Web Appendix to
Bayesian methods for non-ignorable dropout in joint models in smoking cessation studies

J. T. Gaskins, M. J. Daniels, and B. H. Marcus

A.1. Further details about model specification

In this section of the appendix, we provide further details and proofs for results in Section 3 of the paper. In Section 3.1, we claim that partial ignorability implies that inference on $\mathbf{Y}$ only requires the distribution $p(\mathbf{y}, d)$ instead of $p(\mathbf{y}, \mathbf{r}) = p(\mathbf{y}, d, \mathbf{r})$ (recall D is a many-to-one function of $\mathbf{R}$). Similarly to Harel and Schafer (2009), it follows that

$$\begin{aligned} p(\mathbf{y}_{\text{obs}}, d, \mathbf{r}|\boldsymbol{\theta}) &= \int p(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, d, \mathbf{r}|\boldsymbol{\theta})\, \mathrm{d}\mathbf{y}_{\text{mis}} \\ &= \int p(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}|\boldsymbol{\theta}_1)\, p(d|\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \boldsymbol{\theta}_{2A})\, p(\mathbf{r}|d, \mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \boldsymbol{\theta}_{2B})\, \mathrm{d}\mathbf{y}_{\text{mis}} \\ &= \int p(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}|\boldsymbol{\theta}_1)\, p(d|\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \boldsymbol{\theta}_{2A})\, p(\mathbf{r}|d, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}_{2B})\, \mathrm{d}\mathbf{y}_{\text{mis}} \qquad \text{(A.1)} \\ &= p(\mathbf{r}|d, \mathbf{y}_{\text{obs}}, \boldsymbol{\theta}_{2B}) \times \int p(\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}|\boldsymbol{\theta}_1)\, p(d|\mathbf{y}_{\text{obs}}, \mathbf{y}_{\text{mis}}, \boldsymbol{\theta}_{2A})\, \mathrm{d}\mathbf{y}_{\text{mis}}, \end{aligned}$$

where (A.1) is a consequence of partial ignorability. Our goal for inference is $\boldsymbol{\theta}_1$, the parameter that describes the full data $\mathbf{y}$. Due to the integration over the missing data, identification of $\boldsymbol{\theta}_1$ requires the model for dropout time $p(d|\mathbf{y})$ but not the model for the missingness indicators $p(\mathbf{r}|d, \mathbf{y}_{\text{obs}})$. Hence, we base inference on the distribution of $p(\mathbf{y}, d)$, which we then factor as $p(d)p(\mathbf{y}|d)$ as in the PMM.

Next, we show that the joint distribution of $\overline{\mathbf{Y}}_{i,D_i+1} = (\mathbf{Y}_{i1}^\top, \ldots, \mathbf{Y}_{i,D_i}^\top)^\top$ given $D_i = d$ is $\text{N}_{2d}\left(\boldsymbol{\Phi}_d^{-1}\boldsymbol{\zeta}_d, \boldsymbol{\Phi}_d^{-1}\boldsymbol{\Omega}_d\boldsymbol{\Phi}_d^{-\top}\right)$ as discussed in Section 3.2. We can write $\mathbf{Y}_{it} \sim$

$\mathrm{N}_2\left(\boldsymbol{\zeta}_{d;t} + (\boldsymbol{\Phi}_{d;1t}, \ldots, \boldsymbol{\Phi}_{d;t-1,t})\overline{\mathbf{y}}_{it}, \boldsymbol{\Omega}_{d;t}\right)$ as

$$\mathbf{Y}_{it} = \boldsymbol{\zeta}_{d;t} + (\boldsymbol{\Phi}_{d;1t}, \ldots, \boldsymbol{\Phi}_{d;t-1,t})\overline{\mathbf{y}}_{it} + \boldsymbol{\epsilon}_{it},$$

where the residual is $\boldsymbol{\epsilon}_{it} \sim \mathrm{N}_2(\mathbf{0}_2, \boldsymbol{\Omega}_{d;t})$. By stacking these conditional regressions ($t = 1, \ldots, D_i$), we have

$$\overline{\mathbf{Y}}_{i,D_i+1} = \boldsymbol{\zeta}_d + \mathbf{T}_d\overline{\mathbf{Y}}_{i,D_i+1} + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{\epsilon}_{it} \sim \mathrm{N}_{2d}(\mathbf{0}_{2d}, \boldsymbol{\Omega}_d)$ and $\mathbf{T}_d$ is a $2d \times 2d$ lower-triangular block matrix with (t, j) $(j < t)$ block $\boldsymbol{\Phi}_{d;jt}$ and the zero matrix block otherwise. Rearranging shows $(\mathbf{I}_{2d} - \mathbf{T}_d)\overline{\mathbf{Y}}_{i,D_i+1} = \boldsymbol{\zeta}_d + \boldsymbol{\epsilon}_i$, and noting that $\mathbf{I}_{2d} - \mathbf{T}_d$ is our definition of $\boldsymbol{\Phi}_d$, gives $\overline{\mathbf{Y}}_{i,D_i+1} = \boldsymbol{\Phi}_d^{-1}\boldsymbol{\zeta}_d + \boldsymbol{\Phi}_d^{-1}\boldsymbol{\epsilon}_i$, which is a multivariate normal with mean $\boldsymbol{\Phi}_d^{-1}\boldsymbol{\zeta}_d$ and covariance matrix $\boldsymbol{\Phi}_d^{-1}\boldsymbol{\Omega}_d\boldsymbol{\Phi}_d^{-\top}$.

As discussed in Section 3.2, the contribution of patient i to the observed data likelihood is

$$p(d_i, \overline{\mathbf{y}}_{i,d_i+1}|\boldsymbol{\theta}) = \pi_{d_i}\, f_{d_i;1}(\mathbf{y}_{i1}) \prod_{t=2}^{d_i} f_{d_i;t}(\mathbf{y}_{it}|\overline{\mathbf{y}}_{it}), \tag{A.2}$$

assuming no intermittent missingness and that we know the values of the latent quit propensities $\mathbf{Z}$. As this is not the case, the observed data likelihood is better expressed by

$$p(d_i, \mathbf{q}_{i,\text{obs}}, \mathbf{w}_{i,\text{obs}}|\boldsymbol{\theta}) = \pi_{d_i} \int I(q_{it}z_t \geq 0\,\forall t)\, f_{d_i;1}(\mathbf{y}_{i1}) \prod_{t=2}^{d_i} f_{d_i;t}(\mathbf{y}_{it}|\overline{\mathbf{y}}_{it})\,\mathrm{d}\mathbf{z}\,\mathrm{d}\mathbf{w}_{\text{int}},$$

acknowledging that we must integrate over the latent $\mathbf{z} = (z_1, \ldots, z_{d_i})$ and the intermittently-missed weight change measures $\mathbf{w}_{\text{int}} = \{w_t : t \leq d_i, r_{it} = 0\}$ (see also (3)). Except to calculate DIC, we work with (A.2) and use data augmentation to sample values $\mathbf{z}$ and $\mathbf{w}_{\text{int}}$.

The (data-augmented) posterior is given by $p(\boldsymbol{\theta}|d, \mathbf{y}_{\text{obs}}) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{N} p(d_i, \overline{\mathbf{y}}_{i,d_i+1}|\boldsymbol{\theta})$, where the joint prior is

$$\begin{aligned} p(\boldsymbol{\theta}) \;=\; & p(\tau_\zeta^2)\, p(\tau_\phi^2)\, p(\tau_\rho^2)\, p(\tau_\omega^2)\, p(\sigma_\zeta^2)\, p(\lambda)\, p(\gamma_0)\, p(\xi)\, p(\lambda_1)\, p(\lambda_2) \\ & \times \prod_{a=0}^{1}\prod_{t=1}^{T} \left\{ p(\boldsymbol{\zeta}_t^\star|\sigma_\zeta^2)\, p(\rho_t^\star)\, p(\omega_t^\star|\lambda_1, \lambda_2) \prod_{j=1}^{t-1} p(\boldsymbol{\Phi}_{jt}^\star|\lambda, \gamma_0, \xi) \right\} \\ & \times \prod_{a=0}^{1} \left\{ p(\boldsymbol{\pi}) \prod_{d=1}^{T}\prod_{t=1}^{d} \left[p(\boldsymbol{\zeta}_{d;t}|\boldsymbol{\zeta}_t^\star, \tau_\zeta^2)\, p(\rho_{d;t}|\rho_t^\star, \tau_\rho^2)\, p(\omega_{d;t}|\omega_t^\star, \tau_\omega^2) \prod_{j=1}^{t-1} p(\boldsymbol{\Phi}_{d;jt}|\boldsymbol{\Phi}_{jt}^\star, \tau_\phi^2) \right] \right\}. \end{aligned}$$

Note that the hyperparameters on the first line are common across the treatment groups $a = 0$ and $a = 1$. Parameters drawn on the second and third lines are treatment-specific although we continue to suppress dependence on a in the notation. We finally note that even under MAR our model is non-ignorable as the second and third conditions (the parameters of the MDM and marginal response model are separable and a priori independent) are violated. This is clear as our model is parametrized through the marginal distribution of dropout and the response conditionally on dropout.

