Biometrics 71, 1150–1159
December 2015

DOI: 10.1111/biom.12358

# Handling Missing Data in Matched Case-Control Studies Using Multiple Imputation

**Shaun R. Seaman[1,*] and Ruth H. Keogh[2]**

[1]MRC Biostatistics Unit, Cambridge, CB2 0SR, U.K.
[2]London School of Hygiene and Tropical Medicine, London, WC1E 7HT, U.K.
[*]*email:* shaun.seaman@mrc-bsu.cam.ac.uk

Summary. Analysis of matched case-control studies is often complicated by missing data on covariates. Analysis can be restricted to individuals with complete data, but this is inefficient and may be biased. Multiple imputation (MI) is an efficient and flexible alternative. We describe two MI approaches. The first uses a model for the data on an individual and includes matching variables; the second uses a model for the data on a whole matched set and avoids the need to model the matching variables. Within each approach, we consider three methods: full-conditional specification (FCS), joint model MI using a normal model, and joint model MI using a latent normal model. We show that FCS MI is asymptotically equivalent to joint model MI using a restricted general location model that is compatible with the conditional logistic regression analysis model. The normal and latent normal imputation models are not compatible with this analysis model. All methods allow for multiple partially-observed covariates, non-monotone missingness, and multiple controls per case. They can be easily applied in standard statistical software and valid variance estimates obtained using Rubin's Rules. We compare the methods in a simulation study. The approach of including the matching variables is most efficient. Within each approach, the FCS MI method generally yields the least-biased odds ratio estimates, but normal or latent normal joint model MI is sometimes more efficient. All methods have good confidence interval coverage. Data on colorectal cancer and fibre intake from the EPIC-Norfolk study are used to illustrate the methods, in particular showing how efficiency is gained relative to just using individuals with complete data.

Key words: Chained equations; Compatibility; MICE; Multilevel MI; Restricted general location model.

## 1. Introduction

Case-control studies are used to investigate associations between disease and putative risk factors. Confounding of observed associations can be handled at the design stage by matching cases and controls on confounders, at the analysis stage by adjusting for confounders using a regression model, or by a combination of these. In matched case-control studies, each case is individually matched with one or more controls on a subset of confounders and the (usual) analysis uses conditional logistic regression (CLR) to control for the remaining confounders.

Often, the analysis is complicated by missing data on covariates (i.e., exposures and remaining confounders). A common solution is to restrict analysis to individuals with complete data. Although appealing for its simplicity, this "complete-case analysis" ("case" here means any individual, rather than an individual with disease) is inefficient and may be biased. In particular, where exclusion of a case or control leaves a matched set in which remaining members are either all cases or all controls, the whole set ceases to contribute information to the CLR estimating equations.

To improve efficiency and reduce bias, several alternatives have been proposed. Lipsitz et al. (1998) allow for one partially observed covariate. They assume data are missing at random (MAR) and fit a missingness model, i.e., a model for the probability that an individual is a complete case. Functions of the fitted probabilities are then used as offsets in CLR. This consistently estimates odds ratios (ORs) when the missingness model is correctly specified, but is inefficient as it only uses data on complete cases. Paik and Sacco (2000) also allow for just one partially observed covariate and assume MAR. They assume a model for the distribution of the partially observed covariate given the other covariates, matching variables and binary disease status. When this covariate model is correctly specified, consistent estimation of the ORs can be achieved by CLR after imputing the missing covariate as its fitted value when the disease status variable is set to 0.5. Rathouz (2003) notes that this method implicitly assumes missingness does not depend on disease status, and generalizes it to allow for such dependence, as well as for multiple missing covariates. His method assumes the partially observed covariates are all observed or all missing on each individual. Sinha and Wang (2009) take a similar approach, but instead of a parametric covariate model, kernel density estimation is used for those functions in the estimating equations that depend on the distribution of the partially observed covariate. They find their OR estimator is less biased than that of Paik and Sacco (2000) when the latter's covariate model is misspecified. A drawback is that categorical variables are handled by stratifying individuals on these variables and performing kernel density estimation separately in each stratum, which limits the feasible number of categorical variables (and categories). Paik (2004) extends Paik and Sacco's (2000) method to allow for data missing not at random (MNAR).

The forementioned methods all reduce to standard CLR when there are no missing data: the assumed missingness or covariate model then becomes irrelevant. Other methods for missing data derive information from an assumed covariate model even when data are complete. These methods may be more efficient but at the cost of possible bias when the covariate model is misspecified. Satten and Carroll (2000) propose such a method. This allows for multiple partially observed covariates, but assumes these are all observed or all missing on each individual. Ahn et al. (2011) generalize it to allow for MNAR and multiple disease states. Rathouz et al. (2002) elaborate Lipsitz et al.'s (1998) method to use a covariate model and so gain efficiency. The resulting estimator is doubly robust but difficult to implement. They also propose a more practical approximation which, though not doubly robust, still gains efficiency. Liu et al. (2013) use empirical likelihood to develop a semiparametric-efficient competitor to Rathouz et al.'s (2002) estimator. Gebregziabher and DeSantis (2010) assume all covariates are categorical and carry out multiple imputation (MI) using a latent-class model. A drawback is that imputation of a individual's missing value makes no use of data on matching variables, covariate values of other individuals in the same matched set, or disease status, which may cause bias (Moons et al., 2006) and inefficiency.

The methods described so far assume the distribution of the partially observed covariate(s) given fully observed covariates, disease status, and matching variables can be modeled parametrically. Sometimes this is not feasible. For example, if cases are matched with controls from the same family, from the same postcode area, or from the set of patients attending the same general practice, it could be difficult to model parametrically the matching via explicit matching variables, while the alternative of allowing a separate nuisance parameter for each matched set may cause problems with model fitting and induce bias and even inconsistency of estimators. Even when matching could, in principle, be modeled parametrically, this is only possible if the analyst has data on matching variables, which is not always so, and some analysts may prefer to avoid modeling effects of matching variables, since CLR makes no assumptions about the association between disease and matching variables. One solution, adopted by Sinha et al. (2005), is to allow each matched set to have its own parameter in the covariate model but treat these as random effects. They assume a single partially observed covariate and that the random effects are generated by a Dirichlet process. They fit their Bayesian model using a Hastings–Metropolis algorithm with specially written computer code.

Though useful, these methods have limitations. Many assume only one partially observed covariate or that partially observed covariates are collectively observed or missing on each individual. Many require bespoke computer code. Most require parametric modeling of matching variables. In this article, we advocate the use of MI, proposing, and comparing six MI methods suitable for matched case-control data that can be easily implemented in commonly used statistical packages. MI has several advantages. First, it is increasingly being used to handle missing data and many researchers are familiar with the technique. Second, MI software is readily available and easy to use. Third, MI allows for multiple partially observed covariates without needing them to be collectively observed or missing. Fourth, MI can easily incorporate information on variables that are not included in the CLR model but are predictive of missing covariates in that model. This can increase efficiency and can also reduce bias when these extra variables are required to make the MAR assumption more plausible. Fifth, we propose both methods that parametrically model matching variables and methods in which this is not required. Arguably, a sixth advantage is that, unlike some of the methods proposed earlier, MI reduces to standard CLR when there are no missing data. Although this means MI does not offer the potential efficiency gain associated with methods that make use of a covariate model even when data are complete, it should make it more robust to misspecification of that model.

We illustrate the use of MI for matched case-control data on a study of association between fibre intake and colerectal cancer nested within the European Prospective Investigation of Cancer (EPIC) Norfolk cohort. This is one of the studies in the UK Dietary Cohort Consortium, which combines case-control studies nested within several cohorts. Results from this study have been described elsewhere (Dahm et al., 2010). Cases were individuals in the EPIC Norfolk cohort diagnosed with colorectal cancer between recruitment to the cohort (1993–1998) and the end of 2006. Seven-day diet diaries were completed by participants shortly after recruitment to the underlying cohort and stored for later use. The diet diaries were processed for individuals selected for the case-control sample to obtain measures of average daily intake of foods and nutrients (Dahm et al., 2010). There were 318 colorectal cancer cases and each was matched with four controls on sex, age within 3 years, and date of diary completion within 3 months. Controls had to be alive and have not been diagnosed with colorectal cancer at the end of 2006. In the original analysis, the association between fibre intake and colorectal cancer was adjusted for several potential confounders using CLR : smoking status (three categories), education (four categories), social class (six categories), and physical activity level (four categories), and height, weight, exact age, alcohol intake, folate intake, intake of energy from fat, and intake of energy from nonfat (all continuous). We wished also to adjust for aspirin use (two categories). Many other studies have adjusted for aspirin use (Aune et al., 2011), which is known to be associated with reduced risk of colorectal cancer (Asano and McLeod, 2004; National Cancer Institute, 2014). It was omitted from the original analysis (Dahm et al., 2010) because it was not measured in some of the contributing studies. Of the 1590 individuals in the study, 328 (78 cases and 250 controls) were missing one or more adjustment variables, most commonly aspirin use or social class; the main exposure, fibre intake, and the matching variables were fully observed. A complete-case analysis uses only 240 (75%) matched sets and 1012 (64%) individuals.

