HHS Public Access

Author manuscript

J Surv Stat Methodol. Author manuscript; available in PMC 2019 April 02.

Published in final edited form as:

J Surv Stat Methodol. 2016 September ; 4(3): 316–338. doi:10.1093/jssam/smw002.

MULTIPLE IMPUTATION FOR MISSINGNESS DUE TO NONLINKAGE AND PROGRAM CHARACTERISTICS:

A CASE STUDY OF THE NATIONAL HEALTH INTERVIEW SURVEY LINKED TO MEDICARE CLAIMS

GUANGYU ZHANG[*] **[senior service fellow]**,
National Center for Health Statistics, Hyattsville, MD 20782, USA

JENNIFER D. PARKER [senior statistician], and
National Center for Health Statistics, Hyattsville, MD 20782, USA

NATHANIEL SCHENKER [Deputy Director and Acting Associate Director for Research and Methodology]
National Center for Health Statistics, Hyattsville, MD 20782, USA

Abstract

Record linkage is a valuable and efficient tool for connecting information from different data sources. The National Center for Health Statistics (NCHS) has linked its population-based health surveys with administrative data, including Medicare enrollment and claims records. However, the linked NCHS-Medicare files are subject to missing data; first, not all survey participants agree to record linkage, and second, Medicare claims data are only consistently available for beneficiaries enrolled in the Fee-for-Service (FFS) program, not in Medicare Advantage (MA) plans. In this research, we examine the usefulness of multiple imputation for handling missing data in linked National Health Interview Survey (NHIS)–Medicare files. The motivating example is a study of mammography status from 1999 to 2004 among women aged 65 years and older enrolled in the FFS program. In our example, mammography screening status and FFS/MA plan type are missing for NHIS survey participants who were not linkage eligible. Mammography status is also missing for linked participants in an MA plan. We explore three imputation approaches: (i) imputing screening status first, (ii) imputing FFS/MA plan type first, (iii) and imputing the two longitudinal processes simultaneously. We conduct simulation studies to evaluate these methods and compare them using the linked NHIS-Medicare files. The imputation procedures described in our paper would also be applicable to other public health–related research using linked data files with missing data issues arising from program characteristics (e.g., intermittent enrollment or data collection) reflected in administrative data and linkage eligibility by survey participants.

[*]Address correspondence to Guangyu Zhang, 3311 Toledo road, Hyattsville, MD 20782, USA; vha1@cdc.gov.

Supplementary Materials

Supplementary materials are available online at http://www.oxfordjournals.org/our_journals/jssam/. The online supplementary material consists of two sections. Section 1 contains tables 1 to 4. Table 1 shows the NHIS variables we included in the MI models and the percentages and means of these variables by linkage, FFS/MA status, and mammography status. Tables 2 and 3 contain results of simulations 1 and 2, respectively, including biases, standard deviations, and root mean square errors of the estimated percentages of FFS participants with mammography screening. Table 4 shows ranges of lag 1 to lag 3 and above ϕ coefficients by different MI methods for the two simulation studies ($\beta = 7$) and the linked NHIS-Medicare data. Section 2 contains details of the simulation 2 setup.

Keywords

Imputation; Missing data; Record linkage

1. INTRODUCTION

Record linkage, also known as de-duplication and entity resolution, is the task of identifying, matching, and merging lists of possibly distorted records referring to the same individual, often without unique identifier, from several data sources (Christen 2012). It is different from statistical matching, in which the purpose is not to link records for the same individual (Rodgers 1984; Moriarity and Scheuren 2001). Rather, in statistical matching, the linking variables are statistically and scientifically related, and the different data sources being matched may contain disjoint samples. In record linkage, when highly discriminative or unique identifiers exist in different data sources, deterministic linkage methods, in which pairs of records are classified as links and nonlinks based on certain predetermined rules, can be used (Harron, Goldstein, and Dibben 2015). Typically, deterministic linkage requires exact agreement on identifiers and matching variables.

The National Center for Health Statistics (NCHS) has developed a record linkage program to link the center's population-based health surveys with administrative data, including Medicare enrollment and claims records collected from the Centers for Medicare and Medicaid Services (CMS). Although Medicare data include detailed cost and service use information, these data are created for billing purposes and often lack demographic, health, and risk factor information useful for a health study; hence the utility of linking them to survey data. The last release of linked NCHS-Medicare data was in 2010, with data files available for health research in the NCHS Research Data Center (RDC). This linkage included Medicare data through 2007 and National Health Interview Survey (NHIS) data through 2005, and it was produced using a deterministic linkage method described below.

