CAUSAL EFFECTS FROM PANEL DATA IN RANDOMIZED EXPERIMENTS WITH PARTIAL COMPLIANCE

Siddhartha Chib and Liana Jacobi

ABSTRACT

We present Bayesian models for finding the longitudinal causal effects of a randomized two-arm training program when compliance with the randomized assignment is less than perfect in the training arm (but perfect in the non-training arm) for reasons that are potentially correlated with the outcomes. We deal with the latter confounding problem under the principal stratification framework of Sommer and Zeger (1991) and Frangakis and Rubin (1999), and others. Building on the Bayesian contributions of Imbens and Rubin (1997), Hirano et al. (2000), Yau and Little (2001) and in particular Chib (2007) and Chib and Jacobi (2007, 2008), we construct rich models of the potential outcome sequences (with and without random effects), show how informative priors can be reasonably formulated, and present tuned computational approaches for summarizing the posterior distribution. We also discuss the computation of the marginal likelihood for comparing various versions of our models. We find the causal effects of the observed intake from the predictive distribution of each potential outcome for

Bayesian Econometrics
Advances in Econometrics, Volume 23, 183–215

ISSN: 0731-9053/doi:10.1016/S0731-9053(08)23006-9

compliers. These are calculated from the output of our estimation procedures. We illustrate the techniques and ideas with data from the 1994 JOBS II trial that was set up to test the efficacy of a job training program on subsequent mental health outcomes.

1. INTRODUCTION

We present Bayesian models for finding the longitudinal causal effects of a randomized two-arm training program when compliance with the randomized assignment is less than perfect in the training arm (but perfect in the non-training arm) for reasons that are potentially correlated with the outcomes. We deal with the latter confounding problem under the principal stratification framework of Sommer and Zeger (1991) and Frangakis and Rubin (1999), further discussed and applied in Imbens and Rubin (1997), Hirano, Imbens, Rubin, and Zhou (2000), Jo (2002), Yau and Little (2001), Ten Have, Joffe, and Cary (2003), Frangakis et al. (2004), Levy, O'Malley, and Normand (2004), and Mealli, Imbens, Ferro, and Biggeri (2004). In this framework, as explained in detail in Chib and Jacobi (2008), the confounder is assumed to be a (partially observable) latent variable that represents subject type, where subject type can take one of the four values – complier, never-taker, always-taker, and defier – defined in terms of the potential intake for each level of the assignment. Under certain assumptions, most importantly, the absence of always-takers (because these cannot be identified in our partial compliance setup where subjects in the control arm have no possibility of getting the training), the absence of defiers (the monotonicity assumption), and the exclusion restriction (that the assignment variable is a proper instrumental variable that has no direct affect on the outcomes), it becomes possible to find the effect of the actual intake on the outcome for the subclass (or strata) of compliers.

In this paper we discuss how this framework can be modified to the case of panel outcomes. We take a Bayesian approach because there is much that the Bayesian perspective can offer in this context, following the developments reported in Chib (2007) and Chib and Jacobi (2007, 2008). In particular, the Bayesian perspective offers the means to develop rich (parameter-heavy) models of the potential outcomes conditioned on subject type. In this modeling it is also possible to include random effects that vary by subject type. One reason that it is possible to specify rich models of the potential outcomes is because one can include prior information about the parameters in the analysis. For instance, we discuss how information from

another sample of subjects can be used to formulate beliefs about the time-varying (intake and subject-type specific) regression coefficients and intake and subject-type specific covariance matrices. Another reason that the Bayesian perspective is helpful is because it provides a well-established way of dealing with the mixture model that emerges for subjects in the control arm (mixed over the two possible types of subjects in that case, compliers and never-takers). Mixture models are particularly well handled from the Bayesian perspective by simply including the latent subject type of each subject as an additional parameter in the prior-posterior analysis. The label-switching problem that arises in mixture models does not occur in this problem because subject type (under our assumptions) is observed for subjects in the treatment arm who forgo the treatment (these being the never-takers) and for those in the treatment arm who take the treatment (these being the compliers). In contrast, frequentist fitting of the same model is more difficult because mixture models (even with latent type partially observed for some subjects) are not as easy to deal with, especially when there are many parameters (as in our problem) and random effects. Yet another appeal of the Bayesian approach is that it provides the means to calculate the causal effect from a predictive perspective. This perspective is particularly helpful because it leads to various summaries of the causal effects, for instance quantile casual effects, that are not as easily obtained by either a non-predictive formulation or non-Bayesian methods. Finally, the Bayesian approach provides a coherent procedure for comparing various versions of our models through the computation of marginal likelihoods and Bayes factors. We use this method to compare two versions of our panel data causal models, one with random effects and one without. Comparisons of this type are more difficult from the frequentist tradition.

The only previous discussion of the principal stratification framework in the panel context is by Yau and Little (2001). This paper is also from the Bayesian perspective and is motivated by the same data that we analyze in this paper. But apart from those connections, the treatment in this paper is different on the following dimensions:

1. *Modeling*: Our modeling of the potential outcome allows for subject and time specific shocks, whereas the modeling in Yau and Little does not. In their case, therefore, there is no issue about modeling the joint distribution of the potential outcomes since the potential outcomes are generated from the same shocks. In our case, this issue is relevant. However, the recent work of Chib (2007) has shown that the joint distribution of the potential outcomes does not have to be modeled in

causal models. This leads to a considerable simplification in the modeling especially in the context of panel data and type specific distributions where the joint distribution of the potential outcomes can be very high dimensional. This complication can be bypassed as we show here and simplifies both the modeling and the subsequent estimation of the model. This same point of Chib (2007) is utilized to advantage in the panel data model of Chib and Jacobi (2007).

2. *Prior*: Whereas Yau and Little (2001) use diffuse, improper priors, we adopt informative priors that are constructed in a reasoned way from another sample of subjects that were exposed to the same experiment.
3. *Random effects*: We propose and estimate models with random effects. Such models were not analyzed by Yau and Little (2001) but are natural in the context of panel data for dealing with individual specific influences.
4. *Inference*: Although we also proceed by Bayesian means, and summarize the posterior distribution by MCMC methods, the actual fitting approaches we develop are quite different from those used in Yau and Little (2001).
5. *Model comparisons*: Unlike Yau and Little (2001), we go beyond the problem of estimation and consider the question of model comparisons by marginal likelihoods and Bayes factors that we estimate by the method of Chib (1995).
6. *Causal effects*: Finally, our calculation of the causal effects is different and is based on a predictive perspective that provides a more complete summary of these effects than the complier-average causal effects that are reported by Yau and Little (2001).

The rest of the paper is organized as follows. In Section 2 we briefly discuss the data set that we analyze in this paper. This helps to fix the context for the developments we then provide in the remainder of the paper. In Section 3 we present the Bayesian formulation of the principal stratification approach for the panel context and describe two models that we think are useful. For each model, we also discuss our prior distribution. In Section 4 we discuss how the posterior distribution from each of our models can be summarized by MCMC methods, and how the marginal likelihood of the models can be computed. We then present results for various versions of our models that are defined by different assumptions about the error distributions. Section 5 deals with our predictive approach for calculating the causal effects, while Section 6 has our conclusions. Details of the fitting methods are given in the appendix.

2. DATA

As motivation for the model and problem we are going to consider, we discuss a data set that we will analyze below. The data comes from the 1994 JOBS II trial that was set up to test the efficacy of a job training program (see Vinokur, Price, & Schul, 1995 for a detailed description) on subsequent mental health outcomes. In the experiment, recently unemployed subjects were randomized to participate in a job training program with specific components to promote self-esteem and sense of control, job search skills and inoculation against setbacks. Those randomized into the control arm of the experiment received a booklet on job search skills that was also distributed among the treatment arm subjects after the training program. One question to be addressed by the trial was whether a training program can alleviate the negative mental health effects that are commonly associated with job loss (Clark & Oswald, 1994). The mental health of all subjects was evaluated through questionnaires at the start of the experiment and then again 2 months, 6 months, and 2 years after the start of the experiment. Subjects rated various stress symptoms from an 18-item index, each on a scale from 1 to 5. This information was used to construct a continuous outcome variable for the change in the mental health over time, measured in terms of the change in the depression score at each follow-up period compared to the baseline score.

Table 1 gives the sample means and standard deviations for the changes in depressions scores for the three periods. The table also provides a sample summary in terms of other health related variables and personal characteristics

Table 1. Sample Means and Standard Deviations of Our Study Data from the JOBS II Intervention Project.

Variable	Explanation	Mean	Standard Deviation
y_1	Change in depression score ($t = 1$)	−0.36	0.71
y_2	Change in depression score ($t = 2$)	−0.47	0.76
y_3	Change in depression score ($t = 3$)	−0.49	0.78
Depress_0	Baseline depression score	2.44	0.30
Risk_0	Baseline risk score	1.67	0.21
Age	Age in years	37.16	10.27
Motivate	Motivation to attend	5.30	0.80
Edu	School grade completed	13.43	2.05
Assert	Assertiveness	2.98	0.91
Marr	Marriage indicator	0.60	
Econ	Economic hardship	3.54	0.87
Nonw	Indicator for non-white	0.17	

that are used in the modeling of the outcome data. The information in the table refers to a sample of 387 subjects that were classified as being at high risk of depression at the start of the experiment and were observed in all follow-up periods. We have excluded subjects with a low risk of depression since no training effects were found for this group in previous studies.

3. BAYESIAN MODELING

We need the following notation. For each subject i ($i \leq n$) in the sample, let

- $z_i = l$ ($l = 0, 1$) denote the random assignment indicator, with $l = 0$ indicating assignment into the no-training or control arm and $l = 1$ assignment into the training or treatment arm.
- $x_{li} = j$ ($j = 0, 1$) denote the potential intake when $z_i = l$, with $j = 0$ indicating the no-training intake and $j = 1$ indicating receipt of training.
- $x_i = j$ ($j = 0, 1$) denote the actual intake given by

$$x_i = x_{0i}(1 - z_i) + x_{1i}z_i$$

  In the partial compliance setup we are dealing with, $x_i = 0$ if $z_i = 0$, whereas x_i can be 0 or 1 when $z_i = 1$.
- $\mathbf{y}_{ji} = (y_{ji1}, y_{ji2}, y_{ji3})$ denote the vector of potential outcomes when the treatment intake at baseline is j.
- $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3})$ denote the actual response given by

$$\mathbf{y}_i = \mathbf{y}_{0i}(1 - x_i) + \mathbf{y}_{1i}x_i$$

Now let s_i be an unobserved binary confounder that takes the values $k \in \{0, 1\}$, where $k = 0$ represents a never-taker and $k = 1$ a complier. Formally, a subject is a never-taker if $x_{0i} = x_{1i} = 0$, and a complier if $x_{0i} = 0$ and $x_{1i} = 1$. Under the assumption that no other subject types exist, a person with ($z_i = 0$, $x_i = 0$) is either a never-taker or a complier, a person with ($z_i = 1$, $x_i = 0$) is a never-taker, and a person with ($z_i = 1$, $x_i = 1$) is a complier. Table 2 gives the distribution of these types in our sample by assignment and intake. Only 159 of the 260 subjects randomized into the treatment actually participated in the training program. These numbers reflect the general compliance problem that is common in such trials.

Following Chib and Jacobi (2008), the modeling of this problem requires a specification of the joint distribution

$$p(\mathbf{y}_i, x_i = j|\mathbf{W}_i, z_i = l, s_i = k) \equiv p(\mathbf{y}_{ji}, x_i = j|\mathbf{W}_i, z_i = l, s_i = k) \quad (1)$$

$$= p_j(\mathbf{y}_i|\mathbf{W}_i, s_i = k)\Pr(x_i = j|\mathbf{y}_{ij}, \mathbf{W}_i, z_i = l, s_i = k) \quad (2)$$

Table 2. Distribution of the Sample Subjects and their Types by Treatment Assignment and Intake.