A.2. Sampling distributions

Sampling from the model under MAR as in Section 4 proceeds by sampling from each of the conditional distributions from the data-augmented posterior. Many of the distributions may be updated conjugately, and we use slice sampling (Neal, 2003) for those that are not. We state the necessary conditional distributions under the SHRINK/SHRINK/SPARSE model.

- Dropout probabilities: For $a = 0, 1$, $\boldsymbol{\pi}_a \sim \text{Dirichlet}(\boldsymbol{\alpha})$, where α_t is the sum of the prior concentration parameter and the number of patients in treatment group a that dropout at time t.

- Data augmentation: For each patient i, we sample the missing components of $\mathbf{y}$. As stated in Section 3.2, the distribution of the observable data $\overline{\mathbf{Y}}_{i,D_i+1}$ is $\text{N}_{2D_i}(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$ where $\boldsymbol{\mu}_d = \boldsymbol{\Phi}_d^{-1}\boldsymbol{\zeta}_d$ and $\boldsymbol{\Sigma}_d = \boldsymbol{\Phi}_d^{-1}\boldsymbol{\Omega}_d\boldsymbol{\Phi}_d^{-\top}$. The missing data that we need to sample are the components of $\overline{\mathbf{Y}}_{i,D_i+1}$ that correspond to the latent quit status variables Z_{it} ($t = 1, \ldots, D_i$) and any W_{it} that were intermittently missed. Denote these by $\overline{\mathbf{Y}}_{i,D_i+1,\text{mis}}$ and the observed counterpart by $\overline{\mathbf{Y}}_{i,D_i+1,\text{obs}}$. The conditional distribution for $\overline{\mathbf{Y}}_{i,D_i+1,\text{mis}}$ is a restriction of the multivariate normal with mean $\boldsymbol{\mu}_{d,\text{mis}} + \boldsymbol{\Sigma}_{\text{mis,obs}}\boldsymbol{\Sigma}_{\text{obs,obs}}^{-1}(\overline{\mathbf{Y}}_{i,D_i+1,\text{obs}} - \boldsymbol{\mu}_{d,\text{obs}})$ and covariance matrix $\boldsymbol{\Sigma}_{\text{mis,mis}} - \boldsymbol{\Sigma}_{\text{mis,obs}}\boldsymbol{\Sigma}_{\text{obs,obs}}^{-1}\boldsymbol{\Sigma}_{\text{obs,mis}}$. Here $\boldsymbol{\mu}_{d,\text{mis}}$ denote the vector made up of the rows of $\boldsymbol{\mu}$ that correspond to the variables in $\overline{\mathbf{Y}}_{i,D_i+1,\text{mis}}$; $\boldsymbol{\mu}_{d;\text{obs}}, \boldsymbol{\Sigma}_{\text{mis,mis}}, \boldsymbol{\Sigma}_{\text{mis,obs}}, \boldsymbol{\Sigma}_{\text{obs,obs}}$ are defined

similarly. This distribution is restricted by the fact that the elements of $\overline{\mathbf{Y}}_{i,D_i+1,\text{mis}}$ corresponding to an observed quit status must be positive if $Q = 1$ and negative if $Q = 0$. An efficient sampler for this restricted normal can be found in Liu et al. (2009, Proposition 1).

- Conditional intercept: For each treatment group $a = 0, 1$ and each pattern $d = 1, \ldots, T$, $\boldsymbol{\zeta}_d = (\boldsymbol{\zeta}_{d;1}^\top, \ldots, \boldsymbol{\zeta}_{d;d}^\top)^\top$ is multivariate normal with covariance matrix $\boldsymbol{\Sigma}_\zeta = \left(\tau_\zeta^{-1}\mathbf{I}_{2d} + N_d\boldsymbol{\Omega}_d^{-1}\right)^{-1}$ and mean $\boldsymbol{\Sigma}_\zeta^{-1}\left(\tau_\zeta^{-1}\boldsymbol{\zeta}_{1:d}^\star + N_d\boldsymbol{\Omega}_d^{-1}\boldsymbol{\Phi}_d\tilde{\mathbf{Y}}_d\right)$, where N_d is the number of patient in treatment a that dropout at d, $\boldsymbol{\zeta}_{1:d}^\star = (\boldsymbol{\zeta}_1^{\star\top}, \ldots, \boldsymbol{\zeta}_d^{\star\top})^\top$, and $\tilde{\mathbf{Y}}_d$ is the average of $\overline{\mathbf{Y}}_{i,d+1}$ across the patients i with $d_i = d$.

- GARP matrix: For $a = 0, 1$, $d = 2, \ldots, T$, and $t = 2, \ldots, d$, we jointly update the elements associated with the $2 \times 2(d-1)$ matrix $\boldsymbol{\Phi}_{d;t} = [\boldsymbol{\Phi}_{d;1t}\ \boldsymbol{\Phi}_{d;2t} \cdots \boldsymbol{\Phi}_{d;t-1,t}]$. Let $\mathbf{P}_{d;t}$ be the $4(d-1) \times 1$ vector formed by row-wise concatenation. Form $\mathbf{P}_t^\star$ similarly from $\boldsymbol{\Phi}_{1t}^\star, \ldots, \boldsymbol{\Phi}_{t-1,t}^\star$. Then, for patients i in treatment group a and pattern d, let $\mathbf{X}_{it}$ be the $4(d-1) \times 2$ matrix with $\overline{\mathbf{Y}}_{it}$ in the first $2(d-1)$ rows of column 1 and again in the final $2(d-1)$ rows of column 2 with zeroes elsewhere. The distribution of $\mathbf{P}_{d;t}$ is multivariate normal with covariance matrix $\boldsymbol{\Sigma}_P = \left(\tau^{-2}\mathbf{I} + \sum_i \mathbf{X}_{it}\boldsymbol{\Omega}_{d;t}^{-1}\mathbf{X}_{it}^\top\right)^{-1}$ and mean vector $\boldsymbol{\Sigma}_P^{-1}\left[\tau^{-2}\mathbf{P}_t^\star + \sum_i \mathbf{X}_{it}\boldsymbol{\Omega}_{d;t}^{-1}(\mathbf{Y}_{it} - \zeta_{d;t})\right]$.

- Correlation and innovation variance: Under the SHRINK dependence structure we sample these parameters through their transformed values $r_{d;t} = \log[(1+\rho_{d;t})/(1-\rho_{d;t})]$ and $w_{d;t} = \log(\omega_{d;t})$. For $a = 0, 1$, $d = 1, \ldots, T$, and $t = 1, \ldots, d$, the distribution of $(r_{d;t}, w_{d;t})$ is proportional to

$$|\boldsymbol{\Omega}(r_{d;t}, w_{d;t})|^{-N_d/2} \exp\left\{\sum_{i:a_i=a, D_i=d} \frac{-1}{2}\mathbf{e}_{it}^\top\boldsymbol{\Omega}(r_{d;t}, w_{d;t})^{-1}\mathbf{e}_{it}\right.$$

$$\left. + \frac{-1}{2\tau_\rho^2}\left[r_{d;t} - \log\left(\frac{1+\rho_t^\star}{1-\rho_t^\star}\right)\right]^2 + \frac{-1}{2\tau_\omega^2}\left(w_{d;t} - \log(\omega_t^\star)\right)^2\right\},$$

where $\mathbf{e}_{it} = \mathbf{y}_{it} - (\boldsymbol{\Phi}_{d;1t}, \ldots, \boldsymbol{\Phi}_{d;t-1,t})\overline{\mathbf{y}}_{it}$ is the bivariate residual for patient i at time t and $\boldsymbol{\Omega}(r_{d;t}, w_{d;t})$ is the covariance matrix corresponding to the transformed values of $\rho_{d;t}$ and

$\omega_{d;t}$. We apply univariate slice sampling steps (Neal, 2003) to sample $r_{d;t}$ given $w_{d;t}$ and then $w_{d;t}$ given $r_{d;t}$.

- Conditional intercept shrinkage target: For $a = 0, 1$ and $t = 1, \ldots, T$, sample $\boldsymbol{\zeta}_t^\star$ from multivariate normal with covariance $\boldsymbol{\Sigma}_{\zeta\star} = \left(\sigma_\zeta^{-2} + \tau_\zeta^{-2}(T - d + 1)\right)^{-2} \mathbf{I}_2$ and mean vector $\tau_\zeta^{-2}\boldsymbol{\Sigma}_{\zeta\star}^{-1}\left(\sum_{d=t}^T \boldsymbol{\zeta}_{d;t}\right)$.

