Sequential imputation for models with latent variables assuming latent ignorability

Lauren J Beesley, Jeremy M G Taylor, and Roderick J A Little

Department of Biostatistics, University of Michigan

Supplementary Materials

Contents

S1 Ignorability under a joint model (properties 1–5)

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

$$f(D, L, R; \nu) = \prod_{i=1}^{n} f(R_i|Y_i, X_i, L_i; \phi) f(Y_i|X_i, L_i; \theta) f(L_i|X_i; \omega) f(X_i; \psi) \tag{S1.1}$$

where $\nu = (\phi, \theta, \omega, \psi)$ is the set of all model parameters. We assume a flat prior for ν such that ϕ, θ, ω, and ψ are all a priori independent (so they are distinct). The factorization (S1.1) is a form of shared parameter model, where the latent variable is related both to missingness and to the distribution for Y_i (Little and Rubin, 2002).

We can impute missing values of D and L by iteratively drawing the missing values from their posterior predictive distributions, $D^{(mis)} \sim f(D^{(mis)}|D^{(obs)}, L, R)$ and

$L^{(mis)} \sim f(L^{(mis)}|D, L^{(obs)}, R)$. This leads to draws from the joint posterior predictive distribution, $f(D^{(mis)}, L^{(mis)}|D^{(obs)}, L^{(obs)}, R)$ (Little and Rubin, 2002). Define $\rho = (\theta, \omega, \psi)$. We note the following properties of the (conditional) posterior predictive distributions:

Property 1: *Under MAR and LMAR, we can ignore $R = (R^D, R^L)$ when imputing D.*

The missingness mechanism is ignorable for imputing $D^{(mis)}$ if $f(D^{(mis)}|D^{(obs)}, L, R) = f(D^{(mis)}|D^{(obs)}, L)$. Using assumptions (1) and (2) and assuming ϕ and ρ are distinct,

$$\begin{aligned}
f(D^{(mis)}|D^{(obs)}, L, R) =& \frac{1}{f(D^{(obs)}, L, R)} \int\int f(D, L, R, \nu) d\rho d\phi \\
=& \frac{1}{f(D^{(obs)}, L, R)} \int f(R|D, L; \phi) \left[\int f(D, L; \rho) f(\rho) d\rho\right] f(\phi) d\phi \\
=& \frac{f(D^{(mis)}|D^{(obs)}, L) f(D^{(obs)}, L)}{f(D^{(obs)}, L, R)} \int f(R|D^{(obs)}, L; \phi) f(\phi) d\phi \\
=& f(D^{(mis)}|D^{(obs)}, L)
\end{aligned}$$

Therefore, the missingness mechanism is ignorable for imputing D. A similar result for a related latent ignorable missingness setting was shown in Harel (2003). We note that in practice, draws from the posterior predictive distribution are obtained by first drawing the model parameter ρ from its posterior distribution and then drawing $D^{(mis)}$ from $f(D^{(mis)}|D^{(obs)}, L, R; \rho)$. We can perform both of these draws ignoring R.

Property 2: *Under MAR (but not under LMAR), we can ignore $R = (R^D, R^L)$ when imputing L.*

The missingness mechanism is ignorable for imputing $L^{(mis)}$ if $f(L^{(mis)}|L^{(obs)}, D, R) = f(L^{(mis)}|L^{(obs)}, D)$. Again using assumptions (1) and (2) and assuming ϕ and ρ are distinct,

$$\begin{aligned}
f(L^{(mis)}|L^{(obs)}, D, R) =& \frac{1}{f(L^{(obs)}, D, R)} \int\int f(D, L, R, \nu) d\rho d\phi \\
=& \frac{1}{f(L^{(obs)}, D, R)} \int f(R|D, L; \phi) \left[\int f(D, L; \rho) f(\rho)\right] d\rho f(\phi) d\phi \\
=& \frac{f(L^{(mis)}|L^{(obs)}, D) f(L^{(obs)}, D)}{f(L^{(obs)}, D, R)} \int f(R|D^{(obs)}, L; \phi) f(\phi) d\phi
\end{aligned}$$

Suppose first that missingness is MAR. Then, $f(R|D^{(obs)}, L; \phi) = f(R|D^{(obs)}, L^{(obs)}; \phi)$ and $f(L^{(mis)}|L^{(obs)}, D, R) = f(L^{(mis)}|L^{(obs)}, D)$. Therefore, R is ignorable. Under LMAR, however, the term $\int f(R|D^{(obs)}, L; \phi) f(\phi) d\phi$ depends on $L^{(mis)}$, so R is not ignorable.

Property 3: *Suppose that missingness in subset S of $\{D, L\}$ is MAR. We can ignore the corresponding subset of R when imputing L provided a distinctness property holds.*

Let R^S denote the set of missingness indicators for S and R^{-S} denote the missingness indicators for the remaining variables in $\{D, L\}$. Note by assumption (2), $L \subset S$. Let $f(R_i^{-S}|D_i, L_i; \phi) = f(R_i^{-S}|D_i^{(obs)}, L_i; \phi^{-S})$ and $f(R_i^S|D_i, L_i, R_i^{-S}; \phi) = f(R_i^S|D_i^{(obs)}, L_i^{(obs)}; \phi^S)$. Assume also that ϕ^{-S} and ϕ^S are distinct (a priori independent). Then we have

$$\begin{aligned}
f(R|D^{(obs)}, L; \phi) =& f(R^S|D^{(obs)}, L^{(obs)}; \phi^S) f(R^{-S}|D^{(obs)}, L; \phi^{-S}) \implies \\
f(L^{(mis)}|L^{(obs)}, D, R) \propto& f(L^{(mis)}|L^{(obs)}, D) \int f(R^{-S}|D^{(obs)}, L; \phi^{-S}) f(\phi^{-S}) d\phi^{-S}
\end{aligned}$$

The contribution of R^S drops out of the posterior predictive distribution, so R^S is ignorable. A similar result, called "ignorability for submodels", was shown in Harel (2003). For an example of submodel ignorability, see our data application in **Section 6**.

Property 4: *R is ignorable for ρ in a final analysis using only imputed D under MAR*
Suppose we perform our final analysis using the imputed values of D but ignoring the imputed L and again suppose that ϕ and ρ are distinct. In a Bayesian analysis, we want to make inference from the joint posterior of ϕ and ρ:

$$\begin{aligned} f(\nu|D, L^{(obs)}, R) \propto & f(R|L^{(obs)}, D; \nu) f(L^{(obs)}, D; \rho) f(\nu) \\ & \propto \left[\int f(R|L, D; \phi) f(L^{(mis)}|L^{(obs)}, D; \rho) dL^{(mis)}\right] f(L^{(obs)}, D; \rho) f(\phi) f(\rho) \\ & \propto f(R|L^{(obs)}, D^{(obs}; \phi) f(\phi) f(L^{(obs)}, D; \rho) f(\rho) \text{ under MAR} \end{aligned}$$

The posterior distributions of ϕ and ρ separate, and the posterior for ρ is independent of R. Therefore, we can ignore R for inference about ρ under MAR.

Property 5: *A final analysis for making inference about ρ using imputed D (but not imputed L) and ignoring R is valid but not fully efficient under LMAR.*
Under the setting of *Property 4* except assuming LMAR, we again have that

$$f(\nu|D, L^{(obs)}, R) \propto \left[\int f(R|L, D; \phi) f(L^{(mis)}|L^{(obs)}, D; \rho) dL^{(mis)}\right] f(\phi) f(L^{(obs)}, D; \rho) f(\rho)$$

Under LMAR, $f(R|L, D; \phi)$ depends on $L^{(mis)}$, so the contribution of R and ϕ does not factor out of the integral. Therefore, we cannot separate ϕ and ρ in the above equation. We rewrite the above equation as $f(\nu|D, L^{(obs)}, R) \propto h(\nu) f(\rho|D, L^{(obs)})$ where $h(\nu) = \left[\int f(R|L, D; \phi) f(L^{(mis)}|L^{(obs)}, D; \rho) dL^{(mis)}\right] f(\phi)$. Clearly, ν and ρ are not distinct. However, L is MAR given imputed D. Under the ignorability conditions in Little and Rubin (2002) (pg. 119–120), inference ignoring the contribution of R (using $f(\rho|D, L^{(obs)})$) will be valid from a frequency perspective but may not be fully efficient. Intuitively, the loss of efficiency comes from a loss of information about the missing L that comes from ignoring R under LMAR. However, analysis is still valid since missing L is MAR given D.

S2 Motivating the algorithm and performing parameter draws

In this appendix, we provide more details regarding the univariate imputation steps for imputing missing values in D and L. In particular, we discuss distributions we can use to perform the parameter draws within the sequential imputation algorithm. Our proposed method for drawing model parameters within a given univariate imputation step will depend on whether we are performing imputation of the latent variable or a variable in D. Here, we will suppose that L is imputed from the kernel in (2) from the main paper and that missing X and Y are imputed from *working* imputation models that may or may not correspond (3) and (4) from the main paper. Therefore, the following exploration can be applied when outcomes and covariates are imputed using (3) and (4) or using approximations.

First, we will review some notation. Define $D^{(p)}$ to be the p^{th} variable in D and $D^{(-p)}$ to be all variables in D except $D^{(p)}$. Parameter ν represents the parameters for the joint distribution, $f(D, L, R; \nu)$. We partition $\nu = (\phi, \rho)$ where ϕ represents the missingness model parameters and ρ represents all other model parameters. We assume that ρ and ϕ are distinct (a priori independent). Suppose that we specify $\tilde{f}(D_i^{(p)}|D_i^{(-p)}, L_i; \tilde{\rho}_p)$ to be the *working* conditional distribution of $D_i^{(p)}$ *used for imputation.* We can view $\tilde{f}(D_i^{(p)}|D_i^{(-p)}, L_i; \tilde{\rho}_p)$ as an approximation of $f(D_i^{(p)}|D_i^{(-p)}, L_i; \rho)$. If we use the form of the full conditional distribution as in (3) and (4), $\tilde{\rho}_p$ will be a subset of ρ. If we impute using regression models, $\tilde{\rho}_p$ may not be directly related to ρ. We suppose that we impute L from $f(L_i|D_i, R_i; \nu)$ as described in (2).

S2.1 Imputing $D^{(p)}$

In **Section S1**, we show how, when missingness is LMAR, we can impute $D^{(mis)}$ ignoring the contribution of R (assuming some distinctness properties). This is a result of the assumption that missingness is conditionally independent of $D^{(mis)}$. Rather than imputing $D^{(mis)}$ directly from $f(D^{(mis)}|D^{(obs)}, L)$, we instead obtain a draw of $D^{(mis)}$ by iteratively drawing missing values of each $D^{(p,mis)}$ from $f(D^{(p,mis)}|D^{(p,obs)}, D^{(-p)}, L)$ or from an approximated version, $\tilde{f}(D^{(p,mis)}|D^{(p,obs)}, D^{(-p)}, L)$, treating the most recent imputations for the other variables as if they were observed data (including L).

At a given iteration, we want to draw missing values of $D^{(p)}$ under MAR and LMAR from its posterior predictive distribution:

$$\tilde{f}(D^{(p,mis)}|D^{(p,obs)}, D^{(-p)}, L) = \int \tilde{f}(D^{(p,mis)}|D^{(p,obs)}, D^{(-p)}, L; \tilde{\rho}_p)\tilde{f}(\tilde{\rho}_p|D^{(p,obs)}, D^{(-p)}, L)d\tilde{\rho}_p$$

This integral suggests an approach for drawing from the posterior predictive distribution. Assuming that the data $D_i^{(p)}$ across subjects i are conditionally independent given L and $D_i^{(-p)}$, we can obtain a draw from the posterior predictive distribution by performing the following (Little and Rubin, 2002):
1) Draw $\tilde{\rho}_p$ from $\tilde{f}(\tilde{\rho}_p|D^{(p,obs)}, D^{(-p)}, L)$
2) Draw missing $D_i^{(p)}$ from $\tilde{f}(D_i^{(p,mis)}|D_i^{(p,obs)}, D_i^{(-p)}, L_i; \tilde{\rho}_p) = \tilde{f}(D_i^{(p)}|D_i^{(-p)}, L_i; \tilde{\rho}_p)$.

We note that step 1) involves drawing $\tilde{\rho}_p$ conditioning on $D^{(p,obs)}$ using only the observed part of $D^{(p)}$. This is consistent with chained equations imputation in which we draw parameter values using only the observed values of $D^{(p)}$ (Van Buuren et al., 2006). The step for drawing $\tilde{\rho}_p$ conditioning only on the observed data can be accomplished by using the data with observed values for $D^{(p)}$ and prior $\tilde{f}(\tilde{\rho}_p)$. If we assume the prior distribution is proportional to 1, we can draw $\tilde{\rho}_p$ by fitting model $\tilde{f}(D^{(p)}|D^{(-p)}, L; \tilde{\rho}_p)$ to a bootstrap sample of the data with observed values for $D^{(p)}$. We note that while this step for drawing $\tilde{\rho}_p$ does not use the most recent imputation of $D^{(p)}$, it does use the imputed values for L.

An alternative to the above is to draw $\tilde{\rho}_p$ using a Gibbs-type approach. In Gibbs sampling-type imputation algorithms, parameter values are drawn using all of the most recent imputed data, including imputed values for $D^{(p)}$ from the previous iteration. This approach is also used in SMC-FCS, a modified chained equations approach proposed in Bartlett et al. (2014). If preferred, we can obtain valid parameter draws using this approach as well. We note that we can write

$$\tilde{f}(\tilde{\rho}_p|D^{(p,obs)}, D^{(-p)}, L) = \int \tilde{f}(\tilde{\rho}_p|D^{(p)}, D^{(-p)}, L)\tilde{f}(D^{(p,mis)}|D^{(p,obs)}, D^{(-p)}, L)dD^{(p,mis)}$$

The above integral suggests that we can obtain a draw from $\tilde{f}(\tilde{\rho}_p|D^{(p,obs)}, D^{(-p)}, L)$ by drawing $\tilde{\rho}_p$ from $\tilde{f}(\tilde{\rho}_p|D^{(p)}, D^{(-p)}, L)$ using the drawn value of $D^{(p,mis)}$ from the previous iteration, which was drawn from $\tilde{f}(D^{(p,mis)}|D^{(p,obs)}, D^{(-p)}, L)$. Rather than drawing parameter values using the complete case data as is in the usual implementation of chained equations, we can alternatively draw parameters conditioning on the imputed values of $D^{(p)}$ from the last iteration. We use this approach for drawing parameters in our simulations and in our presentation of the proposed method in the main paper.

Rather than approximating the distributions for each variable with missingness with a regression model for imputation, suppose that we impute all variables using the kernel forms in (2), (3) and (4). In this case, $\tilde{\rho}_p$ is is a subset of ρ. For simplicity, we might choose to perform only a single set of parameter draws per iteration of the sequential imputation algorithm and use that set of parameter draws for imputing all of the variables in that iteration. This approach is used in Gibbs sampling-type algorithms. In this case, we might perform a set of parameter draws for ρ in the step for imputing L, which involved drawing ρ using methods treating L as latent as described in the following section. Then, we can use that same drawn value for ρ for imputing the covariate/outcome values. We note that the above derivations above suggest that we should draw ρ conditioning on the imputed values of L when we are imputing covariates/outcomes. In our experience, however, a single draw of ρ using the above approach generally produces good results when we perform our final analysis using only the imputed values of D. When we perform our final analysis using the imputed values of D and L, drawing ρ before each imputation can sometimes produce improved parameter coverage.

S2.2 Imputing the latent variable

In the imputation step for L at a given iteration of the sequential algorithm, we aim to draw missing values from the posterior predictive distribution:

$$f(L^{(mis)}|L^{(obs)}, D, R) = \int f(L^{(mis)}|L^{(obs)}, D, R; \nu) f(\nu|L^{(obs)}, D, R) d\nu$$

under LMAR and the posterior predictive distribution:

$$f(L^{(mis)}|L^{(obs)}, D) = \int f(L^{(mis)}|L^{(obs)}, D; \rho) f(\rho|L^{(obs)}, D) d\rho$$

under MAR. Here, we treat the most recent imputations for D as if they were the observed data. As before, this integral suggests an approach for drawing from the posterior predictive distribution. We can obtain a draw of the posterior predictive distribution by performing the following:

1) Under LMAR, draw ν from $f(\nu|L^{(obs)}, D, R)$.
   Under MAR, draw ρ from $f(\rho|L^{(obs)}, D)$.
2) Under LMAR, draw missing L_i from $f(L_i^{(mis)}|L_i^{(obs)}, D_i, R_i; \nu) = f(L_i|D_i, R_i; \nu)$.
   Under MAR, draw missing L_i from $f(L_i^{(mis)}|L_i^{(obs)}, D_i; \rho) = f(L_i|D_i; \rho)$

We note here that we are assuming that L_i values are conditionally independent across different values of i. Suppose our outcome model is a linear mixed model with a random intercept, L. Then i here would index the clusters (rather than the units within clusters), and a single value of L would be drawn for all units within the cluster.

