JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 00, NO. 0, 1–11: Precision Medicine
https://doi.org/10.1080/01621459.2020.1785476

Personalized Policy Learning Using Longitudinal Mobile Health Data

Xinyu Hu, Min Qian, Bin Cheng, and Ying Kuen Cheung

Department of Biostatistics, Columbia University, New York, NY

ABSTRACT

Personalized policy represents a paradigm shift one decision rule for all users to an individualized decision rule for each user. Developing personalized policy in mobile health applications imposes challenges. First, for lack of adherence, data from each user are limited. Second, unmeasured contextual factors can potentially impact on decision making. Aiming to optimize immediate rewards, we propose using a generalized linear mixed modeling framework where population features and individual features are modeled as fixed and random effects, respectively, and synthesized to form the personalized policy. The group lasso type penalty is imposed to avoid overfitting of individual deviations from the population model. We examine the conditions under which the proposed method work in the presence of time-varying endogenous covariates, and provide conditional optimality and marginal consistency results of the expected immediate outcome under the estimated policies. We apply our method to develop personalized push ("prompt") schedules in 294 app users, with the goal to maximize the prompt response rate given past app usage and other contextual factors. The proposed method compares favorably to existing estimation methods including using the R function "glmer" in a simulation study. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2019
Accepted June 2020

KEYWORDS
Contextual bandits; Endogenous variables; Generalized linear mixed model; Individualized decision rule; Push notifications

1. Introduction

Mobile technologies such as smartphones and wearables enable continuous monitoring of exposure to environmental stressors and ecological assessment of health-relevant data over an extended period of time, thereby facilitating the delivery of tailored interventions in an adaptive manner (Riley et al. 2011). Examples of utilization abound. Heron and Smyth (2010) reviewed the use of tailored interventions based on momentary assessments to support management of a variety of health behaviors and symptoms such as smoking, diabetes, and weight loss. Depp et al. (2010) studied the efficacy of personalized pushed engagement based on real-time data in mental illness patients. Mohr et al. (2013) envisioned a continuous evaluation system of health apps based on evidence generated by routinely collected data. To illustrate, we consider a suite of smartphone apps, called IntelliCare, that serves users with anxiety or depression using different psychological treatment strategies including cognitive behavioral therapy, positive psychology, and physical activity-based interventions (Mohr et al. 2017). The suite consists of a Hub app that helps users navigate apps within the IntelliCare ecosystem and coordinates their experience, with a specific function to provide links and recommendations for other IntelliCare apps so as to maximize user engagement based on a user's app usage history (Cheung et al. 2018). In this article, we are motivated by a sub-study of the IntelliCare suite, in which the Hub app would send push notifications to prompt a user to complete a short four-item patient health questionnaire repeatedly at 7-day intervals at a random time during the day. The purpose of these prompts is to remind users to assess their depression and anxiety symptoms.

The objective of the study is to learn the best time to push a prompt, a factor that is known to impact mobile application usage (Bohmer et al. 2011), to maximize a user's immediate response rate given his or her past engagement and other contextual factors experienced. There are three main reasons to focus on immediate outcomes. First, it may facilitate timely adjustment to a user's other intervention decisions. In the IntelliCare example, information on a prompt responder's health status can assist the Hub app to determine the best app(s) to recommend for the user in the following week, while such information is not available for prompt nonresponders. Second, an immediate outcome often serves as a measure of a user's short-term engagement, which, once sustained, is expected to translate into a long-term engagement, a prelude for behavior changes. Engagement on mobile health apps is reportedly declining quickly over time in the literature (Christensen, Griffiths, and Farrer 2009; Helander et al. 2014). Therefore, constant attentions and prompt actions are required to maintain a user's engagement to materialize any behavior change. Third, optimizing immediate outcome, given the rapid decline in engagement, is most practically feasible, as compared to long-term outcome. While long-term outcome is often desirable, its practical benefit often requires long-term engagement, which is particularly challenging in mobile health applications. Aiming for immediate outcome faces a minimal risk in practical implementation.

CONTACT Min Qian mq2158@cumc.columbia.edu Department of Biostatistics, Columbia University, 722 West 168th Street, 6th Floor, New York, NY 10032.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

Developing decision policies to prescribe evidence-based mobile-delivered intervention is challenging because it must incorporate both shared and individual information. On one hand, decisions on intervention for a given cohort are often based on some shared characteristics, making it possible to learn by borrowing strength from users sharing similar features. On the other hand, any sensible policy on human behavior, mobile health policies included, is necessarily individualized due to unmeasured heterogeneities which should be accounted for when deciding upon an appropriate intervention (Ohrnberger, Fichera, and Sutton 2017). For example, whether and how a user will respond to a push notification depends vastly on his or her circumstance at that moment, which is either difficult or impossible to characterize. Numerous policy learning methods that support decision making using medical data and mobile health data have been proposed. For example, there is a large statistical literature on reinforcement learning (RL) algorithms that estimate optimal policies under a nomothetic model (Murphy 2003; Qian and Murphy 2011; Zhang et al. 2012; Zhao et al. 2012, 2015; Laber, Linn, and Stefanski 2014; Song et al. 2015; Ertefaie and Strawderman 2018; Luckett et al. 2020). A nomothetic approach assumes that a population model captures all between-subject heterogeneity and facilitates estimation by pooling data across participants. While this approach may address user heterogeneity and allow for estimation of personalized policies by incorporating appropriate interactions with actions, it often requires the untestable assumption of no unobserved confounders. Alternatively, an ideographic approach achieves personalization using an "N-of-1" method whereby each person's decision model is estimated using his or her own data only (Lillie et al. 2011; Kravitz and Duan 2014; Lei, Tewari, and Murphy 2017). Although in principle this approach allows for insights about individuals without assumptions on any reference population, its practicality relies on how long a user can be followed. In general, the efficiency of this approach may suffer, especially in situations where an action exhibits similar effects on all individuals.

In this article, we aim to develop policies, one for each user, that provide personalized feedback through their interaction with the IntelliCare apps. We consider estimating personalized policy under a generalized contextual bandits framework. Contextual bandits are popular approaches to personalizing interventions in mobile health applications (Lei, Tewari, and Murphy 2017; Tewari and Murphy 2017). It typically assumes that the contexts are independent and identically distributed, or more generally, that future contexts are independent of current actions. We extend conventional stochastic contextual bandits in two ways: first, we allow the current contextual factors to be endogenous, that is, dependent on the outcome process and previous treatment assignments; and second, we adopt the generalized linear mixed model (GLMM) framework with the outcome as dependent variable and the time-varying contextual factors, action and their interactions as predictors. For instance, in the IntelliCare "Prompt" sub-study, the outcome of interest is a binary response and the action is the time period during a day when a prompt is pushed. The estimated policy aims to recommend an action that maximizes the predicted outcome based on the contextual factors experienced by a user and his or her past engagement. In addition to tailoring, each user will have a personalized policy through the estimation of random effects, which account for individual departures from the population model due to unobserved heterogeneity.

While the GLMM is one of the most popular methods to handle longitudinal outcome data, GLMM-based estimation methods are largely designed for settings where the covariates are exogenous with respect to the outcome process. When the time-varying covariates are allowed to be endogenous (e.g., depending on the outcome process, previous treatment assignments, and possibly random effects parameters), estimation of the GLMM fixed effect coefficients based on likelihood or generalized estimating equations may lead to bias because it no longer corresponds to the conditional interpretation of the parameters (see, e.g., Pepe and Anderson 1994; Diggle et al. 2002). For linear mixed models (LMM) where the conditional interpretation of fixed effects is consistent with the scientific interest in predicting person-specific effects, Qian, Klasnja, and Murphy (2019) showed that standard software can be used to obtain a valid estimate of the fixed effects if the time-varying covariates are independent of the random effects parameters conditional on past history. In this article, we further explore the behavior of both fixed effects and random effects parameter estimates in GLMM in the presence of endogeneity.

We note some previous work on studying personalized treatment effects using GLMM. For example, Cho, Wang, and Qu (2017) used GLMM to predict individual outcome under each treatment arm with a random slope on the treatment indicator, and built a random forest model to predict random slope using patients baseline covariates. Personalized treatment can then be implemented by selecting the treatment with the maximal estimated random effects. Qian, Klasnja, and Murphy (2019) studied the bias and variance of parameter estimates in LMM. However, neither of them considered constructing personalized policy through including random effects on treatment-by-covariate interactions in their models. Allowing for random effects on treatment-by-covariate interactions presents a key computational challenge due to the large number of random effects. To address this challenge, we propose a novel algorithm that jointly estimates the fixed effects and the random effects under a ridge-type penalty on the latter. In addition, to avoid overfitting of the individual deviations from the population mean, we propose a group lasso penalty on the random effects of each covariate (Yuan and Lin 2006). Intuitively, we view the random effects of a covariate as its interactions with the unobserved confounders. If such interactions do not exist, then we expect the corresponding random effects for that covariate shrink to zero simultaneously for all users (hence the *group* lasso); otherwise, the random effects yield the desired heterogeneous policies across users.

In this article, we examine conditions under which the proposed method yields consistent estimators in the presence of endogeneity in GLMM. Compared with Qian, Klasnja, and Murphy (2019), our conditions allow the contextual factors to depend directly on the outcome process or latent random effect parameters. Awaring that a personalized policy reflects both the population and the individual, we propose to use an average error bound to quantify its asymptotic performance. Specifically, we provide an asymptotic averaged error bound of the estimated policies for the conditional expected immediate outcome given current contextual factors and a consistency result of the marginal expected immediate outcome. Furthermore, as shown in Section 3, our method does not require a full conditional distribution of outcome or random effects to be correctly specified, but relies on a much weaker assumption that the conditional mean outcome model is correctly specified.

