doi: 10.1111/j.1467-842X.XXX

# Sequential imputation for models with latent variables assuming latent ignorability

Lauren J. Beesley,[1*] Jeremy M. G. Taylor,[1] and Roderick J. A. Little[1]

*Department of Biostatistics, University of Michigan*

**Summary**

Models that involve an outcome variable, covariates, and latent variables are frequently the target for estimation and inference. The presence of missing covariate or outcome data presents a challenge, particularly when missingness depends on the latent variables. This missingness mechanism is called *latent ignorable* or *latent missing at random* and is a generalization of missing at random. Several authors have previously proposed approaches for handling latent ignorable missingness, but these methods rely on prior specification of the joint distribution for the complete data. In practice, specifying the joint distribution can be difficult and/or restrictive. We develop a novel sequential imputation procedure for imputing covariate and outcome data for models with latent variables under *latent ignorable missingness*. The proposed method *does not require a joint model*; rather, we use results under a joint model to inform imputation with less restrictive modeling assumptions. We discuss identifiability and convergence-related issues, and simulation results are presented in several modeling settings. The method is motivated and illustrated by a study of head and neck cancer recurrence.

*Key words:* multiple imputation; substantive model compatible imputation; chained equations; latent missing at random; latent ignorability

## 1. Introduction

Models that involve latent or partially latent variables in addition to an outcome variable and covariates are frequently the target for estimation and inference. For example, in the Cox proportional hazards mixture cure model, partially latent cure status describes whether individuals are at risk for the event of interest. Cure status is only partially latent because subjects with observed events are known to be non-cured. Another popular model with latent variables is the linear mixed model, where fully latent random effects account for correlation within clusters.

Additional considerations arise when dealing with missing covariates and/or outcomes

*Author to whom correspondence should be addressed.
[1] Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA
Email: lbeesley@umich.edu
*Acknowledgment.* The authors cite the many investigators in the University of Michigan Head and Neck Specialized Program of Research Excellence for their contributions to patient recruitment, specimen collection, and study conduct. This research was partially supported by the National Institutes of Health.

in the presence of latent variables. Many authors have explored the issue of missing data for models with latent variables under assumptions that missingness is independent of the latent variable given the observed data (e.g. Beesley et al. 2016). In this paper, we explore a generalization of this missingness mechanism that allows covariate/outcome missingness to depend on the latent variable, which is a *missing not at random* (MNAR) mechanism (Little & Rubin 2002). Previous examples of such mechanisms are called *latent ignorable* or *latent missing at random* (LMAR) missingness (Frangakis & Rubin 1999; Harel 2003; Harel & Schafer 2009). For example, suppose we model a longitudinal outcome using a mixed model. One common LMAR scenario in the literature relates dropout to the random effect, which can be viewed as a measure of an individual's propensity to drop out.

In general, the underlying missingness mechanism can never be determined from the data alone, and inference under MNAR may be sensitive to unverifiable assumptions about the missingness mechanism. Additionally, inference under MNAR is susceptible to underidentification or weak identification of the model parameters (Little 1995; Molenberghs, Beunckens & Sotto 2008). In this paper, we consider a particular MNAR missingness mechanism (LMAR) in which missingness depends on unknown information *only* through the latent variable, which by assumption has a structured relationship with the observed variables. Therefore, we may view LMAR missingness as a somewhat mild departure from MAR. Still, we must keep these issues in mind when handling missing data under LMAR.

One approach for handling missing data is to analyze only the fully observed subset of the data (complete case analysis). When missingness is LMAR, this approach will generally produce biased results (Little & Rubin 2002). Several authors have discussed likelihood-based approaches for linear mixed models with missingness dependent on the random effect (e.g. Little 1995; Wu & Carroll 1988). These methods often involve an EM algorithm or a likelihood that has integrated out the latent variable.

Multiple imputation is a common general approach for dealing with missing data. One approach to multiple imputation requires one to specify a joint distribution for all the variables and use that joint distribution for imputation, usually in a Gibbs sampling-type algorithm. Each variable with missing values can be sequentially imputed using its conditional distribution, which is determined by the joint distribution. The distribution of the sampled parameters can then be used for inference. Several authors have proposed approaches for handling latent ignorable missingness in specific joint modeling settings (Jung 2007; Yang, Lu & Shoptaw 2008; Lu, Zhang & Lubke 2011). Harel (2003) proposes a non-iterative imputation approach for dealing with general latent-dependent missingness under a joint model.

Existing imputation methods under latent ignorability, however, are limited in their applicability. The main drawback of the joint modeling approach to imputation is that specification of the joint distribution may be difficult or too restrictive, particularly when we have many covariates of different types. Indeed, Gelman (2004) argues that 'having a joint distribution in the imputation is less important than incorporating information from other variables and unique features of the dataset (e.g. bounds, skip patterns, nonlinearity, interactions, etc.).' As such, there is a need to consider methods for imputing variables under latent ignorability that incorporate less restrictive assumptions about the joint model.

Chained equations imputation is an alternative to joint modeling in which variables are imputed iteratively in a series of univariate imputation steps (Raghunathan 2001; Van Buuren et al. 2006). These steps are usually accomplished using standard regression models that can incorporate additional features of the data, and these regressions as a set usually do not correspond to a valid joint distribution. This approach is simple and flexible, but it is less coherent than joint modeling and may not incorporate assumptions about the outcome model directly. Most literature on chained equations assumes that missingness is independent of all unobserved information, called *missing at random* (MAR) (Little & Rubin 2002), and some authors have explored particular limited MNAR settings (e.g. Van Buuren 2007; Little 2009a; Giusti & Little 2011). An alternative approach proposed in Bartlett et al. (2014) called substantive model compatible (SMC) imputation incorporates the outcome model into the chained equations imputation procedure but does not require the user to specify a valid joint distribution for the covariates, leading to improved properties over conventional chained equations but additional flexibility over joint modeling. Similar findings are given in White & Royston (2009) and Beesley et al. (2016). Beesley et al. (2016) explores SMC imputation for a particular modeling setting with latent variables, but we have not found any literature exploring chained equations or SMC imputation under latent ignorable missingness in general.

In this paper, we develop a novel sequential imputation method that can handle MAR and LMAR covariate and outcome missingness for models with latent or partially latent variables and that *does not* require a joint model. The proposed method imputes the latent variable as part of the missing data, allowing the latent variable to be directly used when imputing the missing covariate/outcome values. We first consider the more restrictive setting where the joint model is fully specified. We use results under a joint model to *inform the structure* of the imputation distributions and the method for drawing parameters in the proposed algorithm *without requiring specification of the joint model.* The proposed approach is very flexible and can accommodate either a chained equations-type approach to imputation or a SMC imputation approach that is more strongly informed by the outcome model.

Many works have explored MAR-based imputation in settings with latent variables under a joint model (e.g. Schafer 1997; Schafer & Yucel 2002; Chung, Flaherty & Schafer 2006) or using less restrictive assumptions Beesley et al. (2016). While the proposed method can be applied under MAR or LMAR, the primary novelty consists of the application to the LMAR setting. Existing methods for handling missing data in the LMAR setting assume there is a fully-specified joint model, and this work serves as an extension of these existing methods with less restrictive modeling assumptions. This work develops novel statistical methodology for handling latent ignorability *without requiring joint modeling assumptions*. This is a departure from the existing literature in terms of allowing for much greater flexibility in model specification over joint modeling while still allowing for the incorporation of the outcome model structure into the imputation procedure. The SMC imputation approach to imputation has been previously explored in the context of MAR covariate imputation in Bartlett et al. (2014), but a general imputation algorithm for handling missingness in multiple variables and particularly under MNAR assumptions has not previously been considered. Additionally, the LMAR setting presents a range of identifiability-related difficulties that is not present in the usual MAR setting.

This work is motivated by a study of cancer recurrence in patients treated for head and neck cancer. In this study, many covariates of interest have substantial missingness; in particular, HPV status (human papillomavirus) has roughly 50% missingness. Previous work has explored imputation of these data under MAR assumptions (Beesley et al. 2016), but there is a belief that an induced association between missingness in HPV status and an underlying latent variable (cancer cure status) may be present. Existing statistical methods for latent-dependent missingness are undesirable due to the emphasis on specification of a joint model for the covariates, which in this setting is too restrictive. In particular, relationships between the covariates may be difficult to capture in a standard joint model. As such, there is a need to develop statistical methodology for performing imputation of the missing covariate values under latent-dependent missingness without the restrictive joint modeling assumptions. While this work is motivated by this particular problem, the statistical methods can be applied in a wide range of modeling settings.

In Section 2, we define latent ignorability. In Sections 3 and 4, we describe the proposed imputation approach. In Section 5, we present simulations that evaluate the performance of our method under a variety of scenarios. In Section 6, we apply the proposed methods to the motivating study of time to recurrence in patients with head and neck cancer. In Section 7, we present a discussion.

