The Annals of Applied Statistics
2020, Vol. 14, No. 2, 768–788
https://doi.org/10.1214/20-AOAS1323

SEMIPARAMETRIC BAYESIAN MARKOV ANALYSIS OF PERSONALIZED BENEFIT–RISK ASSESSMENT

BY DONGYAN YAN[1], SUBHARUP GUHA[2], CHUL AHN[3,*] AND RAM TIWARI[3,†]

[1]*Discovery & Development Statistics, Eli Lilly and Company, dyyr2@mail.missouri.edu*
[2]*Department of Biostatistics, University of Florida, s.guha@ufl.edu*
[3]*Division of Biostatistics, Center for Devices and Radiological Health, Office Surveillance and Biometrics, U.S. Food and Drug Administration, *chul.ahn@fda.hhs.gov; †ram.tiwari@fda.hhs.gov*

The development of systematic and structured approaches to assess benefit–risk of medical products is a major challenge for regulatory decision makers. Existing benefit–risk methods depend only on the frequencies of mutually exclusive and exhaustive categories in which the subjects fall, and the responses of individuals are allowed to belong to any of the other categories during their postwithdrawal visits. In this article we introduce a semiparametric Bayesian Markov model (SBMM) that treats the withdrawal category as an absorbing state and analyzes subject-level data for multiple visits, accounting for any within-patient dependencies in the response profiles. A log-odds ratio model is used to model the subject-level effects by assuming a ratio of transition probabilities with respect to a "reference" category. A Dirichlet process is used as a semiparametric model for the subject-level effects to flexibly capture the underlying distributions of the personalized response profiles without making strong parametric assumptions. This also allows the borrowing of strength between the patients and achieves dimension reduction by allocating similar response profiles patterns into an unknown number of latent clusters. We analyze a motivating clinical trial dataset to assess the personalized benefit–risks in each arm and evaluate the aggregated benefits and risks associated with the drug Exalgo.

1. Introduction. Benefit–risk (BR) assessment of treatments or medical products plays an essential role in regulatory decision making in pre-market and post-market review processes. Assessment typically considers comprehensive information on safety and effectiveness and involves quantitative analyses as well as more subjective, qualitative weighing of evidence. A key challenge is the characterization of uncertainty associated with benefit and risk evaluation of medical products. The Institute of Medicine (IOM) has convened a workshop to advance the development of systematic and structured approaches for regulatory decision making (Claiborne, English and Caruso (2014)).

Our main statistical goal is the development of effective statistical methodology for evaluating the benefits and risks of medical products. For a motivating application, we explored the clinical trial data described in Norton (2011). This randomized, double-blinded, placebo-control clinical trial studied the benefits and risks associated with Exalgo, an extended release hydromorphone product approved in 2010 for the management of moderate to severe pain in opioid-tolerant patients. The clinical trial recruited 268 subjects and randomly assigned them equally to the treatment and control arms. Each subject was followed up for eight visits, and their outcomes were evaluated by a team of medical doctors.

In the Norton (2011) clinical trial, *benefit* is defined as a clinically important improvement, as determined by medical officers of at least 30% daily pain reduction in a patient on a given visit. *Risk* is defined as the occurrence of serious adverse events (AE), such as disability or

Received February 2019; revised January 2020.
Key words and phrases. Clinical trials, Dirichlet process, Exalgo, Gibbs sampling, log-odds ratio model, Metropolis–Hastings, SBMM.

TABLE 1
Outcomes of a clinical trial with binary response data

	Benefit		No benefit
No AE	Category 1		Category 3
AE	Category 2		Category 4
Withdrawal		Category 5	

death, or of moderately adverse events that "may be of sufficient severity to make patient uncomfortable; performance of daily activities may be influenced; intervention may be needed." With these definitions the individual responses at each visit were grouped by medical officers into one of five categories, as shown in Table 1. The categories range from most desirable (category 1) to least desirable (category 5).

Subject withdrawals are not unusual in clinical trials such as Norton (2011) with withdrawal rates of 50.7% and 67.9% for the treatment and control arm, respectively, at the last visit. Reasons for withdrawal included loss of contact, inadequate benefit from the drug, and occurrence of adverse events. In general, benefit and risk are not necessarily mutually exclusive characteristics; a high-efficacy drug may also be associated with harmful side effects. Furthermore, the BR tradeoff may vary over the course of the trial. For example, in the earlier stages of a clinical trial some subjects report overall benefit but experience increasingly adverse events as the trial progresses. This strongly suggests that BR associations should be longitudinally evaluated for each patient.

Several quantitative measures for BR assessment have been proposed in the literature. Payne and Loken summarized the BR ratio at the population level (Payne and Loken (1975)). Gelber, Gelman and Goldhirsch (1989) combined population level BR measures into a single quantity called Time Without Symptoms of Disease and Toxic Effects (TWiST), and Glasziou, Simes and Gelber (1990) further generalized this to Quality-adjusted TWiST (Q-TWiST). Norton (2011) developed a longitudinal visualization technique for BR evaluation over the course of a clinical trial. Holden, Juhaeri and Dai (2003) proposed using the ratio of the number of subjects needed to treat for benefit to the number of subjects needed to treat for risks. Pritchett and Tamura (2008) evaluated the robustness of the definitions of "benefit" and "risk" and made suggestions on selected prespecified weights. Multiple-criteria decision analysis (MCDA) has been utilized in assessing BR scores, especially for continuous endpoints. This approach explicitly evaluates multiple conflicting criteria in decision making and is the most commonly used quantitative framework in case studies (Gøtzsche and Jørgensen (2011), Hughes et al. (2016)). Bayesian inference, with its capability for incorporating different sources of information and uncertainty, along with its links to optimal decision theory provides a paradigm for quantitative analysis of BR tradeoff. Costa et al. (2017) provided an overview of the state-of-the-art in Bayesian methodologies for quantitative BR assessment and emphasized the importance of profiling subgroup-specific BR measures. Costa and Drury (2018) proposed two Bayesian joint modeling approaches to account for the potential dependence between efficacy and safety outcomes at the subject level.

The patient outcome categories given in Table 1 were first introduced by Chuang-Stein, Mohberg and Sinkula (1991) to capture the BR profiles of each individual. Compared with the aforementioned BR measures, the Global Benefit–Risk (GBR) scores proposed by Chuang-Stein, Mohberg and Sinkula (1991) are relatively easy and straightforward to implement, provided the BR categories are well defined in an application. Chuang-Stein, Mohberg and Sinkula (1991) also proposed three new measures of BR assessment: one based on a weighted linear combination of the estimated probabilities associated with the five categories, and the

other two based on ratios of those probabilities. GBR scores have been applied to studies for antidepressant drugs (Entsuah and Gorman (2002)). Zhao et al. (2014) proposed a different BR measure using a generalized indicator function for each of the five categories and implemented the analysis using Bayesian power priors (Ibrahim and Chen (2000)) that facilitated a longitudinal BR assessment using various global measures.

A drawback of the above methods is that they ignore subject-level variability in the responses, often resulting in biased inferences for the BR measures. Cui, Zhao and Tiwari (2016) proposed a Bayesian approach that incorporates subject-specific categorical effects. Although it is an important step in the right direction, a drawback of this work is that it does not realistically model subjects who are lost to subsequent follow-ups as a result of withdrawal. The responses of these individuals are allowed to belong with positive probability to any of the Table 1 categories even during their post-withdrawal visits. In clinical trials with high withdrawal rates, which is a common feature of many studies, this unrealistic assumption results in poor model fit. For example, applying the method of Cui, Zhao and Tiwari (2016) to the Norton (2011) clinical trial, Figure 1 displays the 95% credible intervals of the categorical probabilities for subject 86 belonging to the control arm. This subject withdrew from the study at visit 3. However, as seen in the figure, after accounting for uncertainty, the probability estimates for the nonwithdrawal categories are strictly positive during the postwithdrawal visits. In fact, the estimated Category 3 probabilities for subject 86 are approximately 0.2 for postwithdrawal visits 4–8.

As the above analysis illustrates, a limitation of existing statistical approaches is that they fail to account for dependencies between the longitudinal responses of a subject. Motivated

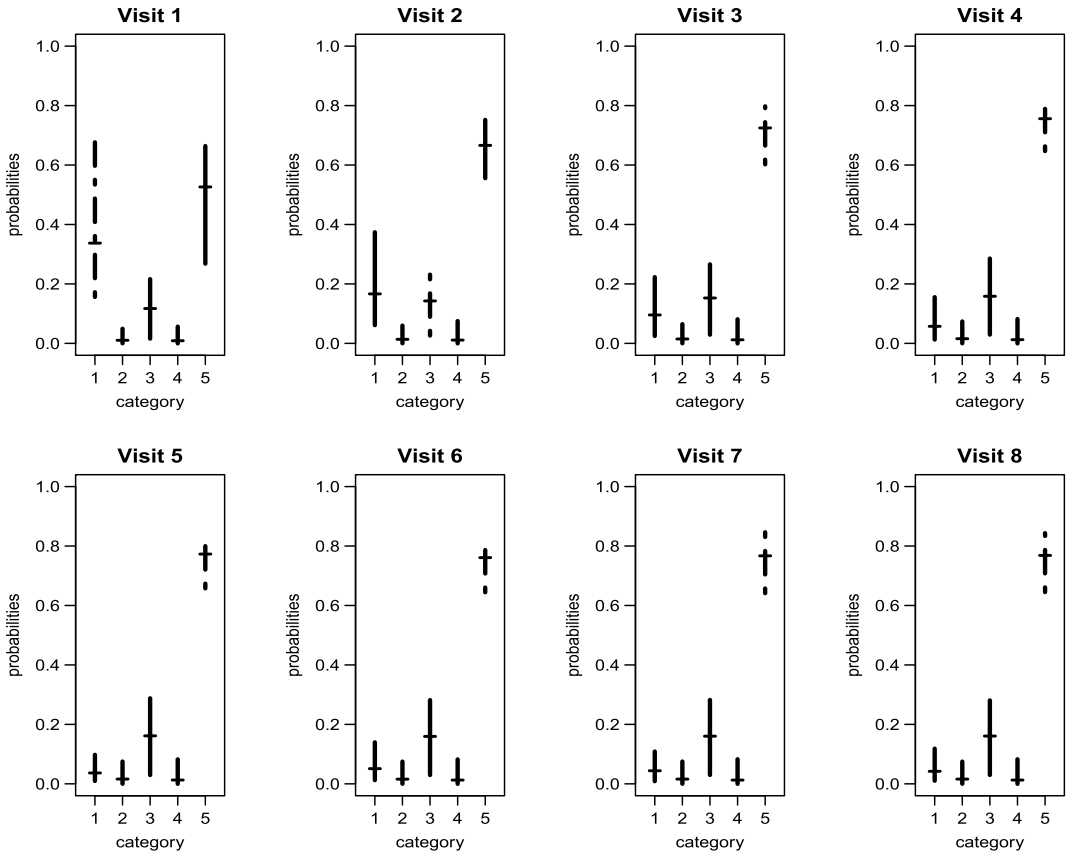

FIG. 1. *95% credible intervals of marginal categorical probabilities for subject* 86 *from the control arm. Dashed line indicates the observed category at each visit*; *this subject withdrew from the study at visit* 3.