This article is organized as follows. In Section 2, we set up the formulation of the personalized policy learning problem, and present a new policy estimation method. We then study the theoretical properties of the proposed method in Section 3, and compare it with some existing approaches in Section 4. We will revisit the IntelliCare Prompt study in Section 5 and apply the proposed method to develop personalized policies in the study. We end this article with some concluding remarks in Section 6. Details of computational algorithms, technical derivations, and proofs are provided as separate supplementary materials.

2. Personalized Policy Learning

2.1. Notations and Problem Formulation

Suppose we have collected data from n mobile application users. Each user i is tracked longitudinally over m_i time intervals. At each time interval $t = 1, \ldots, m_i$, let $S_{it} \in \mathcal{S}$ be the vector of contextual factors containing the user's observation prior to the decision point t (e.g., demographics, app usage prior to decision, etc.), A_{it} be the decision action taking values in a prespecified finite discrete space $\mathcal{A}$ that is randomized to the user, and Y_{it} be the scalar outcome of interest observed after decision, with the convention that larger values of Y_{it} are preferred. We note that the covariates S_{it} may include endogenous variables that depend on previous outcomes and actions, as well as other exogenous and contextual factors. Here we assume that the decision point is fixed. For example, in the IntelliCare sub-study, the decision action is the time period of sending push notification during a day for every seven days since installation of the Hub app; contextual factors may contain app usage pattern in the previous seven days, whether the decision day is a weekday or weekend, user's response to previous action, etc.; and the outcome of interest is whether a user responds to the push notification. In summary, the observed trajectory of each user is denoted by the triplets $\{(S_{it}, A_{it}, Y_{it}) : t = 1, \ldots, m_i\}$. We further denote the entire history up to t by $\underline{S}_{it} = (S_{i1}, \ldots, S_{it})$ and $\underline{A}_{it} = (A_{i1}, \ldots, A_{it})$.

Our objective is to estimate for a given user i a personalized policy π_{0i}, which when implemented will result in the maximal conditional expected outcome, $E_{\pi_{0i}}(Y_{it}|\underline{S}_{it}, \underline{A}_{i,t-1})$, where the expectation is taken with respect to the conditional distribution of Y_{it} given the history $(\underline{S}_{it}, \underline{A}_{i,t-1})$ and action $A_{i,t}$ is consistent with π_{0i}. We further make the commonly used assumption that the conditional distribution of Y_{it} given $\underline{S}_{it}, \underline{A}_{it}$ is Markovian, so that $E_{\pi_{0i}}(Y_{it}|\underline{S}_{it}, \underline{A}_{i,t-1}) = E_{\pi_{0i}}(Y_{it}|S_{it}) = E\{Y_{it}|S_{it}, A_{it} = \pi_{0i}(S_{it})\}$. We note that S_{it} can include lagged variables at previous time points (e.g., $Y_{i,t-2}$). Further let $Q_{0i}(s, a) = E(Y_{it}|S_{it} = s, A_{it} = a)$ so as to make explicit the conditional expectation is user-specific. Then $\pi_{0i}(s) \in \arg\max_{a \in \mathcal{A}} Q_{0i}(s, a)$. Once π_{0i} is estimated by $\hat{\pi}_i$ (say), the estimated policy will be used to guide decision making for the user in the future time points. While this formulation of the problem assumes a stationary policy in that the function Q_{0i} is time-invariant, the policy decisions can be time-dependent by including time in the covariate state S_{it}. In our application, this assumption is aligned with the fact that mobile application usage is habitual given other contextual factors.

We facilitate the learning problem under the GLMM framework, and postulate

$$Q_{0i}(S_{it}, A_{it}) = g^{-1}\left\{h_1(S_{it}, A_{it})^\top \boldsymbol{\beta}_0 + h_2(S_{it}, A_{it})^\top \boldsymbol{\alpha}_{0i\cdot}^\top\right\} := Q(S_{it}, A_{it}; \boldsymbol{\beta}_0, \boldsymbol{\alpha}_{0i\cdot}), \quad (1)$$

for $i = 1, \ldots, n$ and $t = 1, \ldots, m_i$, where $g(\cdot)$ is a known strictly monotone increasing link function. For example, the canonical forms of $g(\cdot)$ are, respectively, the identity function for continuous outcome, logit for binary outcome, and logarithmic for counts. Here $\boldsymbol{\beta}_0$ is a p-dimensional vector of unknown parameters, and $h_1(S_{it}, A_{it}) \in \mathbb{R}^p$ is a prespecified vector function of (S_{it}, A_{it}) so that $h_1(S_{it}, A_{it})^\top \boldsymbol{\beta}_0$ is the fixed effects component; for example, $h_1(S_{it}, A_{it})^\top \boldsymbol{\beta} = \beta_0 + \beta_1 S_{it} + \beta_2 A_{it} + \beta_3 S_{it} A_{it}$. The random effects are denoted by $\boldsymbol{\alpha}_0$, an $n \times q$ $(q \le p)$ matrix with the ith row, $\boldsymbol{\alpha}_{0i\cdot}$, denoting the random effects parameters for the ith user, and $h_2(S_{it}, A_{it}) \in \mathbb{R}^q$ is a sub-vector of $h_1(S_{it}, A_{it})$ chosen so that $h_2(S_{it}, A_{it})^\top \boldsymbol{\alpha}_{0i\cdot}$ models subject-specific deviations from the mean model. Under model (1) and a monotone increasing $g(\cdot)$, the optimal policy π_{0i} can be expressed as

$$\pi_{0i}(S_{it}) \in \arg\max_{a \in \mathcal{A}} \left(h_1(S_{it}, a)^\top \boldsymbol{\beta}_0 + h_2(S_{it}, a)^\top \boldsymbol{\alpha}_{0i\cdot}^\top\right). \quad (2)$$

Note that $\boldsymbol{\alpha}_{0i\cdot}$ plays dual roles in our proposed method. On one hand, it defines the individual deviation from the mean model of the ith user, and can be viewed as a fixed parameter to be estimated and to be acted upon. This role operationalizes the personalized policy decisions (2). On the other hand, $\{\boldsymbol{\alpha}_{0i\cdot}\}$ can be viewed as a random sample of the population. This viewpoint motivates some degree of "smoothness" in the estimation of $\boldsymbol{\alpha}_{0i\cdot}$'s, which is described next.

2.2. Policy Estimation

Let $\{\ell(Y_{it}, S_{it}, A_{it}; \boldsymbol{\beta}, \boldsymbol{\alpha}_{i\cdot}, \phi) : \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\alpha}_{i\cdot}^\top \in \mathbb{R}^q\}$ denote the working conditional log-likelihood of Y_{it} under a fully specified GLMM with the systematic component (1). For example, with a continuous Y_{it}, we may set $\ell(\cdot)$ to be the Gaussian log-likelihood with mean $Q(S_{it}, A_{it}; \boldsymbol{\beta}, \boldsymbol{\alpha}_{i\cdot})$, variance σ^2, and an identity link. When Y_{it} is binary, we may choose $\ell(\cdot)$ to be the Bernoulli log-likelihood with probability $Q(S_{it}, A_{it}; \boldsymbol{\beta}, \boldsymbol{\alpha}_{i\cdot})$ and a logit link. However, the theoretical results described in Section 3 will hold for any choice of $\ell(\cdot)$ that satisfies

$$E\Big[\sum_{i=1}^{n}\sum_{t=1}^{m_i} \nabla_{\boldsymbol{\beta}} \ell(Y_{it}, S_{it}, A_{it}; \boldsymbol{\beta}, \boldsymbol{\alpha}_{i\cdot}, \phi)\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0, \boldsymbol{\alpha}=\boldsymbol{\alpha}_0}\Big] = \mathbf{0} \quad \text{and}$$
$$E\Big[\sum_{t=1}^{m_i} \nabla_{\boldsymbol{\alpha}_{i\cdot}} \ell(Y_{it}, S_{it}, A_{it}; \boldsymbol{\beta}, \boldsymbol{\alpha}_{i\cdot}, \phi)\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0, \boldsymbol{\alpha}=\boldsymbol{\alpha}_0}\Big] = \mathbf{0} \quad \text{for}$$
$$i = 1, \ldots, n, \quad (3)$$

where $\boldsymbol{\alpha}_{i\cdot}$ is the ith row of $\boldsymbol{\alpha}$, ϕ is a nuisance parameter in the working log-likelihood, and $\nabla_{\boldsymbol{\beta}}\ell$ and $\nabla_{\boldsymbol{\alpha}_{i\cdot}}\ell$ denote the partial derivatives of ℓ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_{i\cdot}$, respectively. It is easy to verify that the Gaussian and the Bernoulli log-likelihoods satisfy (3); and since they are often the practical choices for continuous and binary outcomes, they may be used as pseudo-log-likelihood in many applications. Correspondingly, we define the penalized pseudo-log-likelihood

$$L_{ppl}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{t=1}^{m_i} \ell(Y_{it}, S_{it}, A_{it}; \boldsymbol{\beta}, \boldsymbol{\alpha}_{i\cdot}, \phi) - \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{\alpha}_{i\cdot}\boldsymbol{D}^{-}\boldsymbol{\alpha}_{i\cdot}^{\intercal} - \lambda\sum_{l=1}^{q} w_l\|\boldsymbol{\alpha}_{\cdot l}\|, \quad (4)$$

where $\boldsymbol{D} \in \mathbb{R}^{q\times q}$ is a symmetric positive semidefinite matrix, $\boldsymbol{D}^{-}$ is the Moore–Penrose generalized inverse of $\boldsymbol{D}$, w_l is the weight for the th random effects term and $\lambda \geq 0$ is a tuning parameter.