## 2. Latent ignorability

Suppose that the goal is to make inference about a model for outcome $Y$ given covariates $X$ and a latent (or partially latent) mixing variable, $L$. For example, the outcome model may be a linear mixed model with a latent random intercept. We may also be interested in the model for $L|X$. We restrict our attention to situations in which, if all of the covariate and outcome information were observed, the outcome model would be fully identified, and estimation using likelihood-based methods would be possible and lead to consistent parameter estimates. We consider missingness in $X$ and/or $Y$, and we allow missingness to be related to the latent variable, $L$.

Let vector $\boldsymbol{D}_i^\top = (\boldsymbol{X}_i^\top, \boldsymbol{Y}_i^\top)$ represent the (possibly incomplete) data for subject $i$. We assume $\boldsymbol{D}_i$ and $L_i$ are independent across subjects. Let $\boldsymbol{R}_i^D$ be a vector corresponding to whether each element of $D_i$ is observed and $R_i^L$ be an indicator for whether $L_i$ is known (can be 0 for all subjects). Define $\boldsymbol{R}_i^\top = (\boldsymbol{R}_i^{D\top}, R_i^L)$. For any vector $\boldsymbol{V}_i$, let $\boldsymbol{V}_i^{(\mathrm{obs})}$ and $\boldsymbol{V}_i^{(\mathrm{mis})}$ be the observed and missing elements of $\boldsymbol{V}_i$. Assume we have independence of $(D, L, R)$ across $i$.

We assume that missingness in $D_i$ is independent of $D_i^{(\mathrm{mis})}$ and $R_i^L$ such that

$$f(\boldsymbol{R}_i^D | D_i, L_i, R_i^L; \phi^D) = f(\boldsymbol{R}_i^D | D_i^{(\mathrm{obs})}, L_i; \phi^D) \quad (1)$$

We assume that $\phi^D$ is distinct from all other model parameters. We call assumption (1) the *latent missing at random* (LMAR) or *latent ignorability* assumption. This missingness mechanism was first studied in Frangakis & Rubin (1999) and is a special case of latent ignorability explored in Harel (2003) and Harel & Schafer (2009). In longitudinal data analysis, a similar mechanism relating missingness in $Y$ to latent random effects in a linear mixed model has been explored by many authors including Wu & Carroll (1988), Follmann & Wu (1995), Little (1995), and McCulloch, Neuhaus & Olin (2016). Since $L_i$ is latent or partially latent by definition, the mechanism in (1) is a type of MNAR, and when (1) does not depend on $L_i$, the mechanism reduces to MAR. We can view LMAR as a generalization of MAR with less restrictive assumptions.

We now consider assumptions regarding missingness in $L$, which may be latent or partially latent. We make a subtle distinction between *partially latent* and *partially missing* variables. Latent variable $L$ can be viewed as a modeling construct representing unobserved or perhaps unobservable quantities. The *observed* values of the partially latent $L$ are just a function of the observed data, $D^{(\mathrm{obs})}$, and therefore contain no additional information. For example, known values of the partially latent cure status in a Cox proportional hazards cure model are entirely determined by the event indicator and the event/censoring time for each subject. In this way, partially latent variables are different from partially missing variables, which may contain additional information in their observed values. However, we will treat latent and partially latent variables as if they were missing data for the purposes of this method.

When $L_i$ is fully latent, we can view missingness in $L_i$ as missing completely at random (MCAR) with probability of missingness equal to 1. When $L_i$ is partially latent, we allow missingness in $L_i$ to depend on $D_i^{(obs)}$ (so $L$ is MAR) such that

$$f(R_i^L|\boldsymbol{D}_i, L_i, \boldsymbol{R}_i^D; \phi^L) = f(R_i^L|\boldsymbol{D}_{\boldsymbol{i}}^{\mathbf{(obs)}}; \phi^L) \tag{2}$$

Figure 1 shows the assumed relationships between variables. The arrows represent dependence. For example, $R^L$ may depend on $\boldsymbol{X}^{(\text{obs})}$ and $\boldsymbol{Y}^{(\text{obs})}$.

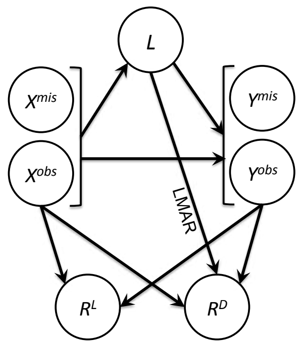


Figure 1. Variable relationships under latent ignorability

*Example 1, linear mixed model with a random intercept:* Suppose our model for multivariate outcome $\boldsymbol{Y}_i$ is a linear mixed model with a latent random intercept, $b_i$, and covariates $\boldsymbol{X}_i$. This model is commonly used for longitudinal data, where the outcome is measured within individuals over time. In such a setting, outcome missingness is particularly common due to dropout. Many authors have described scenarios in which dropout may be related to the random effects (Wu & Carroll 1988; Little 1995; Yang, Lu & Shoptaw 2008, e.g.). In this example, $b_i$ represents an individual's propensity to drop out. This is a LMAR mechanism with $L_i = b_i$. Covariate missingness may also be LMAR.

*Example 2, Cox proportional hazards mixture cure model:* The Cox proportional hazards (CPH) mixture cure model is used in event time analysis when some (cured) subjects are unable to experience the event of interest (Sy & Taylor 2000). For subjects with events, cure status is known, and it is unknown for censored subjects. Therefore, cure status is partially latent. Missingness in cure status is entirely determined by observed information, so its missingness can be viewed as MAR. Suppose we have covariate missingness. We can imagine scenarios in which covariate missingness may depend on the underlying cure status. For example, suppose covariate information is collected through a patient survey. Cured subjects may be more or less likely to answer certain survey questions, resulting in an association between missingness and cure status. Additionally, cure status may be related to an unmeasured confounder that is related to missingness. This will induce a dependence between missingness and cure status. We consider a similar LMAR mechanism in our data application.

*Example 3, mixture of generalized linear models:* Suppose our outcome $\boldsymbol{Y}$ is generated from a mixture of $K$ generalized linear models (GLMs). Let $C_i$ be a fully latent mixing variable indicating which element of the mixture distribution generated the observation for subject $i$. Missingness in $C_i$ can be viewed as MCAR with probability 1. If covariate or outcome missingness is related to $C$, missingness is LMAR. For example, suppose our data are collected using $K$ different populations. For example, we may collect data and multiple different locations and not record the location. The covariate/outcome missingness rates may vary by population, resulting in LMAR missingness.

## 3. Imputation of missing data

In this section, we develop an imputation algorithm for dealing with ignorable and latent ignorable covariate and outcome missingness. First, we explore imputation under a joint model for all the variables. We treat the latent variable as part of the missing data, and we use the form of the joint model to determine how each variable with missing values should be imputed. In particular, we determine which variables need to be included as predictors for each imputation model and describe the components of the joint model (e.g. outcome model, missingness model, covariate model) that are used for imputing each variable. We then use these results to guide our choice of sequential imputation models when a joint model is not specified.

### 3.1. Imputation under a joint model

Suppose that the data are directly modeled using a fully-specified joint model as follows:

$$f(\boldsymbol{D}, L, \boldsymbol{R}; \nu) = \prod_{i=1}^{n} f(\boldsymbol{R}_i|\boldsymbol{Y}_i, \boldsymbol{X}_i, L_i; \phi) f(\boldsymbol{Y}_i|\boldsymbol{X}_i, L_i; \theta) f(L_i|\boldsymbol{X}_i; \omega) f(\boldsymbol{X}_i; \psi) \quad (1)$$

where $\nu = (\phi, \theta, \omega, \psi)$ is the set of all model parameters. We assume a flat prior for $\nu$ such that $\phi$, $\theta$, $\omega$, and $\psi$ are all a priori independent (so they are distinct). The factorization (1) is a form of shared parameter model, where the latent variable is related both to missingness and to the distribution for $Y_i$ (Little & Rubin 2002).

We can impute missing values of $\boldsymbol{D}$ and $L$ by iteratively drawing the missing values from their posterior predictive distributions, $\boldsymbol{D}^{(\text{mis})} \sim f(\boldsymbol{D}^{(\text{mis})}|\boldsymbol{D}^{(\text{obs})}, L, \boldsymbol{R})$ and $L^{(\text{mis})} \sim f(L^{(\text{mis})}|\boldsymbol{D}, L^{(\text{obs})}, \boldsymbol{R})$. This leads to draws from the joint posterior predictive distribution, $f(\boldsymbol{D}^{(\text{mis})}, L^{(mis)}|\boldsymbol{D}^{(\text{obs})}, L^{(\text{obs})}, \boldsymbol{R})$ (Little & Rubin 2002). Define $\rho = (\theta, \omega, \psi)$. In the Supplementary Materials, we formally show the following ignorability properties under a joint model:

*Property 1:* Under MAR and LMAR, we can ignore $\boldsymbol{R} = (\boldsymbol{R}^{\boldsymbol{D}}, R^L)$ when imputing $\boldsymbol{D}$ from $f(\boldsymbol{D}^{(\text{mis})}|\boldsymbol{D}^{(\text{obs})}, L, \boldsymbol{R})$

*Property 2:* Under MAR (but not under LMAR), we can ignore $\boldsymbol{R} = (\boldsymbol{R}^{\boldsymbol{D}}, R^L)$ when imputing $L$ from $f(L^{(\mathrm{mis})}|\boldsymbol{D}, L^{(\mathrm{obs})}, \boldsymbol{R})$

*Property 3:* Suppose that missingness in subset $S$ of $\{\boldsymbol{D}, L\}$ is MAR. Let $\boldsymbol{R}^S$ denote the set of missingness indicators for $S$ and $\boldsymbol{R}^{-S}$ denote the missingness indicators for the remaining variables in $\{\boldsymbol{D}, L\}$. We can ignore $\boldsymbol{R^S}$ when imputing $L$ from $f(L^{(\mathrm{mis})}|\boldsymbol{D}, L^{(\mathrm{obs})}, \boldsymbol{R})$ provided a parameter distinctness property holds.