A few issues affect the use of linked survey data. The NCHS-Medicare data released in 2010 were produced under an interagency agreement among the Office of the Assistant Secretary for Planning and Evaluation, the Social Security Administration (SSA), CMS, and NCHS. To decrease disclosure risks for both survey participants and program beneficiaries, Social Security number (SSN) (or health insurance claim [HIC] number), sex, and date of birth were used for linkage, and these fields needed to match exactly (i.e., a deterministic linkage). The SSNs were verified for accuracy by the SSA. NCHS survey participants who refused to provide their SSNs or HIC numbers for linkage were considered to have refused record linkage and were, as a result, not eligible for linkage (National Center for Health Statistics 2011). The percentage of NHIS participants refusing to provide their SSN reached an overall high of more than 50 percent in the mid-2000s (Miller, Gindi, and Parker 2011). Date of birth was missing for approximately 0.2 percent of participants who had provided an SSN; no records were missing sex. In general, approximately 98 percent of linkage-eligible NHIS survey participants (i.e., those who did not refuse linkage and who had complete data for sex and date of birth) aged 65 years and older at interview were successfully linked to Medicare data.

In addition, studies based on Medicare records often are limited to a subset of Medicare beneficiaries because medical claims data are consistently available only for those in the Fee-for-Service (FFS) components of the Medicare program; information for beneficiaries enrolled in one of the managed care plans, currently known as Medicare Advantage (MA) plans, is less consistently available because beneficiaries enrolled in MA receive Medicare services through their plan and individual claims are not paid through FFS.

Finally, the addition of longitudinal administrative records can add missingness and other complexities to analyses based on linked data files due to many factors, including intermittent or changing program enrollment and eligibility (Simon and Schoendorf 2014), changes in program characteristics (such as FFS or MA plan type), and mobility (Miller, Miller, Judson, He, Day et al. 2014).

In this paper we compare three multiple imputation (MI) procedures for handling missing data due to two main sources of missingness in the NHIS-Medicare files. First, as described above, survey participants who were not linkage eligible (e.g., refused to provide a SSN or HIC number) could not be linked; as a result, information from the Medicare files is not available for these participants. Second, of the linked survey participants, detailed medical information is not currently available for most MA enrollees in the linked data files. The motivating example is a study of annual mammography status using 2004–2005 NHIS data linked to 1999–2004 Medicare data. Only a small number of women aged 65 years or older who were linkage eligible in our data had not been linked to Medicare, and addressing missing data due to the deterministic linkage method was not considered in this study.

Recent reviews of methods for handling missing data can be found in Horton and Kleinman (2007), Little (2008), Ibrahim and Molenberghs (2009), Andridge and Little (2010), White, Royston, and Wood (2011), and Cheema (2014). Two commonly used imputation strategies are joint modeling and sequential regression multivariate imputation (SRMI) (Schafer 1997; Raghunathan, Lebkowski, VanHoewyk, and Solenberger 2001; Van Buuren 2007, 2012; He 2010; Van Buuren and Karin 2011). The joint modeling approach assumes the complete data (i.e., if there were no missingness) follow a joint distribution with unknown parameters. Pairing this assumption with assumptions about the missing data mechanism implies a predictive distribution from which the missing values can be drawn. This approach is theoretically sound but hard to implement for high-dimensional data with different distributional forms for the variables. Under the SRMI approach, imputation models are constructed for each individual variable separately, without explicit consideration of a joint model for the complete data. Sequential regression modeling is flexible and can incorporate variables of different types and distributions. The imputation methods we describe in this paper follow the SRMI approach. To account for imputation uncertainty, multiple datasets can be created by replacing the missing values with independent sets of draws from the predictive distribution. With such MI, data users can perform statistical analysis separately for each imputed dataset and, using Rubin's MI combining rules, derive the final results (Rubin 1978, 1987, 1996; Rubin and Schenker 1986; Barnard and Rubin 1999).

The paper is organized as follows. In section 2, we describe the data files and the motivating example. In section 3, we describe our three MI procedures in detail. In section 4, we show

simulation results on the performance of the MI procedures for data generated under two different scenarios. In section 5, we apply the three MI methods to the linked NHIS-Medicare data. In section 6, we compare the observed data and the imputed data for the simulation studies and the linked NHIS-Medicare data. Section 7 contains concluding remarks.

2. DATA AND MOTIVATING EXAMPLE

2.1 The NHIS and Medicare Data

The NHIS is a cross-sectional survey that was initiated in 1957 (National Center for Health Statistics 2005, 2006). The sampling plan of the NHIS follows a multistage probability design that permits representative sampling of the civilian, noninstitutionalized U.S. population. The current questionnaire contains a basic module and various supplements. The basic module contains questions on health, demographic, and socioeconomic characteristics. The supplements are used to obtain additional information on subjects already covered in the basic module or on different topics. Self-reported mammography screening is available in the supplements but only for selected years of the NHIS.