We propose to estimate $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ by maximizing (4). The maximum penalized-pseudo-likelihood estimator is denoted by

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\alpha}\in\mathbb{R}^{n\times q}} L_{ppl}(\boldsymbol{\beta}.\boldsymbol{\alpha}), \quad (5)$$

and the corresponding personalized policy for user i is estimated by

$$\hat{\pi}_i(s) \in \arg\max_{a\in\mathcal{A}} \left(h_1(s,a)^{\intercal}\hat{\boldsymbol{\beta}} + h_2(s,a)^{\intercal}\hat{\boldsymbol{\alpha}}_{i\cdot}^{\intercal}\right),$$

analogously to π_{0i} in (2).

The second term on the right-hand side of (4) puts a ridge-type penalty to shrink and stabilize the estimation of the random effects $\boldsymbol{\alpha} \in \mathbb{R}^{n\times q}$. Under the viewpoint that $\{\boldsymbol{\alpha}\}$ is a random sample of a population, it is natural to choose $\boldsymbol{D}$ to reflect the variance-covariance matrix of $\boldsymbol{\alpha}_{0i\cdot}^{\intercal}$, although it is not required for the asymptotic properties to hold (see Section 3). The third term in (4) is the group lasso penalty, where each group l contains the random effects parameter of the lth term in $h_2(S_{it}, A_{it})$ for all n users. Under a similar viewpoint, it is intuitive to set the group-specific weight $w_l \geq 0$ to be inverse proportional to the variance of $\boldsymbol{\alpha}_{il}$.

In practice, we propose to update $\boldsymbol{D}$, ϕ, and w_l's iteratively, in conjunction with the trust region Newton (TRON) algorithm in the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Briefly, the TRON algorithm combines the trust region method (Steihaug 1983) and the truncated Newton method (Nash 2000) to solve an unconstrained convex optimization problem. At each iteration, TRON defines a trust region and approximates the objective function using a quadratic model within the region. If a prespecified change of the objective function is achieved in the current iteration, the updated direction is accepted and the region is expanded; the region will be shrunk otherwise. The approximation subproblem is solved via the conjugate gradient method. Since TRON solves the inverse of a potentially large Hessian matrix by iteratively updating the parameters, convergence can be achieved quickly with a large and dense Hessian. Overall, the computational cost per iteration is of the order of the number of nonzero elements in the design matrix. In addition, we propose to choose the tuning parameter λ for the group lasso penalty using an AIC-type criterion. The details are given in Section S1 of the supplementary materials.

3. Theoretical Remarks

In this section, we study the asymptotic behaviors of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$, and conditional and marginal performance of estimated policies $\hat{\pi}_i$'s under the following assumptions. All proofs are given in Section S2 of the supplementary materials.

(C1) There exists a positive constant c_1, such that the treatment randomization probability $P(A_{it} = a_t | \underline{S}_{it} = \underline{s}_t, \underline{A}_{i,t-1} = \underline{a}_{t-1}) \geq c_1$ for all possible values of $(\underline{s}_t, \underline{a}_t)$ at any time point $t \geq 1$.

(C2) The random vectors $h_1(S_{it}, A_{it})$ and $h_2(S_{it}, A_{it})$ and outcome Y_{it} are square integrable under the data generative distribution for $t \geq 1$ and $i = 1, \ldots, n$.

(C3) The latent random effects $\boldsymbol{\alpha}_{0i\cdot}$, $i = 1, \ldots, n$, are independent and identically distributed with mean $\mathbf{0}$ and finite variance $\boldsymbol{\Sigma}$.

(C4) There exists $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$ such that (1) holds, and $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$ is $P_{\boldsymbol{\alpha}_0}$-almost surely an interior point of a compact set $\Omega \in \mathbb{R}^{p+nq}$.

(C5) The pseudo-log-likelihood $\ell(Y_{it}, S_{it}, A_{it}; \boldsymbol{\beta}, \boldsymbol{\alpha}_{i\cdot}, \phi)$ is concave in $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, satisfies condition (3), and its expected second-order derivative is continuous in $(\boldsymbol{\beta}, \boldsymbol{\alpha})$.

(C6) Denote $\ell_1(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{t=1}^{m_i} \ell(Y_{it}, S_{it}, A_{it}; \boldsymbol{\beta}, \boldsymbol{\alpha}_{i\cdot}, \phi)$. We need the following regularity conditions:

   (i) As $N := \sum_{i=1}^{n} m_i \to \infty$, $\boldsymbol{\beta}_0$ satisfies $N^{-1}E\left\{\left[\nabla_{\boldsymbol{\beta}}\ell_1(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)\right]^{\intercal}\nabla_{\boldsymbol{\beta}}\ell_1(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)\right\} = O(1)$; $\sup_{(\boldsymbol{\beta},\boldsymbol{\alpha})\in\Omega}\left\|N^{-1}\nabla_{\boldsymbol{\beta}}^{\intercal}\nabla_{\boldsymbol{\beta}}\ell_1(\boldsymbol{\beta},\boldsymbol{\alpha}) - E\left\{N^{-1}\nabla_{\boldsymbol{\beta}}^{\intercal}\nabla_{\boldsymbol{\beta}}\ell_1(\boldsymbol{\beta},\boldsymbol{\alpha})|\boldsymbol{\alpha}_0\right\}\right\|_{\mathrm{F}} = o_P(1)$ $P_{\boldsymbol{\alpha}_0}$-almost surely, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm; and $\boldsymbol{M}_{\boldsymbol{\beta}\boldsymbol{\beta}} \triangleq -\liminf_{N\to\infty} E\left\{N^{-1}\nabla_{\boldsymbol{\beta}}^{\intercal}\nabla_{\boldsymbol{\beta}}\ell_1(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)\right\}$ is positive definite with all eigenvalues greater than $\delta_0 > 0$.

   (ii) $\sup_i m_i^{-1}E\left\{\nabla_{\boldsymbol{\alpha}_{i\cdot}}\ell_1(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)^{\intercal}\nabla_{\boldsymbol{\alpha}_{i\cdot}}\ell_1(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)\right\} = O(1)$.

   (iii) For $i = 1, \ldots, n$, as $m_i \to \infty$, $\boldsymbol{\alpha}_{0i\cdot}$ satisfies $\sup_{(\boldsymbol{\beta},\boldsymbol{\alpha})\in\Omega}\|m_i^{-1}\nabla_{\boldsymbol{\alpha}_{i\cdot}}^{\intercal}\nabla_{\boldsymbol{\alpha}_{i\cdot}}\ell_1(\boldsymbol{\beta},\boldsymbol{\alpha}) - E\{m_i^{-1}\nabla_{\boldsymbol{\alpha}_{i\cdot}}^{\intercal}\nabla_{\boldsymbol{\alpha}_{i\cdot}}\ell_1(\boldsymbol{\beta},\boldsymbol{\alpha})|\boldsymbol{\alpha}_0\}\|_{\mathrm{F}} = o_P(1)$ $P_{\boldsymbol{\alpha}_0}$-almost surely; and $\boldsymbol{M}_{\boldsymbol{\alpha}_{i\cdot}\boldsymbol{\alpha}_{i\cdot}} \triangleq -\liminf_{m_i} E\{m_i^{-1}\nabla_{\boldsymbol{\alpha}_{i\cdot}}^{\intercal}\nabla_{\boldsymbol{\alpha}_{i\cdot}}\ell_1(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_0)\}$ is positive definite with all eigenvalues greater than $\delta_0 > 0$.

(C7) The weights satisfy $\max_{\{l\in\{1,\ldots,q\}:\sigma_l^2>0\}} |w_l| = O_P(1)$, where σ_l^2 is the lth diagonal element of $\boldsymbol{\Sigma}$, the variance-covariance matrix of $\boldsymbol{\alpha}_{0i\cdot}$.

(C8) The tuning parameter λ satisfies $\lambda = o\left\{n^{-1}(\sum_{i=1}^{n} m_i^{-1})^{1/2}\right\}$.

Theorem 1. Suppose Assumptions C1-C8 hold. As $n, \min_i\{m_i\} \to \infty$, $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ satisfies $||\hat{\boldsymbol{\beta}}_{\lambda} - \boldsymbol{\beta}_0|| = O_p(N^{-1/2})$ and $||\hat{\boldsymbol{\alpha}}_{\lambda i} - \boldsymbol{\alpha}_{0i\cdot}|| = O_P(m_i^{-1/2})$, $i = 1, \ldots, n$.

Remarks. Condition (C6) is similar to the regularity conditions required in maximum likelihood estimation. In particular, when the covariates S_{it}'s are exogenous, it is easy to verify that (C6) holds under the regularity conditions used in GLMM. Interestingly, Condition (C6) also holds under many situations when S_{it}'s are endogenous; and importantly, these situations can be verified. For illustration purposes, we verify this condition in the Appendix in two quite common scenarios: (i) when Y_{it} is binary and the distribution of S_{it} directly depends on the latent random effects $\boldsymbol{\alpha}_{0i\cdot}$; (ii) when Y_{it} follows Gaussian distribution and $S_{it} = Y_{i,t-1}$.

Theorem 1 characterizes the asymptotic behavior of every $\hat{\boldsymbol{\alpha}}_{i\cdot}$ under the condition that $\min_i m_i \to \infty$. This condition, however, can be relaxed if we are only interested in the asymptotic behavior of $\hat{\boldsymbol{\alpha}}_{i\cdot}$ on average. Specifically, we only require that the proportion of m_i's that do not go to infinity goes to zero. Without loss of generality, suppose $m_1 \le m_2 \le \cdots \le m_n$. Let k_n be the index so that m_{k_n} is bounded, and $m_{k_n+1} \to \infty$.