Rather than drawing $\boldsymbol{D}^{(\mathrm{mis})}$ and $L^{(\mathrm{mis})}$ from their posterior predictive distributions directly, we can instead impute each variable with missingness sequentially through a series of univariate imputation steps. Each time we impute a given variable, we treat the most recent imputations of the other variables as observed data. In practice, we specify the full conditional distribution of missing variable $V$ given all other variables (with parameter $v$) and obtain a draw from the posterior predictive distribution of $V$ by 1) drawing $v$ from its posterior distribution and 2) drawing missing values of $V$ from its full conditional distribution at the drawn $v$. After iteration, the imputations will approximate draws of $\boldsymbol{D}^{(\mathrm{mis})}$ and $L^{(\mathrm{mis})}$ from their posterior predictive distributions. Below, we present the form of the imputation distribution (step 1) for imputing different types of variables using the above ignorability properties.

**Predictive distribution of the latent variable**

Define $\boldsymbol{R}^S$ and $\boldsymbol{R}^{-S}$ as in *Property 3* and assume the distinctness property expressed in the Supplementary Materials holds. Then, we can ignore $\boldsymbol{R}^S$ when imputing $L$. Using assumptions (1)–(2) and joint model (1) and treating terms that do not depend on $L_i$ as constants, we have

$$f(L_i|\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{R}_i^{-S}; \nu) \propto f(\boldsymbol{R}_i^{-S}|\boldsymbol{Y}_i^{(\mathrm{obs})}, \boldsymbol{X}_i^{(\mathrm{obs})}, L_i; \phi^{-S}) \times f(\boldsymbol{Y}_i|\boldsymbol{X_i}, L_i; \theta) f(L_i|\boldsymbol{X}_i; \omega) \quad (2)$$

Under MAR, (2) simplifies to

$$f(L_i|\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{R}_i^{-S}; \nu) \propto f(\boldsymbol{Y}_i|\boldsymbol{X}_i, L_i; \theta) f(L_i|\boldsymbol{X}_i; \omega) \propto f(L_i|\boldsymbol{X}_i, \boldsymbol{Y}_i; \rho)$$

When treated as a function of $L_i$, expression (2) is proportional to the desired imputation distribution. We will call the distribution known up to proportionality the kernel. The kernel in (2) involves the distribution of $\boldsymbol{R}_i^{-S}$ under LMAR but not under MAR. In order to impute $L_i$ under LMAR using (2), we need to specify a model for $\boldsymbol{R}_i^{-S}$.

In some particular settings (for example, when $L_i$ is binary), we can use (2) to directly derive the full conditional distribution. When $L_i$ is continuous, the distribution may only be known up to a proportionality constant. In this case, we may need to use more advanced techniques to impute $L_i$ using (2). Many methods exists in the literature for drawing from a distribution knowing only the kernel. These include the Metropolis-Hastings algorithm and rejection sampling. For examples of such methods applied in the context of imputation, see Bartlett et al. (2014) and the Supplementary Materials.

**Predictive distributions of covariates and outcome**

In *Property 1*, we show that we can impute missing values of $\boldsymbol{D}$ ignoring the missingness mechanism under MAR and LMAR. We can similarly impute missing values of individual variables in $\boldsymbol{D}$ from their full conditional distributions without conditioning on $\boldsymbol{R}$.

We first determine the distribution for imputing missing outcome values. We note that $\boldsymbol{Y}$ may be uni- or multivariate. Suppose that we are imputing the $t^{th}$ element of $\boldsymbol{Y}_i$, denoted $Y_i^{(t)}$. Let $\boldsymbol{Y_i^{(-t)}}$ represent the terms in $\boldsymbol{Y}_i$ excluding $Y_i^{(t)}$. Using joint model (1), we can write the conditional distribution for imputing $Y_i^{(t)}$ under MAR and LMAR as

$$f(Y_i^{(t)}|\boldsymbol{Y_i^{(-t)}}, \boldsymbol{X}_i, L_i; \rho) \propto f(\boldsymbol{Y}_i, \boldsymbol{X}_i, L_i; \rho) \propto f(\boldsymbol{Y}_i|\boldsymbol{X}_i, L_i; \theta) \tag{3}$$

When $Y_i^{(t)} = \boldsymbol{Y}_i$, the conditional distribution is equal to $f(\boldsymbol{Y}_i|\boldsymbol{X}_i, L_i; \theta)$.

Suppose that we are imputing the $t^{th}$ covariate in $\boldsymbol{X}_i$, denoted $X_i^{(t)}$. Let $f(X_i^{(t)}|\boldsymbol{X}_i^{(-t)}; \psi)$ be the conditional distribution of $X_i^{(t)}$ given all other variables in $\boldsymbol{X}_i$. Under joint model (1), we can write the conditional distribution for imputing $X_i^{(t)}$ under MAR and LMAR as

$$f(X_i^{(t)}|\boldsymbol{X}_i^{(-t)}, \boldsymbol{Y}_i, L_i; \rho) \propto f(\boldsymbol{Y}_i|\boldsymbol{X}_i, L_i; \theta) f(L_i|\boldsymbol{X}_i; \omega) f(X_i^{(t)}|\boldsymbol{X}_i^{(-t)}; \psi) \tag{4}$$

Expressions in (3) and (4) provide the kernels of the distributions we can use to impute outcomes and covariates in $\boldsymbol{D}$. The kernels take the same form under MAR and LMAR, and they do not involve a model $R$ directly. As with the latent variable imputation, distributions (3) and (4) may only be known up to proportionality, requiring more advanced statistical methods to draw imputations.

### 3.2. Relaxing the modeling assumptions

The imputation distributions derived previously were developed assuming a fully-specified joint model as in (1), but often we will not want to specify such a joint model in practice. Specification of the joint model may be particularly difficult or restrictive in the setting with missingness in covariates of different types. Rather than specifying an explicit joint distribution, we propose imputing missing values using (2)–(4) to *inform* the distributions we use in practice either used an SMC imputation-type approach or a chained equations-type approach. In practice, the resulting conditional distributions for either method may not together correspond to a valid joint distribution for all the variables.

Following SMC imputation proposed in Bartlett et al. (2014), we may specify only the modeling components needed for each imputation. Imputation of missing values of $\boldsymbol{Y}$ using (3) requires a model for $\boldsymbol{Y}|\boldsymbol{X}, L$, and imputation of missing $L$ using (2) further requires a model for $L|\boldsymbol{X}$ and, under LMAR, a model for missingness. Imputation of missing covariate $X_i^{(t)}$ using (4) requires us to specify $f(X_i^{(t)}|\boldsymbol{X}_i^{(-t)}; \psi)$. This approach involves incorporating the outcome model structure (in this case, models for $\boldsymbol{Y}|\boldsymbol{X}, L$ and $L|\boldsymbol{X}$

(and possibly missingness)) to do the imputation, but we can avoid specifying $f(\boldsymbol{X}|\psi)$ by instead specifying $f(X_i^{(t)}|\boldsymbol{X}_i^{(-t)};\psi)$ for covariates with missingness using simple regression models. An additional appealing feature of SMC imputation is that it has additional flexibility over joint modeling in terms of imputation model specification, and it also involves imputing with a model that is congenial with the final analysis model. By uncongeniality, we mean that the imputation model and the final data analysis model are incompatible (Meng 1994). Since SMC imputation directly uses the final analysis model in the imputation procedure, it is attractive from a congeniality point of view.

Imputation using SMC imputation may be difficult when distributions are known only up to proportionality. An alternative, simpler chained equations imputation approach involves using (2)–(4) solely to define what predictors are needed for each imputation. Specifically, (2) suggests that some function of $\boldsymbol{Y}$, $\boldsymbol{X}$, and possibly $\boldsymbol{R}$ (under LMAR) should be used as predictors when imputing $L$. The expression in (3) suggests we need $\boldsymbol{X}$, $L$, and $\boldsymbol{Y}^{(-t)}$ when imputing $Y^{(t)}$, and (4) suggests we need $\boldsymbol{Y}$, $L$, and $\boldsymbol{X}^{(-t)}$ when imputing $X^{(t)}$. We can then perform imputation (by specifying a regression model for imputing each variable) using standard software for chained equations imputation (Raghunathan 2001; Van Buuren et al. 2006). Such an approach would allow for increased flexibility in model specification (for example, by including quadratic or interaction terms) while still allowing $L$ to be used in the imputation. We may view the *working* model actually used for imputation as an approximation to the tru conditional model as in (2)–(4). We recommend imputing $L$ using the kernel form in (2) if possible, and our proposed algorithm will use this method.

