Biometrics 000, 000–000 DOI: 000
000 0000

Fully Bayesian inference under ignorable missingness in the presence of auxiliary covariates

M.J. Daniels [1], C. Wang [2], B.H. Marcus [3]

[1]Division of Statistics & Scientific Computation and Section of Integrative Biology, University of Texas at Austin, Austin, TX 78712
[2]The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD 21287
[3]Department of Family and Preventive Medicine, University of California, San Diego, La Jolla, CA 92093

This paper has been submitted for consideration for publication in *Biometrics*

Abstract In order to make a missing at random (MAR) or ignorability assumption realistic, auxiliary covariates are often required. However, the auxiliary covariates are not desired in the model for inference. Typical multiple imputation approaches do not assume that the imputation model marginalizes to the inference model. This has been termed 'uncongenial' (Meng, 1994). In order to make the two models congenial (or compatible), we would rather not assume a parametric model for the marginal distribution of the auxiliary covariates, but we typically do not have enough data to estimate the joint distribution well non-parametrically. In addition, when the imputation model uses a non-linear link function (e.g., the logistic link for a binary response), the marginalization over the auxiliary covariates to derive the inference model typically results in a difficult to interpret form for effect of covariates. In this article, we propose a fully Bayesian approach to ensure that the models are compatible for incomplete longitudinal data by embedding an interpretable inference model within an imputation model and that also addresses the two complications described above. We evaluate the approach via simulations and implement it on a recent clinical trial.

Key Words: Congenial imputation; Multiple imputation; Marginalized models; Auxiliary variable MAR.

1. Introduction

In clinical studies, investigators often have a primary research question (with an associated model). To make inference in the presence of incomplete data, a common approach is to use multiple imputation (Lavori et al., 1995; Burns et al., 2011). To do multiple imputation and make an assumption of missing at random more realistic, it is not uncommon to use additional information (e.g., auxiliary covariates, $\boldsymbol{v}$) that are not desired in the model for the primary research question (particularly, for randomized studies). We denote as $\boldsymbol{x}$ the covariates desired in the model for the primary research questions; we will call the model $p(\boldsymbol{y}|\boldsymbol{x})$ the *inference model*. In recent years, it has become common practice for an investigator to build both the *imputation model* (to do multiple imputation), $p(\boldsymbol{y}|\boldsymbol{v},\boldsymbol{x})$, which contains both auxiliary covariates, $\boldsymbol{v}$, and model covariates, $\boldsymbol{x}$ and the *inference model*, $p(\boldsymbol{y}|\boldsymbol{x})$, which only contains model covariates, $\boldsymbol{x}$.

There is a long literature on multiple imputation starting with Rubin (Rubin, 1978; Little and Rubin, 1987) in the context of surveys. The two most common approaches for the imputation model include a joint multivariate distribution for all missing variables (Rubin and Schafer, 1990; Liu, 1995; Schafer, 1997) or specification of a series of conditional models for each variable one at a time (Van Buuren et al., 1999; Raghunathan et al., 2001; Gelman and Raghunathan, 2001). The latter, typically does not correspond to a valid joint distribution for the variables to be imputed (Gelman and Raghunathan, 2001). For a recent review of multiple imputation, see Kenward and Carpenter (2007).

To understand the relevant issues more clearly, we will introduce some additional notation. Define the full data longitudinal response as $\boldsymbol{y}$ and the full data as $(\boldsymbol{y},\boldsymbol{r})$, where $\boldsymbol{r}$ are indicators informing which components of $\boldsymbol{y}$ are observed. The observed data is $(\boldsymbol{y}_{\text{obs}},\boldsymbol{r})$ and the missing data is $\boldsymbol{y}_{\text{mis}}$. We define the missing data mechanism (mdm) as $p(\boldsymbol{r}|\boldsymbol{y})$ (suppressing dependence on any covariates for now). The full data model, with parameters $\boldsymbol{\omega}$, is defined as $p(\boldsymbol{y},\boldsymbol{r}|\boldsymbol{\omega})$. The full data response model is $p(\boldsymbol{y}|\boldsymbol{\omega})$; this is obtained from the full data model after marginalizing over $\boldsymbol{r}$.

We define missingness to be *ignorable* (Rubin, 1976) if the following three conditions hold:

(1) The missing data mechanism is MAR (i.e., $p(\boldsymbol{r}|\boldsymbol{y}) = p(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}})$)

(2) The full data parameter $\boldsymbol{\omega}$ can be decomposed as $\boldsymbol{\omega} = (\boldsymbol{\beta}, \boldsymbol{\psi})$, where

- $\boldsymbol{\beta}$ indexes the full-data response model $p(\boldsymbol{y} \mid \boldsymbol{\beta})$, and
- $\boldsymbol{\psi}$ indexes the missing data mechanism $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{\psi})$.

(3) The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$ are a-priori independent; i.e.,

$$p(\boldsymbol{\beta}, \boldsymbol{\psi}) = p(\boldsymbol{\beta})p(\boldsymbol{\psi}).$$

A major advantage of ignorability is that there is no need to specify (explicitly) the form of the mdm.

However, MAR may not hold when conditioning on observed data response and covariates of interest for inference, $\boldsymbol{x}$, but may hold if we also condition on auxiliary covariates. That is,

$$p(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \boldsymbol{x}) \neq p(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{x}) \text{ but}$$

$$p(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}, \boldsymbol{x}, \boldsymbol{v}) = p(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{x}, \boldsymbol{v}).$$

The latter condition has been called auxiliary variable MAR (A-MAR) (Daniels and Hogan, 2008).

As stated above, the advantage of incorporating $\boldsymbol{v}$ to yield A-MAR is that the mdm $p(\boldsymbol{r} \mid \boldsymbol{y}, \boldsymbol{v})$ can be ignored, under slightly modified conditions (given in Section 2). However, even though auxiliary covariates are frequently used in the imputation model, it is typically the case that

$$p(\boldsymbol{y}|\boldsymbol{x}) \neq \int p^{\star}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{v})p(\boldsymbol{v}|\boldsymbol{x})d\boldsymbol{v}.$$

When the imputation model, $p^{\star}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{v})$ is not chosen to match the inference model (i.e., the above equality does not hold), the imputation model has been termed 'uncongenial' (Meng, 1994). Such an approach is *not* principled and has no mathematical justification in terms of compatible probability models and Bayesian inference, particularly, in situations where the same research group specifies both models.

Here, we propose a simple fully Bayesian modeling framework such that

$$p(\boldsymbol{y}|\boldsymbol{x}) = \int p^{\star}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{v})p(\boldsymbol{v}|\boldsymbol{x})d\boldsymbol{v}. \quad (1)$$

i.e., the imputation model, $p^{\star}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{v})$ does marginalize to the inference model. In what follows, we specify the imputation model for which (1) holds as $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{v})$. Our approach does not require two separate model fits and explicit multiple imputation. We will formulate a full Bayesian imputation model that has the desired inference model embedded within it.

Unfortunately, two practical complications arise in implementing such an approach:

(1) Specification of $p(\boldsymbol{v}|\boldsymbol{x})$: This distribution is (basically) a nuisance parameter. As such, we would prefer to not have incorrect inferences from the inference model due to a potentially mis-specified parametric model for $p(\boldsymbol{v}|\boldsymbol{x})$, but we do not typically have enough data to construct an efficient nonparametric estimate of $p(\boldsymbol{v}|\boldsymbol{x})$.

(2) Analytical form of $E[\boldsymbol{Y}|\boldsymbol{x}]$: Deriving the full data response model, $p(\boldsymbol{y}|\boldsymbol{x})$ (inference model) from the imputation model, $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{v})$ often results in a messy, hard to interpret form for inference on the covariates in $E[\boldsymbol{Y}|\boldsymbol{x}]$, especially for generalized linear models with a non-identity link function.