Corollary 1. Suppose (C6)(iii) holds for $i = k_n + 1, \ldots, n$, and the remaining assumptions in C1–C8 continue to hold. As $n, \min_{i>k_n}\{m_i\} \to \infty$, suppose $k_n/n \to 0$. Then, $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| = O_P(N^{-1/2})$ and

$$\frac{1}{n}\sum_{i=1}^{n} \|\hat{\boldsymbol{\alpha}}_{i\cdot} - \boldsymbol{\alpha}_{0i\cdot}\|^2 = O_P\left(\frac{k_n}{n} + \frac{1}{n}\sum_{i=k_n+1}^{n} m_i^{-1}\right).$$

Next, we present the properties of the estimated personalized policies $\hat{\pi}_i$. Specifically, we consider both the conditional expected outcome under $\hat{\pi}_i$ at each time point t given $S_{it} = s_t$, and the marginal expected outcome assuming $\hat{\pi}_i$ is used to make decision for user i from the beginning to time point t. The results are stated in the theorem below.

Theorem 2. Assume all conditions in Corollary 1 hold. Suppose the inverse link function of the corresponding exponential family distribution, $g^{-1}(\cdot)$, is Hölder continuous. That is, for any η_1, η_2 in the domain, $|g^{-1}(\eta_1) - g^{-1}(\eta_2)| \le L|\eta_1 - \eta_2|^{\gamma}$, where L is a positive constant and $0 < \gamma \le 1$. Then for any $t \ge 1$, as $n, \min_{i>k_n}\{m_i\} \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n}[E_{\pi_{0i}}(Y_{it}|S_{it} = s_t, \boldsymbol{\alpha}_{0i\cdot}) - E_{\hat{\pi}_i}(Y_{it}|S_{it} = s_t, \boldsymbol{\alpha}_{0i\cdot})]$$
$$= O_P\left(\left[\frac{k_n}{n} + \frac{1}{n}\sum_{i=k_n+1}^{n} m_i^{-1}\right]^{\gamma/2}\right). \quad (6)$$

In addition, assume

$$P_{\boldsymbol{\alpha}_{0i\cdot}}\left[\left\{\boldsymbol{\alpha}_{0i\cdot} : P\left(\arg\max_{a\in\mathcal{A}} Q(S_{i,t-1}(\underline{\pi}_{0i,t-2}), a; \boldsymbol{\beta}_0, \boldsymbol{\alpha}_{0i\cdot}) \text{ is unique } \middle| \boldsymbol{\alpha}_{0i\cdot}\right) = 1\right\}\right] = 1, \quad (7)$$

where $S_{i,t-1}(\underline{\pi}_{0i,t-2})$ is the potential outcome of $S_{i,t-1}$ that would have been observed were π_{0i} used to make decision up to time point $t-2$. Then we have,

$$\frac{1}{n}\sum_{i=1}^{n} \left|E_{\pi_{0i}}(Y_{it}|\boldsymbol{\alpha}_{0i\cdot}) - E_{\hat{\pi}_i}(Y_{it}|\boldsymbol{\alpha}_{0i\cdot})\right| = o_P(1). \quad (8)$$

Remarks.

1. Equation (6) measures the difference in conditional expected immediate outcome given current state between π_{0i} and the estimated policies $\hat{\pi}_i$'s. The personalized policy π_{0i} is optimal in the conditional sense, in that it yields the maximal expected outcome if treatment assignment A_{it} is consistent with π_{0i} given S_{it}. As such, Equation (6) describes the conditional optimality of estimated policies $\hat{\pi}_i$'s given the current information.
2. We note that π_{0i} may not necessarily be optimal in a marginal sense after integrating out S_{it}, because the distribution of S_{it} may depend on previous treatment assignment. Therefore, the marginal result equation (8) implies consistency rather than optimality.
3. Condition (7) implies that the optimal decision at time $t-1$ is unique almost surely, given that π_{0i} were used to make decision at previous time points. This assumption is not needed to show consistency when $t = 1$.

4. Simulation Study

4.1. Setup

In this section, we examine the estimation properties of the maximum penalized-pseudo-likelihood estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ in (5) and the performance of the personalized policy $\hat{\pi}_i$ using simulation.

In a simulated trial, each user would be followed for $m = 10, 20, 30$ time points for training purposes, with 10 additional subsequent testing time points. At time point t, user i would receive one of three possible actions with equal probability, that is, the actions were generated randomly with probabilities $(1/3, 1/3, 1/3)$; the actions A_{it}'s were then coded using two dummy variables and were centered. The covariate process $S_{it} = (X_{it}, t)$ included a binary endogenous variable $X_{it} \in \{-1, 1\}$, which would depend on the previous outcome $Y_{i,t-1}$, the previous action $A_{i,t-1}$ and the random effects $\boldsymbol{\alpha}$. Specifically, we set $P(X_{i1} = 1) = \alpha_{0i0}$, and

$$\begin{aligned} P(X_{it} &= 1|A_{i,t-1}, S_{i,t-1}, \boldsymbol{\alpha}_{0i\cdot}) \\ &= \text{expit}((-3Y_{i,t-1} + 2X_{i,t-1} - A_{i,t-1})/ \\ &\quad 10 + \alpha_{0i4} - \alpha_{0i5} + \alpha_{0i6} - \alpha_{0i7}), \end{aligned}$$

for $t \ge 2$, where expit$(\cdot)$ is the expit function, $\alpha_{0i0} \sim U(0, 1)$, and α_{0ij} is the jth component of $\boldsymbol{\alpha}_{0i\cdot}$ for $j = 1, \ldots, q$. We considered both binary and continuous outcomes. The conditional mean of the outcome was defined according to (1) where $h_1(S_{it}, A_{it}) = h_2(S_{it}, A_{it}) = (1, S_{it}, A_{it}, S_{it} \otimes A_{it})$ and $\otimes$ denotes the Kronecker product, with logit and identity links, respectively, for the binary and continuous outcomes. The continuous outcomes were generated with an independent Gaussian noise with standard deviation 1.5. The true fixed effects were specified by

$$\boldsymbol{\beta}_0 = (-1, 0.2, -1.5, 0.8, 0.7, 0.1, 0.2, -1.2, -1.4)^{\mathsf{T}}.$$

We considered two scenarios for the random effects $\boldsymbol{\alpha}_0$, which were generated from mean zero Gaussian: we set variance-covariance matrix to be diag(2, 0.1, 0.1, 3, 4, 4, 5, 10, 12) to represent a scenario with non-sparse random effects, and diag(2, 0.1, 0.1, 3, 0, 0, 5, 10, 12) with sparse random effects. We generated 200 simulated trials, each having $n = 50$ users. Once the random effects were sampled, they were treated as fixed parameters in the 50 users.

The estimation properties of the policy parameters based on the training data were evaluated using mean squared error (MSE), defined as $\sum_{i=1}^{n} ||\hat{\boldsymbol{\beta}}(\pi) + \hat{\boldsymbol{\alpha}}_{i\cdot}^{\mathsf{T}}(\pi) - (\boldsymbol{\beta}_0(\pi) + \boldsymbol{\alpha}_{0i\cdot}^{\mathsf{T}}(\pi))||_2^2/(n \times \dim(\boldsymbol{\beta}_0(\pi)))$, where $\boldsymbol{\beta}(\pi)$ is the sub-vector of $\boldsymbol{\beta}$ involved in policy π (i.e., coefficients of A_{it} and $S_{it} \otimes A_{it}$). The quality of decisions at the testing time points by the estimated policies was evaluated in terms of the expected conditional outcome under $\hat{\boldsymbol{\pi}} = \{\hat{\pi}_i\}$ at each testing time point t:

$$V^{\hat{\boldsymbol{\pi}}}(s_t) \triangleq \frac{1}{n}\sum_{i=1}^{n} E^{\hat{\pi}_i}(Y_{it}|S_{it} = s_t)$$
$$= \frac{1}{n}\sum_{i=1}^{n} Q(S_{it} = s_t, A_{it} = \hat{\pi}_i(s_t); \boldsymbol{\beta}_0, \boldsymbol{\alpha}_{0i\cdot}),$$

$t = m+1, m+2, \ldots, m+10$. To facilitate comparison across scenarios, we standardized the expected outcome against the optimal policy $\boldsymbol{\pi}_0 = \{\pi_{0i}\}$ and the worst policy $\boldsymbol{\pi}_{\text{worst}} = \{\pi_{\text{worst},i}\}$ and obtained the value ratio (VR) for the estimated policy $\hat{\boldsymbol{\pi}}$:

$$\text{VR}^{\hat{\boldsymbol{\pi}}}(s_t) = \frac{V^{\hat{\boldsymbol{\pi}}}(s_t) - V^{\boldsymbol{\pi}_{\text{worst}}}(s_t)}{V^{\boldsymbol{\pi}_0}(s_t) - V^{\boldsymbol{\pi}_{\text{worst}}}(s_t)}.$$

4.2. Comparison Methods

In the simulation, we considered some existing methods as alternatives to the proposed personalized policy learning method, which shall be denoted as PPL in the followings.

Under the GLMM framework, instead of using the proposed algorithm described in Section 2.2, we used the "glmer" function in the lme4 package in R (Bates et al. 2014). This method shall be denoted as glmer. The function "glmer" would involve approximating the marginal likelihood by integrating over the random effects. This could be problematic in situations with a large number of random effects (thus having a high-dimensional integrals) and endogenous covariates.

