RUBIN[3], AMANDA B. SPENCE[4] AND YANXUN XU[1,*]

[1]*Department of Applied Mathematics and Statistics, Johns Hopkins University, yli193@jhu.edu;* [*]*yanxun.xu@jhu.edu*

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

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

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

Access and adherence to antiretroviral therapy (ART) has transformed the face of HIV infection from a fatal to a chronic disease. However, ART is also known for its side effects. Studies have reported that ART is associated with depressive symptomatology. Large-scale HIV clinical databases with individuals' longitudinal depression records, ART medications, and clinical characteristics offer researchers unprecedented opportunities to study the effects of ART drugs on depression over time. We develop BAGEL, a Bayesian graphical model to investigate longitudinal effects of ART drugs on a range of depressive symptoms while adjusting for participants' demographic, behavior, and clinical characteristics, and taking into account the heterogeneous population through a Bayesian nonparametric prior. We evaluate BAGEL through simulation studies. Application to a dataset from the Women's Interagency HIV Study yields interpretable and clinically useful results. BAGEL not only can improve our understanding of ART drugs effects on disparate depression symptoms, but also has clinical utility in guiding informed and effective treatment selection to facilitate precision medicine in HIV.

1. Introduction. Antiretroviral therapy (ART) has transformed HIV into a manageable, chronic illness despite its known central nervous system (CNS) side effects, which can complicate disease management. Some of the most commonly reported ART-related CNS adverse effects include depression and anxiety, suicidal ideation, developmental disorders, and neurological toxicities (Nanni et al., 2015; Zash et al., 2018). The presence of these CNS symptoms is a major public health concern as they are associated with ART discontinuation and increases in the likelihood of HIV transmission. Therefore, a major component of care for people living with HIV includes management and prevention of the long-term adverse effects of ART. In this paper, we focus on ART-related effects on depressive symptomatology. Depression is one of the leading mental health comorbidities in people with HIV (Bengtson et al., 2016), and is associated with poor ART adherence, HIV disease progression, and engagement in risky taking behaviors (Chattopadhyay et al., 2017; Ironson et al., 2017; Brickman et al., 2017). The high prevalence and harmful effects of depression among people with HIV highlight the need of effective clinical management and adequate treatment for depression.

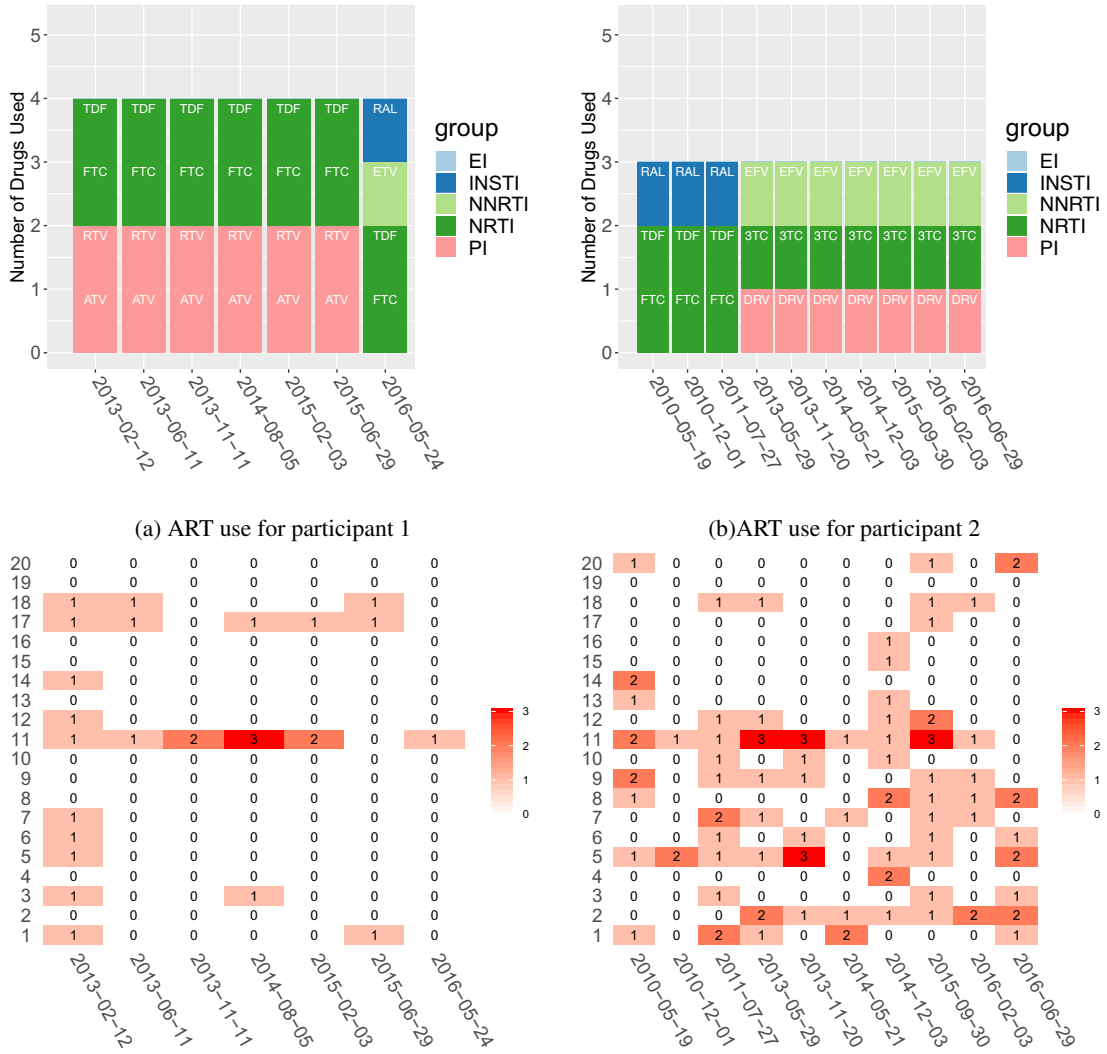
At present, there are more than 30 U.S. Food and Drug Administration-approved ART drugs to treat people with HIV or at risk for HIV. Common ART drugs fall into five drug classes, including nucleotide reverse transcriptase inhibitor (NRTI), non-nucleotide reverse transcriptase inhibitor (NNRTI), protease inhibitor (PI), entry inhibitor (EI), and integrase

Keywords and phrases: Bayesian nonparametrics, Depression, Graphical model, Longitudinal cohort study, Precision medicine.

inhibitor (INSTI). Although ART may alleviate depressive symptoms through suppressing viral load and improving physical health for people with HIV, ART can also exacerbate depressive symptoms through several possible mechanisms, including direct effects on neuronal and mitochondrial functions, astrocyte metabolism, and interference with neurotransmitters (Underwood et al., 2015; Shah et al., 2016; Cohen et al., 2017). The ART drugs most commonly associated with depression include efavirenz (EFV; NNRTI) and dolutegravir (DTG; INSTI) (Bengtson et al., 2017; Elzi et al., 2017; Borghetti et al., 2017). However, little is known about the effects of most ART drugs and ART combinations on depressive symptoms. Furthermore, ART effects may be confounded by other factors such as sociodemographic, clinical, and behavioral characteristics. Therefore, investigating ART effects on depressive symptoms to optimize health outcomes and facilitate personalized medicine remains a major challenge in HIV studies.

The Women's Interagency HIV Study (WIHS) is a large prospective, observational, longitudinal, multicenter study designed to investigate the progression of HIV disease in women with HIV or at risk for HIV infection in the United States (Bacon et al., 2005). At semiannual visits, physical examinations, laboratory testing, and questionnaires regarding sociodemographics, medication use, mental health (including depression), and self-reported clinical diagnoses are performed and data is registered in local and national WIHS repositories. Such electronic health records provide us unprecedented opportunities to examine the longitudinal effects of ART drugs on depression after adjusting for socio-demographic, behavioral, and clinical factors. Figure 1(a, b) present two womens' ART medication history at each of their study visits (shown as calendar dates). They were followed for different time periods with distinct visit dates and drug usages. At each visit, participants' depressive symptoms were also measured and recorded using the Center for Epidemiologic Studies Depression Scale (CES-D), a 20-item self-administered questionnaire (Lewinsohn et al., 1997); see Figure 1(c, d) as an example for the same two participants. Each question regarding depressive symptoms has an ordinal score in $\{0, 1, 2, 3\}$, with 0 being the least severe and 3 being the most severe. The complexity of data including longitudinal laboratory observations, heterogeneous participant population, and dynamic ART assignments, presents analytic and modeling challenges.

Prior studies examining ART-related effects on depressive symptoms in people with HIV have used a number of self-report instruments to assess depressive symptoms including the 15-item depression subscale of the Hopkins Symptom Checklist (D- HSCL) (Derogatis et al., 1974), the 9-item Patient Health Questionnaire (PHQ-9) (Kroenke et al., 2001), and the CES-D (Lewinsohn et al., 1997). However, these studies have several limitations. First, instead of considering each individual depressive symptom, these studies use a sum-score to represent severity, and a somewhat arbitrary threshold to distinguish individuals with depression from healthy ones. For example, with the CES-D, people with a total score higher than 16 are regarded as having moderate or severe depression. However, sum-score is an over-simplified measurement for highly heterogeneous diseases like depression. Two participants with the same score may share no common symptoms or report different symptoms being severe. Moreover, a sum-score assumes equal contributions of all symptoms which are inconsistent with psychometric literature (Fried et al., 2016). Therefore, lumping distinct symptoms to a single sum-score and grouping participants with similar sum-scores but different symptoms into one category may result in significant information loss. We will learn all depression items simultaneously, which is conceptually related to the multi-task learning frameworks in the machine learning literature (Liu et al., 2019). Second, as shown in Figure 1, data in HIV studies is often longitudinal with changing drug assignments and depression scores over the course of treatment, and the drug effects on depression may change over time and depend on individuals' unique characteristics. Existing methods only examined the cross-sectional associations between ART drugs and depression. In other application domains, methods for



(a) ART use for participant 1

(b)ART use for participant 2

(a) Depression score for participant 1

(b) Depression score for participant 2

FIG 1. *(a, b) The number of ART drugs taken by two randomly selected participants from the WIHS versus their visit dates, respectively. Drug names are recorded and different colors represent different drug groups. (c, d) Their depression scores of 20 questions across visits.*

predicting drug responses from longitudinal observation data have been developed. For example, Dey et al. (2019) incorporated a patient similarity graph as network regularization in a linear system to model individual responses of each patient and solved it as a convex optimization problem. However, it does not directly model how drug effects are modified by covariates over time. Therefore, there is a critical need to develop novel statistical models that are capable of comprehensively examining non-stationary longitudinal effects of all ART drugs on disparate depression symptoms in a heterogeneous HIV population.