For the first complication, since the goal of including auxiliary covariates is to make MAR hold, it is advisable to include a rich set of auxiliary covariates. Similar ideas are also applied for multiple imputation in general (Kenward and Carpenter, 2007). To specify a fully non-parametric $p(\boldsymbol{v}|\boldsymbol{x})$ is thus even more challenging when the dimension of $\boldsymbol{v}$ is not small. Of course, there is a trade-off between ensuring that MAR holds and avoiding multicollinearity and computational issues in the imputation model.

The second complication can be seen in the simple setting of a cross-sectional binary response, $\boldsymbol{Y}$. Consider the following glm for the imputation model

$$g(E[\boldsymbol{Y}|\boldsymbol{x},\boldsymbol{v}]) = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{v}\boldsymbol{\alpha}.$$

The (induced) mean regression for the inference model is

$$E[\boldsymbol{Y}|\boldsymbol{x}] = \int g^{-1}(\boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{v}\boldsymbol{\alpha})p(\boldsymbol{v}|\boldsymbol{x})d\boldsymbol{v}.$$

If $g(\cdot)$ is not the identify function, $E[\boldsymbol{Y}|\boldsymbol{x}]$ will typically not be available in closed form and will result in an inconvenient and complex interpretation of the regression effects of inferential covariates $\boldsymbol{x}$. For example, it may not be possible to interpret the effect of an individual covariate conditional on a fixed value for the other covariates in the model.

In this paper, we propose an approach to ensure that (1) holds but that also addresses these two complications. We will rely on shrinkage (Wang et al., 2010) for complication 1 and marginalized models (Heagerty, 2002; Roy and Daniels, 2008) for complication 2. In Section 2, we describe our general approach under A-MAR (for which the missingness need not be monotone), discuss model characteristics and properties, and discuss posterior sampling strategies. In Section 3, we explore the model performance through a simulation study. Section 4 implements the proposed approach for the analysis of a recent smoking cessation clinical trial. Section 5 concludes with a wrap-up and discussion.

2. Approach

2.1 *Auxiliary variable ignorability*

We define missingness to be *ignorable* in the presence of auxiliary covariates if the following three conditions hold:

(1) The missing data mechanism is A-MAR (i.e., $p(\boldsymbol{r}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{v};\boldsymbol{\psi}) = p(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}},\boldsymbol{x},\boldsymbol{v};\boldsymbol{\psi})$)

(2) The full data parameter $\boldsymbol{\omega}$ can be decomposed as $\boldsymbol{\omega} = (\boldsymbol{\alpha},\boldsymbol{\psi},\boldsymbol{\theta})$, where

- $\boldsymbol{\alpha}$ indexes the full-data response model conditional on auxiliary covariates $p(\boldsymbol{y} \mid \boldsymbol{x},\boldsymbol{v};\boldsymbol{\alpha})$,
- $\boldsymbol{\psi}$ indexes the missing data mechanism $p(\boldsymbol{r} \mid \boldsymbol{y},\boldsymbol{x},\boldsymbol{v};\boldsymbol{\psi})$, and
- $\boldsymbol{\theta}$ indexes the marginal distribution of the auxiliary covariates $p(\boldsymbol{v} \mid \boldsymbol{x};\boldsymbol{\theta})$,

(3) The two sets of parameters $(\boldsymbol{\alpha},\boldsymbol{\theta})$ and $\boldsymbol{\psi}$ are a-priori independent; i.e.,

$$p(\boldsymbol{\alpha},\boldsymbol{\theta},\boldsymbol{\psi}) = p(\boldsymbol{\alpha},\boldsymbol{\theta})p(\boldsymbol{\psi}).$$

2.2 *Specification of compatible imputation and inference models*

For our framework, we extend and modify recent models (Heagerty, 2002; Roy and Daniels, 2008) to the context of longitudinal data with ignorable (under A-MAR) missingness.

Denote Y_{it} as a binary observation of subject i at time t. Denote $\boldsymbol{X}^*_{it}$ as covariates of interest for subject i observed at time t; these include baseline covariates, $\boldsymbol{X}_i$ (e.g., treatment) and functions of time along with their potential interactions with $\boldsymbol{X}_i$. The marginal mean is specified as

$$\text{logit}\{E\left(Y_{it}|\boldsymbol{x}^\star_{it};\boldsymbol{\beta}\right)\} = \boldsymbol{x}^{\star T}_{it}\boldsymbol{\beta}. \tag{2}$$

This is the *inference model* . The parameters $\boldsymbol{\beta}$ are of primary interest and are a function of $(\boldsymbol{\alpha}, \boldsymbol{\theta})$ from the definition of A-MAR ignorability.

The *imputation model* conditions on previous $\boldsymbol{Y}$'s, on baseline inference covariates, $\boldsymbol{X}_i$ and baseline auxiliary covariates, $\boldsymbol{V}_i$,

$$\text{logit}\left\{E\left(Y_{it}|\boldsymbol{V}_i, \boldsymbol{X}_i, Y_{i,t-1};\boldsymbol{\alpha},\boldsymbol{\gamma}\right)\right\} = \Delta_{it} + g(\boldsymbol{V}_i, \boldsymbol{X}_i;\boldsymbol{\alpha}) + \boldsymbol{\gamma}^\star_{it}y_{i,t-1}, \tag{3}$$

where $\boldsymbol{\gamma}^\star_{it} = \boldsymbol{Z}_{it}\boldsymbol{\gamma}$, where $\boldsymbol{Z}_{it} \in \boldsymbol{X}^\star_{it}$; if we set $Z_{it} = 1$, there is a first order dependence that does not depend on covariates or time. Δ_{it} is a subject specific intercept at each time that ensures the imputation and inference models match; more detail can be found below. An example of the function $g(\boldsymbol{V}_i, \boldsymbol{X}_i;\boldsymbol{\alpha})$ would be $\boldsymbol{V}_i^T\boldsymbol{\alpha}_1 + (\boldsymbol{V}_i * \boldsymbol{X}_i)^T\boldsymbol{\alpha}_2$. Note this model also accounts for longitudinal dependence via the Markov term, $\boldsymbol{\gamma}_{it}y_{i,t-1}$. The parameters $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ represent the dependence of Y_{it} on $\boldsymbol{V}_i$ and are generally *not* of primary interest.

The subject specific intercept at time t, Δ_{it} is determined by the following equality,

$$E\left(Y_{it}|\boldsymbol{x}_i;\boldsymbol{\beta}\right) = \int E\left(Y_{it}|\boldsymbol{v},\boldsymbol{x}_i;\boldsymbol{\alpha},\boldsymbol{\gamma}\right)p(\boldsymbol{v}|\boldsymbol{x}_i)d\boldsymbol{v}, \tag{4}$$

where

$$E\left(Y_{it}|\boldsymbol{v},\boldsymbol{x}_i;\boldsymbol{\alpha},\boldsymbol{\gamma}\right) = \sum_{y=0}^{1} E\left(Y_{it}|\boldsymbol{v},\boldsymbol{x}_i,Y_{i,t-1}=y;\boldsymbol{\alpha},\boldsymbol{\gamma}\right)p\left(Y_{i,t-1}=y|\boldsymbol{v},\boldsymbol{x}_i;\boldsymbol{\alpha},\boldsymbol{\gamma}\right). \tag{5}$$

(5) is the same as the marginalization relation from Heagerty (2002). So given $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$, we can solve for Δ_{it}; in the MCMC algorithm in Section 2.4, we update Δ_{it} each time we update any of these parameters. The condition in (4) is such that (1) holds and the marginal mean,

$E(\boldsymbol{Y}|\boldsymbol{x})$ has a directly specified and interpretable form. To specify priors when there is a lack of prior information, we use diffuse normal priors, $N(0, \boldsymbol{A})$ for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$; here $\boldsymbol{A} = 10^3 I$.