In addition, we considered the regularized penalized quasi-likelihood (rPQL) approach developed by Hui, Müller, and Welsh (2017) for exogenous covariates as yet another alternative to estimating $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ under the GLMM framework. While rPQL also imposed a group lasso penalty, our proposed algorithm took a different computational approach: First, we adopted the novel trust region method to solve the optimization problem; second, we updated the weights w_l's iteratively whereas rPQL would keep the weights at their initial values throughout the computation.

While the methods above would prescribe personalized policies, we also considered using generalized estimating equations (GEE) to estimate a population-level effect, and developed a non-personalized policy by choosing actions maximizing the estimated population mean. We used an independence working correlation structure, so as to avoid bias under linear models with endogenous variables; see Boruvka et al. (2018).

Finally, we examined the performance of an "N-of-1" approach whereby each user's personalized policy was estimated by fitting a generalized linear model to the user's own data only. That is, there was no borrowing information from across users in this method with multiple generalized linear model (MGLM). We anticipated that MGLM would have difficulties when m was small, especially with Bernoulli outcomes.

4.3. Simulation Results

Table 1 compares the MSE of the policy parameters in the simulation scenario with non-sparse random effects. Overall, the proposed PPL has the smallest MSE when $m = 20, 30$. Its superior performance to the other two GLMM-based methods (glmer and rPQL) indicates the computational advantages of using the trust region algorithm with iterated weights. These three methods, as expected, improve with large m, that is, having more data points.

The "N-of-1" MGLM performs poorly with binary outcome and when $m = 10$ with continuous outcome. Even with a moderate-to-large $m = 30$, the method remains inferior to the other methods. This signifies the importance of borrowing information from across users, even though our goal is to produce different policies for different users.

Interestingly, GEE has the smallest MSE when $m = 10$ and performs relatively well with the larger m's. While it is somewhat surprising at first glance, we note that by avoiding estimating the random effects ($\boldsymbol{\alpha}$ is estimated with $\mathbf{0}$), GEE will induce the least variability and hence the MSE. It is illuminating that the method's MSE does not improve as m increases, when bias becomes dominating in the bias-variance tradeoff.

Table 2 compares the methods under the scenario with sparse random effects. The relative performance of the methods is similar to that in Table 1, although the bias induced by GEE becomes more apparent as the variability in the data is smaller in this scenario. In particular, PPL and glmer has substantially smaller MSE in this scenario than when random effects are not as sparse.

To compare the decision quality of the five methods, Figures 1 and 2 plot the simulated mean VR at the testing time points following $m = 10$ training time points from each user, respectively, under non-sparse random effects and sparse random effects.

The proposed PPL has the largest VR for each possible state X_t for both binary and continuous outcomes. That GEE

Table 1. Estimation properties under scenario with non-sparse random effects (average MSE (SD) over 200 simulation trials).

	Binary			Continuous		
Method	$m = 10$	$m = 20$	$m = 30$	$m = 10$	$m = 20$	$m = 30$
PPL	8.22(3.31)	5.41(0.61)	4.70(0.58)	8.99(3.67)	3.69(0.47)	2.37(0.22)
glmer	43.39(34.16)	9.65(2.72)	6.39(1.17)	13.94(4.79)	4.87(0.65)	3.03(0.38)
GEE	7.87(2.92)	6.08(0.57)	6.17(0.52)	8.38(2.83)	5.94(0.38)	5.74(0.21)
MGLM	>1E10	>1E10	>1E10	272.35(93.00)	36.24(35.00)	8.10(4.33)
rPQL	8.71(3.9)	5.87(0.73)	5.23(0.65)	7.73(2.32)	5.31(0.30)	4.44(0.26)

Table 2. Estimation properties under scenario with sparse random effects (average MSE (SD) over 200 simulation trials).

	Binary			Continuous		
Method	$m = 10$	$m = 20$	$m = 30$	$m = 10$	$m = 20$	$m = 30$
PPL	7.75(3.57)	4.41(0.80)	3.69(0.69)	7.17(3.09)	2.44(0.46)	1.33(0.20)
glmer	44.56(38.85)	8.80(2.50)	5.73(1.50)	11.75(4.08)	3.48(0.64)	1.86(0.33)
GEE	7.24(3.06)	5.04(0.71)	5.11(0.64)	7.15(2.98)	4.85(0.37)	4.67(0.23)
MGLM	>1E10	>1E10	>1E10	274.92(105.00)	34.73(31.90)	7.19(1.78)
rPQL	8.04(3.86)	4.89(0.92)	4.22(0.82)	6.31(1.96)	4.42(0.40)	3.08(0.32)

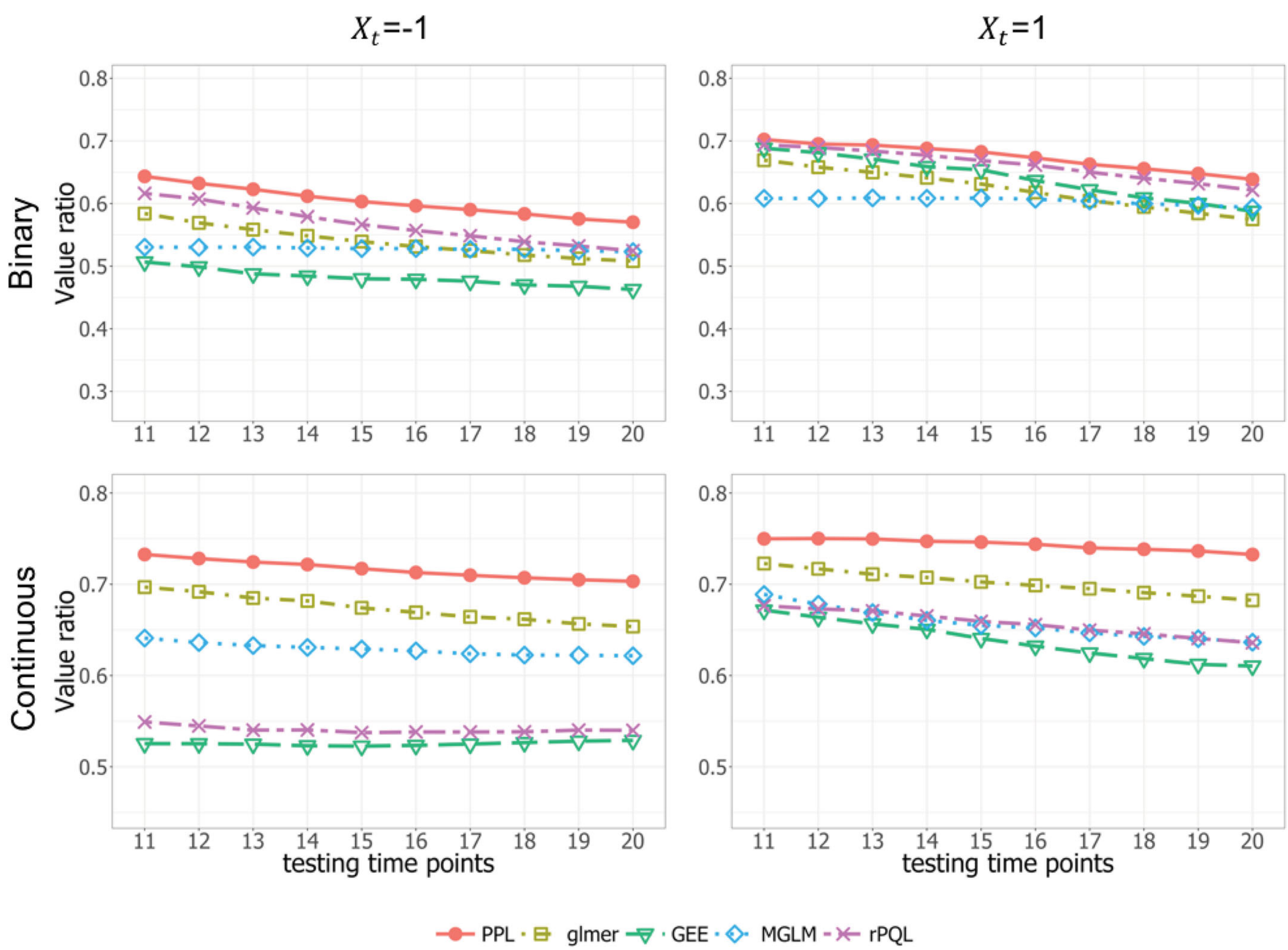

Figure 1. Value ratio at each testing time point in the simulation with $m = 10$ under scenario with non-sparse random effects by different X_t.

producing the smallest MSE when $m = 10$ does not translate into good decision quality, as the method has the smallest VR uniformly in our simulation, when compared to all other personalized policy methods. This serves as an important illustration how simply considering personalized policy, as opposed to personalized decisions (which GEE also prescribes), could lead to potentially radical gain. It is interesting to note that methods that induce large variability in estimation can be quite competitive; for example, MGLM and glmer for continuous outcome when $X_t = 1$. It is due to the fact that the decision quality largely relies on correctly estimating the sign of the random effects, not the magnitude. Therefore, one ought to examine both the estimation properties and decision quality in the comparison of methods. Overall, our simulation results indicate the proposed PPL win in these terms. The relative performance of the methods is similar when $m = 20, 30$, and the results are presented in Section S3 of the supplementary materials.