To this end, we develop BAGEL, a Bayesian graphical model that can simultaneously study personalized effects of ART drugs on item-level depression symptoms while adjusting for participants' longitudinal covariates, understand the participant-dependent dynamic drug effects, and take into account the heterogeneous HIV population by clustering participants into subgroups through a Bayesian nonparametric prior. Specifically, we use a latent bipartite

graph with one group of vertices representing ART drugs and the other group of vertices representing depressive symptoms. The presence of an edge between a drug and a depressive symptom indicates a significant drug-depression effect, the size of which is represented by the edge weight. Importantly, the effect sizes can vary across different clinical visits and different participants. Moreover, for broader dissemination and reproducibility, the code implementing BAGEL is available at https://github.com/YanxunXu/BAGEL.

The rest of paper is organized as follows. In Section 2 we present modeling details of BAGEL and the posterior inference. The performance of BAGEL is evaluated through simulation studies with comparison to alternative methods in Section 3. In Section 4, we apply BAGEL to a dataset from WIHS to study the longitudinal ART drugs effects on depression items and demonstrate the clinical utility of BAGEL. We conclude with a brief discussion in Section 5.

2. Model and Inference.

2.1. *Probability model.* Let $U_{ijq} \in \{0,1,2,3\}$ be the ordinal score of depression item q for participant i at visit j, $i = 1,\ldots,n$, $j = 1,\ldots,J_i$, and $q = 1,\ldots,Q$. In WIHS, we have $Q = 20$ depression items, which measure different depressive symptoms such as "bothered by things", "appetite was poor", and "fearful." A detailed description of all 20 symptoms is provided in Section 4. A higher score indicates a worse depressive symptom. Following Albert and Chib (1993), we introduce a latent continuous random variable Y_{ijq} such that $U_{ijq} = k$ if and only if $Y_{ijq} \in (a_k, a_{k+1}]$, where $k = 0,\ldots,3$, $a_0 = -\infty$, and $a_4 = \infty$. For identifiability, we set $a_1 = 0$. Let Z_{ijd} be a binary indicator to represent whether participant i at visit j uses drug d, $d = 1,\ldots,D$, and denote $\boldsymbol{Z}_{ij} = (Z_{ij1},\ldots,Z_{ijD})$. Let $\boldsymbol{X}_{ij}$ be an S-dimensional row vector including an intercept, time-invariant covariates (e.g., race), and time-varying covariates (e.g., BMI, CD4 count).

We regress the multivariate latent depression scores $\boldsymbol{Y}_{ij} = (Y_{ij1},\ldots,Y_{ijQ})$ on covariates $\boldsymbol{X}_{ij}$ and drug usage $\boldsymbol{Z}_{ij}$,

$$\boldsymbol{Y}_{ij} = \underbrace{\boldsymbol{X}_{ij}\boldsymbol{\beta}_i}_{\text{covariate effects}} + \underbrace{\boldsymbol{Z}_{ij}\boldsymbol{B}_{ij}}_{\text{drug effects}} + \underbrace{\boldsymbol{\omega}_{ij}}_{\text{correlation}} + \boldsymbol{\epsilon}_{ij}, \quad i = 1,\ldots n, \quad j = 1,\ldots,J_i, \tag{2.1}$$

with $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1},\ldots,\epsilon_{ijQ})$. The first term $\boldsymbol{X}_{ij}\boldsymbol{\beta}_i$ captures the dependence of the latent outcome $\boldsymbol{Y}_{ij}$ on the covariates $\boldsymbol{X}_{ij}$, where $\boldsymbol{\beta}_i$ is an $S \times Q$ matrix representing the covariate effects for participant i. The second term $\boldsymbol{Z}_{ij}\boldsymbol{B}_{ij}$ models the drug effects where $\boldsymbol{B}_{ij}$ is a $D \times Q$ matrix with each element $B_{ij,dq}$ being the contribution of drug d to depression item q for participant i at visit j; the modeling details of drug effects $\boldsymbol{B}_{ij}$ will be introduced later. By study design, depression items are correlated. For example, in our data, the pairwise rank correlations of observed depression ordinal scores range from 0.11 to 0.66. To capture such dependencies, we assume $\boldsymbol{\omega}_{ij} \in \mathbb{R}^Q$ follows a centered multivariate normal distribution $\mathrm{MN}(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{C}_\omega)$ with correlation matrix $\boldsymbol{C}_\omega$. The sampling model is completed by assuming independent normal errors $\boldsymbol{\epsilon}_{ij} \sim \mathrm{MN}(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I})$, where $\sigma_\epsilon^2 = 1$ for identifiability.

Modeling drug effects. The main contribution of the proposed model is the formulation of the drug effects $\boldsymbol{B}_{ij}$, which are the key parameters of interest. Since there are 20 depression items and more than 20 drugs, to encourage parsimony, we assume $\boldsymbol{B}_{ij}$ to be a sparse matrix, i.e., not all drugs are associated with every depression item. Letting $\circ$ denote the Hadamard product of two matrices of the same dimension, we decompose $\boldsymbol{B}_{ij} = \boldsymbol{R}_i \circ \boldsymbol{\Lambda}_{ij}$ multiplicatively into two components. The first component $\boldsymbol{R}_i = (R_{i,dq})$ is a $D \times Q$ binary matrix that induces sparsity in $\boldsymbol{B}_{ij}$ with $R_{i,dq} = 1$ if drug d is significantly associated with depression item q for participant i and $R_{i,dq} = 0$ otherwise. The matrix $\boldsymbol{R}_i$ can be also viewed as a directed acyclic graph adjacency matrix of a bipartite network consisting of two sets of nodes:

drugs and depression items; see Figure 2 for example. The second component $\mathbf{\Lambda}_{ij}$ is a $D \times Q$ matrix that quantifies the strength of the (non-zero) drug-depression associations over time with each element $\Lambda_{ij,dq}$ being the effect of drug d on depression item q for participant i at visit j. Note that although the presence or absence of the drug-depression links $\boldsymbol{R}_i$ for participant i do not change over time, we allow the strength of the links $\mathbf{\Lambda}_{ij}$ to vary with time and covariates through the following model,

$$\Lambda_{ij,dq} = \widetilde{\boldsymbol{X}}_{ij}\boldsymbol{\alpha}_{i,dq} + s(t_{ij}), \tag{2.2}$$

where $\widetilde{\boldsymbol{X}}_{ij}\boldsymbol{\alpha}_{i,dq}$ accounts for the dependence of the drug effect on participants' covariates $\widetilde{\boldsymbol{X}}_{ij}$, t_{ij} denotes the time of visit j since the first visit for participant i, and $s(\cdot)$ is a nonlinear smooth function. In general, $\widetilde{\boldsymbol{X}}_{ij}$ could be different from $\boldsymbol{X}_{ij}$ but we assume $\widetilde{\boldsymbol{X}}_{ij} = \boldsymbol{X}_{ij}$ hereafter. In summary, (2.2) allows the non-zero drug effects on depression to vary with participant-specific covariates and visit times. An alternative interpretation[1] of (2.2) is that it models the interaction effects of drug usage $\boldsymbol{Z}_{ij}$, and participant-specific covariates and visit times on depression.

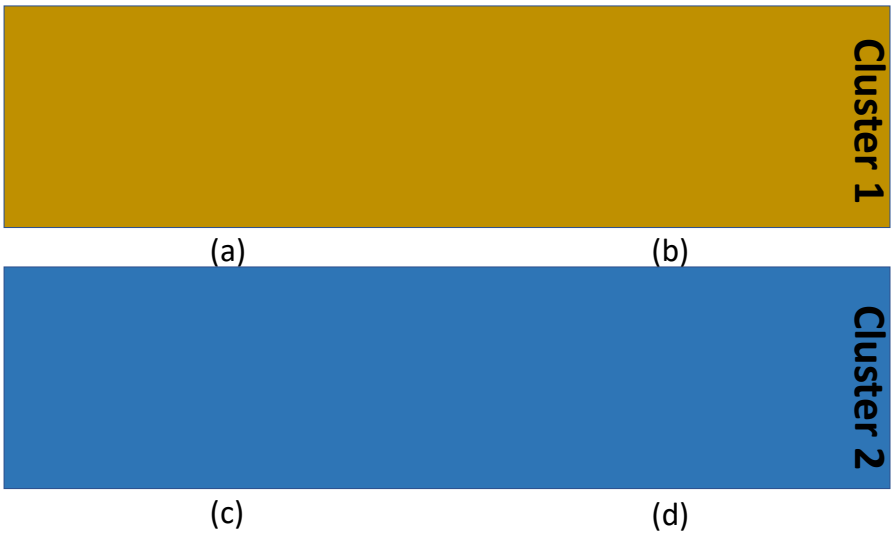


FIG 2. *Drug-depression relationships across visits for four participants (a)-(d) in two different clusters. Different colors encode different visits for one participant, and the width of each edge represents the weight of the association between the corresponding drug (circles) and depression (squares).*

We use splines to model the nonlinear function $s(\cdot)$ by letting $s(t_{ij}) = \tilde{\boldsymbol{t}}_{ij}\boldsymbol{\gamma}_{i,dq}$, where $\tilde{\boldsymbol{t}}_{ij} = (\tilde{t}_{ij1}, \ldots, \tilde{t}_{ijB})$ denotes the cubic B-spline basis expansion of t_{ij} and B is the number of bases. To prevent overfitting, we leverage the P-spline (Eilers and Marx, 1996) that penalizes finite differences of adjacent B-spline coefficients to encourage smoothness. As shown in Lang and Brezger (2004), P-splines can be written as a Bayesian hierarchical model by assuming $\boldsymbol{\gamma}_{i,dq} \sim \mathrm{MN}(0, \sigma^2_\gamma \boldsymbol{K}^-)$, where $\boldsymbol{K}$ is a singular penalty matrix constructed from the second-order differences of the adjacent spline coefficients and $\boldsymbol{K}^-$ is the pseudo-inverse of the matrix $\boldsymbol{K}$. In summary, the term $\boldsymbol{Z}_{ij}\boldsymbol{B}_{ij}$ represents (i) the drug main effect, (ii) drug-covariate interaction, and (iii) nonlinear drug-time interaction.

[1] It can be seen by plugging (2.2) and $\boldsymbol{B}_{ij} = \boldsymbol{R}_i \circ \mathbf{\Lambda}_{ij}$ into (2.1).