Medicare is a national insurance program, administered by CMS since 1965. Medicare provides health insurance for people aged 65 years and older, people younger than 65 years with disabilities, and people of all ages with end-stage renal disease. Administration of the Medicare program leads to multiple data files each year, many of which have been linked to the NHIS, including the Denominator file, which includes information on enrollment and FFS/MA plan type, and the Carrier file, where the mammography claims are recorded.

As described above, the linkage of the NHIS with Medicare data was conducted as part of the general NCHS-CMS linkage activity to create data files for health research.

2.2 A Motivating Example

For this paper, we used 2004–2005 NHIS data linked to 1999–2004 Medicare data to estimate annual percentages of women aged 65 years and older in the FFS program who have mammography screening. Medicare data contain mammography claims information that have been used to study mammography screening among women, particularly those aged 65 years and older (e.g., Townsend-Rocchiccioli and Steele 2002; Braithwaite, Zhu, Hubbard, O'Meara, Miglioretti et al. 2013). All women aged 40 years and older with Medicare coverage can be reimbursed for a screening mammogram every 12 months.

Women who were aged 65 years or older in 2004 were included in the study; we assumed these women were eligible for Medicare for at least one year from 1999 through 2004 because of their age. Based on this criterion, 12,137 women were included from the 2004–2005 NHIS. Among these women, 6,939 (57 percent) were not eligible to be linked, or did not link, to any of the Medicare data for the years 1999–2004. For the linked women, we assume that mammography screening percentages can be estimated by Medicare claims for FFS enrollees but, as described above, not for MA enrollees (although a small number of claims were submitted for these women); of the women linked to Medicare, approximately 20 percent were in MA each year. For the unlinked survey participants, MI methods,

described below, used demographic-, health-, and income-related variables from the NHIS to impute the mammography status and FFS/MA plan type.

Supplementary table 1 (please see the supplementary data online) shows the NHIS variables we included in the MI models. These variables were found to be statistically significantly different between the linked and unlinked survey participants and related to mammography status and/or FFS/MA plan type among the linked survey participants. The missingness percentages for these variables are listed in the table as well. Missing values for these variables were imputed as part of our MI process. In addition, we used previously imputed family income information, which was released by NCHS for public use, due to the high percentage of missingness in this variable (Schenker, Raghunathan, Chiu, Makuc, Zhang et al. 2006). As a result, the missingness percentage for family income shown on the table is zero. The table also shows means and percentages of these NHIS variables by linkage status (linked vs. not linked), plan type, and mammography status (mammography vs. without mammography).

We assumed the data were missing at random (MAR) because we could identify a large number and variety of covariates from the NHIS related to linkage eligibility, plan type, and mammography status (supplementary table 1; please see the supplementary data online). Little and Rubin (2002) recommend conditioning on as many covariates as possible related to the missingness and the response variable(s) to increase the plausibility of the MAR assumption, which we did.

3. MULTIPLE IMPUTATION WITH SEQUENTIAL REGRESSION MULTIVARIATE IMPUTATION

We conducted MI using sequential regression multivariate imputation (Raghunathan et al. 2001) as implemented by IVEware. For the problem of estimating the percentage of women in FFS Medicare with mammography screening using the linked NHIS-Medicare data, mammography information and FFS/MA plan type are missing for survey participants who are not linkage eligible and for the few who are eligible but did not link. Mammography information is also missing for most women in MA. Although we do not directly estimate mammography screening percentages for this group, mammography data for women in MA are used, when available, for some MI models. Because there are two main variables with missingness, MI could be applied in one of three ways: (i) by imputing for plan type first and then mammography status; (ii) by imputing for mammography status first and then plan type; or (iii) by imputing for both variables simultaneously.

We now express the above ideas more formally. Let $\underline{\mathbf{X}}_{\mathrm{NHIS}}$ be a vector of covariates from the NHIS. Define $\underline{\mathbf{P}} = (P_1, \ldots, P_J)$ to be a vector of the plan types (MA/FFS) over J years, j = 1, …, J, with

$P_j = 1$ if a participant enrolls in MA at time j,

$P_j = 0$ if a participant enrolls in FFS at time j, and

P_j = NA (not applicable) if a participant is aged less than 65 years at time j.

Let $\underline{\mathbf{S}} = (S_1, \ldots, S_J)$ represent a vector of mammography screening claim statuses over J years of a participant, with

$S_j = 1$ if a participant files mammography claims through Medicare at time j,

$S_j = 0$ if she doesn't file mammography claims through Medicare at time j, and

$S_j = \text{NA}$ if a participant is aged less than 65 years at time j, j = 1, …, J.