In the development here, we assume the inference and auxiliary covariates are fully observed. Auxiliary covariates that are MAR could easily be addressed by adding an augmentation step to the posterior computations described in Section 2.4.

In terms of the modelling, the remaining component is specification of the joint distribution of the auxiliary covariates, $p(\boldsymbol{v}|\boldsymbol{x})$. We detail this in the next section.

2.3 *Specification of, and priors for, the distribution of the auxiliary covariates*

In the following, we focus on the case where all components of $\boldsymbol{v}$ are categorical. In situations where there are continuous covariates, we assume there is a natural discretization (e.g., see the example in Section 4). In a setting of a randomized trial, $\boldsymbol{x}$ is often just a treatment indicator. As such, modeling can be done separately for each treatment (i.e., each value of $\boldsymbol{x}$) or be done by assuming $p(\boldsymbol{v}|\boldsymbol{x}) \equiv p(\boldsymbol{v})$ (e.g. for randomized trials). As such, for ease of notation, we will not condition on $\boldsymbol{x}$ in what follows. We assume $\boldsymbol{v}$ has p components and the jth component has L_j levels/categories. Given that we do not want to impose strong parametric assumptions on the joint distribution of $\boldsymbol{v}$, we start out with a saturated model, which here corresponds to a multinomial with $\prod_{j=1}^{p} L_j$ categories and $(\prod_{j=1}^{p} L_j) - 1$ parameters, $\boldsymbol{\theta}$. In most cases, the data will be too sparse to estimate the probability of each category well. To overcome this obstacle, we will shrink the saturated model to a simpler model in a computationally efficient way (see, e.g., Wang et al., 2010). In particular, we assume a Dirichlet distribution on $\boldsymbol{\theta}$ with precision parameter η and prior expectation of a simple form. For example, here we specify the prior expectation of the joint distribution of $\boldsymbol{v}$ as the product of the marginals, $p(\boldsymbol{v}) = p(v_1)p(v_2)\cdots p(v_p)$, resulting in $\sum_j (L_j - 1)$ parameters in the Dirichlet prior. We denote the set of expectation parameters as $\boldsymbol{\pi}$ and the full set of parameters as $\boldsymbol{\theta}^\star = (\boldsymbol{\theta}, \boldsymbol{\pi}, \eta)$. We provide more modelling details next.

Let $\underline{l} = \{l_1, l_2, \ldots, l_p\}$ denote a realization of $\boldsymbol{V}$ with each of the p categorical covariates being $l_j \in \{1, \ldots, L_j\}$ for all j. Let $\boldsymbol{N} = \{N_{\underline{l}}; \forall \underline{l}\}$ be a realization of the entire vector of

auxiliary covariates, $\boldsymbol{V}$, with $N_{\underline{l}}$ the number of subjects with $\boldsymbol{V} = \underline{l}$. Let $\theta_{\underline{l}}$ be $P(\boldsymbol{V} = \underline{l})$ and $\boldsymbol{\theta} = \{\theta_{\underline{l}}; \forall \underline{l}\}$ is thus the collection of $\theta_{\underline{l}}$. $[\boldsymbol{N}|\boldsymbol{\theta}]$ follows a (saturated) multinomial distribution.

We assign a shrinkage prior on $\boldsymbol{\theta}$ as follows. First, we assume that $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{a})$, with $\boldsymbol{a} = \{a_{\underline{l}} : \forall \underline{l}\}$ a function of $(\boldsymbol{\pi}, \eta)$ as follows

$$f(\boldsymbol{\theta}) \propto \prod_{l} \theta_{\underline{l}}^{a_{\underline{l}}-1}.$$

where

$$a_{\underline{l}} = \frac{1}{\eta} \prod_{j=1}^{p} \pi_{j,l_j}$$

with $\sum_{k=1}^{L_j} \pi_{j,k} = 1$ for all j; thus, η and $\boldsymbol{\pi}$ are the hyperparameters of this Dirichlet prior. This prior has as expectation the product of the marginal probabilities of the p components of $\boldsymbol{V}$.

The full model specification is then given as

$$\boldsymbol{N}|\boldsymbol{\theta} \sim \text{Multinomial}(\text{number of subjects}; \boldsymbol{\theta})$$
$$\theta_{\underline{l}}|\pi_1, \ldots, \pi_p, \eta \sim \text{Dir}(\{\frac{1}{\eta} \prod_{j=1}^{p} \pi_{j,l_j}\})$$

For each realization of $\boldsymbol{V}$, $\underline{l}$, the prior expectation and variance of $\theta_{\underline{l}} = P(\boldsymbol{V} = \underline{l})$ are given by

$$E(\theta_{\underline{l}}|\pi_1, \ldots, \pi_p, \eta) = \prod_{j=1}^{p} \pi_{j,l_j}$$

and

$$\text{Var}(\theta_{\underline{l}}|\pi_1, \ldots, \pi_p, \eta) = \frac{\eta}{\eta + 1} \prod_{j=1}^{p} \pi_{j,l_j}(1 - \prod_{j=1}^{p} \pi_{j,l_j}).$$

As $\eta \to 0$, the variance goes to zero. Hence, η is a shrinkage parameter and controls the amount of shrinkage (toward marginal independence of the categorical covariates); when $\eta = 0$, there is complete shrinkage toward the mean of the Dirichlet prior (which corresponds to marginal independence).

For the hyperparameters of the Dirichlet prior, we assign a $\text{Dir}(\mathbf{1})$ as a hyperprior on

$\pi_j = \{\pi_{j,1}, \ldots, \pi_{j,L_j}\}$ for all j and a uniform shrinkage prior on η (Daniels, 1999). That is,

$$\pi_j \sim \text{Dir}(\mathbf{1}_{1\times L_j}) \quad \forall j$$
$$\eta \sim p(\eta) = \frac{\sum_{\underline{l}} N_{\underline{l}}}{(\eta \sum_{\underline{l}} N_{\underline{l}} + 1)^2}.$$

The derivation of the prior for η is given in Web Appendix A. The uniform shrinkage prior for η has several desirable properties. It is a proper prior and it is flat on the shrinkage factor (see Web Appendix A) and thus can be viewed as noninformative. The prior median is $\frac{1}{\sum N_{\underline{l}}}$. Thus there is less shrinkage as the sample size increases.

2.4 *Posterior Computations*

We now provide some details on posterior computations.

Likelihood: Define the entire parameter vector as $\boldsymbol{\omega} = (\boldsymbol{\omega}^\star, \boldsymbol{\theta}^\star)$, where $\boldsymbol{\omega}^\star = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ and $\boldsymbol{\theta}^\star = (\boldsymbol{\theta}, \boldsymbol{\pi}, \eta)$. The likelihood is given by

$$\begin{aligned} L\left(\boldsymbol{\omega}|\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n, \boldsymbol{x}, \boldsymbol{v}\right) &= \prod_i p\left(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n|\boldsymbol{v}, \boldsymbol{x}\right) p(\boldsymbol{v}|\boldsymbol{x}) \\ &= \prod_i \prod_{t=2}^{n_i} \{p\left(y_{it}|\boldsymbol{v}_i, \boldsymbol{x}^\star_{it}, y_{i,t-1}\right)\} p\left(y_{i1}|\boldsymbol{v}_i, \boldsymbol{x}^\star_{i1}\right) p\left(\boldsymbol{v}_i|\boldsymbol{x}_i\right), \end{aligned}$$

where n_i is the number observations for subject i (assuming monotone missingness); note for intermittent missingness, the observed data likelihood would appropriately average over the missing responses. The posterior distribution of interest is $p(\boldsymbol{\omega}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{v}, \boldsymbol{x})$ can be factored as

$$p(\boldsymbol{\omega}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{v}, \boldsymbol{x}) = p(\boldsymbol{\omega}^\star|\boldsymbol{y}_{\text{obs}}, \boldsymbol{v}, \boldsymbol{x}) p(\boldsymbol{\theta}^\star|\boldsymbol{v}, \boldsymbol{x})$$

since $p(\boldsymbol{\theta}^\star|\boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}_{\text{obs}}) = p(\boldsymbol{\theta}^\star|\boldsymbol{v}, \boldsymbol{x})$.