2.2. *Priors.* The proposed model is parameterized by $(\{\Theta_i\}_{i=1}^n, \boldsymbol{C}_\omega)$, where $\Theta_i = (\boldsymbol{\beta}_i, \boldsymbol{R}_i, \{\boldsymbol{\alpha}_{i,dq}, \boldsymbol{\gamma}_{i,dq}\}_{d=1,q=1}^{D,Q})$. Although one can apply model (2.1) independently to one participant at a time, it is deemed inefficient and hard to interpret, especially for individuals with few visits. In the other extreme, we could have assumed all effects are the same across all participants, i.e., $\Theta_i = \Theta$ for all i, which however overlooks the heterogeneity of the population. We take a comprise between these two extremes by proposing a joint modeling approach through a Bayesian nonparametric (BNP) prior on Θ_i's. The BNP prior borrows information across "similar" participants for estimating personalized drug-depression relationships over time. Particularly, we consider a Dirichlet Process (DP) prior (Ferguson, 1974), $\Theta_i \overset{\text{i.i.d.}}{\sim} G$ and $G \sim DP(m_0, G_0)$ with a concentration parameter m_0 and a baseline measure $G_0(\cdot)$. DP has been widely used in various biomedical applications (Müller and Quintana, 2004; Xu et al., 2016) to account for sampling heterogeneity. Sampling from a DP prior can generate the same value of Θ_i for different participants due to its almost surely discreteness, which elegantly defines a natural partition of participants without a pre-specified number of clusters — i and i' belong to the same cluster if $\Theta_i = \Theta_{i'}$. In other words, on the one hand, for participants that are deemed similar (i.e., if they belong to the same cluster), their covariate and drug effects on depression are identical and hence estimated jointly. On the other hand, the estimation would be done independently for dissimilar participants that do not belong to the same clusters. Because of the presence of ties, let $\{\tilde{\Theta}_h\}_{h=1}^H$ with $\tilde{\Theta}_h = (\tilde{\boldsymbol{\beta}}_h, \tilde{\boldsymbol{R}}_h, \{\tilde{\boldsymbol{\alpha}}_{h,dq}, \tilde{\boldsymbol{\gamma}}_{h,dq}\}_{d=1,q=1}^{D,Q})$ denote the unique values of $\{\Theta_i\}_{i=1}^n$. Let e_i denote the clustering membership indicator such that $e_i = h$ indicates that participant i belongs to cluster h. Then $\Theta_i = \tilde{\Theta}_h$ if $e_i = h$. Given e_i, the DP prior induces the following prior on Θ_i,

$$\Theta_i = \sum_{h=1}^{H} \tilde{\Theta}_h I(e_i = h) \text{ and } \tilde{\Theta}_h \sim G_0.$$

In essence, the DP prior allows for clustering of participants by grouping participant-specific parameters $\{\Theta_i\}$ into $\{\tilde{\Theta}_h\}$ and makes inference for cluster-specific parameters $\tilde{\Theta}_h$ using participants within cluster h. Figure 2(a-d) illustrates the drug-depression relationships for four participants in two different clusters with three visits recorded for participants (a) and (c) and two visits recorded for participants (b) and (d). Participants belonging to the same cluster share the same graph structure (encoded in $\boldsymbol{R}_i$), although the edge weights can differ across participants and visits (quantified by $\boldsymbol{B}_{ij}$). Furthermore, the BNP prior on Θ_i yields a flexible nonparametric response surface for $\boldsymbol{Y}_{ij}$ capturing complex drug effect dynamics.

Let $\tilde{\boldsymbol{\beta}}_{hq}$ be the q-th column of $\tilde{\boldsymbol{\beta}}_h$. We assume a conjugate base measure G_0 of DP, $G_0(\tilde{\Theta}_h) = \prod_{q=1}^{Q} p(\tilde{\boldsymbol{\beta}}_{hq}) \prod_{d=1,q=1}^{D,Q} p(\tilde{R}_{h,dq}) p(\tilde{\boldsymbol{\alpha}}_{h,dq}) p(\tilde{\boldsymbol{\gamma}}_{h,dq})$ where $p(\tilde{\boldsymbol{\beta}}_{hq}) = \text{MN}(0, \sigma_\beta^2 \boldsymbol{I})$, $p(\tilde{\boldsymbol{\alpha}}_{h,dq}) = \text{MN}(0, \sigma_\alpha^2 \boldsymbol{I})$, $p(\tilde{\boldsymbol{\gamma}}_{h,dq}) = \text{MN}(0, \sigma_\gamma^2 \boldsymbol{K}^-)$, and $p(\tilde{R}_{h,dq}) = \text{Bernoulli}(\rho)$. We consider a non-informative Unif$(-1, 1)$ prior for the off-diagonal elements of $\boldsymbol{C}_\omega$, although alternative priors could be adopted if certain prior knowledge on the correlation structure of depression items is available. We complete the model construction by assigning the following conjugate hyperpriors: $\sigma_\beta^2 \sim$ inv-Gamma(a_β, b_β), $\sigma_\alpha^2 \sim$ inv-Gamma(a_α, b_α), $\sigma_\gamma^2 \sim$ inv-Gamma(a_γ, b_γ), and $\rho \sim$ Beta$(\alpha_\rho, \beta_\rho)$. The posterior inference is carried out using Markov chain Monte Carlo (MCMC) simulations. Details of the sampling procedure are described in Section A of the Supplementary Material (Li et al., 2021).

3. Simulation Study. In this section, we evaluated the performance of BAGEL through simulation studies by comparing posterior inference with simulation truth. Furthermore, to demonstrate the advantages of (i) imposing a DP prior to account for participants' heterogeneity and (ii) encouraging sparsity in the drug-depression effects $\boldsymbol{B}_{ij}$, we compared

BAGEL to two alternative methods. The first method does not take into account participants' heterogeneity by assuming that all participants shared the same parameter, i.e., $\Theta_i = \Theta$ and $\Theta \sim G_0$. We call this method Homogeneous. The second method, called NoSparsity, dropped the sparse binary matrix $\boldsymbol{R}_i$ from BAGEL, i.e., assuming $\boldsymbol{B}_{ij} = \boldsymbol{\Lambda}_{ij}$.

3.1. *Simulation setup.* We considered two scenarios, one with $n = 200$ participants and the other with $n = 500$ participants. In both scenarios, we assumed that there were $D = 5$ drugs and $Q = 3$ depression items, with the true number of clusters being $H_0 = 3$. All participants were randomized to the three clusters with equal probabilities. For each participant i, we generated the number of visits J_i from a Poisson distribution with mean 10. The duration between two consecutive visits (i.e., $t_{i,j+1} - t_{ij}$) was generated from a normal distribution with the mean 1 year and standard deviation 0.2. Let $\mathrm{N}(\mu, \sigma^2)$ denote a normal distribution with mean μ and variance σ^2. For participants' covariates $\boldsymbol{X}_{ij}$, we considered two time-invariant covariates with one being generated from $\mathrm{Bernoulli}(0.5)$ and the other being generated from $\mathrm{N}(0, 4)$, and three time-variant covariates mimicking age, body mass index (BMI), and smoking status. We randomly sampled participants' ages at their initial visits from the motivating WIHS dataset with replacement, then computed their ages at followup visits based on the generated t_{ij}'s. The BMI and smoking status were generated from $\mathrm{N}(28, 8)$ and $\mathrm{Bernoulli}(0.6)$ independently across participants and their visits, respectively. The drug usage Z_{ijd} for participant i at visit j was generated from $\mathrm{Bernoulli}(0.8)$ independently.

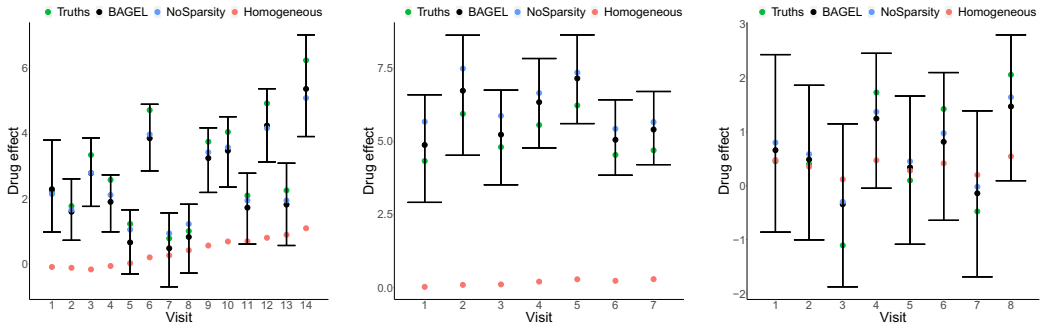
Conditional on clustering memberships, the simulated true values of $\boldsymbol{\beta}_i$'s and $\boldsymbol{\alpha}_i$'s in both scenarios are tabulated in Supplementary Tables T1 and T2. For the matrix $\boldsymbol{R}_i$ that induces sparsity in drug-depression effects, we generated each element $R_{i,dq}$ independently from $\mathrm{Bernoulli}(0.4)$. The simulated true $s(t_{ij})$ was assumed to be $s(t_{ij}) = 6t_{ij}$. Setting $\boldsymbol{C}_\omega$ to be $\begin{bmatrix} 1 & 0.3 & 0.35 \\ 0.3 & 1 & 0.4 \\ 0.35 & 0.4 & 1 \end{bmatrix}$ and the threshold to be $(a_1, a_2, a_3) = (0, 8, 16)$, we generated the latent continuous Y_{ijq} from (2.1) as well as the corresponding ordinal depression scores U_{ijq}, $i = 1, \ldots, n; j = 1, \ldots, J_i; q = 1, \ldots, Q$.

We applied BAGEL to the simulated dataset with 100 repeated simulations in both scenarios. The hyperparameters were set to be $m_0 = 1, a_\beta = a_\alpha = a_\gamma = 3, b_\beta = b_\alpha = b_\gamma = 10, \alpha_\rho = 1$, and $\beta_\rho = 10$. In each analysis, we ran 20,000 MCMC iterations with an initial burn-in of 10,000 iterations, thinned by 10. It took $\sim$20 minutes for each analysis using a personal laptop equipped with Processor 2GHz Intel Core i5, Memory 8GB. We assessed the convergence using Geweke's diagnostic in R package *coda* (Plummer et al., 2006), showing no sign of lack of convergence. Additional simulation studies with larger dimensions are provided in Section D of the Supplementary Material, which show similar results.

3.2. *Simulation results.* We first evaluated the clustering performance of BAGEL. BAGEL successfully identified $\hat{H} = 3$ clusters in both scenarios, since the marginal posterior probabilities of the simulated true number of clusters averaged over 100 repeated simulations were $p(H = 3 \mid \mathrm{data}) = 99.6\%$ when $n = 200$, and $p(H = 3 \mid \mathrm{data}) = 89.1\%$ when $n = 500$. We calculated the posterior co-clustering probabilities of participants based on the empirical proportion of participants being clustered together over the post-burn-in MCMC samples. The simulated true clustering schemes and the co-clustering probability matrices for one randomly selected simulation under both scenarios are depicted in Supplementary Figure F1, showing that BAGEL assigns participants to their simulated true clusters with high probabilities. The adjusted rand indices corresponding to Supplementary Figure F1 are 0.995 and 1 for $n = 200$ and $n = 500$, respectively.