The article is structured as follows. Section 2 discusses CLR with complete data. Section 3 describes MI in general. For matched case-control studies, Section 4 proposes three MI methods that parametrically model the matching variables, and Section 5 three analogous methods that avoid this. Section 6 contains a simulation study comparing the methods. Section 7 describes their application to the EPIC study. We end with a discussion in Section 8.

## 2. Analysis of Matched Case-Control Studies with Complete Data

For each individual in the population, let $D = 1$ if he/she has disease and $D = 0$ otherwise. So, $D = 1$ for cases and $D = 0$ for controls. Let $\boldsymbol{S}$ denote the variables used to match controls with cases. Let $\boldsymbol{X}^{\text{cat}}$ and $\boldsymbol{X}^{\text{con}}$ denote categorical and continuous covariates, respectively. A categorical variable with $m > 2$ levels is coded as $m - 1$ dummy variables. Assume

$$P(D = 1 \mid \boldsymbol{X}^{\text{cat}}, \boldsymbol{X}^{\text{con}}, \boldsymbol{S}) = \frac{\exp\{\boldsymbol{\beta}_{\text{cat}}^{\top}\boldsymbol{X}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\top}\boldsymbol{X}^{\text{con}} + q(\boldsymbol{S})\}}{1 + \exp\{\boldsymbol{\beta}_{\text{cat}}^{\top}\boldsymbol{X}^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\top}\boldsymbol{X}^{\text{con}} + q(\boldsymbol{S})\}}, \quad (1)$$

where $q(\boldsymbol{S}) = \text{logit } P(D = 1 \mid \boldsymbol{X}^{\text{cat}} = \boldsymbol{0}, \boldsymbol{X}^{\text{con}} = \boldsymbol{0}, \boldsymbol{S})$. Let $M$ denote the number of controls matched with each case. We use subscript $j$ ($j = 1, \ldots, M + 1$) to index individual within set and assume cases and controls have been ordered so that $D_1 = 1$ and $D_2 = \cdots = D_{M+1} = 0$.

In ordinary logistic regression the log ORs $\boldsymbol{\beta}_{\text{cat}}$ and $\boldsymbol{\beta}_{\text{con}}$ are estimated by maximizing the likelihood based on expression (1) and the data on the sampled individuals. This requires that either $q(\boldsymbol{S})$ is modeled or a separate baseline parameter is included for each matched set. The former corresponds to breaking the matching and adjusting for $\boldsymbol{S}$, which there is often a reluctance to do, because it requires a functional form to be specified for the effect of matching variables on disease risk. The alternative of including a baseline parameter for each set yields inconsistent maximum likelihood estimates (Breslow and Day, 1980). For this reason, CLR is often used instead. CLR includes a baseline parameter for each set, but then eliminates these from the likelihood by conditioning on the number of cases and controls in each set. Let $G(\boldsymbol{x}_1^{\text{cat}}, \boldsymbol{x}_1^{\text{con}}, \ldots, \boldsymbol{x}_{M+1}^{\text{cat}}, \boldsymbol{x}_{M+1}^{\text{con}})$ denote the conditional probability that $(\boldsymbol{X}_1^{\text{cat}}, \boldsymbol{X}_1^{\text{con}}) = (\boldsymbol{x}_1^{\text{cat}}, \boldsymbol{x}_1^{\text{con}})$ given that $(\boldsymbol{X}_1^{\text{cat}}, \boldsymbol{X}_1^{\text{con}}) = (\boldsymbol{x}_1^{\text{cat}*}, \boldsymbol{x}_1^{\text{con}*}), \ldots, (\boldsymbol{X}_{M+1}^{\text{cat}}, \boldsymbol{X}_{M+1}^{\text{con}}) = (\boldsymbol{x}_{M+1}^{\text{cat}*}, \boldsymbol{x}_{M+1}^{\text{con}*})$ for some permutation $(\boldsymbol{x}_1^{\text{cat}*}, \boldsymbol{x}_1^{\text{con}*}), \ldots, (\boldsymbol{x}_{M+1}^{\text{cat}*}, \boldsymbol{x}_{M+1}^{\text{con}*})$ of $(\boldsymbol{x}_1^{\text{cat}}, \boldsymbol{x}_1^{\text{con}}), \ldots, (\boldsymbol{x}_{M+1}^{\text{cat}}, \boldsymbol{x}_{M+1}^{\text{con}})$ and given that $D_1 = 1$ and $D_2 = \ldots = D_{M+1} = 0$ and $\boldsymbol{S}_1 = \ldots = \boldsymbol{S}_{M+1}$. Equation (1) implies

$$G(\boldsymbol{x}_1^{\text{cat}}, \boldsymbol{x}_1^{\text{con}}, \ldots, \boldsymbol{x}_{M+1}^{\text{cat}}, \boldsymbol{x}_{M+1}^{\text{cat}}) = \frac{\exp(\boldsymbol{\beta}_{\text{cat}}^{\top}\boldsymbol{x}_1^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\top}\boldsymbol{x}_1^{\text{con}})}{\sum_{j=1}^{M+1}\exp(\boldsymbol{\beta}_{\text{cat}}^{\top}\boldsymbol{x}_j^{\text{cat}} + \boldsymbol{\beta}_{\text{con}}^{\top}\boldsymbol{x}_j^{\text{con}})}, \quad (2)$$

and vice versa (Web Appendix A). CLR finds the values of $\boldsymbol{\beta}_{\text{cat}}$ and $\boldsymbol{\beta}_{\text{con}}$ that maximize the product of expression (2) over the matched sets; these consistently estimate the log ORs .

## 3. Joint Model MI and Full-Conditional Specification (FCS) MI

We briefly review the most commonly used forms of MI: joint model MI and FCS MI (Web Appendix B has more detail). In joint model MI, a Bayesian model with non-informative priors is specified for the distribution of the partially observed variables given fully observed variables. This "imputation model" is fitted to the observed data, and values for missing variables are then sampled from their joint posterior predictive distribution. The model of interest ("analysis model") is fitted to each resulting complete (or "imputed") dataset separately, and the parameter and variance estimates obtained are combined using simple equations called Rubin's Rules. When the imputation model is correctly specified and is *compatible with the analysis model*, i.e., there exists a model for the joint distribution of all the variables that implies the analysis and imputation models as submodels, and data are MAR, joint model MI gives consistent parameter and variance estimates for the analysis model. Thus, compatibility, if possible, is desirable. The first of the methods described in each of Sections 4 and 5 are based on imputation models that are compatible with the CLR analysis model.

Instead of requiring a joint model for the partially observed variables, FCS MI involves specifying a model for the conditional distribution of each partially observed variable given all other variables. The FCS algorithm cycles through these models, sampling missing values for the dependent variable in the current model given the observed and most recently sampled values of all the other variables, until convergence is achieved. This may be easier than specifying and fitting a joint model. In special cases, FCS corresponds to joint model MI (Hughes et al., 2014). Otherwise, FCS is less theoretically justified, but there is much evidence that it works well in terms of approximate unbiasedness of parameter and variance estimates and coverage of confidence intervals (van Buuren, 2012; Hughes et al., 2014; Lee and Carlin, 2010). An important theoretical result was given by Liu et al. (2014). They defined the set of conditional models to be *compatible with a joint model* if, for each conditional model and every possible set of parameter values for that model, there exists a set of parameter values for the joint model such that the conditional and joint models imply the same distribution for the dependent variable of that conditional model. They showed that when this compatibility holds, the distribution of the data imputed by FCS MI converges, as sample size tends to infinity, to the posterior predictive distribution of the missing data under that joint model. Hence, FCS MI is asymptotically equivalent to joint model MI in this case. The first of the MI methods in each of Sections 4 and 5 use this asymptotic result.