by this, we propose modeling the transition probabilities between categories, treating the withdrawal category as an absorbing state as well as taking subject-level category effects into account. The idea of transition probabilities follows from the Markov property, where the conditional probability distribution of future states of the process (conditional on both past and present states) depends upon the present state. In a clinical trial, subject's status at any visit is highly informative in determining the status of the next visit. The transition probabilities are modeled by a log-odds ratio model, the subject-level category effects are then modeled through a Dirichlet process (Ferguson (1973)). BR measures are thereby defined based on the transition probabilities.

The rest of this article is organized as follows. Section 2 reviews the benefit–risk categories and quantitative measures introduced in Chuang-Stein, Mohberg and Sinkula (1991) and modifies the traditional definitions to accommodate Markov transition probabilities. Section 3 describes the semiparametric Bayesian Markov model (SBMM). Section 4 presents the MCMC sampling scheme. In Section 5 we present simulation studies to evaluate the proposed model's performance in estimating the subject-level transition probabilities. In Section 6 we analyze the motivating clinical trial data using SBMM.

2. Benefit–risk categories and measures. In a longitudinal clinical trial setting, suppose that the patients visit sites at K different time points. Suppose that the patient outcomes at each visit can be classified into one of the five mutually exclusive and exhaustive categories: "benefit without AE, benefit with AE, no benefit, no AE, no benefit with AE" and "withdrawal," as listed in Table 1. These categories are progressively ordered from the most desirable to the least desirable for the patient.

We consider the patient's outcome at a given time point as beneficial if he or she has transited to a "better" category since the last visit and consider the outcome to be risky if the patient has transited to a "worse" category. Specifically, transitions from Category 1 to 1, Category 2 to 1 or 2, Category 3 to 1 or 2 and Category 4 to 1, 2 or 3 are all considered favoring benefit. On the other hand, transitions from Category 1 to others, Category 2 to 3, 4 or 5, Category 3 to 3, 4 or 5 and Category 4 to 4 or 5 are considered favoring risk.

To evaluate the benefits and risks for each arm of the clinical study as well as compare the two arms with respect to benefits and risks, we consider three measures first proposed in Chuang-Stein, Mohberg and Sinkula (1991), namely, linear ($\mathrm{BR_L}$), ratio ($\mathrm{BR_R}$) and composite ratio ($\mathrm{BR_{CR}}$) measures, after adapting each to the graded BR interpretation associated with the five outcome categories. Conditional on the event that the previous visit belonged to category s, let p_{sj} denotes the unknown probability of observing category j. Let $\boldsymbol{P} = ((p_{sj}))$ be the $J \times J$ transition probability matrix, with the last row corresponding to the withdrawal state given by $(p_{51}, \ldots, p_{55}) = (0, 0, 0, 0, 1)$. With these definitions the BR measures of Chuang-Stein, Mohberg and Sinkula (1991), adapted to the graded outcome categories of Table 1, become

$$\begin{aligned}
\mathrm{BR_L}(\boldsymbol{P}) &= w_{11}p_{11} + \sum_{s=2}^{3}\sum_{j=1}^{2} w_{sj}p_{sj} + \sum_{j=1}^{3} w_{4j}p_{4j} \\
&\quad - \sum_{j=2}^{5} w_{1j}p_{1j} - \sum_{s=2}^{3}\sum_{j=3}^{5} w_{sj}p_{sj} - \sum_{j=4}^{5} w_{4j}p_{4j}, \\
\mathrm{BR_R}(\boldsymbol{P}) &= \frac{(w_{11}p_{11} + \sum_{s=2}^{3}\sum_{j=1}^{2} w_{sj}p_{sj} + \sum_{j=1}^{3} w_{4j}p_{4j})^{\psi}}{\sum_{j=2}^{5} w_{1j}p_{1j} + \sum_{s=2}^{3}\sum_{j=3}^{5} w_{sj}p_{sj} + \sum_{j=4}^{5} w_{4j}p_{4j}}, \quad \psi > 0, \\
\mathrm{BR_{CR}}(\boldsymbol{P}) &= \frac{\sum_{s=1}^{4} w_{s1}p_{s1}}{\sum_{s=1}^{4} w_{s5}p_{s5}} \left(\frac{\sum_{s=2}^{3} w_{s2}p_{s2} + \sum_{j=2}^{3} w_{4j}p_{4j}}{\sum_{j=2}^{4} w_{1j}p_{1j} + \sum_{s=2}^{3}\sum_{j=3}^{4} w_{sj}p_{sj} + w_{44}p_{44}} \right)^{\phi}, \quad \phi > 0,
\end{aligned} \tag{2.1}$$

where w_{sj} is a user-specified positive score associated with the ordered category pair (s, j) for consecutive visits. The scores are chosen by the investigator to reflect the relative importance of the transitions between categories to a researcher, clinician, patient or care giver, when evaluating a treatment. Because the acceptability of risk depends on the achievable benefit as well as available treatments, the scores may change with different diseases or symptoms. In the motivating application we assume prespecified sets of weights, $\boldsymbol{w}_1 = (4, 0.5, 0.5, 1, 2)$, $\boldsymbol{w}_2 = (2, 1, 1, 1, 3)$, $\boldsymbol{w}_3 = (3, 0.5, 0, 0.5, 2)$ and $\boldsymbol{w}_4 = (3, 2, 1, 1, 5)$, representing the scores assigned to the five BR categories during the current visit when the previous visit's state equals the subscript s of vector $\boldsymbol{w}_s$.

The score $\mathrm{BR_L}$ is a linear combination of the probabilities of the five benefit–risk categories. The ratio of benefit and risk with nonnegative exponent ψ reflecting the relative importance of benefit to risk forms the basis of the ratio score, $\mathrm{BR_R}$. Composite ratio score $\mathrm{BR_{CR}}$ is based on a composite ratio of benefit and risk, where exponent ϕ is a nonnegative constant and is used to give a different score to different benefit or risk categories. If the constant ϕ equals zero, it represents the ratio of the best categories (transitions to benefit without risk) over the worst category (withdrawal from any category). In this paper we assumed $\psi = 1$ and $\phi = 1$ in equation (2.1).

Similarities or differences in the posterior inferences provided by the different BR measures are expected to vary across datasets. For the motivating clinical trial data, the scores $\mathrm{BR_L}$ and $\log(\mathrm{BR_R})$ give qualitatively similar results in Section 6, where the high withdrawal rate of the study is shown to account for the slightly different results for score $\log(\mathrm{BR_{CR}})$.

Chuang-Stein, Mohberg and Sinkula (1991) assumed that transition matrix $\boldsymbol{P}$ is shared by all individuals belonging to a study arm (treatment and control) and are, respectively, denoted by $\boldsymbol{P}^{(T)}$ and $\boldsymbol{P}^{(C)}$. For two-arm randomized trials, Chuang-Stein, Mohberg and Sinkula (1991) proposed an asymptotic test for equality of benefit–risk measures between the arms. Specifically, for the BR measures (2.1) evaluated for each arm, the absolute BR difference (ABRD) between the treatment and control arms is defined as

$$\begin{aligned}
\mathrm{ABRD_L} &= \mathrm{BR_L}\bigl(\boldsymbol{P}^{(T)}\bigr) - \mathrm{BR_L}\bigl(\boldsymbol{P}^{(C)}\bigr),\\
\mathrm{ABRD_R} &= \log\biggl\{\frac{\mathrm{BR_R}(\boldsymbol{P}^{(T)})}{\mathrm{BR_R}(\boldsymbol{P}^{(C)})}\biggr\},\\
\mathrm{ABRD_{CR}} &= \log\biggl\{\frac{\mathrm{BR_{CR}}(\boldsymbol{P}^{(T)})}{\mathrm{BR_{CR}}(\boldsymbol{P}^{(C)})}\biggr\}.
\end{aligned} \tag{2.2}$$

The ABRD measures compare the differential BR measures between the study arms and are, therefore, the primary parameters of interest; a positive (negative) value indicates that the benefit is greater (less) than the risk for the treatment arm relative to the control arm. It can be shown that the $\mathrm{ABRD_L}$ values range from -24 to 24 for the previously mentioned user-specified weights $\boldsymbol{w}_s$, $s = 1, \ldots, 5$, for the Table 1 categories. Both $\mathrm{ABRD_R}$ and $\mathrm{ABRD_{CR}}$ are on the log scale with values ranging from $-\infty$ to ∞.

Additonally, Chuang-Stein, Mohberg and Sinkula (1991) derived 95% approximate, large-sample frequentist confidence intervals (CIs) for the ABRD measures and used these CIs to perform approximate hypothesis tests for the efficacy of the treatment. However, their assumption of a common transition matrix for all individuals within a study arm may not be unrealistic, and this is one of the motivations for the proposed method of this paper.

The Bayesian procedures developed in this paper are exact and are applicable irrespective of the number of subjects participating in the clinical trial. They avoid making unrealistic assumptions about the underlying distributions associated with the treatment and control arms, including the existence of a common transition probability matrix for all subjects in an arm. Additionally, the techniques are able to accommodate high withdrawal rates and adjust for

dependencies between within-patient responses. The model details are presented in the next section.