5. Application

We apply the proposed PPL to estimate the best personalized push schedule in 294 users, who have received at least 20 prompts to complete the patient-health questionnaire since they downloaded the Hub app. Since the prompts were scheduled on 7-day intervals, this would represent a subsample of users with at least 20 weeks of app use. The distribution of the number of prompts in these users is shown in Figure 3. In the data, we tracked the timestamp of when a prompt was sent. For the purpose of this analysis, we grouped the time of prompt into four periods: Night (a_1): from midnight to 6:00 a.m.; Morning

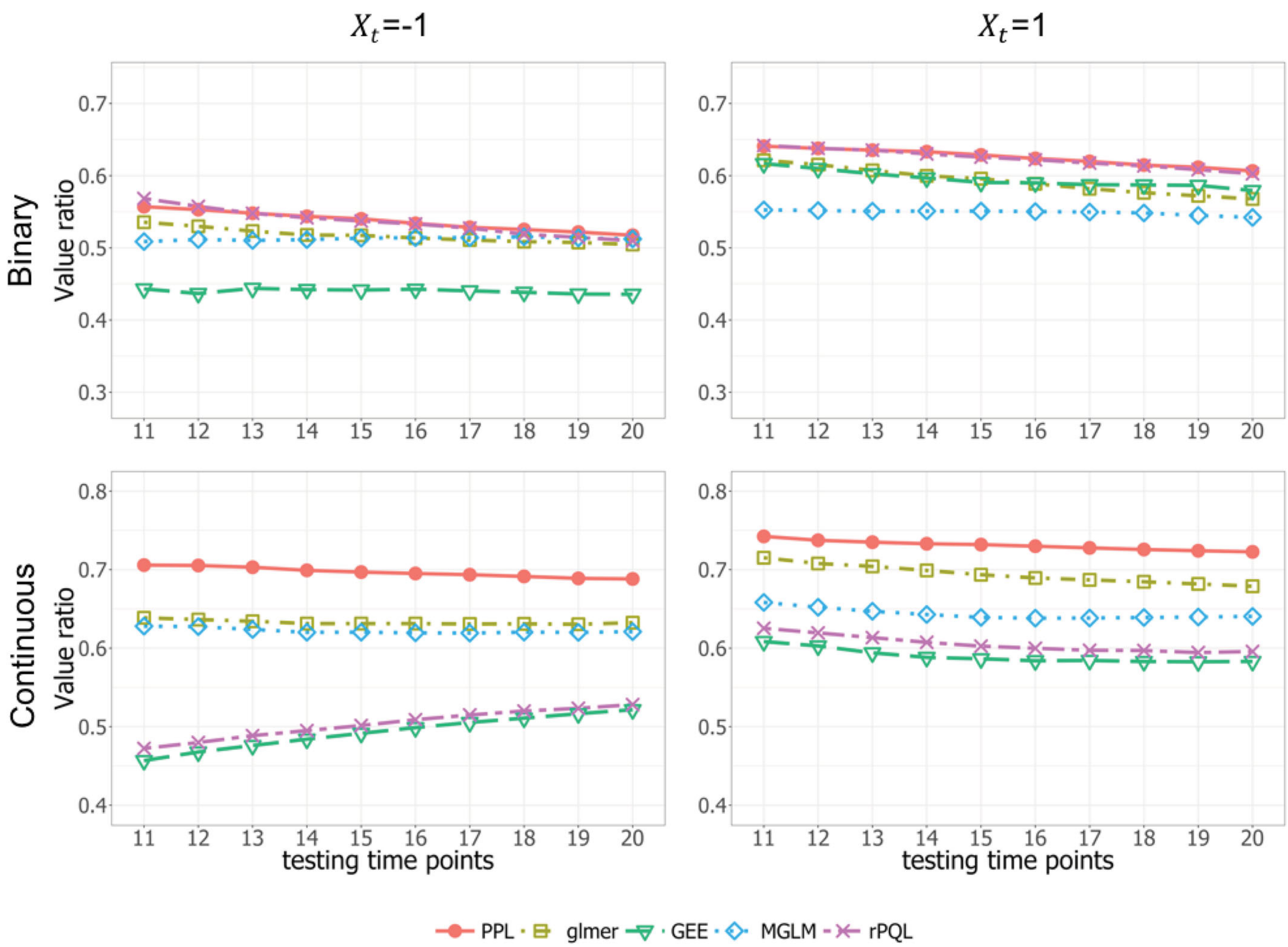

Figure 2. Value ratio at each testing time point in the simulation with $m = 10$ under scenario with sparse random effects by different X_t.

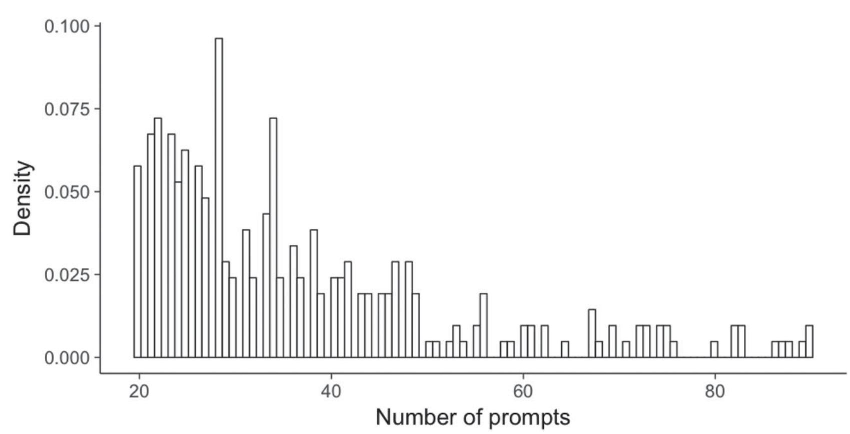

Figure 3. The distribution of the number of prompts in 294 IntelliCare users.

(a_2): from 6:00 a.m. to noon; Afternoon (a_3): from noon to 6:00 p.m.; Evening (a_4): from 6:00 p.m. to midnight. The observed proportions of the four periods were, respectively, 0.10, 0.23, 0.35, and 0.32. Using a_1 as the reference group, we used three dummy variables, centered by the observed proportions, to code the actions a_2, a_3, and a_4 in model fitting.

The state S_{it} at each time point consisted of three variables. First, the number of times the Hub was launched (*launches*) in the week prior to the prompt was recorded. Second, the timestamp indicated whether a prompt was sent on a weekday (*weekday*). Third, the time point t of the prompt was included as a predictor in the covariate process S_{it}. With a binary response outcome, we estimated $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ under model (1) with a logit link, $h_1(S, A) = (1, launches, t, weekday, A, launches \otimes A, weekday \otimes A, t \otimes A)$ and $h_2(S, A) = (1, A, launches \otimes A, weekday \otimes A, t \otimes A)$ using the first 80% of the time points of each user as training data. Since each user had at least 20 prompts, we had $m_i \geq 16$ in the training data for all 294 users.

Table 3. Model fit using the training data: $\hat{\boldsymbol{\beta}}$ is the coefficients of the fixed effects, and SD($\hat{\boldsymbol{\alpha}}$) is the standard deviation of the fitted individual random effects coefficients.

Variables	$\hat{\boldsymbol{\beta}}$	SD($\hat{\boldsymbol{\alpha}}$)
Intercept	−1.80	1.31
weekday	0.01	—
launches (per 5 times)	1.52	—
t (per 5 time points)	−0.20	—
Morning (a_2)	1.65	1.13
Afternoon (a_3)	1.57	0.95
Evening (a_4)	1.06	0.78
weekday : a_2	0.73	0.34
weekday : a_3	0.16	0.62
weekday : a_4	0.66	0.52
launches : a_2	−2.46	0.39
launches : a_3	−1.40	0.21
launches : a_4	−1.15	0.43
$t : a_2$	−1.25	0.66
$t : a_3$	−0.96	0.47
$t : a_4$	−0.93	0.44

Table 3 summarizes the results of the model fit. The positive fixed effects for a_2, a_3, a_4 suggest prompts in the morning, afternoon, and evening tend to induce better response rate than those sent during the night (midnight to 6:00 a.m.). The effects associated with these non-night periods are even greater on weekdays, indicated by the positive (fixed) interaction between weekday and these periods. While this result is not surprising, we also note substantial heterogeneity of the period effects and the *weekday*:period interactions, whose SD($\hat{\boldsymbol{\alpha}}$)'s have comparable magnitude to $\hat{\boldsymbol{\beta}}$. This supports the needs for personalizing push schedule in our application.

In contrast, for the *launches*:period interactions and the t:period interactions, the fixed effects ($\hat{\boldsymbol{\beta}}$) dominate the random effects; heterogeneity of the random effects coefficients are measured by SD($\hat{\boldsymbol{\alpha}}$). Based on the fixed effects, the response rate decreases over time, by 0.20 in log-odds over $t = 5$ time points. This is in line with findings in the literature (see, e.g., Helander et al. 2014). In addition, every five additional *launches* of the Hub in the prior week improves the log-odds of response to a night prompt by 1.52. Based on the negative coefficients of *launches*:period interactions, a large number of *launches* also seems to attenuate or even negate the effects of the time of prompts. This suggests that for active users who engage the Hub often, their response pattern is less sensitive to the time of the prompt.

The quality of these personalized policies in the testing data is evaluated by the mean response rate under the policies estimated via inverse probability treatment weighted method averaged over all test time points. The mean response rate according to PPL would have been 23%, which compares favorably to other studies in light of the fact that all testing points are at least 16 weeks from first download. It has been reported that user engagement is in the range of 3%–15% in the third month after download (Helander et al. 2014). As a reference point, the observed response rate in the testing data is 11%. In addition, we analyzed the prompt response data using the other methods with the same 80%–20% split of training and testing data, and obtained the mean response rate 14%, 17%, 14%, and 8%, respectively, for glmer, GEE, MGLM, and rPQL.