	No Training $x = 0$	Training $x = 1$
Control arm $z = 0$	$n_{00} = 127$ (compliers and never-takers)	–
Treatment arm $z = 1$	$n_{10} = 101$ (never-takers)	$n_{11} = 159$ (compliers)

where $p_j(\mathbf{y}_i|\mathbf{W}_i, s_i = k)$ is the density of $\mathbf{y}_{ji}$ conditional on the latent subject type and the second term is the conditional mass function of $x_i = j$. The former density does not involve $z_i = l$ on account of the so-called exclusion restriction. Notice, too, that the second term is either 0 or 1, for any value of $\mathbf{y}_i$ or $\mathbf{W}_i$. For example, if $z_i = 0$ and $s_i = 1$, then $x_i = 0$, so that $\Pr(x_i = 0|\mathbf{y}_i, \mathbf{W}_i, z_i = 0, s_i = 1) = 1$. In addition, if $z_i = 1$ and $s_i = 0$, then $x_i = 0$, implying that $\Pr(x_i = 0|\mathbf{y}_i, \mathbf{W}_i, z_i = 1, s_i = 0) = 1$. Thus, given $z_i = l$ and $s_i = k$, the intake is fully determined. It is important to keep in mind that there is no need to model the joint density

$$p(\mathbf{y}_{0i}, \mathbf{y}_{1i}, x_i = j|\mathbf{W}_i, z_i = l, s_i = k)$$

which is actually unidentified because the potential outcomes $(\mathbf{y}_{0i}, \mathbf{y}_{1i})$ are not observed simultaneously. That the modeling and subsequent estimation of the model can proceed without this joint distribution is due to Chib (2007).

To model the joint density of the outcome and the intake, let $I_{lj} = \{i: z_i = l \text{ and } x_i = j\}$ denote the sample indices of the subjects in each of the three non-empty cells of Table 2. Also, let $\Pr(s_i = 1|\mathbf{v}_i) = q_{ci}$ denote the probability that a subject is of type c, which we assume is a function of the $q \times 1$ vector of pre-treatment variables $\mathbf{v}_i$ that is a subset of $\mathbf{W}_i$. This probability is independent of z_i because of the random assignment of subjects to the treatment arms. However, since we do not observe the subject type in the control arm, the joint density of y_i and $x_i = j$ conditional on $z_i = l$ is given by appropriately averaging over possible types:

$$p(\mathbf{y}_i, x_i = j|\mathbf{W}_i, z_i = l) = \begin{cases} (1 - q_{ci})p_0(\mathbf{y}_i|\mathbf{W}_i, s_i = 0) + q_{ci}p_0(\mathbf{y}_i|\mathbf{W}_i, s_i = 1) & \text{if } i \in I_{00} \\ (1 - q_{ci})p_0(\mathbf{y}_i|\mathbf{W}_i, s_i = 0) & \text{if } i \in I_{10} \\ q_{ci}p_1(\mathbf{y}_i|\mathbf{W}_i, s_i = 1) & \text{if } i \in I_{11} \end{cases} \tag{3}$$

This expression does not involve the mass function of the intake due to the discussion surrounding Eq. (1). Note also that z_i neither appears in the conditioning set of the exogenous type probability due the randomization

argument nor in that of the outcome distribution due to the exclusion restriction.

From expression (3) we see that the modeling of $(\mathbf{y}_i, x_i = j)$ requires three type and treatment state specific multivariate distributions for the health outcomes, $p_0(\mathbf{y}_i|\mathbf{W}_i, s_i = 0)$ and $p_j(\mathbf{y}_i|\mathbf{W}_i, s_i = 1)$, for $j = 0, 1$, and a model for the type probabilities q_{ci}. In the next section we introduce two model specifications that are based on different formulations for the intake and type specific distributions of the health outcomes. In each case we assume that the probability of being a complier, q_{ci}, is generated by a probit model. Previous papers such as Hirano et al. (2000), Jo (2002), Frangakis et al. (2004), and Chib and Jacobi (2008) have found that it is important to model the compliance probability in terms of baseline predictors. Here we follow Yau and Little (2001) and let

$$q_{ci} = \Phi(\mathbf{w}_{i0}'\boldsymbol{\alpha})$$

where $\mathbf{w}_{i0} = (1, \text{Age}, \text{Edu}, \text{Marr}, \text{Nonw}, \text{Assert}, \text{Motivate}, \text{Econ})$.

Our modeling is completed with a prior distribution on the parameters of the preceding distributions. We use standard distributional forms to compose the prior distribution. For example, we choose the normal distribution for the regression parameters and the Wishart distribution for the covariance matrices. A challenging component of the prior specification is the choice of hyperparameters. We deal with this problem by constructing a prior distribution that is reasonable for the sample of low-risk subjects that is excluded from our analysis. Our strategy is to set the hyperparameters, sample the prior, and then simulate the outcome distributions. We do this many times and see whether the simulated distribution of the outcomes is similar to the empirical distribution of the outcomes in the low-risk sample. If the two distributions are quite different, we revise our hyperparameters somewhat and repeat the process.

3.1. Model 1

One choice (which we call Model 1) is to let

$$\begin{aligned} p_0(\mathbf{y}_i|\mathbf{W}_i, s_i = 0) &= t_\nu(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n}, \boldsymbol{\Omega}_{0n}) \\ p_j(\mathbf{y}_i|\mathbf{W}_i, s_i = 1) &= t_\nu(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{jc}, \boldsymbol{\Omega}_{jc}), \quad j = 0, 1 \end{aligned} \tag{4}$$

where $\boldsymbol{\beta}_{0n}$ and $\boldsymbol{\beta}_{jc}$ are intake and type specific regression parameters, Ω_0 and $\boldsymbol{\Omega}_{jc}$ are the corresponding full (3×3) dispersion matrices and $t_\nu(\cdot|\mu, \Omega)$ is the

multivariate Student's t density function with v degrees of freedom, mean μ, and variance matrix $v\boldsymbol{\Omega}/(v-2)$, $v>2$. Equivalently, under the common representation of the Student's t distribution as a scale mixture of normal distributions, the latter model can be expressed as

$$\begin{aligned} p_0(\mathbf{y}_i|\mathbf{W}_i, s_i=0, \lambda_i) &= \mathcal{N}(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n}, \lambda_i^{-1}\boldsymbol{\Omega}_{0n}) \\ p_j(\mathbf{y}_i|\mathbf{W}_i, s_i=1, \lambda_i) &= \mathcal{N}(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{jc}, \lambda_i^{-1}\boldsymbol{\Omega}_{jc}), \quad j=0,1 \end{aligned} \tag{5}$$

where λ_i is distributed as gamma

$$\lambda_i \sim \mathcal{G}\left(\frac{v}{2}, \frac{v}{2}\right)$$

In our application, we parameterize the matrix $\mathbf{W}_i$ in a way to allow for time-varying effects for each of the covariates:

$$\mathbf{W}_i = \begin{pmatrix} 1 & \text{depress}_{i0} & \text{risk}_{i0} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \text{depress}_{i0} & \text{risk}_{i0} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \text{depress}_{i0} & \text{risk}_{i0} \end{pmatrix}$$

and denote the covariate vector as $\boldsymbol{\beta}_{jk} = (\boldsymbol{\beta}_{jk,1}, \boldsymbol{\beta}_{jk,2}, \boldsymbol{\beta}_{jk,3})$: 9×1 so that $\boldsymbol{\beta}_{jk,1}$ is the effect of the three predictors in the first time period, $\boldsymbol{\beta}_{jk,2}$ in the second time period, and $\boldsymbol{\beta}_{jk,3}$ in the third time period.

We specify the prior distribution for the vector of model parameters $\boldsymbol{\theta}$ as

$$\pi(\boldsymbol{\theta}) = \mathcal{N}_p(\boldsymbol{\alpha}|\boldsymbol{\alpha}_0, \mathbf{A}_0) \prod_{j=0}^{1} \prod_{k \in K_j} \mathcal{N}_{3k}(\boldsymbol{\beta}_{jk}|\boldsymbol{\beta}_{jk,0}, \mathbf{B}_{jk,0}) \mathcal{W}(\boldsymbol{\Omega}_{jk}^{-1}|\rho_{jk,0}, \mathbf{R}_{jk,0}) \tag{6}$$

and fix the prior means for $\boldsymbol{\beta}_{jk}$ and $\boldsymbol{\alpha}$:

$$\begin{aligned} \boldsymbol{\beta}_{0c,0} &= (0.8, -1.1, 0.8, 1.5, -1.5, 1.0, 1.4, -1.2, 0.6) \\ \boldsymbol{\beta}_{0n,0} &= (0.7, -1.1, 0.8, 1.5, -1.5, 1.0, 1.2, -1.2, 0.6) \\ \boldsymbol{\beta}_{1c,0} &= (0.8, -1.1, 0.8, 0.6, -1.5, 1.0, 1.6, -1.2, 0.6) \\ \boldsymbol{\alpha}_0 &= (-5, .03, .5, .1, 0, 0, 0, 0) \end{aligned}$$

As mentioned above we want a prior distribution that generates outcomes that are reasonable in relation to those seen in the low-risk sample. For this, we set $\mathbf{B}_{jk,0} = 9\mathbf{I}$ and $\mathbf{A}_0 = 9\mathbf{I}$ and set the hyperparameters of the Wishart prior for $\boldsymbol{\Omega}_{jk}^{-1}$ to imply a full covariance matrix with 0.5 on the diagonal and 0.25 for all off-diagonal elements. This seems a reasonable choice given our outcome variable that is measured in terms of the change in depression scores, each restricted to values between 0 and 5. To show what these

Table 3. Model 1 – Means and Quantiles of the Empirical Distribution of the Change in Depression Score Implied by the Assumed Prior.

	$y^s_{0.05}$	$y^s_{0.25}$	$y^s_{0.50}$	$y^s_{0.75}$	$y^s_{0.95}$
$t=1$	−6.74	−3.89	−0.59	2.73	5.48
$t=2$	−6.94	−4.05	−0.90	2.44	5.12
$t=3$	−6.63	−3.75	−0.47	2.77	5.47

assumptions imply for the outcomes, we provide in Table 3 the lower and upper quantiles of the simulated outcome distributions under this prior. We conclude that our prior assumptions are reasonably flexible.

3.2. Model 2

Another option, which we call Model 2, is to derive $p_0(\mathbf{y}_i|\mathbf{W}_i,\ s_i=0)$ and $p_j(\mathbf{y}_i|\mathbf{W}_i,\ s_i=1)$ from a random effects formulation. Let $\mathbf{V}_i$ be a $3\times k$ $(k<3)$ matrix of covariates whose effect on the outcome is individual specific. In this particular application, where all the covariates in the outcome model are measured at baseline, $\mathbf{V}_i$ is a vector of constants. To allow for flexibility in the covariance structure, as in Model 1, we assume that the random effects are intake and type specific. We denote these by $\mathbf{b}_{i0c}$, $\mathbf{b}_{i1c}$, and $\mathbf{b}_{in}$, one for compliers under no intake, another for compliers under treatment intake, and finally one for never-takers. Conditioned on the random effects, we now let

$$\begin{aligned} p_0(\mathbf{y}_i|\mathbf{W}_i, s_i=0, \mathbf{b}_{in}) &= t_\nu(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n}+\mathbf{V}_i\mathbf{b}_{in}, \mathrm{diag}(\sigma_{0n})) \\ p_j(\mathbf{y}_i|\mathbf{W}_i, s_i=1, \mathbf{b}_{ijc}) &= t_\nu(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{jc}+\mathbf{V}_i\mathbf{b}_{ijc}, \mathrm{diag}(\sigma_{jc})), \quad j=0,1 \end{aligned} \tag{7}$$

where the dispersion matrices are in diagonal form for identification reasons. Once again with the introduction of positive latent scale variables $\lambda_i \sim \mathcal{G}(\nu/2, \nu/2)$, we can express this model as

$$\begin{aligned} p_0(\mathbf{y}_i|\mathbf{W}_i, s_i=0, \mathbf{b}_{in}, \lambda_i) &= \mathcal{N}_3(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n}+\mathbf{V}_i\mathbf{b}_{in}, \lambda_i^{-1}\mathrm{diag}(\sigma_{0n})) \\ p_j(\mathbf{y}_i|\mathbf{W}_i, s_i=1, \mathbf{b}_{ijc}, \lambda_i) &= \mathcal{N}_3(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{jc}+\mathbf{V}_i\mathbf{b}_{ijc}, \lambda_i^{-1}\mathrm{diag}(\sigma_{jc})), \quad j=0,1 \end{aligned} \tag{8}$$