- GARP matrix shrinkage target: For $a = 0, 1$ and $t = 2, \ldots, T$, we update $\mathbf{P}_t^\star$ (using the notation from the GARP matrix step) by sampling from a multivariate normal with covariance matrix $\boldsymbol{\Sigma}_{P\star} = \left(\boldsymbol{\Sigma}_{NG}^{-1} + \tau_\phi^2(T - t + 1)\mathbf{I}\right)^{-1}$ and mean vector $\tau_\phi^{-2}(T - t + 1)^{-1}\boldsymbol{\Sigma}_{P\star}^{-1}\left(\sum_{d=t}^T \mathbf{P}_{d;t}\right)$, where $\boldsymbol{\Sigma}_{NG}$ is the $4(d-1) \times 4(d-1)$ diagonal matrix containing the elements $\sigma_{jt;k}^2$ in the appropriate location to correspond with the $\phi_{jt;k}^\star$ element of $\mathbf{P}_t^\star$.

- Correlation and innovation variance shrinkage target: For $a = 0, 1$ and $t = 1, \ldots, T$, we use univariate slice sampling to draw $(\rho_t^\star, \omega_t^\star)$ from the distribution proportional to

$$(\omega_t^\star)^{-\lambda_1 - 1}\mathrm{e}^{-\lambda_2/\omega_t^\star} \prod_{d=t}^T \exp\left\{\frac{-1}{2\tau_\rho^2}\left[\log\left(\frac{1+\rho_{d;t}}{1-\rho_{d;t}}\right) - \log\left(\frac{1+\rho_t^\star}{1-\rho_t^\star}\right)\right]^2 + \frac{-1}{2\tau_\omega^2}\left[\log(\omega_{d;t}) - \log(\omega_t^\star)\right]^2\right\}.$$

- Shrinkage variance: For $k \in \{\zeta, \phi, \rho, \omega\}$, the conditional distribution for τ_k given the other parameters is proportional to $p(\tau_k|-) \propto \tau_k^{-a_k/2} \exp\left\{\frac{-S_k^2}{2\tau_k^2}\right\}[\tau_k^2 + \gamma_k^2]^{-1}$, where a_k is the number of parameters and S_k^2 is a sum of squares term associated with shrinkage of parameter k. These distributions are drawn by slice sampling, and as needed we subscript parameters by the treatment group $a = 0, 1$.

$$\begin{aligned}
a_\zeta &= 2T(T+1), \quad & S_\zeta^2 &= \textstyle\sum_{a=0}^1 \sum_{d=1}^T \sum_{t=1}^d (\boldsymbol{\zeta}_{a;d;t} - \boldsymbol{\zeta}_{a;t}^\star)^\top (\boldsymbol{\zeta}_{a;d;t} - \boldsymbol{\zeta}_{a;t}^\star), \\
a_\phi &= \tfrac{4}{3}T(T^2-1), \quad & S_\phi^2 &= \textstyle\sum_{a=0}^1 \sum_{d=1}^T \sum_{t=1}^d \sum_{j=1}^{t-1} \\
& & & \left(\mathrm{vec}(\boldsymbol{\Phi}_{a;d;jt}) - \mathrm{vec}(\boldsymbol{\Phi}_{a;jt}^\star)\right)^\top \left(\mathrm{vec}(\boldsymbol{\Phi}_{a;d;jt}) - \mathrm{vec}(\boldsymbol{\Phi}_{a;jt}^\star)\right), \\
a_\rho &= T(T+1), \quad & S_\rho^2 &= \textstyle\sum_{a=0}^1 \sum_{d=1}^T \sum_{t=1}^d \left(\log\left[\frac{1+\rho_{a;d;t}}{1-\rho_{a;d;t}}\right] - \log\left[\frac{1+\rho_{a;t}^\star}{1-\rho_{a;t}^\star}\right]\right)^2, \\
a_\omega &= T(T+1), \quad & S_\omega^2 &= \textstyle\sum_{a=0}^1 \sum_{d=1}^T \sum_{t=1}^d \left(\log(\omega_{a;d;t}) - \log(\omega_{a;t}^\star)\right)^2
\end{aligned}$$

- Variance of conditional intercept shrinkage targets: When the prior is InvGamma(h_1, h_2), σ_ζ^2 is sampled from InvGamma $\left(2T + h_1, 0.5 \sum_{a=0}^{1} \sum_{t=1}^{T} \boldsymbol{\zeta}_{a;t}^{\star\top} \boldsymbol{\zeta}_{a;t}^{\star} + h_2\right)$.

- Shape parameter of innovation variance shrinkage targets: With prior $\lambda_1 \sim$ Gamma(h_1, h_2), we update λ_1 by slice sampling from the distribution proportional to

$$\Gamma(\lambda_1)^{-2T} \lambda_2^{2T\lambda_1} \lambda_1^{h_1 - 1} \exp\left\{-\lambda_1 \left[h_2^{-1} + \sum_{a=0}^{1} \sum_{t=1}^{T} \log(\omega_{a;t}^{\star})\right]\right\}.$$

- Scale parameter of innovation variance shrinkage targets: With prior $\lambda_2 \sim$ Gamma(h_1, h_2), we sample λ_2 from a Gamma distribution with shape $2T\lambda_1 + h_1$ and scale $h_2^{-1} + \sum_{a=0}^{1} \sum_{t=1}^{T} (\omega_{a;t}^{\star})^{-1}$.

- GARP shrinkage factors: For $a = 0, 1$, $t = 1, \ldots, T$, $j = 1, \ldots, t-1$, and $k = 1, \ldots, 4$, the distribution of $\sigma_{jt;k}^2$ is proportional to the generalized inverse Gaussian distribution with kernel

$$(\sigma_{jt;k}^2)^{(\lambda - \frac{1}{2}) - 1} \exp\left\{\frac{-1}{2}\left[\frac{\sigma_{jt;k}^2}{\gamma_0 \xi^{t-j}} + \frac{(\phi_{jt;k}^{\star})^2}{\sigma_{jt;k}^2}\right]\right\}.$$

- Parameter for GARP shrinkage factors: We sample (λ, ξ, γ_0) from univariate slice sampling steps according to the distribution proportional to

$$\lambda^{h_1 - 1} \mathrm{e}^{-\lambda/h_2} \gamma_0^{-h_3 - 1} \mathrm{e}^{-h_4/\gamma_0} \prod_{a,j,t,k} \frac{1}{\Gamma(\lambda)\,(2\gamma_0 \xi^{t-j})^{\lambda}} (\omega_{a;jt;k}^{\star})^{\lambda - 1} \exp\left\{\frac{-\sigma_{a;jt;k}^2}{2\gamma_0 \xi^{t-j}}\right\},$$

where the priors are $\lambda_1 \sim$ Gamma(h_1, h_2), $\gamma_0 \sim$ InvGamma(h_3, h_4) and $\xi \sim$ Unif$(0, 1)$.

A.3. Simulation Studies on Parameter Estimation and Model Comparison

In this section we provide further details about the simulation study from Section 4. Recall that we consider data generating model (A) where the "true" model is specified using the parameter estimates from the SHRINK/EQUAL/SPARSE model applied to the CTQ2 data. We also consider

data generating model (B) where the mean follows the MVN-MAR assumption, and the parameter values are chosen by adjusting the parameters in (A) toward this assumption. Data generating model (C) is chosen by adjusting parameters so that the $\zeta_{d;t}$ differ more substantially across patterns, relative to model (A), so as to favor the PATTERN mean model or SHRINK with a large value of τ_ζ. In contrast to the choices (A)–(C) which all assume common dependence structures across patterns, we include model (D) which adjusts the parameters from the CTQ2 data with the SHRINK/SHRINK/SPARSE model to allow GARPs, correlation, and innovation variances to vary across patterns. Figure A.1 displays parameter values for the wellness treatment under each of these generating choices.