We follow the steps below at each iteration for posterior sampling:

(1) Use slice sampling to sample each π_{j,l_j} ($j = 1, \ldots, p$, $l_j = 1, \ldots, L_j$) of $\boldsymbol{\pi}$ from $p(\boldsymbol{\pi}|\eta, \boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{x})$ and sample η from $p(\eta|\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{v}, \boldsymbol{x})$.

(2) Sample $\boldsymbol{\theta}$ from a Dirichlet distribution $p(\boldsymbol{\theta}|\boldsymbol{\pi}, \eta, \boldsymbol{v}, \boldsymbol{x})$.

(3) Use Gibbs sampling with the Metropolis-Hastings steps for sampling $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ from $p(\boldsymbol{\beta}|\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{y}_{\text{obs}})$, $p(\boldsymbol{\alpha}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{y}_{\text{obs}})$, and $p(\boldsymbol{\gamma}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{y}_{\text{obs}})$, respectively. To do this, for each

parameter update, we use the Newton method to solve for Δ_{it} in (3) for all i and all t in order to compute the observed data likelihood $p(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{v}, \boldsymbol{x})$.

See further details in Web Appendix B.

Note that our approach does not explicitly require multiple imputation but only requires one posterior based on the proposed model. In our data example, we only have dropout (monotone missingness) thus we do not need data augmentation. If we had intermittent missingness, we would fill in those responses at each iteration using data augmentation. See Web Appendix B for specifics.

2.5 *Properties*

The specification of the *inference* (marginal mean) model is usually determined by the research objectives. Misspecification of the inference model is generally not a major issue. However, there is a less clear understanding about the impact of covariates on the missing data mechanism, especially when the number of the covariates is large. As a principle, we may at times include extra covariates in the imputation model to make it more likely that A-MAR holds.

The proposed modeling approach has the following properties if extra un-needed auxiliary covariates are included in the imputation model. For each, we assume *all necessary* auxiliary covariates are included in the model.

- Property 1: The interpretation of $\boldsymbol{\beta}$ is invariant to unnecessary auxiliary covariates.
- Property 2: The posterior distribution of $\boldsymbol{\beta}$ is consistent in the presence of unnecessary auxiliary covariates.

In the following, we denote as $\boldsymbol{v}$ the necessary auxiliary covariates and $\boldsymbol{v}^\star$ as the unnecessary auxiliary covariates. We provide a heuristic argument below for why these two properties hold.

Property 1 can be understood via the following expression:

$$\begin{aligned} E(\boldsymbol{Y}|\boldsymbol{x}) &= \int\int \boldsymbol{y} p(\boldsymbol{y}|\boldsymbol{v},\boldsymbol{x}) p(\boldsymbol{v}|\boldsymbol{x}) d\boldsymbol{v} d\boldsymbol{y} \\ &= \int\int \boldsymbol{y} p(\boldsymbol{y}|\boldsymbol{v},\boldsymbol{v}^\star,\boldsymbol{x}) p(\boldsymbol{v}^\star|\boldsymbol{v},\boldsymbol{x}) p(\boldsymbol{v}|\boldsymbol{x}) d\boldsymbol{v}^\star d\boldsymbol{v} d\boldsymbol{y}. \end{aligned}$$

That is, the marginal mean of $\boldsymbol{Y}$ (and its interpretation) is invariant to the inclusion of $\boldsymbol{v}^\star$.

Property 2 can be understood by noting that the observed data (i.e., $\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}$) distribution conditional on needed auxiliary covariates, $\boldsymbol{v}$ is the same as that when including un-needed covariates, $\boldsymbol{v}^\star$

$$\begin{aligned} p(\boldsymbol{y}_{\text{obs}}, \boldsymbol{r}|\boldsymbol{x},\boldsymbol{v}) &= \int p(\boldsymbol{y},\boldsymbol{r}|\boldsymbol{x},\boldsymbol{v}) d\boldsymbol{y}_{\text{mis}} \\ &= p(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}},\boldsymbol{x},\boldsymbol{v}) \int p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{v}) d\boldsymbol{y}_{\text{mis}} \\ &= p(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}},\boldsymbol{x},\boldsymbol{v}) \int p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{v},\boldsymbol{v}^\star) p(\boldsymbol{v}^\star|\boldsymbol{v},\boldsymbol{x}) d\boldsymbol{v}^\star d\boldsymbol{y}_{\text{mis}} \\ &= p(\boldsymbol{r}|\boldsymbol{y}_{\text{obs}},\boldsymbol{x},\boldsymbol{v},\boldsymbol{v}^\star) \int p(\boldsymbol{y}_{\text{obs}}|\boldsymbol{x},\boldsymbol{v},\boldsymbol{v}^\star) p(\boldsymbol{v}^\star|\boldsymbol{x},\boldsymbol{v}) d\boldsymbol{v}^\star \\ &= \int p(\boldsymbol{y}_{\text{obs}},\boldsymbol{r}|\boldsymbol{x},\boldsymbol{v},\boldsymbol{v}^\star) p(\boldsymbol{v}^\star|\boldsymbol{x},\boldsymbol{v}) d\boldsymbol{v}^\star. \end{aligned}$$

The second to last equality holds since $\boldsymbol{v}^\star$ are unnecessary auxiliary covariates and as such they are not present in the missing data mechanism once we condition on $\boldsymbol{y}_{\text{obs}}$, $\boldsymbol{x}$, and $\boldsymbol{v}$. Thus if the estimators of the mean parameters, $\boldsymbol{\beta}$ are consistent without unnecessary auxiliary covariates, they will also be consistent with unnecessary auxiliary covariates based on the expressions above and Property 1.

3. Simulations

We conduct simulations to better understand the finite sample properties of our approach. In particular, we design the simulations to examine three scenarios.

- Situation where the MNAR coefficient in the mdm is much smaller (or zero) when the appropriate auxiliary covariates, $\boldsymbol{v}$ are included.
- Comparison of shrinkage prior for the distribution of $p(\boldsymbol{v}|\boldsymbol{x})$ to a non-informative prior.

- Robustness of marginal mean parameters, $\boldsymbol{\beta}$ to inclusion of $\boldsymbol{V}$'s that are not necessary for A-MAR (as discussed in Section 2.5).

We provide details on the setup next.

3.1 *Simulation Setup*

Auxiliary Covariates

To simulate the auxiliary covariates, we draw samples from

$$\boldsymbol{V}^*_{p\times 1} \sim \mathrm{N}(0_{p\times 1}, \boldsymbol{\Sigma}_{p\times p}),$$

where $p = 8$, $\mathrm{Var}(V_j^*) = 1$ for $j = 1, \ldots, p$ and $\mathrm{Cov}(V_j^*, V_{j'}^*) = 0.4$ for $j \neq j'$. We then dichotomize $\boldsymbol{V}^*$ and define $V_j = \mathrm{I}(V_j^* > \kappa_j)$, where

$$\kappa = (-0.6, -0.8, -0.7, -0.8, -0.5, -0.9, -0.9, -0.7).$$

Inference (Marginal mean) and Imputation (Conditional) Models

The inference and imputation models for the simulation study are specified as

$$\mathrm{logit}\, P(Y_{it} = 1) = \beta_0 + \beta_1(t - \bar{t}). \tag{6}$$

$$\mathrm{logit}\, P(Y_{it}|\boldsymbol{V}_i, Y_{i,t-1}) = \Delta_{i,t} + \sum_{j=1}^{p} \alpha_j V_{ij} + \gamma Y_{i,t-1} \tag{7}$$

for $t = 0, \ldots, T$. Here, we set $T = 3$. For ease of notation, we let $Y_{-1} \equiv 0$.