The imputation distributions, therefore, can be easily modified to accommodate settings without a joint distribution. Indeed, Gelman (2004) argues that "having a joint distribution in the imputation is less important than incorporating information from other variables and unique features of the dataset (e.g. zero/nonzero features in income components, bounds, skip patterns, nonlinearity, interactions)." The SMC imputation and chained equations approaches allow these unique features of the data to be directly incorporated in the imputation models. This approach allows for greater flexibility in the specification of the imputation distributions compared to joint modeling.

When we replace the true predictive distributions under a joint model with a *working* imputation model, the corresponding parameters may no longer correspond to the parameters under the joint model. In the next section, we will describe how we can perform imputation using these *working* imputation distributions in practice.

217 ### 3.3. Sequential imputation method

We propose a sequential imputation method in which each variable with missingness is imputed one-by-one in an iterative algorithm. At each step, we obtain a single imputation of a variable $V$ from the *working* posterior predictive distribution of $V$ (with parameter $v$) by 1) drawing $v$ from its posterior distribution and 2) drawing missing values of $V$ from its full conditional distribution at the drawn $v$.

Just before the imputation step for each variable, we draw the parameters necessary for the imputation from a current estimate of the parameters' *working* posterior predictive distribution. Let $X^{(t)}$ and $Y^{(t)}$ be defined as before. Let $\tilde{f}$ indicate a working distribution (usually a regression model) *used for imputation* that may not necessarily be equal to the distribution under a joint model. In the imputation step for each variable, we treat the most recent imputations of the other variables as observed. At each iteration, we draw missing data and parameters using one of the two following algorithms. An in-depth description and motivation for our proposed parameter draw methods is included in the Supplementary Materials. In describing how to perform the parameter draws, we assume flat priors for all parameters.

**SMC imputation algorithm:**

$$\begin{aligned} \text{Impute } L: \quad & [\theta,\omega] \sim f(\theta,\omega|\boldsymbol{D},L^{(\text{obs})}) \qquad \phi^{-S} \sim f(\phi^{-S}|\boldsymbol{D},L,\boldsymbol{R}^{-S}) \\ & L_i^{(\text{mis})} \propto f(\boldsymbol{R}_i^{-S}|\boldsymbol{Y}_i^{(\text{obs})},\boldsymbol{X}_{\boldsymbol{i}}^{(\textbf{obs})},L_i;\phi^{-S})f(\boldsymbol{Y}_i|\boldsymbol{X}_i,L_i;\theta)f(L_i|\boldsymbol{X}_i;\omega) \\ \text{Impute } Y^{(t)}: \quad & \theta \sim f(\theta|\boldsymbol{D},L) \qquad Y_i^{(t,\text{mis})} \propto f(\boldsymbol{Y}_i|\boldsymbol{X}_i,L_i;\theta) \\ \text{Impute } X^{(t)}: \quad & [\theta,\omega] \sim f(\theta,\omega|\boldsymbol{D},L) \qquad \tilde{\psi}_t \sim \tilde{f}(\tilde{\psi}_t|\boldsymbol{X}) \\ & X_i^{(t,\text{mis})} \propto f(\boldsymbol{Y}_i|\boldsymbol{X}_i,L_i;\theta)f(L_i|\boldsymbol{X}_i;\omega)\tilde{f}(X_i^{(t)}|\boldsymbol{X}_i^{(-t)};\tilde{\psi}_t) \end{aligned} \tag{5}$$

When imputing $L$, we can obtain a (approximate) draw $[\theta,\omega]$ by fitting our outcome model to a bootstrap sample of $[\boldsymbol{D},L^{(obs)}]$ using methods that treat $L$ as latent. For example, suppose our outcome model is a linear mixed model. We can obtain this draw by fitting a linear mixed model to a bootstrap sample of the data. We can obtain a draw of $\phi^{-S}$ by fitting a model for $\boldsymbol{R}^{(-\boldsymbol{S})}$ to a bootstrap sample of the most recent imputed data (including imputed $L$). When imputing $Y^{(t)}$, we can obtain a draw of $\theta$ by fitting a model for $\boldsymbol{Y}|\boldsymbol{X},L$ using a bootstrap sample of the most recently imputed data. When imputing $X^{(t)}$, we can obtain a draw of $\tilde{\psi}_t$ by fitting the corresponding model to a bootstrap sample of $\boldsymbol{X}$. In the Supplementary Materials, we provide details regarding how we can perform each of the imputation steps for the examples discussed Section 2.

**Chained equations imputation algorithm:**

$$\begin{aligned}\text{Impute } L: \quad & [\theta,\omega] \sim f(\theta,\omega|\boldsymbol{D},L^{(\mathrm{obs})}) \qquad \phi^{-S} \sim f(\phi^{-S}|\boldsymbol{D},L,\boldsymbol{R}^{-S}) \\ & L_i^{(\mathrm{mis})} \propto f(\boldsymbol{R}_i^{-S}|\boldsymbol{Y}_i^{(\mathrm{obs})},\boldsymbol{X}_i^{(\mathrm{obs})},L_i;\phi^{-S})f(\boldsymbol{Y}_i|\boldsymbol{X}_i,L_i;\theta)f(L_i|\boldsymbol{X}_i;\omega) \\ \text{Impute } Y^{(t)}: \quad & \tilde{\theta}_t \sim \tilde{f}(\tilde{\theta}_t|\boldsymbol{D},L) \qquad Y_i^{(t,\mathrm{mis})} \propto \tilde{f}(\boldsymbol{Y_i}|\boldsymbol{X}_i,L_i;\tilde{\theta}_t) \\ \text{Impute } X^{(t)}: \quad & \tilde{\psi}_t \sim \tilde{f}(\tilde{\psi}_t|\boldsymbol{X},\boldsymbol{Y},L) \qquad X_i^{(t,\mathrm{mis})} \propto \tilde{f}(X_i^{(t)}|\boldsymbol{X}_i^{(-t)},\boldsymbol{Y}_i,L_i;\tilde{\psi}_t)\end{aligned} \tag{6}$$

We can impute $L$ as before. When imputing $\boldsymbol{Y}$ and $\boldsymbol{X}$, we draw the parameters of interest by fitting corresponding models to bootstrap versions of the most recently imputed data.

Iteration of the above algorithms is required even if we have only one variable in $\boldsymbol{D}$ with missing values. We can ignore the imputation steps for each fully observed variable. We initialize the missing values for each variable in $\boldsymbol{D}$ by drawing from the observed values with equal probability. We can initialize missing $L$ using the distribution $f(L|\boldsymbol{X})$ obtained from a fit to the data with fully observed $\boldsymbol{D}$ (using methods that treat $L$ as latent).

For both of the above algorithms, we assume that missingness is LMAR. Suppose instead that we know that missingness is MAR. We can apply the above algorithms but using that $f(L_i|\boldsymbol{X}_i,\boldsymbol{Y}_i,\boldsymbol{R}_i^{-S};\nu) \propto f(\boldsymbol{Y}_i|\boldsymbol{X}_i,L_i;\theta)f(L_i|\boldsymbol{X}_i;\omega)$ instead to impute $L_i$ and without drawing values for $\phi^{-S}$. In this way, the above development also gives us an imputation algorithm for dealing with missing data for models with latent variables under MAR.

We perform the imputation procedure $m$ times to construct $m$ filled-in datasets (with $m$ different initializations). We then estimate $\rho$ by fitting our model of interest to each of the imputed datasets *ignoring* $\boldsymbol{R}$. When we perform this analysis, we may choose to use only imputed $\boldsymbol{D}$, only imputed $L$, or both. We can then use Rubin's combining rules to obtain a single set of parameter estimates and errors from which we make the desired inference (Rubin 1987).

It is important to consider the impact of ignoring $\boldsymbol{R}$ for each one of these final analysis strategies. Harel & Schafer (2009) shows that when imputed $L$ is included in the final analysis, we can ignore $\boldsymbol{R}$. This result holds true under MAR and LMAR and whether or not imputed $\boldsymbol{D}$ is included in the final analysis. In *Properties 4–5* in the Supplementary Materials, we explore the ignorability of $\boldsymbol{R}$ when performing a final analysis using only the imputed $\boldsymbol{D}$. We show that $\boldsymbol{R}$ is ignorable under MAR and that such an analysis ignoring $\boldsymbol{R}$ under LMAR is valid but not fully efficient. Even with a potential loss of efficiency, we may still choose to perform our final analysis ignoring the imputed $L$ as this may provide improved numerical stability of the algorithm and more robustness to misspecification of the imputation models, and we may have little loss of efficiency in practice.

## 4. Identifiability and convergence

As with all missing data methods involving MNAR assumptions, one big concern is how to model the missingness mechanism (which will be unverifiable) (Molenberghs, Beunckens & Sotto 2008). Another concern is whether the resulting model parameters are identifiable (Little 1995). Even when the parameters are technically identified, weak identifiability may also have implications on the numerical convergence of the proposed imputation algorithm. In this section, we briefly comment on some identifiability- and convergence-related issues that arise in the application of the proposed imputation algorithm.