In addition to the risk analysis described in Section 4 of the manuscript, we also consider the performance of a common model selection criterion to distinguish models fit to our simulated data. We use the deviance information criterion (DIC; Spiegelhalter et al., 2002). The DIC is composed of the deviance at a parameter estimate $\hat{\boldsymbol{\theta}}$, given by $Dev = -2\sum_{i=1}^{N} \log p(d_i, \mathbf{q}_{i,\text{obs}}, \mathbf{w}_{i,\text{obs}}|\hat{\boldsymbol{\theta}})$, and a term p_D measuring the model complexity, the difference of the posterior expected deviance $\mathrm{E}_{\text{post}}\left(-2\sum_{i=1}^{N} \log p(d_i, \mathbf{q}_{i,\text{obs}}, \mathbf{w}_{i,\text{obs}}|\boldsymbol{\theta})\right)$ and Dev. p_D is often interpreted as the effective number of model parameters, and models with smaller DIC $= Dev + 2p_D$ are considered to better balance the model fit and complexity. Note that we use the observed data likelihood (3) in this calculation. While the integral in (3) is not available in closed form, it can be estimated using importance sampling; see Gaskins et al. (2014, with appendix). It is well known that computation of Dev is not invariant to the choice of $\hat{\boldsymbol{\theta}}$, and the standard estimator of the covariance matrix $\left[\mathrm{E}_{\text{post}}\left(\boldsymbol{\Phi}_d \boldsymbol{\Omega}_d^{-1} \boldsymbol{\Phi}_d'\right)\right]^{-1}$ fails to maintain the structure of ones required by the unidentifiability of the scale of Z_{it}. Hence, our estimator $\hat{\boldsymbol{\theta}}$ uses the posterior means of $\boldsymbol{\zeta}_d$, $\boldsymbol{\Phi}_d$, and $\rho_{d;t}$, and for $\hat{\omega}_{d;t}$ we use $\left[\mathrm{E}_{\text{post}}(\omega_{d;t}^{-1})\right]^{-1}$. These are the same parameter estimators we use in the risk analysis.

For each of the four model scenarios (A)–(D), we compute the DIC value for the five estimation models and rank the models in each of the 100 data sets. Table A.1 provides the proportion of times each model is selected. When the true model is (A), the correct generating model

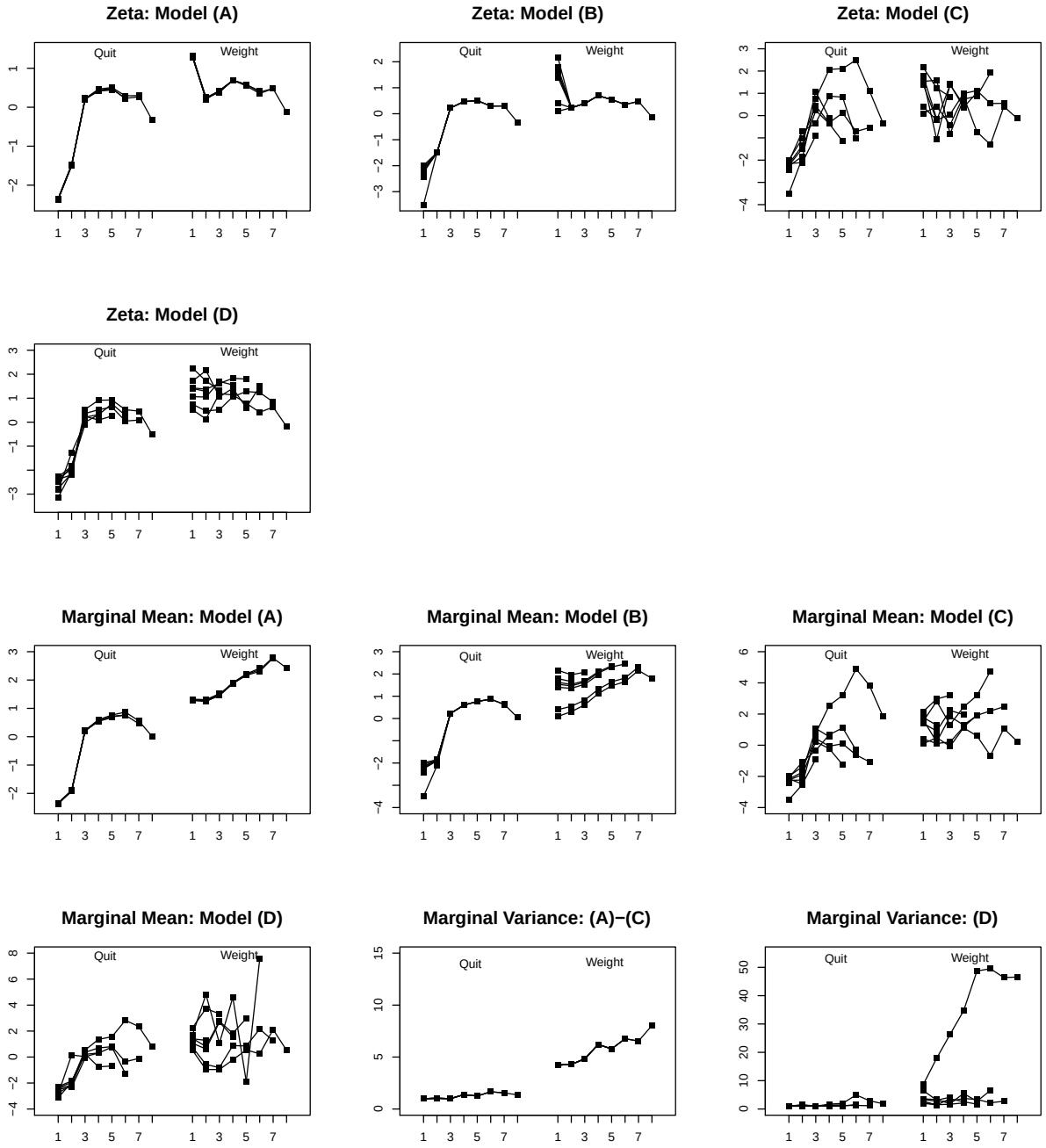

Figure A.1: Wellness treatment parameter values for data generating models in simulation study. Zeta refers to the condition intercepts $\boldsymbol{\zeta}_{d;t}$, marginal mean is $\boldsymbol{\Phi}_d^{-1}\boldsymbol{\zeta}_d$, and marginal variance are the diagonal elements from $\boldsymbol{\Phi}_d^{-1}\boldsymbol{\Omega}_d\boldsymbol{\Phi}^{-\top}$ (see section A.1).

DIC Ranking	Estimation Model				
Mean:	MVN-MAR	PATTERN	SHRINK	SHRINK	PATTERN
Dependence:	EQUAL	EQUAL	EQUAL	SHRINK	PATTERN
GARP:	SPARSE	SPARSE	SPARSE	SPARSE	NON-SPARSE
		Data Generating Model (A)			
Best	8	3	89	0	0
2nd	79	14	4	3	0
3rd	13	63	6	18	0
4th	0	17	1	78	4
Worst	0	3	0	1	96
		Data Generating Model (B)			
Best	11	9	80	0	0
2nd	79	10	11	3	0
3rd	10	70	8	12	0
4th	0	8	1	86	5
Worst	0	3	0	2	95
		Data Generating Model (C)			
Best	3	33	63	0	1
2nd	9	57	29	4	1
3rd	63	6	3	27	1
4th	21	3	2	67	7
Worst	4	1	3	2	90
		Data Generating Model (D)			
Best	14	51	33	1	1
2nd	7	35	54	4	0
3rd	70	9	11	9	1
4th	7	2	1	86	4
Worst	2	3	1	0	94

Table A.1: Performance of DIC model selection statistic

(SHRINK/EQUAL/SPARSE) is selected 89 times out 100, followed by MVN-MAR and PATTERN mean structures. In scenario (B) where the true model has the MVN-MAR structure, the SHRINK mean structure is still selected most often by DIC although the MVN-MAR structure is consistently second. Recall from the risk simulations that even though MVN-MAR is the correct model, the SHRINK model produces parameter estimates with lower risk. Hence, it may be unfair to call the selection of SHRINK/EQUAL/SPARSE the "wrong model" since it has better estimates on average. In scenario (C), the mean parameters differ significantly across patterns, and so both PATTERN and SHRINK (with large τ_ζ) are correct models. For 96 of the 100 data sets, one of these two are the chosen model, with SHRINK mean winning 63 times.

For scenario (D) which has distinct dependence structures for each pattern, we find that the DIC favors the models with EQUAL dependence structure. While this is clearly the wrong model choice, it is important to remember that DIC is designed to balance the model fit to the data relative the model complexity (number of parameters). Even though the models fit using the SHRINK and PATTERN dependence parameters are expected to be better, these models are much more complex (more parameters) and have much larger p_D values. With only 208 partially observed patients, there is not enough information in the data to support the estimation of this many parameters, and the model that is estimated using EQUAL dependence is relatively close to the fit from the more complex models. This is especially the case when the dropout patterns are as unbalanced as those that we observe (Table 1). With fewer than 10 patients in the non-completer groups, we do not reasonably expect to estimate the dependence well. As the sample size increases, one would expect DIC to favor the SHRINK and PATTERN choices asymptotically. But given the complexity of these models, it is unclear how much larger the data would need to be.