## 4. MI Using Matching Variables

Let $\boldsymbol{R}$ denote the missingness pattern in $(\boldsymbol{X}^{\text{cat}\top}, \boldsymbol{X}^{\text{con}\top})^{\top}$. Assume $\boldsymbol{D}$ and $\boldsymbol{S}$ are fully observed and the data are MAR. In this section, we propose multiply imputing missing $(\boldsymbol{X}^{\text{cat}\top}, \boldsymbol{X}^{\text{con}\top})^{\top}$ from its conditional distribution given $\boldsymbol{S}$ and $\boldsymbol{D}$. We call this "MI using matching variables." It is analogous to breaking the matching and adjusting for the matching variables. However, matching is broken only to impute missing data; matching is then restored and the imputed data analyzed using CLR. Most methods reviewed in Section 1 effectively break the matching for the individuals with missing data. In Section 5, we describe an alternative ("MI using matched set"), which imputes without breaking the matching. We now propose three ways of modeling the distribution of $(\boldsymbol{X}^{\text{cat}\top}, \boldsymbol{X}^{\text{con}\top})^{\top}$ given $\boldsymbol{S}$ and $\boldsymbol{D}$.

The first model for $(\boldsymbol{X}^{\text{cat}\top}, \boldsymbol{X}^{\text{con}\top})^{\top}$ given $\boldsymbol{S}$ and $\boldsymbol{D}$ is a restricted general location model (Schafer, 1997). This has a log-linear model for $\boldsymbol{X}^{\text{cat}}$ and normal model for $\boldsymbol{X}^{\text{con}}$ given

$\boldsymbol{X}^{\text{cat}}$:

$$P(\boldsymbol{X}^{\text{cat}} = \boldsymbol{x}^{\text{cat}} \mid \boldsymbol{S}, D) = \frac{\exp\{a(\boldsymbol{x}^{\text{cat}}, \boldsymbol{S};\ \boldsymbol{\zeta}) + D\boldsymbol{\lambda}^{\top}\boldsymbol{x}^{\text{cat}}\}}{\sum_{\boldsymbol{x}^{\text{cat}'}} \exp\{a(\boldsymbol{x}^{\text{cat}'}, \boldsymbol{S};\ \boldsymbol{\zeta}) + D\boldsymbol{\lambda}^{\top}\boldsymbol{x}^{\text{cat}'}\}} \tag{3}$$

$$\boldsymbol{X}^{\text{con}} \mid \boldsymbol{X}^{\text{cat}}, \boldsymbol{S}, D \sim N(\boldsymbol{\alpha} + \boldsymbol{\phi}D + \boldsymbol{\gamma}\boldsymbol{X}^{\text{cat}} + \boldsymbol{\delta}\boldsymbol{S},\ \boldsymbol{\Sigma}) \tag{4}$$

where $a(\boldsymbol{x}^{\text{cat}}, \boldsymbol{S};\ \boldsymbol{\zeta})$ includes a main effect for $\boldsymbol{X}^{\text{cat}}$ and all pairwise interactions between $\boldsymbol{X}^{\text{cat}}$ and $\boldsymbol{S}$ and between pairs of elements of $\boldsymbol{X}^{\text{cat}}$. Vectors $\boldsymbol{\lambda}$, $\boldsymbol{\zeta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$ and matrices $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$ are unknown parameters. In Web Appendix C, we prove that (3)–(4) imply that equation (2) holds with $\boldsymbol{\beta}_{\text{cat}} = \boldsymbol{\lambda} - \boldsymbol{\gamma}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}$ and $\boldsymbol{\beta}_{\text{con}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}$. Hence, this model is compatible with the CLR analysis model.

Bayesian modeling software, such as WinBUGS (Lunn et al., 2000), can be used to impute missing $(\boldsymbol{X}^{\text{cat}\top}, \boldsymbol{X}^{\text{con}\top})^{\top}$ from its posterior predictive distribution implied by joint model (3)–(4). However, such software requires specialist programming skills. Instead, we propose using FCS MI with a set of conditional models that is compatible with this joint model, and hence is asymptotically equivalent to joint model MI. FCS MI is widely available in general-purpose statistical packages, e.g., Stata, R, and SAS. In Web Appendix D, we show that a compatible conditional model for a partially observed continuous covariate (an element of $\boldsymbol{X}^{\text{con}}$) is a linear regression of this covariate on $\boldsymbol{S}$, $D$, $\boldsymbol{X}^{\text{cat}}$ and the remaining elements of $\boldsymbol{X}^{\text{con}}$. Likewise, a compatible conditional model for one of the partially observed categorical covariates making up $\boldsymbol{X}^{\text{cat}}$ is a multinomial logistic regression of this categorical covariate on $\boldsymbol{S}$, $D$, $\boldsymbol{X}^{\text{con}}$ and those elements of $\boldsymbol{X}^{\text{cat}}$ that are not dummy indicators for this categorical covariate. Conveniently, these conditional models are the default options in many MI packages.

Although asymptotically equivalent, in finite samples this FCS MI method may be inefficient compared to joint model MI, because it estimates the parameters of the conditional model for $\boldsymbol{X}^{\text{cat}}$ using only part of the available data on $\boldsymbol{X}^{\text{con}}$ (Hughes et al., 2014). Our second proposed model for $(\boldsymbol{X}^{\text{cat}\top}, \boldsymbol{X}^{\text{con}\top})^{\top}$ given $\boldsymbol{S}$ and $D$ is a latent normal model (Carpenter and Kenward, 2013). This is not compatible with the CLR analysis model, but it has the advantage that it can be used for joint model MI without needing specialist Bayesian software. For simplicity, suppose that all the categorical covariates are binary (see Carpenter and Kenward (2013) for general case). The latent normal model is

$$(\boldsymbol{X}^{\text{con}\top}, \boldsymbol{W}^{\text{cat}\top})^{\top} \mid \boldsymbol{S}, D \sim N(\boldsymbol{\alpha}_{\text{LN}} + \boldsymbol{\phi}_{\text{LN}}D + \boldsymbol{\delta}_{\text{LN}}\boldsymbol{S},\ \boldsymbol{\Sigma}_{\text{LN}}) \tag{5}$$

where $\boldsymbol{W}^{\text{cat}}$ is a vector of latent variables (each with unit variance), one for each element of $\boldsymbol{X}^{\text{cat}}$, and such that an element of $\boldsymbol{X}^{\text{cat}}$ equals 1 if its corresponding element of $\boldsymbol{W}^{\text{cat}}$ is positive and 0 otherwise. $\boldsymbol{\alpha}_{\text{LN}}$, $\boldsymbol{\phi}_{\text{LN}}$, $\boldsymbol{\delta}_{\text{LN}}$, and $\boldsymbol{\Sigma}_{\text{LN}}$ are unknown parameters. Joint model MI using (5) can be done using the software REALCOM-MI or the jomo package in R. The realcomImpute program provides an interface between Stata and REALCOM-MI.

When all partially observed covariates are continuous, our FCS MI method and joint model MI using (5) both reduce to joint model MI using the normal model (4). Use of this normal model for MI even when some partially observed variables are categorical was originally promoted by Schafer (1997) and has become common. Although the model is obviously misspecified, this method has been found to work well in many situations and software is widely available, e.g., mi mvn impute in Stata and norm in R. Thus, our third proposed model for $(\boldsymbol{X}^{\text{cat}\top}, \boldsymbol{X}^{\text{con}\top})^{\top}$ given $\boldsymbol{S}$ and $D$ is expression (5) with $\boldsymbol{W}^{\text{cat}}$ replaced by $\boldsymbol{X}^{\text{cat}}$. Following Bernaards et al. (2007), we use "adaptive rounding" after imputation to handle non-integer imputed values of $\boldsymbol{X}^{\text{cat}}$.