6. Discussion

This article proposed PPL for mobile health applications under the GLMM framework. The advantage of using GLMM is that, when data from each user are limited, it allows for estimation of the shared features (i.e., fixed effects) as well as the individual features (i.e., random effects), two indispensable components of a personalized policy. We have shown that personalized policies lead to higher value than non-personalized policy in simulation studies, and have clearly demonstrated substantial heterogeneity of the action effects in the prompt response data. We have also provided theoretical justifications of PPL by examining its asymptotic properties under a fairly general set of assumptions. In particular, we have established a conditional regret bound and marginal consistency in the presence of endogenous covariate process, where the covariates may depend on previous outcomes, actions, and even the latent random effects. As endogeneity is ubiquitous in longitudinal mobile application usage (how many times a user launched the Hub app would likely depend on how he/she had interacted with the Hub in the past), these theoretical results have broadened the applicability of PPL to many practical situations.

As an initial attempt to tackle personalized policy learning problem, we have limited our attention to optimize immediate outcomes. This can be viewed as an RL problem with a discount factor of zero. Our choice to maximize immediate outcomes is primarily based on the unique nature of mobile health study where adherence declines drastically. On the other hand, when adequate engagement can be achieved, say, through education or compensation, a long-term outcome would also be desirable. To optimize long-term outcomes, we need to take into account the delayed effect of current action on future contexts and outcomes. In RL, a standard approach is to treat each user's longitudinal data as a Markovian decision process (MDP). An interesting research direction is to extend the GLMM framework to estimate the MDP parameters to learn personalized policies that target on long-term outcomes.

This article investigated learning of optimal personalized policies for existing users after actions have been observed in batched data. Our ultimate goal is to create an evidence-based mobile app intervention delivery system in which a key element is to set up a mechanism that provides feedback so as to enhance decision quality for existing and new users over time based on data previously accrued. The proposed PPL algorithm, coupling with the adaptive randomization (AR), can be used to build an online decision support system. Briefly speaking, the proposed PPL algorithm can be used to construct an initial personalized policy for each existing user. Using this as a warm start, we can then tilt the allocation of empirically superior interventions (actions) to the user with high probabilities using ϵ-greedy or Thompson-type AR scheme. The policy can be further updated using PPL algorithm as more data are collected. Note that personalized algorithms depend on individual-level parameters, as well as the population-level parameters. For a new user, an initial personalized policy can be obtained by randomly drawing an individual-level parameter from the existing pool. The principle of specifying an AR scheme can be extended similarly. To implement AR in practice, we need to keep track and model the enrollment time of mobile app users. Cheung, Chakraborty, and

Davidson (2015) introduced the implementation of real-time AR in the context of a population multi-stage policy setting. A future research direction is to incorporate real-time AR in PPL framework in the mobile decision support system.

Appendix: Examples of Endogenous Covariates

In this section, we verify condition (C6) in two examples with endogenous covariates. In the first example, Y_{it} is binary , and the distribution of S_{it} directly depends on the random effects parameters $\boldsymbol{\alpha}_{0i\cdot}$. In the second example, Y_{it} is Gaussian, and $S_{it} = Y_{i,t-1}$. For simplicity, we verify the condition with $n = 1$ (since individuals are iid), and omit subscript i from the notations. In both examples, we consider a scalar mean zero random effects parameter $\alpha_0 \in \mathbb{R}$, and the treatment $A_t \in \{-1, 1\}$ is randomly assigned with $P(A_t = 1) = P(A_t = -1) = 1/2$ for $t \geq 1$.

Example 1. For binary outcome $Y_t \in \{0, 1\}$, suppose

$$g(E(Y_t|S_t, A_t)) = (\beta_0 + \alpha_0)S_tA_t,$$

where $g(\cdot)$ is the logit link. Conditioning on α, $S_t, t = 1, \dots, m$, are iid $N(\alpha_0, \tau^2)$.

Let $\ell(Y_t, S_t, A_t; \beta, \alpha)$ be the log-likelihood of Bernoulli distribution with mean $\frac{e^{(\beta+\alpha)S_tA_t}}{1+e^{(\beta+\alpha)S_tA_t}}$. Then $\ell_1(\beta, \alpha) = \sum_{t=1}^m \ell(Y_t, S_t, A_t; \beta, \alpha)$ satisfies

$$\ell_1(\beta, \alpha) = \sum_{t=1}^m Y_t(\beta + \alpha)S_tA_t - \log\left\{1 + e^{(\beta+\alpha)S_tA_t}\right\},$$

$$\nabla_\beta \ell_1(\beta, \alpha) = \nabla_\alpha \ell_1(\beta, \alpha) = \sum_{t=1}^m S_tA_t\left\{Y_t - \frac{e^{(\beta+\alpha)S_tA_t}}{1 + e^{(\beta+\alpha)S_tA_t}}\right\},$$

and

$$\nabla_\beta^\intercal \nabla_\beta \ell_1(\beta, \alpha) = \nabla_\alpha^\intercal \nabla_\alpha \ell_1(\beta, \alpha) = -\sum_{t=1}^m \frac{S_t^2 e^{(\beta+\alpha)S_tA_t}}{\left\{1 + e^{(\beta+\alpha)S_tA_t}\right\}^2}.$$

It is easy to verify that

$$E\left\{\nabla_\beta \ell_1(\beta_0, \alpha_0)\right\} = E\left\{\nabla_\alpha \ell_1(\beta_0, \alpha_0)\right\} = 0,$$

$$\frac{1}{m}E\left[\left\{\nabla_\beta \ell_1(\beta_0, \alpha_0)\right\}^2\right] = \frac{1}{m}E\left[\{\nabla_\alpha \ell_1(\beta_0, \alpha_0)\}^2\right]$$
$$= \frac{1}{m}\sum_{t=1}^m E\left\{\frac{S_t^2 e^{(\beta_0+\alpha_0)S_tA_t}}{\left\{1 + e^{(\beta_0+\alpha_0)S_tA_t}\right\}^2}\right\} = E\left\{\frac{S_1^2 e^{(\beta_0+\alpha_0)S_1A_1}}{\left\{1 + e^{(\beta_0+\alpha_0)S_1A_1}\right\}^2}\right\} < \infty,$$

and

$$\boldsymbol{M}_{\beta\beta} = \boldsymbol{M}_{\alpha\alpha} = E\left\{\frac{S_1^2 e^{(\beta_0+\alpha_0)S_1A_1}}{\left\{1 + e^{(\beta_0+\alpha_0)S_1A_1}\right\}^2}\right\} > 0.$$

Finally, conditioning on α_0, $\left\{\frac{S_t^2 e^{(\beta+\alpha)S_tA_t}}{\{1+e^{(\beta+\alpha)S_tA_t}\}^2}, t = 1, \dots, m\right\}$ are iid, where $\alpha = \alpha_0 + u$. By the uniform law of large numbers theorem,

$$\sup_{(\beta,\alpha)\in\Omega}\left|m^{-1}\nabla_\beta^\intercal \nabla_\beta \ell_1(\beta, \alpha) - E\left\{\frac{S_1^2 e^{(\beta+\alpha)S_1A_1}}{\left\{1 + e^{(\beta+\alpha)S_1A_1}\right\}^2}\middle|\alpha_0\right\}\right| = o_P(1).$$

Example 2. Suppose

$$Y_t|S_t, A_t, \beta_0, \alpha_0 \sim N\{(\beta_0 + \alpha_0)S_tA_t, \sigma^2\},$$

where $S_t = Y_{t-1}$, and $Y_0 \equiv \mu_0$ is a constant. We consider $l(\cdot)$ to be the log-likelihood of Gaussian distribution. Below we show that condition (C6) holds when

$$P_{\alpha_0}(|\beta_0 + \alpha_0| < 1) = 1. \tag{9}$$

Note that condition (9) is a sufficient condition for an AR(1) process to be stationary.

$$\ell_1(\beta, \alpha) = -\frac{1}{2\sigma^2}\sum_{t=1}^m \left\{Y_t - (\beta + \alpha)Y_{t-1}A_t\right\}^2,$$
$$\nabla_\beta \ell_1(\beta, \alpha) = \nabla_\alpha \ell_1(\beta, \alpha)$$
$$= \frac{1}{\sigma^2}\sum_{t=1}^m \left\{Y_t - (\beta + \alpha)Y_{t-1}A_t\right\} Y_{t-1}A_t,$$

and

$$\nabla_\beta^\intercal \nabla_\beta \ell_1(\beta, \alpha) = \nabla_\alpha^\intercal \nabla_\alpha \ell_1(\beta, \alpha) = -\frac{1}{\sigma^2}\sum_{t=1}^m Y_{t-1}^2.$$

We can verify that

$$E\left\{\nabla_\beta \ell_1(\beta_0, \alpha_0)\right\} = E\left\{\nabla_\alpha \ell_1(\beta_0, \alpha_0)\right\} = 0.$$

Noticing that $E(Y_t^2) = \sum_{k=1}^{t-1} \sigma^2 E\left\{(\beta_0 + \alpha_0)^{2(k-1)}\right\}$, we have,

$$\frac{1}{m}E\left[\left\{\nabla_\beta \ell_1(\beta_0, \alpha_0)\right\}^2\right] = \frac{1}{m}E\left[\left\{\nabla_\alpha \ell_1(\beta_0, \alpha_0)\right\}^2\right]$$
$$= \frac{1}{m\sigma^2}\sum_{t=1}^m E(Y_t^2) = \frac{1}{m}\sum_{t=1}^m\sum_{k=1}^{t-1} E\left\{(\beta + \alpha)^{2(k-1)}\right\}$$
$$= \frac{1}{m}\sum_{t=1}^m (m - t + 1)E\left\{(\beta_0 + \alpha_0)^{2(t-1)}\right\}$$

which is $O(1)$ when (9) holds.