3. A semiparametric Bayesian Markov model. Consider a clinical trial consisting of categorical outcomes Y_{ik} falling into J possible categories, or states, for subject $i = 1, \ldots, N$, and visit $k = 1, \ldots, K$. In the motivating application we have $J = 5$ categories as shown in Table 1. For the first visit the initial states of the patients are assumed to be i.i.d. categorical (i.e., generalized Bernoulli) distribution on J categories with probabilities $\boldsymbol{q} = (q_1, \ldots, q_J)$. The unknown probability vector is given a Dirichlet distribution on the unit simplex in $\mathcal{R}^J$ with concentration parameter 1. That is,

$$Y_{i1} \mid \boldsymbol{q} \overset{\text{i.i.d.}}{\sim} \mathcal{C}_J(\boldsymbol{q}), \quad i = 1, 2, \ldots, N, \tag{3.1}$$
$$\boldsymbol{q} \sim \mathcal{D}_J\left(\frac{1}{J}, \ldots, \frac{1}{J}\right).$$

Markov dependence. To account for longitudinal dependencies between the patient responses, we assume that outcome Y_{ik} depends on the history, $Y_{i1}, \ldots, Y_{i,k-1}$, only through the outcome $Y_{i,k-1}$, for visit $k = 2, \ldots, K$. The outcomes rely on an underlying set of *subject-specific* transition probabilities, $P(Y_{ik} = j \mid Y_{i,k-1} = s)$, that may possibly depend on subject i. The probabilities are stationary because they do not depend on visit k. This framework provides the flexibility of potentially allowing a different set of transition probabilities for every individual. However, as we shall later see, dimension reduction in the large number of subjects is achieved via Dirichlet process latent clusters for the subject-specific sets of probabilities.

Labeling the withdrawal (absorbing) state as category J, we trivially obtain $P(Y_{ik} = j \mid Y_{i,k-1} = J) = \delta_J(j)$, the Dirac delta function. In other words, once a patient has withdrawn from the study, they would persist in the absorbing state throughout their remaining visits. On the other hand, if the $(k-1)$ visit belongs to a nonabsorbing state $s \neq J$, then the log-odds of category j at visit k, relative to the reference category 1 are given by

$$\log\left\{\frac{P(Y_{ik} = j \mid Y_{i,k-1} = s)}{P(Y_{ik} = 1 \mid Y_{i,k-1} = s)}\right\} = \beta_{isj}, \quad j = 2, \ldots, J, \text{ and provided } s \neq J, \tag{3.2}$$

where subject-specific log-odds vector $\boldsymbol{\beta}_i = (\beta_{i12}, \ldots, \beta_{i1J}, \ldots, \beta_{i(J-1)2}, \ldots, \beta_{i(J-1)J})$ represents the categorical effects.

Equivalently, $\boldsymbol{P}_i = ((p_{isj}))$ is the $J \times J$ subject-specific transition probability uniquely identified by vector $\boldsymbol{\beta}_i$. The last row of this matrix corresponds to the previous visit belonging to the absorbing Jth state and is, therefore, equal to the vector $(0, \ldots, 0, 1)$ of length J. In conjunction with equation (3.2), the elements of transition matrix $\boldsymbol{P}_i$ are given by

$$p_{isj} = \begin{cases} 1\Big/\left[1 + \sum_{t=2}^{J} \exp(\beta_{ist})\right], & s = 1, \ldots, J-1, \text{ and } j = 1, \\ \exp(\beta_{isj})\Big/\left[1 + \sum_{t=2}^{J} \exp(\beta_{ist})\right], & s = 1, \ldots, J-1, \text{ and } j = 2, \ldots, J, \\ \delta_J(j) & s = J, \text{ and } j = 1, \ldots, J. \end{cases} \tag{3.3}$$

Likelihood function. Let y_{ik} be the *observed* category of subject i on visit k. Let Θ represent the set of model parameters $(\boldsymbol{q}, \boldsymbol{P}_1, \ldots, \boldsymbol{P}_N)$; equivalently, the set of parameters

$(\boldsymbol{q}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N)$. With $[\cdot]$ denoting densities, assumptions (3.1) and (3.3) imply that the likelihood contribution of subject i is

$$\mathcal{L}_i(\Theta \mid \boldsymbol{y}) = \{y_{i1}, \ldots, y_{iK} \mid \Theta\} = q_{y_{i1}} \prod_{s=1}^{J-1} \prod_{j=1}^{J} p_{isj}^{y_{isj}},$$

where $n_{isj} = \sum_{k=2}^{K} \mathcal{I}(y_{i,k-1} = s, y_{ik} = j)$ is the total number of transitions from category s to category j by subject i over the K visits. The joint likelihood function is then $\mathcal{L}(\Theta \mid \boldsymbol{y}) = \prod_{i=1}^{N} \mathcal{L}_i(\Theta \mid \boldsymbol{y})$, with the corresponding log-likelihood function given by

$$\begin{aligned} l(\Theta \mid \boldsymbol{y}) = {} & \sum_{i=1}^{N} \log q_{y_{i1}} + \sum_{i=1}^{N} \sum_{s=1}^{J-1} \sum_{j=2}^{J} n_{isj} \beta_{isj} \\ & - \sum_{i=1}^{N} \sum_{s=1}^{J-1} \sum_{j=1}^{J} n_{isj} \log \left\{ 1 + \sum_{t=2}^{J} \exp(\beta_{ist}) \right\}. \end{aligned} \tag{3.4}$$

Modeling the patient population. Let the random assignment of the subjects to the treatment or control arm of the clinical trial be recorded in the variable $h_i \in \{1, 2\}$, for subject $i = 1, \ldots, N$, with the label 1 (2) representing the control (treatment) arm. Let N_1 and N_2, respectively, denote the number of subjects assigned to the control and treatment arms. Within each arm the population of patients may be regarded as an admixture of latent subpopulations consisting of shared probability transition matrices. Subjects belonging to the same subpopulation cluster in an arm have identical probability transition matrices $\boldsymbol{P}_i$ (equivalently, identical log-odds vector $\boldsymbol{\beta}_i$), whereas subjects belonging to different subpopulations have different matrices and log-odds vectors. Consequently, in each arm subjects within latent clusters also have identical BR measures, defined in equation (2.1).

Since the number of latent subpopulations and the common distribution of the a priori exchangeable log-odds vectors are unknown, we assume that the vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$ are random draws from unknown arm-specific distributions which are themselves distributed as a Dirichlet process (Ferguson (1973)). Specifically, let the N subject-specific log-odds vectors, each of length $(J-1)^2$, follow either $(J-1)^2$-variate distribution G_1 or G_2, depending on whether subjects are assigned to the control ("1") or treatment ("2") arm, with $P(G_1 = G_2) = 0$ a priori. These distributions follow a Dirichlet process with concentration parameter α and $(J-1)^2$-variate normal base distribution, G_0:

$$\begin{aligned} & \boldsymbol{\beta}_i \overset{\text{indep}}{\sim} G_{h_i}, \quad i = 1, \ldots, N \quad \text{where} \\ & G_1, G_2 \overset{\text{i.i.d.}}{\sim} \text{DP}(\alpha, G_0) \quad \text{with} \\ & G_0 = N_{(J-1)^2}(\mathbf{0}, \boldsymbol{\Sigma}_0) \end{aligned} \tag{3.5}$$

for a suitably chosen positive-definite matrix, $\boldsymbol{\Sigma}_0$. The assumption that $p(G_1 = G_2) = 0$ a priori is reasonable because it implies that control and treatment arms have no shared subpopulations for their probability transition matrices. In general, however, one might use a nested Dirichlet process (Rodríguez, Dunson and Gelfand (2008)) for the random distribution of $\boldsymbol{\beta}_i$. More recently, Camerlenghi et al. (2019) discussed a generalization to a wider class of nonparametric Bayesian models that are also applicable in the general situation.

Random distributions G are almost surely discrete for Dirichlet processes (Ferguson (1973)). Furthermore, theoretical results (Ghosal, Ghosh and Ramamoorthi (1999)) guarantee that any pair of true (discrete or continuous) arm-specific distributions for the $\boldsymbol{\beta}_i$'s belong to the prior support of the Dirichlet process and can be consistently inferred a posteriori as the number of subjects in each arm grows.

We notice that the terms related to the set of vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$ and to parameter $\boldsymbol{q}$ are mutually separable in the log-likelihood expression (3.4). Consequently, since the priors in expressions (3.1) and (3.5) are independent, these two sets of model parameters are a posteriori independent.

Allocation variable. Let us focus first on the control arm of the clinical trial; an exact analogy holds for the treatment arm. As previously mentioned, the Dirichlet process allocates the N_1 subjects of the control arm to an unknown number, M_1, of latent clusters so that $M_1 \leq N_1$. Imagine that the patient-cluster allocation occurs through a variable, $c_i^{(1)}$, that equals m, if the ith subject is assigned to the mth cluster of the control arm, so that $c_i^{(1)} \in \{1, \ldots, M_1\}$. Let $\boldsymbol{c}^{(1)} = (c_1^{(1)}, \ldots, c_{N_1}^{(1)})$ denote the allocation vector and the common values of the log-odds vectors associated with the M_1 clusters be $\boldsymbol{\varphi}_1^{(1)}, \ldots, \boldsymbol{\varphi}_{M_1}^{(1)}$. Thus, $\boldsymbol{\beta}_i^{(1)} = \boldsymbol{\varphi}_{c_i^{(1)}}^{(1)}$. Furthermore, since the cluster labels are arbitrary, we assume without loss of generality that $c_1^{(1)} = 1$, that is, the first subject is assigned to cluster 1 of the control arm.