## 5. MI Using Matched Set

Now, we propose three models for $\boldsymbol{X}^{\text{set}} = (\boldsymbol{X}_1^{\text{cat}\top}, \boldsymbol{X}_1^{\text{con}\top}, \ldots, \boldsymbol{X}_{M+1}^{\text{cat}\top}, \boldsymbol{X}_{M+1}^{\text{con}\top})^{\top}$ unconditional on $\boldsymbol{S}$, thus allowing imputation without using the matching variables. These are analogous to the models in Section 4 but involve a matched-set-specific random effect, $\boldsymbol{u}$.

The first is a restricted general location model. Assume that for each matched set,

$$\begin{aligned} &P(\boldsymbol{X}_1^{\text{cat}} = \boldsymbol{x}_1^{\text{cat}}, \ldots, \boldsymbol{X}_{M+1}^{\text{cat}} = \boldsymbol{x}_{M+1}^{\text{cat}} \mid D_1 = 1, D_2 = \cdots = D_{M+1} = 0) \\ &= \frac{\exp\{b(\boldsymbol{x}_1^{\text{cat}}, \ldots, \boldsymbol{x}_{M+1}^{\text{cat}};\ \boldsymbol{\nu}) + \boldsymbol{\tau}^{\top}\boldsymbol{x}_1^{\text{cat}}\}}{\sum_{\boldsymbol{x}_1^{\text{cat}'}, \ldots, \boldsymbol{x}_{M+1}^{\text{cat}'}} \exp\{b(\boldsymbol{x}_1^{\text{cat}'}, \ldots, \boldsymbol{x}_{M+1}^{\text{cat}'};\ \boldsymbol{\nu}) + \boldsymbol{\tau}^{\top}\boldsymbol{x}_1^{\text{cat}'}\}} \end{aligned} \tag{6}$$

$$\begin{aligned} \text{with } b(\boldsymbol{x}_1^{\text{cat}}, \ldots, \boldsymbol{x}_{M+1}^{\text{cat}}) &= \sum_{j=1}^{M+1} b_1(\boldsymbol{x}_j^{\text{cat}};\ \boldsymbol{\nu}) \\ &\quad + \sum_{j=1}^{M} \sum_{k=j+1}^{M+1} b_2(\boldsymbol{x}_j^{\text{cat}}, \boldsymbol{x}_k^{\text{cat}}; \boldsymbol{\nu}) \end{aligned} \tag{7}$$

where $b_1(\boldsymbol{x}_j^{\text{cat}};\ \boldsymbol{\nu})$ includes a main effect of each element of $\boldsymbol{X}_j^{\text{cat}}$ and an interaction between each pair of these elements, and $b_2(\boldsymbol{x}_j^{\text{cat}}, \boldsymbol{x}_k^{\text{cat}}; \boldsymbol{\nu})$ includes all pairwise interactions between one element of $\boldsymbol{X}_j^{\text{cat}}$ and one element of $\boldsymbol{X}_k^{\text{cat}}$. This allows correlation between $\boldsymbol{X}^{\text{cat}}$ of members of the same matched set. Also assume that for $j = 1, \ldots, M+1$ independently,

$$\begin{aligned} &\boldsymbol{X}_j^{\text{con}} \mid D_1 = 1, D_2 = \cdots = D_{M+1} = 0, \boldsymbol{X}_1^{\text{cat}}, \boldsymbol{X}_2^{\text{cat}}, \ldots, \boldsymbol{X}_{M+1}^{\text{cat}}, \boldsymbol{u} \\ &\quad \sim N(\boldsymbol{\eta} + \boldsymbol{\xi}I(j=1) + \boldsymbol{\rho}\boldsymbol{X}_j^{\text{cat}} + \boldsymbol{\psi}\bar{\boldsymbol{X}}^{\text{cat}} + \boldsymbol{u}, \boldsymbol{\Lambda}) \end{aligned} \tag{8}$$

$$\begin{aligned} &\text{and } \boldsymbol{u} \mid D_1 = 1, D_2 = \cdots = D_{M+1} \\ &\quad = 0, \boldsymbol{X}_1^{\text{cat}}, \boldsymbol{X}_2^{\text{cat}}, \ldots, \boldsymbol{X}_{M+1}^{\text{cat}} \sim N(\boldsymbol{0}, \boldsymbol{\Omega}) \end{aligned} \tag{9}$$

where $\bar{\boldsymbol{X}}^{\text{cat}} = (M+1)^{-1}\sum_{j=1}^{M+1}\boldsymbol{X}_j^{\text{cat}}$. Note that $\boldsymbol{\psi}$ and $\boldsymbol{\Omega}$ allow correlation between one individual's $\boldsymbol{X}^{\text{con}}$ and the $\boldsymbol{X}^{\text{cat}}$ and $\boldsymbol{X}^{\text{con}}$ of other members of the same matched set. In Web Appendix E, we show this model implies equation (2) holds with $\boldsymbol{\beta}_{\text{cat}} = \boldsymbol{\tau} - \boldsymbol{\rho}^{\top}(\boldsymbol{F} - \boldsymbol{C})\boldsymbol{\xi}$ and $\boldsymbol{\beta}_{\text{con}} = (\boldsymbol{F} - \boldsymbol{C})\boldsymbol{\xi}$, where $\boldsymbol{C}^{-1} = \boldsymbol{\Lambda} + \boldsymbol{\Omega} - \boldsymbol{\Omega}(\boldsymbol{\Lambda} + M\boldsymbol{\Omega})^{-1}M\boldsymbol{\Omega}$ and $\boldsymbol{F} = -(\boldsymbol{\Lambda} + M\boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}\boldsymbol{C}$.

As in Section 4, we propose using FCS MI with conditional models compatible with this joint model. In Web Appendix F, we show that a compatible conditional model for a partially observed element of $\boldsymbol{X}_j^{\text{con}}$ is a linear regression of that element on $\boldsymbol{X}_j^{\text{cat}}$, $\sum_{k\neq j}\boldsymbol{X}_k^{\text{cat}}$, $\sum_{k\neq j}\boldsymbol{X}_k^{\text{con}}$ and all the remaining elements

of $\boldsymbol{X}_j^{\text{con}}$. Likewise, a compatible conditional model for one of the partially observed categorical variables making up $\boldsymbol{X}_j^{\text{cat}}$ is a multinomial logistic regression of this categorical variable on $\boldsymbol{X}_j^{\text{con}}$, $\sum_{k\neq j}\boldsymbol{X}_k^{\text{con}}$, $\sum_{k\neq j}\boldsymbol{X}_k^{\text{cat}}$ and those elements of $\boldsymbol{X}_j^{\text{cat}}$ that are not dummy indicators for this categorical variable. These conditional models are not the default options in MI software, because some predictors in the regression are sums of conditioning variables, e.g., $\sum_{k\neq j}\boldsymbol{X}_k^{\text{cat}}$. However, specification of non-default conditional models is straightforward (see Web Appendix H).

As with the FCS method in Section 4, this method is asymptotically equivalent to joint model MI, but in finite samples may be inefficient. Our second proposed model for $\boldsymbol{X}^{\text{set}}$ is a latent normal model with random effects (Carpenter and Kenward, 2013). Like the latent normal model of Section 4, this is not compatible with equation (2), but its use may improve efficiency. The latent normal model is the same as model (5), but with $\boldsymbol{\delta}_{\text{LN}}\boldsymbol{S}$ replaced by $\boldsymbol{u}$ and now conditioning on all of $D_1,\ldots,D_{M+1}$: for $j=1,\ldots,M+1$ independently,

$$(\boldsymbol{X}_j^{\text{con}\top},\boldsymbol{W}_j^{\text{cat}\top})^\top \mid D_1=1,D_2=\cdots=D_{M+1}=0,\boldsymbol{u} \sim N(\boldsymbol{\alpha}_{\text{LN}}+\boldsymbol{\phi}_{\text{LN}}D_j+\boldsymbol{u},\ \boldsymbol{\Sigma}_{\text{LN}}) \quad (10)$$

where $\boldsymbol{u}$ is normally distributed with mean zero and unstructured variance given $D_1=1,D_2=\cdots=D_{M+1}=0$. Again, joint model MI can be done using REALCOM-MI or jomo.

As in Section 4, there is a normal version of this model. This assumes (10) but with $\boldsymbol{W}_j^{\text{cat}}$ replaced by $\boldsymbol{X}_j^{\text{cat}}$. Joint model MI with this model can be done using the pan package of R.