### 4.1. Modeling the missingness mechanism

Under LMAR, we must specify a model for $\boldsymbol{R^D}$ (or some subset $\boldsymbol{R}^{-S}$ following *Property 3*). While we can conceive of many different models for $\boldsymbol{R^D}$, the model parameter $\nu = (\phi, \rho)$ may not always be identifiable. In some specific settings (e.g. Wu & Carroll 1988; Miao, Ding & Geng 2016), identifiability has been demonstrated analytically, but exploring identifiability can be difficult in general. Wang, Shao & kwang Kim (2014) relates identifiability to the existence of instrumental variables. We explore identifiability in several particular modeling settings in the Supplementary Materials. In this paper, we will not attempt to prove identifiability properties for general LMAR mechanisms. Instead, we will provide some guidance for applying the proposed methods in the presence of possible identifiability issues.

In order to reduce the potential for identifiability issues, many authors (e.g. Little 2009b) recommend that we avoid overburdening the missingness model with extra variables. However, if we leave out variables that should be in the model, we may introduce bias in estimating the parameter of interest as seen in our simulations. In our simulations, imputation with LMAR *outcome* missingness tended to be more susceptible to identifiability problems than *covariate* missingness. Some authors recommend performing a sensitivity analysis in which we specify the form of the missingness model and carry out analysis using fixed values for $\phi^D$ (e.g. Little 2009b). We can then perform the desired analysis many times using different values for $\phi^D$. This approach allows us to directly study the impact of $\phi^D$ on inference and avoid estimating the parameters of the missingness model. Additionally, MNAR missingness mechanisms are known to be particularly sensitive to assumptions about the structure of the missingness mechanism, and we could perform a sensitivity analysis using different missingness model structures (Little 1995). We take this approach in our head and neck cancer example. These sensitivity approaches allow the proposed methods to be applied while avoiding some of the pitfalls of MNAR settings.

## 4.2. A note on convergence

When the conditional models used for imputation correspond to a well-defined joint distribution with identified parameters, our imputation algorithm is expected to converge to draws of the joint posterior distribution for the missing data (Liu et al. 2013; Hughes et al. 2014; Bartlett et al. 2014). When the imputation models do not correspond to a valid joint distribution (called incompatibility), our imputation method is not guaranteed to converge. However, several works have demonstrated that we can often still obtain good inference under incompatible imputation models (Van Buuren et al. 2006; Van Buuren 2007).

We will not attempt to prove convergence or consistency properties for the proposed algorithm beyond what exists in the chained equations and SMC imputation literature. Instead, we will use simulation and some minor analytical exploration to identify settings that may be particularly susceptible to concerns about convergence. In particular, identifiability concerns related to the missingness model have implications on the convergence of the algorithm. When parameters are not identifiable (in terms of the observed data likelihood having a unique maximiser), we may not expect the imputation algorithm to converge properly. Even when the parameters are all identifiable, we may run into numerical issues if the observed data likelihood is nearly flat. These issues appear to be of greater concern for outcome missingness. We note that in our experience, even when we have numerical convergence issues for $\phi$ (missingness model) and $\omega$ (model for $L|\boldsymbol{X}$), the draws for $\theta$ (model for $\boldsymbol{Y}|\boldsymbol{X}, L$) may still converge to reasonable values. In such cases, the identifiability-related numerical problems may not strongly impact the draws for the primary parameter of interest, $\theta$. It is important to monitor the convergence of all model parameters, and we may still be able to make inference about $\theta$ in the presence of some mild identifiability-related convergence issues for $\phi$. We explore identifiability-related convergence issues further in Section 5 and the Supplementary Materials.

# 5. Simulation study

In this section, we present a simulation study with four parts. In the first three parts, we evaluated how the proposed algorithm performs in terms of bias, empirical variance, and coverage for outcome model parameters in linear mixed models (Simulation 1), CPH cure models (Simulation 2), and normal mixture models (Simulation 3). In Simulation 4, we explored convergence under a variety of modeling scenarios. Details can be found in the Supplementary Materials. A fifth/sixth set of simulations included in the Supplementary Materials (1) explored the impact of including or ignoring the imputed $L$ in the final analysis and (2) assessed the impact of ignoring latent-dependent missingness in the CPH cure model setting in more detail, but we will not discuss these simulations further here. Unless otherwise specified, imputations were drawn using the SMC imputation method rather than the chained equations method.

### 5.1. Simulations 1–3: exploring bias, variance, and coverage

In Simulation 1, we simulated 1500 datasets with 500 subjects each under a linear mixed model with a random intercept. Each dataset contained two binary covariates, $X_1$ and $X_2$. We drew random intercept $b_i \sim N(0,1)$ for each individual and generated $\boldsymbol{Y}$ for each individual at each of three time-points using the model $Y_{ij} = \beta_{\text{Intercept}} + \beta_{X_1} X_{i1} + \beta_{X_2} X_{i2} + \beta_{\text{Time}} \text{Time}_{ij} + b_i + e_{ij}$ for $j = 1, 2, 3$ with independent $N(0,1)$ errors, $\beta_{\text{Intercept}} = \beta_{X_1} = \beta_{X_2} = 0.5$, and $\beta_{\text{Time}} = 0.2$. Additional simulation details are available in the Supplementary Materials. We imposed $\sim 50\%$ missingness in $X_2$ under four different mechanisms: (A) MAR dependent on baseline outcome value $Y_1$, (B) LMAR with moderate dependence between missingness and the random intercept, (C) LMAR with strong dependence on the random intercept, and (D) LMAR with dependence on the random intercept and the baseline outcome value $Y_1$.

We then imputed values of $X_2$ and $b$ using methods discussed in this paper under various working models. When we imputed under a LMAR working model, we modeled the covariate missingness indicator $R_i^D$ using a logistic regression with different functions of $b$, $X_1$, and $\boldsymbol{Y}$ as predictors. When we assumed MAR, we imputed $L$ ignoring the missingness mechanism. For each simulated dataset, we created 10 imputed datasets. We then fit a linear mixed model to each of the imputed datasets and use Rubin's rules to obtain a single set of parameter estimates and their corresponding variances for each simulation. We then computed the bias, empirical variance, and coverage rates across the 1500 simulations. To improve readability, we list coverage rates in Table S1 in the Supplementary Materials. We note that the APPROX simulations take a chained equations imputation approach in which we impute $X_2$ conditional on $X_1$, $L$ and $\boldsymbol{Y}$ using a logistic regression form, so the imputation distributions for $X_2$ and $L$ in this case do not correspond to a coherent joint distribution.

Table 1 shows the results for Simulation 1. Simulations 2–3 included in the Supplementary Materials are similar. Simulations 1–3 generally demonstrated that the proposed imputation approach can result in essentially unbiased estimates of outcome model parameters with nominal (or perhaps slightly conservative) coverage when the working missingness model contains the true model. We demonstrated that complete case analysis and imputation assuming MAR can sometimes result in biased parameter estimates when missingness is at least moderately associated with the latent variable. The bias created by incorrectly assuming MAR appears larger when $L$ is a fully latent compared to partially

Table 1. Linear mixed model estimates using proposed imputation methods.