While we acknowledge that the DIC does not perform as well as anticipated, we are unable to find a model selection criteria that performs better in our scenario. The log pseudo-marginal likelihood criterion based on the condition predictive ordinate (Geisser and Eddy, 1979; Christensen et al., 2011) selects the true generating model with similar or worse probability. Further, because

the conditional predictive ordinates are estimated using a harmonic mean of the likelihood, we observe major instability in computation that may indicate the estimates have infinite variance. Simulations that perform model selection using posterior predictive measures (Ibrahim and Laud, 1994; Daniels et al., 2012) are highly dependent on the choice of summary statistics and tend to produce very similar values across all models. Further research into model selection methods tailored specifically for mixed outcome data with missingness is needed to develop criteria that more consistently choose the correct model. This is beyond the scope of the current work.

A.4. MCMC for MNAR

Here, we provide details behind the assertion in Section 6.1 that it is not necessary to rerun the MCMC analysis when using a sensitivity parameter. We are interested in the posterior distribution of the full parameter vector $(\boldsymbol{\theta}_O, \boldsymbol{\theta}_E)$ conditional on the observed data $\mathbf{y}_{\text{obs}}, \mathbf{r}$.

$$\begin{aligned}
\pi(\boldsymbol{\theta}_O, \boldsymbol{\theta}_E | \mathbf{y}_{\text{obs}}, \mathbf{r}) &\propto \pi(\boldsymbol{\theta}_O)\, \pi(\boldsymbol{\theta}_E | \boldsymbol{\theta}_O)\, p(\mathbf{y}_{\text{obs}}, \mathbf{r} | \boldsymbol{\theta}_O, \boldsymbol{\theta}_E) \\
&= \pi(\boldsymbol{\theta}_O)\, \pi(\boldsymbol{\Delta} | \boldsymbol{\theta}_O)\, p(\mathbf{y}_{\text{obs}}, \mathbf{r} | \boldsymbol{\theta}_O, \boldsymbol{\Delta}) \qquad \text{(A.3)} \\
&= \pi(\boldsymbol{\theta}_O)\, \pi(\boldsymbol{\Delta} | \boldsymbol{\theta}_O)\, p(\mathbf{y}_{\text{obs}}, \mathbf{r} | \boldsymbol{\theta}_O, \boldsymbol{\Delta} = 0) \qquad \text{(A.4)} \\
&\propto \pi(\boldsymbol{\Delta} | \boldsymbol{\theta}_O)\, \pi(\boldsymbol{\theta}_O | \mathbf{y}_{\text{obs}}, \mathbf{r}, \boldsymbol{\Delta} = 0).
\end{aligned}$$

As the extrapolation parameter $\boldsymbol{\theta}_E$ is a function of the observed data parameter $\boldsymbol{\theta}_O$ and the sensitivity parameter $\boldsymbol{\Delta}$, we may replace $\boldsymbol{\theta}_E$ by $\boldsymbol{\Delta}$ in (A.3). Since the sensitivity parameter plays no role in the observed data likelihood, (A.4) follows, and $\pi(\boldsymbol{\theta}_O | \mathbf{y}_{\text{obs}}, \mathbf{r}, \boldsymbol{\Delta} = 0)$ is precisely the posterior corresponding to non-ignorable MAR. Hence, we obtain a posterior sample of $(\boldsymbol{\theta}_O, \boldsymbol{\theta}_E)$ (equivalently, $(\boldsymbol{\theta}_O, \boldsymbol{\Delta})$) by drawing $\boldsymbol{\theta}_O$ from the MAR sample and drawing $\boldsymbol{\Delta}$ from the elicited prior.

Table A.2 displays the algorithm used for drawing pseudo-patients from the PMM under MNAR. We draw the full data of $N = 5000$ observations for each of the $G = 1500$ values of $\boldsymbol{\theta}_0$ in our MAR posterior sample. To draw samples under MAR, the sensitivity parameters are

Table A.2: Pseudo-code for sampling pseudo-patients under MNAR

for g in $1, \ldots, G$ (indexing the MAR posterior samples $\boldsymbol{\theta}_O^{(g)}$) **do**
  for i in $1, \ldots, N$ (indexing the pseudo-patients within g) **do**
    Sample $D_i \sim$ Multinomial($\boldsymbol{\pi}^{(g)}$)
    for t in $1, \ldots, D_i$ **do**
      Sample $\mathbf{Y}_{it}|\overline{\mathbf{Y}}_{it} \sim f_{D_i,t}(\cdot|\overline{\mathbf{Y}}_{it}, \boldsymbol{\theta}_O^{(g)})$
    end for
    if $D_i < T$ **then**
      $t \leftarrow D_i + 1$
      Compute $\hat{p} = P(Q_{it} = 0|\overline{\mathbf{Y}}_{it}, D \geq t, \boldsymbol{\theta}_O^{(g)})$ from (15) with $\Delta_1 = 0$
      Given $\hat{p}$, draw $\tilde{p} \sim \frac{1}{2}\text{Unif}\left(\psi_{\text{LB}}(\hat{p}), \psi_{\text{med}}(\hat{p})\right) + \frac{1}{2}\text{Unif}\left(\psi_{\text{med}}(\hat{p}), \psi_{\text{UB}}(\hat{p})\right)$
      Given $\tilde{p}$, solve (15) for Δ_1
      Sample $\Delta_2 \sim \frac{1}{2}\text{Unif}(\delta_{\text{LB}}, \delta_{\text{med}}) + \frac{1}{2}\text{Unif}(\delta_{\text{med}}, \delta_{\text{UB}})$
      Sample $S \sim$ Multinomial($P(D = s|\overline{\mathbf{Y}}_{it}, D \geq t, \boldsymbol{\theta}_O^{(g)})$)
      Sample $\mathbf{Y}_{it}|\overline{\mathbf{Y}}_{it} \sim f_{S,t}(\cdot|\overline{\mathbf{Y}}_{it}, \boldsymbol{\theta}_O^{(g)})$
      Shift $\mathbf{Y}_{it} \leftarrow \mathbf{Y}_{it} + (\Delta_1, \Delta_2)$
    end if
    if $D_i < T - 1$ **then**
      for t in $D_i + 2, \ldots, T$ **do**
        $(\Delta_1, \Delta_2) \leftarrow (0, 0)$
        Sample $S \sim$ Multinomial($P(D = s|\overline{\mathbf{Y}}_{it}, D \geq t - 1, \boldsymbol{\theta}_O^{(g)})$)
        Sample $\mathbf{Y}_{it}|\overline{\mathbf{Y}}_{it} \sim f_{S,t}(\cdot|\overline{\mathbf{Y}}_{it}, \boldsymbol{\theta}_O^{(g)})$
        if $S = t - 1$ **then**
          Compute $\hat{p} = P(Q_{it} = 0|\overline{\mathbf{Y}}_{it}, D \geq t, \boldsymbol{\theta}_O^{(g)})$ from (15) with $\Delta_1 = 0$
          Given $\hat{p}$, draw $\tilde{p} \sim \frac{1}{2}\text{Unif}\left(\psi_{\text{LB}}(\hat{p}), \psi_{\text{med}}(\hat{p})\right) + \frac{1}{2}\text{Unif}\left(\psi_{\text{med}}(\hat{p}), \psi_{\text{UB}}(\hat{p})\right)$
          Given $\tilde{p}$, solve (15) for Δ_1
          Sample $\Delta_2 \sim \frac{1}{2}\text{Unif}(\delta_{\text{LB}}, \delta_{\text{med}}) + \frac{1}{2}\text{Unif}(\delta_{\text{med}}, \delta_{\text{UB}})$
          Shift $\mathbf{Y}_{it} \leftarrow \mathbf{Y}_{it} + (\Delta_1, \Delta_2)$
        end if
      end for
    end if
  end for
end for

identically zero, and so lines 9–12, 15, 19, 22–28 are ignored. To solve (15) for Δ_1 in lines 11 and 25, we use a naive bisection algorithm which has been found to work well in practice. For each iteration g, the quantities of interest are the cessation probability $N^{-1}\sum_i I(Z_{iT} \geq 0)$, the mean weight change $N^{-1}\sum_i W_{iT}$, and the correlation between $Q_{iT} = I(Z_{iT} \geq 0)$ and W_{iT} at the final week T. We use 5000 pseudo-patients per iteration so that the Monte Carlo error associated with estimating the quit probability given $\boldsymbol{\theta}_O^{(g)}$, the observed data parameter at iteration g, by $N^{-1}\sum_i I(Z_{iT} \geq 0)$ is negligible. Estimates for $P(Q_{iT} = 1)$, $\mathrm{E}(W_{iT})$, and $\mathrm{corr}(Q_{iT}, W_{iT})$ are averaged over iterations, credible intervals formed by taking the sample quantiles from the per-iteration estimates, and posterior probabilities are estimated by the proportion of iterations where the event of interest occurs.