If we now assume that the random effects are distributed as

$$\begin{aligned} \mathbf{b}_{ijc}|\mathbf{D}_c &\sim \mathcal{N}_k(0, \mathbf{D}_{jc}), \quad j=0,1 \\ \mathbf{b}_{in}|\mathbf{D}_n &\sim \mathcal{N}_k(0, \mathbf{D}_n) \end{aligned}$$

where the matrices $\mathbf{D}_{jc}$ and $\mathbf{D}_n$ are unknown, it follows that marginalized over the random effects (but conditioned on λ_i) the distributions of the outcome by intake and type are given by

$$p_j(\mathbf{y}_i|\mathbf{W}_i, s_i = 1, \lambda_i) = \mathcal{N}_3(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{jc}, \sum\nolimits_{jc} = \{\lambda_i^{-1}\mathrm{diag}(\sigma_{jc}) + \mathbf{V}_i\mathbf{D}_{jc}\mathbf{V}_i'\}$$
$$p_0(\mathbf{y}_i|\mathbf{W}_i, s_i = 0, \lambda_i) = \mathcal{N}_3(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n}, \sum\nolimits_{0n} = \{\lambda_i^{-1}\mathrm{diag}(\sigma_{0n}) + \mathbf{V}_i\mathbf{D}_n\mathbf{V}_i'\})$$

whereas marginalized over λ_i these are

$$p_j(\mathbf{y}_i|\mathbf{W}_i, s_i = 1) = \int_0^\infty \mathcal{N}_3\left(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{jc}, \boldsymbol{\Sigma}_{jc}\right)\mathcal{G}\left(\lambda_i\middle|\frac{v}{2}, \frac{v}{2}\right)\mathrm{d}\lambda_i$$
$$p_0(\mathbf{y}_i|\mathbf{W}_i, s_i = 0) = \int_0^\infty \mathcal{N}_3\left(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n}, \boldsymbol{\Sigma}_{0n}\right)\mathcal{G}\left(\lambda_i\middle|\frac{v}{2}, \frac{v}{2}\right)\mathrm{d}\lambda_i \tag{9}$$

which differ from the ones in Eq. (4).

As in Model 1 we specify the prior distribution of the model parameters $\pi(\boldsymbol{\theta})$ as

$$\mathcal{N}_{\mathrm{p}}(\boldsymbol{\alpha}|\boldsymbol{\alpha}_0, \mathbf{A}_0)\prod_{j=0}^{1}\prod_{k\in K_j}\mathcal{N}_{3k}(\boldsymbol{\beta}_{jk}|\boldsymbol{\beta}_{jk,0}, \mathbf{B}_0)\left(\prod_{t=1}^{3}\mathcal{IG}\left(\sigma_{jk,t}\middle|\frac{v_{jk,0}}{2}, \frac{\delta_{jk,0}}{2}\right)\right)\mathcal{W}(\mathbf{D}_{jk}^{-1}|\rho_{jk,0}, \mathbf{R}_{jk,0})$$

where $K_0 = \{c, n\}$ and $K_1 = \{c\}$. The prior means and variances of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_{jk}$ are fixed at the same values as in Model 1, so that

$$\begin{aligned}
\boldsymbol{\beta}_{0c,0} &= (0.0, -1.0, 0.6, 1.0, -1.5, 1.0, 1.8, -1.2, 0.6)\\
\boldsymbol{\beta}_{0n,0} &= (0.0, -1.0, 0.6, 0.8, -1.5, 1.0, 0.5, -1.2, 0.6)\\
\boldsymbol{\beta}_{1c,0} &= (0.0, -1.0, 0.6, 0.7, -1.5, 1.0, 0.5, -1.2, 0.6)\\
\boldsymbol{\alpha}_0 &= (-5, .03, .5, .1, 0, 0, 0, 0)
\end{aligned}$$

and $\mathbf{B}_{jk,0} = 9\mathbf{I}$ and $\mathbf{A}_0 = 9\mathbf{I}$. The parameters of the inverse gamma prior on the scalar variances are set to imply means and standard errors of 0.5 and 3, respectively ($v_{jk,0} = 4.04$, $\delta_{jk,0} = 1.03$). Finally, for the Wishart prior on the inverse of the variances of the random effects we let $\rho_{jk,0} = 5$ and $\mathbf{R}_{jk,0} = 0.66\mathbf{I}$, which implies a prior mean of $0.5\mathbf{I}$ for $\mathbf{D}_{jk}$. As we had done in the case of Model 1, we simulate the outcomes under this prior. The resulting lower and upper quantiles of these outcome distributions are given in Table 4 and again appear to be reasonable and sufficiently flexible.

Table 4. Model 2 – Means and Quantiles of the Empirical Distribution of the Change in Depression Score Implied by the Assumed Prior.

	$y^s_{0.05}$	$y^s_{0.50}$	$y^s_{0.25}$	$y^s_{0.75}$	$y^s_{0.95}$
$t = 1$	−7.49	−4.26	−0.81	2.62	5.68
$t = 2$	−7.17	−3.90	−0.47	2.98	6.04
$t = 3$	−7.50	−4.26	−0.81	2.60	5.64

4. PRIOR-POSTERIOR ANALYSIS

4.1. Model 1

We now turn to the prior-posterior analysis of the first model. Our modeling assumptions imply that $p(\mathbf{y},\mathbf{x}|\boldsymbol{\theta},\{\lambda_i\})$, the joint density of the observed health outcomes $\mathbf{y} = \{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n\}$ and training intake data $\mathbf{x} = \{x_1,x_2,\ldots,x_n\}$ given the vector of model parameters and the scale parameters, is of the form

$$\prod_{i=1}^{N} \mathcal{G}\left(\lambda_i \middle| \frac{\nu}{2}, \frac{\nu}{2}\right) \prod_{i\in I_{00}} [(1-\Phi(\mathbf{w}'_{i0}\boldsymbol{\alpha}))\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n}, \boldsymbol{\Omega}_{0n}) + \Phi(\mathbf{w}'_{i0}\boldsymbol{\alpha})\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0c}, \boldsymbol{\Omega}_{0c})] \times \prod_{i\in I_{10}} (1-\Phi(\mathbf{w}'_{i0}\boldsymbol{\alpha}))\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n}, \boldsymbol{\Omega}_{0n}) \times \prod_{i\in I_{11}} \Phi(\mathbf{w}'_{i0}\boldsymbol{\alpha})\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{1c}, \boldsymbol{\Omega}_{1c}) \tag{10}$$

This joint density has three distinct components that correspond to the three non-empty cells in Table 2. The first term gives the likelihood contributions for the 127 subjects in the control arm, while the second and the third product terms provide the likelihood contributions for the 101 never-takers and the 159 compliers in the treatment arm, respectively. As the type is not observed for the first group, the likelihood contributions take the form of mixture distributions over compliers and never-takers. It may be noted that the mixture component is only present in the control arm since subject type is otherwise observed.

We handle the mixture terms in the control arm by including the latent subject type of each subject as an additional parameter in the prior-posterior analysis. The label-switching problem that arises in mixture models does not occur in this problem because subject type is observed for subjects in the

treatment arm who forgo the treatment and for those in the treatment arm who take the treatment. Let $\mathbf{s}_{00}$ denote the type indicators for control arm subjects. Then, our target posterior density of interest is $\pi(\boldsymbol{\theta}, \mathbf{s}_{00}, \{\lambda_i\}|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z})$, which is proportional to the prior density specified in Eq. (6) times the function

$$\begin{aligned}
&\prod_{i=1}^{N} \mathcal{G}\left(\lambda_i \Big| \frac{\nu}{2}, \frac{\nu}{2}\right) \prod_{i \in I_{00}} \{I[s_i = 0](1 - \Phi(\mathbf{w}'_{i0}\boldsymbol{\alpha}))\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n}, \lambda_i^{-1}\boldsymbol{\Omega}_{0n}) \\
&\quad + I[s_i = 1]\Phi(w'_{i0}\boldsymbol{\alpha})\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0c}, \lambda_i^{-1}\boldsymbol{\Omega}_{0c})\} \\
&\quad \times \prod_{i \in I_{10}} I[s_i = 0](1 - \Phi(\mathbf{w}'_{i0}\boldsymbol{\alpha}))\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n}, \lambda_i^{-1}\boldsymbol{\Omega}_{0n}) \\
&\quad \times \prod_{i \in I_{11}} I[s_i = 1]\Phi(\mathbf{w}'_{i0}\boldsymbol{\alpha})\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{1c}, \lambda_i^{-1}\boldsymbol{\Omega}_{1c})
\end{aligned}$$

We summarize this density by tuned MCMC methods (see Chib & Greenberg, 1995 for details on the Metropolis–Hastings algorithm and Chib, 2001 for an extended summary of MCMC methods). The sampling scheme involves three blocks and is summarized next. Full details are supplied in the appendix.

1. Sample $(\mathbf{s}_{00}, \boldsymbol{\alpha}, \{\lambda_i\}|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \Omega)$ by sampling
   (a) s_i for $i \in I_{00}$ with $\Pr(s_i = 1|\mathbf{y}_i, x_i, \boldsymbol{\beta}_{0c}, \boldsymbol{\beta}_{0n}, \boldsymbol{\alpha}, \Omega_{0c}, \Omega_{0n})$
   (b) $\boldsymbol{\alpha}|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \Omega, \mathbf{s}_{00}$ with a Metropolis–Hastings step
   (c) $\lambda_i|\mathbf{y}_i, x_i, \boldsymbol{\beta}, \Omega, \mathbf{s}_{00}$ for $i \in N$ from a gamma density
2. Sample $\{\boldsymbol{\beta}_{jk}\}$ by drawing $\boldsymbol{\beta}_{jk}|\mathbf{y}, \mathbf{x}, \mathbf{s}_{00}, \boldsymbol{\lambda}, \Omega, \boldsymbol{\alpha}$ from a normal density
3. Sample $\{\boldsymbol{\Omega}_{jk}^{-1}\}$ by drawing $\boldsymbol{\Omega}_{jk}^{-1}|\mathbf{y}, \mathbf{x}, \mathbf{s}_{00}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{D}$ from a Wishart density

In the first block, to produce a well mixing chain, the type indicators $\mathbf{s}_{00}$, $\boldsymbol{\alpha}$ and the scale parameters λ_i are sampled jointly by the method of composition. It is also possible to proceed by sampling the s_i's under the framework of Albert and Chib (1993). In the second block we update the coefficients $\boldsymbol{\beta}_{0c}$, $\boldsymbol{\beta}_{0n}$, and $\boldsymbol{\beta}_{0n}$. Under our model setup, the $\boldsymbol{\beta}_{jk}$'s depend on distinct subsets of the population, $(\mathbf{y}_{jk}$: $\{x_i = j, s_i = k\}$, $\mathbf{x}_{jk}$: $\{x_i = j, s_i = k\})$. We can therefore sample $\boldsymbol{\beta}_{jk}|\mathbf{y}_{jk}, \mathbf{x}_{jk}, \mathbf{s}_{00}, \boldsymbol{\alpha}, \Omega_{jk}$ separately from their respective normal posterior distribution. We proceed in a similar fashion to update Ω_{jk} in the last block of the chain and sample $\boldsymbol{\Omega}_{jk}^{-1}|\mathbf{y}_{jk}, \mathbf{x}_{jk}, \mathbf{s}_{00}, \boldsymbol{\alpha}, \boldsymbol{\beta}_{jk}$ from Wishart distributions.