| Method | Contains Truth# | Intercept Bias (Var) † | $X_1$ Bias (Var) | $X_2$ Bias (Var) | Time Bias (Var) |
|---|---|---|---|---|---|
| Full Data | - | 0 (1.2) | 0 (1.0) | 0 (1.1) | 0 (0.10) |
| Missingness in $X_2$ dependent on $Y_1$ and independent of $b$ (Mechanism A) | | | | | |
| Complete Case | - | −78 (2.0) | −9 (1.8) | −9 (1.9) | 19 (0.20) |
| MAR Imputation | Yes | 0 (1.8) | 0 (1.1) | 0 (2.8) | 0 (0.10) |
| LMAR Imputation: $b$* | No | 6 (1.4) | 2 (1.1) | −9 (1.9) | 0 (0.10) |
| LMAR Imputation: $b, X_1, b \times X_1$ | No | 6 (1.4) | 1 (1.1) | −9 (2.0) | 0 (0.10) |
| LMAR Imputation: $b, Y_1$ | Yes | 0 (1.8) | 0 (1.1) | 0 (2.8) | 0 (0.10) |
| LMAR Imputation: $\mathbb{I}(b > 0), Y_1$ | Yes | 0 (1.9) | 0 (1.1) | 0 (2.8) | 0 (0.10) |
| LMAR Imputation: $b, X_1, b \times X_1, Y_1$ | Yes | 0 (1.9) | 0 (1.1) | 0 (2.8) | 0 (0.10) |
| LMAR Imputation: $b, Y_2$ | No | 7 (1.4) | 2 (1.1) | −11 (1.8) | 0 (0.10) |
| MAR APPROX Imputation | Yes | −1 (1.9) | 0 (1.1) | 0 (3.0) | 0 (0.10) |
| LMAR APPROX Imputation: $b$ | No | 5 (1.5) | 1 (1.1) | −8 (2.1) | 0 (0.10) |
| Missingness in $X_2$ moderately dependent on $b$ (Mechanism B) | | | | | |
| Complete Case | - | −24 (2.4) | 0 (2.1) | 0 (2.2) | 0 (0.19) |
| MAR Imputation | No | −2 (1.7) | 0 (1.1) | 2 (2.4) | 0 (0.10) |
| LMAR Imputation: $b$ | Yes | 0 (1.6) | 0 (1.1) | 0 (2.2) | 0 (0.10) |
| LMAR Imputation: $b, X_1, b \times X_1$ | Yes | 0 (1.6) | 0 (1.1) | 0 (2.2) | 0 (0.10) |
| LMAR Imputation: $b, Y_1$ | Yes | 0 (1.6) | 0 (1.1) | 0 (2.2) | 0 (0.10) |
| LMAR Imputation: $\mathbb{I}(b > 0), Y_1$ | No | 0 (1.6) | 0 (1.1) | 0 (2.2) | 0 (0.10) |
| LMAR Imputation: $b, X_1, b \times X_1, Y_1$ | Yes | 0 (1.6) | 0 (1.1) | 0 (2.2) | 0 (0.10) |
| MAR APPROX Imputation | No | −3 (1.7) | 0 (1.1) | 3 (2.4) | 0 (0.10) |
| LMAR APPROX Imputation: $b$ | Yes | 0 (1.6) | 0 (1.1) | 0 (2.2) | 0 (0.10) |
| Missingness in $X_2$ strongly dependent on $b$ (Mechanism C) | | | | | |
| Complete Case | - | −48 (2.5) | 0 (1.8) | 0 (2.0) | 0 (0.22) |
| MAR Imputation | No | −7 (2.0) | 0 (1.1) | 8 (2.8) | 0 (0.10) |
| LMAR Imputation: $b$ | Yes | 0 (1.5) | 0 (1.1) | 0 (2.0) | 0 (0.10) |
| LMAR Imputation: $b, X_1, b \times X_1$ | Yes | 0 (1.5) | 0 (1.1) | 0 (2.0) | 0 (0.10) |
| LMAR Imputation: $b, Y_1$ | Yes | 0 (1.6) | 0 (1.1) | 0 (2.1) | 0 (0.10) |
| LMAR Imputation: $\mathbb{I}(b > 0), Y_1$ | No | 0 (1.6) | 0 (1.1) | 0 (2.1) | 0 (0.10) |
| LMAR Imputation: $b, X_1, b \times X_1, Y_1$ | Yes | 0 (1.5) | 0 (1.1) | 0 (2.1) | 0 (0.10) |
| MAR APPROX Imputation | No | −8 (2.0) | 0 (1.1) | 9 (2.8) | 0 (0.10) |
| LMAR APPROX Imputation: $b$ | Yes | 0 (1.5) | 0 (1.1) | 0 (2.1) | 0 (0.10) |
| Missingness in $X_2$ dependent on $b$ and $Y_1$ (Mechanism D) | | | | | |
| Complete Case | - | −73 (2.0) | −5 (1.6) | −5 (1.6) | 8 (0.21) |
| MAR Imputation | No | −8 (2.0) | −1 (1.1) | 10 (2.8) | 0 (0.10) |
| LMAR Imputation: $b$ | No | 3 (1.4) | 0 (1.1) | −5 (1.7) | 0 (0.10) |
| LMAR Imputation: $b, X_1, b \times X_1$ | No | 3 (1.4) | 0 (1.1) | −5 (1.6) | 0 (0.10) |
| LMAR Imputation: $b, Y_1$ | Yes | 0 (1.5) | 0 (1.1) | 0 (2.0) | 0 (0.10) |
| LMAR Imputation: $\mathbb{I}(b > 0), Y_1$ | No | 0 (1.6) | 0 (1.1) | 0 (2.0) | 0 (0.10) |
| LMAR Imputation: $b, X_1, b \times X_1, Y_1$ | Yes | 0 (1.6) | 0 (1.1) | 0 (2.0) | 0 (0.10) |
| LMAR Imputation: $b, Y_2$ | No | 3 (1.4) | 0 (1.1) | −6 (1.7) | 0 (0.10) |
| MAR APPROX Imputation | No | −9 (2.1) | −1 (1.1) | 11 (2.9) | 0 (0.10) |
| LMAR APPROX Imputation: $b$ | No | 3 (1.4) | 0 (1.1) | −4 (1.7) | 0 (0.10) |

*Variables after colon represent linear predictors in working model for $R_i^D$

† All values in table multiplied by 100. Var indicates empirical variance.

# Indicates whether working missingness model contains true model.

APPROX: Imputation of $X_2$ uses a logistic regression with predictors $X_1, b, Y_1, Y_2, Y_3$ (instead of kernel (4))

Complete Case: Analysis excluding subjects with missing $X_2$

latent. Imputation under LMAR assumptions can correct this bias when we use a working model containing the truth and can sometimes reduce the bias compared to imputation assuming MAR when the working model is *close* to the truth. When missingness was truly MAR, simulations suggested that imputation under a LMAR model that did not contain the true model can create bias. However, simulations showed that LMAR methods with working models containing the true MAR model can still be applied with little or no loss of efficiency (when the LMAR model is well-identified) in this setting. Very complicated working missingness models can sometimes result in a loss of efficiency, but this loss was generally small.

### 5.2. Simulation 4: exploring identifiability and convergence

Even if the model parameters are technically identifiable, one additional concern is that the likelihood surface near the maximizer may be nearly flat, which can lead to issues with model fitting and convergence of the imputation algorithm. In order to better understand possible identifiability-related convergence issues, we performed a set of simulations evaluating convergence of the imputation algorithm under a variety of modeling scenarios.

We simulated data under a linear mixed model, cure model, and mixture of normals respectively as described in the Supplementary Materials. We imposed $\sim 50\%$ covariate or outcome missingness (but not both) using MAR or LMAR mechanisms. For each simulated dataset, we performed imputation using a correct working missingness model structure. For each outcome model parameter, we evaluated parameter convergence using the Gelman-Rubin convergence statistic (Gelman & Rubin 1992).

Simulations demonstrated good convergence properties under LMAR/MAR covariate and MAR outcome missingness. Under LMAR outcome missingness, the outcome model parameters appeared to converge, but missingness model parameter (in particular, for the latent variable) showed some evidence of convergence problems. The drawn values of the outcome model parameters appeared reasonable (with small or no bias) even when the missingness model parameters do not converge, but this may not be true in general. When we fixed the value of the parameter related to the latent variable in the missingness model, we saw a large improvement in the convergence properties of the imputation algorithm.

## 6. Application to head and neck cancer data

We consider data from a cohort study of $N$=1226 patients treated for head and neck squamous cell carcinoma (HNSCC). This study was conducted by the University of Michigan Head and Neck Specialized Program of Research Excellence (SPORE) and followed patients who were treated at the University of Michigan Cancer Center for HNSCC between Nov. 2003 and July 2013. Details about this study can be found in Duffy et al. (2008) and

Peterson et al. (2016). After treatment, patients were followed for recurrence. Covariate information was also collected at baseline. We are interested in studying the association between covariates and the time to HNSCC recurrence after treatment. For head and neck cancer, it has been established that some patients can be cured by treatment, and these patients will never experience a recurrence (Taylor 1995). We model the time to HNSCC recurrence using a Cox proportional hazards cure model.

HPV status was unavailable for $55.8\%$ of the subjects, and small amounts of missingness was present in other study variables. Beesley et al. (2016) explores imputation-based approaches for dealing with the missing covariate data for this study. The analysis in Beesley et al. (2016), however, assumes that covariate missingness is MAR and does not depend on underlying cure status. An induced LMAR association between missingness in HPV status and cure status (denoted $G$) could occur if HPV missingness is related to an unmeasured variable that is also related to the cure probability. In this study, the HPV missingness rate is related to calendar time (in a nonlinear way), and calendar time may be associated with the cure rate. Additionally, a more experienced doctor may be more likely to recommend HPV testing and to have cured patients. We cannot control for this effect due to a lack of detailed information about treating doctors for each patient. Given the large rate of missingness in HPV status, we are interested to explore the robustness of model inference to our assumptions about the missingness mechanism.

We are interested in comparing model inference assuming MAR to inference obtained when missingness in HPV is assumed to be LMAR. We assume missingness in all other variables is MAR. We consider three working assumptions for HPV status missingness: (A) MAR, (B) missingness dependent only on cure status, and (C) missingness dependent on cure status, age at diagnosis, cancer site, and (grouped) enrollment year. Assumptions (B) and (C) are modeled using logistic regression.