Drawing ρ under MAR

When L is partially observed, we can draw ρ from $f(\rho|L^{(obs)}, D) \propto f(L^{(obs)}, D; \rho) f(\rho)$ using only the observed values of L and prior $f(\rho)$ using methods that treat L as latent or partially latent and ignoring R. For example, suppose our outcome model is a mixture of GLMs and we use $f(\rho) \propto 1$. Then, we can draw the parameter for the outcome model by fitting a latent class model to a bootstrap sample of the data treating L as fully latent.

Drawing ρ and ϕ under LMAR

We note that

$$f(\nu|L^{(obs)}, D, R) = f(\rho|L^{(obs)}, D, R, \phi)f(\phi|L^{(obs)}, D, R) \tag{S2.2}$$

When L is partially latent (so it is partially observed), we can draw values of ν using only the subjects with L observed. When L is fully latent, however, drawing from (S2.2) may not be so simple. Therefore, we will propose an alternative approach that can be applied for latent and partially latent L. We will consider how to draw ϕ and ρ separately using the factorization in (S2.2).

We first consider how to draw values for ρ from $f(\rho|L^{(obs)}, D, R, \phi)$. We have that

$$\begin{aligned} f(\rho|L^{(obs)}, D, R, \phi) &\propto f(L^{(obs)}, D, R; \rho, \phi)f(\rho) \\ &\propto f(R|D, L^{(obs)}; \nu)f(L^{(obs)}, D; \rho)f(\rho) \end{aligned}$$

This kernel separates into two factors: one that depends on ϕ and R and one that does not. We note that L is treated as MCAR when L is fully latent and is assumed to be MAR when L is partially latent, so the missingness in L is ignorable given $D^{(obs)}$. When we condition on the imputed D, we can make valid inference about ρ (in a frequentist sense) without conditioning on R and ϕ (Little and Rubin, 2002). However, R does contain some information about the value of L under LMAR (ν and ρ are clearly not distinct) and therefore would contribute some information about ρ. Ignoring R when drawing ρ, therefore, may result in a loss of efficiency. We can validly (but with some potential loss of efficiency) ignore the contribution of R and ϕ to $f(\rho|L^{(obs)}, D, R, \phi)$ and instead draw ρ from $f(\rho|L^{(obs)}, D)$. This is important because it may be difficult to draw from $f(\rho|L^{(obs)}, D, R, \phi)$, but a draw from $f(\rho|L^{(obs)}, D)$ can be obtained using standard methods that treat L as latent or partially latent and ignoring R.

We now consider how to draw values for ϕ. The distribution $f(\phi|L^{(obs)}, D, R)$ may be difficult to draw from under LMAR assumptions since this distribution does not condition on $L^{(mis)}$. We instead use the integral decomposition:

$$f(\phi|L^{(obs)}, D, R) = \int f(\phi|L, D, R)f(L^{(mis)}|L^{(obs)}, D, R)dL^{(mis)}$$

We can obtain a valid draw from $f(\phi|L^{(obs)}, D, R)$ by instead drawing from $f(\phi|L, D, R)$ using the most recent imputation of L, which was drawn from $f(L^{(mis)}|L^{(obs)}, D, R)$. Therefore, we can draw values of ϕ directly using the most recent imputed values of L. This is easier than drawing from $f(\phi|L^{(obs)}, D, R)$ because it can directly incorporate the working LMAR model for the missingness mechanism without integrating out missing values of L. We do not choose to use this same integral decomposition approach for drawing ρ as our proposed approach (which does not condition on the most recent imputation of L) tends to result in more stable convergence properties in our experience (for fully latent L).

S3 Bias of complete case analysis under LMAR

In the main paper, we claim that we may expect bias in complete case analysis when missingness in a covariate or outcome depends on the latent variable, L. Here, we provide some justification for this claim, which may initially seem unintuitive. In usual regression analysis, complete case analysis is valid (but not fully efficient) when missingness in the covariates/outcome does not depend on the outcome value conditional on X. However, missingness can depend on other missing covariate values. The same does not apply when missingness in covariates or the outcome depend on latent L.

Let's consider the simple setting in which Y and X are univariate. Suppose first that L was *fully observed*, so covariate and outcome missingness is then missing at random. The two models of interest are $f(L|X)$ and $f(Y|X, L)$. If L is fully observed, we do not expect bias in estimating parameters related to $f(Y|X, L)$ unless sampling is directly related to Y. However, we may run into bias in estimating parameters related $f(L|X)$. Suppose $f(L|X)$ is a logistic regression model as it is in the Cox proportional hazards mixture cure model. In this case, if missingness only depends on L, then would not expect bias in estimating covariate effects, but we would expect bias in the intercept of the logistic regression. This result comes from literature related to case-control sampling. Suppose however that missingness depends on L and X. In this case, even with L fully observed, complete case analysis with respect to LMAR missing X values could result in biased inference for $f(L|X)$.

In reality, L is *fully or partially unobserved.* In this case, LMAR missingness in X_i or Y_i is MNAR. For the sake of simplicity, let's assume that L is fully latent, so L is never observed for any subject. In this case, we define complete case analysis as analysis of the subjects with Y and X fully observed, but all of these subjects will still have L unobserved. The outcome distribution given R_i^D can be written as

$$f(Y_i|R_i^D = 1, X_i) = \frac{1}{P(R_i^D = 1|X_i)} \int P(R_i^D = 1|L_i, X_i, Y_i; \phi^D) f(Y_i|L_i, X_i; \theta) f(L_i|X_i; \omega) dL_i$$

$$= \frac{f(Y_i|X_i)}{P(R_i^D = 1|X_i)} \int P(R_i^D = 1|L_i, X_i, Y_i; \phi^D) f(L_i|Y_i X_i; \omega) dL_i$$

When missingness depends on L_i, the missingness mechanism does not factor out of the integral, and therefore R_i^D is not ignorable. If we perform a complete case analysis ignoring the missingness mechanism, we could have biased inference. We contrast this with the result earlier on, which states that likelihood inference about $\rho = (\theta, \omega)$ given the full imputed D and ignoring the missingness mechanism is valid in a frequentist sense but with a possible loss of efficiency.

Instead, we can view this problem in terms of the sampling probability given the observed data directly. We have that:

$$P(R_i^D = 1|X_i, Y_i) = \int P(R_i^D = 1|L_i, X_i, Y_i) f(L_i|X_i, Y_i) dL$$

Assuming missingness depends only on L_i and possibly X_i (otherwise, we expect complete case analysis to be biased anyway), we have

$$P(R_i^D = 1|X_i, Y_i) = \frac{1}{f(Y_i|X_i)} \int P(R_i^D = 1|L_i, X_i) f(Y_i|L_i, X_i) f(L_i|X_i) dL$$

When missingness depends on L_i, then the marginal sampling probability $P(R_i^D = 1|X_i, Y_i)$ does depend on Y_i. Therefore, the association between L_i and sampling induces an association between Y_i and sampling when L_i is a latent variable. Therefore, we can have bias in the complete case analysis when complete cases are defined in terms of observed values in Y and X. This explains why we see bias in the complete case analysis intercept terms and occasionally for regression coefficients when sampling depends on the latent variable in our simulations.

S4 Simulation study

In the main paper, we summarize overall results for a simulation study. Here, we present details of a simulation study in five parts. In Simulation 1–3, we explore bias, coverage, and empirical variance of outcome model parameters after imputation in the linear mixed model, Cox proportional hazards mixture cure model, and mixture of two normals settings respectively. In Simulation 4, we explore convergence of the imputation algorithm in several missingness scenarios. In Simulation 5, we explore the impact of different types of final analysis on efficiency. Unless otherwise specified, imputations are drawn using the SMC imputation method in the main paper rather than the chained equations method. Simulations denoted 'APPROX' correspond to the chained equations method. The majority of the simulations focus on the SMC imputation method.

Simulations 1–3 explore properties of both the SMC imputation and chained equations imputation methods. Simulations 4-5 focus on the SMC imputation setting, but the overall results are expected to be similar for the chained equations method.

S4.1 Simulation 1: linear mixed model with random intercept

We simulate 1500 datasets with 500 subjects each under a linear mixed model with a random intercept. Each dataset contains two binary covariates, X_1 and X_2. X_1 takes the value 1 with a probability of 0.5, and X_2 is generated using $\text{logit}(P(X_2 = 1|X_1)) = 0.5X_1$. We draw random intercept $b_i \sim N(0, 1)$ for each individual and then generate Y for each individual at each of three time-points using the model

$$Y_{ij} = \beta_{Intercept} + \beta_{X_1}X_{i1} + \beta_{X_2}X_{i2} + \beta_{Time}Time_{ij} + b_i + e_{ij}, \qquad j = 1, 2, 3$$

with independent $N(0, 1)$ errors and with $\beta_{Intercept} = \beta_{X_1} = \beta_{X_2} = 0.5$ and $\beta_{Time} = 0.2$. In this simulation setting, $Y = (Y_1, Y_2, Y_3)$, $X = (X_1, X_2)$, and $L = b$. We impose $\sim 50\%$ missingness in X_2 using each of the following mechanisms:

(A) MAR with $\text{logit}(P(X_2 \text{ missing}|X_1, b, Y)) = -1.1 + Y_1$
(B) LMAR with $\text{logit}(P(X_2 \text{ missing}|X_1, b, Y)) = 0.5b$
(C) LMAR with $\text{logit}(P(X_2 \text{ missing}|X_1, b, Y)) = 0.1 + 1.2b$.
(D) LMAR $\text{logit}(P(X_2 \text{ missing}|X_1, b, Y)) = -0.5 + 1.2b + 0.5Y_1$

Mechanism (A) is MAR dependent on Y_1, the baseline value of Y. Mechanism (B) is LMAR with a moderate dependence between missingness and b, mechanism (C) is LMAR with a strong dependence on b, and mechanism (D) is LMAR with dependence on both b and Y_1. Y and X_1 are fully observed.

We then impute values of X_2 and b using methods discussed in the main paper under various working models. When we impute under a LMAR working model, we model the covariate missingness indicator R_i^D using a logistic regression with different functions of b, X_1, and Y as predictors. When we assume a MAR working model, we impute L ignoring the missingness mechanism. For each simulated dataset, we create 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 compute the bias, empirical variance, and coverage rates across the 1500 simulations. We note that the APPROX simulations involve **chained equations-type** imputation of X_2 conditional on X_1, L and Y using a logistic regression form rather than using kernel (4), so the imputation distributions for X_2 and L in this case do not correspond to a coherent joint distribution. Since we only have missingness in a single covariate (X_2), the imputation distributions for the other simulation settings do correspond to a valid joint model, although that joint model was never specified directly. We compare the results to complete case analysis. Since X is time-invariant in this model, complete case analysis involves excluding all subjects with missing X.

Table S1 shows the simulation results. Complete case analysis produced biased parameter estimates in all four underlying missingness mechanisms considered. Under MAR missingness mechanism (A), the MAR-based imputation approach produces unbiased parameter estimates. LMAR imputation under mechanism (A) produces biased parameter estimates when an incorrect working missingness model is used. When the working model contains the underlying missingness model, however, the LMAR method results in essentially unbiased parameter estimates. Under mechanism (A), the MAR-based imputation approach and the LMAR imputation approach with the correct working model result in very similar coverage and relative variance. APPROX Imputation using a logistic regression model for imputing X_2 had similar performance to imputation using kernel (4). This suggests that the LMAR-based imputation approach can be applied when the true missingness model is MAR as long as the missingness model contains the true model.

Under mechanism (B), all imputation approaches produce essentially unbiased or low bias parameter estimates. The LMAR approaches, however, result in small increases in coverage and reductions in variance and bias compared to the MAR imputation approach. Under mechanism (C), the MAR-based imputation approach produces noticeable bias in estimating the mixed model intercept and parameter associated with the imputed covariate. We see corresponding reductions in coverage for these parameters. In contrast, the LMAR-based imputation approaches produce unbiased parameter estimates. For mechanisms (B) and (C), the working model that uses $\mathbb{I}(b > 0)$ instead of b in the working model still shows good performance despite the fact that the working model does not contain the true model. We do not see evidence of problems arising from lack of identifiability or lack of convergence under any of the working models considered here. MAR-based imputation using a logistic regression model for imputing X_2 resulted in slightly greater bias than MAR imputation using kernel (4).

Under mechanism (D), MAR-based imputation was substantially biased, and all imputation settings assuming LMAR-based imputation resulted in reduced bias. Notably, even in imputation settings where the missingness model was mis-specified (truth depends linearly on b and Y_1), we can see a reduction in bias compared to MAR-based imputation. Taken as a whole, this set of simulations suggests that our imputation approach can induce bias when the missingness model is far off the true model, but we can often see good properties when the working model contains or is somewhat "close" to the true model.

Table S1: Linear mixed model estimates using proposed imputation methods

		Parameters			
	Contains	Intercept	X_1	X_2	Time
Method	Truth#	Bias (Var) CI†	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI
Full Data	-	0 (1.2) 95	0 (1.0) 95	0 (1.1) 95	0 (0.10) 95
Missingness dependent on Y_{i1}, independent of b_i (Mechanism A)					
Complete Case	-	-78 (2.0) 0	-9 (1.8) 91	-9 (1.9) 90	19 (0.20) 1
MAR Imputation	Y	0 (1.8) 94	0 (1.1) 95	0 (2.8) 94	0 (0.10) 95
LMAR Imputation: b*	N	6 (1.4) 94	2 (1.1) 95	-9 (1.9) 94	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1$	N	6 (1.4) 93	1 (1.1) 96	-9 (2.0) 95	0 (0.10) 95
LMAR Imputation: b, Y_1	Y	0 (1.8) 94	0 (1.1) 96	0 (2.8) 94	0 (0.10) 95
LMAR Imputation: $\mathbb{I}(b > 0), Y_1$	Y	0 (1.9) 94	0 (1.1) 96	0 (2.8) 93	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1, Y_1$	Y	0 (1.9) 94	0 (1.1) 95	0 (2.8) 92	0 (0.10) 95
LMAR Imputation: b, Y_2	N	7 (1.4) 92	2 (1.1) 95	-11 (1.8) 93	0 (0.10) 95
MAR APPROX Imputation	Y	-1 (1.9) 94	0 (1.1) 96	0 (3.0) 94	0 (0.10) 95
LMAR APPROX Imputation: b	N	5 (1.5) 95	1 (1.1) 95	-8 (2.1) 96	0 (0.10) 95
Missingness moderately dependent on b_i (Mechanism B)					
Complete Case	-	-24 (2.4) 66	0 (2.1) 95	0 (2.2) 94	0 (0.19) 95
MAR Imputation	N	-2 (1.7) 93	0 (1.1) 95	2 (2.4) 92	0 (0.10) 95
LMAR Imputation: b	Y	0 (1.6) 95	0 (1.1) 95	0 (2.2) 93	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1$	Y	0 (1.6) 95	0 (1.1) 95	0 (2.2) 94	0 (0.10) 95
LMAR Imputation: b, Y_1	Y	0 (1.6) 95	0 (1.1) 95	0 (2.2) 94	0 (0.10) 95
LMAR Imputation: $\mathbb{I}(b > 0), Y_1$	N	0 (1.6) 95	0 (1.1) 95	0 (2.2) 94	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1, Y_1$	Y	0 (1.6) 95	0 (1.1) 95	0 (2.2) 94	0 (0.10) 95
MAR APPROX Imputation	N	-3 (1.7) 94	0 (1.1) 95	3 (2.4) 93	0 (0.10) 95
LMAR APPROX Imputation: b	Y	0 (1.6) 95	0 (1.1) 95	0 (2.2) 95	0 (0.10) 95
Missingness strongly dependent on b_i (Mechanism C)					
Complete Case	-	-48 (2.5) 11	0 (1.8) 95	0 (2.0) 94	0 (0.22) 95
MAR Imputation	N	-7 (2.0) 90	0 (1.1) 95	8 (2.8) 90	0 (0.10) 95
LMAR Imputation: b	Y	0 (1.5) 96	0 (1.1) 96	0 (2.0) 96	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1$	Y	0 (1.5) 96	0 (1.1) 96	0 (2.0) 97	0 (0.10) 95
LMAR Imputation: b, Y_1	Y	0 (1.6) 95	0 (1.1) 95	0 (2.1) 95	0 (0.10) 95
LMAR Imputation: $\mathbb{I}(b > 0), Y_1$	N	0 (1.6) 95	0 (1.1) 96	0 (2.1) 96	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1, Y_1$	Y	0 (1.5) 96	0 (1.1) 96	0 (2.1) 96	0 (0.10) 95
MAR APPROX Imputation	N	-8 (2.0) 90	0 (1.1) 95	9 (2.8) 89	0 (0.10) 95
LMAR APPROX Imputation: b	Y	0 (1.5) 97	0 (1.1) 96	0 (2.1) 97	0 (0.10) 95
Missingness dependent on b_i and Y_{i1} (Mechanism D)					
Complete Case	-	-73 (2.0) 0	-5 (1.6) 93	-5 (1.6) 94	8 (0.21) 56
MAR Imputation	N	-8 (2.0) 91	-1 (1.1) 96	10 (2.8) 89	0 (0.10) 95
LMAR Imputation: b	N	3 (1.4) 96	0 (1.1) 96	-5 (1.7) 98	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1$	N	3 (1.4) 96	0 (1.1) 96	-5 (1.6) 98	0 (0.10) 95
LMAR Imputation: b, Y_1	Y	0 (1.5) 97	0 (1.1) 96	0 (2.0) 97	0 (0.10) 95
LMAR Imputation: $\mathbb{I}(b > 0), Y_1$	N	0 (1.6) 96	0 (1.1) 96	0 (2.0) 97	0 (0.10) 95
LMAR Imputation: $b, X_1, b \times X_1, Y_1$	Y	0 (1.6) 96	0 (1.1) 96	0 (2.0) 96	0 (0.10) 95
LMAR Imputation: b, Y_2	N	3 (1.4) 96	0 (1.1) 95	-6 (1.7) 97	0 (0.10) 95
MAR APPROX Imputation	N	-9 (2.1) 90	-1 (1.1) 95	11 (2.9) 88	0 (0.10) 95
LMAR APPROX Imputation: b	N	3 (1.4) 97	0 (1.1) 96	-4 (1.7) 98	0 (0.10) 95