We simulate the complete response data $\boldsymbol{Y}$ using the parameter values given in Table 1.

Specification of Missing Data Mechanism

We specify the missing data mechanism with the following form,

$$\mathrm{logit}\, P(R_{it} = 0|R_{i,t-1} = 1, \boldsymbol{Y}_i, \boldsymbol{V}_i) = \psi_0 + \sum_{j=1}^{p} \psi_j V_{ij} + \phi_0 Y_{i,t-1} + \phi_1 Y_{it}, \tag{8}$$

where $R_{it} = \mathrm{I}(Y_{it}$ is observed) and $\phi_1 = 0$. The values of $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$ are given in Table 1. With these values, the missing data rate at $T = 3$ is about 49%. To evaluate the impact of the auxiliary covariates on the MDM, we fit the following model

$$\mathrm{logit}\, P(R_{it} = 0|R_{i,t-1} = 1, \boldsymbol{Y}_i, \boldsymbol{V}_i) = \psi_0^* + \phi_0^* Y_{i,t-1} + \phi_1^* Y_{it}$$

to $20,000$ complete responses $\boldsymbol{Y}$, generated by (6) and (7) and missing data indicators $\boldsymbol{R}$

generated by (8). With the auxiliary covariates (incorrectly) ignored, we obtain $\hat{\phi}_1^* = 0.36$ which indicates a strong dependence of R_{it} on Y_{it} and missing not at random.

Inclusion of $\boldsymbol{V}$'s unnecessary for A-MAR

To evaluate the robustness to inclusion of $\boldsymbol{v}$'s that are unnecessary for A-MAR, but predictive of $\boldsymbol{y}$, we set the coefficients of $V_5, \ldots, V_8$ $(\psi_5, \ldots, \psi_8)$ to zero in the MDM in (8). We examine the efficiency in terms of estimation of $\boldsymbol{\beta}$ in this context.

Models considered

We consider five different models: 1) model without auxiliary covariates; 2) model with the necessary auxiliary covariates plus unnecessary ones and the shrinkage prior for their distribution; 3) model with the necessary auxiliary covariates plus unnecessary ones and the noninformative prior for their distribution; 4) model with the necessary auxiliary covariates and the shrinkage prior for their distribution; 5) model with the necessary auxiliary covariates and a noninformative prior for their distribution. For the noninformative prior for the distribution of the auxiliary covariates, we use a Dir($\mathbf{1}$) prior on $\boldsymbol{\theta} = \{\theta_{\underline{l}}\}$.

3.2 *Simulation Results*

We simulated 400 datasets each for sample sizes of 100, 200, and 20000. For posterior sampling, we used 10000 iterations after a burn-in of 2500 iterations. To evaluate the estimation of the distribution of the auxiliary covariates, $p(\boldsymbol{v})$, we report a Pearson's Chi-square of $p(\boldsymbol{v})$,

$$\sum_{\boldsymbol{v}} \frac{(P(\boldsymbol{V}=\boldsymbol{v}) - EP(\boldsymbol{V}=\boldsymbol{v}|\boldsymbol{N}))^2}{P(\boldsymbol{V}=\boldsymbol{v})},$$

where $\boldsymbol{v}$ are categories of $\boldsymbol{V}$. For the other parameters, we report bias and mean square error (MSE).

Table 2 shows point estimates for the five models considered and for the three sample sizes. The results show that when the missing data mechanism is A-MAR, ignoring V can result in significant bias (as expected); this is seen (in particular) in the slope, β_1 which then propagates to large bias in the response probabilities, $P(Y_t = 1|X)$, especially for $t = 2, 3$. On the other hand, when the "correct" auxiliary covariates are conditioned upon and A-MAR

holds, adding unnecessary auxiliary covariates to the imputation model will not introduce bias for $\boldsymbol{\beta}$, which is important given that the inference model is obtained by integrating all of the auxiliary covariates out of the imputation model. Such a feature is desirable in practice as it implies that when there are no issues with collinearity, researchers may choose to be conservative and include extra auxiliary covariates as opposed to focusing on model selection for the auxiliary covariates. However, when the sample size is small, we see that there is some loss of efficiency from including extra, unnecessary auxiliary covariates (in terms of MSE of $\boldsymbol{\beta}$), but this goes away as the sample size increases. Also, for small sample sizes, we see some loss of efficiency and small biases for fully observed responses (e.g., Y_0), but not for responses with missingness (e.g., Y_3).

The results also show that the shrinkage approach for the joint distribution of the auxiliary covariates performs better for the estimation of $P(Y_j)$ and $p(\boldsymbol{v})$ than the non-informative prior approach especially when sample size is not large (and the dimension of $\boldsymbol{v}$ is not small); for the latter this often results in sparsity in the joint distribution of the auxiliary covariates. We note that here that even though the marginal independence assumption on the categorical auxiliary covariates is not correct, gains were still seen from shrinking toward this simple structure.

Table 3 shows the 90% credible interval coverage rates. Note that precision of the coverage estimates in the table are $\pm.04$ (i.e., ± 2 SE's). The coverage for the shrinkage approach using only the necessary auxiliary covariates is very good (though a bit conservative); a slight benefit (in terms of coverage) is seen over the noninformative approach for small sample sizes. The model without auxiliary covariates has coverage that gets worse as the sample size increases. The model with the extra auxiliary covariates has coverage a bit below the nominal level; the importance of shrinkage for estimating the distribution of the auxiliary covariates is very clear in this case as the coverage under the noninformative prior is not very good.

4. Smoking Cessation Trial

The Commit to Quit (CTQ) study was a randomized, controlled, prospective trial designed to evaluate the effect of exercise on smoking cessation (Marcus et al., 1999). The inclusion criteria required the participants to be healthy women aged 18 to 65, who smoked ten or more cigarettes per day for more than three years and exercised less than 90 minutes per week for at least 6 months.

All participants joined a 12-week, group-based, cognitive-behavioral smoking cessation program. In addition, the treatment arm participants were required to exercise three sessions per week with exercise specialist supervision. To eliminate the potential bias of treatment effect due to added staff time, the control arm participants were given three supervised health education lectures per week. Smoking cessation status was measured weekly by self-report of the number of cigarettes smoked daily and confirmed by analyzing cotinine in saliva and carbon monoxide in end-expiratory air. The target quit day was week 4 following randomization.

A total of 134 female smokers were randomized to the treatment exercise arm ($X = 1$) and 147 were randomized to control ($X = 0$). The dropout was substantial: at the end of the 12-week follow up, the dropout rates were 30.6% and 34.7% for the treatment and control arm respectively. However, these rates were less than in many cessation trials.

The baseline covariates collected on the CTQ trial included demographic and psychosocial predictors and smoking histories. As auxiliary covariates that significantly predict smoking cessation and drop out, we consider in the following analysis: education, length of previous quit attempts, and age (Borrelli et al., 2002). The discretization of the covariates was based on their clinical interpretation (Table 4).