We apply our proposed SMC imputation method to impute the missing data. In this setting, $G$ is the partially latent cure status, $Y$ is the censored event time data (time and indicator), and $\boldsymbol{X}$ is the set of covariates. Here, the model $Y|G = 1, \boldsymbol{X}$ is a Cox regression and the model for $G|\boldsymbol{X}$ is a logistic regression. We impute cure status $G$ using (2). As suggested in Beesley et al. (2016), we impute missing values of each $p$th covariate $X^{(p)}$ using a standard regression model with $\boldsymbol{X}^{(-p)}$, $G$, $G \times \hat{H}_0(T)$, and $G \times \hat{H}_0(T) \times \boldsymbol{X}^{(-p)}$ as predictors. Here, $\hat{H}_0(T)$ is an estimate of the cumulative baseline hazard of having an event in the non-cured group. As in Beesley et al. (2016), we will draw values for the regression model's parameter without conditioning on the imputed $X^{(p)}$ (as is done in usual chained equations). Variables included in $X^{(p)}$ for the imputation include log-transformed number of sexual partners, PNI, comorbidities, smoking habits, alcohol use, age, cancer site, cancer stage, gender, and enrollment period (2003-2008, 2009-2011, 2012-2013).

Table 2 presents the cure model fit under different assumptions about the missingness mechanism. We see that the fits are nearly identical. The largest difference between the fits is in the estimate for the HPV effect on the time to recurrence in the non-cured group. We estimate a slightly stronger effect of HPV status under LMAR assumptions than under MAR assumptions, and the strongest effect is estimated when missingness is assumed to be LMAR dependent on $G$ and other covariates. However, the HPV effect is not significant in any of the fits. We cannot make conclusions about the correct missingness mechanism, but regardless of the true missingness model, the CPH cure model inference appears to be very robust to different specifications of the working missingness model.

Table 2. Cure model fits to head and neck data under different missing model assumptions.

| Missingness model: | MAR* | LMAR1* | LMAR2* |
|---|---|---|---|
| | Logistic regression, odds ratio (95% CI) | | |
| Age/10 | 1.14 (1.00, 1.31)† | 1.14 (0.99, 1.32) | 1.14 (1.00, 1.30)† |
| Cancer Stage | | | |
| I/Cis (ref) | | | |
| II | 1.25 (0.57, 2.74) | 1.25 (0.54, 2.89) | 1.25 (0.57, 2.74) |
| III | 2.36 (1.18, 4.72)† | 2.32 (1.16, 4.61)† | 2.33 (1.18, 4.63)† |
| IV | 3.32 (1.74, 6.33)† | 3.30 (1.74, 6.26)† | 3.30 (1.80, 6.03)† |
| Cigarette Use | | | |
| Never (ref) | | | |
| Current | 1.46 (0.97, 2.18) | 1.47 (0.96, 2.24) | 1.46 (0.98, 2.16) |
| Former | 1.27 (0.85, 1.90) | 1.28 (0.85, 1.93) | 1.28 (0.84, 1.95) |
| HPV Status | | | |
| Negative (ref) | | | |
| Positive | 0.34 (0.19, 0.58)† | 0.35 (0.19, 0.64)† | 0.34 (0.20, 0.56)† |
| Comorbidities | | | |
| None (ref) | | | |
| Mild | 1.14 (0.77, 1.69) | 1.15 (0.79, 1.68) | 1.15 (0.79, 1.68) |
| Moderate | 1.66 (1.08, 2.56)† | 1.66 (1.07, 2.58)† | 1.66 (1.07, 2.55)† |
| Severe | 1.94 (1.10, 3.43)† | 1.94 (1.08, 3.48)† | 1.97 (1.08, 3.57)† |
| Cancer Site | | | |
| Larynx (ref) | | | |
| Hypopharynx | 1.93 (0.88, 4.22) | 1.93 (0.86, 4.30) | 1.99 (0.91, 4.33) |
| Oral Cavity | 1.24 (0.81, 1.90) | 1.24 (0.81, 1.89) | 1.24 (0.81, 1.90) |
| Oropharynx | 1.68 (0.94, 3.02) | 1.64 (0.90, 2.97) | 1.68 (0.95, 2.96) |
| | Cox proportional hazards, hazard ratio (95% CI) | | |
| Age/10 | 1.08 (0.98, 1.19) | 1.08 (0.98, 1.18) | 1.08 (0.98, 1.19) |
| Cancer Stage | | | |
| I/Cis (ref) | | | |
| II | 1.67 (0.70, 3.95) | 1.62 (0.69, 3.82) | 1.61 (0.66, 3.88) |
| III | 2.42 (1.22, 4.79)† | 2.40 (1.24, 4.66)† | 2.42 (1.21, 4.84)† |
| IV | 2.76 (1.48, 5.16)† | 2.76 (1.47, 5.18)† | 2.77 (1.45, 5.29)† |
| Cigarette Use | | | |
| Never (ref) | | | |
| Current | 0.98 (0.70, 1.38) | 0.97 (0.70, 1.35) | 0.97 (0.70, 1.33) |
| Former | 0.94 (0.66, 1.33) | 0.94 (0.67, 1.32) | 0.94 (0.67, 1.32) |
| HPV Status | | | |
| Negative (ref) | | | |
| Positive | 0.91 (0.55, 1.48) | 0.85 (0.51, 1.40) | 0.81 (0.52, 1.28) |
| Comorbidities | | | |
| None (ref) | | | |
| Mild | 0.89 (0.65, 1.23) | 0.89 (0.65, 1.22) | 0.89 (0.65, 1.22) |
| Moderate | 1.10 (0.75, 1.61) | 1.09 (0.73, 1.61) | 1.09 (0.75, 1.58) |
| Severe | 1.07 (0.63, 1.80) | 1.06 (0.64, 1.74) | 1.06 (0.63, 1.80) |
| Cancer Site | | | |
| Larynx (ref) | | | |
| Hypopharynx | 1.43 (0.77, 2.67) | 1.42 (0.78, 2.60) | 1.42 (0.78, 2.58) |
| Oral Cavity | 1.33 (0.90, 1.97) | 1.32 (0.92, 1.90) | 1.32 (0.92, 1.89) |
| Oropharynx | 1.02 (0.62, 1.68) | 1.06 (0.66, 1.70) | 1.09 (0.69, 1.72) |

*Corresponds to working model for Prob(HPV missing). LMAR1 includes $G$ only. LMAR2 includes $G$ and covariates includes main effects for cancer site, age, and enrollment year group.
† Significant at p = 0.05

## 7. Discussion

We present a novel sequential imputation algorithm that can handle both missing at random (MAR) and latent missing at random (LMAR) covariate and outcome missingness for models with latent or partially latent variables. Unlike existing methods, the proposed approach does not require specification of the full joint distribution of the complete data. The proposed algorithm imputes the latent variable as part of the missing data, allowing the latent variable to be directly used to help impute the other variables.

We first consider the more restrictive setting where the joint model is fully specified. We use results under a joint model to *inform the structure* of the imputation distributions and the method for drawing parameters in the proposed algorithm *without requiring specification of the joint model.* The proposed approach is very flexible and can accommodate either a chained equations-type approach to imputation or a substantive model compatible (SMC) imputation approach that is more strongly informed by the outcome model.

Several authors have previously proposed approaches for handling latent ignorable missingness in specific joint modeling settings (Jung 2007; Yang, Lu & Shoptaw 2008; Lu, Zhang & Lubke 2011), and Harel (2003) proposes a non-iterative imputation approach for dealing with general latent-dependent missingness under a joint model. These methods, however, all rely on the prior specification of a joint model for the complete data. In practice, however, such a joint model may be difficult or too restrictive. Therefore, there is a need to consider methods for imputing variables under latent ignorability that incorporate less restrictive assumptions about the joint model.

Therefore, we consider two departures for joint model-based imputation: SMC imputation and chained equations imputation. It is worth noting the distinction between the SMC imputation method and joint modeling. The primary distinction in our setting is in the specification (or lack thereof) of the joint distribution for $\boldsymbol{X}$. The imputation distributions for $L$ and $\boldsymbol{Y}$ are similar to the distributions obtained under a joint model. However, the ability to avoid specifying the joint distribution of the covariates provides a large advantage in terms of modeling—the covariate distribution is the hard one to specify. We often have many covariates of different types and with different restrictions, and specification of a valid joint distribution can be very challenging. Therefore, replacing the need to specify the joint distribution of $\boldsymbol{X}$ with specification of the conditional distribution for only the variables with missingness does present a clear advantage over joint modeling in many settings, and the statistical properties of the resulting algorithm can be quite different. This motivates a separate treatment of SMC imputation from joint modeling. The proposed chained equations imputation method, where only the latest variable is imputed using assumptions about the outcome model, takes an additional step away from joint modeling; the other variables are imputed using regression models specified separately for each variable with missingness. It is worth noting that, while the proposed methods can be applied under MAR or under LMAR missingness assumptions, the primary novelty lies in handling imputation under LMAR. We are not aware of any literature developing SMC imputation or chained equations imputation methods to handle latent ignorable missingness for a *general class* of latent variable models.