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

† All values in table multiplied by 100. CI indicates coverage of 95% confidence intervals. 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

S4.2 Simulation 2: Cox proportional hazards mixture cure model

We simulate 500 datasets of 500 subjects under a CPH mixture cure model. Covariates X_1 and X_2 are simulated as in Simulation 1. We simulate an underlying cure status using the relation $\text{logit}(P(\text{Not Cured}|X_{i1}, X_{i2})) = 0.5 + 0.5X_{i1} + 0.5X_{i2}$. This results in an average cure rate of 26%. For the non-cured group (G=1), we simulate an event time using the hazard function $\lambda(t) = 0.0005t^{0.3}e^{0.5X_{i1}+0.5X_{i2}}$. For cured subjects (G=0), the event time is taken to be infinity. We generate censoring times using the relation $\lambda_C(t) = 0.00015t^{0.5}$ for the first 400 subjects and impose administrative censoring at 3000 for the remaining 100 subjects. The observed event/censoring time T_i is taken as the minimum of the censoring and event time, and δ_i represents the event indicator. In this simulation setting, $Y = (T, \delta)$, $X = (X_1, X_2)$, and $L = G$. For the estimation, we assume subjects with T_i greater than a late cut-point are cured. We choose a cut-point of 50 as the Kaplan-Meier plots demonstrate a clear plateau by that point. We impose ∼50% missingness in X_2 using each of the following mechanisms:

(A) MCAR with missingness probability of 0.5
(B) LMAR with $\text{logit}(P(X_2 \text{ missing}|X_1, G, T, \delta)) = -0.2 + 0.3G$
(C) LMAR with $\text{logit}(P(X_2 \text{ missing}|X_1, G, T, \delta)) = -0.9 + 1.2G$.
(D) LMAR with $\text{logit}(P(X_2 \text{ missing}|X_1, G, T, \delta)) = -1.1 + 1.2G + 0.5X_1$.

Mechanism (A) is MCAR, mechanism (B) is LMAR with a moderate dependence on cure status (G), mechanism (C) is LMAR with a strong dependence on cure status, and mechanism (D) depends on both cure status and X_1.

We assume a Weibull baseline hazard in the non-cured group for imputation. For each imputed dataset, we fit a CPH cure model, which consists of a logistic regression for the probability of not being cured and a Cox regression for the hazard of events in the not cured group. We fit this model using the package *smcure* in R (Cai et al., 2012). Variances were estimated using 100 bootstrap samples.

Table S2 shows the simulation results for the Cox proportional hazards mixture cure model. As expected, complete case analysis is essentially unbiased under covariate missingness mechanism (A) (MCAR), but the imputation-based methods are more efficient than the complete case analysis. When missingness depends on the underlying cure status, however, complete case analysis is biased. We see comparatively little bias in the imputation-based estimates across missingness mechanisms and imputation models using kernel (4). We note that even when we specify the correct missingness model, we sometimes see bias in estimating the intercept parameter in the logistic regression. This parameter is the most difficult to estimate due to identifiability issues with the CPH cure model, and these biases will reduce with larger sample sizes (simulated sample size = 500). As such, we should not over-interpret bias in this parameter. APPROX Imputation using a logistic regression model for imputing X_2 resulted in increased bias in all scenarios. For missingness mechanisms (A) and (B) and using kernel (4) for imputation, we see very little difference between the MAR and LMAR imputation approaches in terms of bias, coverage, and relative variance. APPROX imputation under LMAR resulted in slightly larger variances than APPROX imputation under MAR.

In mechanism (C) (when missingness depends strongly on cure status) and mechanism (D) (when missingness depends on cure status and X_1), we still see little difference between MAR and LMAR imputation methods using kernel (4) in terms of bias. Larger bias differences between MAR-based and LMAR-based imputation can be seen when covariate imputation uses a logistic regression instead of kernel (4) in mechanism (c). The LMAR imputation approaches using kernel (4) (which differ only in terms of the working missingness model) produce essentially unbiased estimates for all model parameters (except for the intercept of the logistic regression). LMAR imputation using G, X_1, and

$G \times X_1$ in the working model resulted in some numerical convergence issues for several of the simulations (15 simulations failed), which may indicate issues with model identifiability (possibly due to collinearity). We included only the converging simulations (485 of them) in **Table S2**.

Overall, these simulations suggest a large degree of stability in CPH cure model inference when we impute assuming MAR and the true mechanism is LMAR. A greater degree of bias is introduced in the APPROX simulations, where the covariate imputation distribution assumed to follow a simple regression model form rather than the form in (4).

Table S2: CPH cure model estimates using proposed imputation methods

		- - - - - - - - Logistic Regression	- - - - - - - -		- - - - Cox Regression - - - -	
	Contains	Intercept	X_1	X_2	X_1	X_2
Method	Truth#	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI
Full Data	-	1 (6.5) 94	1 (7.9) 95	0 (8.4) 95	0 (2.0) 95	0 (2.3) 95
		MCAR missingness independent of G_i (Mechanism A)				
Complete Case	-	2 (12.7) 97	1 (14.9) 97	1 (18.5) 96	1 (4.3) 95	0 (5.1) 94
MAR Imputation	Y	3 (9.1) 94	1 (8.4) 96	0 (18.0) 95	0 (2.1) 96	0 (4.8) 93
LMAR Imputation: G*	Y	3 (9.4) 94	1 (8.3) 96	0 (19.5) 95	0 (2.1) 96	0 (4.9) 94
LMAR Imputation: G, X_1, Y, δ	Y	3 (9.3) 94	1 (8.3) 96	0 (18.8) 95	0 (2.1) 96	0 (4.8) 95
LMAR Imputation: $G, X_1, G \times X_1$	Y	3 (9.5) 94	1 (8.4) 96	0 (18.9) 95	0 (2.1) 96	0 (4.8) 95
MAR APPROX Imputation	Y	6 (9.4) 93	1 (8.2) 95	-6 (19.6) 91	0 (2.1) 96	-1 (4.4) 93
LMAR APPROX Imputation: G	Y	6 (9.6) 93	1 (8.4) 96	-6 (21.1) 91	0 (2.1) 96	-2 (4.5) 91
		Missingness moderately dependent on G_i (Mechanism B)				
Complete Case	-	-13 (12.1) 93	3 (16.0) 97	0 (15.4) 97	0 (4.8) 95	1 (5.4) 93
MAR Imputation	N	3 (8.8) 95	0 (8.5) 96	0 (16.9) 96	0 (2.2) 96	0 (4.8) 93
LMAR Imputation: G	Y	3 (8.9) 96	0 (8.5) 96	0 (16.7) 96	0 (2.2) 95	0 (4.8) 94
LMAR Imputation: G, X_1, Y, δ	Y	3 (8.8) 94	1 (8.5) 96	0 (16.0) 96	0 (2.2) 95	1 (4.7) 94
LMAR Imputation: $G, X_1, G \times X_1$	Y	3 (9.0) 94	0 (8.6) 96	0 (16.3) 95	0 (2.2) 95	0 (4.8) 93
		Missingness strongly dependent on G_i (Mechanism C)				
Complete Case	-	-50 (9.9) 62	2 (12.9) 97	0 (12.2) 96	1 (5.4) 95	0 (5.8) 95
MAR Imputation	N	4 (8.3) 96	1 (8.6) 96	-1 (15.2) 95	0 (2.2) 96	0 (5.1) 94
LMAR Imputation: G	Y	2 (7.9) 95	0 (8.5) 97	0 (13.3) 96	0 (2.3) 96	0 (5.5) 93
LMAR Imputation: G, X_1, Y, δ	Y	2 (7.8) 95	1 (8.4) 97	1 (13.0) 96	0 (2.2) 95	0 (5.4) 93
LMAR Imputation: $G, X_1, G \times X_1$	Y	1 (7.5) 93	1 (8.1) 94	0 (13.0) 93	0 (2.2) 92	0 (5.4) 91
MAR APPROX Imputation	N	8 (8.4) 93	1 (8.5) 97	-12 (16.1) 91	0 (2.3) 96	-1 (5.3) 90
LMAR APPROX Imputation: G	Y	6 (8.0) 94	1 (8.4) 97	-9 (14.2) 93	1 (2.3) 96	-1 (5.2) 91
		Missingness dependent on G_i and X_{i1} (Mechanism D)				
Complete Case	-	-45 (9.1) 69	-12 (13.1) 96	0 (12.0) 96	1 (5.6) 95	0 (6.5) 95
MAR Imputation	N	4 (8.4) 95	0 (8.5) 97	-2 (15.5) 94	0 (2.3) 96	0 (5.6) 94
LMAR Imputation: G, X_1	Y	3 (7.8) 95	0 (8.5) 98	0 (13.2) 96	0 (2.3) 96	0 (6.0) 93

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

† All values multiplied by 100. CI indicates coverage of 95% confidence intervals. Var indicates empirical variance.

# Indicates whether working missingness model contains true model.

APPROX: Imputation of X_2 uses a logistic regression with predictors $X_1, G, G \times H_0(Y), G \times H_0(Y) \times X_1$.

S4.3 Simulation 3: mixture of normals

We simulate 500 datasets of 500 subjects under a normal mixture model with two binary covariates and two latent classes. Covariates X_1 and X_2 are simulated as in Simulation 1. We generate the mixing variable C_i with $P(C_i = 1) = 0.62$ for each individual. We draw $N(0, 1)$ errors e_i and then generate Y using the model $Y_i = 0.5 + 0.5X_{i1} + 0.5X_{i2} + e_i$ if $C_i = 1$ and $Y_i = 2 + 3X_{i1} + 2X_{i2} + e_i$ if $C_i = 0$. In this simulation setting, $X = (X_1, X_2)$ and $L = C$. We then impose missingness in X_2 using each of the following mechanisms:

(A) MAR with $P(X_2 \text{ missing}|X_1, C, Y) = -0.5 + 0.2Y$
(B) LMAR with $P(X_2 \text{ missing}|X_1, C, Y) = -0.3 + 0.5C$
(C) LMAR with $P(X_2 \text{ missing}|X_1, C, Y) = -1.1 + 1.7C$.

Mechanism (A) is MAR dependent on Y. Mechanism (B) is LMAR with a moderate dependence on the latent class variable (C), and mechanism (C) is LMAR with a strong dependence on the latent class.

For each imputed dataset, we fit a latent class model (with two classes) using the package 'flexmix' in R to estimate θ through an EM algorithm (Leisch, 2004). The package 'flexmix' estimates the variance for $\hat{\theta}$ for each dataset by fitting a GLM weighted by estimated class membership probabilities for each individual. When parameters are drawn using latent class modeling, we may not be able to determine which value of C belongs to which subclass identified by the latent class modeling. In other words, we may not be able to differentiate which subset of θ belongs to which value of C. We can circumvent this issue by placing an additional assumption to differentiate between classes. We impose an identifying restriction that defines class $C_i = 1$ to be the cluster determined by the latent class modeling with a smaller intercept value. We note that the two clusters are well-separated in this example. We predict that we may encounter greater identifiability issues (in differentiating the clusters) when the clusters have parameters that are very close together.

Table S3 shows the simulation results for a mixture of normal distributions. Complete case analysis results in biased parameter values for mechanism (A) and mild or no bias for mechanisms (B) and (C). For mechanism (A), the MAR-based imputation approach produces essentially unbiased parameter estimates. The LMAR imputation approaches with working missingness models containing the true missingness model also produce very small bias. Mild increases in bias can be seen for the LMAR imputation approach using an incorrect working model. Compared to the MAR approach, the LMAR approach using the correct working model resulted in similar or slightly larger variances for all parameter estimates.

For mechanism (B), little bias can be seen across all of the imputation approaches. Similar coverage rates can be seen across imputation approaches. In this example, we see slightly smaller variances for the LMAR approaches with the more complicated working models. Under mechanism (C), we see increases in bias and small decreases in coverage for estimating mixture model parameters using the MAR-based imputation method (either using kernel (4) or logistic regression for imputing X_2). The LMAR-based imputation method using only C in the working missingness model produces essentially unbiased parameter estimates for all parameters. Compared to the approaches using the more complicated working model, the simpler LMAR approach using kernel (4) results in smaller variances for estimating model parameters.