A.5. Alternative choice for eliciting Δ_1 for one-component model

As mentioned in Section 6.2, in the special case where the MAR distribution is a single component, as in the MCAR and MVN-MAR models, a simpler method is available for joining the location shift parameter Δ_1 to the change in smoking rates after dropout elicited by the expert. Because the standard normal distribution function can be approximated by a scaled logistic distribution, $F(x) = [1 + \exp(-kx)]^{-1}$ at $k = 1.749$ (Savalei, 2006), the change in the log-odds of $Q_{it} = 1$ for those who drop out at t and those who remain in the study is approximately $k\Delta_1$; that is,

$$\log\left[\frac{P\left(Q_{it}=1|\overline{\mathbf{y}}_{it}, D=t-1\right)}{P\left(Q_{it}=0|\overline{\mathbf{y}}_{it}, D=t-1\right)}\right] - \log\left[\frac{P\left(Q_{it}=1|\overline{\mathbf{y}}_{it}, D\geq t\right)}{P\left(Q_{it}=0|\overline{\mathbf{y}}_{it}, D\geq t\right)}\right] \approx k\Delta_1.$$

Thus, if the assumption of constant log-odds is reasonable and the subject-matter expert is comfortable with the log-odds scale, we can elicit the distribution for Δ_1 in terms of the change in the odds of smoking after dropout instead of using Table 3. We rescale the expert's values by k^{-1} and form a prior directly on Δ_1 using the mixture of two uniforms. In this case equation (15) no longer needs to be solved, leading to faster computing.

A.6. Simulation Study on Treatment Effect Estimation

Here, we explore the properties of our treatment effect estimates using the first simulation study (A) from Section 4. Again, the parameter values for the data generating model come from the CTQ2 data analysis using the SHRINK/EQUAL/SPARSE model. When the data are generated, the missingness is taken to match the missingness pattern from the CTQ2 data, but the complete data is still generated (conditional on the fixed dropout time). Hence, we have access to the full data (both observed and unobserved) that we can use to compare to both our model-based estimates and a naive choice that only use the study completers. When generating this data, we assume the intermittent missingness is partially ignorable, missingness after dropout is MNAR but does satisfy non-future dependence (NFD), and that the first post-dropout distribution (14) is a location-shift of the MAR distribution where the true value of $\boldsymbol{\Delta}$ is the median-elicited value in Table 3 of the manuscript.

As in the manuscript, we are interested in estimating $P(Q_{iT} = 1)$, $\mathrm{E}(W_{iT})$, and $\mathrm{corr}(Q_{iT}, W_{iT})$. For simplicity we only consider the estimates from the wellness treatment arm. The true values are determined by appending 50,000 data sets drawn from the model and obtaining Monte Carlo estimates of the desired quantities with negligible standard error. Using the MCMC outputs from the SHRINK/EQUAL/SPARSE model for each of the 100 data sets, we obtain 5 estimators. First, we consider the naive moment estimator that use only those patients who complete the study (66 out of 104 patients). We consider three model-based estimators: MAR which uses $\boldsymbol{\Delta} = \mathbf{0}$ as in Section 5; $\boldsymbol{\Delta}$-known which fixes $\boldsymbol{\Delta}$ at the median-elicited value used to generate the data; and $\boldsymbol{\Delta}$-prior which uses the elicited prior distribution for $\boldsymbol{\Delta}$. As a benchmark estimator, we use the moment estimates based on the full data of all 104 patients. In real data situations, we do not have access to this full data estimator, but we use it as an ideal case comparison. Table A.3 provide the mean point estimate, the bias, and the mean squared error (MSE) for each of the quantities of interest. and Figure A.2 provide box plots of the parameter estimates.

Estimator	$P(Q_{iT}=1)$			$\mathrm{E}(W_{iT})$			$\mathrm{corr}(Q_{iT}, W_{iT})$		
	Estimate	Bias	MSE	Estimate	Bias	MSE	Estimate	Bias	MSE
Completer	0.497	0.019	4.8×10^{-3}	2.44	0.17	18.1×10^{-2}	0.130	-0.010	20.0×10^{-3}
MAR	0.496	0.018	4.2×10^{-3}	2.37	0.10	13.4×10^{-2}	0.106	-0.034	13.8×10^{-3}
$\boldsymbol{\Delta}$-known	0.481	0.003	3.9×10^{-3}	2.20	-0.06	12.7×10^{-2}	0.109	-0.031	13.4×10^{-3}
$\boldsymbol{\Delta}$-prior	0.481	0.002	3.9×10^{-3}	2.28	0.02	12.4×10^{-2}	0.103	-0.037	14.0×10^{-3}
Full Data	0.481	0.003	2.0×10^{-3}	2.28	0.02	8.0×10^{-2}	0.135	-0.005	8.7×10^{-3}

Table A.3: Parameter estimation in the simulation study.

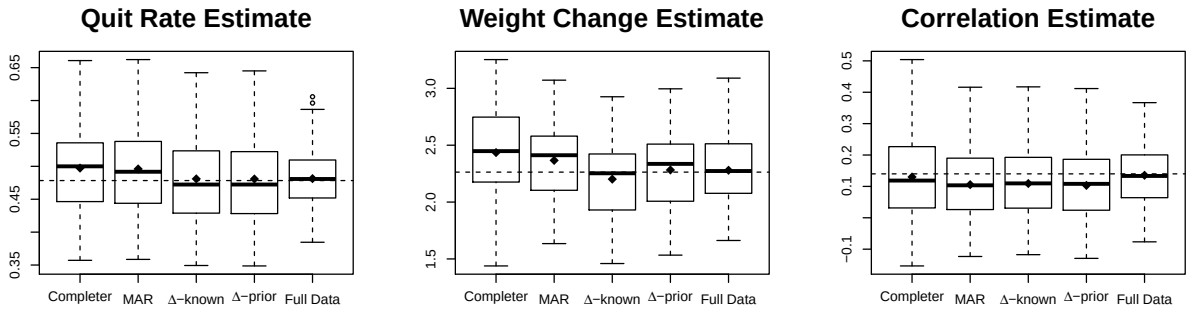

Figure A.2: Box plots of parameter estimates under each of the five estimation methods. The true parameter value is given by the dashed line.

As expected, estimates based on the full data do best, but the full data is not available in practice. We find that our model-based methods do a very good job of estimating quit rate and mean weight change when there is informative missingness. The quit rate estimates under MAR and complete case analysis are biased high by around 2 percentage points, whereas our MNAR estimates are essentially unbiased and reduce mean squared error by around 10% over MAR and 20% over completer-only. Similarly, MAR and completer-only overstate the expected weight gain, and our MNAR model provides the necessary correction reducing bias and MSE. While the model-based correlation estimates appear to exhibit a minor downward bias (probably due to the sparse GARP prior favoring low correlation), MSE show good performance relative to full data, indicating we are trading minor bias for stability (low variance) in the estimate.

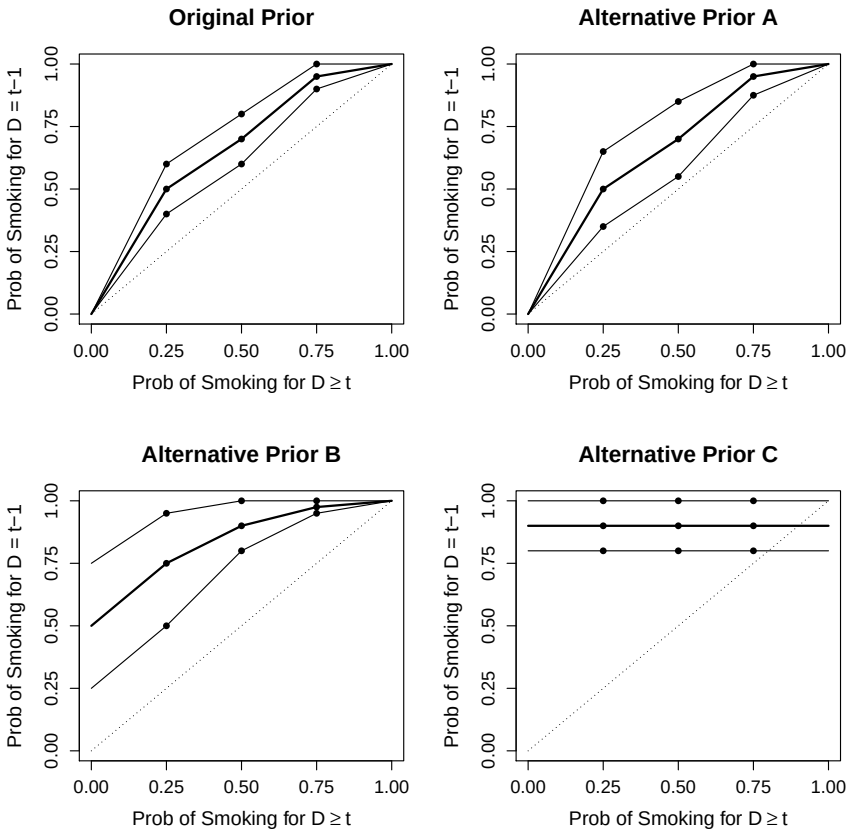

Figure A.3: Prior distribution of $P\left(Q_{it} = 0|\overline{\mathbf{Y}}_{it}, D = t-1\right)$ given $P\left(Q_{it} = 0|\overline{\mathbf{Y}}_{it}, D \geq t\right)$ for each choice of Δ_1 prior. The bold line represents the median value, the solid lines the lower and upper bounds, and the dotted line is $P\left(Q_{it} = 0|\overline{\mathbf{Y}}_{it}, D \geq t\right)$.