Reporting cluster-specific parameters is challenging due to the issue of label switching. Following Dahl (2006), let e_i denote the clustering membership indicator of participant i and $\mathbf{V}$ be an $n \times n$ matrix whose entry $V_{i_1,i_2} = \mathbb{P}(e_{i_1} = e_{i_2})$ represents the posterior probability of participants i_1 and i_2 being clustered together, which can be computed based on posterior samples. In each MCMC iteration, we computed an $n \times n$ binary matrix $\mathbf{V}^{\boldsymbol{e}}$, whose entry $V^{\boldsymbol{e}}_{i_1,i_2} = \mathbb{I}(e_{i_1} = e_{i_2})$ indicates whether participants i_1 and i_2 are clustered together at that iteration. The clustering that minimizes the Frobenius distance between $\mathbf{V}^{\boldsymbol{e}}$ and $\mathbf{V}$, given by $\hat{\boldsymbol{e}} = \arg\min_{\boldsymbol{e}} \|\mathbf{V}^{\boldsymbol{e}} - \mathbf{V}\|$, is called the least-square summary of clustering. Then we relabeled the cluster memberships at each MCMC iteration by minimizing its Frobenius distance to the least-square summary as a post-processing step (Li et al., 2020). Supplementary Figures F2 and F3 plot posterior means and 95% credible intervals (CIs) for $\tilde{\boldsymbol{\beta}}_{hq}$'s and $\tilde{\boldsymbol{\alpha}}_{h,dq}$'s averaged over 100 repeated simulations, respectively, showing that all 95% CIs are centered around the simulated true values. To assess the ability of BAGEL in recovering the sparsity $\boldsymbol{R}_i$ of drug effects, we plotted the posterior probabilities of $\tilde{R}_{h,dq}$ being equal to the simulated true values in Supplementary Figure F4 under both scenarios, indicating satisfactory performance. We also computed the true positive rates (TPR) and false discovery rates (FDR) for recovering $\boldsymbol{R}_i$ under both scenarios: when $n = 200$, TPR=0.96, FDR=0.08; when $n = 500$, TPR=0.98, FDR=0.02, indicating good recovery.

Next, we examine whether BAGEL can recover individualized longitudinal drug effects $\boldsymbol{B}_{ij}$, $j = 1, \ldots, J_i$. We randomly selected one participant from each of the three clusters in one simulated dataset when $n = 200$. These participants had 14, 7, and 8 visits, respectively. As drug effects for one participant are different across different drugs and depression items, we computed the posterior summaries of $B_{ij,dq}$ for the randomly selected d and q such that $R_{i,dq} \neq 0$ for illustration (note that when $R_{i,dq} = 0$, the corresponding $B_{ij,dq} = 0$). Figure 3 plots the estimated posterior means of $B_{ij,dq}$'s with 95% CIs for these three participants, indicating that BAGEL can successfully recover the simulated true drug effects.



(a) participant in cluster 1 (b) participant in cluster 2 (c) participant in cluster 3

FIG 3. *Simulated true values, posterior means with 95% CIs of the estimated $B_{ij,dq}$ under BAGEL, and posterior means under two alternative methods: NoSparsity and Homogeneous for three participants randomly selected from each of the three clusters, respectively. The green dots represent simulation truths; the black dots represent posterior estimates under BAGEL; the blue and pink dots represent posterior estimates under NoSparsity and Homogeneous, respectively.*

Furthermore, we compared BAGEL with two alternative methods: Homogeneous and NoSparsity. Figure 3 shows the estimated posterior means of $B_{ij,dq}$ under the two alternatives. Both BAGEL and NoSparsity estimated the drug effects much better than the Homogeneous method, especially for participants from clusters 1 and 2 partly because the drug effects may be averaged out in a heterogeneous population and hence greatly biased towards zero. We also computed the mean squared errors $\mathrm{MSE}^{(b)}_{dq}$ of the estimated

$B_{ij,dq}$ averaged over all visits of all participants for each drug and depression item, i.e., $\frac{1}{\sum_{j=1}^{J_i}}\sum_{i=1}^{n}\sum_{j=1}^{J_i}(\hat{B}_{ij,dq}^{(b)} - B_{ij,dq}^{o})^2$ under each method, where $\hat{B}_{ij,dq}^{(b)}$ is the estimated $B_{ij,dq}$ in MCMC iteration b, $B_{ij,dq}^{o}$ is the simulated true value. Then we reported the mean and standard deviation of $\mathrm{MSE}_{dq}^{(b)}$'s for each drug and depression item across all post-burn-in iterations under the three methods in Supplementary Table T6. Both BAGEL and NoSparsity yielded much smaller means of $\mathrm{MSE}_{dq}^{(b)}$'s across drugs and depression items compared to the Homogeneous. BAGEL slightly outperformed NoSparsity with substantially better interpretability due to sparsity. In addition, as shown in Supplementary Figures F5 and F6, the CIs for the model parameters of NoSparsity tend to be wider than those of BAGEL. We also compared the three methods in terms of Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010; Gelman et al., 2014). BAGEL yielded the smallest WAIC, which was 3688.73, while the WAIC values of NoSparsity and Homogeneous were 3816.89 and 9602.72, respectively.

4. The WIHS Data Analysis.

4.1. *Data.* The WIHS is a multi-center, longitudinal study developed to characterize the natural and treated history of HIV infection and clinical outcomes in women residing in the United States (Barkan et al., 1998; Bacon et al., 2005; Adimora et al., 2018). WIHS participants complete "core" visits approximately every 6 months. At each visit, women undergo clinical examination, structured medical and psychosocial interviews, and laboratory testing to assess HIV status/viral load. For the present analysis, we included all visits after January 2012 from women in the Washington, D.C. WIHS site resulting in a total of 214 participants with 1,240 visits. The primary outcome of interest was the item-level responses on CES-D, a 20-item self-administered questionnaire, which is commonly used to assess depression in HIV studies (Moore et al., 1999; Ickovics et al., 2001; Maki et al., 2012). The CES-D measures the frequency of depressive symptoms (e.g., "people dislike me") during the week prior to the visit, where "0" indicates no symptom or the duration of symptom is less than one day, "1" indicates the duration of symptom is between one and two days, "2" indicates the duration of symptom is between three and four days, and "3" indicates the duration of symptom is longer than five days. Figure 5 lists all 20 symptoms, which can be categorized in three types: somatic, affect, and interpersonal symptoms. Somatic symptoms include unpleasant or worrisome mood such as "restless", "appetite", and "concentration." Affect symptoms include the lack of positive affect (e.g., "enjoyed life", "hopeful of future") and/or the presence of negative affect ("fearful", "lonely", "failure"). Interpersonal symptoms include"people disliked me" and "people unfriendly". The items reflecting positive affect were reversed scored so that higher values on each item reflected more negative symptoms.

We included the following sociodemographic, behavioral, and clinical covariates as risk factors (Cook et al., 2002; Rubin et al., 2011; Maki et al., 2012) for depressive symptoms: age, body mass index (BMI), CD4 count, nadir CD4 count (CD4 Nadir), viral load (VLOAD), race, smoking, substance abuse (e.g., marijuana, cocaine, and heroin), and education. Table 1 summarizes these characteristics of participants at their first visits in the dataset. There were 23 drugs used in the dataset, falling into five drug classes: NRTI, NNRTI, PI, EI, and INSTI. The drug use frequency varies significantly among different drugs from hundreds of visits to only few visits, the details of which are reported in Supplementary Table T3.

4.2. *Results.* We applied BAGEL to the dataset with the same hyperparameters as in the simulation study. The least-square summary of clustering estimated four clusters with the number of participants in each cluster being 94, 105, 6, 9, respectively. Table 1 summarizes

Variables	Overall (n=199) n(%)	Cluster 1 (n=94) n(%)	2 (n=105) n(%)	3 (n=6) n(%)	4 (n=9) n(%)
Demographics					
Age (years)					
≤ 35	10 (5)	4 (4)	6 (6)	0 (0)	0 (0)
36-45	66 (33)	28 (30)	38 (36)	3 (50)	5 (56)
46-55	80 (40)	40 (43)	40 (38)	2 (33)	3 (33)
> 55	43 (22)	22 (23)	21 (20)	1 (17)	1 (11)
BMI					
< 18.5	6 (3)	3 (3)	3 (3)	0 (0)	0 (0)
18.5-29.9	107 (54)	50 (53)	57 (54)	5 (83)	4 (44)
30-39.9	59 (30)	25 (27)	34 (32)	1 (17)	4 (44)
≥ 40	27 (13)	16 (17)	11 (11)	0 (0)	1 (12)
Race					
White	35 (18)	12 (13)	24 (23)	0 (0)	0 (0)
Black	150 (75)	76 (81)	74 (70)	6 (100)	8 (89)
Others	13 (7)	6 (6)	7 (7)	0 (0)	1 (11)
HIV-related clinical characteristics					
CD4					
≤ 250	18 (9)	10 (11)	8 (8)	1 (17)	0 (0)
251-500	57 (29)	32 (34)	25 (24)	2 (33)	3 (33)
501-1000	107 (54)	44 (47)	63 (60)	3 (50)	6 (67)
≥ 1001	17 (8)	8 (8)	9 (8)	0 (0)	0 (0)
CD4NADIR					
≤ 250	74 (37)	31 (33)	43 (41)	2 (33)	3 (33)
251-500	98 (49)	48 (51)	50 (48)	4 (67)	4 (45)
> 500	27 (14)	15 (16)	12 (11)	0 (0)	2 (22)
VLOAD					
≤ 500	168 (85)	81 (87)	87 (83)	5 (83)	8 (89)
501-5000	14 (7)	6 (6)	8 (8)	0 (0)	1 (11)
5001-50000	15 (7)	6 (6)	9 (8)	1 (17)	0 (0)
> 50001	2 (1)	1 (1)	1 (1)	0 (0)	0 (0)
Lifestyle					
Smoking					
Yes	72 (36)	49 (52)	23 (22)	2 (33)	3 (33)
No	127 (64)	45 (48)	78 (78)	4 (67)	6 (67)
Substance abuse					
Yes	24 (12)	15 (16)	9 (9)	0 (0)	1 (11)
No	175 (88)	79 (84)	96 (91)	6 (100)	8 (89)
Education					
<High school	49 (25)	23 (24)	26 (25)	0 (0)	1 (11)
High school	121 (61)	62 (66)	59 (56)	5 (83)	8 (89)
$\geq$College	29 (14)	9 (10)	20 (19)	1 (17)	0 (0)