## 6. Simulation Study

One thousand datasets were generated for each of 24 scenarios resulting from considering two sample sizes ($N=100$ or 500 matched sets), two numbers of matching controls ($M=1$ or $M=4$), three missingness mechanisms, and two proportions of missing data. Each dataset was generated using the model defined by expressions (3)–(4). Specifically, there were two matching variables, one binary ($S^{\text{cat}}$) and one continuous ($S^{\text{con}}$), and three covariates, one categorical ($X^{\text{cat}}$) and two continuous ($X^{\text{conA}}$ and $X^{\text{conB}}$). We assumed $P(S^{\text{cat}}=1\mid D=1)=0.6$ and $S^{\text{con}}\mid S^{\text{cat}},D=1\sim N(0,1)$. These could represent, for example, sex and standardized age. Among cases, the sex with greater risk would be more common, while age might be approximately normal if risk increases with age but total population size diminishes due to all-cause mortality. We assumed logit $P(X^{\text{cat}}=1\mid S^{\text{cat}},S^{\text{con}},D)=-2.5+0.5S^{\text{cat}}+0.5S^{\text{con}}+0.75D$ (so about 10% of controls and 20% of cases have $X^{\text{cat}}=1$), and $(X^{\text{conA}},X^{\text{conB}})$ given $X^{\text{cat}},S^{\text{cat}},S^{\text{con}},D$ is bivariate normal with univariate marginal distributions $N(0.5X^{\text{cat}}+0.5S^{\text{cat}}+0.5S^{\text{con}}+0.5D,1)$ and covariance 0.5. From $\boldsymbol{\beta}_{\text{cat}}=\boldsymbol{\lambda}-\boldsymbol{\gamma}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}$ and $\boldsymbol{\beta}_{\text{con}}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\phi}$, the true log ORs of $X^{\text{cat}}$, $X^{\text{conA}}$, and $X^{\text{conB}}$ are $\beta_{\text{cat}}=5/12$, $\beta_{\text{conA}}=1/3$, and $\beta_{\text{conB}}=1/3$.

Missingness was imposed on $X^{\text{cat}}$ and $X^{\text{conA}}$ assuming either missing completely at random (MCAR) or one of two MAR mechanisms. For MCAR data, each individual's $X^{\text{cat}}$ and $X^{\text{conA}}$ variables were independently missing with probability $p_{\text{miss}}$. Two values, $p_{\text{miss}}=0.1$ and $p_{\text{miss}}=0.25$, were considered. Thus, either 19 or 44% of individuals had at least one missing variable. For the first MAR mechanism (MAR-A), each individual's $X^{\text{cat}}$ and $X^{\text{conA}}$ variables were independently missing with logit probability $c_{\text{miss}}+0.25(X^{\text{conB}}+S^{\text{cat}}+S^{\text{con}}+D)$. For the second (MAR-B), it was $c_{\text{miss}}+0.25(X^{\text{conB}}+S^{\text{cat}}+S^{\text{con}}+D+X^{\text{conB}}D)$. In both cases, $c_{\text{miss}}$ was chosen to give $p_{\text{miss}}=0.1$ or $p_{\text{miss}}=0.25$ missingness in each of $X^{\text{cat}}$ and $X^{\text{conA}}$.

Each dataset was analyzed using CLR with $X^{\text{cat}}$, $X^{\text{conA}}$ and $X^{\text{conB}}$ as covariates. Missing data were handled in seven ways: complete-case analysis; FCS MI using matching variables or matched set (using ice in Stata); latent normal MI using matching variables or matched set (jomo in R); and normal MI using matching variables (mi impute in Stata) or matched set (pan in R). We used 25 imputed datasets when $p_{\text{miss}}=0.1$, and 50 when $p_{\text{miss}}=0.25$. In addition, the complete data were analyzed before imposing missingness on the covariates.

Tables 1 and 2 show results for the MCAR mechanism with 1:1 matching ($M=1$) and 1:4 matching ($M=4$) when $N=500$. The results from these scenarios also give a good indication of the general patterns observed for MAR-A, MAR-B and $N=100$ (see Web Tables 1–10). We shall focus on $\beta_{\text{cat}}$ and $\beta_{\text{conA}}$, since for $\beta_{\text{conB}}$ (the fully-observed covariate), differences between the three MI methods using matched variables were small, as were differences between those using matched set. To ease comparison of the six MI methods, Tables 3 and 4 show, for $p_{\text{miss}}=0.25$, the bias, ratio of empirical SEs, ratio of mean estimated SEs, and relative efficiency (i.e., ratio of mean squared errors, MSE) of each method, averaged over the three missingness mechanisms, separately for $\beta_{\text{cat}}$ and $\beta_{\text{conA}}$, for $N=100$ and 500, and for $M=1$ and $M=4$. Unsurprisingly, differences between methods were smaller when $p_{\text{miss}}=0.1$; we focus on $p_{\text{miss}}=0.25$ below.

The two FCS methods are approximately unbiased when $N=500$ and usually when $N=100$. Exceptions are when $N=100$ and $M=1$, where the complete-data method is also biased (with biases similar to those of FCS MI), and when $N=100$ and $M=4$, where there is bias for $\beta_{\text{cat}}$ when using matched set. Normal MI has some negative bias for $\beta_{\text{cat}}$, especially when using matching variables (except when $N=100$ and $M=1$, where its negative bias cancels out the positive bias of the complete-data estimator). Latent normal MI has some positive bias for $\beta_{\text{cat}}$ when $M=1$; latent normal MI using matched set also has negative bias for $\beta_{\text{conA}}$. The complete-case estimators are generally approximately unbiased, but note that the estimator of $\beta_{\text{conB}}$ is severely biased under MAR-B (Web Tables 2, 4, 7, and 10).

Empirical standard errors (SEs) from MI are almost always smaller when using matching variables than when using matched set, and negatively biased estimators tend to have smaller SEs. For $\beta_{\text{cat}}$, the SEs from FCS MI and latent normal MI are usually similar (when using matched set with $N=100$, FCS MI has the smaller SE); the smallest SEs come from normal MI. For $\beta_{\text{conA}}$, latent normal MI has the smallest SEs; the SEs from normal MI are similar to those from FCS MI when using matching variables and larger when using matched set. These differences are less marked when $M=4$.

Efficiency (mean square error, MSE) is a function of bias and SE. For $\beta_{\text{cat}}$, normal MI is most efficient, despite its bias; FCS MI and latent normal MI are usually about equally

**Table 1**

*Results from 1000 simulated datasets of $N = 500$ cases and $M = 1$ control per case with MCAR missingness mechanism. "LOR" is mean estimated log odds ratio, "SE" is empirical standard error, "estSE" is mean estimated standard error, "MSE" is mean-squared error $\times 1000$, and "cv" is coverage of 95% confidence interval. True log odds ratios are 0.417, 0.333, and 0.333 for $X^{\text{cat}}$, $X^{\text{conA}}$, and $X^{\text{conB}}$, respectively.*