Random quantity M_1 is asymptotically equivalent to $\alpha \log N_1$ (Ishwaran and Zarepour (2002)). In other words, the number of latent clusters is much smaller than the number of patients, resulting in dimension reduction in datasets with large N_1. The clustering feature of Dirichlet process and the almost surely discreteness of random distribution G_1 are evident from the stick-breaking representation of Sethuraman and Tiwari (1982) and Sethuraman (1994): If $G_1 \sim \mathrm{DP}(\alpha, G_0)$, then random distribution G_1 takes the form

$$G_1 = \sum_{t=1}^{\infty} \pi_t \delta_t \quad \text{where}$$

$$\boldsymbol{\xi}_t \overset{\text{i.i.d.}}{\sim} G_0,$$

$$\pi_t = \begin{cases} \nu_1 & \text{if } t = 1, \\ \nu_t \prod_{u=1}^{t-1} (1 - \nu_u) & \text{otherwise,} \end{cases}$$

where $\delta_{\boldsymbol{\xi}_t}$ denotes a point mass at the "atom" $\boldsymbol{\xi}_t$. The distinct vectors $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{M_1}$ are a subset of the infinite number of atoms, and the labels of the two sets of M_1 parameters are permutations of each other.

Neal (2000) and Shahbaba and Neal (2009) have shown that a Dirichlet process is the limit of a finite mixture model as the number of mixture components grows to ∞. An equivalent representation of Dirichlet processes is the Polya urn scheme of Blackwell and MacQueen (1973), which characterizes the predictive distribution of allocation variable $c_i^{(1)}$ conditional on the history $c_1^{(1)}, \ldots, c_{i-1}^{(1)}$, as i increases from 2 to N_1. For subject i, suppose there exist $N^{(i-1)}$ distinct clusters among $c_1^{(1)}, \ldots, c_{i-1}^{(1)}$, with the mth cluster containing $N_m^{(i-1)}$ number of subjects. The predictive (prior) probability of allocation variable $c_i^{(1)}$ is then

$$P\big(c_i^{(1)} = m \mid c_1^{(1)}, \ldots, c_{i-1}^{(1)}\big) \propto \begin{cases} N_m^{(i-1)} & \text{if } m = 1, \ldots, N^{(i-1)}, \\ \alpha & \text{if } m = N^{(i-1)} + 1. \end{cases}$$

The first line corresponds to the event that subject i joins one of the clusters occupied by the first $i - 1$ subjects. The second line corresponds to subject i opening a new cluster and belonging to a different subpopulation than the previous $i - 1$ subjects.

By replacing superscript 1 by superscript 2 throughout, similar remarks apply to the treatment arm. Prior (3.5) for the subject-specific vectors $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N)$ implies that these vectors

are equivalent to the parameter set, $(\boldsymbol{c}^{(1)}, \boldsymbol{\varphi}_1^{(1)}, \ldots, \boldsymbol{\varphi}_{M_1}^{(1)}, \boldsymbol{c}^{(2)}, \boldsymbol{\varphi}_1^{(2)}, \ldots, \boldsymbol{\varphi}_{M_2}^{(2)})$. The model parameters are, therefore, regarded as $\Theta = (\boldsymbol{q}, \boldsymbol{c}^{(1)}, \boldsymbol{\varphi}_1^{(1)}, \ldots, \boldsymbol{\varphi}_{M_1}^{(1)}, \boldsymbol{c}^{(2)}, \boldsymbol{\varphi}_1^{(2)}, \ldots, \boldsymbol{\varphi}_{M_2}^{(2)})$. By our earlier discussion, parameter $\boldsymbol{q}$ is a posteriori independent of the remaining model parameters of set Θ. The aforementioned model is referred to as the semiparametric Bayesian Markov model (SBMM).

4. Inference. The Bayesian model for the parameters Θ is complex, necessitating simulated-based techniques such as MCMC for posterior inferences. We briefly outline the iterative procedure in Section 4.1 for SBMM. Subsequently, the MCMC sample is processed to address the scientific questions of interest as described in Section 4.

4.1. *MCMC procedure.* Starting from initial parameters values obtained from ad hoc estimates, the model parameters are iteratively updated as described below.

Vector $\boldsymbol{q}$. Due to conjugate prior (3.1) of parameter $\boldsymbol{q}$ characterizing the initial visit and its a posteriori independence from the remaining model parameters, it can be shown that the marginal posterior of $\boldsymbol{q}$ follows a known Dirichlet distribution on the unit simplex in $\mathcal{R}^J$ with concentration parameter $N + 1$.

For these reasons, if this parameter vector is of interest, inferences are straightforward. However, this is typically not the case, since parameter $\boldsymbol{q}$ is unrelated to the parameters associated with primary questions of interest, which is the comparison of the treatment and control arms via ABRD measures (2.2); the property of posterior independence then allows us to simply ignore parameter $\boldsymbol{q}$ and update the remaining model parameters.

Parameters $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$. The full conditionals of subject-level category effects $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$ are analytically intractable. Consequently, the Metropolis–Hastings algorithm is applied to conditionally update these parameters, along with the associated cluster allocations. Due to the posterior independence of the parameter sets $(\boldsymbol{c}^{(1)}, \boldsymbol{\varphi}_1^{(1)}, \ldots, \boldsymbol{\varphi}_{M_1}^{(1)})$ and $(\boldsymbol{c}^{(2)}, \boldsymbol{\varphi}_1^{(2)}, \ldots, \boldsymbol{\varphi}_{M_2}^{(2)})$, we can independently update these parameters. For the clinical trial arm indexed by $h = 1, 2$:

1. We apply the nonconjugate "auxiliary parameter" algorithm of Neal (2000) to iteratively update the subject-cluster allocations, $c_1^{(h)}, \ldots, c_{N_h}^{(h)}$, along with the distinct log-odds vectors associated with any *newly opened* latent clusters.

2. Given the subject-cluster allocations, the distinct log-odds vectors $\boldsymbol{\varphi}_1^{(h)}, \ldots, \boldsymbol{\varphi}_{M_h}^{(h)}$ are iteratively updated. Due to the conditional nonconjugacy of the model for these parameters, Gibbs sampling cannot be applied to generate draws from the full conditionals. We therefore apply the Laplace approximation to derive a multivariate normal approximation for making proposals for the distinct log-odds vectors. Refer to Zeger and Karim (1991), Chib and Greenberg (1994), Winkelmann (1994) and Guha (2008) for the details.

Specifically, for latent cluster $m = 1, \ldots, M_h$, a new value is proposed from the $(J-1)^2$-variate normal distribution obtained by the Laplace approximation and the current value of $\boldsymbol{\varphi}_m$. The normal proposal is appropriately scaled so that the overall Metropolis–Hastings acceptance rates lie between 25% and 40%, ensuring proper mixing of the MCMC chain. The proposed new value is accepted or rejected by a Metropolis–Hastings probability to compensate for the difference between the true and approximate full conditional of $\boldsymbol{\varphi}_m$.

The postburn-in MCMC draws are used for posterior inferences on the BR measures.

4.2. *Posterior inference on BR and ABRD measures.* Bayes estimates for the ABRD measures comparing the benefit and risk in the treatment and control arms, and previously defined in equation (2.2), are obtained as follows:

1. Each iteration of the MCMC sample is processed to compute the overall BR measures, defined in (2.1), for each study arm.

Specifically, for convenience let the three BR measures in equation (2.1) be generically denoted by $\mathrm{BR}(\boldsymbol{P}_i)$ in subject $i = 1, \ldots, N$. We notice that, for MCMC iteration l, the probability transition matrix $\boldsymbol{P}_i^{(l)}$ of subject i satisfies the relation, $\boldsymbol{P}_i^{(l)} = \boldsymbol{Q}_{c_i^{(h_i)(l)}}^{(h_i)(l)}$, where $\boldsymbol{Q}_m^{(h)(l)}$ denotes the cluster-specific transition matrix corresponding to the cluster-specific log-odds vector, $\boldsymbol{\varphi}_m^{(h)(l)}$, for latent cluster m in arm h at MCMC iteration l. Let $N_{mh}^{(l)} = \sum_{i=1}^{N} \mathcal{I}(c_i^{(h)(l)} = m, h_i = h)$ be the number of patients belonging to the mth latent cluster and hth study arm, for $m = 1, \ldots, M_h$ and $h = 1, 2$. Notice that $N_h = \sum_{m=1}^{M_h} N_{mh}^{(l)}$.

Then, based on the parameter values generated by MCMC iteration l, the overall BR measure for all patients belonging to a study arm has the following expression:

$$\mathrm{BR}\big(\boldsymbol{P}_{(h)}^{(l)}\big) = \frac{1}{N_h} \sum_{m=1}^{M_h} N_{mh}^{(l)} \cdot \mathrm{BR}\big(\boldsymbol{Q}_m^{(h)(l)}\big), \quad h = 1, 2 \tag{4.1}$$

for MCMC iteration $l = 1, \ldots, L$.

Averaging over the MCMC sample, these values provide estimates for the BR measures for each arm, as explained below.

2. The results are averaged over the MCMC sample of size L to obtain the overall BR measure estimates for each study arm

$$\mathrm{BR}(\boldsymbol{P}_{(h)}) = \frac{1}{L} \sum_{l=1}^{L} \mathrm{BR}\big(\boldsymbol{P}_{(h)}^{(l)}\big) \quad \text{for arm } h = 1, 2. \tag{4.2}$$

Standard errors are also estimated in a similar manner. The MCMC sample values could also be utilized to make various other kinds of inferences. For example, marginal posteriors (via histograms or density plots) and joint posteriors (via scatterplots or joint density plots) can be estimated for the BR measures associated with the study arms.

3. Referring back to equation (2.1), we identify the measure $\mathrm{BR}(\boldsymbol{P}_{(1)})$ as measure $\mathrm{BR}(\boldsymbol{P}^{(C)})$ and the measure $\mathrm{BR}(\boldsymbol{P}_{(2)})$ as measure $\mathrm{BR}(\boldsymbol{P}^{(T)})$. Then, Bayes estimates for the three ABRD measures of equation (2.2), along with their estimated standard errors and 95% posterior credible intervals, are immediately available by postprocessing the MCMC sample.

5. Simulation study. To assess the proposed method's performance, we simulated datasets from the Section 3 model to match key characteristics of the motivating Norton (2011) clinical trial. For example, the responses were assumed to fall into $J = 5$ categories with the last category representing the absorbing (withdrawal) state. The two arms correspond to treatment and control, and $N_1 = N_2 = 134$ patients were assigned to each arm, totaling $N = 268$ patients. Outcomes for each patient were generated for K number of visits, where K belonged to the set $\{8, 12, 16\}$.