Table S3: Mixture of normals model estimates using proposed imputation methods

		- - - - - - C = 1 - - - - - -			- - - - - - C = 0 - - - - - -		
	Contains	Intercept	X_1	X_2	Intercept	X_1	X_2
Method	Truth#	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI	Bias (Var) CI
Full Data	-	0 (1.6) 95	0 (1.9) 92	0 (1.5) 97	0 (3.6) 94	0 (3.3) 94	0 (3.4) 94
Missingness dependent on Y_i, independent of C_i (Mechanism A)							
Complete Case	-	2 (2.8) 94	-7 (3.5) 92	-4 (2.7) 95	3 (12.6) 93	-12 (10.3) 90	-4 (10.3) 93
MAR Imputation	Y	1 (1.9) 95	0 (4.3) 94	1 (4.6) 95	-1 (6.1) 94	-2 (7.8) 94	0 (5.5) 95
LMAR Imputation: C*	N	3 (1.9) 94	1 (4.6) 94	-2 (4.5) 95	-3 (6.2) 94	-4 (7.8) 94	2 (5.6) 94
LMAR Imputation: C, Y, X_1	Y	1 (2.0) 94	0 (4.3) 93	1 (4.6) 95	-1 (6.4) 95	-2 (7.7) 94	0 (5.3) 95
LMAR Imputation: $C, Y, C \times Y$	Y	1 (2.0) 94	1 (4.6) 93	1 (4.6) 96	-2 (6.2) 94	-2 (8.1) 94	0 (6.0) 95
LMAR Imputation: $C, Y, C \times Y, X_1$	Y	1 (2.0) 94	0 (4.5) 93	1 (4.9) 95	-2 (6.3) 94	-2 (8.0) 93	0 (6.0) 94
MAR APPROX Imputation	Y	1 (2.0) 94	0 (2.9) 93	0 (3.2) 92	0 (6.3) 92	0 (6.4) 93	-3 (5.5) 94
LMAR APPROX Imputation: C	N	3 (1.9) 93	0 (3.4) 93	-3 (3.8) 93	-1 (6.3) 93	-2 (6.9) 92	0 (5.3) 93
Missingness moderately dependent on C_i (Mechanism B)							
Complete Case	-	1 (5.4) 93	-1 (5.4) 93	0 (4.1) 95	-1 (8.6) 91	0 (7.9) 92	0 (6.9) 91
MAR Imputation	N	1 (2.3) 94	0 (2.7) 94	-1 (4.0) 93	-2 (5.8) 94	-2 (4.8) 95	2 (4.7) 93
LMAR Imputation: C	Y	0 (2.3) 94	1 (4.7) 94	1 (5.1) 93	-1 (5.6) 95	-2 (6.4) 95	0 (5.3) 94
LMAR Imputation: C, Y, X_1	Y	0 (2.4) 95	1 (5.1) 94	1 (5.5) 93	-1 (5.8) 94	-2 (7.1) 95	0 (5.7) 93
LMAR Imputation: $C, Y, C \times Y$	Y	0 (2.4) 94	1 (3.8) 94	1 (4.9) 92	-1 (5.7) 93	-2 (5.7) 94	0 (5.0) 94
LMAR Imputation: $C, Y, C \times Y, X_1$	Y	1 (2.4) 95	1 (3.6) 95	1 (4.7) 92	-1 (5.8) 94	-2 (5.6) 94	0 (5.0) 94
Missingness strongly dependent on C_i (Mechanism C)							
Complete Case	-	3 (8.2) 93	-2 (7.6) 92	-1 (6.4) 93	-2 (5.8) 92	1 (6.1) 92	1 (5.3) 92
MAR Imputation	N	5 (2.5) 93	2 (7.2) 93	-7 (7.4) 91	-5 (4.9) 95	-4 (7.6) 94	5 (7.0) 92
LMAR Imputation: C	Y	1 (2.8) 94	1 (4.6) 94	0 (6.4) 92	-1 (4.5) 96	-2 (5.9) 95	0 (5.2) 94
LMAR Imputation: C, Y, X_1	Y	1 (2.9) 94	2 (5.9) 94	0 (7.2) 92	-2 (4.7) 96	-3 (6.9) 95	0 (5.9) 94
LMAR Imputation: $C, Y, C \times Y$	Y	1 (2.8) 94	1 (5.4) 94	0 (7.3) 92	-2 (4.5) 96	-2 (6.7) 95	0 (5.8) 93
LMAR Imputation: $C, Y, C \times Y, X_1$	Y	1 (2.8) 95	2 (6.6) 94	0 (7.8) 92	-2 (4.8) 94	-3 (8.0) 94	0 (6.1) 92
MAR APPROX Imputation	N	5 (2.5) 92	2 (5.7) 93	-7 (6.6) 88	-6 (4.9) 94	-3 (6.8) 94	6 (6.2) 91
LMAR APPROX Imputation: C	Y	1 (2.8) 92	2 (6.6) 94	0 (7.3) 87	-1 (4.4) 95	-3 (7.3) 95	0 (6.4) 94

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

† All values in table multiplied by 100. CI indicates coverage of 95% confidence intervals.Var indicates empirical variance.

# Indicates whether working missingness model contains true model.

APPROX: Imputation of X_2 uses a logistic regression with predictors $C, X_1, X_1 \times C, Y, Y \times C$ (instead of kernel (4))

S4.4 Simulation 4: Exploring identifiability and convergence

One criticism of the selection model factorization in (1) is that it is often difficult to determine whether the parameters of the working missingness model are identifiable (Little, 2009). By "identifiable," we mean that the observed data likelihood has a unique maximizer. Even if the model parameters are technically identifiable, one additional concern is that the likelihood surface near the maximizer may be nearly flat. These identifiability concerns can lead to issues with model fitting and convergence of the imputation algorithm. In order to better understand possible identifiability-related convergence issues, we perform a set of simulations evaluating convergence of the imputation algorithm under a variety of modeling scenarios.

We simulate 500 complete datasets under a linear mixed model, cure model, and mixture of normals as in Simulations 1-3. We impose $\sim 50\%$ covariate or outcome missingness (but not both) under a variety of missingness models.

For covariate missingness, we generate MAR and LMAR missingness using missingness mechanisms (A) and (C) from Simulations 1-3. For both the linear mixed model and mixture of normals model, we generate outcome missingness under MCAR and LMAR using mechanism (C) from Simulations 1 and 3 applied to the outcome instead of the covariate. We also impose LMAR outcome missingness for the mixture of normals model using the relation $\text{logit}(P(Y \text{ missing}|X, C)) = -1.1 + 0.5X_1 - 0.5X_2 + 1.7C$. This results in $\sim 50\%$ outcome or covariate missingness in each scenario. We note that in each case in Simulation 4, we only have missingness in a single covariate (X_2) or a single outcome variable (Y). Therefore, the SMC imputation distributions do correspond to a valid joint model, although that joint model was never specified directly. The primary purpose of this simulation is to explore identifiability-related convergence issues, which would be similarly present in the joint modeling and SMC imputation settings.

For each outcome model parameter, we estimate the fraction of missing information as described in (Little and Rubin, 2002). We also calculate the Gelman-Rubin convergence statistic (the potential scale reduction factor) for the outcome and missingness model parameter draws across imputation streams. The Gelman-Rubin statistic is a measure of the relative between and within-chain variance, and values less than 1.1 generally indicate satisfactory convergence (Gelman and Rubin, 1992). We also calculate a multivariate version of the Gelman-Rubin statistic to evaluate convergence overall across different model parameters (Brooks and Gelman, 1998).

Table S4 shows the simulation results. Under covariate missingness, the fractions of missing information tend to be generally small, particularly for parameters related to X_1, the fully-observed covariate. We see larger estimates for the fraction of missing information when we impose similar rates of missingness in the outcome. Additionally, we see good Gelman-Rubin convergence properties under covariate missingness and MAR outcome missingness. Under LMAR outcome missingness, the outcome model parameters appear to converge, but the parameters in the missingness model (in particular, the parameter attached to the latent variable) show some evidence of convergence problems. The drawn values of the outcome model parameters appear 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 fix the value of the parameter related to the latent variable in the missingness, we see a large improvement in the convergence properties of the imputation algorithm.

Table S4: Fraction of missing information and convergence properties[†]

	— Fraction of Missing Information —						——Gelman-Rubin Statistic ——								Overall
	Outcome Model Parameters*						Outcome Model Parameters*						$\phi_0^{\dagger}$	$\phi_1^{\ddagger}$	Gelman-Rubin
				Covariate missingness											
Linear Mixed Model, MAR	0.27	0.07	0.54	0			1.01	1.00	1.02	1.01			-	-	1.03
Linear Mixed Model, LMAR: $b^{\#}$	0.24	0.06	0.52	0			1.01	1.00	1.01	1.01			1.00	1.03	1.04
Cure Model, MCAR	0.20	0.03	0.46	0.06	0.45		1.01	1.00	1.02	1.00	1.01		-	-	1.04
Cure Model, LMAR: G	0.16	0.02	0.36	0.07	0.52		1.01	1.00	1.01	1.00	1.02		1.00	1.00	1.04
Mixture of Normals, MAR	0.17	0.06	0.39	0.39	0.32	0.32	1.00	1.02	1.01	1.00	1.01	1.01	-	-	1.02
Mixture of Normals, LMAR: C	0.33	0.10	0.62	0.22	0.14	0.14	1.00	1.02	1.01	1.00	1.01	1.00	1.00	1.01	1.02
				Outcome missingness											
Linear Mixed Model, MCAR	0.18	0.22	0.21	0.49			1.00	1.01	1.01	1.03			-	-	1.05
Linear Mixed Model, LMAR: b	0.19	0.22	0.20	0.57			1.01	1.01	1.01	1.07			1.00	1.10	1.14
Linear Mixed Model, LMAR: b	0.18	0.21	0.22	0.52			1.01	1.01	1.01	1.04			1.01	FIXED	1.06
Mixture of Normals, MCAR	0.52	0.54	0.53	0.54	0.54	0.54	1.02	1.03	1.02	1.02	1.02	1.01	-	-	1.06
Mixture of Normals, LMAR: C	0.57	0.57	0.57	0.46	0.46	0.47	1.02	1.02	1.01	1.01	1.03	1.01	1.55	1.65	1.66
Mixture of Normals, LMAR: C, X	0.58	0.62	0.57	0.47	0.48	0.47	1.02	1.02	1.02	1.01	1.03	1.01	1.38	1.64	1.65
Mixture of Normals, LMAR: C	0.68	0.68	0.66	0.35	0.35	0.35	1.04	1.04	1.02	1.01	1.03	1.01	1.05	FIXED	1.13
Mixture of Normals, LMAR: C, X	0.67	0.71	0.65	0.35	0.37	0.34	1.04	1.03	1.02	1.01	1.03	1.01	1.01	FIXED	1.08

† Imputations drawn using kernels (2)–(3)

*For each model, these are the parameters from the outcome model (same as **Tables S1 - S3**):

— Linear mixed model: intercept, X_1, X_2, and time

— Cure model: intercept, X_1, and X_2 in the logistic regression and X_1 and X_2 in the Cox regression

— Mixture of Normals: intercept, X_1, and X_2 for the $C = 1$ and $C = 0$ classes respectively

‡ ϕ_0 is the intercept in the missingness model. ϕ_1 is the parameter for the latent variable.

# Notation: True and working missingness models depend on variables after colon

S4.5 Simulation 5: Comparison of final analysis with and without imputed L

After imputation, we have several choices as to what combination of the imputed L and D we want to include in the final analysis. We first suppose that we will perform our final analysis ignoring the contribution of R. When both imputed D and L are included in the final analysis, R is ignorable. In *Property 4*, we show that R is also ignorable if only imputed D is included in the final analysis when missingness is MAR. When missingness is LMAR, we show in *Property 5* that final analysis using only the imputed D and ignoring R will be valid but not fully efficient. In this section, we want to briefly explore the practical impact of including or excluding the imputed values of L (assuming we are ignoring R) in the final analysis through simulation.

We generate simulated data under a linear mixed model, mixture of normals model, and Cox proportional hazards model as described for Simulations 1-3. We impose either MAR or LMAR (Strong Dependence) missingness in X_2 as in Simulations 1-3 and impute using a working missingness model with the correct structure (either MAR or LMAR dependent only on the latent variable) and kernels (2)–(4). After imputation, we perform the final analysis using the imputed values for X_2 and either ignoring or using the imputed values for the latent variable (and in both cases ignoring R). Additionally, in the course of our simulations, we observed that some simulations under the mixture of normals model had estimated variances that were very large when we used the imputed latent variable in the final model fit. This may be an indicator of inadequate convergence of the model fit. Therefore, we present the mixture of normals results 1) for all 500 simulations and 2) restricting to simulations in which the estimated variances were all less than 0.2 (20 in the scale presented in the table). This issue did not arise for the linear mixed model simulations. In **Tables S1-S3**, we perform all final analyses ignoring the imputed latent variable and without restricting to simulations that have variance < 0.2, and the corresponding rows in this table are the same as the results in **Tables S1-S3**.

Table S5 shows the simulation results. We first consider the results for the mixture of normals model. We first notice that analyses using the imputed latent variable in the final analysis result in substantial bias when we include all simulations in our estimation of bias. This is the result of just a few simulations with parameter estimates far from the true value. This suggests some instability or lack of convergence in the model fitting. However, when we restrict our focus to simulations that appear to have convergence (reasonable standard errors), we see that final analyses including and excluding the imputed latent variable perform similarly well. For some simulation settings, the variance estimates using C are slightly larger, and the reverse is true for other simulations, so there is not a clear trend in efficiency including or excluding the latent variable in the final analysis in these simulations.

Although not included in our results, it is worth mentioning that analysis including and ignoring the imputed L may be associated with different fractions of missing information, which could have implications on the number of imputations needed for good inference. Let $\bar{U}$ represent the average of the variance estimators for parameter θ across the m imputed datasets and B represent the sample variance of the estimates of θ across the m imputed datasets. Then, we can express the relative increase in variation due to the missing data (r) and the fraction of missing information (λ) as (Schafer, 1999):

$$r = \frac{(1+\frac{1}{m})B}{\bar{U}} \qquad \lambda = \frac{r}{1+r}$$

The relative efficiency of an estimate θ based on m imputations compared to the estimate based of in infinite number of imputations is:

$$RE = \frac{1}{1 + \frac{\lambda}{m}}$$

We may expect an analysis that conditions on the imputed L in the final analysis to have larger relative between imputation variance vs. within imputation variance (r) compared to an analysis that does not condition on L in the final analysis for some parameters. This is because, when we include L in the final analysis, each fit treats the imputed L as known, resulting in substantially reduced "within imputation" standard error estimates for some parameters. This leads to larger values for the fraction of missing information, λ, for the same value of m when we include L in the final analysis compared to an analysis that ignores imputed L. In simulations (not shown), a final analysis using L did result in larger fractions of missing information compared to an analysis ignoring imputed L in the random intercept linear mixed model setting. We note that in practice this may translate into only a very small difference in relative efficiency between the two methods of analysis. However, several authors have noted practical issues regarding estimation of p-values and confidence intervals when a small number of imputations are used and the fraction of missing information is moderate to large (e.g. White and Royston, 2011; Bodner, 2008). Therefore, we may prefer to perform our final analysis using only the imputed D in the final analysis in an attempt to reduce the potential negative impact of larger fractions of missing information.

Table S5: Bias and variance of parameter estimates under different final analyses

	Linear mixed model							
Model$^{\#}$	Analysis	Intercept	X_1	X_2	Time			SIMS
		Bias (Var)†	Bias (Var)	Bias (Var)	Bias (Var)			
MAR	Ignoring b	0 (1.81)	0 (1.10)	-1 (2.58)	0 (0.1055)			500
MAR	Using b	0 (1.84)	0 (1.11)	-1 (2.59)	0 (0.1055)			500
LMAR	Ignoring b	0 (1.49)	0 (1.15)	-1 (2.03)	0 (0.1055)			500
LMAR	Using b	0 (1.50)	0 (1.08)	-1 (2.05)	0 (0.1055)			500
	Mixture of normals							
		- - - - - - - - - - C = 1- - - - - - - - - -			- - - - - - - - - C = 0 - - - - - - - - -			SIMS
Model$^{\#}$	Analysis	Intercept	X_1	X_2	Intercept	X_1	X_2	
		Bias (Var)	Bias (Var)	Bias (Var)	Bias (Var)	Bias (Var)	Bias (Var)	
Variance Unrestricted								
MAR	Ignoring C	1 (1.98)	0 (4.35)	1 (4.64)	-1 (6.18)	-2 (7.80)	0 (5.53)	500
MAR	Using C	2 (2.22)	6 (7.55)	5 (6.38)	-3 (7.09)	-8 (10.38)	-3 (6.49)	500
LMAR	Ignoring C	1 (2.89)	1 (4.63)	0 (6.42)	-1 (4.56)	-2 (5.91)	0 (5.29)	500
LMAR	Using C	1 (3.28)	9 (9.71)	5 (7.76)	-3 (5.18)	-9 (10.03)	-4 (7.18)	500
Variance Restricted*								
MAR	Ignoring C	1 (2.04)	-1 (2.09)	0 (2.88)	-1 (6.06)	0 (5.34)	0 (5.02)	483
MAR	Using C	1 (2.07)	-1 (1.95)	0 (2.94)	0 (5.73)	-1 (5.14)	0 (4.25)	404
LMAR	Ignoring C	0 (2.79)	0 (2.07)	0 (5.39)	0 (4.28)	0 (3.77)	1 (3.94)	477
LMAR	Using C	0 (2.95)	0 (2.11)	0 (5.44)	0 (4.35)	-1 (3.75)	0 (3.95)	418

† All values in table multiplied by 100

# Indicates true and working missingness model

* Ignoring simulations in which the estimated variance was greater than 0.2 (20 in the scale of this table) for at least one parameter.

S4.6 Simulation 6: more explorations for the CPH cure model

Suppose missingness is LMAR but that we impute incorrectly assuming MAR missingness. Since missingness is MNAR, we may have bias in estimating the resulting model parameters when we impute assuming MAR. This bias can be seen for missingness mechanisms strongly dependent on the latent variable in Simulations 1 (linear mixed model) and 3 (mixture of normals). In Simulation 2, however, MAR-based imputation under true LMAR missingness does not appear to create much bias in the resulting outcome model parameter estimates. In this section, we provide some intuition as to why this might be the case.