TABLE 1

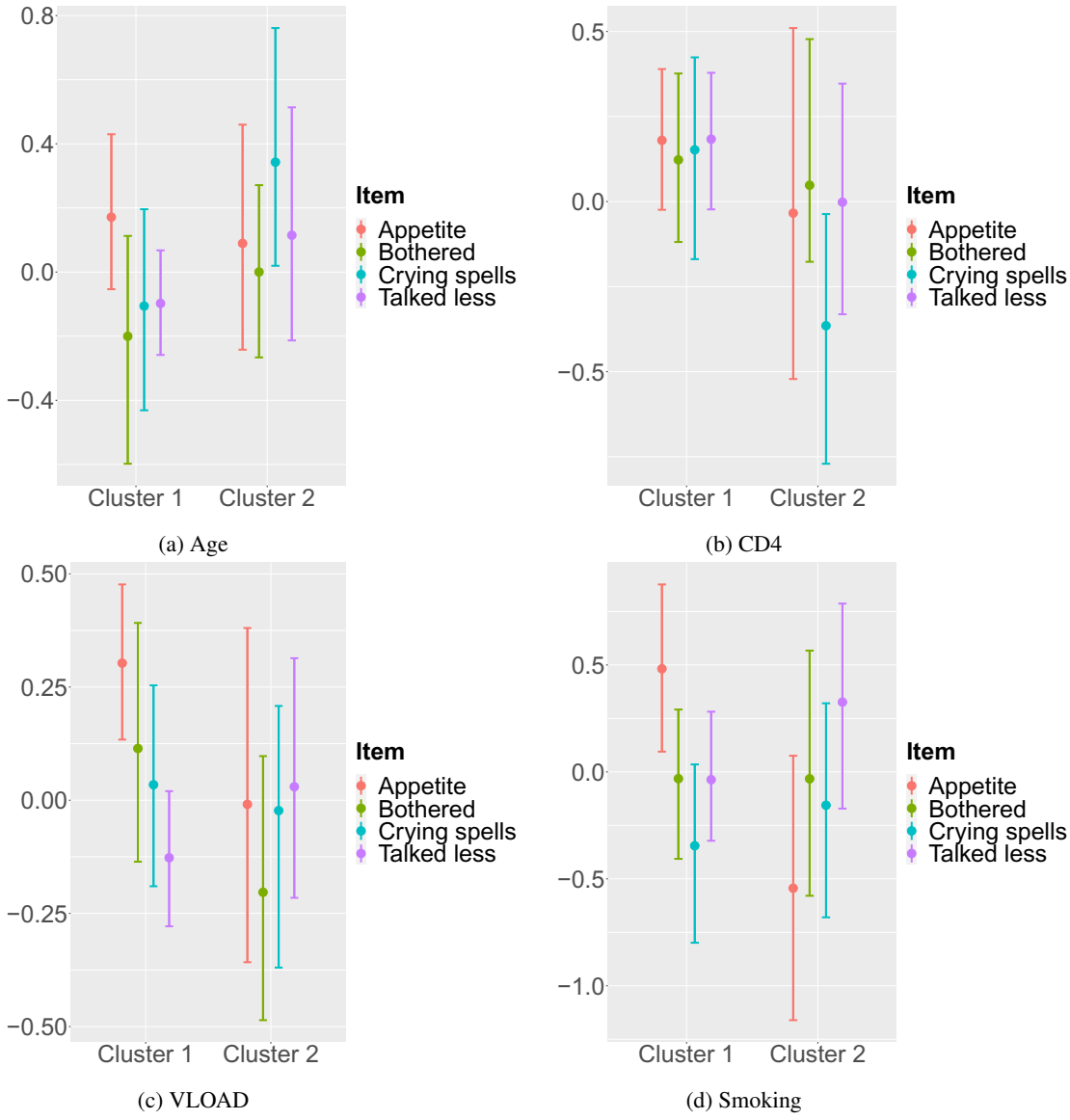
Demographic, clinical, and behavioral characteristics of participants at their first visits in the overall sample and in the four clusters.

demographic, clinical, and behavioral characteristics of participants in the four clusters at their first visits, and Supplementary Table T3 reports the frequency of the ART drugs used in the four clusters. Since clusters 3 and 4 only had few participants, our subsequent analyses and clinical interpretation focused on clusters 1 and 2.

Figure 4 plots posterior means and the corresponding 95% CIs of the estimated effect sizes of age, CD4, viral load, and smoking for four randomly selected depressive symptoms: "appetite", "bothered", "crying spells", and "talked less." As shown in Figure 4, covariates have distinct effects in different clusters on different depression items. Panel (a) shows that the age effects were not significant on appetite in both clusters. In contrast, the effect of age on "crying spells" was significant in cluster 2, but not in cluster 1, highlighting the heterogeneity among participants. Panel (b) indicates that a higher CD4 was associated with less "crying spells" in cluster 2. Panels (c) and (d) show that a higher viral load and smoking were associated with more lack of appetite in cluster 1, but not in cluster 2. As shown in Table 1, the proportion of smoking participants in cluster 1 (52%) is significantly higher than that in cluster 2 (22%), indicating that women who smoke are more likely to experience lack of appetite when viral load is high. These findings are consistent with the literature (Brink et al., 2010; Taniguchi et al., 2014; Clubreth et al., 2016; Williams et al., 2020).

Then we report the estimated $\tilde{R}_{h,dq}$, which represents if drug d is significantly associated with depression item q for participants in cluster h. Note that $\tilde{R}_{h,dq}$ only indicates whether the significant association exists. The sign and magnitude are represented by $\boldsymbol{B}_{ij}$ through $\boldsymbol{\Lambda}_{ij}$, which will be summarized later. Figure 5 plots the posterior probabilities of $\{\tilde{R}_{h,dq} = 1\}$ across all drugs and depression items in clusters 1 and 2, showing that the drug-depression associations are different between clusters. The most frequently used NRTI in the dataset, TDF, was associated with the symptom "people unfriendly" in cluster 1, and "restless" and "effort" in cluster 2 with high probabilities. Several clinical trials have reported the associations between TDF and depression (Squires et al., 2003; Mills et al., 2016). The most frequently used NNRTI drug, EFV, was associated with nearly half of the depression items in cluster 1 and three items in cluster 2 with high probabilities. This is consistent with studies demonstrating associations between EFV and depressive symptoms, such as suicidal behavior (Mollan et al., 2014; Bengtson et al., 2017; Arenas-Pinto et al., 2018). INSTI drugs have also been linked to psychiatric symptoms including depression and are among primary reasons for discontinuing INSTI treatment (Hoffmann et al., 2017; Borghetti et al., 2018; Revuelta-Herrero et al., 2018). For example, Harris et al. (2008) reported several treatment-experienced HIV-seropositive patients had significant exacerbation of pre-existing depression after starting to take RAL, an INSTI drug. The proposed BAGEL identified significant associations between two of the INSTI drugs (EVG and RAL) and several depression items in the two clusters. Compared to previous studies, BAGEL not only determines whether a certain drug is significantly associated with depression, but also which specific depression item that drug is associated with. Since two people with the same sum-score (total depression score) may have distinct symptoms that have different consequences and require different treatments, these more precise findings demonstrate the potential of BAGEL to facilitate precision medicine on treating HIV and its comorbidities including depression.

Next we report the estimation of drug effects $\boldsymbol{B}_{ij}$ on depressive symptoms for participant i at visit j. For illustration, we randomly selected two participants, called participant 1 and participant 2 hereafter. Figure 6 depicts the associations between ART drugs and item-level depressive symptoms over time for participant 1. Blue lines indicate that the ART drug is associated with less symptomatology, indicating favorable effects in reducing depressive symptoms, and red lines indicate that the ART drug is associated with more symptomatology, indicating unfavorable effects on symptoms. The width of the line is proportional to the magnitude of the association (i.e., the estimated posterior mean of $B_{ij,dq}$). Supplementary Table T4 reports the posterior means with 95% CIs of these associations in Figure 6. The drug effects on depressive symptoms change over time, in both magnitude and sign. Particularly, participant 1 had 5 visits and used two NRTIs (TDF+FTC) and one NNRTI (NVP) in all 5 visits. FTC was associated with less symptomatology on "effort", and the beneficial effect



(a) Age

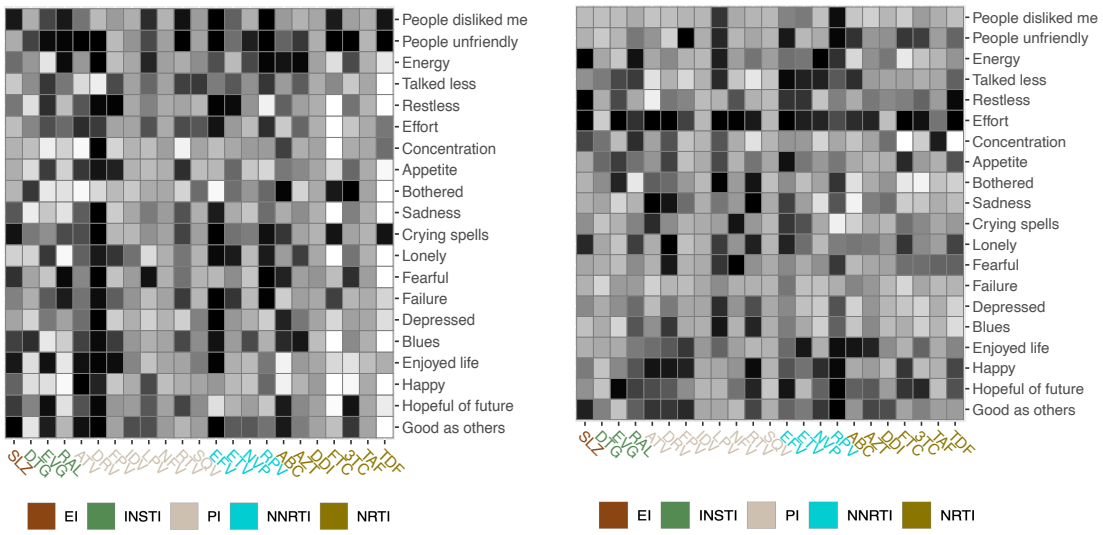
(b) CD4

(c) VLOAD

(d) Smoking

FIG 4. *Posterior means and 95% CIs for the estimated coefficients corresponding to age, CD4, viral load, and smoking. The dots represent the posterior means.*

was magnified over time: -0.864 (95% CI: -2.926, 1.169) at the first visit, then -1.412 (95% CI: -2.662, -0.013), -1.848 (95% CI: -2.875, -0.783), -2.250 (95% CI: -3.087, -1.395) at visits 2-4, and finally -3.006 (95% CI: -4.345, -2.025) at the last visit, indicating that the long-term use of FTC could be beneficial on "effort" for people with HIV. NVP was also beneficial for "energy" in all visits. By contrast, TDF was associated with more symptomatology on "restless", and the negative effect was magnified over time: 0.055 at the first visit, then 0.211, 0.309, 0.47 at visits 2-4, and finally 0.693 at the last visit, indicating that the long-term use of TDF could cause lasting and worse sleeping issue for people with HIV. However, cautions need to be taken to interpret the effect sizes of TDF as they were much smaller than those of FTC and, consequently, were not as statistically significant (see Supplementary Table T4). TDF was also associated with less symptomatology on "effort" at the first visit with the corresponding posterior mean of $B_{ij,dq}$ being -1.472, but the beneficial effect kept decreasing at