Generation strategy. The artificially generated patient population consisted of $M_1 = M_2 = 3$ latent clusters having unique characteristics with respect to their transition probabilities. Although it is not entirely accurate, these characteristics can be described in a loose manner as follows. In cluster 1 subjects tend to move to "good" categories 1 or 2. In cluster 2 subjects tend to stay at their current categories. In cluster 3 subjects tend to move to "bad" categories 3, 4 or 5. The three cluster-specific probability transition matrices were

$$\boldsymbol{P}_1^* = \begin{pmatrix} 0.638 & 0.180 & 0.090 & 0.075 & 0.015 \\ 0.480 & 0.323 & 0.102 & 0.078 & 0.015 \\ 0.422 & 0.186 & 0.268 & 0.084 & 0.038 \\ 0.413 & 0.179 & 0.125 & 0.239 & 0.040 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\boldsymbol{P}_2^* = \begin{pmatrix} 0.770 & 0.080 & 0.064 & 0.064 & 0.020 \\ 0.083 & 0.746 & 0.075 & 0.075 & 0.018 \\ 0.081 & 0.074 & 0.751 & 0.074 & 0.018 \\ 0.080 & 0.072 & 0.072 & 0.755 & 0.018 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

and

$$\boldsymbol{P}_3^* = \begin{pmatrix} 0.245 & 0.109 & 0.249 & 0.366 & 0.029 \\ 0.084 & 0.241 & 0.245 & 0.398 & 0.030 \\ 0.065 & 0.065 & 0.347 & 0.474 & 0.046 \\ 0.062 & 0.062 & 0.179 & 0.646 & 0.049 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

where the last row in each matrix corresponds to the absorbing state.

Within each study arm, 134 patients were randomly assigned to the three latent clusters by mixture probabilities denoted by $\boldsymbol{\pi}_h = (\pi_{h1}, \pi_{h2}, \pi_{h3})$ for arm $h = 1, 2$. In other words, for arm h the patients were independently assigned to the three latent clusters by a finite mixture model with parameter $\boldsymbol{\pi}_h$. Two candidate sets of mixture probabilities were considered for each arm: $\boldsymbol{\pi}_B^* = (0.5, 0.3, 0.2)$, representing the situation where the benefit outweighed the risk for the arm, and $\boldsymbol{\pi}_R^* = (0.2, 0.3, 0.5)$, representing the situation where the risk outweighed the benefit for the arm.

Four scenarios for data generation were obtained in this manner. In Scenario 1 the benefit outweighed the risk equally in the treatment and control arms. In Scenario 2 the risk outweighed the benefit equally in the treatment and control arms. We refer to Scenarios 1 and 2 as *balanced*. The true ABRD measures, defined in expression (2.2), are 0 in the balanced scenarios. In Scenario 3 the benefit outweighed the risk for the treatment arm compared to the control arm. In Scenario 4 the risk outweighed the benefit for the treatment arm compared to the control arm. We refer to Scenarios 3 and 4 as *unbalanced*. The situation is illustrated in Table 2. We observe that the true ABRD measures, defined in expression (2.2), are 0 in Scenarios 1 and 2. The true ABRD measures are positive (negative) in Scenario 3 (4).

In this manner the subject-specific transition probability matrices $\boldsymbol{P}_i$, for $i = 1, \ldots, N$, were obtained in each scenario. For generating the first visit states of the patients, we assumed a categorical distribution on the first *four* (nonabsorbing) states with probabilities following

TABLE 2
Scenarios for generating artificial datasets. The scenarios on the diagonal are "balanced," and those on the off-diagonal are "unbalanced"

	Treatment mixture $\boldsymbol{\pi}_1$	
Control mixture $\boldsymbol{\pi}_2$	$\boldsymbol{\pi}_B^*$	$\boldsymbol{\pi}_R^*$
$\boldsymbol{\pi}_B^*$	Scenario 1	Scenario 4
$\boldsymbol{\pi}_R^*$	Scenario 3	Scenario 2

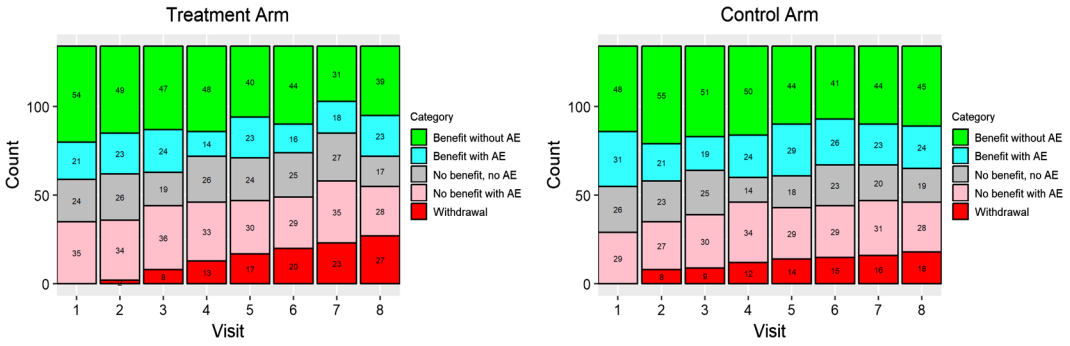

FIG. 2. *For $K = 8$ visits, simulated data for Scenario* 1 (*benefit outweighing risk in both study arms*).

the Dirichlet distribution, $\mathcal{D}_4(1, 1, 1, 1)$. The remaining $K - 1$ visit categories for each patient were generating via a Markov model relying on their subject-specific transition probability matrix, $\boldsymbol{P}_i$.

Independent replications. For each of the four scenarios of Table 2 and number of visits $K \in \{8, 12, 16\}$, 100 datasets were independently generated in the aforementioned manner. The simulated outcomes for $K = 8$ visits for a randomly selected dataset under Scenario 1 are shown in Figure 2. We observe that the better categories dominate the worse categories equally in the two study arms. Similarly, the outcomes for a randomly selected dataset under Scenario 3 are shown in Figure 3; the better categories dominate the worse categories in the treatment arm but not in the control arm.

We fit each of the 100 artificial datasets for each scenario and number of visits using the Section 3 model and Section 4 inference procedure. For the Dirichlet process we assumed concentration parameter $\alpha = 1$, which corresponds to a "prior sample size" of one observation, and base distribution $G_0 = N_{16}(\mathbf{0}, 9\boldsymbol{I}_{16})$. We applied the MCMC procedures in Section 4.1 to obtain posterior samples for the model parameters.

The SBMM methodology accommodates subject-by-subject similarities, accounts for longitudinal effect by modeling transition matrix at subject level and is able to capture the personalized response profiles, instead of merely calculating an overall average profile for each study arm. For Scenario 1 and $K = 8$ visits, the accuracy with which the personalized BR measures are detected is illustrated for two randomly chosen subjects in Figure 4. For a broader evaluation of the accuracy with which the personalized BR measures are inferred, for each scenario and summarizing over on the 100 datasets, Table 3 evaluates the proportion of the 268 subjects (i.e., in the treatment and control arms combined) for whom the true BR

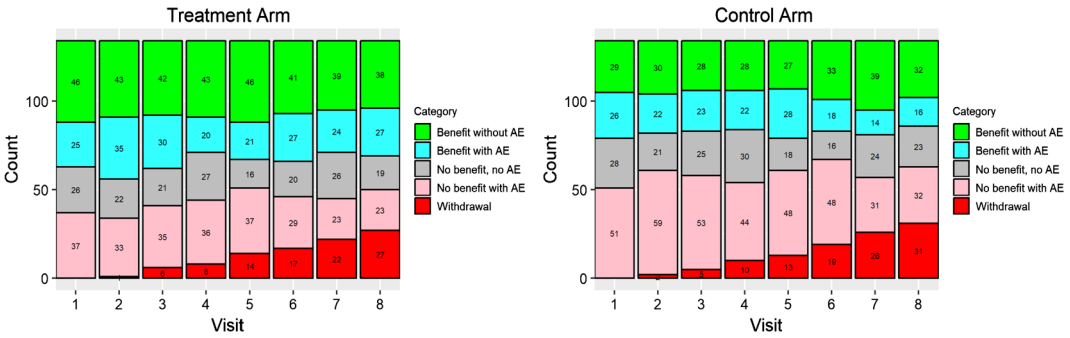

FIG. 3. *For $K = 8$ visits, simulated data for Scenario* 3 (*benefit outweighing risk in the treatment but not in the control arm*).

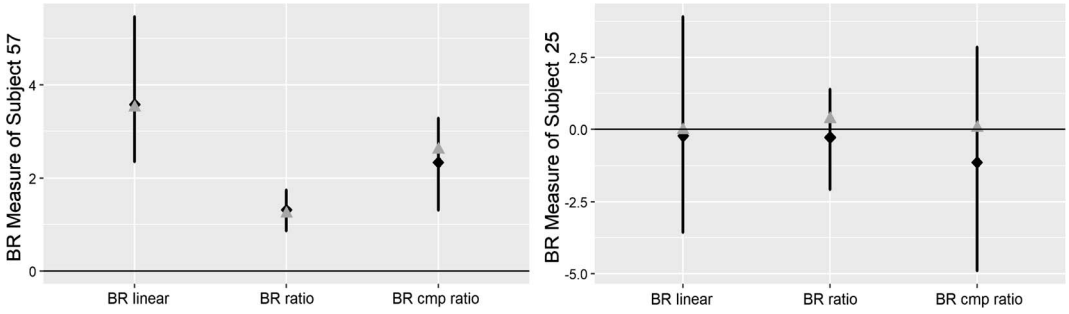

FIG. 4. *For $K = 8$ visits, BR measure of expression (2.1) for subject $i = 57$ from the treatment arm (left panel) and subject $i = 25$ from the control arm (right panel) under Scenario 1. Grey triangles depict the true BR measures, black diamonds depict the posterior means and vertical bars indicate 95% posterior credible intervals.*

measures were covered by the 95% posterior credible intervals of the inferred BR measures. We find that, as the number of visits increases, greater accuracies are achieved.