One reason for this result has to do with the form of the imputation distribution for G under LMAR. Under MAR and using notation from **Section S7.3**, we impute G from

$$\text{logit}(P(G_i = 1|X_i, T_i, \delta_i = 0; \rho)) = \omega_0 + \omega_1 X_i - \Lambda_0(T_i)e^{\theta X_i}$$

and we impute G from the following under LMAR:

$$\text{logit}(P(G_i = 1|X_i, T_i, \delta_i = 0, R_i; \nu)) = \omega_0 + \omega_1 X_i - \Lambda_0(T_i)e^{\theta X_i} + \log\left[\frac{f(R_i^{-S}|T_i, \delta_i = 0, X_i^{(obs)}, G_i = 1; \phi^{-S})}{f(R_i^{-S}|T_i, \delta_i = 0, X_i^{(obs)}, G_i = 0; \phi^{-S})}\right]$$

These two distributions differ only by the offset term, $\log\left[\frac{f(R_i^{-S}|T_i,\delta_i=0,X_i^{(obs)},G_i=1;\phi^{-S})}{f(R_i^{-S}|T_i,\delta_i=0,X_i^{(obs)},G_i=0;\phi^{-S})}\right]$. This offset term follows the familiar form of the offset term under case-control sampling dependent on G.

Suppose first that missingness depends only on G. In this case, the two distributions (imputation under MAR vs. LMAR) differ by a term depending only on R and ϕ^{-S}. Since R and X are independent given G in this setting, exclusion of this offset term will impact the intercept but may not appreciably impact the estimated covariate effects, ω_1. Therefore, the imputation distributions under MAR and LMAR are really only different in terms of the population cure rate, which is associated with the intercept in the logistic regression part of the Cox proportional hazards cure model.

Suppose instead that missingness depends on both G and observed X. In this case, the offset term will be correlated with X, so exclusion of the offset term by incorrectly assuming MAR could more appreciably impact the imputation distribution for G. This in turn may more strongly impact the resulting inference. This is loosely supported by results from mechanism (D) in Simulation 2, but we still don't see much difference between MAR-based and LMAR-based imputation under LMAR in this setting.

An additional reason for similarity between the MAR-based and LMAR-based imputation results under true LMAR mechanisms may be the actual rate of missingness in G (the partially latent cure status). Since the only difference between MAR-based and LMAR-based imputation is the distribution used to impute G, we might expect the imputation distribution to have a bigger impact on inference when we have a larger degree of missingness in G. Recall, subjects having an observed recurrence are known to be non-cured ($G = 1$), and we assume subjects with long follow-up and no recurrence are cured ($G = 0$). Therefore, it is the non-recurring subjects who are censored early that have unknown cure status. We might expect that a heavier degree of censoring (resulting in a greater proportion of subjects with missing G) may produce greater bias resulting from imputing incorrectly assuming MAR. We performed an additional set of simulations (not shown) to explore how the degree of censoring impacted the relative performance of MAR-based and LMAR-based imputation when missingness is truly LMAR. We found that the bias did not appreciably increase with greater degrees of censoring as we had predicted. Indeed, large amounts of censoring (which produce greater degrees of missingness in G) are associated with difficulty in estimating cure model parameters. In settings where we might expect a lot of bias, therefore, we couldn't even feasibly fit the model of interest, and settings in which the censoring mechanism was such that we did not run into numerical issues produced little difference between MAR-based and LMAR-based imputation.

One subtle reason for this lack-of-bias phenomenon is that LMAR-based missingness is truly a small step from MAR missingness in the cure model setting. Intuitively, we might expect some bias from incorrectly assuming MAR when missingness is LMAR. However, one distinguishing feature of the cure model is that G is partially observed. When G is observed, it is always equal to the observed event status indicator, δ. Suppose we observed recurrences for every single non-cured subject. In this case, LMAR missingness is actually MAR, since δ represents the true cure status. This will usually not be the case, but the close relationship between G and δ may be enough to protect against bias resulting from ignoring LMAR missingness. We might think of δ as a messy measure of G. By conditioning on δ, we might make the MAR assumption more reasonable.

While it is possible to have induced bias due to ignoring LMAR missingness in the Cox proportional hazards cure model setting, this bias resulting from ignoring the latent-dependent missingness mechanism is therefore generally expected to be somewhat small. This is demonstrated in Simulation 2 and in the analysis of the head and neck cancer data in the main paper. Although not shown, additional simulations suggest that greater amounts of censoring either do not produce much bias or create numerical issues with estimating the cure model parameters (so the data themselves are not well-suited for modeling using a cure model). We also considered different degrees of dependence between missingness and cure status and fully observed covariates. In all settings explored, we still saw relatively little bias created by incorrectly assuming MAR under LMAR missingness. Overall, we may be less worried about the impact of ignoring latent-dependent missingness in the Cox proportional hazards cure model setting (and possibly other settings with partially observed latent variables) compared do settings in which the latent variable is never observed.

S5 Example 1: identifiability for joint normal models

In this paper, we restrict applications of the proposed methods to cases in which the model parameters would be identified had the missing data been observed. Here, we present an example in which parameters identified in the LMAR-based model would not be identified if the missing data had been observed. In particular, we first explore assumptions required to achieve identifiability for a measurement error model. Then, we compare the measurement error model to linear mixed models and explain how the linear mixed model is able to attain identifiability of all outcome model parameters.

S5.1 Example 1.1: Measurement Error Model with Covariates

Suppose we have a noisy version (Y) of an underlying variable of interest, L. L is never observed, and Y is observed at least for some subjects. We suppose Y and L are univariate and related to fully measured covariates, X. Suppose we model

$$Y_i = \alpha_0 + \alpha_1 L_i + \alpha_2 X_i + e_i, \quad L_i \sim N(\beta_0 + \beta_1 X_i, \Sigma_L), \quad e_i \sim N(0, \sigma^2), \quad e_i \perp L_i$$

This is an example of a measurement error model. This model contains 7 parameters. This implies the following:

$$\begin{pmatrix} Y_i \\ L_i \end{pmatrix} |X_i = N\left(\begin{pmatrix} \alpha_0 + \alpha_1(\beta_0 + \beta_1 X_i) + \alpha_2 X_i \\ \beta_0 + \beta_1 X_i \end{pmatrix}, \begin{pmatrix} \sigma^2 + \alpha_1^2 \Sigma_L & \alpha_1 \Sigma_L \\ \alpha_1 \Sigma_L & \Sigma_L \end{pmatrix} \right)$$

$$L_i | Y_i, X_i \sim N\left(\beta_0 + \beta_1 X_i + \frac{\alpha_1 \Sigma_L}{\sigma^2 + \alpha_1^2 \Sigma_L} \left[Y - \alpha_0 - \alpha_1(\beta_0 + \beta_1 X_i) - \alpha_2 X_i \right], \right.$$
$$\left. \Sigma_L - \frac{\alpha_1^2 \Sigma_L^2}{\sigma^2 + \alpha_1^2 \Sigma_L} \right)$$

Suppose we have no missingness in Y. In this case, the observed data likelihood can be expressed as follows:

$$Lik_{NoMissing}^{(obs)} = \prod_{i=1}^{n} f(Y_i|X_i) = \prod_{i=1}^{n} N\left(Y_i; \alpha_0 + \beta_0\alpha_1 + [\alpha_2 + \alpha_1\beta_1]X_i, \sigma^2 + \alpha_1^2 \Sigma_L\right)$$

where $N(a; b, c)$ indicates the normal density evaluated at a with mean b and variance c. In order for the model to be identified, **we must fix 4 of the 7 parameters** in this model $(\alpha_0, \alpha_1, \alpha_2, \sigma^2, \beta_0, \beta_1, \Sigma_L)$, so we can identify the 3 remaining parameters.

Suppose instead that we have LMAR missingness in Y is follows: Probit$(P(R_i^Y = 1|L_i, Y_i, X_i)) = \phi_0 + \phi_1 L_i$, so we assume that missingness in Y only depends on L. This scenario is a simple case of the Heckman (1976) selection model if $\alpha_1 = 0$ with a modified missingness model (Little and Rubin, 2002; Heckman, 1976). The observed data likelihood can be expressed as follows:

$$Lik^{(obs)} = \prod_{i=1}^{n} \left[\int \Phi(\phi_0 + \phi_1 L_i) f(Y_i, L_i|X_i) dL_i \right]^{R_i^Y} \left[\int (1 - \Phi(\phi_0 + \phi_1 L_i)) f(L_i|X_i) dL_i \right]^{1-R_i^Y}$$

$$= \prod_{i=1}^{n} \left[f(Y_i|X_i) \int \Phi(\phi_0 + \phi_1 L_i) f(L_i|Y_i, X_i) dL_i \right]^{R_i^Y} \left[1 - \int \Phi(\phi_0 + \phi_1 L_i) f(L_i|X_i) dL_i \right]^{1-R_i^Y}$$

$$= \prod_{i=1}^{n} \left[f(Y_i|X_i) E_{L|Y,X}\left(\Phi(\phi_0 + \phi_1 L_i)\right) dL_i \right]^{R_i^Y} \left[1 - E_{L|X}\left(\Phi(\phi_0 + \phi_1 L_i)\right) \right]^{1-R_i^Y}$$

We will make use of the following identity:
Let $U \sim N(\mu_1, \sigma_1^2)$ and $V \sim N(\mu_2, \sigma_2^2)$ be independent random variables. Now, $U - V \sim$

$N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$.

$$\Phi\left(\frac{-(\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) = P(U \leq V) = \int \Phi\left(\frac{v - \mu_1}{\sigma_1}\right) f_V(v) dv = E_V\left(\Phi\left(\frac{v - \mu_1}{\sigma_1}\right)\right)$$

Using this identity and setting $\sigma_1 = 1/\phi_1$ and that $\mu_1 = -\phi_0/\phi_1$, we have that

$$Lik^{(obs)} = \prod_{i=1}^{n} \left[1 - \Phi\left(\frac{\phi_0 + \phi_1(\beta_0 + \beta_1 X_i)}{\sqrt{\phi_0^2 + \phi_1^2 \Sigma_L}}\right)\right]^{1-R^Y}$$

$$\times \left[f(Y_i|X_i)\Phi\left(\frac{\phi_0 + \phi_1(\beta_0 + \beta_1 X_i) + \phi_1\alpha_1\Sigma_L\left[\sigma^2 + \alpha_1^2\Sigma_L\right]^{-1}(Y_i - \alpha_0 - \alpha_1(\beta_0 + \beta_1 X_i) - \alpha_2 X_i)}{\sqrt{\phi_0^2 + \phi_1^2(\Sigma_L - \alpha_1^2\Sigma_L^2\left[\sigma^2 + \alpha_1^2\Sigma_L\right]^{-1})}}\right)\right]^{R^Y}$$

This expression contains 9 parameters, but we cannot simultaneously identify all parameters. Suppose we set

$$\begin{array}{llll} A = \phi_1\alpha_1\Sigma_L & B = \sigma^2 + \alpha_1^2\Sigma_L & C = \alpha_0 + \alpha_1\beta_0 & D = \alpha_1\beta_1 + \alpha_2 \\ E = \phi_0 + \phi_1\beta_0 & F = \phi_1\beta_1 & G = \phi_0^2 + \phi_1^2\Sigma_L & \end{array}$$

Then we can rewrite the observed data likelihood as:

$$Lik^{(obs)} = \prod_{i=1}^{n} \left[N(Y_i; C + DX_i, B)\Phi\left(\frac{E + FX_i + \frac{A}{B}(Y_i - C - DX_i)}{\sqrt{G - \frac{A^2}{B}}}\right)\right]^{R^Y} \left[1 - \Phi\left(\frac{E + FX_i}{\sqrt{G}}\right)\right]^{1-R^Y}$$

Therefore, we can represent the 9 parameters as 7 parameters in the expression for the observed data likelihood, and the 7 parameters are estimable. We must fix 2 parameters in order for the remaining parameters to be (weakly) identified.

Suppose we fix ϕ_0 and ϕ_1. Then we can (weakly) identify all 7 remaining parameters under LMAR. However, suppose that we had observed Y for all subjects. In this case, we would need fix 4 parameters out of $(\alpha_0, \alpha_1, \alpha_2, \sigma^2, \beta_0, \beta_1, \Sigma_L)$ in order for the remaining 3 parameters to be identified. Therefore, the model fit without any outcome missingness requires some parameters to be fixed that do not need to be fixed in the LMAR-based model in order to achieve (weak) identifiability. Curiously, we have more information about the parameter set under LMAR than if we had observed Y for all subjects. It is worth noting that when we instead fix four parameters in $(\alpha_0, \alpha_1, \alpha_2, \sigma^2, \beta_0, \beta_1, \Sigma_L)$, the resulting parameters $A - G$ will be overidentified, but this should not present any problems.

It is important to note that we cannot verify the form of the missingness model, and here assumed missingness model results in additional parameters becoming identifiable under LMAR. Therefore, the identification is a direct result of unverifiable assumptions, and an analysis that relies on the missingness model being correct such that the outcome model parameters would not be identified if the model were incorrect seems untrustworthy. This provides further justification for excluding situations in which the parameters would not be identifiable if there was not covariate or outcome missingness.

While technically identified, our imputation algorithm leads to convergence problems when imputing under this LMAR model with only two fixed parameters (simulations not shown). If we fix additional parameters, the proposed imputation algorithm has better performance. In general, we do not expect our imputation algorithm to perform well in settings where the model would not be identified or would be very weakly identified if there were no covariate/outcome missingness. In such settings, we recommend fixing additional parameters to achieve good identification properties before performing the proposed imputation algorithm.

S5.2 Example 1.2: linear mixed model example

We notice that the form of the measurement error model in the previous section is similar to the usual structure of a linear mixed model with a random intercept except that the outcome in the linear mixed model case is multivariate. Suppose we observe $K > 1$ values of Y for each subject and we assume that elements of Y within subjects are independent conditional that subject's covariates and the random intercept. We model:

$$Y_i | X_i, L_i \sim N_K(\alpha_0 + \mathbf{1}_K b_i + \alpha_2 X_i, \sigma^2 \mathbb{I}_K), \qquad b_i \sim N(0, \Sigma_L)$$

Here, $\mathbf{1}_K$ corresponds to α_1 in the previous measurement error model. Additionally, this model assumes that $\beta_0 = \beta_1 = 0$. Therefore, three parameters from the model in the previous section are fixed by design. The modeling assumptions imply the following joint distribution:

$$\begin{pmatrix} Y_i \\ L_i \end{pmatrix} = N\left(\begin{pmatrix} \alpha_0 + \alpha_2 X_i \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 \mathbb{I}_K + \mathbf{1}_K \Sigma_L \mathbf{1}_K^T & \mathbf{1}_K \Sigma_L \\ \mathbf{1}_K^T \Sigma_L & \Sigma_L \end{pmatrix}\right)$$

Suppose we have no missingness in Y. In this case, the observed data likelihood can be expressed as follows:

$$Lik^{(obs)}_{NoMissing} = \prod_{i=1}^{n} MVN_K\left(Y_i; \alpha_0 + \alpha_2 X_i, \sigma^2 \mathbb{I}_K + \mathbf{1}_K \Sigma_L \mathbf{1}_K^T\right)$$

We can identify all four of these model parameters. We compare this to the situation with the measurement error model with covariates in which 4 out of the 7 parameters needed to be fixed in order to achieve identifiability. In this case, three of the 7 parameters are fixed by design ($\alpha_2 = \mathbf{1}_K, \beta_0 = \beta_1 = 0$), and we can identify an additional parameter due to the compound symmetric structure of the variance for $Y|X$ resulting from the repeated measures within individuals. In this case, the model under no outcome or covariate missingness is well-identified, and the proposed imputation approach can perform well under some MAR and LMAR missingness scenarios.

S6 Example 2: identifiability under LMAR for a mixture of GLMs

In this section, we explore issues of identifiability for another simple modeling scenario. Unlike the measurement error example, this example demonstrates a situation in which the model is fully identified under no covariate/outcome missingness but has issues with identifiability under a simple LMAR missingness mechanism. We present simulations demonstrating evidence of identifiability-related numerical issues.

Suppose our model for outcome Y is a mixture of two GLMs and let C represent the fully latent mixing variable. Within each latent class, we model the relationship between Y and covariates X using a GLM. We will assume that $C \perp X$ with $P(C_i = 1|X_i) = \omega$. We first suppose there is no covariate/outcome missingness. The observed data likelihood takes the following form:

$$Lik^{(obs)}_{NoMissing} \propto \prod_{i=1}^{n} [\omega f(Y_i, |X_i, C_i = 1; \theta) + (1-\omega) f(Y_i, |X_i, C_i = 2; \theta)]$$

Assuming the distribution of $Y|X, C$ depends on C and is an identifiable GLM in its own right, then θ and ω are both identifiable.