A.7. Further Analyses Using Alternative Missingness Assumptions

In addition to the results described in the main article, we discuss further sensitivity analyses to our assumptions about the missingness here. All of our beliefs about the behavior of the patients after they drop out are determined by the the distributional assumptions for $\mathbf{\Delta}$ and the NFD assumptions. First, we consider the sensitivity of these results to varying choices in the distribution of the sensitivity parameter. Figure A.3 displays the original distributional choice for Δ_1 as determined by the subject matter expert (Table 3, Figure 2) along with three new, alternative choices. To align

with the investigator's original belief, we continue to use the same sensitivity prior for Δ_1 for both control and exercise treatments.

- The alternative prior A considers a more dispersed version of the original prior. The minimum and maximum values for $\tilde{p}$ are adjusted to be twice as far from the median as the original choice (up to boundary conditions). Similarly, we adjust the prior on the weight change sensitivity parameter Δ_2 to have twice the width of the original choice. As a referee pointed out, it is commonly the case that investigator-derived priors are overly confident, so adaptations that increase the variance should be considered in the sensitivity analysis phase.

- Alternative prior B was chosen to provide a stronger assumption about the patients who have dropped out. Under this choice, they are more likely to be smoking than in the original prior and much more likely than under MAR (dotted line). We assume the weight change is ignorable and fix $\Delta_2 = 0$.

- Alternative prior C assumes that the probability a patient is smoking after dropout does not depend on how likely a patient under observation with common history is to be smoking. The probability of smoking is uniformly distributed between 80% and 100%. Although there is no longer the connection between $\tilde{p}$ and $\hat{p}$, Δ_1 is still found by solving (15). Weight change is ignorable ($\Delta_2 = 0$).

Table A.4 displays estimates of the quantities of interest under MAR, our original MNAR analysis, and each of these alternative choices.

As the alternative priors B and C represent a stronger belief in a higher smoking rate for patients who leave the study, the overall quit rate decreases relative to the original MNAR and MAR analyses. The more dispersed prior A leads to inference that is relatively unchanged from the original prior. As this choice is centered at the same value as the original prior, this is not surprising. For priors B and C which assume weight change is ignorable, we see slightly lower weight change

Quantity of interest	Treatment	Missing data assumption				
		Z_{it} MAR W_{it} MAR	Original Prior W_{it} MNAR	Prior A Dispersed MNAR	Prior B W_{it} MAR	Prior C W_{it} MAR
$P(Q_{iT} = 1)$	Wellness	0.53 (0.40, 0.65)	0.50 (0.37, 0.63)	0.50 (0.37, 0.63)	0.47 (0.34, 0.60)	0.46 (0.34, 0.59)
	Exercise	0.47 (0.35, 0.59)	0.44 (0.32, 0.56)	0.44 (0.32, 0.56)	0.40 (0.29, 0.52)	0.39 (0.28, 0.51)
	Post. prob.	0.25	0.24	0.23	0.22	0.21
$\mathrm{E}(W_{iT})$	Wellness	3.0% (2.3, 3.8)	2.9% (2.2, 3.6)	3.0% (2.2, 3.7)	2.9% (2.2, 3.7)	2.9% (2.2, 3.7)
	Exercise	3.0% (2.3, 3.7)	2.8% (2.0, 3.4)	2.8% (2.1, 3.5)	2.9% (2.2, 3.6)	2.9% (2.2, 3.6)
	Post. prob.	0.51	0.61	0.61	0.50	0.49
$\mathrm{corr}(Q_{iT}, W_{iT})$	Wellness	0.18 (- 0.08, 0.42)	0.18 (- 0.08, 0.42)	0.16 (- 0.10, 0.41)	0.18 (- 0.07, 0.41)	0.18 (- 0.07, 0.40)
	Exercise	0.13 (- 0.14, 0.36)	0.13 (- 0.13, 0.36)	0.12 (- 0.14, 0.35)	0.13 (- 0.13, 0.35)	0.13 (- 0.12, 0.34)
	Post. prob.	0.61	0.59	0.58	0.61	0.63

Table A.4: Posterior mean and 95% credible interval for the quantities of interest under varying assumptions for the distribution of the sensitivity parameters. See Figure A.3 and the bulleted list on the previous page. The posterior probability row gives the probability that the exercise treatment is superior: higher cessation rate, lower weight change, lower correlation.

than under MAR (also Z-MNAR/W-MAR from Table 4) due to the positive correlation between Q and W and the lower quit rate.

In addition to the sensitivity to our prior for $\mathbf{\Delta}$, it is also worthwhile to explore sensitivity to the non-future dependence assumption. As we discuss in Section 6.1, NFD is often intuitively appealing since it assumes that the probability of dropping out at week d is independent of the responses beyond the next week and because it identifies all extrapolation distributions beyond time $d+1$. Recall that under NFD, we assumed that the distribution of the first missed observation is a location-shifted version of the MAR distribution given by $f_{d;t}(\mathbf{y}_{it}|\overline{\mathbf{y}}_{it}) = \sum_{s=t}^{T} \alpha(s, \overline{\mathbf{y}}_{it}) \tilde{f}_{s;t}(\mathbf{y}_{it}|\overline{\mathbf{y}}_{it})$ for $d = t-1$ where $\tilde{f}_{s;t}(\mathbf{y}_{it}|\overline{\mathbf{y}}_{it})$ is the distribution $f_{s;t}(\mathbf{y}_{it}|\overline{\mathbf{y}}_{it})$ after the location shift $\mathbf{\Delta}$. The distributions $f_{d;t}(\mathbf{y}_{it}|\overline{\mathbf{y}}_{it})$ at $t > d+1$ are given by (14). Without NFD we must specify all distributions $f_{d;t}(\mathbf{y}_{it}|\overline{\mathbf{y}}_{it})$ for $d < t$.

As a simple alternative that does not require NFD, we assume a location-shift distribution for all observations after dropout. That is, $f_{d;t}(\mathbf{y}_{it}|\overline{\mathbf{y}}_{it}) = \sum_{s=t}^{T} \alpha(s, \overline{\mathbf{y}}_{it}) \tilde{f}_{s;t}(\mathbf{y}_{it}|\overline{\mathbf{y}}_{it})$ for all $d < t$, not only $d = t-1$. Under this belief, we consider three choices for the distribution on $\mathbf{\Delta}$. We use the original priors on Δ_1 and Δ_2 from Table 3, the original prior on Δ_1 with $\Delta_2 = 0$ (weight change

Treatment	Missing data assumption					
	NFD Original Prior W_{it} MNAR	Without NFD Original Prior W_{it} MNAR	NFD Original Prior W_{it} MAR	Without NFD Original Prior W_{it} MAR	NFD Prior C W_{it} MAR	Without NFD Prior C W_{it} MAR
			Quantity of Interest: $P(Q_{iT} = 1)$			
Wellness	0.50 (0.37, 0.63)	0.40 (0.28, 0.51)	0.50 (0.37, 0.63)	0.39 (0.28, 0.51)	0.46 (0.34, 0.59)	0.36 (0.27, 0.45)
Exercise	0.44 (0.32, 0.56)	0.36 (0.25, 0.47)	0.43 (0.32, 0.56)	0.36 (0.25, 0.47)	0.39 (0.28, 0.51)	0.33 (0.25, 0.43)
Post. prob.	0.24	0.32	0.23	0.31	0.21	0.35
			Quantity of Interest: $\mathrm{E}(W_{iT})$			
Wellness	2.9% (2.2, 3.6)	2.5% (1.7, 3.3)	3.0% (2.2, 3.7)	2.9% (2.1, 3.7)	2.9% (2.2, 3.7)	2.8% (1.9, 3.6)
Exercise	2.8% (2.0, 3.4)	2.0% (1.2, 2.8)	3.0% (2.3, 3.6)	2.8% (2.1, 3.5)	2.9% (2.2, 3.6)	2.8% (2.0, 3.5)
Post. prob.	0.61	0.79	0.51	0.52	0.49	0.49
			Quantity of Interest: $\mathrm{corr}(Q_{iT}, W_{iT})$			
Wellness	0.18 (- 0.08, 0.42)	0.20 (- 0.03, 0.42)	0.18 (- 0.07, 0.42)	0.18 (- 0.06, 0.40)	0.18 (- 0.07, 0.40)	0.18 (- 0.00, 0.34)
Exercise	0.13 (- 0.13, 0.36)	0.22 (- 0.02, 0.42)	0.13 (- 0.14, 0.36)	0.14 (- 0.10, 0.36)	0.13 (- 0.12, 0.34)	0.14 (- 0.07, 0.32)
Post. prob.	0.59	0.45	0.61	0.57	0.63	0.61