Inferences on ABRD measures. For each combination of scenario and number of visits K, Table 4 displays posterior credible intervals for the three ABRD measures of expression (2.2) for a *randomly selected dataset*. We find that the intervals for balanced Scenarios 1 and 2 contain the true ABRD measure values of 0. Similarly, all the intervals for unbalanced Scenario 3 (4) belong entirely on the positive (negative) part of the real line, corresponding to the fact that the true ABRD measures are positive (negative). For $K = 8$ visits, Figure 5 plots the posterior credible intervals for (balanced) Scenario 1 and (unbalanced) Scenario 3 for the same dataset displayed in Table 4. We find that, as the number of visits increases, the posterior credible intervals in Table 4 tend to become more precise.

TABLE 3

For each combination of scenario and number of visits and averaging over the 100 datasets, mean proportion of subjects (268 subjects from both arms) for which the true BR measures are contained within the respective Bayesian credible intervals. The standard errors are displayed in the parentheses

	$K = 8$		
Scenario	BR_L	BR_R	BR_{CR}
1	0.59 (0.05)	0.64 (0.05)	0.76 (0.04)
2	0.56 (0.05)	0.62 (0.05)	0.60 (0.05)
3	0.57 (0.05)	0.63 (0.05)	0.67 (0.05)
4	0.58 (0.05)	0.64 (0.05)	0.69 (0.05)
		$K = 12$	
1	0.77 (0.04)	0.79 (0.04)	0.81 (0.04)
2	0.78 (0.04)	0.81 (0.04)	0.82 (0.04)
3	0.76 (0.04)	0.79 (0.04)	0.80 (0.04)
4	0.79 (0.04)	0.81 (0.04)	0.83 (0.04)
		$K = 16$	
1	0.81 (0.04)	0.84 (0.04)	0.85 (0.04)
2	0.86 (0.03)	0.79 (0.04)	0.87 (0.03)
3	0.83 (0.04)	0.80 (0.04)	0.84 (0.04)
4	0.84 (0.04)	0.83 (0.04)	0.88 (0.03)

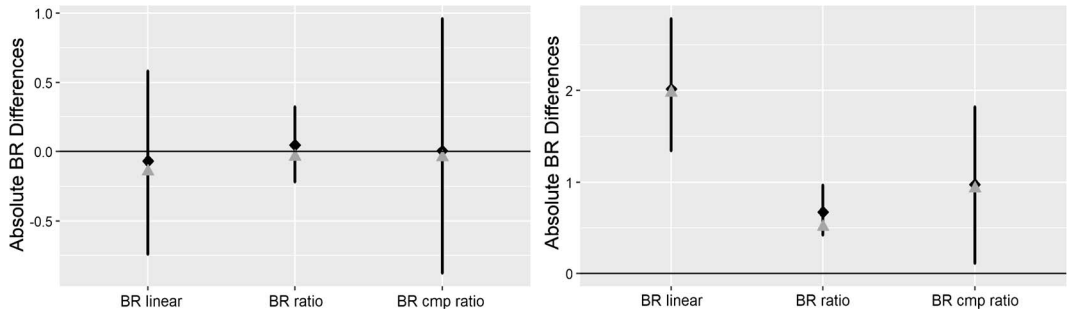

FIG. 5. *For $K = 8$ visits, 95% posterior credible intervals for the three ABRD measures of expression (2.2). The left panel corresponds to the (balanced) Scenario 1, and the right panel corresponds to the (unbalanced, positive) Scenario 3. Grey triangles depict the true BR measures; black diamonds depict the posterior means.*

Summarizing over the 100 datasets for each combination of scenario and number of visits, Table 5 displays the 95% Wilson confidence interval of the proportion of datasets for which the true ABRD measures are contained within the Bayesian credible intervals. As the number of visits increases, the Wilson confidence interval gets narrower. We observe that, although there is variability over the scenarios and ABRD measures, the lower bound of the Wilson confidence intervals are fairly close to 1, especially for $K = 12$ and $K = 16$. The findings in Tables 4 and 5 are a consequence of the fact that the subject-specific transition probability matrices, $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_{268}$, are detected with increasing accuracy as K grows.

Clustering accuracy. Point estimates for the cluster-subject allocations, $\hat{c}_1, \ldots, \hat{c}_N$, can be obtained by applying the technique of Dahl (2006). Ideally, a pair of subjects should be placed in the same estimated cluster by the Dirichlet process if and only if they belong to the same true cluster. To investigate the accuracy of the inferred allocations, we computed the

TABLE 4
95% posterior credible intervals for the different ABRD measures of expression (2.2) for a randomly selected datasets corresponding to each combination of scenario and number of visits, K

	$K = 8$		
Scenario	ABRD_L	ABRD_R	ABRD_CR
1	$(-0.75, 0.58)$	$(-0.25, 0.34)$	$(-0.94, 0.97)$
2	$(-0.54, 1.10)$	$(-0.24, 0.44)$	$(-0.43, 1.06)$
3	$(1.43, 2.81)$	$(0.49, 0.99)$	$(0.15, 1.83)$
4	$(-2.13, -0.75)$	$(-0.80, -0.23)$	$(-1.73, -0.17)$
	$K = 12$		
1	$(-1.00, 0.17)$	$(-0.38, 0.10)$	$(-1.09, 0.32)$
2	$(-0.40, 0.70)$	$(-0.24, 0.28)$	$(-0.67, 0.44)$
3	$(1.37, 2.52)$	$(0.44, 0.90)$	$(0.15, 1.46)$
4	$(-2.75, -1.59)$	$(-1.03, -0.55)$	$(-2.01, -0.67)$
	$K = 16$		
1	$(-0.73, 0.26)$	$(-0.31, 0.11)$	$(-0.82, 0.31)$
2	$(-0.14, 0.87)$	$(-0.10, 0.34)$	$(-0.64, 0.46)$
3	$(1.29, 2.31)$	$(0.38, 0.82)$	$(0.03, 1.19)$
4	$(-2.14, -1.21)$	$(-0.77, -0.36)$	$(-1.50, -0.38)$

TABLE 5
95% Wilson intervals for the proportion of datasets for which the true ABRD measures are contained within the respective Bayesian credible intervals, for each combination of scenario and number of visits

	$K = 8$		
Scenario	ABRD_L	ABRD_R	ABRD_CR
1	(0.80, 0.93)	(0.84, 0.96)	(0.95, 1.00)
2	(0.75, 0.90)	(0.75, 0.90)	(0.80, 0.93)
3	(0.80, 0.93)	(0.80, 0.93)	(0.62, 0.80)
4	(0.95, 1.00)	(0.84, 0.96)	(0.66, 0.83)
		$K = 12$	
1	(0.89, 0.99)	(0.89, 0.99)	(0.95, 1.00)
2	(0.89, 0.99)	(0.89, 0.99)	(0.95, 1.00)
3	(0.95, 1.00)	(0.95, 1.00)	(0.71, 0.87)
4	(0.89, 0.99)	(0.89, 0.99)	(0.75, 0.90)
		$K = 16$	
1	(0.89, 0.99)	(0.89, 0.99)	(0.89, 0.99)
2	(0.89, 0.99)	(0.95, 1.00)	(0.84, 0.96)
3	(0.95, 1.00)	(0.95, 1.00)	(0.71, 0.87)
4	(0.95, 1.00)	(0.95, 1.00)	(0.80, 0.93)

proportion of correctly coclustered subject pairs,

$$\hat{\varkappa} = \frac{1}{\binom{N}{2}} \sum_{i_1 \neq i_2 \in (1,\ldots,N)} \mathcal{I}\{\mathcal{I}(\hat{c}_{i_1} = \hat{c}_{i_2}) = \mathcal{I}(c_{i_1}^{(0)} = c_{i_2}^{(0)})\},$$

where $c_1^{(0)}, \ldots, c_N^{(0)}$ are the true allocations. Table 6 displays summaries of these proportions for the mixture probabilities $\boldsymbol{\pi}_B^*$ and $\boldsymbol{\pi}_R^*$, previously used to define the four scenarios (e.g., see Table 2). Greater accuracy is achieved as K increases, suggesting that the true allocations are detected with greater precision along with the subject-specific transition probabilities.

6. Data analysis. We return to the motivating clinical trial data for the drug Exalgo (Norton (2011)), previously described in Section 1. The data were analyzed using the proposed model and inference procedures. We assume concentration parameter $\alpha = 1$, and baseline measure $G_0 = N_{16}(\mathbf{0}, 9\boldsymbol{I}_{16})$ for the Dirichlet process prior (3.5).

The performance of the proposed SBMM was evaluated using the log pseudomarginal likelihood (LPML) which has been extensively used in Bayesian model selection problems (Chen, Shao and Ibrahim (2000), Brown and Ibrahim (2003), Ghosh, Basu and Tiwari (2009), Ho et al. (2013)). For comparison, the LPML for the reference Bayesian method of Cui, Zhao and Tiwari (2016) was also evaluated. For the two methods and each study arm, Table 7

TABLE 6
Mean proportion of correctly co-clustered subject pairs with the standard error shown in parentheses

	$K = 8$	$K = 12$	$K = 16$
$\boldsymbol{\pi}_B^*$	0.775 (0.042)	0.854 (0.035)	0.889 (0.031)
$\boldsymbol{\pi}_R^*$	0.756 (0.042)	0.846 (0.036)	0.892 (0.031)

TABLE 7
LPML values for the Norton (2011) clinical trial data

	Treatment arm		Control arm	
	Nonwithdrawal	Withdrawal	Nonwithdrawal	Withdrawal
SBMM	−182.6	−204.2	−114.0	−251.7
Cui, Zhao and Tiwari (2016)	−217.4	−425.1	−143.5	−557.6

presents the LPML values for two groups of subjects: those who withdrew and those who did not withdraw from the study. Larger values of LPML indicate better model fit. We find that SBMM has a larger LPML in every combination of group and study arm. However, SBMM most dramatically outperforms the reference method for the subjects who withdrew from the study. The results in Table 7 convincingly demonstrate the advantages of the SBMM technique.