Suppose now that we have latent-dependent missingness in the outcome for some subjects. Let R^Y be a vector of indicators representing the response of Y. Let ϕ be the parameter attached to the missingness model. We define $p_j(\phi) = P(R_i = 1|X_i, C_i = j; \phi)$ for latent classes $j = 1, 2$. We can write the observed data likelihood as follows:

$$\begin{aligned} Lik^{(obs)}(\nu) &\propto \prod_{i=1}^{n} \int\int f(R_i^Y|X_i, L_i; \phi) f(Y_i, |X_i, C_i; \theta) f(C_i|X_i; \omega) dY_i^{(mis)} dC_i^{(mis)} \\ &\propto \prod_{i=1}^{n} [p_1(\phi) f(Y_i, |X_i, C_i = 1; \theta)\omega + p_2(\phi) f(Y_i, |X_i, C_i = 2; \theta)(1-\omega)]^{R_i^Y} \\ &\quad \times [(1 - p_1(\phi))\omega + (1 - p_2(\phi))(1-\omega)]^{1-R_i^Y} \end{aligned}$$

S6.1 Example 2.1: R^Y is Independent of X (Nonidentifiable Model)

First, we assume that R^Y is independent of X, so it only depends on C. Define $p_1(\phi) = \text{expit}(\phi_0 + \phi_1)$ and $p_2(\phi) = \text{expit}(\phi_0)$. We can write the observed data likelihood as:

$$\begin{aligned} Lik^{(obs)}(\nu) \propto \prod_{i=1}^{n} &\left[\frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}} \omega f(Y_i, |X_i, C_i = 1; \theta) + \frac{e^{\phi_0}}{1+e^{\phi_0}}(1-\omega) f(Y_i, |X_i, C_i = 2; \theta) \right]^{R_i^Y} \\ &\left[-\frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}} \omega + 1 - \frac{e^{\phi_0}}{1+e^{\phi_0}}(1-\omega) \right]^{1-R_i^Y} \end{aligned}$$

This likelihood can be reparameterized using $A = \frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}}\omega$ and $B = \frac{e^{\phi_0}}{1+e^{\phi_0}}(1-\omega)$, so we can represent three of the model parameters using just two parameters. Therefore, we will not be able to identify all three of ϕ_1, ϕ_0, and ω, but A and B can be identified. We suppose that θ is of primary interest. In this example, we can still identify θ even though we cannot identify ϕ_1, ϕ_0, and ω. We note that under MAR, $\phi_1 = 0$, and both ϕ_0 and ω are identified.

Under LMAR, we can identify A and B, but we cannot identify ϕ_1, ϕ_0, and ω. We want to know whether A and B are enough to perform the imputation of C and Y. In order to impute Y, we will draw from $f(Y_i|X_i, C_i)$, which does not involve ω or ϕ. We would impute C using:

$$P(C_i = 1|X_i, Y_i) = \left[\frac{\frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}}\omega f(Y_i|X_i, C_i = 1)}{\frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}}\omega f(Y_i|X_i, C_i = 1) + \frac{e^{\phi_0}}{1+e^{\phi_0}}(1-\omega) f(Y_i|X_i, C_i = 2)}\right]^{R_i^Y}$$
$$\times \left[\frac{\omega(1 - \frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}}) f(Y_i|X_i, C_i = 1)}{\omega(1 - \frac{e^{\phi_0+\phi_1}}{1+e^{\phi_0+\phi_1}}) f(Y_i|X_i, C_i = 1) + (1-\omega)(1 - \frac{e^{\phi_0}}{1+e^{\phi_0}}) f(Y_i|X_i, C_i = 2)}\right]^{1-R_i^Y}$$

When we impute C and Y was observed, we are imputing using only functions of the parameters that ARE identifiable. However, imputation when Y is missing requires parameters that are not strictly identifiable. This may result in numerical issues within the imputation algorithm.

While we cannot identify all three of ϕ_1, ϕ_0, and ω, we can identify the other two parameters if we hold one parameter fixed. This provides a suggestion for imputation under this unidentifiable model. We can fix values of one of the parameters and then perform imputation. We can repeat this for different values of the fixed parameter and explore the impact of the fixed parameter on model inference.

S6.2 Example 2.2: R^Y Depends on X (Identifiable Model)

Now, we assume that R^Y is not independent of X. Suppose we model $p_1(\phi) = \text{expit}(\phi_0 + \phi_1 X_i + \phi_2)$ and $p_2(\phi) = \text{expit}(\phi_0 + \phi_1 X)$. We can write

$$Lik(\nu) \propto \prod_{i=1}^{n} \left[\frac{e^{\phi_0+\phi_1 X_i+\phi_2}}{1+e^{\phi_0+\phi_1 X_i+\phi_2}}\omega f(Y_i, |X_i, C_i = 1; \theta) + \frac{e^{\phi_0+\phi_1 X_i}}{1+e^{\phi_0+\phi_1 X_i}}(1-\omega) f(Y_i, |X_i, C_i = 2; \theta)\right]^{R_i^Y}$$
$$\times \left[-\frac{e^{\phi_0+\phi_1 X_i+\phi_2}}{1+e^{\phi_0+\phi_1 X_i+\phi_2}}\omega + 1 - \frac{e^{\phi_0+\phi_1 X_i}}{1+e^{\phi_0+\phi_1 X_i}}(1-\omega)\right]^{1-R_i^Y}$$

When ϕ_1 is nonzero, we can identify the model parameters. Therefore, additional complexity in the missingness mechanism results in an identifiable model.

S6.3 Simulation using nonidentifable model

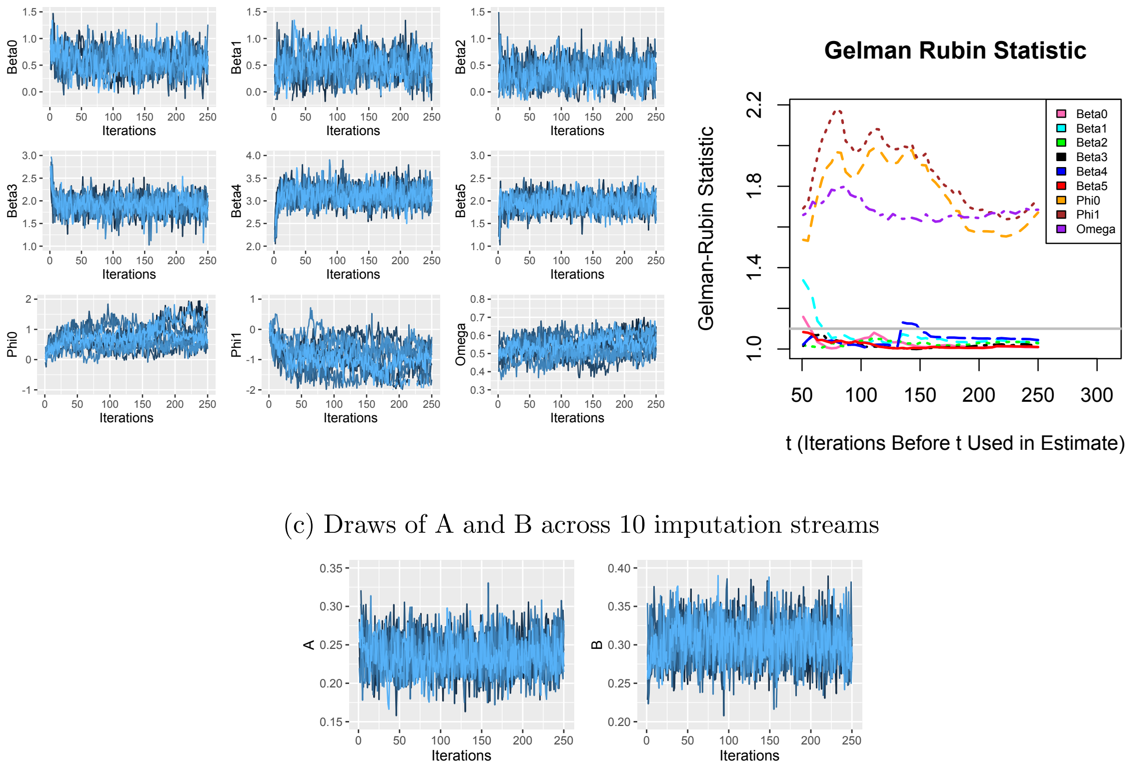
We simulate a single dataset under a mixture of linear regressions model as in Simulation 3. We impose outcome missingness using the relation $\text{logit}(P(Y \text{ is observed}|X_1, X_2, C, Y)) = \phi_0 + \phi_1 C$ where $\phi_0 = 1.1$ and $\phi_1 = -1.7$. Therefore, we have that $p_1(\phi) = \text{expit}(-0.6)$ and $p_2(\phi) = \text{expit}(1.1)$ (using notation from **Section S6.1**). This is a LMAR mechanism. Define β to be the parameters of $f(Y|X, C)$ and $\omega = P(C = 1|X)$.

We first perform our imputation algorithm using a correct working model structure but without fixing values for ϕ_0 and ϕ_1. Previously, we showed in **Section S6.1** that the parameters ϕ_0, ϕ_1, and ω are not all identifiable. However, at each iteration of the imputation algorithm, we can draw values of these three parameters. We perform 10 streams of our imputation algorithm in which we impute values of Y and L. **Figure S1(a)** shows the parameter draws for each iteration of the imputation algorithm. Different imputation streams are shown with differently colored lines.

Figure S1: Drawn parameters in nonidentifiable model with no fixed parameters

(a) Parameter draws across 10 imputation streams

(b) Gelman-Rubin diagnostics



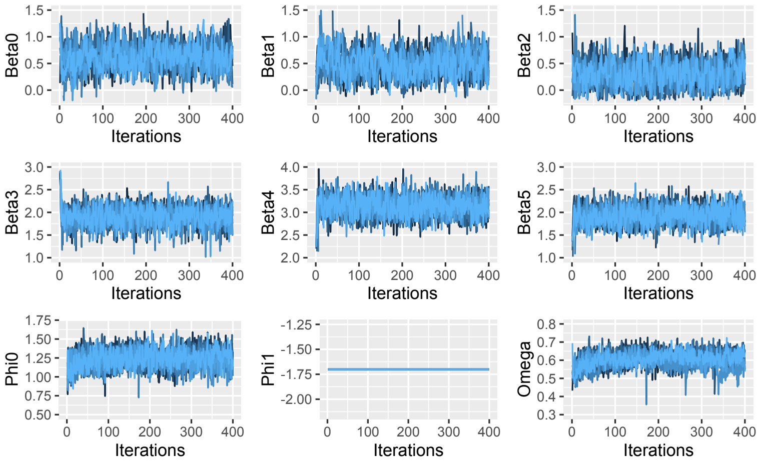
(c) Draws of A and B across 10 imputation streams

Visually, we can see that we have some issues with convergence for ϕ_1, ϕ_0, and ω. However, the draws for the β parameters (the parameters ultimately of interest) appear to converge. One criterion for evaluating the convergence is the Gelman-Rubin statistic Gelman and Rubin (1992). This statistic is calculated by comparing the variation of the parameter draws within each stream to the variation between streams. For good algorithms, the value of this statistic should move toward 1 as the number of iterations increases, and values greater than 1.1 are generally considered to represent insufficient convergence. **Figure S1(b)** shows the estimated Gelman-Rubin statistic for several model parameters across iterations of the imputation algorithm. We do not include the first 50 iterations in the calculations. The gray line represents a Gelman-Rubin statistic of 1.1. While the draws for the β parameters are converging, we do not see convergence for ϕ_0, ϕ_1, and ω. While we cannot identify ϕ_0, ϕ_1, and ω, we previously showed (with a different parameterization) that functions A and B of these parameters are identifiable. **Figure S1(c)** shows the parameter draws for A and B, and we can see that these parameters appear to converge nicely even though ϕ_0, ϕ_1, and ω do not.

Even though ϕ_0, ϕ_1, and ω are not all simultaneously identifiable, the parameter related to the outcome model can be identified. In terms of the practical implications of identifiability issues on inference, this hints that we may still be able to obtain reasonable inference about the outcome model parameter in some cases. In this simulation, the β parameters do appear to converge to values that are very close to the true values even in the presence of convergence issues for the other parameters.

While we cannot identify ϕ_0, ϕ_1, and ω simultaneously, we can identify two of the parameters if we fix values of the third. Fixing ϕ_1, we perform imputation drawing values for all other parameters. **Figure S2** shows the resulting parameter draws across the 10 streams of imputation. When we fix ϕ_1, we see good numerical convergence properties for the other model parameters.

Figure S2: Parameter draws across 10 imputation streams when ϕ_1 is fixed



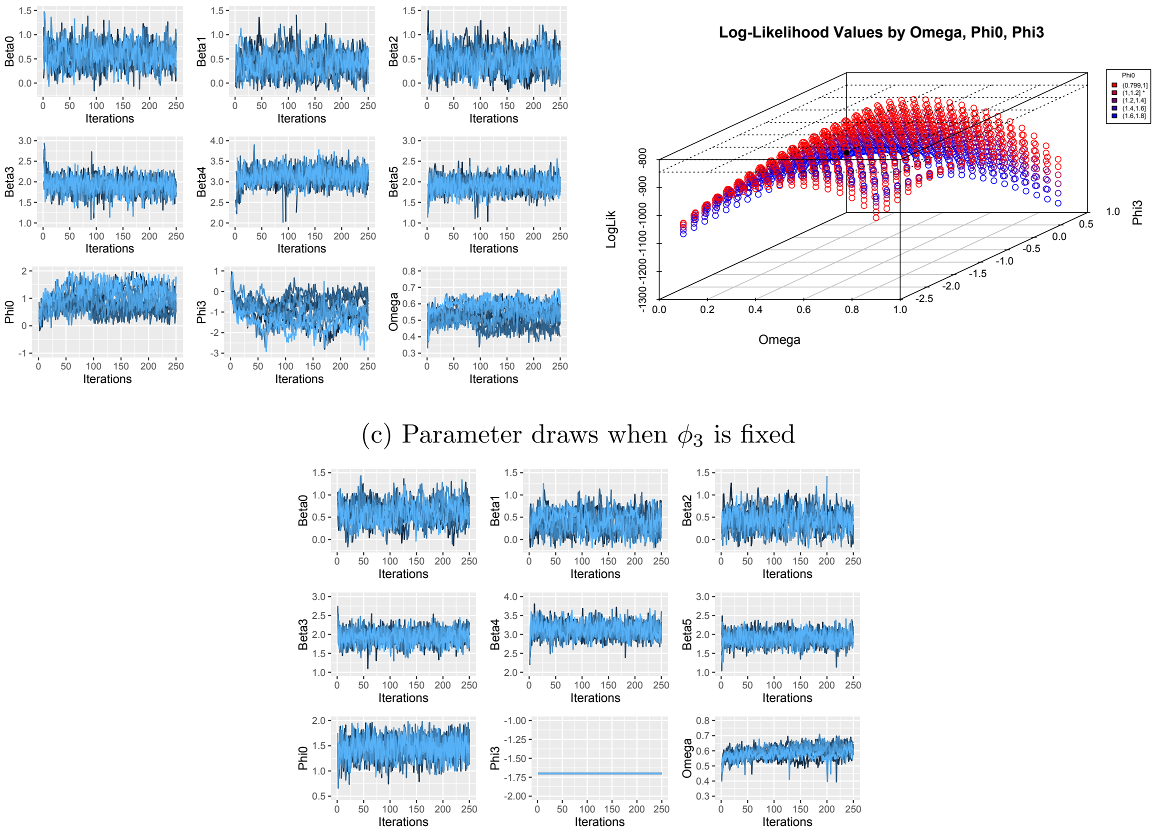
S6.4 Simulation using identifiable model

We now consider the setting where missingness in the outcome is generated using the relation $\text{logit}(P(Y \text{is observed}|X_1, X_2, C, Y)) = \phi_0 + \phi_1 X_1 + \phi_2 X_2 + \phi_3 C$ where $\phi_0 = 1.1$, $\phi_1 = 0.5$, $\phi_2 = -0.5$, and $\phi_3 = -1.7$. Again, this is a LMAR mechanism.