Table A.5: Posterior mean and 95% credible interval for the quantities of interest with and without the NFD assumption. The posterior probability row gives the probability that the exercise treatment is superior: higher cessation rate, lower weight change, lower correlation.

is ignorable), and Prior C for Δ_1 with $\Delta_2 = 0$. The estimated quantities are displayed in Table A.5 with their corresponding estimates under the NFD assumption.

Clearly, NFD has a strong impact on estimation, as the models without NFD produce quit rate estimates between 6 and 11 points lower. However, a word of caution is in order. Recall that our sensitivity specification compares a Patient A who is not observed to a Patient B who is observed and has the same history $\overline{\mathbf{y}}_{it}$. For a patient who dropped out at time d, her values for the first missed observation $\mathbf{y}_{i,d+1} = (Z_{i,d+1}, W_{i,d+1})$ will be lower than a patient with the same history $\overline{\mathbf{y}}_{i,d+1}$ ($\boldsymbol{\Delta}$ is mostly negative values). But at time $d + 2$, the new values $\mathbf{y}_{i,d+2}$ are lower than a patient who is observed with the history $(\overline{\mathbf{y}}_{i,d+1}, \mathbf{y}_{i,d+1})$, which is already lower than typical. Hence, for patients who drop out early, the distribution of the (unobserved) value at the final week may be concentrated on values that are unreasonably low. This is seen in the estimates for $\mathrm{E}(W_{iT})$ with the MNAR prior on weight change. In all other analyses, the expected weight change is incredibly stable, but using our original model without NFD we see a dramatic drop. We note that this issue of $\mathbf{Y}_{iT}$ drifting

to unreasonable values may be minimized by the choice of Prior C which does not specify the probability of smoking in terms of a patient with common history.

We discuss this simply as a first step in understanding the role of the NFD assumption. When we are not comfortable assuming NFD, further models may need to be developed to represent a more appropriate mechanism for the behavior of patients after they drop out. In particular, one could imagine specifying a distribution for the first missed response (along the lines of our original prior) and a different model (perhaps, Prior C) for the later missed measurement that will not lead to this drifting. Our focus in the current article is modeling under the NFD assumption, and additional research is needed to move beyond this.

Finally, we return to the partial ignorability assumption. Recall that partial ignorability assumes that the intermittent missing values (missingness prior to dropout) provide no information beyond that obtained from the observed data and the dropout time. In this context, partial ignorability implies that a patient who is missing is no more likely to be smoking that an observed patient who will dropout at the same time d and has the same (past and future) observed values. Most longitudinal models with intermittent missingness that do not use partial ignorability lead to identified parameters in the extrapolation distribution (e.g., Daniels and Hogan, 2008, Section 10.3). We do not wish to explore such models here. As a simple competitor we adjust the data so that all intermittently missed observations are assumed to be smoking. While this is more extreme than is reasonable, this choice will provide the extent to which the intermittently missed values impact inference. Intermittently missed weight change is assumed ignorable. Additionally, we return to analysis from Section 6.3 that assumed $Q_{it} = 0$ for all missed values. Now we set $Q_{it} = 0$ only after dropout to provide a partially ignorable version of this analysis.

The results in Table A.6 indicate that the impact of partial ignorability under MAR, our original MNAR prior, and the $Q = 0$ assumption. We see the impact is minimal, leading to changes in quit rate of around one percentage point. As our assumptions about the behavior after dropout becomes more extreme (as depicted by the $Q = 0$ case), the partial ignorability assumption has a slightly

Treatment	Missing data assumption					
	Part. Ign. Z_{it} MAR W_{it} MAR	Without Part. Ign. Z_{it} MAR W_{it} MAR	Part. Ign. Original Prior W_{it} MNAR	Without Part. Ign. Original Prior W_{it} MNAR	Part. Ign. $Q_{it} = 0$ if missing W_{it} MAR	Without Part. Ign. $Q_{it} = 0$ if missing W_{it} MAR
	Quantity of Interest: $P(Q_{iT} = 1)$					
Wellness	0.53 (0.40, 0.65)	0.52 (0.39, 0.64)	0.50 (0.37, 0.63)	0.49 (0.36, 0.61)	0.34 (0.24, 0.44)	0.30 (0.21, 0.39)
Exercise	0.47 (0.35, 0.59)	0.46 (0.34, 0.59)	0.44 (0.32, 0.56)	0.43 (0.30, 0.55)	0.31 (0.22, 0.41)	0.28 (0.20, 0.37)
Post. prob.	0.25	0.25	0.24	0.24	0.34	0.42
	Quantity of Interest: $\mathrm{E}(W_{iT})$					
Wellness	3.0% (2.3, 3.8)	3.1% (2.3, 3.8)	2.9% (2.2, 3.6)	2.9% (2.2, 3.7)	2.6% (1.9, 3.4)	2.6% (1.8, 3.3)
Exercise	3.0% (2.3, 3.7)	3.0% (2.4, 3.7)	2.8% (2.0, 3.4)	2.7% (2.1, 3.4)	2.7% (2.1, 3.4)	2.7% (2.1, 3.4)
Post. prob.	0.51	0.52	0.61	0.63	0.42	0.38
	Quantity of Interest: $\mathrm{corr}(Q_{iT}, W_{iT})$					
Wellness	0.18 (- 0.08, 0.42)	0.19 (- 0.06, 0.42)	0.18 (- 0.08, 0.42)	0.19 (- 0.05, 0.41)	0.18 (- 0.04, 0.37)	0.21 (0.02, 0.39)
Exercise	0.13 (- 0.14, 0.36)	0.14 (- 0.11, 0.37)	0.13 (- 0.13, 0.36)	0.15 (- 0.09, 0.37)	0.16 (- 0.04, 0.35)	0.17 (- 0.04, 0.35)
Post. prob.	0.61	0.61	0.59	0.59	0.55	0.62

Table A.6: Posterior mean and 95% credible interval for the quantities of interest with and without the partial ignorability assumption. The posterior probability row gives the probability that the exercise treatment is superior: higher cessation rate, lower weight change, lower correlation.

greater impact, reducing the quit rate by 4 and 3 points, respectively.

In conclusion, we find that the results as presented in the main article are robust to the partial ignorability assumption (Table A.6), while the choice of the distribution of the sensitivity parameter $\boldsymbol{\Delta}$ has a moderate impact on inference (Table A.4). Of the three assumption we consider, the choice of non-future dependence leads to larger changes in the quit rates (Table A.5). The decision to assume NFD must be made by balancing the appropriateness of the mathematical simplification and intuition about the MDM against the difficulties of specifying the unidentified extrapolation distribution.

References

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. (2011). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. CRC Press.

Daniels, M. J., Chatterjee, A., and Wang, C. (2012). Bayesian model selection for incomplete data using the posterior predictive distribution. *Biometrics*, 68:1055–1063.

Daniels, M. J. and Hogan, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Chapman & Hall.

Gaskins, J. T., Daniels, M. J., and Marcus, B. H. (2014). Sparsity inducing prior distributions for correlation matrices of longitudinal data. *Journal of Computational and Graphical Statistics*, 23(4):966–984.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Harel, O. and Schafer, J. L. (2009). Partial and latent ignorability in missing-data problems. *Biometrika*, 96(1):37–50.

Ibrahim, J. G. and Laud, P. W. (1994). A predictive approach to the analysis of designed experiments. *Journal of the American Statistical Association*, 89(425):309–319.

Liu, X., Daniels, M. J., and Marcus, B. (2009). Joint models for the association of longitudinal binary and continuous processes with application to a smoking cessation trial. *Journal of the American Statistical Association*, 104(486):429–438.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3):705–767.

Savalei, V. (2006). Logistic approximation to the normal: The KL rationale. *Psychometrika*, 71(4):763–767.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64(4):583–639.