By randomization, $\boldsymbol{V}$ is independent of treatment assignment X (i.e. $p(\boldsymbol{v}|X = x) = p(\boldsymbol{v})$ for $x = 0$ and 1) for CTQ data analysis. We focus on the CTQ data after the target quit time point, and define the 'baseline' response Y_0 as the last observed smoking status between week 1 to week 4. That is, $Y_0 = Y_{t'}$ where $t' = \max_{t\in 1,\ldots,4}\{Y_t \text{ observed}\}$. The inference model

is specified as:

$$\text{logit}\{E\left(Y_{i,0}|X=x_i;\boldsymbol{\beta}\right)\}=\beta_{0,x_i}$$

$$\text{logit}\{E\left(Y_{it}|X=x_i;\boldsymbol{\beta}\right)\}=\beta_{1,x_i},\quad 5\leq t\leq 12$$

and the imputation model is specified as

$$\text{logit}\left\{E(Y_{i,5}|\boldsymbol{V}_i,Y_{i,0},X_i;\boldsymbol{\alpha},\boldsymbol{\gamma})\right\}=\Delta_{i,5}+g(\boldsymbol{V}_i,X_i;\boldsymbol{\alpha})+\gamma_0 y_{i,0}$$

$$\text{logit}\left\{E(Y_{it}|\boldsymbol{V}_i,Y_{i,t-1},X_i;\boldsymbol{\alpha},\boldsymbol{\gamma})\right\}=\Delta_{i,t}+g(\boldsymbol{V}_i,X_i;\boldsymbol{\alpha})+\gamma_1 y_{i,t-1},\quad 6\leq t\leq 12$$

with

$$g(\boldsymbol{V}_i,X_i;\boldsymbol{\alpha})=\alpha_1*\text{YrsEduc}_i+\alpha_2*\text{RectQuit}_i+\alpha_3*\text{LongQuit}_i+\alpha_4*\text{Age}_i,$$

and not depending on X_i.

In our analysis here, we assume A-MAR (conditional on observed responses, treatment, and auxiliary covariates); we expect this assumption to be more reasonable than MAR (which does not condition on any auxiliary covariates). The posterior sampling is based on 20000 iterations with a thinning factor of 2 and a burn-in of 4000. Multiple chains and trace plots were used to verify the convergence (not shown).

Figure 1 presents various estimates of $p(\boldsymbol{v})$, including the observed frequencies of $\boldsymbol{V}$ as the empirical results, the posterior mean of $p(\boldsymbol{v})$ using the shrinkage method, the posterior mean of $p(\boldsymbol{v})$ using non-informative priors, and the empirical estimation of $p(\boldsymbol{v})$ assuming auxiliary covariates are independent. From the results, we see that the estimates under the shrinkage prior are shrunk toward the estimated $p(\boldsymbol{v})$ assuming the auxiliary covariates are independent. Thus, the shrinkage method allows information sharing by collapsing over $\boldsymbol{V}$ categories, which may be preferable especially when the data are extremely sparse.

The posterior means and 95% credible intervals for parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are reported in Table 5. We define *significance* as the 95% credible intervals excluding the null value (here, zero). We note the differences in $\beta_{1,\text{Exercise}}$ and in $\beta_{1,\text{Control}}$ between the model without auxiliary covariates and the models with auxiliary covariates. We also point out the slightly narrower confidence intervals for these two parameters for the shrinkage approach versus

the noninformative prior approach. Also, in terms of the auxiliary covariates, the covariate LongQuit (days of longest quit before the program) was significant with those with longer previous quit attempts less likely to be smoking.

Figure 2 shows the posterior density of the difference of smoking cessation rates between the exercise and control group and displays the (slight) shift in the posterior from *excluding* the auxiliary covariates. Posterior summaries for the difference of smoking cessation rates between the exercise and control group corresponding to the different approaches is reported in Web Appendix C. The biggest difference is between the no auxiliary covariate approach vs. the three auxiliary covariate approaches.

5. Discussion

We have proposed a fully Bayes approach to allow the imputation and inference models to be compatible that provides for a simple interpretation of the coefficients of covariates in the inference model (which is embedded within the imputation model). Simulations show that the approach is robust to inclusion of unnecessary auxiliary covariates and shrinkage estimation of the (marginal) distribution of auxiliary covariates leads to more efficient inferences. In the CTQ data analysis, we incorporated clinically relevant covariates such as length of previous quit attempts in the imputation model of the proposed approach. As the result, the difference of the smoking cessation rate between the control and the exercise arm was smaller when the auxiliary covariates were taken into account, indicating the evidence for the positive effect of the intervention may be weaker than shown under an (incorrect) MAR assumption. The objective here was to move closer to an MAR assumption being correct by including auxiliary covariates. However, the missingness may still be MNAR. We are working on extending these models to MNAR and allowing for sensitivity parameters.

Numerous other extensions to this approach are apparent. In terms of covariates, extending our approach to continuous (i.e., not discretizing) and time-varying covariates (possibly with missingness) would be very useful. For the shrinkage priors for the distribution of auxiliary covariates, alternative shrinkage targets based on parsimonious log linear models would allow

more flexibility; computations might be facilitated by estimating the hyperparameters using maximum likelihood and then using an empirical Bayes type approach.

In addition, further work is needed on the best approaches to decide which auxiliary covariates to include, which depends on being needed for MAR but also, being predictive of the response. Approaches like those in (Wang et al., 2012) could be useful here. To allow for the possibility of many auxiliary covariates, we might put priors on the coefficients in the imputation model that are shrunk towards zero with an unknown variance and/or use a spike and slab prior (Ishwaran and Rao, 2005).

We are currently working on comparing the proposed approach to approximate Bayesian and frequentist approaches that are not congenial. We will explore the extent of bias as a function of how 'uncongenial' particular inference and imputation models are to obtain a better understanding of the practical implications of not having congenial models.

6. Supplementary Materials

Web Appendices, Tables referenced in Sections 2 and 4 and the R code implementing our algorithm are available with this paper at the Biometrics website on Wiley Online Library. The R code is also available at www.sbs.utexas.edu/mjdaniels.

Acknowledgments

This work was partially supported by NIH grant CA85295. We thank Shira Dunsiger for help with data including clarifications, Minzhao Liu and Qin Li for some computing assistance at the various stages in the project.

References

Borrelli, B., Hogan, J., Bock, B., Pinto, B., Roberts, M., and Marcus, B. (2002). Predictors of quitting and dropout among women in a clinic-based smoking cessation program. *Psychology of Addictive Behaviors* **16,** 22–27.

Burns, R. A., Butterworth, P., Kiely, K. M., Bielak, A. A., Luszcz, M. A., Mitchell, P., Christensen, H., Von Sanden, C., and Anstey, K. J. (2011). Multiple imputation was an efficient method for harmonizing the mini-mental state examination with missing item-level data. *Journal of clinical epidemiology* **64,** 787–793.

Daniels, M. (1999). A prior for the variance in hierarchical models. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* **27,** 567–578.

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

Gelman, A. and Raghunathan, T. (2001). Using conditional distributions for missing-data imputation. *Discussion of" Conditionally Specified Distributions," by Arnold et al., Statistical Science* **3,** 268–269.

Heagerty, P. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics* **58,** 342–351.

Ishwaran, H. and Rao, J. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics* **33,** 730–773.

Kenward, M. G. and Carpenter, J. (2007). Multiple imputation: current perspectives. *Statistical Methods in Medical Research* **16,** 199–218.

Lavori, P., Dawson, R., and Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in medicine* **14,** 1913–1925.

Little, R. and Rubin, D. (1987). *Statistical analysis with missing data*. Wiley New York.

Liu, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of multivariate analysis* **53,** 139–158.

Marcus, B., Albrecht, A., King, T., Parisi, A., Pinto, B., Roberts, M., Niaura, R., and Abrams, D. (1999). The efficacy of exercise as an aid for smoking cessation in women: a randomized controlled trial. *Archives of Internal Medicine* **159,** 1229–1234.

Meng, X. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* **9,** 538–558.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate

technique for multiply imputing missing values using a sequence of regression models. *Survey methodology* **27,** 85–96.