Let $f(\underline{\mathbf{S}}, \underline{\mathbf{P}} | \underline{\mathbf{X}}_{\text{NHIS}})$ denote the joint distribution of $\underline{\mathbf{S}}$ and $\underline{\mathbf{P}}$, conditional on $\underline{\mathbf{X}}_{\text{NHIS}}$. One way we can factorize the joint distribution is as follows:

$$f(\underline{\mathbf{S}}, \underline{\mathbf{P}} | \underline{\mathbf{X}}_{\text{NHIS}}) = f(\underline{\mathbf{S}} | \underline{\mathbf{X}}_{\text{NHIS}}) f(\underline{\mathbf{P}} | \underline{\mathbf{S}}, \underline{\mathbf{X}}_{\text{NHIS}}) \quad (1)$$

Imputation based on (1) imputes the missing mammography statuses first and then imputes the missing plan types given the mammography statuses. This is the first imputation procedure we explored in this paper, and we call it the screening-first approach.

With the screening-first approach, for survey participants aged 65 years or older in each year, we imputed mammography status for the unlinked participants. After this imputation was completed for all years, we imputed plan type using the imputed mammography statuses; for example, S_j is imputed based on the conditional distribution $f(S_j | S_1, \ldots, S_{j-1}, S_{j+1}, \ldots, S_J, \underline{\mathbf{X}}_{\text{NHIS}})$, and P_j is imputed based on the conditional distribution $f(P_j | P_1, \ldots, P_{j-1}, P_{j+1}, \ldots, P_J, \underline{\mathbf{S}}, \underline{\mathbf{X}}_{\text{NHIS}})$, j = 1, …, J. The missing values in $\underline{\mathbf{X}}_{\text{NHIS}}$ are also imputed during the first step of imputation, that is, the same step that imputes for $\underline{\mathbf{S}}$.

Another way we can factorize the joint distribution is as follows:

$$f(\underline{\mathbf{S}}, \underline{\mathbf{P}} | \underline{\mathbf{X}}_{\text{NHIS}}) = f(\underline{\mathbf{P}} | \underline{\mathbf{X}}_{\text{NHIS}}) f(\underline{\mathbf{S}} | \underline{\mathbf{P}}, \underline{\mathbf{X}}_{\text{NHIS}}) \quad (2)$$

This is the reverse of (1). With this plan-first approach, after imputation of plan status, we imputed the screening status for FFS participants ($P_j = 0$, j = 1, …, J) but not for MA participants ($P_j = 1$, j = 1, …, J), consistent with our objective to estimate screening for FFS beneficiaries only.

Theoretically, when we have complete data, (1) and (2) are the same for estimating the joint distribution $f(\underline{\mathbf{S}}, \underline{\mathbf{P}} | \underline{\mathbf{X}}_{\text{NHIS}})$. However, when missing values exist, the order of imputation could affect imputation results. For example, (2) works better when P_j can be effectively predicted by P_1, …, P_{j-1}, P_{j+1}, …, P_J and $\underline{\mathbf{X}}_{\text{NHIS}}$ and S_j can be effectively predicted by S_1, …, S_{j-1}, S_{j+1}, …, S_J, $\underline{\mathbf{P}}$ and $\underline{\mathbf{X}}_{\text{NHIS}}$.

The third approach we explored is called one-step imputation. For this approach, we do not impute the mammography screening status and plan type separately. As above, we started with participants who were aged 65 years or older in a given year and limited imputation to

the unlinked participants. Imputation of the mammography claim status $\underline{\mathbf{S}}$ and plan type $\underline{\mathbf{P}}$ across years is iterated between $f\left(S_j|S_1,\ldots,S_{j-1},S_{j+1},\ldots,S_J,\underline{\mathbf{P}},\underline{\mathbf{X}}_{NHIS}\right)$ and $f\left(P_j|P_1,\ldots,P_{j-1},P_{j+1},\ldots,P_J,\underline{\mathbf{S}},\underline{\mathbf{X}}_{NHIS}\right)$, j = 1, …, J.

Among these three types of methods, the one-step type is widely used in other applications and usually yields satisfactory results. However, factorization of the joint likelihood has certain advantages, such as saving computing time and producing relatively consistent imputed data. Section 7 contains more information on strengths and weaknesses of the three methods.

4. SIMULATION STUDIES

We conducted two simulation studies. Following the imputation approaches described in section 3, for simulation 1, we simulated the data based on (1) by generating the mammography screening status first and then generating the FFS/MA enrollment status. For simulation 2, we created simulated data based on (2).