4.2. Model 2

For the posterior analysis of Model 2, we augment the parameter space with the random effects $\{\mathbf{b}_{i0c}\}$, $\{\mathbf{b}_{i1c}\}$, and $\{\mathbf{b}_{in}\}$. To improve the tractability of the posterior distribution further we follow the same strategy as in Model 1 and include the type indicators $\mathbf{s}_{00} = \{s_i\colon i \in I_{00}\}$, and the latent scale parameters $\boldsymbol{\lambda} = \{\lambda_i\}$. The posterior density of interest is then $\pi(\boldsymbol{\theta}, \mathbf{s}_{00}, \{\mathbf{b}_{ijk}\}, \{\lambda_i\}|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z})$, which is proportional to the prior density times

$$
\begin{aligned}
&\prod_{i\in I_{00}} \big\{ I[s_i = 0](1 - \Phi(\mathbf{w}_{i0}'\boldsymbol{\alpha}))\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n} + \mathbf{V}_i\mathbf{b}_{in}, \lambda_i^{-1}\mathrm{diag}(\sigma_{0n}))\mathcal{N}_k(\mathbf{b}_{in}|0, \mathbf{D}_n) \\
&\quad + I[s_i = 1]\Phi(\mathbf{w}_{i0}'\boldsymbol{\alpha})\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0c} + \mathbf{V}_i\mathbf{b}_{ic}, \lambda_i^{-1}\mathrm{diag}(\sigma_{0c}))\mathcal{N}_k(\mathbf{b}_{i0c}|0, \mathbf{D}_{0c}) \big\} \\
&\quad \times \prod_{i\in I_{10}} I[s_i = 0](1 - \Phi(\mathbf{w}_{i0}'\boldsymbol{\alpha}))\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n} + \mathbf{V}_i\mathbf{b}_{in}, \lambda_i^{-1}\mathrm{diag}(\sigma_{0n}))\mathcal{N}_k(\mathbf{b}_{in}|0, \mathbf{D}_n) \\
&\quad \times \prod_{i\in I_{11}} I[s_i = 1]\Phi(\mathbf{w}_{i0}'\boldsymbol{\alpha})\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{1c} + \mathbf{V}_i\mathbf{b}_{ic}, \lambda_i^{-1}\mathrm{diag}(\sigma_{1c}))\mathcal{N}_k(\mathbf{b}_{i1c}|0, \mathbf{D}_{1c}) \\
&\quad \times \prod_{i=1}^{N} \mathcal{G}\left(\lambda_i \Big| \frac{v}{2}, \frac{v}{2}\right)
\end{aligned}
$$

In the appendix we provide a detailed description of the MCMC algorithm we have developed to generate draws from the posterior distribution. One important point is that in Step 1a we sample the compliance indicators marginalized over the random effects, which avoids having to sample the complier and never-taker random effects for each subject in the control arm. This reduces the computational burden considerably and improves the mixing of the MCMC chain. A short version of the algorithm is given here:

1. Sample $(\mathbf{s}_{00}, \boldsymbol{\alpha}, \{\boldsymbol{\beta}_{jk}\}, \{\mathbf{b}_{ijk}\}|\mathbf{y}, \mathbf{x}, \{\sigma_{jk}\}, \{\lambda_i\}, \{\mathbf{D}_{jk}\})$ by sampling
   (a) s_i for $i \in I_{00}$ with $\Pr(s_i = 1|\mathbf{y}_i, x_i, \boldsymbol{\beta}_{0c}, \boldsymbol{\beta}_{0n}, \boldsymbol{\alpha}, \sigma_{0c}, \sigma_{0n}, \{\lambda_i\}, \{\mathbf{D}_{jk}\})$
   (b) $\boldsymbol{\alpha}|\mathbf{y}_{jk}, \mathbf{x}_{jk}, \mathbf{s}_{00}$ by a MH step
   (c) $\boldsymbol{\beta}_{jk}|\mathbf{y}, \mathbf{x}, \mathbf{s}_{00}, \boldsymbol{\lambda}, \sigma_{jk}, \boldsymbol{\lambda}, \{\mathbf{D}_{jk}\}$ from a normal density
   (d) $\mathbf{b}_{jki}|\mathbf{y}_i, x_{\mathrm{i}}, \boldsymbol{\beta}_{jk}, \mathbf{s}_{00}, \lambda_i, \sigma_{jk}, \sigma_{jk}, \mathbf{D}_{jk}$ for $i \in I_{jk}$ from a normal density
2. Sample $\lambda_i|\mathbf{y}, \mathbf{x}, \{\boldsymbol{\beta}_{jk}\}, \{\sigma_{jk}\}, \boldsymbol{\alpha}, \mathbf{s}_{00}, \{\mathbf{b}_{jk}\}$ for $i \in N$ from a gamma density
3. Sample $\sigma_{jk}|\mathbf{y}, \mathbf{x}, \mathbf{s}_{00}, \boldsymbol{\lambda}, \boldsymbol{\beta}_{jk}, \boldsymbol{\alpha}, \{\mathbf{b}_{tjk}\}, \{\mathbf{D}_{jk}\}$ from an inverse gamma density
4. Sample $D_{jk}^{-1}|\mathbf{y}, \mathbf{x}, \mathbf{s}_{00}, \boldsymbol{\lambda}, \boldsymbol{\beta}_{jk}, \boldsymbol{\alpha}, \{\mathbf{b}_{ijk}\}, \sigma_{jk}$ from a Wishart distribution

4.3. Model Comparison

In practice one would be interested in comparing Models 1 and 2 and variations of these models to see which model is best supported by the data. We do this comparison from the marginal likelihood/Bayes factor perspective. Following Chib (1995), the log marginal likelihood of a given model can be expressed in terms of the logs of the likelihood and the prior and posterior distribution evaluated at $\boldsymbol{\theta}^*$ as

$$ln\, m(\mathbf{y}, \mathbf{x}) = \ln\, f(\mathbf{y}, \mathbf{x}|\mathbf{z}, \mathbf{W}, \boldsymbol{\theta}^*) + \ln\, \pi(\boldsymbol{\theta}^*) - \ln\, \pi(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, M)$$

where $\boldsymbol{\theta}^*$ is a vector of the model parameters given by (say) the posterior mean. The prior ordinate at $\boldsymbol{\theta}^*$ for models 1 and 2 can, of course, be computed directly from the respective prior densities. The likelihood ordinate for Model 1 can also be computed directly from the expression in Eq. (4). However, the likelihood of Model 2, marginalized over $\{\mathbf{b}_i\}$ and $\{\lambda_i\}$, is not available in closed form. Since the likelihood contribution conditional on $\{\lambda_i\}$ is in closed form, we employ an importance sampling approach to get $p'(\mathbf{y}_i|\mathbf{W}_i, s_i = k)$.

We now turn to the estimation of the posterior ordinates. For Model 1, with the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}_{0c}, \boldsymbol{\beta}_{0n}, \boldsymbol{\beta}_{1c}, \Omega_{0c}, \Omega_{0n}, \Omega_{1c}, \boldsymbol{\alpha})$, we employ the decomposition

$$\pi(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{x}, \mathbf{W}) = \pi(\boldsymbol{\Omega}^{-1*}|\mathbf{y}, \mathbf{x}, \mathbf{W})\pi(\boldsymbol{\alpha}^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, \boldsymbol{\Omega}^*)\pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, \boldsymbol{\Omega}^*, \boldsymbol{\alpha}^*)$$

where the first expression can be obtained via Rao–Blackwell methods as

$$\hat{\pi}(\boldsymbol{\Omega}^{-1*}|\mathbf{y}, \mathbf{x}, \mathbf{W}) = M^{-1} \sum_{g=1}^{M} \left(\prod_{j=0,1} \prod_{k \in K_j} \pi\left(\boldsymbol{\Omega}_{jk}^{-1*}|\mathbf{y}_{jk}, \mathbf{x}_{jk}, \mathbf{s}_{jk}^{(g)}, \mathbf{W}_{jk}, \boldsymbol{\alpha}^{(g)}, \boldsymbol{\beta}_{jk}^{(g)}, \boldsymbol{\lambda}_{jk}^{(g)}\right) \right)$$

For the second, ordinate we use the result from Chib and Jeliazkov (2001) that

$$\pi(\boldsymbol{\alpha}^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\Omega}^*) = \frac{\mathbb{E}_1[\boldsymbol{\alpha}(\boldsymbol{\alpha}^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\Omega}^*, \mathbf{z})q(\boldsymbol{\alpha}^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\Omega}^*, \boldsymbol{\beta})]}{\mathbb{E}_2[\alpha(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\Omega}^*, \mathbf{z})]} \tag{11}$$

where the expectation $\mathbb{E}_1$ in the numerator is with respect to $\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}|\mathbf{y}, \mathbf{x}, \Omega^*)$ and the expectation $\mathbb{E}_2$ in the denominator is with respect to $\pi(\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}, \boldsymbol{\alpha}^*, \Omega^*)$ $q(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \Omega^*)$. Each expectation can be estimated from the output of suitable reduced runs (Chib, 1995). To estimate the numerator, we fix Ω at $\boldsymbol{\Omega}^*$ and continue the MCMC iterations with the quantities $\boldsymbol{\theta}_{-\Omega}$ and $\mathbf{z} = (\boldsymbol{\lambda}, \mathbf{s}_{00})$, and then average $\alpha(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \Omega^*, \mathbf{z})q(\boldsymbol{\alpha}^*|\mathbf{y}, \mathbf{x}, \Omega^*, \boldsymbol{\beta})$ over the

resulting draws. To estimate the denominator, we fix $(\Omega, \boldsymbol{\alpha})$ at $(\Omega^*, \boldsymbol{\alpha}^*)$ and continue the MCMC iterations; in each cycle of this run, we also draw $\boldsymbol{\alpha}$ from $q(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{x}, \Omega^*, \boldsymbol{\beta})$. We then average $\alpha(\boldsymbol{\alpha}^*, \boldsymbol{\alpha}|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \Omega^*, \mathbf{z})$ over the draws on $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ from this run. Simultaneously, from the output of the latter run we estimate $\pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{x}, \Omega^*, \boldsymbol{\alpha}^*)$ as

$$\hat{\pi}(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{x}, \mathbf{W}) = M^{-1} \sum_{g=1}^{M} \left(\prod_{j=0,1} \prod_{k \in K_j} \pi\left(\boldsymbol{\beta}_{jk}^* | \mathbf{y}_{jk}, \mathbf{x}_{jk}, \mathbf{s}_{jk}^{(g)}, \mathbf{W}_{jk}, \boldsymbol{\alpha}^*, \boldsymbol{\Omega}_{jk}^*, \boldsymbol{\lambda}_{jk}^{(g)}\right) \right)$$