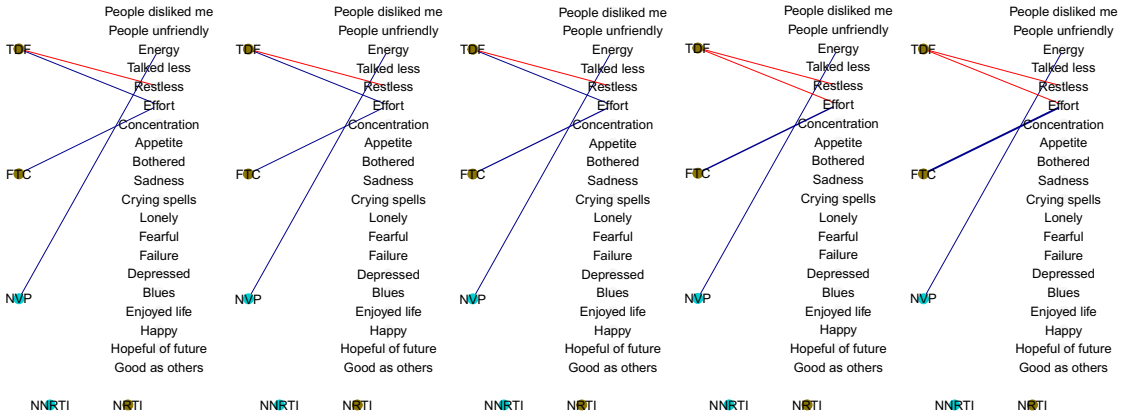


(a) Cluster 1 (b) Cluster 2

FIG 5. *Heatmaps of the posterior probabilities of* $\{\tilde{R}_{h,dq} = 1\}$ *across all drugs and depression items in clusters 1 and 2. A darker cell color represents a higher posterior probability.*

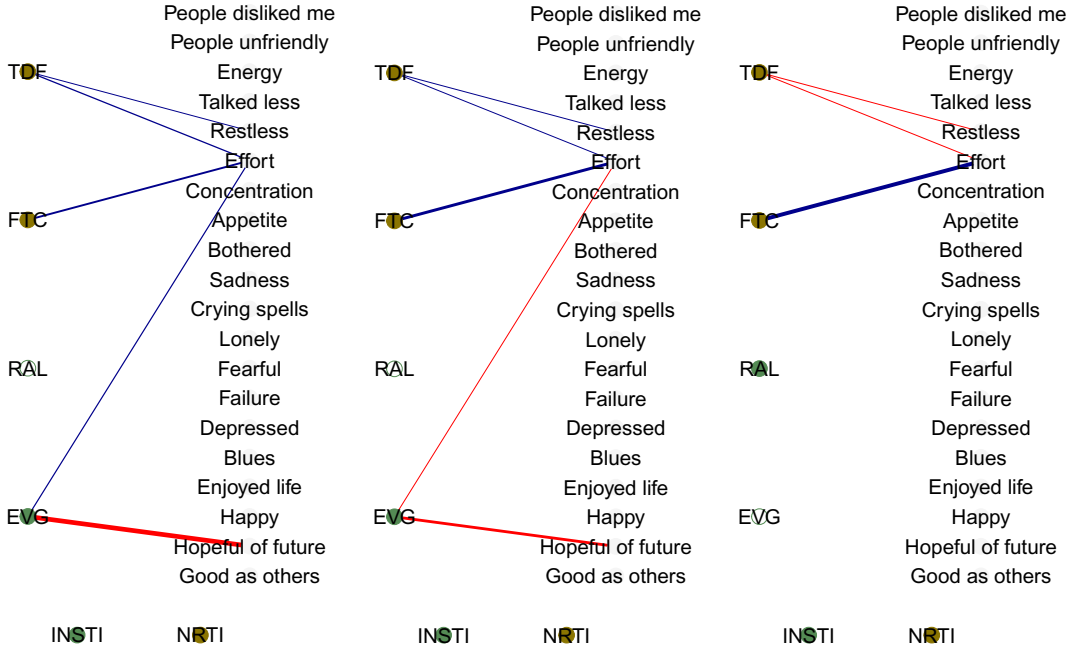
the second and third visits with effects being -0.795 and -0.379, respectively. At the fourth visit, the beneficial effect of TDF on "effort" became harmful with the effect being 0.013, then became worse at the fifth visit with the effect being 0.941. In general, failure to suppress viral load and severe side effects are two major reasons for treatment switch. This participant's viral loads were undetectable in all five visits and she was consistently reported use of TDF+FTC+NVP, which agreed with our finding that TDF+FTC+NVP was well tolerated in terms of depressive symptoms. However, if sleeping problems were a major concern for this woman, then this symptom might not be reflected by the CESD sum-score since the overall score could be dominated by positive effects of drugs on other symptoms. Instead, BAGEL could guide the physician to switch her treatment by considering another NRTI to replace TDF.

Figure 7 plots the associations between ART drugs and item-level depressive symptoms over time for participant 2 who belongs to cluster 2. Supplementary Table T5 reports the posterior means with 95% CIs of these associations in Figure 7. This participant had three visits: at the first two visits, she reported using two NRTIs (TDF and FTC) and one INSTI (EVG); at her last visit, she reported using the same two NRTIs but switched to another INSTI (RAL). Since RAL is not significantly associated with any depressive symptom in cluster 2, there is no edge connecting RAL with symptoms. TDF was associated with less symptomatology in two somatic symptoms ("restless" and "effort") at her first visit, then the beneficial effects decreased at her second visit. As she continued to take TDF, the effects of TDF turned negative on these two depressive symptoms at her last visit. TDF was reported to have limited neurotoxic effects due to its low central nervous system penetrance (Best et al., 2012). However, our finding highlights the need of evaluating the long-term effect of TDF on depressive symptoms in people with HIV, especially somatic symptoms. EVG had a positive effect on "effort" and a negative effect on "hopeful of future" at her first visit, then the effect on "effort" became negative at her second visit. Previous clinical trials (Cohen et al., 2011; Abers et al., 2014) reported psychiatric adverse events for participants treated with EVG. As shown in Figure 7, this participant switched from EVG to RAL, indicating that those negative effects of EVG on depression may be the major reason for this switch because her viral loads were well suppressed across all visits and hence unlikely caused the switch.



(a) Visit 1 (b) Visit 2 (c) Visit 3 (d) Visit 4 (e) Visit 5

FIG 6. *Drug effect on depression symptoms for participant 1. Blue lines indicate that the ART drug is associated with less symptomatology and red lines indicate that the ART drug is associated with more symptomatology. The width of the line indicates the magnitude of the association.*



(a) Visit 1 (b) Visit 2 (c) Visit 3

FIG 7. *Drug effect on depression symptoms for participant 2. Blue lines indicate that the ART drug is associated with less symptomatology and red lines indicate that the ART drug is associated with more symptomatology. The width of the line indicates the magnitude of the association.*

Lastly, we demonstrate how BAGEL can guide more informed and effective individualized ART drug selection for better depression control in HIV clinical practice. Assume that we have observed data for participant i at her first J_i visits, denoted by $\mathcal{D}_i = \{U_{ijq}, \boldsymbol{Z}_{ij}\}_{q=1,j=1}^{Q,J_i}$, and we aim to find an optimal drug combination at her (J_i+1)-st visit. For example, pretend-

ing that participant 2 has data reported up to her second visit, i.e., $J_i = 2$, the analysis with BAGEL can assist physician in drug management at her third visit. The US department of Health and Human Service has recommended two NRTI drugs as backbone with an additional INSTI drug as first-line therapy (https://aidsinfo.nih.gov/guidelines). Assume that at her third visit, the physician wants to choose two NRTIs from TDF, 3TC, and FTC, and choose one INSTI from RAL, EVG, and DTG, resulting in a total of 9 possible combinations. Denote the posterior predictive probability of the q-th depression score at her third visit being larger than 0 under drug combination $\boldsymbol{z}$ by $\pi_{iq}(\boldsymbol{z}) = \Pr(U_{i3q} > 0 \mid \boldsymbol{Z}_{i3} = \boldsymbol{z}, \mathcal{D}_i)$, $q = 1, \ldots, Q$ which is shown in Figure 8. As an example, we define an individualized utility score $\Delta_i(\boldsymbol{z})$ as the criteria of choosing an optimal drug combination based on the sum of posterior predictive probabilities for all depression items $\Delta_i(\boldsymbol{z}) = \sum_q \pi_{iq}(\boldsymbol{z})$. Clearly, a drug combination is less desirable if it has a higher score. For simplicity, we first compare two candidate drug combinations that she took over her three visits (Figure 7), $\boldsymbol{z}_1$=TDF+FTC+EVG and $\boldsymbol{z}_2$=TDF+FTC+RAL. We found $\Delta_i(\boldsymbol{z}_1) = 2.988$ with the top three adverse effects being "effort" (0.439), "talked less" (0.290), and "concentration" (0.188) whereas $\Delta_i(\boldsymbol{z}_2) = 2.714$ with the top three adverse effects being "effort" (0.307), "restless" (0.210), and "happy" (0.202). Since $\boldsymbol{z}_2$=TDF+FTC+RAL leads to a smaller utility score, we claim that $\boldsymbol{z}_2$=TDF+FTC+RAL is better than $\boldsymbol{z}_1$=TDF+FTC+EVG in controlling overall depressive symptoms for this participant and hence would recommend a drug switch from EVG to RAL at her third visit. As shown in Figure 7, the physician's actual choice was indeed RAL, which agreed with our prediction. However, if DTG was considered in the comparison, we would recommend the drug combination TDF+FTC+DTG since it is expected to cause the least depressive symptomatology: $\Delta_i(\boldsymbol{z}) = 2.646$ with the top three adverse effects being "effort" (0.261), "restless" (0.224), and "concentration" (0.215).

Note that in our simple illustrative example, the defined individualized utility score $\Delta_i(\boldsymbol{z})$ simply sums over all depression items with equal weights. In clinical practice, different categories of depression symptoms might impact clinical utility differently. For example, somatic symptoms including "energy", "talked less", "restless", "effort", "concentration", "appetite", and "bothered", have been shown more likely to cause discontinuation of treatment for people with HIV (Kapfhammer, 2006; Taibi, 2013). Therefore, the utility score $\Delta_i(\boldsymbol{z})$ can be adjusted to impose more weights on somatic symptoms chosen according to the physician's opinion or individual preference of people with HIV. For example, let Ξ denote the set of somatic symptoms, the utility score can be alternatively defined as $\widetilde{\Delta}_i(\boldsymbol{z}) = \sum_{q \in \Xi} \pi_{iq}(\boldsymbol{z})$ if the physician wants to focus on somatic symptoms. Again, take participant 2 for illustration. We computed $\widetilde{\Delta}_i(\boldsymbol{z})$ for all 9 drug combinations and concluded that TDF+FTC+DTG was still the optimal choice. Finally, we remark that the flexibility of choosing utility score is enabled by one of the prominent features of BAGEL, i.e., precise probabilistic characterization of ART drug effects on item-level depressive symptoms.