| | $X^{\text{cat}}$ | | | | | $X^{\text{conA}}$ | | | | | $X^{\text{conB}}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LOR | SE | estSE | MSE | cv | LOR | SE | estSE | MSE | cv | LOR | SE | estSE | MSE | cv |
| Complete data | 0.426 | 0.213 | 0.206 | 45.3 | 94 | 0.336 | 0.078 | 0.082 | 6.15 | 96 | 0.337 | 0.083 | 0.082 | 6.88 | 95 |
| **10% missing** | | | | | | | | | | | | | | | |
| Complete cases | 0.431 | 0.264 | 0.256 | 70.0 | 96 | 0.337 | 0.100 | 0.102 | 10.1 | 96 | 0.340 | 0.103 | 0.102 | 10.6 | 96 |
| Match var: FCS | 0.429 | 0.224 | 0.219 | 50.2 | 94 | 0.335 | 0.083 | 0.087 | 6.92 | 96 | 0.338 | 0.085 | 0.084 | 7.17 | 94 |
| Normal | 0.410 | 0.214 | 0.218 | 45.7 | 96 | 0.336 | 0.083 | 0.087 | 6.85 | 96 | 0.339 | 0.084 | 0.083 | 7.16 | 95 |
| Latent norm | 0.435 | 0.223 | 0.218 | 50.0 | 95 | 0.330 | 0.081 | 0.087 | 6.63 | 96 | 0.340 | 0.084 | 0.084 | 7.12 | 94 |
| Match set: FCS | 0.429 | 0.225 | 0.219 | 51.0 | 95 | 0.334 | 0.084 | 0.087 | 7.05 | 96 | 0.338 | 0.085 | 0.084 | 7.23 | 94 |
| Normal | 0.420 | 0.221 | 0.223 | 49.0 | 96 | 0.340 | 0.086 | 0.089 | 7.44 | 96 | 0.335 | 0.086 | 0.084 | 7.34 | 95 |
| Latent norm | 0.437 | 0.226 | 0.220 | 51.6 | 95 | 0.320 | 0.082 | 0.087 | 6.83 | 96 | 0.340 | 0.085 | 0.084 | 7.35 | 95 |
| **25% missing** | | | | | | | | | | | | | | | |
| Complete cases | 0.449 | 0.379 | 0.377 | 145 | 96 | 0.341 | 0.144 | 0.149 | 20.8 | 97 | 0.342 | 0.150 | 0.149 | 22.4 | 96 |
| Match var: FCS | 0.431 | 0.240 | 0.241 | 57.6 | 96 | 0.336 | 0.090 | 0.096 | 8.06 | 96 | 0.337 | 0.087 | 0.086 | 7.51 | 95 |
| Normal | 0.386 | 0.215 | 0.235 | 47.1 | 97 | 0.338 | 0.090 | 0.095 | 8.04 | 97 | 0.341 | 0.086 | 0.086 | 7.49 | 95 |
| Latent norm | 0.446 | 0.238 | 0.241 | 57.7 | 96 | 0.322 | 0.085 | 0.095 | 7.31 | 97 | 0.343 | 0.085 | 0.086 | 7.39 | 95 |
| Match set: FCS | 0.430 | 0.247 | 0.243 | 61.0 | 95 | 0.335 | 0.094 | 0.097 | 8.81 | 96 | 0.339 | 0.088 | 0.086 | 7.69 | 95 |
| Normal | 0.407 | 0.238 | 0.251 | 56.8 | 96 | 0.350 | 0.098 | 0.101 | 9.93 | 96 | 0.329 | 0.090 | 0.088 | 8.08 | 94 |
| Latent norm | 0.455 | 0.249 | 0.247 | 63.7 | 95 | 0.300 | 0.085 | 0.095 | 8.27 | 95 | 0.344 | 0.088 | 0.088 | 7.89 | 95 |

efficient, with neither uniformly better than the other. For $\beta_{\text{conA}}$, latent normal MI is more efficient than FCS and normal MI when using matching variables; FCS and normal MI are equally efficient. When using matched set, FCS MI is more efficient for $\beta_{\text{conA}}$ than normal MI; latent MI is more efficient than FCS MI when $M = 1$, but is the least efficient of all the methods when $N = 500$ and $M = 4$, where its bias dominates its smaller SE. All MI methods are more efficient than the complete-case analysis.

The MI methods show a tendency to slightly overestimate SEs. Mostly, this is fairly mild, but is more severe for normal MI with $\beta_{\text{cat}}$ when $M = 1$, and for latent normal MI with $\beta_{\text{conA}}$ when $M = 1$ or 4. Thus, although normal and latent normal MI are most efficient for $\beta_{\text{cat}}$ and $\beta_{\text{conA}}$, respectively,

**Table 2**

*Results from 1000 simulated datasets of $N = 500$ cases and $M = 4$ controls per case with MCAR missingness mechanism. "LOR" is mean estimated log odds ratio, "SE" is empirical standard error, "estSE" is mean estimated standard error, "MSE" is mean-squared error $\times 1000$, and "cv" is coverage of 95% confidence interval. True log odds ratios are 0.417, 0.333, and 0.333 for $X^{\text{cat}}$, $X^{\text{conA}}$, and $X^{\text{conB}}$, respectively.*

| | $X^{\text{cat}}$ | | | | | $X^{\text{conA}}$ | | | | | $X^{\text{conB}}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | LOR | SE | estSE | MSE | cv | LOR | SE | estSE | MSE | cv | LOR | SE | estSE | MSE | cv |
| Complete data | 0.418 | 0.150 | 0.144 | 22.4 | 94 | 0.334 | 0.058 | 0.061 | 3.37 | 97 | 0.337 | 0.062 | 0.061 | 3.89 | 94 |
| **10% missing** | | | | | | | | | | | | | | | |
| Complete cases | 0.419 | 0.179 | 0.169 | 32.1 | 94 | 0.333 | 0.069 | 0.071 | 4.72 | 96 | 0.338 | 0.073 | 0.071 | 5.38 | 94 |
| Match var: FCS | 0.418 | 0.158 | 0.153 | 25.1 | 95 | 0.333 | 0.062 | 0.065 | 3.82 | 97 | 0.337 | 0.064 | 0.062 | 4.07 | 94 |
| Normal | 0.407 | 0.154 | 0.152 | 23.8 | 96 | 0.335 | 0.062 | 0.064 | 3.82 | 96 | 0.339 | 0.064 | 0.062 | 4.06 | 94 |
| Latent norm | 0.424 | 0.158 | 0.153 | 25.1 | 94 | 0.329 | 0.061 | 0.065 | 3.69 | 97 | 0.339 | 0.063 | 0.062 | 4.05 | 94 |
| Match set: FCS | 0.415 | 0.159 | 0.153 | 25.4 | 94 | 0.332 | 0.062 | 0.065 | 3.80 | 96 | 0.338 | 0.064 | 0.062 | 4.08 | 94 |
| Normal | 0.411 | 0.157 | 0.154 | 24.7 | 95 | 0.336 | 0.062 | 0.065 | 3.89 | 97 | 0.337 | 0.064 | 0.062 | 4.08 | 94 |
| Latent norm | 0.424 | 0.159 | 0.153 | 25.2 | 95 | 0.320 | 0.060 | 0.065 | 3.83 | 97 | 0.340 | 0.064 | 0.062 | 4.15 | 94 |
| **25% missing** | | | | | | | | | | | | | | | |
| Complete cases | 0.411 | 0.242 | 0.225 | 58.5 | 93 | 0.337 | 0.089 | 0.093 | 7.90 | 97 | 0.338 | 0.095 | 0.093 | 8.99 | 95 |
| Match var: FCS | 0.417 | 0.176 | 0.168 | 30.8 | 94 | 0.335 | 0.069 | 0.071 | 4.75 | 97 | 0.337 | 0.066 | 0.064 | 4.36 | 94 |
| Normal | 0.389 | 0.164 | 0.165 | 27.7 | 95 | 0.338 | 0.069 | 0.071 | 4.73 | 96 | 0.340 | 0.066 | 0.064 | 4.37 | 94 |
| Latent norm | 0.432 | 0.174 | 0.168 | 30.4 | 94 | 0.323 | 0.066 | 0.070 | 4.42 | 97 | 0.343 | 0.065 | 0.064 | 4.37 | 94 |
| Match set: FCS | 0.409 | 0.176 | 0.168 | 31.0 | 94 | 0.333 | 0.068 | 0.071 | 4.68 | 96 | 0.338 | 0.066 | 0.064 | 4.41 | 94 |
| Normal | 0.398 | 0.170 | 0.170 | 29.3 | 95 | 0.343 | 0.070 | 0.072 | 5.04 | 95 | 0.335 | 0.067 | 0.064 | 4.47 | 94 |
| Latent norm | 0.431 | 0.177 | 0.170 | 31.4 | 94 | 0.302 | 0.065 | 0.070 | 5.23 | 95 | 0.344 | 0.067 | 0.065 | 4.59 | 94 |