The differences between the performance of the two methods can be explained as follows. As previously demonstrated in Figure 1, a serious shortcoming of the Cui, Zhao and Tiwari (2016) approach is that it assigns high probabilities to the nonwithdrawal categories for subjects who have withdrawn from the study. This results in poor model fit and low LPML values in clinical trials with high withdrawal rates; for the motivating study, the withdrawal rates were 50.7% and 67.9% for the treatment and control arms, respectively. Table 8 displays the percentage of subjects in each group for whom the SBMM method had a higher LPML than the method of Cui, Zhao and Tiwari (2016). We conclude that the large proportion of subjects who withdrew from the study, along with the poorness of model fit for these subjects by the competing technique, is the key reason why SBMM convincingly outperforms the technique of Cui, Zhao and Tiwari (2016) in the motivating study. Since clinical trials often have high withdrawal rates, this feature of SBMM offers an important advantage in these investigations.

The proposed SBMM flexibly analyzes subject-level category effects and is thus able to provide personalized BR measures defined in equation (2.1). For example, Figure 6 presents 95% posterior credible intervals of the measures BR_{L}, $\log(\mathrm{BR}_{\mathrm{R}})$ and $\log(\mathrm{BR}_{\mathrm{CR}})$, aggregated over the eight visits, for subject $i = 122$ belonging to the control arm.

To evaluate whether the drug Exalgo provides improved benefits compared to the control, we evaluated the posterior distributions of the BR measures in each arm and of the ABRD measures, as described in Section 4.2. The left panel of Figure 7 plots the posterior means and 95% posterior intervals for benefit–risk measures BR_{L}, $\log(\mathrm{BR}_{\mathrm{R}})$ and $\log(\mathrm{BR}_{\mathrm{CR}})$ for each study arm. BR_{L} and $\log(\mathrm{BR}_{\mathrm{R}})$ give similar results for both arms; however, $\log(\mathrm{BR}_{\mathrm{CR}})$ gives negative posterior means. This is probably due to high withdrawal rate of the study, since $\log(\mathrm{BR}_{\mathrm{CR}})$ emphasizes a ratio between the most beneficial (transitions to Category 1) and the most risky transitions (transitions to Category 5). The plot shows that the benefits are substantially greater in the treatment arm. To compare the benefit–risk measures between the

TABLE 8
Percentage of subjects for whom SBMM had higher LPML values than Cui, Zhao and Tiwari (2016)

Treatment arm		Control arm	
Nonwithdrawal	Withdrawal	Nonwithdrawal	Withdrawal
40.91%	100%	41.86%	97.80%

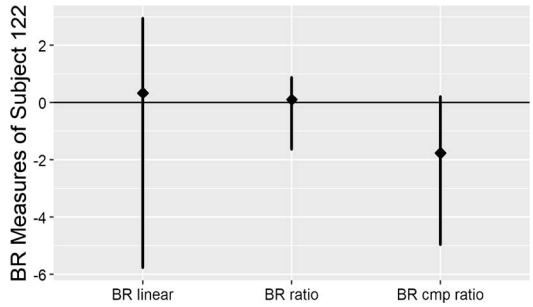

FIG. 6. *For subject* 122 *belonging to the control arm*, 95% *posterior credible intervals for personalized BR measures of expression* (2.1).

treatment and control, we constructed 95% posterior credible intervals for $\mathrm{ABRD_L}$, $\mathrm{ABRD_R}$ and $\mathrm{ABRD_{CR}}$ in the right panel of Figure 7. All three intervals are above zero, providing convincing evidence of greater benefits for the treatment arm.

The Dirichlet process specification of the SBMM model postulates that the population in each arm is an admixture of latent subpopulations consisting of shared probability transition matrices. The inference procedure discovered $M_1 = 4$ latent clusters in the control arm and $M_2 = 5$ latent clusters in the treatment arm. For the control arm the estimated mixture probabilities of the latent clusters are 0.515, 0.380, 0.045 and 0.060. The estimated mixture probabilities of the first four latent clusters of the treatment arm are 0.687, 0.172, 0.097 and 0.037, collectively accounting for 99.3% of the treatment population. The following discussion focuses on these eight cluster–arm combinations.

To better assist the interpretation of the detected clusters, the BR measures, defined in expression (2.1), were calculated for each latent cluster. In Figure 8 the measures $\mathrm{BR_L}$, $\log(\mathrm{BR_R})$ and $\log(\mathrm{BR_{CR}})$ averaged over the cluster members are plotted for the detected clusters in each arm. For treatment arm cluster 1, constituting 68.7% of the treatment arm patient population, the benefit strongly outweighed the risk for measures $\mathrm{BR_L}$ and $\log(\mathrm{BR_R})$. For treatment arm clusters 2, 3 and 4, the differences were less pronounced for linear measure $\mathrm{BR_L}$. In contrast, for control arm clusters 2 and 4 the risks outweighed the benefits; for control arm clusters 1 and 3 the BR measures are negative but less evidently different from 0.

These results provide strong evidence that the benefits exceed the risks in the treatment arm. Along with the uncertainty estimates, the results are reliable because the proposed method appropriately adjusts for departures from parametric forms and borrows strength among the subjects via the Dirichlet process. Additionally, SBMM accounts for within-

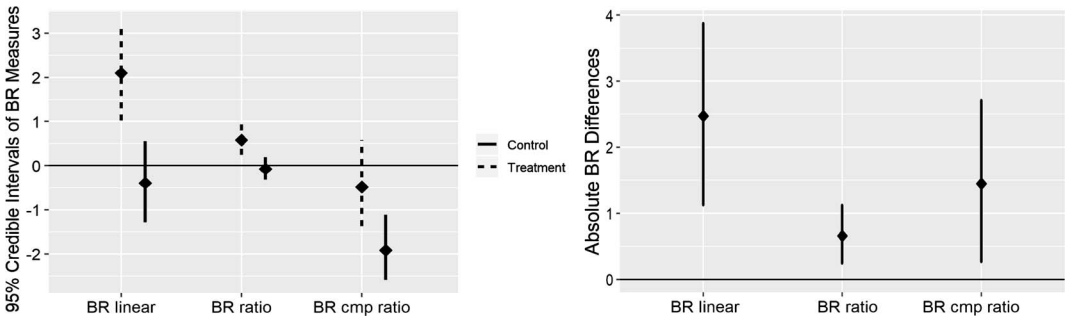

FIG. 7. *The left panel displays* 95% *posterior credible intervals for BR measures of expression* (2.1) *marginalized over all subjects in each arm. Black diamonds indicate posterior means. The right panel displays* 95% *posterior credible intervals for the three ABRD measures of expression* (2.2).

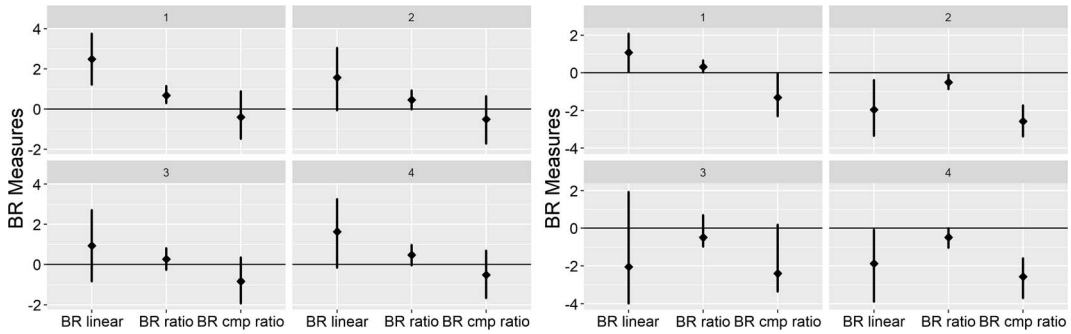

FIG. 8. *BR measures for the four large latent clusters in the treatment arm* (*left panel*) *and the four clusters in the control arm* (*right panel*).

subject longitudinal dependencies via the Markov process to provide accurate inferences about the benefits and risks of the drug Exalgo.

7. Discussion. This paper proposes an innovative semiparametric framework incorporating flexible unsupervised learning of the patient population in a clinical trial. It thereby fosters a paradigm for future studies of personalized benefit–risk assessment. Specifically, we present a novel application of a Markov model and the Dirichlet process by modeling the transition probabilities and exploiting the clustering property, for estimating personalized benefit–risk measures and for comparing aggregated benefit–risk measures in randomized clinical trials. The analysis is carried out by modeling homogeneous transition probabilities at the logit scale in a generalized linear mixed model. Common features of subject-level transition probabilities are characterized by the Dirichlet process. This model can be further generalized to modeling nonhomogeneous transition probabilities, such as including additional visit effects. The model can also consider including baseline covariates and other factors, such as regional effects, if the trial was conducted in multiple regions.

The modeling of transition probabilities incorporates longitudinal dependence among benefit–risk categories at subject's level and takes the nature of withdrawal category into account by treating it as an absorbing state. The use of the Dirichlet process to model the subject-level transition probabilities facilitates the handling of overparameterization by putting the subjects in distinct but smaller number of clusters. To let the data play a bigger role, we chose the mass parameter α to be 1; however, one can let α have its own prior distributions, such as an independent inverse-gamma or truncated normal on the positive real line.

The proposed SBMM method is based on the benefit–risk scores using categorical outcome which take a set of inherently continuous measurements of both benefit and risk endpoints and collapse all that into benefit–risk categories. This potentially could lead to loss of information, primarily due to the boundaries between benefit and no-benefit and between risk and no-risk. One may define benefit and risk in a finer gradation, for example, large benefit, reasonable benefit and no-benefit as benefit category, and severe risk, less severe risk and no-risk as risk category. Nevertheless, a Bayesian Markov model can be applied in a similar fashion.

The choice of scores remains one of the most challenging areas of the quantitative benefit–risk evaluation. Scores should be appropriately adjusted to the disease, the patient population and the treatments under consideration. In this article we used the scores for benefit–risk measures as suggested in Chuang-Stein, Mohberg and Sinkula (1991). The selection of scores may be subjective and requires insights from the clinicians and other subject experts. As suggested by Ho et al. (2016), which discussed an integrated benefit–risk measure, scores

TABLE 9
Effect of transitions when ties are not allowed. *"+" indicates improvement and "−" indicates deterioration*

	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5
Cat. 1	+	−	−	−	−
Cat. 2	+	+	−	−	−
Cat. 3	+	+	−	−	−
Cat. 4	+	+	+	−	−

may be determined based on respective preference and measures for uncertainty aversion. One may also use data-dependent scores or treat scores as random quantities.