5. Conclusion. To better understand the long-term effects of ART or ART drug switches on item-level depressive symptoms longitudinally and facilitate HIV precision medicine, we developed BAGEL, a novel Bayesian graphical model that estimates longitudinal drug effects on depression while accounting for participants' heterogeneity as well as demographic, clinical, and behavior characteristics. Through simulation studies and analysis of the WIHS dataset, we have demonstrated that BAGEL accurately estimates the longitudinal drug effects, yields meaningful and interpretable results, and has the potential to assist physicians' decisions on personalized ART drug prescriptions. In addition, we have made the code that implements BAGEL publicly available so that users can apply BAGEL to datasets in a similar setup.

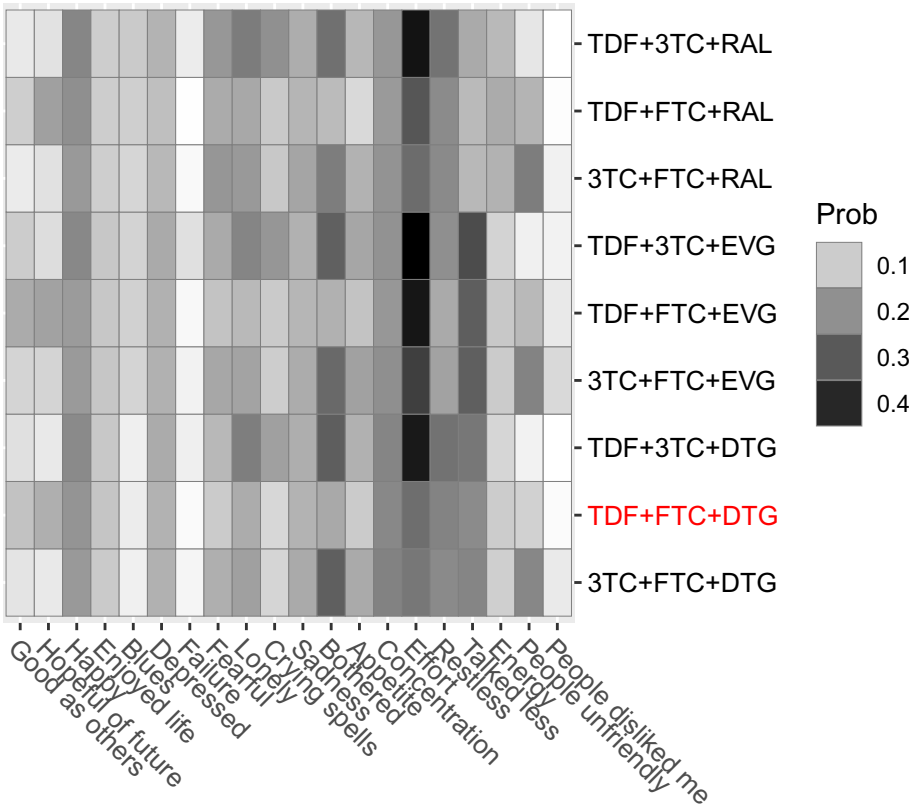


FIG 8. *Posterior predictive probabilities of depression items at participant 2's third visit given the data from her first two visits and various combinations of ART drugs. TDF+FTC+DTG is the optimal drug combination based on the utility score since it leads to the smallest* $\Delta(\boldsymbol{z})$.

We focus on the effects of ART drugs on depressive symptomatology in this paper, but BAGEL ca also helpful to understand other ART-related complications such as cognitive impairment using the model of $\boldsymbol{Y}_{ij}$ directly since cognition outcomes are continuous. BAGEL can also be easily extended to incorporate genetic polymorphisms, an important factor in studies related to mental health and HIV as the use of certain ART drugs in the setting of specific genetic polymorphisms can increase risk for adverse effects on mental health. For example, EFV combined with polymorphisms in CYP2B6 and CYP2A6 increases the risk of suicidality (Bengtson et al., 2017; Mollan et al., 2017). DTG combined with polymorphisms in SLC22A2 is associated with psychiatric symptoms (Borghetti et al., 2018). In the proposed model, polymorphisms can be incorporated as time-invariant covariates, i.e., being part of $\boldsymbol{X}_{ij}$. With this setup, we include the main effects of polymorphisms (via $\boldsymbol{X}_{ij}\boldsymbol{\beta}_i$) as well as their interactions with ART drugs (via $\boldsymbol{Z}_{ij}\boldsymbol{B}_{ij}$). In addition, the information of five drug classes were not used in BAGEL. To incorporate such known clinical knowledge on ART drugs, a hierarchical prior can be considered for modeling the drug effects, which would allow information to be shared among the drugs within the same class. Moreover, to further account for the correlation of depression items (currently only through $\boldsymbol{\omega}_{ij}$), one can impose a low-rank structure/prior on the drug effects, which will potentially facilitate the investigation of how certain drugs may have similar effects on correlated depression items.

BAGEL is based on a longitudinal directed acyclic graphical model. Although specifically motivated by the WIHS application, it can be applied to other biomedical longitudinal

studies or to other fields. For example, in sports medicine, individual monitoring of stress and recovery provides useful information to prevent injuries and illnesses in athletes, and a number of longitudinal studies have been conducted in sports such as soccer and basketball (Jones et al., 2017). BAGEL can be applied to such longitudinal datasets to study the impact of psychosocial stress on the risk of sports injuries adjusting for physical stress effect.

Acknowledgment. This work was supported by the Johns Hopkins University Center for AIDS Research NIH/NIAID fund (P30AI094189) 2019 faculty development award to Dr. Xu, National Science Foundation 1940107 to Dr. Xu, National Science Foundation DMS1918854 to Drs. Xu and Rubin, and National Science Foundation DMS1918851 to Dr. Ni.

SUPPLEMENTARY MATERIAL

The supplement contains all computational details, supplementary figures and tables, and more simulation studies.

REFERENCES

Abers, M. S., Shandera, W. X., and Kass, J. S. (2014). Neurological and psychiatric adverse effects of antiretroviral drugs. *CNS drugs*, 28(2):131–145.

Adimora, A. A., Ramirez, C., Benning, L., Greenblatt, R. M., Kempf, M.-C., Tien, P. C., Kassaye, S. G., Anastos, K., Cohen, M., Minkoff, H., et al. (2018). Cohort profile: the womens interagency hiv study (wihs). *International journal of epidemiology*, 47(2):393–394i.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

Arenas-Pinto, A., Grund, B., Sharma, S., Martinez, E., Cummins, N., Fox, J., Klingman, K. L., Sedlacek, D., Collins, S., Flynn, P. M., et al. (2018). Risk of suicidal behavior with use of efavirenz: results from the strategic timing of antiretroviral treatment trial. *Clinical Infectious Diseases*, 67(3):420–429.

Bacon, M. C., Von Wyl, V., Alden, C., Sharp, G., Robison, E., Hessol, N., Gange, S., Barranday, Y., Holman, S., Weber, K., and Young, M. A. (2005). The women's interagency hiv study: an observational cohort brings clinical sciences to the bench. *Clinical and diagnostic laboratory immunology*, 12(9):1013–1019.

Barkan, S. E., Melnick, S. L., Preston-Martin, S., Weber, K., Kalish, L. A., Miotti, P., Young, M., Greenblatt, R., Sacks, H., and Feldman, J. (1998). The women's interagency hiv study. *Epidemiology*, pages 117–125.

Bengtson, A. M., Pence, B. W., Crane, H. M., Christopoulos, K., Fredericksen, R. J., Gaynes, B. N., Heine, A., Mathews, W. C., Moore, R., Napravnik, S., Safren, S., and Mugavero, M. J. (2016). Disparities in depressive symptoms and antidepressant treatment by gender and race/ethnicity among people living with hiv in the united states. *PloS one*, 11(8):e0160738.

Bengtson, A. M., Pence, B. W., Mollan, K. R., Edwards, J. K., Moore, R. D., O'CLEIRIGH, C., Eaton, E. F., Eron, J. J., Kitahata, M. M., Mathews, W. C., et al. (2017). The relationship between efavirenz as initial antiretroviral therapy and suicidal thoughts among HIV-infected adults in routine care. *Journal of acquired immune deficiency syndromes (1999)*, 76(4):402.

Best, B. M., Letendre, S. L., Koopmans, P., Rossi, S. S., Clifford, D. B., Collier, A. C., Gelman, B. B., Marra, C. M., McArthur, J. C., McCutchan, J. A., et al. (2012). Low csf concentrations of the nucleotide hiv reverse transcriptase inhibitor, tenofovir. *Journal of acquired immune deficiency syndromes (1999)*, 59(4):376.

Borghetti, A., Baldin, G., Capetti, A., SterrantCohen ino, G., Rusconi, S., Latini, A., Giacometti, A., Madeddu, G., Picarelli, C., De Marco, R., Cossu, M. V., Lagi, F., Cauda, R., De Luca, A., Giambenedetto, D., and Group, S. O. S. (2017). Efficacy and tolerability of dolutegravir and two nucleos (t) ide reverse transcriptase inhibitors in HIV-1-positive, virologically suppressed patients. *AIDS*, 31(3):457–459.

Borghetti, A., Calcagno, A., Lombardi, F., Cusato, J., Belmonti, S., D?avolio, A., Ciccarelli, N., La Monica, S., Colafigli, M., Delle Donne, V., Marco, R. D., Tamburrini, E., Visconti, E., Perri, G. D., Luca, A. D., Bonora, S., and Giambenedetto, S. D. (2018). SLC22A2 variants and dolutegravir levels correlate with psychiatric symptoms in persons with HIV. *Journal of Antimicrobial Chemotherapy*.

Brickman, C., Propert, K. J., Voytek, C., Metzger, D., and Gross, R. (2017). Association between depression and condom use differs by sexual behavior group in patients with hiv. *AIDS and Behavior*, 21(6):1676–1683.

Brink, M. S., Visscher, C., Arends, S., Zwerver, J., Post, W. J., and Lemmink, K. A. (2010). Monitoring stress and recovery: new insights for the prevention of injuries and illnesses in elite youth soccer players. *British Journal of Sports Medicine*, 44(11):809–815.

Chattopadhyay, S., Ball, S., Kargupta, A., Talukdar, P., Roy, K., Talukdar, A., and Guha, P. (2017). Cognitive behavioral therapy improves adherence to antiretroviral therapy in hiv-infected patients: a prospective randomized controlled trial from eastern india. *HIV & AIDS Review. International Journal of HIV-Related Problems*, 16(2):89–95.