To estimate the posterior ordinate for Model 2, where $\boldsymbol{\theta} = (\{\boldsymbol{\beta}_{jk}\}, \{\sigma_{jk}\}, \boldsymbol{\alpha}, \{\mathbf{D}_{jk}\})$, we proceed in a similar way using the decomposition

$$\begin{aligned} \pi(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{x}, \mathbf{W}) = {} & \pi(\mathbf{D}_{0c}^{-1*}, \mathbf{D}_{1c}^{-1*}, \mathbf{D}_n^{-1*}|\mathbf{y}, \mathbf{x}, \mathbf{W}) \pi(\boldsymbol{\alpha}^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{D}_{0c}^*, \mathbf{D}_{1c}^*, \mathbf{D}_n^*) \\ & \times \pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{D}_{0c}^*, \mathbf{D}_{1c}^*, \mathbf{D}_n^*, \boldsymbol{\alpha}^*) \\ & \times \pi(\boldsymbol{\sigma}^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{D}_{0c}^*, \mathbf{D}_{1c}^*, \mathbf{D}_n^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \end{aligned}$$

where $\hat{\pi}(\mathbf{D}_{0c}^{-1*}, \mathbf{D}_{1c}^{-1*}, \mathbf{D}_n^{-1*}|\mathbf{y}, \mathbf{x}, \mathbf{W})$ is estimated via Rao–Blackwell methods from

$$\begin{aligned} M^{-1} \sum_{g=1}^{M} \Bigg(& \prod_{i \in N_{0c}} \pi\left(\mathbf{D}_{0c}^{-1*}|\mathbf{y}_{jk}, \mathbf{x}_{jk}, \mathbf{s}_{jk}^{(g)}, \mathbf{W}_{jk}, \boldsymbol{\alpha}^{(g)}, \boldsymbol{\beta}_{jk}^{(g)}, \boldsymbol{\lambda}_{jk}^{(g)}, \mathbf{b}_{jk}^{(g)}\right) \\ & \prod_{i \in N1} \pi\left(\mathbf{D}_{1c}^{-1*}|\mathbf{y}_{jk}, \mathbf{x}_{jk}, \mathbf{s}_{jk}^{(g)}, \mathbf{W}_{jk}, \boldsymbol{\alpha}^{(g)}, \boldsymbol{\beta}_{jk}^{(g)}, \boldsymbol{\lambda}_{jk}^{(g)}, \mathbf{b}_{jk}^{(g)}\right) \\ & \prod_{i \in N_n} \pi\left(\mathbf{D}_n^{-1*}|\mathbf{y}_{jk}, \mathbf{x}_{jk}, \mathbf{s}_{jk}^{(g)}, \mathbf{W}_{jk}, \boldsymbol{\alpha}^{(g)}, \boldsymbol{\beta}_{jk}^{(g)}, \boldsymbol{\lambda}_{jk}^{(g)}, \mathbf{b}_{jk}^{(g)}\right) \Bigg) \end{aligned}$$

The reduced ordinates for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are updated in the same manner as in Model 1 in two reduced runs. Here the first reduced run is done conditional on $(\mathbf{D}_{0c}^*, \mathbf{D}_{1c}^*, \mathbf{D}_n^*)$ and the second reduced run, which also yields the posterior estimate of $\boldsymbol{\beta}^*$, is done conditional on $(\mathbf{D}_{0c}^*, \mathbf{D}_{1c}^*, \mathbf{D}_n^*, \boldsymbol{\beta}^*)$. A final third reduced run with $(\mathbf{D}_{0c}, \mathbf{D}_{1c}, \mathbf{D}_n, \boldsymbol{\alpha}, \boldsymbol{\beta})$ fixed at $(\mathbf{D}_{0c}^*, \mathbf{D}_{1c}^*, \mathbf{D}_n^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ is required to estimate the posterior ordinate $\hat{\pi}(\boldsymbol{\sigma}^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{D}_{0c}^*, \mathbf{D}_{1c}^*, \mathbf{D}_n^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ from

$$M^{-1} \sum_{g=1}^{M} \left(\prod_{j=0,1} \prod_{k \in K_j} \pi\left(\boldsymbol{\sigma}_{jk}^* | \mathbf{y}_{jk}, \mathbf{x}_{jk}, \mathbf{s}_{jk}^{(g)}, \mathbf{W}_{jk}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}_{jk}^*, \boldsymbol{\lambda}_{jk}^{(g)}\right) \right)$$

4.4. Results

In this section we present the key results from fitting Models 1 and 2 to the data on high-risk respondents. For each model we consider three different values of the degrees of freedom parameter ($v = 5$, 10, 20) and the model with the highest marginal likelihood in each model class is then studied more intensively. All our results are based on 20,000 MCMC iterations following a burn-in of 1,000 iterations. Table 5 contains the estimated log marginal likelihoods for our six contending models. As can be seen, the models with $v = 5$ provide the best fit to the data. We now discuss the fitting results in more detail.

Table 6 summarizes the prior-posterior analysis for the covariance matrices in Model 1. One point to note is that our MCMC algorithm is well behaved as indicated by the low inefficiency factors that are reported for all the parameters. The inefficiency factors are computed as $1 + 2\sum_{l=1}^{L}\rho_k(l)$, where $\rho_k(l)$ is the autocorrelation of the kth parameter at lag l and L is chosen as the value at which the autocorrelation function tapers off. The inefficiency factors approximate the ratio of the numerical variance of the posterior mean from the MCMC chain relative to that from hypothetical iid draws. As is evident from Tables 6–9, the inefficiency factors for the covariance and slope parameters of Model 1 (and 2) are small and in some case quite close to the ideal value of 1.

An interesting point is that even though the prior on the covariances matrices Ω_{jk} is the same, the posterior mean of these matrices is quite different. In this connection it may be observed that the largest variances and covariances occur for compliers in the no-training state. To illustrate the differences we show image plots of the covariance matrices (see Fig. 1). To plot the posterior means we have used a gray scale that is set at black for 0 and white for 0.55. The results suggest that our extension of Yau and Little's

Table 5. Estimates of the Log Marginal Likelihoods for Models 1 and 2 for Different Degrees of Freedom.

Model	Degrees of Freedom		
	$v = 5$	$v = 10$	$v = 20$
Log marginal likelihoods			
M1	*−1339.07*	−1345.11	−1362.13
M2	*−1332.96*	−1341.32	−1351.26

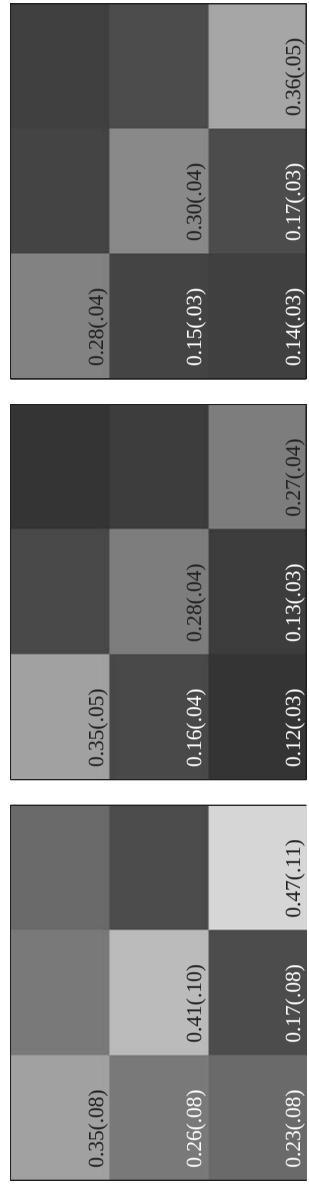
Table 6. Model 1 – Prior-Posterior Analysis for the Covariance Matrices: Prior Means, Posterior Means and Standard Deviations (in Parentheses).

Ω_{0c}			Ω_{0n}			Ω_{1c}		
Prior	Post.	Ineff.	Prior	Post.	Ineff.	Prior	Post.	Ineff.
0.50	0.35 (0.08)	2.21	0.50	0.35 (0.05)	1.86	0.50	0.28 (0.04)	1.32
0.25	0.26 (0.08)	3.34	0.25	0.16 (0.04)	2.12	0.25	0.15 (0.03)	1.30
0.50	0.40 (0.10)	3.02	0.50	0.28 (0.04)	2.88	0.50	0.30 (0.04)	1.83
0.25	0.23 (0.08)	2.27	0.25	0.12 (0.03)	1.95	0.25	0.14 (0.03)	1.24
0.25	0.17 (0.08)	2.34	0.25	0.13 (0.03)	2.23	0.25	0.17 (0.03)	1.27
0.50	0.46 (0.11)	2.52	0.50	0.27 (0.04)	2.05	0.50	0.36 (0.05)	1.86

basic model to allow for type and treatment specific random shocks in the distributions of the health outcomes is useful in the context of these data.

Table 7 summarizes the prior-posterior analysis for the elements in the diagonal covariance matrices σ_{jk} and the random effects variances $\mathbf{D}_{jk}$ from the fitting of Model 2. As in Model 1, we observe that the estimates vary by intake and type. For a better comparison of the results with Model 1 we consider the covariance matrix of the Student's t outcome distribution marginalized over the random effects (see Eq. (9)). We obtain estimates of the posterior means (and standard deviations) by computing $\sum_{jk,i} = \{\lambda_i^{-1} diag(\sigma_{jk}) + \mathbf{V}_i \mathbf{D}_{jk} \mathbf{V}_i'\}$ at each iteration of the MCMC algorithm for Model 2. The plots of the posterior means in Fig. 2 use a gray scale that is set to black at 0 and white at 0.9. As in the case of Model 1, the largest variances/covariances are observed for compliers under no training. In general, the random effects specification yields higher variances than Model 1.

We now turn to the inferences about the remaining parameters in Models 1 and 2. We focus on the coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_{0c}$, and $\boldsymbol{\beta}_{1c}$, which play a key role in the determination of the causal training effects on the mental health outcomes discussed in Section 5. Table 8 summarizes the prior-posterior analysis for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_{0c}$, and $\boldsymbol{\beta}_{1c}$ from the fitting of Model 1. For each parameter we report the prior and posterior means and standard deviations. We also provide the inefficiency factors as a measure of the autocorrelation of the



(a) Ω_{0c} (b) Ω_{0n} (c) Ω_{1c}

Fig. 1. Image Plots of the Posterior Means of the Covariance Matrices in Model 1. The Gray Scale is Set at Black for 0 and White for 0.5. The Posterior Standard Deviations are Given in Parentheses.

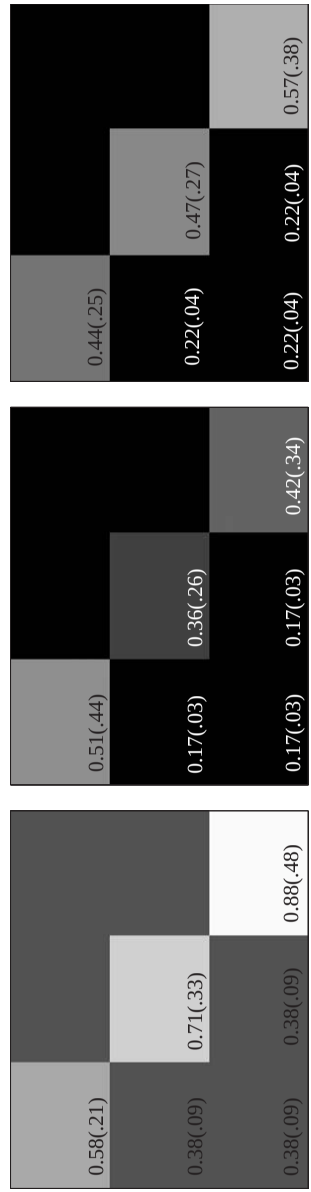
Table 7. Model 2 – Prior-Posterior Analysis for the Variance Parameters: Prior Means, Posterior Means and Standard Deviations (in Parentheses).

	Prior	$j = 0, k = c$		$j = 0, k = n$		$j = 1, k = c$	
		Post.	Ineff.	Post.	Ineff.	Post.	Ineff.
σ_{jk}	0.50	0.12 (0.04)	4.14	0.20 (0.03)	2.39	0.14 (0.02)	2.59
	0.50	0.22 (0.07)	2.95	0.12 (0.02)	4.04	0.15 (0.03)	2.54
	0.50	0.32 (0.09)	2.52	0.15 (0.03)	2.81	0.21 (0.04)	2.44
$\mathbf{D}_{jk}$	0.50	0.38 (0.09)	2.07	0.17 (0.03)	4.18	0.22 (0.04)	2.52

draws. The second column of the table gives the posterior inference on $\boldsymbol{\alpha}$. The reported posterior means imply a higher compliance probability for subjects that are older, more motivated to attend the program and better educated. Subjects that have a higher level of assertiveness, are married, and non-white and those who experience economic hardship are less likely to be a complier. Column 3 shows that all coefficients are measured with low inefficiency factors.

Columns 4 through 9 in the same table provide results for $\boldsymbol{\beta}_{jc}$. These parameters capture the interaction of the training intake with the coefficients on the constant and the baseline depression and risk scores on the change in the depression scores in the subsequent three time periods. A comparison of the posterior means between $\boldsymbol{\beta}_{0c}$ and $\boldsymbol{\beta}_{1c}$ in columns 5 and 8 reveals that the actual training intake affects the health outcomes after controlling for unobserved confounders through subject type. The differences between the posterior means of compliers in both training intake states are especially pronounced in the last two time periods, shown in the last six rows of the table. The higher posterior standard deviations and inefficiency factors of $\boldsymbol{\beta}_{0c}$, as compared to $\boldsymbol{\beta}_{1c}$, reflect our earlier point that the parameters for compliers under no treatment are the most difficult to estimate as they are identified from a mixture distribution.