For both simulations, we generated data based on the results of logistic regression models of plan type and logistic regression models of mammography screening at each year. Plan type and/or mammography screening at the same and/or at different years are the most important predictors for plan type and mammography screening for a specific year. Based on the real data, plan type across years is highly positively associated, as is mammography across years. Consequently, we used similar coefficients on plan type and mammography screening for the simulated logistic regression models of plan type and mammography screening. Because the covariates from the NHIS used for the MI are mainly binary or categorical (supplementary table 1; please see the supplementary data online), we used ten binary variables $X_1 - X_{10}$ to represent them. We allowed some of the variables to be related to plan type and/or mammography, some of the variables to be related to linkage, and some of the variables to be noise. The X variables in the simulations do not correspond to specific covariates in the NHIS. Details of the simulation setups are in section 4.1. Compared with the real data, our simulation datasets have similar percentages of FFS/MA (results not shown), similar percentages of mammography screening among FFS participants, and perhaps most important, similar longitudinal associations of plan type and mammography screening across years. We generated data for five time points with age at time 1 ranging from 61 years to 70 years. For each simulation, we simulated fifty replicates with sample size of 12,000 each.

4.1 Simulation Study Setup

Let X_1 to X_{10} be ten binary independent variables with parameters ranging from 0.3 to 0.7. For simulation 1, the mammography screening status, S_1 to S_5, is related to X_1 to X_4 and the previous mammography screening status(es), as follows:

$$\text{Logit}(\text{prob}(S_1 = 1)) = -4.8 + 2X_1 + 2X_2 - 4X_3 + 4X_4,$$

$$\text{Logit}(\text{prob}(S_2 = 1)) = -2.8 + X_1 + X_2 - 2X_3 + 2X_4 + 0.5S_1,$$

$$\text{Logit}(\text{prob}(S_j = 1)) = -2.8 + X_1 + X_2 - 2X_3 + 2X_4 + 0.5S_{j-1} + 0.5S_{j-2},\ j = 3, 4, 5.$$

Here $S_j = 1$ means a subject has filed a mammography screening claim at time j and $S_j = 0$ means otherwise. The coefficients are set at 0.5 for S_{j-1} and S_{j-2} to be close to the point estimates based on the linked NHIS-Medicare data, which are around 0.7 for different years. With these coefficients, participants who filed a mammography claim at one of the prior two time points are more likely to file a claim at the current time point.

The FFS/MA status at time j, P_j, relates to the current mammography screening status and previous FFS/MA status, as follows:

$$\text{Logit}(\text{prob}(P_1 = 1)) = -1.75 + 2X_1 + 2X_8 - 3X_9 - 4S_1,$$

$$\text{Logit}(\text{prob}(P_j = 1)) = \text{intercept} + 2X_1 + 2X_8 - 3X_9 - 4S_j + \beta P_{j-1},\ j = 2,\ldots,5.$$

The estimated coefficients for S_j based on the linked NHIS-Medicare data are around −3.5 for different years, so we set the coefficients for S_j at −4 to be close to the real data. In this and the following simulation study, we generated the mammography claim status to mimic the linked NHIS-Medicare data, where a small percentage of beneficiaries with MA coverage filed claims through Medicare. Actual mammography use for MA beneficiaries is not part of these set-ups. Because few beneficiaries in MA file mammography claims, participants with observed mammography claims are more likely to be in the FFS program.

From the linked NHIS-Medicare data, the coefficient estimates for P_{j-1} are around 7 for different years, which suggest FFS/MA enrollment at the current time is highly related to the FFS/MA enrollment of the previous time. We set values of the coefficient β for P_{j-1} at three different levels (0.5, 4, and 7) to represent low, medium, and high association of the current plan type with the previous plan type. We included low and medium levels of association in the simulation study to test whether the imputation results are similar to the case in which the association of the plan type over the years is high. We changed the values of the intercepts under different βs to have about 75 percent of participants in the FFS plan and about 25 percent of participants in MA plans at each year. Based on this setup, around 40 percent of participants in FFS and 3 percent of participants in MA file mammography claims through Medicare; these percentages are close to those observed in the linked NHIS-Medicare data.

For simulation 2, we generated FFS/MA plan type first and then generated screening status. We used similar parameters as in simulation 1; details of the simulation 2 setup can be found in section 2 of the supplementary material (please see supplementary data online).

For both simulations, we generated a linkage status (linked, not linked) for each participant. We let the linkage probability depend on X_4, X_5, X_6, and X_8. The logit of the linkage probability is $0.5 + X_4 - X_5 - X_6 + 2X_8$, with about 55 percent of participants linked. For those who are not linked, the mammography screening and plan statuses are set to missing. For both simulations, X_7 and X_{10} are treated as noise because they are unrelated to mammography screening, plan type, and the probability of linkage.

We conducted 10 imputations of the variables for each replicate, with X_1 to X_{10} and age included as covariates. Parameter estimates after multiple imputation were derived based on Rubin's combination rule (Rubin 1978, 1987, 1996).