We first perform imputation of Y and L using the correct working model without fixing any parameter values. **Figure S3(a)** shows the parameter draws for the 10 imputation streams. We can see evidence of convergence issues for several model parameters. However, we still see that the parameters of interest in θ appear to converge nicely near their true values.

While the parameters may all be technically identifiable, we can sometimes run into problems when the observed data log-likelihood surface is nearly flat with respect to one or more parameters. **Figure S3(b)** shows the value of the observed data log-likelihood for different values of ϕ_0, ϕ_3, and ω using the true values for all other parameters. The plotted plane indicates the maximum of the observed data log-likelihood and the black dot indicates the true values for the parameters. Fixing ϕ_3 and ϕ_0, we can see that the shape of the log-likelihood with ω is fairly concave. However, the log-likelihood surface as a whole is fairly flat across different combinations of ϕ_0, ϕ_3, and ω. When we fix the value of ϕ_3, however, we can do a better job at estimating ω and ϕ_0, resulting in improved convergence performance as shown in **Figure S3(c)**.

Figure S3: Drawn parameters in identifiable model



(a) Parameter draws with no fixed parameters

(b) Log-Lik surface with respect to ϕ_0, ϕ_3, and ω

(c) Parameter draws when ϕ_3 is fixed

S7 Implementation of the SMC imputation algorithm

In this section, we provide specifics for how we can implement the proposed imputation algorithm for the three examples of latent ignorability considered in the main paper. In each case, we will use notation defined in the main paper and use R^{-S} as defined in *Property 3.* We will assume we are using flat priors for all model parameters. This assumption allows us to draw parameter values using maximum likelihood methods on bootstrap samples of the data. In this section, we do not make a distinction between the true and working models for $f(X^{(t)}|X^{(-t)};\psi)$ in the notation, but it should be understood that this may be a working version, and these covariate distributions together may not correspond to a valid joint distribution. Imputation using the proposed chained equations method will replace the steps for imputing the missing covariate and outcome data with steps where we specify and impute from corresponding regression models.

S7.1 Drawing from a Distribution Known up to Proportionality

In the main paper, we present distributions we can use to impute missing values for latent variables, but in some cases these distributions may only known up to proportionality. We call the form of the distribution known up to proportionality the "kernel" of the distribution. Many methods exist in the literature for drawing from a distribution knowing only the kernel. In this section, we will *briefly* describe two such methods.

Rejection sampling
The strategy of rejection sampling is to determine a easy-to-draw-from distribution that dominates a hard-to-draw-from distribution. We can then draw values from the hard-to-draw-from distribution by instead drawing from the easy-to-draw-from distribution distribution many times and accepting the first draw that satisfies a simple inequality. In more concrete terms, rejection sampling algorithms involve determining a simple density, $g(v)$, that dominates the distribution known up to proportionality, $k(v)$, such that we can write

$$k(v) \leq Kg(v) \;\; \forall \;\; v$$

where K is a constant greater than or equal to 1. Once we have specified a density $g(v)$ that dominates $k(v)$, we can obtain a draw V from $k(v)$ by performing the following:

1) Generate V from $g(v)$ and U from $U(0,1)$

2) Accept draw V if $U \leq \frac{k(V)}{Kg(V)}$. Otherwise, we reject draw V and return to 1) (Robert and Casella, 2004).

If $Kg(v)$ is much larger than $k(v)$, the rejection sampling algorithm may require many repetitions in order to accept a draw. Therefore, the choice of $g(v)$ and K is important to the efficiency of the imputation algorithm. In the following sections, we propose possible choices for K and $g(v)$ in specific settings, but more efficient choices may be available.

Rejection sampling methods for imputation knowing the distribution only up to proportionality were considered in Bartlett et al. (2014), which used dominating function $f(X_i^{(t)}|X_i^{(-t)};\psi)$ for covariate imputation. We can use a similar approach for covariate imputation as discussed below.

Metropolis-Hastings
Like the rejection sampling algorithm, the goal of the Metropolis-Hastings algorithm is to obtain a draw values of variable V from a distribution known only up to proportionality, $k(v)$. The strategy is to first specify a proposal distribution, $p(v|u)$, from which we

propose new values for the variable $V = v$ given the most recent drawn value of V, u. We can obtain a draw V from $k(v)$ by performing the following:

1) Generate v^* from $p(v|u)$. Generate $U \sim U(0,1)$

2) Define acceptance probability $\alpha = \min\left(1, \frac{p(u|v^*)k(v^*)}{p(v^*|u)k(u)}\right)$. Accept draw $V = v^*$ if $U \leq \alpha$.

Otherwise, we reject draw $V = v^*$ and keep $V = u$ (Robert and Casella, 2004). One popular choice of proposal distributions is a normal distribution centered at the most recent imputation u and with variance as a tuning parameter.

S7.2 Linear mixed model with random intercept

Suppose our outcome model is a linear mixed model with a latent random intercept, b_i. Let outcome Y_i be a vector of $K > 1$ normal outcomes and X_i be a $K \times d$ matrix containing a column of 1's and covariates for subject i. We model

$$Y_i|X_i, b_i \sim N_K(X_i\theta + 1_K b_i, \Sigma) \quad \text{and} \quad b_i|X_i \sim N(0, \omega^2)$$

We have the following joint distribution:

$$\begin{pmatrix} Y_i \\ b_i \end{pmatrix} = N\left(\begin{pmatrix} X_i\theta \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma + \mathbf{1}_K\omega^2\mathbf{1}_K^T & \mathbf{1}_K\omega^2 \\ \mathbf{1}_K^T\omega^2 & \omega^2 \end{pmatrix}\right)$$

In this modeling framework, random intercept b_i is missing for all subjects. Suppose we also have missingness in Y and X that may be MAR or LMAR. We also suppose that $\Sigma = \sigma^2\mathbb{I}_K$, so the outcomes are independent across subjects given b and X. We can use the imputation algorithm described below to impute missing values in b_i, X, and Y. We can initialize the missing values of the covariates by drawing from the observed values with equal probability. We can initialize the latent random intercept using the Best Linear Unbiased Predictors (BLUPs) from a complete case fit.

Imputation of latent variable

Assuming MAR

Under MAR and using (2), we want to impute missing b_i from

$$f(b_i|X_i, Y_i; \nu) \propto f(Y_i|X_i, b_i; \theta)f(b_i|X_i; \omega) = f(b_i|X_i, Y_i; \rho)$$

Using properties of multivariate normal random variables, we have that

$$f(b_i|X_i, Y_i; \rho) = N(\mathbf{1}_K^T\omega^2 \left[\Sigma + \mathbf{1}_K\omega^2\mathbf{1}_K^T\right]^{-1}(Y_i - X_i\theta), \omega^2 - \mathbf{1}_K^T\omega^2\left[\Sigma + \mathbf{1}_K\omega^2\mathbf{1}_K^T\right]^{-1}\mathbf{1}_K\omega^2)$$

We can draw values of Σ, ω^2, and θ by fitting a linear mixed model to a bootstrap sample of the most recently imputed data and then draw missing b_i from $f(b_i|X_i, Y_i; \rho)$.

Assuming LMAR

Under LMAR and using (2), we want to impute missing b_i from

$$f(b_i|X_i, Y_i, R_i^{-S}; \nu) \propto f(R_i^{-S}|Y_i^{(obs)}, X_i^{(obs)}, b_i; \phi^{-S})f(b_i|X_i, Y_i; \rho) \tag{S7.1}$$

This distribution depends on R_i^{-S}, the subset of R_i corresponding to variables that are LMAR. We must specify a model for R_i^{-S} given $Y_i^{(obs)}$, $X_i^{(obs)}$, and b_i. When R_i^{-S} contains missingness indicators for multiple variables (e.g. outcome at different time-points), this may be a challenging task. Several authors have discussed specification of this missingness model in the context of missingness dependent on random effects, and we will not discuss this choice further here (Wu and Carroll, 1988; Yang et al., 2008).

The distribution in (S7.1) is only known up to proportionality, but we can use one of the two above methods for drawing from a distribution knowing only the kernel. For example, we may use Metropolis-Hastings methods to draw values of b_i with a normal proposal distribution centered at the most recent imputed value of b_i and with some small variance, τ, which will be a tuning parameter. Given τ, the most recent imputation of D, and draws of ρ and ϕ, we can use the above kernel to impute b_i under LMAR.

Another option is to use rejection sampling. We note that $f(R_i^{-S}|Y_i^{(obs)}, X_i^{(obs)}, b_i; \phi^{-S})$ is a probability, so it is less than or equal to 1. We define

$$k(b_i) = f(R_i^{-S}|Y_i^{(obs)}, X_i^{(obs)}, b_i; \phi^{-S})f(b_i|X_i, Y_i; \rho)$$

and can define dominating function $g(b_i) = f(b_i|X_i, Y_i; \rho)$ with $K = 1$. $g(b_i)$ is a normal distribution with mean and variance as functions of model parameters, so this distribution is easy to draw from. We can then perform the following algorithm to impute b_i:

1) Generate V from $g(b_i) = f(b_i|X_i, Y_i; \rho)$ and U from $U(0, 1)$

2) Accept draw $V = b_i$ if $U \leq\ f(R_i^{-S}|Y_i^{(obs)}, X_i^{(obs)}, V; \phi^{-S})$.

Otherwise, we reject draw V and return to 1).

Under LMAR, we can obtain a draw of ρ using the same approach as under MAR. We can obtain a draw of ϕ by fitting our specified model for the missingness given $Y_i^{(obs)}, X_i^{(obs)}$, and b_i to a bootstrap sample of the data and using the most recent imputation of b_i.

Imputation of missing covariates and outcomes

Covariates

We also note that X_i as defined in the above equation is a matrix. In the notation developed in **Section 2**, covariate set X_i represents a vector. Therefore, we have some notation mismatch that we will need to rectify in order to apply (4) for imputation. Let $Z_i^{(t)}$ represent the vector of elements corresponding to covariate t for subject i and $Z_i^{(-t)}$ be a stacked vector containing the remaining elements of X_i that are not in $Z_i^{(t)}$. We note that by assumption, $b_i|X_i$ does not depend on X_i. Using this notation, we can impute missing $Z_i^{(t)}$ (and therefore the missing values for the t^{th} variable in X_i) using:

$$\begin{aligned} f(Z_i^{(t)}|Z_i^{(-t)}, Y_i, b_i; \rho) &\propto f(Y_i|X_i, b_i; \theta) f(b_i|X_i; \omega) f(Z_i^{(t)}|Z_i^{(-t)}; \psi) \\ &\propto f(Y_i|X_i, b_i; \theta) f(Z_i^{(t)}|Z_i^{(-t)}; \psi) \end{aligned}$$

In this case, $f(Z_i^{(t)}|Z_i^{(-t)}; \psi)$ is a *multi-dimensional* distribution. For example, $f(Z_i^{(t)}|Z_i^{(-t)}; \psi)$ may be multivariate normal.

We can obtain imputations of $Z_i^{(t)}$ by performing a block-wise Metropolis-Hastings draw. In settings with where $f(Z_i^{(t)}|Z_i^{(-t)}; \psi)$ is not easy to draw from, we recommend this approach. Alternatively, we could perform the following rejection sampling procedure. Define $k(Z_i^{(t)}) = f(Y_i|X_i, b_i; \theta) f(Z_i^{(t)}|Z_i^{(-t)}; \psi)$ and $g(Z_i^{(t)})) = f(Z_i^{(t)}|Z_i^{(-t)}; \psi)$. We want to find a constant that dominates $f(Y_i|X_i, b_i; \theta)$ across different values of $Z_i^{(t)}$. We note that $f(Y_i|X_i, b_i; \theta)$ is multivariate normal by assumption, and its maximum value across all covariate values will occur when $Y_i = X_i\theta + 1_K b_i$, at which point $f(Y_i|X_i, b_i; \theta) = \frac{1}{\sqrt{|2\pi\Sigma|}}$. Define $K = \frac{1}{\sqrt{|2\pi\Sigma|}}$. We can then impute $Z_i^{(t)}$ jointly using the following rejection sampling algorithm:

1) Generate V from $g(Z_i^{(t)}) = f(Z_i^{(t)}|Z_i^{(-t)}; \psi)$ and U from $U(0, 1)$

2) Accept draw $V = Z_i^{(t)}$ if

$$U \leq \frac{f(Y_i|X_i, b_i; \theta)}{K}|_{Z_i^{(t)}=V} = e^{-\frac{1}{2}(Y_i - X_i\theta - 1_K b_i)^T \Sigma^{-1}(Y_i - X_i\theta - 1_K b_i)}|_{Z_i^{(t)}=V}$$

Otherwise, return to 1).

We note that the above imputation algorithm allows the elements of $Z_i^{(t)}$ to take different values. Suppose the covariate represented by $Z_i^{(t)}$ is time-independent. Then we would want the elements of $Z_i^{(t)}$ to be equal. We can impose this property by defining $f(Z_i^{(t)}|Z_i^{(-t)}; \psi)$ such that it requires all of the elements of $Z_i^{(t)}$ to be equal. In this case, the rejection sampling algorithm would be simple to perform.

Imputation using the above approach requires draws of Σ, θ, and ψ. We can use the drawn values of Σ and θ from the step for imputing the random intercept. However, suppose we want to draw new values for the parameters conditional on the imputed values of b. Since we assumed that $\Sigma = \sigma^2 \mathbb{I}_K$ (so the elements of Y_i are independent given b_i), we can draw Σ and θ by fitting a linear regression model to Y treating the elements of Y_i as independent and using offset term b_i for all elements in Y_i (to a bootstrap sample of the data). We can draw ψ by fitting a model for $Z_i^{(t)}|Z_i^{(-t)}$ to a bootstrap sample.

Outcomes

We note that Y_i is a vector in this case. We can impute the t^{th} element of Y_i using:

$$f(Y_i^{(t)}|Y_i^{(-t)}, b_i; \rho) \propto f(Y_i|X_i, b_i; \theta)$$

Since the elements of Y_i are multivariate normal by assumption, we can easily work out this conditional distribution. This distribution simplifies further when we assume that the elements of Y_i are independent given b_i and X_i. In this case, we can impute $Y_i^{(t)}$ from a normal distribution with mean equal to the t^{th} element of $X_i\theta$ and variance σ^2. We can draw θ and σ^2 as we do for covariate imputation.

Final analysis

We can use the above imputation method to obtain M imputed datasets. We can then fit a model to each of the imputed datasets and use Rubin's combining rules to obtain a single set of parameter estimates and standard errors. As discussed in the main paper, there are several different ways we can perform the final analysis for any given imputed dataset. If we choose to use the imputed random intercept values, we can estimate θ by fitting a linear regression with offset term b_i. For this fit, we can either use or ignore the imputed D. We can estimate ω^2 as the sample variance of the imputed b_i. Alternatively, we can ignore the imputed random intercept values and fit linear mixed model using the imputed values for D. This approach may be simpler and more stable in practice, but it may not be fully efficient in the LMAR setting as shown in *Property 4*.

Brief comparison to some existing methods

Imputation-based approaches for dealing with missing linear mixed model outcome data under MAR and a joint model have been explored extensively in the literature. The proposed approach under MAR is very similar to existing Gibbs Sampler-based approaches (e.g. Schafer and Yucel, 2002). Unlike other Gibbs Sampling approaches, our method for imputing b_i involves drawing parameters from a distribution that does not condition on the imputed values for b_i and imputes missing data sequentially rather than jointly. Additionally, in our application of the proposed methods, we assume flat priors for all model parameters. This assumption substantially simplifies the step for drawing model parameters in practice. However, the greatest distinction between our method and existing methods is the SMC imputation approach to imputing the covariates, which uses the outcome information but does not require a joint model for the covariates.

Yang et al. (2008) describes a two-stage imputation approach for linear mixed models with intermittent MAR outcome missingness and LMAR dropout. Unlike Yang et al. (2008), we propose performing imputation of all outcome missingness (from different causes) in a single stage. Missing outcome values are imputed under the same model regardless of the mechanism generating the missingness, and information about different sources of missingness can be incorporated into the missingness model used to impute the latent variable. Additionally, Yang et al. (2008) takes a Gibbs Sampling approach, and the steps for drawing the parameter values can be complicated and themselves require methods for sampling from distributions known only up to proportionality. In the proposed algorithm, parameter draws under uniform priors can be obtained my fitting models using MLE methods to a bootstrap sample of the data. This substantially simplifies the parameter drawing.