Roy, J. and Daniels, M. (2008). A general class of pattern mixture models for nonignorable dropout with many possible dropout times. *Biometrics* **64,** 538–545.

Rubin, D. (1976). Inference and missing data. *Biometrika* **63,** 581–592.

Rubin, D. (1978). Multiple imputations in sample surveys–a phenomenological bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 20–34.

Rubin, D. and Schafer, J. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. In *Proceedings of the Statistical Computing Section of the American Statistical Association*, pages 83–88.

Schafer, J. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.

Van Buuren, S., Boshuizen, H., and Knook, D. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine* **18,** 681–694.

Wang, C., Daniels, M., Scharfstein, D., and Land, S. (2010). A Bayesian shrinkage model for incomplete longitudinal binary data with application to the breast cancer prevention trial. *Journal of the American Statistical Association* **105,** 1333–1346.

Wang, C., Parmigiani, G., and Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* **68,** 661–671.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

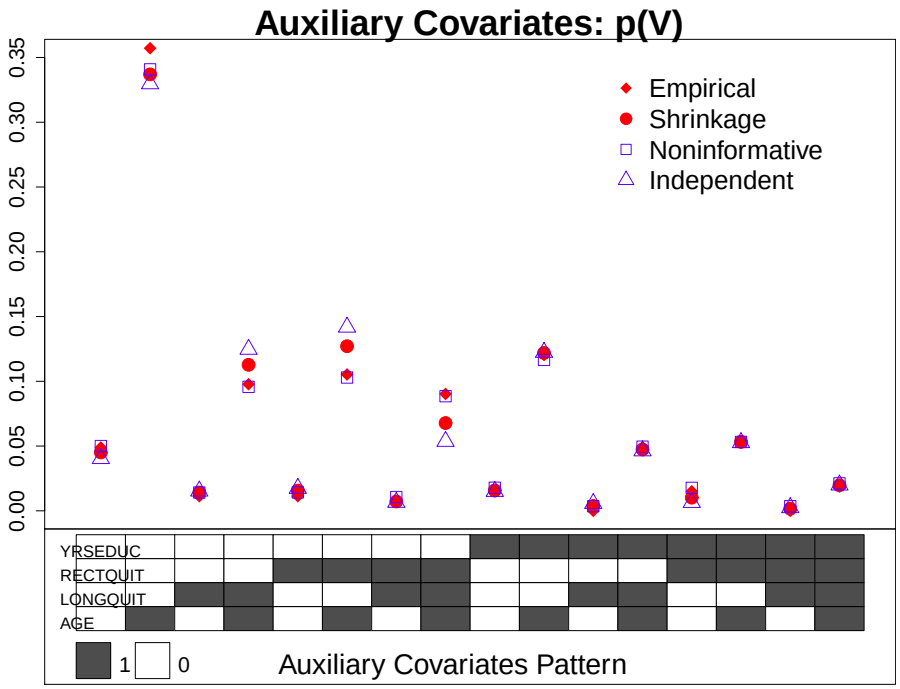

Figure 1. Estimates of $p(v)$ based on the empirical distribution (empirical) and the empirical distribution under independence (independent) and posterior means using the shrinkage and noninformative priors.

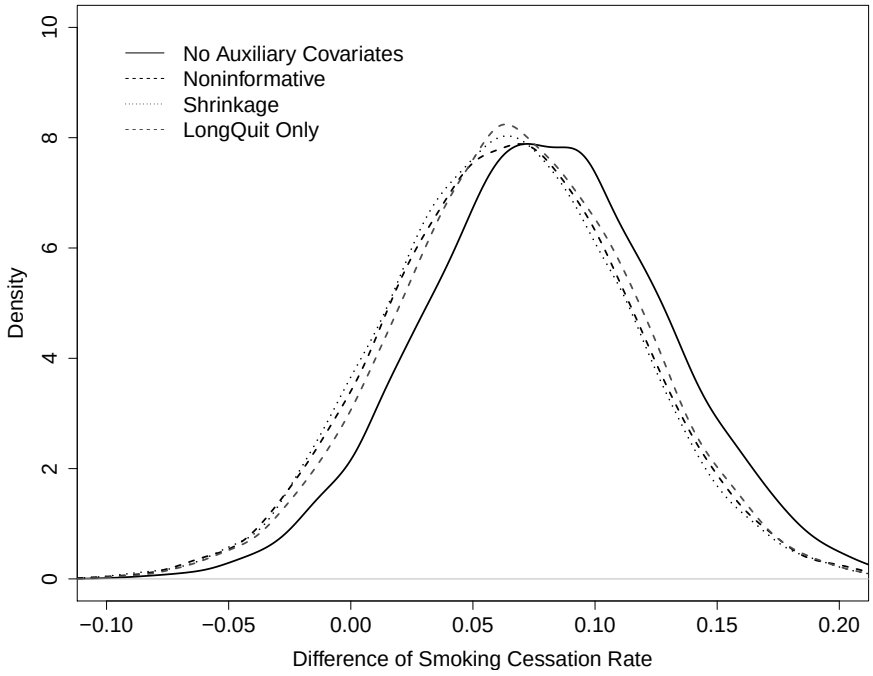

Figure 2. Posterior Density of Difference of Smoking Cessation Rates

Table 1
Simulation Parameter Values.

Marginal		Transition				MDM					
β_0	0.5	α_1	0.4	α_6	0.6	ψ_0	-3.5	ψ_5	0	ϕ_1	0
β_1	0.25	α_2	0.3	α_7	0.3	ψ_1	0.6	ψ_6	0		
		α_3	0.5	α_8	0.7	ψ_2	0.7	ψ_7	0		
		α_4	0.9	γ	0.3	ψ_3	0.5	ψ_8	0		
		α_5	0.8			ψ_4	0.4	ϕ_0	0.5		

Table 2

Simulation results: posterior mean (MCSE) for ignoring all auxiliary covariates (No V), shrinkage (Shrinkage∗) and non-informative (Noninform∗) methods considering extra V, and shrinkage (Shrinkage) and non-informative (Noninform) methods ignoring extra V. $E(MSE\ of\ \beta) = E(\sum(\hat{\beta}-\beta)^2)$, $E(MSE\ of\ P(Y_t)) = E(\hat{P}(Y_t=1)-P(Y_t=1))^2$.