4.2 Simulation 1: Results

In simulation 1, the screening-first imputation approach is consistent with the data-generating process and is expected to perform well. We calculated the mammography screening percentage at each time point among those with FFS coverage (supplementary table 2; please see the supplementary data online). We used before-deletion analysis, which analyzes the simulated data with all of the survey participants linked and thus no missing values, as a gold standard. After creation of missing values, in addition to conducting multiple imputation analyses, we analyzed the available cases without imputation, which included records without missing values at each time point, not considering whether the included records have missing information at other time points to preserve more data. The average absolute relative bias (AARB) (figure 1), the mean across years of the simulated percent absolute relative biases with respect to the before-deletion analysis, is calculated as follows:

$$\mathrm{AARB} = 100 \times \frac{\sum_{j=1}^{J} \left| M_j - BD_j \right| / BD_j}{J} \%,$$

where M_j is the average (over the 50 replicates) estimated percentage from a method at time j, BD_j is the average estimated percentage from before-deletion analysis at time j, and J is the total number of time points.

Compared with the before-deletion analysis, the available-case analysis yielded biased estimates for the percentage of FFS participants with mammography screening (e.g., AARBs were 13.11 percent, 12.92 percent, and 12.46 percent when $\beta = 0.5$, 4, and 7). In contrast, all three imputation methods yielded estimates with much smaller biases (figure 1). The screening-first method, the presumably correct imputation method, yielded small biases for low and medium associations among the FFS/MA plan type across years and slightly larger biases when the FFS/MA plan type association over years was high (e.g., AARBs were 0.11 percent, 0.24 percent, and 0.45 percent when $\beta = 0.5$, 4, and 7). These patterns were close to the results with the one-step approach (AARBs were 0.17 percent, 0.36 percent, and 0.43 percent when $\beta = 0.5$, 4, and 7). On the other hand, the plan-first approach yielded results with a pattern opposite that of the screening-first method. When the association between the FFS/MA plan type was low, the plan-first approach yielded estimates with larger biases (AARB was 0.50 percent when $\beta = 0.5$); however, when the association of the FFS/MA plan type across years increased, the biases of the plan-first method were smaller than those the other methods (AARBs were 0.09 percent and 0.07 percent when $\beta = 4$ and 7).

When the association of the FFS/MA plan type across years is low, the plan-first approach yields more biased results because the plan type is not correctly imputed without the screening information. However, when the association of the FFS/MA plan type across years is high, the plan type of a given year can be accurately predicted from the plan types of other years, even without the screening information included. In the second step, the imputation model for screening status includes all of the variables predicting the response variable and thus is a correctly specified model. As a result, we observed the smallest biases for the plan-

first approach when the association of the FFS/MA plan across years was high. Note that, although there appears to be differences in the performance of the three imputation approaches depending on the parameter values used in the simulation, all three approaches vastly outperformed the available-case analysis with regard to bias.

All three imputation methods yielded similar variances, which were larger than those of the before-deletion analysis (because the latter is based on there being no missing data) but close to those of the available-case analysis (supplementary table 2; please see the supplementary data online). All three imputation methods also yielded similar root mean squared errors, which were smaller than those of the available-case analysis.

4.3 Simulation 2: Results

For simulation 2, we generated data based on (2). We expected the plan-first approach to perform well because plan type was correctly imputed and then screening status was imputed conditional on the imputed plan type.

Again, the available-case analysis yielded estimates with large biases (AARBs were 7.01 percent, 7.02 percent, and 7.16 percent when $\beta = 0.5$, 4, and 7). The screening-first imputation method yielded estimates with relatively larger biases compared with the plan-first and the one-step imputation methods, and the biases increased when the association between FFS/MA plan type across years increased (AARBs were 0.45 percent, 0.45 percent, and 0.9 percent when $\beta = 0.5$, 4, and 7) (figure 2). The plan-first method yielded small biases (AARBs were 0.21 percent, 0.26 percent, and 0.27 percent when $\beta = 0.5$, 4, and 7), and similar results were observed for the one-step procedure (AARBs were 0.28 percent, 0.28 percent, and 0.33 percent when $\beta = 0.5$, 4, and 7). The screening-first method yielded larger biases than the other imputation methods under simulation 2 because the imputation model for the screening status was underfitted without one of the important predictors, the FFS/ MA plan type. The plan-first and the one-step approaches included all of the variables needed and thus yielded smaller biases. As with simulation 1, however, all of the imputation approaches were superior to the available-case analysis with respect to bias.

The variances from the three imputation procedures were similar and close to those of the available-case analysis (supplementary table 3; please see the supplementary data online). The root mean squared errors from the screening-first imputation were slightly larger than those of the plan-first and the one-step approaches, and all of them were smaller than the available-case analysis.

4.4 Summary of Simulation Studies

In summary, when the association of the FFS/MA plan type across years was low, the “correct” imputation methods (screening-first for simulation 1 and plan-first for simulation 2) yielded the estimates with the smallest AARBs, as expected. When the association of the FFS/MA plan type across years was high, the plan-first approach yielded smaller biases compared with the other two imputation methods. In both of these simulation studies, the one-step imputation method reduced the biases of the available-case analysis, and the results were close to the methods consistent with the data-generating process (screening-first for simulation 1 and plan-first for simulation 2).