Clubreth, R., Dube, S., and Maggio, D. (2016). Associations between major depression, health-risk behaviors, and medication adherence among hiv-positive adults receiving medical care in georgia. *Journal of the Georgia Public Health Association*.

Cohen, C., Elion, R., Ruane, P., Shamblaw, D., DeJesus, E., Rashbaum, B., Chuck, S. L., Yale, K., Liu, H. C., Warren, D. R., et al. (2011). Randomized, phase 2 evaluation of two single-tablet regimens elvitegravir/cobicistat/emtricitabine/tenofovir disoproxil fumarate versus efavirenz/emtricitabine/tenofovir disoproxil fumarate for the initial treatment of hiv infection. *Aids*, 25(6):F7–F12.

Cohen, J., D?Agostino, L., Wilson, J., Tuzer, F., and Torres, C. (2017). Astrocyte senescence and metabolic changes in response to hiv antiretroviral therapy drugs. *Frontiers in aging neuroscience*, 9:281.

Cook, J. A., Cohen, M. H., Burke, J., Grey, D., Anastos, K., Kirstein, L., Palacio, H., Richardson, J., Wilson, T., and Young, M. (2002). Effects of depressive symptoms and mental health quality of life on use of highly active antiretroviral therapy among hiv-seropositive women. *JAIDS-HAGERSTOWN MD-*, 30(4):401–409.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In Do, K.-A., Müller, P., Vannucci, MarinaDo, K.-A., Müller, P., and Vannucci, M., editors, *Bayesian Inference for Gene Expression and Proteomics*, chapter 10, pages 201–218. Cambridge University Press.

Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., and Covi, L. (1974). The Hopkins Symptom Checklist (HSCL): A self-report symptom inventory. *Behavioral science*, 19(1):1–15.

Dey, S., Zhang, P., Sow, D., and Ng, K. (2019). Perdrep: Personalized drug effectiveness prediction from longitudinal observational data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1258–1268.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, pages 89–102.

Elzi, L., Erb, S., Furrer, H., Cavassini, M., Calmy, A., Vernazza, P., Günthard, H., Bernasconi, E., and Battegay, M. (2017). Adverse events of raltegravir and dolutegravir. *AIDS (London, England)*, 31(13):1853.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The annals of statistics*, 2(4):615–629.

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., and Borsboom, D. (2016). Measuring depression over time... or not? lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment*, 28(11):1354.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016.

Harris, M., Larsen, G., and Montaner, J. S. (2008). Exacerbation of depression associated with starting raltegravir: a report of four cases. *Aids*, 22(14):1890–1892.

Hoffmann, C., Welz, T., Sabranski, M., Kolb, M., Wolf, E., Stellbrink, H.-J., and Wyen, C. (2017). Higher rates of neuropsychiatric adverse events leading to dolutegravir discontinuation in women and older patients. *HIV medicine*, 18(1):56–63.

Ickovics, J. R., Hamburger, M. E., Vlahov, D., Schoenbaum, E. E., Schuman, P., Boland, R. J., Moore, J., and Group, H. E. R. S. (2001). Mortality, cd4 cell count decline, and depressive symptoms among hiv-seropositive women: longitudinal analysis from the hiv epidemiology research study. *Jama*, 285(11):1466–1474.

Ironson, G., Fitch, C., and Stuetzle, R. (2017). Depression and survival in a 17-year longitudinal study of people with hiv: Moderating effects of race and education. *Psychosomatic medicine*, 79(7):749–756.

Jones, C. M., Griffiths, P. C., and Mellalieu, S. D. (2017). Training load and fatigue marker associations with injury and illness: a systematic review of longitudinal studies. *Sports Medicine*, 47(5):943–974.

Kapfhammer, H.-P. (2006). Somatic symptoms in depression. *Dialogues in clinical neuroscience*, 8(2):227.

Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Lang, S. and Brezger, A. (2004). Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1):183–212.

Lewinsohn, P. M., Seeley, J. R., Roberts, R. E., and Allen, N. B. (1997). Center for Epidemiologic Studies Depression Scale (CES-D) as a screening instrument for depression among community-residing older adults. *Psychology and aging*, 12(2):277.

Li, Y., Bandyopadhyay, D., Xie, F., and Xu, Y. (2020). BAREB: A Bayesian repulsive biclustering model for periodontal data. *Statistics in Medicine*, 39(16):2139–2151.

Li, Y., Ni, Y., Rubin, L. H., Spence, A. B., and Xu, Y. (2021). Supplement to "BAGEL: A Bayesian Graphical Model for Inferring Drug Effect Longitudinally on Depression in People with HIV".

Liu, B., Li, Y., Ghosh, S., Sun, Z., Ng, K., and Hu, J. (2019). Complication risk profiling in diabetes care: A bayesian multi-task and feature relationship learning approach. *IEEE Transactions on Knowledge and Data Engineering*.

Maki, P. M., Rubin, L. H., Cohen, M., Golub, E. T., Greenblatt, R. M., Young, M., Schwartz, R. M., Anastos, K., and Cook, J. A. (2012). Depressive symptoms are increased in the early perimenopausal stage in ethnically diverse HIV+ and HIV- women. *Menopause (New York, NY)*, 19(11):1215.

Mills, A., Arribas, J. R., Andrade-Villanueva, J., DiPerri, G., Van Lunzen, J., Koenig, E., Elion, R., Cavassini, M., Madruga, J. V., Brunetta, J., et al. (2016). Switching from tenofovir disoproxil fumarate to tenofovir alafenamide in antiretroviral regimens for virologically suppressed adults with hiv-1 infection: a randomised, active-controlled, multicentre, open-label, phase 3, non-inferiority study. *The Lancet Infectious Diseases*, 16(1):43–52.

Mollan, K. R., Smurzynski, M., Eron, J. J., Daar, E. S., Campbell, T. B., Sax, P. E., Gulick, R. M., Na, L., O'Keefe, L., Robertson, K. R., et al. (2014). Association between efavirenz as initial therapy for hiv-1 infection and increased risk for suicidal ideation or attempted or completed suicide: an analysis of trial data. *Annals of internal medicine*, 161(1):1–10.

Mollan, K. R., Tierney, C., Hellwege, J. N., Eron, J. J., Hudgens, M. G., Gulick, R. M., Haubrich, R., Sax, P. E., Campbell, T. B., Daar, E. S., Robertson, K. R., Ventura, D., Ma, Q., Edwards, D. R. V., Haas, D. W., and the AIDS Clinical Trials Group (2017). Race/ethnicity and the pharmacogenetics of reported suicidality with efavirenz among clinical trials participants. *The Journal of infectious diseases*, 216(5):554–564.

Moore, J., Schuman, P., Schoenbaum, E., Boland, B., Solomon, L., and Smith, D. (1999). Severe adverse life events and depressive symptoms among women with, or at risk for, hiv infection in four cities in the united states of america. *Aids*, 13(17):2459–2468.

Müller, P. and Quintana, F. A. (2004). Nonparametric bayesian data analysis. *Statistical science*, pages 95–110.

Nanni, M. G., Caruso, R., Mitchell, A. J., Meggiolaro, E., and Grassi, L. (2015). Depression in HIV infected patients: a review. *Current psychiatry reports*, 17(1):530.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: convergence diagnosis and output analysis for mcmc. *R news*, 6(1):7–11.

Revuelta-Herrero, J. L., Chamorro-de Vega, E., Rodríguez-González, C. G., Alonso, R., Herranz-Alonso, A., and Sanjurjo-Sáez, M. (2018). Effectiveness, safety, and costs of a treatment switch to dolutegravir plus rilpivirine dual therapy in treatment-experienced hiv patients. *Annals of Pharmacotherapy*, 52(1):11–18.

Rubin, L. H., Cook, J. A., Grey, D. D., Weber, K., Wells, C., Golub, E. T., Wright, R. L., Schwartz, R. M., Goparaju, L., Cohan, D., et al. (2011). Perinatal depressive symptoms in HIV-infected versus HIV-uninfected women: a prospective study from preconception to postpartum. *Journal of Women's Health*, 20(9):1287–1295.

Shah, A., Gangwani, M. R., Chaudhari, N. S., Glazyrin, A., Bhat, H. K., and Kumar, A. (2016). Neurotoxicity in the post-haart era: caution for the antiretroviral therapeutics. *Neurotoxicity research*, 30(4):677–697.

Squires, K., Pozniak, A. L., Pierone, G., Steinhart, C. R., Berger, D., Bellos, N. C., Becker, S. L., Wulfsohn, M., Miller, M. D., Toole, J. J., et al. (2003). Tenofovir disoproxil fumarate in nucleoside-resistant hiv-1 infection: a randomized trial. *Annals of internal medicine*, 139(5_Part_1):313–320.

Taibi, D. M. (2013). Sleep disturbances in persons living with hiv. *Journal of the Association of Nurses in AIDS Care*, 24(1):S72–S85.

Taniguchi, T., Shacham, E., Önen, N. F., Grubb, J. R., and Overton, E. T. (2014). Depression severity is associated with increased risk behaviors and decreased cd4 cell counts. *AIDS care*, 26(8):1004–1012.

Underwood, J., Robertson, K. R., and Winston, A. (2015). Could antiretroviral neurotoxicity play a role in the pathogenesis of cognitive impairment in treated hiv disease? *Aids*, 29(3):253–261.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

Williams, D. W., Li, Y., Dastgheyb, R., Fitzgerald, K. C., Maki, P. M., Spence, A. B., Gustafson, D. R., Milam, J., Sharma, A., Adimora, A. A., et al. (2020). Associations between antiretroviral drugs on depressive symptomatology in homogenous subgroups of women with hiv. *Journal of Neuroimmune Pharmacology*, pages 1–14.

Xu, Y., Xu, Y., and Saria, S. (2016). A non-parametric Bayesian approach for estimating treatment-response curves from sparse time series. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, pages 282–300.

Zash, R., Makhema, J., and Shapiro, R. L. (2018). Neural-tube defects with dolutegravir treatment from the time of conception. *New England Journal of Medicine*, 379(10):979–981.

Preflight Results

Document Overview

Title: BAGEL: A Bayesian Graphical Model for Inferring Drug Effect Longitudinally in People with HIV
Author: Yuliang Li, Yang Ni, Leah H. Rubin, Amanda B. Spence
Creator: LaTeX with hyperref package
Producer: pdfTeX-1.40.16

Preflight Information

Profile: Convert to PDF/A-2b
Version: Quoppa jPDFPreflight v2020R2.01
Date: Jul 1, 2021 4:28:56 PM

Legend: (X) - Can NOT be fixed by PDF/A-2b conversion.
(!X) - Could be fixed by PDF/A-2b conversion. User chose to be warned in PDF/A settings.

Page 5 Results

(X) Page uses transparency but does not have a device independent Blending Color Space