Since $(m - t + 1)E\left\{(\beta_0 + \alpha_0)^{2(t-1)}\right\} = m$ when $t = 1$, we have

$$\boldsymbol{M}_{\beta\beta} = \boldsymbol{M}_{\alpha\alpha} = \frac{1}{m}\sum_{t=1}^m (m - t + 1)E\left\{(\beta_0 + \alpha_0)^{2(t-1)}\right\} \geq 1.$$

Finally, under condition (9), for any $\epsilon > 0$,

$$P\left\{\sup_{(\beta,\alpha')\in\Omega}\left|\frac{1}{m}\sum_{t=1}^m Y_t^2 - E(Y_t^2|\alpha_0)\right| > \epsilon\middle|\alpha_0\right\}$$
$$= P\left\{\left|\frac{1}{m}\sum_{t=1}^m Y_t^2 - E(Y_t^2|\alpha_0)\right| > \epsilon\middle|\alpha_0\right\}$$
$$\leq (\epsilon m)^{-2}E\left[\sum_{t=1}^m\left\{Y_t^2 - E(Y_t^2|\alpha_0)\right\}\middle|\alpha_0\right]^2$$
$$= (\epsilon m)^{-2}2\sigma^4\sum_{t=1}^m\sum_{k=1}^{t-1}(\beta_0 + \alpha_0)^{4(k-1)}$$
$$+(\epsilon m)^{-2}4\sigma^4\sum_{t=1}^{m-1}\sum_{t'=t+1}^m(\beta_0 + \alpha_0)^{2(t'-t)}\sum_{l=0}^{2t-1}(\beta_0 + \alpha_0)^{2l}$$
$$= (\epsilon m)^{-2}O(m) = o(1),$$

which implies that

$$\sup_{(\beta,\alpha)\in\Omega}\left|m^{-1}\nabla_\beta^\intercal \nabla_\beta \ell_1(\beta, \alpha) - E\left\{m^{-1}\nabla_\beta^\intercal \nabla_\beta \ell_1(\beta, \alpha)|\alpha_0\right\}\right| = o_P(1).$$

Supplementary Materials

Section S1 contains the estimation algorithm of policy parameters.
Section S2 contains proofs of all technical results.
Section S3 contains additional simulation results for decision quality comparison of the five methods when $m = 20, 30$.

Funding

The authors gratefully acknowledge NIH grants R01MH109496 and R21MH108999.

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014), "Fitting Linear Mixed-Effects Models Using lme4," arXiv no. 1406.5823. [6]

Bohmer, M., Hecht, B., Schoning, J., Kruger, A., and Bauer, G. (2011), "Falling Asleep With Angry Birds, Facebook and Kindle—A Large Scale Study on Mobile Application Usage," in *MobileHCI*, pp. 47–56. [1]

Boruvka, A., Almirall, D., Witkiewitz, K., and Murphy, S. A. (2018), "Assessing Time-Varying Causal Effect Moderation in Mobile Health," *Journal of the American Statistical Association*, 113, 1112–1121. [6]

Cheung, K., Ling, W., Karr, C. J., Weingardt, K., Schueller, S. M., and Mohr, D. C. (2018), "Evaluation of a Recommender App for Apps for the Treatment of Depression and Anxiety: An Analysis of Longitudinal User Engagement," *Journal of the American Medical Informatics Association*, 25, 955–962. [1]

Cheung, Y. K., Chakraborty, B., and Davidson, K. W. (2015), "Sequential Multiple Assignment Randomized Trial (SMART) With Adaptive Randomization for Quality Improvement in Depression Treatment Program," *Biometrics*, 71, 450–459. [10]

Cho, H., Wang, P., and Qu, A. (2017), "Personalize Treatment for Longitudinal Data Using Unspecified Random-Effects Model," *Statistica Sinica*, 27, 187–206. [2]

Christensen, H., Griffiths, K., and Farrer, L. (2009), "Adherence in Internet Interventions for Anxiety and Depression," *Journal of Medical Internet Research*, 11, e13. [1]

Depp, C. A., Mausbach, B., Granholm, E., Cardenas, V., Ben-Zeev, D., Patterson, T. L., Lebowitz, B. D., and Jeste, D. V. (2010), "Mobile Interventions for Severe Mental Illness: Design and Preliminary Data From Three Approaches," *The Journal of Nervous and Mental Disease*, 198, 715. [1]

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002), *Analysis of Longitudinal Data*, Oxford: Oxford University Press. [2]

Ertefaie, A., and Strawderman, R. L. (2018), "Constructing Dynamic Treatment Regimes Over Indefinite Time Horizons," *Biometrika*, 105, 963–977. [2]

Helander, E., Kaipainen, K., Korhonen, I., and Wansink, B. (2014), "Factors Related to Sustained Use of a Free Mobile App for Dietary Self-Monitoring With Photography and Peer Feedback: Retrospective Cohort Study," *Journal of Medical Internet Research*, 16, e109. [1,9]

Heron, K. E., and Smyth, J. M. (2010), "Ecological Momentary Interventions: Incorporating Mobile Technology Into Psychosocial and Health Behaviour Treatments," *British Journal of Health Psychology*, 15, 1–39. [1]

Hui, F. K., Müller, S., and Welsh, A. (2017), "Joint Selection in Mixed Models Using Regularized PQL," *Journal of the American Statistical Association*, 112, 1323–1333. [6]

Kravitz, R., and Duan, N. (2014), "Design and Implementation of N-of-1 Trials: A User's Guide," AHRQ Publication No. 13(14)-EHC122-EF, Agency for Healthcare Research and Quality. [2]

Laber, E., Linn, K., and Stefanski, L. (2014), "Interactive Model Building for Q-Learning," *Biometrika*, 101, 831–847. [2]

Lei, H., Tewari, A., and Murphy, S. (2017), "An Actor-Critic Contextual Bandit Algorithm for Personalized Interventions Using Mobile Devices," arXiv no. 1706.09090. [2]

Lillie, E., Patay, B., Diamant, J., Issell, B., Topol, E., and Schork, N. (2011), "The n-of-1 Clinical Trial: The Ultimate Strategy for Individualizing Medicine," *Personalized Medicine*, 8, 161–173. [2]

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2020), "Estimating Dynamic Treatment Regimes in Mobile Health Using V-Learning," *Journal of the American Statistical Association*, 115, 692–706. [2]

Mohr, D. C., Cheung, K., Schueller, S. M., Brown, C. H., and Duan, N. (2013), "Continuous Evaluation of Evolving Behavioral Intervention Technologies," *American Journal of Preventive Medicine*, 45, 517–523. [1]

Mohr, D. C., Tomasino, K. N., Lattie, E. G., Palac, H. L., Kwasny, M. J., Weingardt, K., Karr, C. J., Kaiser, S. M., Rossom, R. C., Bardsley, L. R., and Caccamo, L. (2017), "IntelliCare: An Eclectic, Skills-Based App Suite for the Treatment of Depression and Anxiety," *Journal of Medical Internet Research*, 19, e10. [1]

Murphy, S. A. (2003), "Optimal Dynamic Treatment Regimes," *Journal of the Royal Statistical Society*, Series B, 65, 331–355. [2]

Nash, S. G. (2000), "A Survey of Truncated-Newton Methods," *Journal of Computational and Applied Mathematics*, 124, 45–59. [4]

Ohrnberger, J., Fichera, E., and Sutton, M. (2017), "The Rise of Consumer Health Wearables: Promises and Barriers," *Journal of the Economics of Ageing*, 9, 52–62. [2]

Pepe, M. S., and Anderson, G. L. (1994), "A Cautionary Note on Inference for Marginal Regression Models With Longitudinal Data and General Correlated Response Data," *Communications in Statistics—Simulation and Computation*, 23, 939–951. [2]

Qian, M., and Murphy, S. A. (2011), "Performance Guarantees for Individualized Treatment Rules," *The Annals of Statistics*, 39, 1180–1210. [2]

Qian, T., Klasnja, P., and Murphy, S. A. (2019), "Linear Mixed Models Under Endogeneity: Modeling Sequential Treatment Effects With Application to a Mobile Health Study," arXiv no. 1902.10861. [2]

Riley, W. T., Rivera, D. E., Atienza, A. A., Nilsen, W., Allison, S. M., and Mermelstein, R. (2011), "Health Behavior Models in the Age of Mobile Interventions: Are Our Theories up to the Task?," *Translational Behavioral Medicine*, 1, 53–71. [1]

Song, R., Wang, W., Zeng, D., and Kosorok, M. R. (2015), "Penalized Q-Learning for Dynamic Treatment Regimes," *Statistical Sinica*, 25, 901–920. [2]

Steihaug, T. (1983), "The Conjugate Gradient Method and Trust Regions in Large Scale Optimization," *SIAM Journal on Numerical Analysis*, 20, 626–637. [4]

Tewari, A., and Murphy, S. (2017), "From Ads to Interventions: Contextual Bandits in Mobile Health," in *Mobile Health*, eds. J. Rehg, S. Murphy, and S. Kumar, Cham: Springer, pp. 495–517. [2]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [2]

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012), "A Robust Method for Estimating Optimal Treatment Regimes," *Biometrics*, 68, 1010–1018. [2]

Zhao, Y., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015), "New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *Journal of the American Statistical Association*, 110, 583–598. [2]

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [2]