5. RESULTS ON MAMMOGRAPHY SCREENING FORTHE LINKED NHIS AND MEDICARE DATA

Results of analyzing the NHIS-Medicare linked data are shown in figure 3. We generated 10 imputed datasets for each MI approach. For each imputed dataset, we derived the weighted mammography screening percentages among women with FFS coverage. The variances of the parameter estimates for each imputed dataset were computed based on Taylor linearization, which controls for the complex survey design, using the proc surveyfreq procedure in SAS version 9.3. The final estimates were derived using Rubin's combination rule.

Among those with FFS coverage, the percentages of women having mammography screening ranged from 39.68 percent to 42.03 percent across years based on the available-case analysis. These estimates were 1 percent to 2 percent higher than those of the MI methods (figure 3). The three MI methods yielded percentages close to each other (across years, screening-first: 38.61 percent–40.90 percent; plan-first: 38.26 percent–40.65 percent; one-step: 38.12 percent–40.86 percent), where the differences among them for any single year were within 1 percent. The estimated variances using the screening-first imputation method were smaller than those from the available-case analysis, whereas the plan-first and one-step imputation methods yielded larger variances than those for the available-case analysis for the years 1999–2001 but not for the later years.

To further examine the MI results, we used the self–reported mammogram information available from the 2005 NHIS for both linked and unlinked women in both FFS and MA Medicare programs. Although our analysis examined annual mammography, the NHIS question is, "Have you EVER had a mammogram?" This is a good indicator of overall screening status for the Medicare beneficiaries but does not identify recent screening. Based on this question, 91.34 percent (95 percent CI = 89.89 percent to 92.79 percent) of linked women answered that they had ever had a mammogram compared with 87.33 percent (95 percent CI = 85.72 percent to 88.94 percent) of unlinked women. These estimates suggest that the unlinked subjects may be less likely to obtain a mammogram, which is consistent with our results that the overall percentage of women with mammography screening is lower after inclusion of the unlinked respondents using MI.

6. COMPARISON OF THE OBSERVED DATA ANDTHE COMPLETED (OBSERVED + IMPUTED) DATA FOR THE SIMULATION STUDIES AND THE LINKED NHIS-MEDICARE DATA

Although the MAR assumption for MI is not testable, Abayomi, Gelman, and Levy (2008) suggested that for a specific imputation model fitted to the observed data, the observed data and the completed data (observed + imputed) could be compared to examine the plausibility of the imputed data. One way to compare the observed and the completed data for our study is to compare the longitudinal associations of FFS/MA coverage and mammography screening status among FFS participants across years. We calculated ϕ coefficients (Fleiss 1981), a measure of association for two binary variables, for the simulation studies (with β =

7, which best represents the linked NHIS-Medicare data) and the linked NHIS-Medicare data. Supplementary table 4 (please see the supplementary data online) contains ϕ coefficients for ranges of lag 1 to lag 3 and above for both the simulation studies and the linked NHIS-Medicare data. Figure 4 shows the distributions of ϕ coefficients (i.e., not separated by lags), using boxplots, for the linked NHIS-Medicare data.

In general, for the simulation studies, the ϕ coefficients for the available-case analysis (the observed data) were close to those for the before-deletion analysis because the missing data mechanism was MAR (supplementary table 4; please see the supplementary data online). The ϕ coefficients for the completed data from the three imputation methods were generally similar to those for the available-case analysis for both simulation studies except in the case of the screening-first method in simulation 2. In this scenario, the ϕ coefficients for the screening-first method were slightly lower for FFS/MA coverage and were slightly higher for mammography among FFS participants compared with the other methods.

The results for the NHIS-Medicare data were similar to those of simulation 2 (figure 4). For FFS/MA coverage, the quantiles of ϕ coefficients (minimum, 25 percent, median, 75 percent, and maximum) were lower for the screening-first method compared with those of the other methods. On the other hand, for mammography screening among FFS participants, the range of ϕ coefficients for the screening-first method was wider than those of the other methods, with a relatively larger maximum and a relatively smaller minimum, first quantile, and median. The ϕ coefficients for the plan-first and the one-step methods were similar to each other and close to those for the available-case analysis, suggesting these two methods preserved longitudinal association of plan type and mammography screening after imputation, which was expected under the MAR assumption.

Another way to compare the observed data and completed data is to study the conditional distribution of FFS/MA coverage and mammography given selected covariates (Abayomi et al. 2008). We compared two-way frequency tables of mammography screening by FFS/MA coverage for each year, conditioning on some key covariates such as race and marital status, for the linked NHIS-Medicare data. There were some differences between the observed cases and the completed cases (results not shown), but we did not find any extreme departures that would raise questions about the imputation results.