**Table 3**

*Biases ("bias"), ratios of empirical SEs ("ratio SE"), ratios of mean estimated SEs ("ratio empSE"), and relative efficiencies (%, "rel. eff.") of six MI methods when $N = 500$ and $p_{\text{miss}} = 0.25$. Ratios and relative efficiencies are calculated relative to the corresponding complete-data estimators. Each reported ratio or relative efficiency is the average over three ratios or relative efficiencies: one from each of the MCAR, MAR-A, and MAR-B scenarios. Reported biases are the signed average absolute bias over these three scenarios.*

| | $\beta_{\text{cat}}$ | | | | $\beta_{\text{conA}}$ | | | |
|---|---|---|---|---|---|---|---|---|
| | bias | ratio SE | ratio estSE | rel. eff. | bias | ratio SE | ratio estSE | rel. eff. |
| **M=1** | | | | | | | | |
| Complete data | 0.010 | 1.000 | 1.000 | 100.0 | 0.003 | 1.000 | 1.000 | 100.0 |
| Match var: FCS | 0.014 | 1.136 | 1.188 | 77.4 | 0.002 | 1.142 | 1.161 | 76.7 |
| Normal | −0.046 | 1.018 | 1.156 | 92.2 | 0.006 | 1.137 | 1.157 | 77.2 |
| Latent norm | 0.026 | 1.126 | 1.183 | 78.1 | −0.010 | 1.081 | 1.150 | 84.4 |
| Match set: FCS | 0.013 | 1.181 | 1.207 | 71.7 | −0.002 | 1.196 | 1.184 | 70.0 |
| Normal | −0.015 | 1.132 | 1.245 | 77.9 | 0.014 | 1.251 | 1.223 | 62.7 |
| Latent norm | 0.035 | 1.187 | 1.220 | 69.7 | −0.035 | 1.082 | 1.157 | 73.0 |
| **M=4** | | | | | | | | |
| Complete data | 0.001 | 1.000 | 1.000 | 100.0 | 0.001 | 1.000 | 1.000 | 100.0 |
| Match var: FCS | −0.001 | 1.195 | 1.200 | 70.0 | 0.002 | 1.188 | 1.172 | 70.8 |
| Normal | −0.051 | 1.108 | 1.171 | 73.9 | 0.005 | 1.184 | 1.169 | 70.9 |
| Latent norm | 0.010 | 1.177 | 1.195 | 72.0 | −0.010 | 1.129 | 1.163 | 76.6 |
| Matchset: FCS | −0.006 | 1.220 | 1.210 | 67.3 | −0.002 | 1.190 | 1.187 | 70.5 |
| Normal | −0.036 | 1.149 | 1.206 | 72.0 | 0.007 | 1.215 | 1.196 | 67.0 |
| Latent norm | 0.008 | 1.200 | 1.211 | 69.3 | −0.034 | 1.126 | 1.168 | 62.4 |

**Table 4**

*Biases ("bias"), ratios of empirical SEs ("ratio SE"), ratios of mean estimated SEs ("ratio empSE"), and relative efficiencies (%, "rel. eff.") of six MI methods when $N = 100$ and $p_{\text{miss}} = 0.25$. Ratios and relative efficiencies are calculated relative to the corresponding complete-data estimators. Each reported ratio or relative efficiency is the average over three ratios or relative efficiencies: one from each of the MCAR, MAR-A and MAR-B scenarios. Reported biases are the signed average absolute bias over these three scenarios.*

| | $\beta_{\text{cat}}$ | | | | $\beta_{\text{conA}}$ | | | |
|---|---|---|---|---|---|---|---|---|
| | bias | ratio SE | ratio estSE | rel. eff. | bias | ratio SE | ratio estSE | rel. eff. |
| **M=1** | | | | | | | | |
| Complete data | 0.038 | 1.000 | 1.000 | 100.0 | 0.018 | 1.000 | 1.000 | 100.0 |
| Match var: FCS | 0.048 | 1.199 | 1.227 | 69.7 | 0.025 | 1.164 | 1.192 | 73.5 |
| Normal | −0.017 | 1.068 | 1.186 | 88.3 | 0.027 | 1.150 | 1.185 | 75.2 |
| Latent norm | 0.066 | 1.193 | 1.219 | 69.9 | 0.008 | 1.080 | 1.178 | 86.4 |
| Match set: FCS | 0.041 | 1.221 | 1.277 | 67.3 | 0.022 | 1.234 | 1.231 | 65.7 |
| Normal | 0.029 | 1.160 | 1.309 | 74.7 | 0.051 | 1.321 | 1.295 | 55.6 |
| Latent norm | 0.099 | 1.304 | 1.301 | 57.9 | −0.013 | 1.085 | 1.203 | 85.3 |
| **M=4** | | | | | | | | |
| Complete data | −0.015 | 1.000 | 1.000 | 100.0 | 0.003 | 1.000 | 1.000 | 100.0 |
| Match var: FCS | −0.018 | 1.200 | 1.217 | 69.5 | 0.005 | 1.130 | 1.183 | 78.3 |
| Normal | −0.065 | 1.099 | 1.184 | 80.3 | 0.008 | 1.124 | 1.176 | 78.9 |
| Latent norm | −0.009 | 1.191 | 1.210 | 70.7 | −0.008 | 1.073 | 1.172 | 86.6 |
| Match set: FCS | −0.037 | 1.157 | 1.213 | 74.3 | −0.011 | 1.091 | 1.183 | 83.7 |
| Normal | −0.040 | 1.136 | 1.228 | 76.7 | 0.012 | 1.172 | 1.211 | 72.5 |
| Latent norm | −0.010 | 1.234 | 1.237 | 65.8 | −0.026 | 1.079 | 1.185 | 83.6 |

**Table 5**

*Association between fibre intake and colerectal cancer estimated from EPIC-Norfolk. Log odds ratio is for six-gram per day increase in fibre intake, conditional on smoking, education, social class, physical activity, height, weight, age, alcohol intake, folate intake, intake of energy from fat and non-fat, aspirin use, and the matching variables. Missing data are handled by restriction to complete cases or by MI.*

| Method | log OR | SE | 95% CI | p-value |
|---|---|---|---|---|
| Complete cases | −0.196 | 0.126 | (−0.444, 0.052) | 0.121 |
| MI using matching variables: | | | | |
| FCS | −0.176 | 0.104 | (−0.380, 0.027) | 0.090 |
| Normal | −0.177 | 0.104 | (−0.380, 0.027) | 0.088 |
| Latent normal | −0.176 | 0.104 | (−0.380, 0.027) | 0.089 |
| MI using matched set: | | | | |
| FCS | −0.175 | 0.104 | (−0.378, 0.028) | 0.092 |
| Normal | −0.181 | 0.104 | (−0.384, 0.023) | 0.082 |
| Latent normal | −0.174 | 0.104 | (−0.377, 0.030) | 0.094 |

this advantage is not apparent in the width of the estimated confidence intervals. Indeed, the average estimated SEs of the three MI methods using matching variables were generally rather similar; the same was true of the methods using matched set. Coverage of 95% confidence intervals was between 93% and 97% for all methods.

We also performed two simulation studies using modified data-generating mechanisms that make our imputation models misspecified. In the first, there was an interaction between $S^{\text{cat}}$ and $S^{\text{con}}$; in the second, $X^{\text{conA}}$ and $X^{\text{conB}}$ were log-normally distributed. See Web Appendix G and Web Tables 11–16 for details and results. Briefly, none of the MI methods showed considerable bias for either of these data-generating mechanisms, and all MI methods were much more efficient than the complete-case analysis.

In summary, all the MI methods appear to work well. Using matching variables is more efficient than using matched set. If using matching variables, normal and latent normal MI appear to be preferable to FCS MI, which is less efficient; normal MI is more efficient for $\beta_{\text{cat}}$, but latent normal MI more efficient for $\beta_{\text{conA}}$. Of these, one might prefer latent normal MI, because of the bias in $\beta_{\text{cat}}$ for normal MI. If using matched set, FCS MI might be preferred when $M = 4$, on bias and efficiency grounds. However, when $M = 1$ and using matched set, no method appears better than any other.

## 7. Analysis of EPIC-Norfolk Data

Table 5 shows the estimated adjusted log OR for fibre intake from the complete-case analysis. This analysis excludes all matched sets in which the case had missing data, as well as any controls with missing data. It uses 240 (75%) matched sets consisting of 240 cases and 772 controls. Also shown are the results of the three MI methods using matching variables, including sex, age and date of diary completion as $\boldsymbol{S}$ in the imputation model. The complete-case and MI analyses produce similar log OR estimates (differing by less than 20% of an SE), but the latter are more efficient, because they use all 318 cases and 1272 controls i.e., 33% more matched sets, and this is reflected by a 17% reduction in estimated SE.

MI using matching variables imputes missing values assuming that age, sex, and time of diary completion have linear and additive effects on the logit probability of disease. Furthermore, the way that recruitment took place in the EPIC-Norfolk cohort means that date of diary completion is predictive of which GP surgery the individual was registered with, and hence matching by the former tends to induce some degree of matching by the latter. Treating date of diary completion as a continuous variable will not fully account for this. For these two reasons, one might prefer MI using matched set, or might wish to check that the results from the two methods do not differ substantially.