S7.3 Cox proportional hazards cure model algorithm

We define indicator G_i that takes the value 1 if subject i is not cured and 0 if subject i is cured. Let T_i be the observed event or censoring time and δ_i be the event indicator. We have $Y_i = (T_i, \delta_i)$. Let X_i be a set of covariates. The CPH mixture cure model consists of 1) a logistic regression for the probability of being "not cured" [$\text{logit}(P(G_i = 1|X_i)) = \omega_0 + \omega_1 X_i$] and 2) a Cox proportional hazards model for the event hazard in the "not cured" group $\left[\lambda(t) = \lambda_0(t)e^{\theta X_i}\right]$.

We recall that non-cure status, G_i, is partially latent. For subjects with observed events ($\delta_i = 1$), we know that $G_i = 1$. We may also assume that subjects still at risk by a certain time t are cured ($G_i = 0$). For all other subjects, G_i is unknown. In addition to missingness in cure status, suppose we have ignorable or latent ignorable missingness in covariates X. We can use the imputation algorithm proposed in the main paper to iteratively impute values for the latent variable and the covariates. Below, we present some details for the approach for imputing the latent variable and covariates. We can initialize the missing values of the latent variable and the covariates from drawing from the observed values with equal probability.

Imputation of latent variable

Assuming MAR

We will first assume that missingness in X_i is MAR. In this case, we can impute G_i using the following relation derived from (2):

$$\text{logit}(P(G_i = 1|X_i, T_i, \delta_i = 0; \rho)) = \omega_0 + \omega_1 X_i - \Lambda_0(T_i)e^{\theta X_i}$$

This imputation distribution depends on the most recent imputed values for X_i, parameters ω and θ, and the cumulative baseline hazard function, $\Lambda_0(t)$. An identical imputation distribution was proposed in Beesley et al. (2016) for imputing cure status in the Cox proportional cure model setting under MAR. In Beesley et al. (2016), $\Lambda_0(t)$ is estimated using a weighted Breslow-type estimator at each iteration of the imputation algorithm, and we can use the same estimation approach here. We can draw values for ρ by fitting a Cox proportional cure model to a bootstrap sample of the most recent imputed data or by fitting a cure model to the most recent imputed data and draw ρ from a multivariate normal distribution with mean and variance from the cure model fit.

Assuming LMAR

Now, we assume missingness in X_i is LMAR. From (2), we can impute G_i using

$$\begin{aligned}\text{logit}(P(G_i = 1|X_i, T_i, \delta_i = 0, R_i; \nu)) =& \omega_0 + \omega_1 X_i - \Lambda_0(T_i)e^{\theta X_i} \\ &+ \log\left[\frac{f(R_i^{-S}|T_i, \delta_i = 0, X_i^{(obs)}, G_i = 1; \phi^{-S})}{f(R_i^{-S}|T_i, \delta_i = 0, X_i^{(obs)}, G_i = 0; \phi^{-S})}\right]\end{aligned}$$

This distribution differs from the one used under MAR by an offset term on the logit scale. When the difference in the missingness distribution by cure status is small, the offset term will be near zero. This distribution again depends on the cumulative baseline hazard function, $\Lambda_0(t)$, which can be estimated as in the MAR case. It also depends on ω, θ, and ϕ. We also must specify a model for missingness of the set of indicators that are conditionally dependent on L_i, R_i^{-S}.

We can draw θ and ω using the same approach as in the MAR case (ignoring the most recent imputations of L). We can draw ϕ by fitting a model for R_i^{-S} to a bootstrap sample of the data using the most recent imputation of cure status.

Imputation of missing covariates

By (4), we can impute missing values for covariate $X^{(t)}$ using:

$$f(X_i^{(t)}|X_i^{(-t)},Y_i,G_i;\rho) \propto [P(G_i=1|X_i;\omega)f(Y_i|X_i,G_i;\theta)]^{G_i}\,P(G_i=0|X_i;\omega)^{1-G_i}f(X_i^{(t)}|X_i^{(-t)};\psi)$$
$$\propto \left[\frac{e^{\omega_0+\omega_1 X_i}}{1+e^{\omega_0+\omega_1 X_i}}\left(\lambda_0(T_i)e^{\theta X_i}\right)^{\delta_i} e^{-\Lambda_0(T_i)e^{\theta X_i}}\right]^{G_i}\left[\frac{1}{1+e^{\omega_0+\omega_1 X_i}}\right]^{1-G_i} f(X_i^{(t)}|X_i^{(-t)};\psi) \qquad \text{(S7.2)}$$

When $X_i^{(t)}$ is categorical, we can easily use the above expression to derive the full form of the distribution used for imputation. For example, imputation of a binary covariate. Then imputation can proceed using the following relation:

$$P(X_i^{(t)}=1|X_i^{(-t)},Y_i,G_i;\rho)=\frac{(S7.2)|_{X_i^{(t)}=1}}{(S7.2)|_{X_i^{(t)}=1}+(S7.2)|_{X_i^{(t)}=0}}$$

When $X_i^{(t)}$ has continuous structure, the imputation distribution may only be known up to proportionality. We can use Metropolis-Hastings methods to draw missing $X_i^{(t)}$ from (S7.2) using a proposal distribution centered at the most recent imputation of $X_i^{(t)}$. Alternatively, we could use the following rejection sampling algorithm: Define $k(X_i^{(t)})$ =(S7.2). We note that

$$\begin{aligned} k(X_i^{(t)}) &\leq \left[\left(\lambda_0(T_i)e^{\theta X_i}\right)^{\delta_i} e^{-\Lambda_0(T_i)e^{\theta X_i}}\right]^{G_i} f(X_i^{(t)}|X_i^{(-t)};\psi) \\ &\leq [f(T_i|X_i,G_i=1)]^{\delta_i}\, f(X_i^{(t)}|X_i^{(-t)};\psi) \end{aligned}$$

Suppose we define

$$K=(1-\delta_i)+\delta_i \max_{X_i^{(t)}} f(T_i|X_i,G_i=1)$$

so K takes the value 1 if $\delta_i=0$ and takes the maximum of the event time distribution function across $X_i^{(t)}$ if $\delta_i=1$. This maximum can usually be easily calculated given parameter values when the baseline hazard is parametric. We further define $g(X_i^{(t)})=f(X_i^{(t)}|X_i^{(-t)};\psi)$. Then we have that $k(X_i^{(t)})\leq Kg(X_i^{(t)})$. Then we can obtain a draw of $X_i^{(t)}$ from $k(X_i^{(t)})$ through the following algorithm:

1) Generate V from $g(X_i^{(t)})=f(X_i^{(t)}|X_i^{(-t)};\psi)$ and U from $U(0,1)$

2) Accept draw $V=X_i^{(t)}$ if $U\leq \dfrac{\left[\frac{e^{\omega_0+\omega_1 X_i}}{1+e^{\omega_0+\omega_1 X_i}}\left(\lambda_0(T_i)e^{\theta X_i}\right)^{\delta_i} e^{-\Lambda_0(T_i)e^{\theta X_i}}\right]^{G_i}\left[\frac{1}{1+e^{\omega_0+\omega_1 X_i}}\right]^{1-G_i}}{K}$. Otherwise, return to 1).

Imputation by (S7.2) requires draws of ω, θ, and ψ. We can either use the draws of ω and θ obtained in the imputation step for the latent variable or draw new values. If we draw new values, we should use methods that use the most recent imputation of L. We can then draw θ by fitting a Cox regression to a bootstrap sample of the subjects with imputed $G=1$. We can draw ω by fitting a logistic regression to G for a bootstrap sample of the entire dataset. We can draw ψ by fitting a model for $X_i^{(t)}|X_i^{(-t)}$ to a bootstrap sample.

Final analysis

We can use the above imputation method to obtain M imputed datasets. We can then fit a model to each of the imputed datasets and use Rubin's combining rules to obtain a single set of parameter estimates and standard errors. There are several different ways we can perform the final analysis for any given imputed dataset. If we choose to use the imputed G, we can estimate θ by fitting a Cox regression to the subjects with imputed

$G = 1$, and we can estimate ω by fitting a logistic regression for G. For these fits, we can either use or ignore the imputed D. Alternatively, we can ignore the imputed G and fit cure model using the imputed values for D. We recommend this last approach.

Brief comparison to some existing methods

Beesley et al. (2016) explores SMC covariate imputation for the Cox proportional hazards cure model under MAR assumptions, and our proposed algorithm under MAR is very similar with some small differences in the methods for drawing parameters. We believe we are the first to explore covariate imputation for the Cox proportional hazards cure model under LMAR assumptions.

S7.4 Mixture of GLMs

Suppose our outcome Y is generated from a mixture of K generalized linear models (GLMs) where K is known. 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. We suppose the distribution of $Y_i|X_i, C_i = j$ is modeled using a GLM (e.g. normal, logistic, Poisson) for $j = 1, \ldots, K$ and that the distribution for $C_i|X_i$ is independent of X_i.

We suppose that we have ignorable or latent ignorable missingness in Y and/or X. We can use the proposed methods for imputation. We can initialize the missing values of the covariates from drawing from the observed values with equal probability. We can initialize C based on the estimated probabilities $P(C_i = 1)$ obtained by fitting a latent class model to the complete case data.

Imputation of latent variable

Assuming MAR

The imputation distribution for the latent mixing variable C_i under MAR can be easily worked out based on the kernel in (2) to be multinomial with corresponding probabilities as follows:

$$P(C_i = j|X_i, Y_i, R_i; \nu) = \frac{f(Y_i|X_i, C_i = j; \theta)P(C_i = j; \omega)}{\sum_{l=1}^{K} f(Y_i|X_i, C_i = l; \theta)P(C_i = l; \omega)}$$

We can obtain a draw of θ and ω by fitting a latent class model to a bootstrap sample the most recently imputed data. In R, we can perform this latent class model fit using the package *flexmix* (Leisch, 2004). This package will estimate θ and ω for a specified number of latent classes, but it cannot differentiate between the different class labels. Therefore, we will need to impose a restriction to relate the latent classes identified by *flexmix* to values of C.

Assuming LMAR

Under LMAR, we can impute missing values of C_i using:

$$P(C_i = j|X_i, Y_i, R_i; \nu) = \frac{f(R_i^{-S}|X_i^{(obs)}, C_i = j, Y_i^{(obs)}; \phi^{-S})f(Y_i|X_i, C_i = j; \theta)P(C_i = j|X_i; \omega)}{\sum_{l=1}^{K} f(R_i^{-S}|X_i^{(obs)}, C_i = l, Y_i^{(obs)}; \phi^{-S})f(Y_i|X_i, C_i = l; \theta)P(C_i = l|X_i; \omega)}$$

This imputation distribution requires us to model R_i^{-S}. Draws of θ and ω can be obtained as in the MAR case. We can obtain a draw of ϕ by fitting a model for R_i^{-S} to a bootstrap sample of the data using the most recent imputation of C.

Imputation of missing covariates and outcome

Covariates

By (4) and since $f(C_i|X_i; \omega) = f(C_i; \omega)$ by assumption, we can impute missing values for covariate $X^{(t)}$ using:

$$f(X_i^{(t)}|X_i^{(-t)}, Y_i, C_i; \rho) \propto f(Y_i|X_i, C_i; \theta)f(X_i^{(t)}|X_i^{(-t)}; \psi)$$

When $X_i^{(t)}$ is categorical, we can easily use the above expression to derive the full form of the distribution used for imputation. Otherwise, we can use methods to draw from the above distribution known only up to proportionality. For example, we can use the following rejection sampling algorithm: Define $k(X_i^{(t)}) = f(Y_i|X_i, C_i;\theta)f(X_i^{(t)}|X_i^{(-t)};\psi)$ and $g(X_i^{(t)}) = f(X_i^{(t)}|X_i^{(-t)};\psi)$. Define

$$K = \max_{X_i^{(t)}} f(Y_i|X_i, C_i;\theta)$$

Then we have that $k(X_i^{(t)}) \leq Kg(X_i^{(t)})$. Then we can obtain a draw of $X_i^{(t)}$ from $k(X_i^{(t)})$ through the following algorithm:

1) Generate V from $g(X_i^{(t)}) = f(X_i^{(t)}|X_i^{(-t)};\psi)$ and U from $U(0, 1)$

2) Accept draw $V = X_i^{(t)}$ if $U \leq \frac{f(Y_i|X_i,C_i;\theta)}{K}$.

Otherwise, return to 1).

Imputation using the above method requires draws of ω, θ, and ψ. We can draw θ by fitting a GLM (or multiple GLMS) to a bootstrap sample of subjects using the most recent imputation of C. We can draw ω by looking at the proportion of subjects with $C = j$ for each j in a bootstrap sample of the data. We can draw ψ by fitting a model for $X_i^{(t)}|X_i^{(-t)}$ to a bootstrap sample.

Outcome

We will assume here that Y is univariate. By (3), we can impute missing values for outcome Y using:

$$f(Y_i|X_i, C_i;\rho) = f(Y_i|X_i, C_i;\theta)$$

We can obtain a draw for θ as in covariate imputation and then draw missing values of Y simply using the GLM corresponding to the most recent imputed value for C_i.

Final analysis

We can use the above imputation method to obtain M imputed datasets. We can then fit a model to each of the imputed datasets and use Rubin's combining rules to obtain a single set of parameter estimates and standard errors. There are several different ways we can perform the final analysis for any given imputed dataset. If we choose to use the imputed C, we can estimate θ by fitting a GLM for $f(Y|C, X)$ using the imputed C and either using or ignoring the imputed D. Alternatively, we can ignore the imputed C and fit a latent class model (e.g. using *flexmix*) using the imputed values for D. This second approach would require us to use an identifying assumption to determine which cluster identified by the latent class modeling corresponds to which value of C. We recommend this second approach.

Brief comparison to some existing methods

Many authors have explored similar imputation approaches for mixtures of GLMs under MAR assumptions, but comparatively little work has been done exploring LMAR missingness in this setting (e.g. Vidotto et al., 2015). Jung (2007) considers the case of a multivariate outcome related to a categorical latent mixing variable and proposes a MCMC imputation scheme that iteratively imputes missing values of the outcome and C. Additionally, Jung (2007) assumes the outcome is independent of the covariates given C. Our proposed approach can be viewed as a generalization of the approach in Jung (2007) that can handle LMAR missingness in the outcome and covariates while also allowing for conditional dependence between Y and X. In addition, the approach in Jung (2007) relies on a valid joint distribution for the covariates, while our approach side-steps that issue.

References

Bartlett, J. W., S. R. Seaman, I. R. White, and J. R. Carpenter (2014). Multiple imputation of covariates by fully conditional specification: accomodating the substantive model. *Statistical Methods in Medical Research 24*(4), 462–487.

Beesley, L. J., J. W. Bartlett, G. T. Wolf, and J. M. G. Taylor (2016). Multiple imputation of missing covariates for the Cox proportional hazards cure model. *Statistics in Medicine 35*(26), 4701–4717.

Bodner, T. E. (2008). What Improves with Increased Missing Data Imputations? *Structural equation modeling : a multidisciplinary journal 15*(4), 651–675.

Brooks, S. P. B. and A. Gelman (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics 7*(4), 434–455.

Cai, C., Y. Zou, Y. Peng, and J. Zhang (2012). smcure: An R-package for Estimating Semiparametric Mixture Cure Models. *Computer Methods and Programs in Biomedicine 108*(3), 1255–1260.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–511.

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

Heckman, J. J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement 5*(4), 475–492.

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

Leisch, F. (2004). FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software 11*(8), 1–18.

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

Little, R. J. A. and D. B. Rubin (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley and Sons, Inc.

Robert, C. P. and G. Casella (2004). *Monte Carlo Statistical Methods* (2nd ed.). Springer.

Schafer, J. L. (1999). Multiple imputation : a primer. *Statistical Methods in Medical Research 8*(1), 3–15.

Schafer, J. L. and R. M. Yucel (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics 11*(2), 437–457.

Van Buuren, S., J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation 76*(12), 1049–1064.

Vidotto, D., J. K. Vermunt, and M. C. Kaptein (2015). Multiple Imputation of Missing Categorical Data using Latent Class Models : State of the Art. *Psychological Test and Assessment Modeling 57*(4), 542–576.

White, I. R. and P. Royston (2011). Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. *Statistics in Medicine 30*(4), 377–399.

Wu, M. C. and R. J. Carroll (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics 44*(1), 175–188.

Yang, X., J. Lu, and S. Shoptaw (2008). Imputation-based strategies for clinical trial longitudinal data with nonignorable missing values. *Statistics in Medicine 27*, 2826–2849.