A natural question is whether the proposed methodology is able to accommodate ties among the ordered categories of the outcomes. This can indeed be achieved by minor modifications to the expressions for the BR measures in equation (2.1). For example, the equation of $\mathrm{BR}_{\mathrm{L}}(\boldsymbol{P})$ in expression (2.1) implicitly does not allow ties. Table 9 shows the effect of each transition from one category to another on the BR measures when the ties are not allowed. For example, transitions from Category 1 to 1 are considered "improvement" (or at least, not getting worse) and have a positive coefficient in calculating the BR measures. However, transitions from Category 1 to any other categories are considered "deterioration" and have a negative coefficient in calculating the BR measures.

On the other hand, if we wish to allow ties among the ordered categories, the most ideal candidates would be between Category 2 and 3. These two may be considered similar in terms of BR profile because Category 2 represents "benefit with AE" and Category 3 represents "no benefit, no AE". Table 10 shows the effect of each transition from one category to another on BR measures when ties are allowed between Category 2 to 3. We find that transitions from Category 2 to 3 and from Category 3 to 3, now have a positive effect on BR measures (instead of a negative effect) because Category 2 and 3 are considered similar in terms of BR profile. The linear BR measure is then modified as follows:

$$\mathrm{BR}_{\mathrm{L}}(\boldsymbol{P}) = w_{11}p_{11} + \sum_{s=2}^{3}\sum_{j=1}^{\mathbf{3}} w_{sj}p_{sj} + \sum_{j=1}^{3} w_{4j}p_{4j} - \sum_{j=2}^{5} w_{1j}p_{1j} - \sum_{s=2}^{3}\sum_{j=\mathbf{4}}^{5} w_{sj}p_{sj} - \sum_{j=4}^{5} w_{4j}p_{4j}.$$

Ties among other categories could be accommodated by similar changes to the expressions for the BR measures.

In this article we have treated "minor irritation" and "severe irritation" as the same level of AE. However, we could further characterize AE into a more detailed gradation, according

TABLE 10
Effect of transitions when ties are allowed. *"+" indicates improvement and "−" indicates deterioration*

	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5
Cat. 1	+	−	−	−	−
Cat. 2	+	+	+	−	−
Cat. 3	+	+	+	−	−
Cat. 4	+	+	+	−	−

to the nature of the condition and the treatment. This would introduce more categories in Table 1 and would result in more transition probabilities that have to be estimated from the data.

Finally, the proposed SBMM technique can be extended for the benefit–risk measures based on continuous endpoints, such as mean reduction in cholesterol level from baseline to the end of the trial.

APPENDIX: RUNNING THE R CODE FOR SBMM

The R code for SBMM is available under Supplementary Material (Yan et al. (2020)). To replicate the analysis, simply run "Exalgo.R." This is the master file that calls the necessary functions in the other .R files.

Acknowledgments. D. Yan and S. Guha are supported by the National Science Foundation under Award DMS-1854003.

SUPPLEMENTARY MATERIAL

Supplement to "Semiparametric Bayesian Markov analysis of personalized benefit–risk assessment" (DOI: 10.1214/20-AOAS1323SUPP; .zip). R Code for SBMM.

REFERENCES

BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355. MR0362614

BROWN, E. R. and IBRAHIM, J. G. (2003). Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics* **59** 686–693. MR2004274 https://doi.org/10.1111/1541-0420.00079

CAMERLENGHI, F., LIJOI, A., ORBANZ, P. and PRÜNSTER, I. (2019). Distribution theory for hierarchical processes. *Ann. Statist.* **47** 67–92. MR3909927 https://doi.org/10.1214/17-AOS1678

CHEN, M.-H., SHAO, Q.-M. and IBRAHIM, J. G. (2000). *Monte Carlo Methods in Bayesian Computation. Springer Series in Statistics*. Springer, New York. MR1742311 https://doi.org/10.1007/978-1-4612-1276-8

CHIB, S. and GREENBERG, E. (1994). Bayes inference in regression models with ARMA(p, q) errors. *J. Econometrics* **64** 183–206. MR1310523 https://doi.org/10.1016/0304-4076(94)90063-9

CHIB, S., GREENBERG, E. and WINKELMANN, R. (1998). Posterior simulation and Bayes factors in panel count data models. *J. Econometrics* **86** 33–54.

CHUANG-STEIN, C., MOHBERG, N. R. and SINKULA, M. S. (1991). Three measures for simultaneously evaluating benefits and risks using categorical data from clinical trials. *Stat. Med.* **10** 1349–1359. https://doi.org/10.1002/sim.4780100904

CLAIBORNE, A. B., ENGLISH, R. A. and CARUSO, D. (2014). *Characterizing and Communicating Uncertainty in the Assessment of Benefits and Risks of Pharmaceutical Products*: *Workshop Summary*. National Academies Press, Washington, DC.

COSTA, M. J. and DRURY, T. (2018). Bayesian joint modelling of benefit and risk in drug development. *Pharm. Stat.* **17** 248–263.

COSTA, M. J., HE, W., JEMIAI, Y., ZHAO, Y. and DI CASOLI, C. (2017). The case for a Bayesian approach to benefit–risk assessment: Overview and future directions. *Ther. Innov. Regul. Sci.* **51** 568–574.

CUI, S., ZHAO, Y. and TIWARI, R. C. (2016). Bayesian approach to personalized benefit–risk assessment. *Stat. Biopharm. Res.* **8** 316–324.

DAHL, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics* 201–218. Cambridge Univ. Press, Cambridge.

ENTSUAH, R. and GORMAN, J. M. (2002). Global benefit–risk assessment of antidepressants: Venlafaxine XR and fluoxetine. *J. Psychiatr. Res.* **36** 111–118.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

GELBER, R. D., GELMAN, R. S. and GOLDHIRSCH, A. (1989). A quality-of-life-oriented endpoint for comparing therapies. *Biometrics* **45** 781–795.

GHOSAL, S., GHOSH, J. K. and RAMAMOORTHI, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27** 143–158. MR1701105 https://doi.org/10.1214/aos/1018031105

GHOSH, P., BASU, S. and TIWARI, R. C. (2009). Bayesian analysis of cancer rates from SEER program using parametric and semiparametric joinpoint regression models. *J. Amer. Statist. Assoc.* **104** 439–452. MR2751429 https://doi.org/10.1198/jasa.2009.0038

GLASZIOU, P. P., SIMES, R. J. and GELBER, R. D. (1990). Quality adjusted survival analysis. *Stat. Med.* **9** 1259–1276.

GØTZSCHE, P. C. and JØRGENSEN, A. W. (2011). Opening up data at the European Medicines Agency. *BMJ, Br. Med. J.* **342**.

GUHA, S. (2008). Posterior simulation in the generalized linear mixed model with semiparametric random effects. *J. Comput. Graph. Statist.* **17** 410–425. MR2439966 https://doi.org/10.1198/106186008X319854

HO, M.-W., TU, W., GHOSH, P. and TIWARI, R. C. (2013). A nested Dirichlet process analysis of cluster randomized trial data with application in geriatric care assessment. *J. Amer. Statist. Assoc.* **108** 48–68. MR3174602 https://doi.org/10.1080/01621459.2012.734164

HO, M., SAHA, A., MCCLEARY, K. K., LEVITAN, B., CHRISTOPHER, S., ZANDLO, K., BRAITHWAITE, R. S. and HAUBER, A. B. (2016). A framework for incorporating patient preferences regarding benefits and risks into regulatory assessment of medical technologies. *Value Health* **19** 746–750.

HOLDEN, W. L., JUHAERI, J. and DAI, W. (2003). Benefit–risk analysis: A proposal using quantitative methods. *Pharmacoepidemiol. Drug Saf.* **12** 611–616.

HUGHES, D., WADDINGHAM, E., MT ISA, S., GOGINSKY, A., CHAN, E., DOWNEY, G. F., HALLGREEN, C. E., HOCKLEY, K. S., JUHAERI, J. et al. (2016). Recommendations for benefit–risk assessment methodologies and visual representations. *Pharmacoepidemiol. Drug Saf.* **25** 251–262.

IBRAHIM, J. G. and CHEN, M.-H. (2000). Power prior distributions for regression models. *Statist. Sci.* **15** 46–60. MR1842236 https://doi.org/10.1214/ss/1009212673

ISHWARAN, H. and ZAREPOUR, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statist. Sinica* **12** 941–963. MR1929973

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 https://doi.org/10.2307/1390653

NORTON, J. D. (2011). A longitudinal model and graphic for benefit–risk analysis, with case study. *Drug Inf. J.* **45** 741–747.

PAYNE, J. T. and LOKEN, M. K. (1975). A survey of the benefits and risks in the practice of radiology. *CRC Crit. Rev. Clin. Radiol. Nucl. Med.* **6** 425–439.

PRITCHETT, Y. and TAMURA, R. (2008). The application of global benefit–risk assessment in clinical trial design and some statistical considerations. *Pharm. Stat.* **7** 170–178.

RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2008). The nested Dirichlet process. *J. Amer. Statist. Assoc.* **103** 1131–1144. MR2528831 https://doi.org/10.1198/016214508000000553

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SETHURAMAN, J. and TIWARI, R. C. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Statistical Decision Theory and Related Topics, III, Vol. 2 (West Lafayette, Ind., 1981)* 305–315. Academic Press, New York. MR0705321

SHAHBABA, B. and NEAL, R. (2009). Nonlinear models using Dirichlet process mixtures. *J. Mach. Learn. Res.* **10** 1829–1850. MR2540778

YAN, D., GUHA, S., AHN, C. and TIWARI, R. C. (2020). Suppelment to "Semiparametric Bayesian Markov analysis of personalized benefit–risk assessment." https://doi.org/10.1214/20-AOAS1323SUPP.

ZEGER, S. L. and KARIM, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86. MR1137101

ZHAO, Y., ZALKIKAR, J., TIWARI, R. C. and LAVANGE, L. M. (2014). A Bayesian approach for benefit–risk assessment. *Stat. Biopharm. Res.* **6** 326–337.