We report a similar set of results for Model 2 in Table 9. The posterior means and standard deviations for $\boldsymbol{\alpha}$ reported in column 2 are almost identical with those reported for Model 1 in Table 8. One would expect this result as both models use the same probit specification for the compliance probability. This is not the case for the coefficients in the outcome



(a) Σ_{0c} (b) Σ_{0n} (c) Σ_{1c}

Fig. 2. Image Plots of the Posterior Means of the Covariance Matrices from Model 2. The Posterior Mean and Standard Errors (in Parentheses) are Given for Each Element.

Table 8. Model 1 – Prior-Posterior Analysis for the Coefficient Vectors from the Compliance Probability and Outcome Models for Compliers: Prior Means, Posterior Means, Standard Deviations (in Parentheses) and Inefficiency Factors.

α			β_{0c}			β_{1c}		
Prior	Post.	Ineff.	Prior	Post.	Ineff.	Prior	Post.	Ineff.
−0.5	−3.39	2.13	0.8	0.91	3.51	0.8	0.77	1.00
(5.0)	(0.87)		(5.0)	(0.91)		(5.0)	(0.38)	
0.03	0.04	2.15	−1.1	−1.00	4.72	−1.1	−1.00	1.00
(5.0)	(0.01)		(5.0)	(0.62)		(5.0)	(0.26)	
0.5	0.42	2.05	0.8	0.71	5.18	0.8	0.66	1.00
(5.0)	(0.10)		(5.0)	(0.96)		(5.0)	(0.38)	
0.1	0.13	2.03	1.5	0.77	3.89	0.6	0.58	1.00
(5.0)	(0.04)		(5.0)	(0.93)		(5.0)	(0.41)	
0.0	−0.26	1.99	−1.5	−1.42	3.49	−1.5	−1.14	1.00
(5.0)	(0.10)		(5.0)	(0.58)		(5.0)	(0.27)	
0.0	−0.22	2.89	1.0	1.39	2.65	1.0	0.92	1.00
(5.0)	(0.18)		(5.0)	(0.82)		(5.0)	(0.39)	
0.0	−0.03	2.25	1.4	1.87	1.77	1.6	1.00	1.00
(5.0)	(0.11)		(5.0)	(0.89)		(5.0)	(0.44)	
0.0	−0.25	1.96	−1.2	−1.25	4.47	−1.2	−0.95	1.00
(5.0)	(0.21)		(5.0)	(0.67)		(5.0)	(0.30)	
			0.6	0.52	3.42	0.6	0.31	1.00
			(5.0)	(0.93)		(5.0)	(0.43)	

distributions. The posterior means of $\boldsymbol{\beta}_{0c}$ and $\boldsymbol{\beta}_{1c}$ reported in columns 4 and 6 in the table differ from those reported for Model 1 in Table 8.

As in the case of Model 1 we observe that the posterior means of the elements in $\boldsymbol{\beta}_{0c}$ and $\boldsymbol{\beta}_{1c}$ differ in all time periods. For example, in all periods the coefficient on the intercept is lower for compliers under training. The posterior means of the coefficient on the baseline depression score is negative for all compliers, but more negative for compliers in the no-training state. On the other hand, the posterior means of the coefficients on the baseline risk score are positive and smaller for compliers in the training state. All estimates vary by time. Also note that the coefficients in Model 2 come with lower inefficiency factors than those from Model 1. While we can conclude from these results that training intake affects the mental health outcomes, it is less easy to calculate the size and direction of the training effects from these estimates. In the next section we discuss a predictive approach that allows us to calculate the training effects.

Table 9. Model 2 – Prior-Posterior Analysis for the Coefficient Vectors from the Compliance Probability and Outcome Models for Compliers: Prior Means, Posterior Means, Standard Deviations (in Parentheses) and Inefficiency Factors.

α			β_{0c}			β_{1c}		
Prior	Post.	Ineff.	Prior	Post.	Ineff.	Prior	Post.	Ineff.
−0.5	−3.41	2.07	0.0	1.10	1.56	0.0	0.83	1.00
(5.0)	(0.86)		(5.0)	(0.88)		(5.0)	(0.42)	
0.03	0.04	2.13	−1.0	−1.36	1.73	−1.0	−1.11	1.00
(5.0)	(0.01)		(5.0)	(0.56)		(5.0)	(0.30)	
0.5	0.43	1.40	0.6	1.20	1.85	0.6	0.84	1.00
(5.0)	(0.10)		(5.0)	(0.88)		(5.0)	(0.41)	
0.1	0.13	1.97	1.0	1.26	1.20	0.7	0.67	1.00
(5.0)	(0.04)		(5.0)	(0.95)		(5.0)	(0.43)	
0.0	−0.23	1.97	−1.5	−2.12	1.66	−1.5	−1.28	1.00
(5.0)	(0.10)		(5.0)	(0.61)		(5.0)	(0.30)	
0.0	−0.22	2.12	0.8	2.25	1.57	1.0	1.12	1.00
(5.0)	(0.18)		(5.0)	(0.92)		(5.0)	(0.42)	
0.0	−0.04	2.20	1.4	2.29	1.00	0.5	1.10	1.00
(5.0)	(0.10)		(5.0)	(0.99)		(5.0)	(0.46)	
0.0	−0.21	1.94	−1.2	−1.44	1.20	−1.2	−1.08	1.00
(5.0)	(0.21)		(5.0)	(0.63)		(5.0)	(0.32)	
			0.6	0.62	1.17	0.6	0.56	1.00
			(5.0)	(0.94)		(5.0)	(0.45)	

5. ANALYSIS OF TREATMENT EFFECTS

The treatment effects analysis investigates whether the actual intake of the training program has a positive causal effect on the vector of mental health outcomes for compliers. A natural way to answer this question within our Bayesian modeling framework is to take a predictive perspective. Chib (2007) has shown that the Bayesian predictive approach is useful in drawing inferences about causal treatment effects. In this section we extend the methods discussed in Chib and Jacobi (2008) to the panel case. We also show how these predictive distributions can be used to compute various treatment effects, such as quantile treatment effects, and a predictive version of the complier-average causal effect that was computed in Yau and Little (2001).

We begin our predictive analysis by considering a subject that is randomly drawn from the subpopulation of compliers. We let $\mathbf{y}_{jc,n+1}$,

$j = 0, 1$, denote the subject's potential vector of depression scores under no training and under training, and $p(\mathbf{y}_{jc,n+1}|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z})$ denote the corresponding marginal predictive distribution of interest. These marginal distributions are defined as

$$\int p(\mathbf{y}_{jc,n+1}|\mathbf{W}_{n+1}, \boldsymbol{\theta}) I(s_{n+1} = 1) p(s_{n+1}|\mathbf{w}_{i0}, \boldsymbol{\alpha}) \pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z}) p(\mathbf{W}_{n+1}|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z}) \mathrm{d}s_{n+1} \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{W}_{n+1}$$

where

$$p(s_{n+1}|\mathbf{v}_{n+1}, \boldsymbol{\alpha}) = \Phi(\mathbf{w}'_{i0}\boldsymbol{\alpha})^{s_{n+1}} \{1 - \Phi(\mathbf{w}'_{i0}\boldsymbol{\alpha})\}^{1-s_{n+1}}$$

but are not in closed form. In this expression we take $p(\mathbf{W}_{n+1}|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z})$ as the empirical distribution of the covariates in our sample. From expressions (4) and (9) we know that $p(\mathbf{y}_{jc,n+1}|\mathbf{W}_{n+1}, \boldsymbol{\theta}) = t_v(\mathbf{y}_i|\mathbf{W}_{n+1}\boldsymbol{\beta}_{jc}, \boldsymbol{\Omega}_{jc})$ in Model 1 and $p(\mathbf{y}_{jc,n+1}|\mathbf{W}_{n+1}, \boldsymbol{\theta}) = t_v(\mathbf{y}_i|\mathbf{W}_{n+1}\boldsymbol{\beta}_{jc} + \mathbf{V}_i\mathbf{b}_{ijc}, \mathrm{diag}(\sigma_{jc}))$ in Model 2. The fact that these conditional distributions are easily sampled means that the predictive distributions can be calculated by the method of composition. At each iteration $g = 1, 2, \ldots, M$ of the MCMC chain, we randomly sample $\mathbf{W}^{(g)}_{n+1}$ and $w^{(g)}_{n+1,0}$ from the full set of covariates. Next, we sample $s^{(g)}_{n+1} = I[\mathbf{w}'^{(g)}_{n+1,0}\boldsymbol{\alpha}^{(g)} + u^{(g)}_{n+1} > 0]$, where $u^{(g)}_{n+1} \sim \mathcal{N}(0, 1)$. We then check compliance. If $s^{(g)}_{n+1} = 1$, we draw the potential outcomes $\mathbf{y}^{(g)}_{jc,n+1}$ under each intake state from the Student's t outcome density, conditional on the current sampled draw of the parameters. Otherwise we skip and move to the next step in the chain. The resulting draws $[\mathbf{y}^{(1)}_{jc,n+1}, \ldots, \mathbf{y}^{(g)}_{jc,n+1}, \ldots, \mathbf{y}^{(J)}_{jc,n+1}], J \leq M$ are from the marginal predictive distributions of the potential outcomes.

We summarize these generated draws in various ways. One is in terms of (kernel smoothed) predictive density plots. Another is in terms of the differences in means and quantiles of the sampled draws. For example, the predictive average treatment is calculated as $E(\mathbf{y}_{1c,n+1}|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z}) - E(\mathbf{y}_{0c,n+1}|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z})$, where the means are computed directly as sample averages

$$E(\mathbf{y}_{jc,n+1}|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z}) = \frac{1}{J}\sum_{g=1}^{J} \mathbf{y}^{(g)}_{jc,n+1}$$

5.1. Results

Before providing the predictive treatment effects, we pause to examine which subjects in the control arm are a-posteriori classified as compliers and whether the subjects so classified are similar to the compliers in the treatment arm with respect to their observable characteristics. In Fig. 3 we

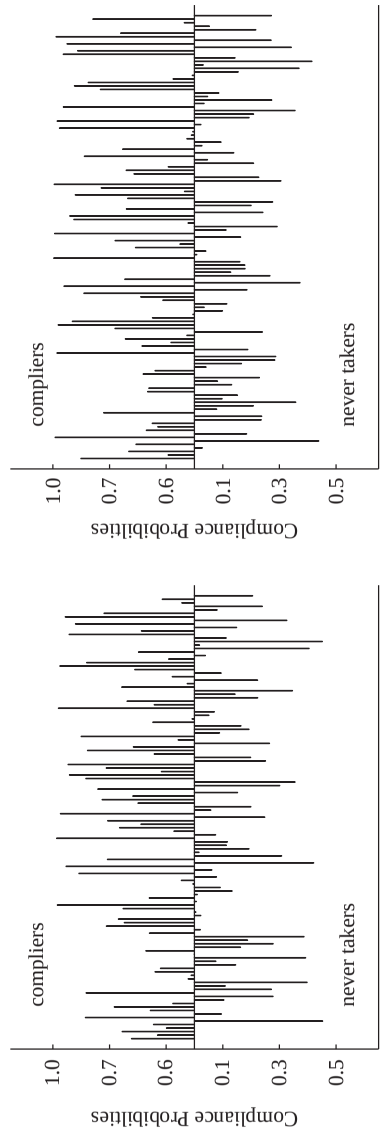


Fig. 3. Posterior Mean of Probability of Compliance for Control Arm Subjects from Models 1 and 2.

present the posterior mean of the compliance probability from each best-fitting model for each of the 127 control arm subjects. Probabilities between 0 and 0.5 (which can be taken to indicate a never-taker) are plotted below the horizontal axis and those above 0.5 (which indicate a complier) are plotted above the horizontal axis. We see that when the compliance probability is less than 0.5, it tends to be generally less than 0.4, suggesting strongly that each of those subjects is a never-taker. Similarly, when the compliance probability is greater than 0.5, it tends to be generally greater than 0.6, suggesting again that for those subjects inference about the type is more or less decisive. There are, however, some subjects whose compliance probabilities are close to 0.5 and therefore for these subjects a precise determination of type is not possible. An interesting point is that these compliance probabilities are almost the same across the two models. That the covariates are balanced for compliers can be seen from Table 10, which reports the sample means of the covariates by intake and type. The first 7 rows refer to the covariates in the model for the compliance probability. Comparing the sample means for compliers in the control arm with those in the treatment arm (columns 2 and 6 for Model 1, columns 3 and 7 for

Table 10. Sample Means of the Baseline Covariates from the Probit Model for the Compliance Probabilities by Type.