Simulations demonstrate that the proposed methods can result in good performance (in terms of bias, coverage, etc) under a variety of modeling scenarios as long as the working missingness model contains the true model. In practice, we will not know the true missingness model. Preliminary simulations in the LMAR setting suggest that this may not always be a problem as long as we posit a working missingness is somewhat *close* to the true model. Suppose missingness is truly LMAR. We demonstrate that imputation incorrectly assuming MAR can result in biased outcome model parameter estimates, and the proposed approach using LMAR assumptions can correct or reduce this bias. Suppose instead that missingness is truly MAR. Simulations demonstrate that imputation under LMAR can produce good results as long as the working model contains the true MAR mechanism. Since associations between missingness and fully observed variables can be directly explored using the observed data, we can often identify observed factors related to sampling to construct a good working model structure for LMAR-based imputation.

Additional simulations explore the numerical convergence properties of the proposed SMC imputation algorithm. We do not see evidence of convergence issues under MAR outcome missingness or MAR/LMAR covariate missingness except in the case where the working missingness model contains many highly correlated predictors. In some scenarios, we see convergence issues when we have LMAR outcome missingness, and parameters of the missingness model were particularly susceptible. Convergence problems can be substantially reduced by fixing parameters related to the latent variable in the missingness model.

We apply the imputation approach to a motivating study of head and neck cancer recurrence. We impute missing values under MAR and LMAR assumptions, and the resulting model fits are very similar. In this application, the model inference is robust to the assumptions about missingness. We also see this phenomenon in the simulations based on the cure model, suggesting that the cure model in particular may be fairly robust to MAR assumptions under cure status-dependent missingness. This issue is discussed in more detail in the Supplementary Materials (Simulation 6). We may be generally less concerned about accounting for latent-dependent missingness in the cure model setting, where the latent variable is always partially observed.

One criticism of methods that do not assume a fully-specified joint distribution is that the algorithm is not guaranteed to converge to draws from a valid joint posterior predictive distribution for the missing values (Van Buuren et al. 2006). Our proposed imputation approach is similarly not guaranteed to converge to a valid joint distribution in general, and convergence can be impacted by identifiability issues. In this paper, we do not prove convergence properties for the proposed algorithm beyond existing properties in the SMC imputation and chained equations literature (Bartlett et al. 2014, e.g.). Instead, we use simulation to identify settings that may be particularly susceptible to concerns about convergence. We demonstrate that the convergence of the proposed algorithm can be impacted by parameter identifiability. Care should be taken to monitor algorithm convergence, particularly in the setting of LMAR outcome missingness or with working missingness models containing many predictors. We similarly do not prove identifiability properties for general LMAR mechanisms. In some settings (e.g. Wu & Carroll 1988; Miao, Ding & Geng 2016), identifiability has been demonstrated analytically, but exploring identifiability can be difficult in general. We view proofs of identifiability for general LMAR mechanisms to be outside the scope of this work. Instead, we provide some guidance for applying the proposed methods in the presence of possible identifiability issues.

The proposed methods can be applied under MAR and LMAR outcome/covariate missingness. Unlike usual MAR-based imputation, the proposed imputation approach requires us to model the data missingness mechanism when missingness is *assumed to be* LMAR. However, this direct dependence on the missingness model provides a convenient framework for studying the sensitivity of outcome model inference to different assumptions about the missingness mechanism (Little 1995; Molenberghs, Beunckens & Sotto 2008). Additionally, we propose an imputation procedure when missingness is *assumed to be* MAR, but this approach is similar to other methods existing in the literature that do not require a joint model. Simulations suggest that the proposed LMAR-based imputation approach can be applied even in MAR settings as long as the working missingness model contains or is close to the true model and the LMAR-based model is well-identified. Since associations between missingness and observed variables can be readily evaluated using observed data, we may often be able to construct a reasonable working missingness model allowing for additional dependence on the latent variable. The proposed method allows us to incorporate the outcome model directly into the imputation of the latent variable (and possibly missing covariate/outcome values), potentially resulting in improved imputations and reduced bias in the downstream analysis compared to usual chained equations. Our proposed method, therefore, provides a flexible and novel generalization of the usual MAR-based imputation that allows us to study a wider class of missingness models, of which MAR is a special case.

## *References*


BARTLETT, J.W., SEAMAN, S.R., WHITE, I.R. & CARPENTER, J.R. (2014). Multiple imputation of covariates by fully conditional specification: accomodating the substantive model. *Statistical Methods in Medical Research* **24**, 462–487.

BEESLEY, L.J., BARTLETT, J.W., WOLF, G.T. & TAYLOR, J.M.G. (2016). Multiple imputation of missing covariates for the Cox proportional hazards cure model. *Statistics in Medicine* **35**, 4701–4717.

CHUNG, H., FLAHERTY, B.P. & SCHAFER, J.L. (2006). Latent class logistic regression: application to marijuana use and attitudes among high school seniors. *Journal of the Royal Statistical Society* **169**, 723–743.

DUFFY, S., TAYLOR, J.M.G., TERRELL, J., ISLAM, M., YUAN, Z., FOWLER, K., WOLF, G. & TEKNOS, T. (2008). IL-6 predicts recurrence among head and neck cancer patients. *Cancer* **113**, 750–757.

FOLLMANN, D. & WU, M.C. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51**, 151–168.

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

GELMAN, A. (2004). Parameterization and bayesian modeling. *Journal of the American Statistical Association* **99**, 537–545.

GELMAN, A. & RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–511.

GIUSTI, C. & LITTLE, R.J.A. (2011). An analysis of nonignorable nonresponse to income in a survey with a rotating panel design. *Journal of Official Statistics* **27**, 211–229.

HAREL, O. (2003). Strategies for data analysis with two types of missing values. Ph.D. thesis, Pennsylvania State University.

HAREL, O. & SCHAFER, J.L. (2009). Partial and latent ignorability in missing-data problems. *Biometrika* **96**, 37–50.

HUGHES, R.A., WHITE, I.R., SEAMAN, S.R., CARPENTER, J.R., TILLING, K. & STERNE, J.A.C. (2014). Joint modeling rationale for chained equations. *BMC Medical Research Methodology* **14**, 1–10.

JUNG, H. (2007). A latent-class selection model for nonignorable missing data. Ph.D. thesis, Pennsylvania State University.

LITTLE, R.J. (2009a). Comments on : Missing data methods in longitudinal studies : a review. *Test* **18**, 47–50.

LITTLE, R.J. (2009b). Selection and pattern-mixture models. In *Longitudinal Data Analysis*, eds. G. Fitzmaurice, M. Davidian, G. Verbeke & G. Molenberghs, chap. 18. New York, NY: Taylor & Francis Group, pp. 409–431.

LITTLE, R.J.A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* **90**, 1112–1121.

LITTLE, R.J.A. & RUBIN, D.B. (2002). *Statistical analysis with missing data.* Hoboken, NJ: John Wiley and Sons, Inc, 2nd edn.

LIU, J., GELMAN, A., HILL, J., SU, Y.S. & KROPKO, J. (2013). On the stationary distribution of iterative imputation. *Biometrika* **101**, 155–173.

LU, Z.L., ZHANG, Z. & LUBKE, G. (2011). Bayesian inference for growth mixture models with latent class dependent missing data. *Multivariate Behavioral Research* **46**, 567–597.

MCCULLOCH, C.E., NEUHAUS, J.M. & OLIN, R.L. (2016). Biased and unbiased estimation in longitudinal studies with informative visit processes. *Biometrics* **72**, 1315–1324.

MENG, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* **9**, 538–573.

MIAO, W., DING, P. & GENG, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association* **111**, 1673–1683.

MOLENBERGHS, G., BEUNCKENS, C. & SOTTO, C. (2008). Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of the Royal Statistical Society (Series B)* **70**, 371–388.

PETERSON, L.A., BELLILE, E.L., WOLF, G.T., VIRANI, S., SHUMAN, A.G. & TAYLOR, J.M.G. (2016). Cigarette use, comorbidities, and prognosis in a prospective head and neck squamous cell carcinoma population. *Head and Neck* **38**, 1810–1820.

RAGHUNATHAN, T.E. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* **27**, 85–95.

RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York, NY: John Wiley and Sons, Inc, 1st edn.

SCHAFER, J.L. (1997). Imputation of missing covariates under a multivariate linear mixed model. *Technical report*, Pennsylvania State University.

SCHAFER, J.L. & YUCEL, R.M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics* **11**, 437–457.

SY, J.P. & TAYLOR, J.M.G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics* **56**, 227–236.

TAYLOR, J.M.G. (1995). Semiparametric estimation in failure time mixture models. *Biometrics* **51**, 899–907.

VAN BUUREN, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* **16**, 219–242.

VAN BUUREN, S., BRAND, J.P.L., GROOTHUIS-OUDSHOORN, C.G.M. & RUBIN, D.B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76**, 1049–1064.

WANG, S., SHAO, J. & KWANG KIM, J. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica* **24**, 1097–1116.

WHITE, I.R. & ROYSTON, P. (2009). Imputing missing covariate values for the Cox model. *Statistics in Medicine* **28**, 1982–1998.

WU, M.C. & CARROLL, R.J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44**, 175–188.

YANG, X., LU, J. & SHOPTAW, S. (2008). Imputation-based strategies for clinical trial longitudinal data with nonignorable missing values. *Statistics in Medicine* **27**, 2826–2849.