	No V	Shrinkage∗	Noninform∗	Shrinkage	Noninform
Sample Size 100					
$10^3 * \chi^2$ of $P(V)$	NA	20.34(0.54)	55.12(0.30)	NA	NA
Bias β_0	0.17(0.01)	0.19(0.01)	0.75(0.01)	0.14(0.01)	0.16(0.01)
Bias β_1	0.11(0.00)	0.09(0.00)	0.09(0.00)	0.09(0.00)	0.09(0.00)
10^3∗MSE β_0	41.33(2.57)	50.51(3.05)	592.59(11.6)	30.44(2.39)	38.10(2.32)
10^3∗MSE β_1	19.95(1.26)	13.72(0.96)	12.14(0.83)	12.96(0.89)	11.84(0.80)
Bias $P(Y_0)$	0.03(0.00)	0.05(0.00)	0.18(0.00)	0.04(0.00)	0.04(0.00)
Bias $P(Y_1)$	0.03(0.00)	0.04(0.00)	0.18(0.00)	0.03(0.00)	0.03(0.00)
Bias $P(Y_2)$	0.04(0.00)	0.04(0.00)	0.17(0.00)	0.03(0.00)	0.04(0.00)
Bias $P(Y_3)$	0.06(0.00)	0.05(0.00)	0.16(0.00)	0.04(0.00)	0.04(0.00)
10^3∗MSE $P(Y_0)$	1.9(0.13)	3.61(0.23)	35.46(0.85)	2.04(0.14)	2.43(0.16)
10^3∗MSE $P(Y_1)$	1.48(0.10)	2.46(0.15)	33.58(0.66)	1.24(0.09)	1.73(0.11)
10^3∗MSE $P(Y_2)$	2.8(0.18)	2.61(0.17)	31.29(0.75)	1.54(0.11)	2.11(0.14)
10^3∗MSE $P(Y_3)$	5.37(0.35)	3.47(0.25)	29.44(1.01)	2.40(0.18)	3.13(0.23)
Sample Size 200					
$10^3 * \chi^2$ of $P(V)$	NA	12.90(0.35)	33.96(0.25)	NA	NA
Bias β_0	0.13(0.00)	0.15(0.01)	0.59(0.01)	0.10(0.00)	0.10(0.00)
Bias β_1	0.09(0.00)	0.06(0.00)	0.06(0.00)	0.06(0.00)	0.06(0.00)
10^3∗MSE β_0	26.51(1.66)	32.5(2.01)	359.52(5.96)	14.47(1.03)	16.93(1.16)
10^3∗MSE β_1	11.9(0.69)	5.54(0.40)	4.87(0.34)	5.29(0.38)	5.04(0.36)
Bias $P(Y_0)$	0.03(0.00)	0.04(0.00)	0.14(0.00)	0.03(0.00)	0.03(0.00)
Bias $P(Y_1)$	0.02(0.00)	0.03(0.00)	0.14(0.00)	0.02(0.00)	0.02(0.00)
Bias $P(Y_2)$	0.04(0.00)	0.03(0.00)	0.13(0.00)	0.02(0.00)	0.02(0.00)
Bias $P(Y_3)$	0.05(0.00)	0.03(0.00)	0.12(0.00)	0.03(0.00)	0.03(0.00)
10^3∗MSE $P(Y_0)$	1.01(0.07)	2.45(0.15)	21.37(0.46)	1.08(0.07)	1.18(0.08)
10^3∗MSE $P(Y_1)$	0.91(0.06)	1.71(0.11)	19.91(0.34)	0.65(0.05)	0.79(0.06)
10^3∗MSE $P(Y_2)$	1.82(0.11)	1.54(0.10)	18.05(0.36)	0.71(0.05)	0.88(0.06)
10^3∗MSE $P(Y_3)$	3.37(0.19)	1.72(0.12)	16.21(0.46)	1.02(0.07)	1.23(0.09)
Sample Size 20000					
$10^3\chi^2$ of $P(V)$	NA	0.06(0.00)	0.06(0.00)	NA	NA
Bias β_0	0.11(0.00)	0.01(0.00)	0.02(0.00)	0.01(0.00)	0.01(0.00)
Bias β_1	0.08(0.00)	0.01(0.00)	0.01(0.00)	0.01(0.00)	0.01(0.00)
10^3∗MSE β_0	12.99(0.13)	0.20(0.01)	0.33(0.02)	0.14(0.01)	0.14(0.01)
10^3∗MSE β_1	6.73(0.06)	0.05(0.00)	0.05(0.00)	0.06(0.00)	0.06(0.00)
Bias $P(Y_0)$	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
Bias $P(Y_1)$	0.02(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
Bias $P(Y_2)$	0.03(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
Bias $P(Y_3)$	0.04(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
10^3∗MSE $P(Y_0)$	0.01(0.00)	0.01(0.00)	0.02(0.00)	0.01(0.00)	0.01(0.00)
10^3∗MSE $P(Y_1)$	0.26(0.00)	0.01(0.00)	0.02(0.00)	0.01(0.00)	0.01(0.00)
10^3∗MSE $P(Y_2)$	0.99(0.01)	0.01(0.00)	0.02(0.00)	0.01(0.00)	0.01(0.00)
10^3∗MSE $P(Y_3)$	1.91(0.02)	0.01(0.00)	0.02(0.00)	0.01(0.00)	0.01(0.00)

Table 3

Simulation results: posterior 90% CI coverage rate for ignoring all auxiliary covariates (No V), shrinkage (Shrinkage$*$) and non-informative (Noninform*) methods considering extra V, and shrinkage (Shrinkage) and non-informative (Noninform) methods ignoring extra V.

	No V	Shrinkage*	Noninform*	Shrinkage	Noninform
Sample Size 100					
β_0	0.86	0.81	0.01	0.94	0.87
β_1	0.89	0.92	0.92	0.94	0.94
$P(Y_0)$	0.95	0.86	0.04	0.95	0.92
$P(Y_1)$	0.91	0.82	0.01	0.94	0.90
$P(Y_2)$	0.82	0.84	0.02	0.93	0.88
$P(Y_3)$	0.84	0.87	0.17	0.93	0.90
Sample Size 200					
β_0	0.80	0.74	0.00	0.93	0.89
β_1	0.85	0.94	0.95	0.96	0.95
$P(Y_0)$	0.94	0.80	0.01	0.92	0.92
$P(Y_1)$	0.88	0.73	0.00	0.93	0.91
$P(Y_2)$	0.77	0.81	0.00	0.94	0.92
$P(Y_3)$	0.77	0.89	0.05	0.95	0.94
Sample Size 20000					
β_0	0.00	0.86	0.71	0.92	0.91
β_1	0.00	0.92	0.89	0.92	0.90
$P(Y_0)$	0.88	0.89	0.78	0.91	0.92
$P(Y_1)$	0.00	0.84	0.71	0.93	0.92
$P(Y_2)$	0.00	0.87	0.73	0.92	0.92
$P(Y_3)$	0.00	0.89	0.79	0.91	0.91

Table 4
CTQ Auxiliary Covariates

Auxiliary Covariate	Definition 0	Definition 1	Description
YrsEduc	≤ 15	> 15	years of education. 0:no advanced degree, 1:advanced degree
RectQuit	≤ 21	> 21	days of most recent quit before program. 0:shorter than 3 weeks, 1:longer than 3 weeks
LongQuit	≤ 180	> 180	days of longest quit before program. 0:shorter than 6 month, 1:longer than 6 month
Age	≤ 30	> 30	age. 0:younger than 30, 1:older than 30

Table 5
Posterior mean (95% CI) of the model parameters

Par	No Auxiliary V	Noninformative	Shrinkage	LongQuit Only
$\beta_{0,\text{Control}}$	-2.68(-3.42,-2.06)	-2.67(-3.39,-2.03)	-2.66(-3.39,-2.02)	-2.67(-3.39,-2.03)
$\beta_{1,\text{Control}}$	-0.85(-1.19,-0.54)	-0.82(-1.14,-0.50)	-0.81(-1.14,-0.49)	-0.83(-1.16,-0.52)
$\beta_{0,\text{Exercise}}$	-2.37(-3.04,-1.78)	-2.40(-3.07,-1.82)	-2.38(-3.03,-1.80)	-2.40(-3.06,-1.82)
$\beta_{1,\text{Exercise}}$	-0.49(-0.80,-0.19)	-0.53(-0.83,-0.22)	-0.52(-0.82,-0.22)	-0.53(-0.83,-0.24)
γ_0	2.40(1.30, 3.67)	2.30(1.18, 3.59)	2.31(1.19, 3.57)	2.30(1.18, 3.58)
γ_1	4.97(4.58, 5.37)	4.93(4.54, 5.34)	4.93(4.54, 5.34)	4.92(4.53, 5.33)
α_1 (YrsEduc)	NA	0.12(-0.23, 0.47)	0.12(-0.24, 0.47)	NA
α_2 (RectQuit)	NA	-0.02(-0.36, 0.33)	-0.02(-0.36, 0.32)	NA
α_3 (LongQuit)	NA	0.58(0.25, 0.92)	0.58(0.24, 0.92)	0.57(0.24, 0.91)
α_4 (Age)	NA	0.06(-0.44, 0.57)	0.06(-0.43, 0.57)	NA