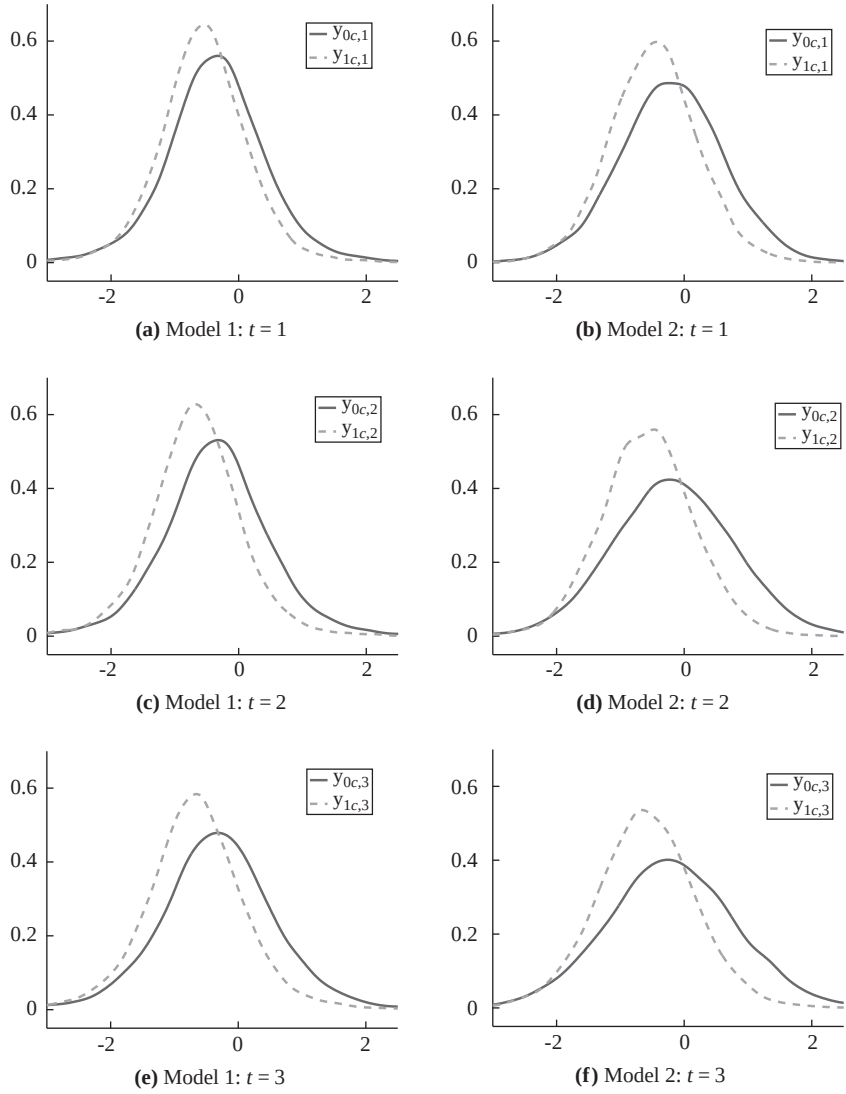
Variable	Sample Means by Intake and Type					
	$j = 0$				$j = 1$	
	Compliers		Never-Takers		Compliers	
	M1	M2	M1	M2	M1	M2
Age (demeaned)	20.70	21.20	13.57	13.50	19.26	19.26
Motivate	5.62	5.61	5.01	5.02	5.46	5.46
Edu	13.69	13.85	13.03	12.98	13.72	13.72
Assert	2.87	2.91	3.13	3.11	2.87	2.87
Marr	0.47	0.48	0.39	0.39	0.37	0.37
Econ	3.41	3.39	3.62	3.62	3.53	3.53
Nonw	0.15	0.12	0.21	0.22	0.14	0.14
Depress	2.49	2.46	2.44	2.46	2.41	2.41
Risk	1.70	1.67	1.67	1.68	1.67	1.67

Note: Compliers and never-takers in the control arm are classified based on the estimated posterior mean of compliance.

Model 2) we see that the two groups look almost identical. The only exception is the marriage indicator covariate which in any case was estimated with a very low precision and mean of 0 (see Tables 8 and 9). A look at columns 4 and 5 shows that never-takers seem different from compliers. Never-takers are younger, less motivated, and less educated. Finally, compliers and never-takers have almost identical sample means of the covariates (baseline depression and risk scores) that are in the outcome model (but not present in the model of compliance).

The three graphs in Fig. 4 show the kernel plots of the marginal predictive densities for compliers, for each of the periods $t = 1,2,3$. The solid lines refer to the potential outcomes under no training and the dashed lines under training participation ($y_{1c,t}$). All plots show an improvement in mental health from participation in the training participation compared to no training. The marginal densities under training participation have more mass for negative values. In comparison, the densities in the no-training case have more mass over positive values. The greatest difference between the two predictive densities occur in the second and third time periods.

To get a better view of the magnitude of the mental health improvements caused by the participation in the training program, we compare the means and quantiles of the predictive densities in each time period. In Table 11 we report the average and quantile treatment effects for the 0.05, 0.25, 0.50, 0.75, and 0.95 quantiles for Models 1 and 2. The entries in the first row of results show that on average the program leads to a decrease in the depression scores. In the case of Model 1 the training program decreases average depression scores between 0.20 points in the first period, 0.32 points in the second period, and 0.36 points in the third period. The analysis for Model 1 suggests higher average treatment effects that range between 0.28 in the first period to -0.41 in the second and third periods. As indicated by the kernel plots in Fig. 4 all quantile treatment effects are negative. For Model 1 the effects vary between -0.02 and -0.16 points at the 5% quantiles and between -0.37 and -0.60 points at the 95% quantiles. For Model 2 the estimated 5% quantile treatment effects lie between -0.04 and -0.07 points. The 95% quantiles range between -0.52 and -0.80 points. Our results for the average complier effect differ from those found by Yau and Little (2001) in that we do not find a decrease in the treatment effect after period 2. Our estimated average complier effect at $t = 2$ under Model 1 is similar in magnitude to that found in the study by Skrondahl and Rabe-Hesketh (2004) that focused on health outcomes in period 2.



(a) Model 1: $t = 1$ **(b)** Model 2: $t = 1$

(c) Model 1: $t = 2$ **(d)** Model 2: $t = 2$

(e) Model 1: $t = 3$ **(f)** Model 2: $t = 3$

Fig. 4. Predictive Marginal Distributions of the Potential Outcomes for Compliers in Models 1 and 2.

Table 11. Predicted Average and Quantile Treatment Effects from Models 1 and 2 for all Time Periods.

Treatment Effect	Treatment Effects Estimates					
	Model 1			Model 2		
	$t=1$	$t=2$	$t=3$	$t=1$	$t=2$	$t=3$
Average	−0.20	−0.32	−0.36	−0.28	−0.41	−0.41
Quantile						
5%	−0.02	−0.13	−0.16	−0.07	−0.04	−0.04
25%	−0.14	−0.25	−0.28	−0.20	−0.24	−0.24
50%	−0.21	−0.32	−0.36	−0.28	−0.40	−0.40
75%	−0.25	−0.40	−0.45	−0.39	−0.57	−0.57
95%	−0.37	−0.55	−0.60	−0.52	−0.78	−0.80

6. CONCLUSION

We have discussed Bayesian models for finding the longitudinal causal effects of a randomized two-arm training program when compliance with the randomized assignment is less than perfect in the training arm for reasons that are potentially correlated with the outcomes. We show how the type approach can be used to calculate interesting causal effects. An important point is that the Bayesian approach is particularly useful in this context because it provides an automatic way of dealing with the mixture outcome distribution in the control arm. The possibility of incorporating real prior information is also another advantage of the Bayesian approach. We discuss how different versions of our models can be compared by marginal likelihoods and Bayes factors and how useful summaries of the causal effects can be determined from a predictive perspective. All of our computations proceed without the joint distribution of the potential outcomes. In addition, the fitting algorithms are efficient and provide detailed information about the compliance status of subjects in the control arm. Because of these strengths of the techniques discussed here, we believe that the methods of this paper will prove useful in practical work.

REFERENCES

Albert, J., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–779.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*, 1313–1321.

Chib, S. (2001). Markov Chain Monte Carlo methods: Computation and inference. In: J. J. Heckman & E. Leamer (Eds), *Handbook of econometrics* (Vol. 5, pp. 3569–3649). Amsterdam: North Holland.

Chib, S. (2007). Analysis of treatment response data without the joint distribution of potential outcomes. *Journal of Econometrics*, *140*, 401–412.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis–Hastings algorithm. *American Statistician*, *49*, 327–335.

Chib, S., & Jacobi, L. (2007). Modeling and calculating the effect of treatment at baseline from panel outcomes. *Journal of Econometrics*, *140*, 781–801.

Chib, S., & Jacobi, L. (2008). Analysis of treatment response data from eligibility designs. *Journal of Econometrics*, *144*, 465–478.

Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, *96*, 270–281.

Clark, A. E., & Oswald, E. J. (1994). Unhappiness and unemployment. *The Economic Journal*, *104*, 648–659.

Frangakis, C. E., Brookmeyer, R. S., Varadhan, R., Safaeian, M., Vlahov, D., & Strathdee, S. A. (2004). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *Journal of the American Statistical Association*, *99*, 239–249.

Frangakis, C. F., & Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, *86*, 365–379.

Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, X. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, *1*, 69–88.

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized trials with noncompliance. *The Annals of Statistics*, *25*, 305–327.

Jo, B. (2002). Statistical power in randomized trials with noncompliance. *Psychological Methods*, *7*, 178–193.

Levy, D. E., O'Malley, J. A., & Normand, S. T. (2004). Covariate adjustment in clinical trials with non-ignorable missing data and non-compliance. *Statistics in Medicine*, *23*, 2319–2339.

Mealli, F., Imbens, G. W., Ferro, S., & Biggeri, A. (2004). Analyzing a randomized trial on breast self-examination with non-compliance and missing outcomes. *Biostatistics*, *5*, 207–222.

Sommer, A., & Zeger, S. (1991). On estimating efficacy in clinical trials. *Statistics in Medicine*, *10*, 45–52.

Skrondahl, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. New York: Chapman & Hall/CRC.

Ten Have, T. R., Joffe, M., & Cary, M. (2003). Causal logistic models for non-compliance under randomized treatment with univariate binary response. *Statistics in Medicine*, *22*, 1255–1283.

Vinokur, A. D., Price, R. H., & Schul, Y. (1995). Impact of JOBS intervention on unemployed workers varying in risk for depression. *American Journal of Community Psychology*, *19*, 543–562.

Yau, L., & Little, R. (2001). Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association*, *96*, 1232–1244.