Table 5 shows that the results from MI using matching variables and MI using matched set are very similar, providing some reassurance about the validity of both sets of results. In this study, both approaches can be used, but had matching been on GP practice itself, MI using matched set might have been the only feasible option.

## 8. Discussion

We have described two broad MI approaches to the analysis of matched case-control studies with missing values in covariates, and three methods within each approach. One approach involves parametric modeling of the association between the matching variables and the partially observed covariates; the other instead treats matched set as a random effect. Our simulation results suggest that the first approach is preferable when it can be done, as it is more efficient. However, in studies where matching is on, e.g., family, GP practice or postcode area of residence, or if data on the matching variables are not available to the analyst, the first approach is not feasible and the second approach can be used instead. The second approach might also be preferred if one were reluctant to specify a form for association between matching variables and covariates in the imputation model, because, for example, there were several matching variables, including continuous ones and potential interactions.

Of the three MI methods within each approach, FCS MI based on a restricted general location model and joint model MI using a multivariate normal distribution can be implemented in many statistical packages, whereas joint model MI using a latent normal distribution is currently limited to R and the specialist software REALCOM-MI. All three methods are easy to use, appear to work well, and are more efficient than the complete-case analysis. They can all handle continuous and nominal categorical covariates, multiple partially-observed covariates, non-monotone missingness patterns, and multiple controls per case. Computer commands to implement the methods are given in Web Appendix H.

FCS MI has the theoretical appeal of being asymptotically equivalent to joint model MI using an imputation model (the restricted general location model) that is compatible with the CLR analysis model. It nearly always gave the least-biased estimates in simulations. However, when using matching variables, normal MI and latent normal MI were more efficient. When using matched set, FCS MI was marginally better than normal and latent normal MI for a 1:4 matched study ($M = 4$); no method was obviously best or worst for 1:1 matching ($M = 1$). A drawback of FCS MI using matched set

when $M > 1$ is that the estimates may depend on the arbitrary order chosen for the $M$ controls in each matched set. Any order produces valid imputations, but one could avoid this dependence by randomly permuting indices of controls within matched sets before generating each imputed dataset. However, that is only likely to be worthwhile if the sample size is small and there are a lot of missing data. Normal MI was, in general, the most biased of the three methods, but even its biases were fairly modest. A slight drawback of normal MI is the need manually to post-process imputed values of categorical variables, e.g., using adaptive rounding. None of the MI methods was uniformly superior to the others in simulations, and we regard use of any of them as entirely acceptable.

All methods can handle the situation where the number of matched controls, $M$, varies between cases, although this is slightly more complicated for FCS MI using matched set. For this method, extra controls with completely missing data would have to be added to those matched sets with fewer than the maximum number of controls, before performing MI, and then deleted again before analyzing the imputed datasets.

Another method, which merits further research, is joint model MI using the restricted general location model. This requires specialist Bayesian software and more advanced programming skills, and the focus of this article is on methods that are easy to implement in standard packages. Nevertheless, it would be worth investigating whether this method is significantly more efficient than our FCS MI method based on the same model. The mix package in R (Schafer, 1997) may also be of interest. This uses a model similar to (3)–(4), but additionally assumes the continuous part of $\boldsymbol{S}$ is normally distributed given $D$, $\boldsymbol{X}^{\text{cat}}$ and the rest of $\boldsymbol{S}$. The MI methods considered in this article assume, like the CLR analysis model, nothing about the distribution of the matching variables. Mix cannot be used for MI using matched set.

Finally, we note that, as always with missing data methods, it is important to consider the plausibility of the assumption about the missing data mechanism. Often, the MAR assumption can be made more plausible by including in the imputation model additional variables that are associated with the partially observed covariates.

## 9. Supplementary Materials

Web Appendices referenced in Sections 3–8, along with computer code, are available with this paper at the *Biometrics* website on Wiley Online Library.

## Acknowledgements

SRS is funded by MRC Grants U105260558 and MC_US_A030_0015. We would like to acknowledge the contribution of the staff and participants of the EPIC-Norfolk Study. EPIC-Norfolk is supported by the MRC programme Grant (G0401527).

## References

Ahn, J., Mukherjee, B., Gruber, S., and Sinha, S. (2011). Missing exposure data in stereotype regression model: Application to matched case-control study with disease subclassification. *Biometrics* **67**, 546–558.

Asano, T. and McLeod, R. (2004). Non steroidal anti-inflammatory drugs (NSAID) and aspirin for preventing colorectal adenomas and carcinomas. *Cochrane Database of Systematic Reviews* **Issue 1**, 1–35, Art. No.: CD004079. DOI: 10.1002/14651858.CD004079.pub2.

Aune, D., Chan, D., Lau, R., Vieira, R., Greenwood, D., Kampman, E., and Norat, T. (2011). Dietary fibre, whole grains, and risk of colorectal cancer: Systematic review and dose-response meta-analysis of prospective studies. *British Medical Journal* **343**, d6617.

Bernaards, C., Belin, T., and Schafer, J. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine* **26**, 1368–82.

Breslow, N. and Day, N. (1980). *Statistical Methods in Cancer Research, Volume I – The Analysis of Case-Control Studies*. Lyon, France: IARC Scientific Publications.

Carpenter, J. and Kenward, M. (2013). *Multiple Imputation and Its Applications*. New Jersey, NJ: Wiley.

Dahm, C., Keogh, R., Spencer, E., Greenwood, D., Key, T., Fentiman, I. et al. (2010). Dietary fiber and colorectal cancer risk: A nested case-control study using food diaries. *Journal of the National Cancer Institute* **102**, 614–626.

Gebregziabher, M. and DeSantis, S. (2010). Latent class based multiple imputation approach for missing categorical data. *Journal of Statistical Planning and Inference* **140**, 3252–3262.

Hughes, R., White, I., Seaman, S., Carpenter, J., Tilling, K., and Sterne, J. (2014). Joint modelling rationale for chained equations. *BMC Medical Research Methodology* **14**, 28.

Lee, K. and Carlin, J. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology* **171**, 624–632.

Lipsitz, S., Parzen, M., and Ewell, M. (1998). Inference using conditional logistic regression with missing covariates. *Biometrics* **54**, 295–303.

Liu, J., Gelman, A., Hill, J., Su, Y., and Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika* **101**, 155–173.

Liu, T., Yuan, X., Li, Z., and Li, Y. (2013). Empirical and weighted conditional likelihoods for matched case-control studies with missing covariates. *Journal of Multivariate Analysis* **119**, 185–199.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* **10**, 325–337.

Moons, K., Donders, R., Stijnen, T., and Harrell, F. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology* **59**, 1092–1101.

National Cancer Institute (2014). Colorectal cancer prevention (PDQ). http://www.cancer.gov/cancertopics/pdq/prevention/colorectal/HealthProfessional/page3. Accessed: 17/06/2014.

Paik, M. (2004). Nonignorable missingness in matched case-control data analyses. *Biometrics* **60**, 306–314.

Paik, M. and Sacco, R. (2000). Matched case-control data analyses with missing covariates. *Applied Statistics* **49**, 145–156.

Rathouz, P. (2003). Likelihood methods for missing covariate data in highly stratified studies. *Journal of the Royal Statistical Society, Series B* **65**, 711–723.

Rathouz, P., Satten, G., and Carroll, R. (2002). Semiparametric inference in matched casecontrol studies with missing covariate data. *Biometrika* **89**, 905–916.

Satten, G. and Carroll, R. (2000). Conditional and unconditional categorical regression models with missing covariates. *Biometrics* **56**, 384–388.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Sinha, S., Mukherjee, B., Ghosh, M., Mallick, B., and Carroll, R. (2005). Semiparametric Bayesian analysis of matched case-control studies with missing exposure. *Journal of the American Statistical Association* **100**, 591–601.

Sinha, S. and Wang, S. (2009). A new semiparametric procedure for matched case-control studies with missing covariates. *Journal of Nonparametric Statistics* **21**, 889–905.

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. London: Chapman & Hall/CRC.

*Received July* 2014. *Revised April* 2015. *Accepted May* 2015.