APPENDIX. MCMC ALGORITHMS IN DETAIL

Model 1

1. Sample $(\mathbf{s}_{00}, \boldsymbol{\alpha}, \{\boldsymbol{\lambda}_i\}|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \Omega)$ by sampling
   (a) $s_i = 1|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \Omega$ for $i \in I_{00}$ with probability $\Pr(s_i = 1|\mathbf{y}_i, x_i, \boldsymbol{\beta}_{0c}, \boldsymbol{\beta}_{0n}, \boldsymbol{\alpha}, \boldsymbol{\Omega}_{0c}, \boldsymbol{\Omega}_{0n})$ given by

$$\frac{q_{ci} t_{T,v}(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0c}, \boldsymbol{\Omega}_{0c})}{q_{ci} t_{T,v}(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0c}, \boldsymbol{\Omega}_{0c}) + (1 - q_{ci}) t_{T,v}(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n}, \boldsymbol{\Omega}_{0n})}$$

   (b) $\boldsymbol{\alpha}|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \Omega, \mathbf{s}_{00}$ for $i \in N$ by a MH step by proposing $\boldsymbol{\alpha}^\dagger$ from $t_{20}(\boldsymbol{\alpha}|\mu, V)$ and accepting the proposal value $\boldsymbol{\alpha}^\dagger$ with probability

$$\min\left\{1, \frac{\pi(\boldsymbol{\alpha}^\dagger) \prod_{i \in N_{0c} \cup N_{1c}} \Phi(s_i|\mathbf{w}'_{i0}\boldsymbol{\alpha}^\dagger) \prod_{i \in N_{0n}} \{1 - \Phi(s_i|\mathbf{w}'_{i0}\boldsymbol{\alpha})\} t_{20}(\boldsymbol{\alpha}|\mu, V)}{\pi(\boldsymbol{\alpha}) \prod_{i \in N_{0c} \cup N_{1c}} \Phi(s_i|\mathbf{w}'_{i0}\boldsymbol{\alpha}) \prod_{i \in N_{0n}} 1 - \Phi(s_i|\mathbf{w}'_{i0}\boldsymbol{\alpha}) t_{20}(\boldsymbol{\alpha}^\dagger|\mu, V)}\right\}$$

   where μ is the approximate mode of

$$\ln\left[\prod_{i \in N_{0c} \cup N_{1c}} \Phi(s_i|\mathbf{w}'_{i0}\boldsymbol{\alpha}) \prod_{i \in N_{0n}} \{1 - \Phi(s_i|\mathbf{w}'_{i0}\boldsymbol{\alpha})\}\right]$$

   and V is the inverse Hessian of the latter expression evaluated at μ.

   (c) $\boldsymbol{\lambda}_i|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \Omega, \mathbf{s}_{00}$ for $i \in N$ from

$$\mathcal{G}\left(\boldsymbol{\lambda}_i \middle| \frac{v + T}{2}, \frac{v + (\mathbf{y}_i - \mathbf{W}_i\boldsymbol{\beta}_{jk})\boldsymbol{\Omega}_{jk}^{-1}(\mathbf{y}_i - \mathbf{W}_i\boldsymbol{\beta}_{jk}}{2}\right)$$

2. Sample $\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}, \mathbf{s}_{00}, \boldsymbol{\lambda}, \Omega, \boldsymbol{\alpha}$ by sampling $\boldsymbol{\beta}_{jk}|\mathbf{y}, \mathbf{x}, \mathbf{s}_{00}, \boldsymbol{\lambda}, \Omega_{jk}, \boldsymbol{\alpha}$ from

$$\mathcal{N}_p\left(\boldsymbol{\beta}_{jk}|\mathbf{B}_{jk}\left\{\mathbf{B}_{jk,0}^{-1}\boldsymbol{\beta}_{jk,0} + \sum_{i \in N_{jk}} \mathbf{W}'_i\boldsymbol{\lambda}_i\boldsymbol{\Omega}_{jk}^{-1}\mathbf{y}_i\right\},\right.$$

$$\left.\mathbf{B}_{jk} = \left\{\mathbf{B}_{jk,0}^{-1} + \sum_{i \in N_{jk}} \mathbf{W}'_i\boldsymbol{\lambda}_i\boldsymbol{\Omega}_{jk}^{-1}\mathbf{W}_i\right\}^{-1}\right)$$

3. Sample $\boldsymbol{\Omega}|\mathbf{y}, \mathbf{x}, \mathbf{s}_{00}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{D}$ by sampling $\boldsymbol{\Omega}_{jk}|\mathbf{y}, \mathbf{x}, \mathbf{s}_{00}, \boldsymbol{\beta}_{jk}, \boldsymbol{\lambda}, \boldsymbol{\alpha}$ from

$$\mathcal{W}_3\left(\boldsymbol{\Omega}_{jk}^{-1}|\rho_{jk,0}+n_{jk}, \left[\mathbf{R}_{jk,0}^{-1}+\sum_{i\in N_{jk}}\boldsymbol{\lambda}_i(\mathbf{y}_i-\mathbf{W}_i\boldsymbol{\beta}_{jk})(\mathbf{y}_i-\mathbf{W}_i\boldsymbol{\beta}_{jk})'\right]^{-1}\right)$$

Model 2

1. Sample $\mathbf{s}_{00}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{b}_{in}, \mathbf{b}_{ic}|\mathbf{y}, \mathbf{x}, \boldsymbol{\lambda}, \sigma, \mathbf{D}_{0c}, \mathbf{D}_{1c}, \mathbf{D}_n$ by sampling
   (a) s_i for $i\in I_{00}$ with probability $\Pr(s_i=1|\mathbf{y}_i, x_i, \boldsymbol{\beta}_{0c}, \boldsymbol{\beta}_{0n}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \sigma_{0c}, \sigma_{0n})$ given by

$$\frac{q_{ci}\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0c}, \sum_{0c,i})}{q_{ci}\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0c}, \sum_{0c,i})+(1-q_{ci})\mathcal{N}_T(\mathbf{y}_i|\mathbf{W}_i\boldsymbol{\beta}_{0n}, \sum_{0n,i})}$$

   where $\sum_{jci}\{\boldsymbol{\lambda}_i^{-1}diag(\sigma_{jc})+\mathbf{V}_i\mathbf{D}_{jc}\mathbf{V}_i'\}$.
   (b) $\boldsymbol{\alpha}|\mathbf{y}, \mathbf{x}, \mathbf{s}_{00}$ by a MH step by proposing $\boldsymbol{\alpha}^{\dagger}$ from $t_{20}(\boldsymbol{\alpha}|\mu, V)$ and accepting the proposal value with probability of move given in the algorithm for Model 1.
   (c) $\boldsymbol{\beta}_{jk}|\mathbf{y}, \mathbf{x}, \mathbf{s}_{00}, \boldsymbol{\lambda}, \sigma_{jk}, \boldsymbol{\lambda}, \mathbf{D}_{0c}, \mathbf{D}_{1c}, \mathbf{D}_n$ from

$$\mathcal{N}_p\left(\boldsymbol{\beta}_{jk}|\mathbf{B}_{jk}\left\{\mathbf{B}_{jk,0}^{-1}\boldsymbol{\beta}_{jk,0}+\sum_{i\in N_{jk}}\mathbf{W}_i'\Sigma_{jk}^{-1}\mathbf{y}_i\right\}, \mathbf{B}_{jk}=\left\{\mathbf{B}_{jk,0}^{-1}+\sum_{i\in N_{jk}}\mathbf{W}_i'\Sigma_{jk}^{-1}\mathbf{W}_i\right\}^{-1}\right)$$

   (d) $\mathbf{b}_{jci}|\mathbf{y}_i, x_i, \boldsymbol{\beta}_{jc}, \mathbf{s}_{00}, \boldsymbol{\lambda}_i, \sigma_{0c}, \sigma_{1c}, \mathbf{D}_{jc}$ for $i\in I_{jc}$, $j=0, 1$, from

$$\mathcal{N}_q(\mathbf{b}_{jci}|\mathbf{B}_{jci}\{\boldsymbol{\lambda}_i\mathbf{V}_i'\boldsymbol{\Omega}_{jc}^{-1}(\mathbf{y}_i-\mathbf{W}_i\boldsymbol{\beta}_{jc})\}, \mathbf{B}_{jci}=\{\mathbf{D}_{jc}^{-1}+\boldsymbol{\lambda}_i\mathbf{V}_i'(\mathrm{diag}(\boldsymbol{\sigma}_{jc}))^{-1}\mathbf{V}_i\}^{-1})$$

   (e) $\mathbf{b}_{in}|\mathbf{y}_i, x_i, \boldsymbol{\beta}_{0n}, \mathbf{s}_{00}, \boldsymbol{\lambda}_i, \sigma_{0n}, \mathbf{D}_n$ for $i\in I_{0n}$ from

$$\mathcal{N}_q(\mathbf{b}_{ni}|\mathbf{B}_{ni}\{\boldsymbol{\lambda}_i\mathbf{V}_i'\boldsymbol{\Omega}_{0n}^{-1}(\mathbf{y}_i-\mathbf{W}_i\boldsymbol{\beta}_{0n})\}, \mathbf{B}_{ni}=\{\mathbf{D}_n^{-1}+\boldsymbol{\lambda}_i\mathbf{V}_i'(\mathrm{diag}(\boldsymbol{\sigma}_{0n}))^{-1}\mathbf{V}_i\}^{-1})$$

2. $\boldsymbol{\lambda}_i|\mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \sigma_{0c}, \sigma_{0n}, \sigma_{1c}, \sigma_{0c}, \boldsymbol{\alpha}, \mathbf{s}_{00}, \mathbf{b}_{in}, \mathbf{b}_{ic}$ for $i\in N$ from

$$\mathcal{G}\left(\boldsymbol{\lambda}_i\middle|\frac{v+T}{2}, \frac{v+(\mathbf{y}_i-\mathbf{W}_i\boldsymbol{\beta}_{jk}-\mathbf{V}_i\mathbf{b}_{ik})(\mathrm{diag}(\boldsymbol{\sigma}_{jk}))^{-1}(\mathbf{y}_i-\mathbf{W}_i\boldsymbol{\beta}_{jk}-\mathbf{V}_i\mathbf{b}_{ik})}{2}\right)$$

3. Sample $\sigma|\mathbf{y}, \mathbf{x}, \mathbf{s}_{00}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \{\mathbf{b}_{in}\}, \{\mathbf{b}_{ic}\}, \mathbf{D}_{0c}, \mathbf{D}_{1c}, \mathbf{D}_n$ by sampling $\sigma_{jk,t}|\mathbf{y}_{jk,t}, \mathbf{x}, \mathbf{s}_{00}, \boldsymbol{\beta}_{jk}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, \{\mathbf{b}_{in}\}, \{\mathbf{b}_{ic}\}, \mathbf{D}_{0c}, \mathbf{D}_{1c}, \mathbf{D}_n$ from

$$\mathcal{IG}\left(\boldsymbol{\sigma}_{jk,t}\middle|\frac{\eta_{jk,t0}+n_{jk}}{2}, \frac{d_{jk,0}+\sum_{i\in N_{jk}}\boldsymbol{\lambda}_i(y_{i,t}-\mathbf{w}'_{it}\boldsymbol{\beta}_{jk}-\mathbf{v}_{it}\mathbf{b}_i)^2}{2}\right)$$

4. Sample $\mathbf{D}_{0c}, \mathbf{D}_{1c}, \mathbf{D}_n|\mathbf{y}, \mathbf{x}, \mathbf{s}_{00}, \boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \{\mathbf{b}_{i0c}\}, \{\mathbf{b}_{i1c}\}, \{\mathbf{b}_{in}\}, \sigma$ from

$$\mathcal{W}_k\left(\mathbf{D}_{jc}^{-1}|\rho_{jc,0}+n_{jc}, \left[\mathbf{R}_{jc,0}^{-1}+\sum_{i\in N_{jc}}\mathbf{b}_{ji}\mathbf{b}'_{ji}\right]^{-1}\right), \quad j=0,1$$

$$\mathcal{W}_k\left(\mathbf{D}_n^{-1}|\rho_{n,0}+n_n, \left[\mathbf{R}_{n,0}^{-1}+\sum_{i\in N_{0n}}\mathbf{b}_i\mathbf{b}'_i\right]^{-1}\right)$$