7. DISCUSSION

We explored three imputation procedures in this paper. The one-step approach is commonly used in other applications. The advantage of this procedure is that all variables except the variable to be imputed are used as predictors, and thus it is less likely to have an underfitted imputation model compared with the two factorization approaches; the drawback is that more variables are included in the model, and thus there is more “noise” in the imputation process, which may lead to larger variances. The screening-first and the plan-first procedures factorize the joint distribution and conduct imputation in two steps. Factorization of the joint distribution saves computation time, especially for large data, because fewer variables are included in the first step of imputation. In addition, using a two-step procedure may yield inferences with reduced variances (Kinney and Reiter 2009).

Moreover, Medicare data contain many health-related longitudinal variables, most often for the FFS beneficiaries, that can be used to study a variety of outcomes, such as heart disease (Chen, Normand, Wang, and Krumholz 2011), pancreatic cancer (Wang, Schrag, Brooks, and Dominici 2014), and hospital readmissions (Gerhardt, Yemane, Apostle, Oelschlaeger, Rollins et al. 2014), to list a few. When adding population survey data to the claims data, using the plan-first method would be especially useful for multiple studies with different outcome variables so that researchers could start with the same set of linked data. After imputation of plan type, different outcomes of interest can be imputed separately or together, as needed. By doing so, researchers would not need to re-impute plan type for different research topics and would have relatively consistent plan type information when comparing different health-related outcomes.

For the linked NHIS-Medicare data, all three imputation methods yielded lower percentages of mammography screening among women with FFS Medicare compared with the available-case analysis. Compared with the screening-first approach, the results of the plan-first and the one-step imputation methods were closer to each other. Because the association of the FFS/MA plan type across years is high (β is around 7) for the linked NHIS-Medicare and the plan-first and the one-step procedures were more robust to changes in the longitudinal associations (figure 4), we recommend using the plan-first and the one-step approaches for linked files subject to missing data from linkage eligibility and program characteristics and enrollment.

The high association of plan type across years for the linked NHIS-Medicare data could cause multicollinearity issues and unstable imputation model fits. To address this possible problem, we analyzed collinearity diagnostics for plan type. Belsley, Kuh, and Welsch (1980) suggested that, when the condition index number is around 10, weak dependencies might exist and start to affect regression estimates. When this number is greater than 100, estimates of regression models might have large numerical error. For the linked NHIS-Medicare data, the largest condition index number for plan type across years was 13.2, so we do not expect the multicollinearity of plan type to have a big impact on the imputation results for our study. In general, if collinearity of the covariates does affect the model fit significantly, we may need to remove some years of data from the imputation models to obtain more stable imputation results.

Missingness due to nonlinkage could lead to biased results for a linked data file. Other potential sources of bias include, but are not limited to, misspecification of the imputation model and/or missing data mechanism, measurement error of survey data and/or administrative records, linkage error, and so on. Sensitivity analysis is recommended for future research to study the impacts of these issues on statistical inference from linked population-based surveys and administrative data.

Record linkage refusal in the NHIS has changed over time, and these changes are described in Miller et al. (2011). In general, NHIS respondents were asked to provide their SSN for linkage to health-related information. Respondents who did not provide the SSN were considered to have refused record linkage. When changes in 2007 required only the last four digits of the SSN, the percentage refusing to provide the SSN for record linkage decreased.

In addition, since 2007, participants can agree to linkage without providing SSN by answering whether they permit to link their survey data with health-related records of other government agencies. This change may bring in new challenges for linking the NHIS data to administrative records; for example, a possible higher rate of missed matches when the survey participants cannot be uniquely linked to administrative records. To address these potential problems, the deterministic linkage method described earlier could be combined with a probabilistic linkage method. Probabilistic linkage methods are commonly used when different data sources do not have complete and accurate unique identifiers, but by comparing variables from both data sources, a linkage probability (a similarity measure) can be derived. To account for uncertainty due to the probabilistic linkage, Bayesian approaches have been developed in recent years (Wu 1995; Gutman, Afendulis, and Zaslavsky 2013; Sadinle 2014; Steorts 2015; Steorts, Hall, and Fienberg 2015). Moreover, methods for linking records across multiple files can further increase the utility of linked data files (Sadinle and Fienberg 2013; Steorts et al. 2015). On the other hand, computational inefficiency is an issue for large-scale probabilistic record linkage. To reduce computational burden, blocking methods have been used to reduce the number of comparisons (Steorts, Ventura, Sadinle, and Fienberg 2014; Miller, Betancourt, Zaidi, Wallach, Steorts 2015). In summary, probabilistic linkage could potentially improve future record linkage of NHIS data to administrative records; nevertheless, multiple imputation and/or other missing data methods will still be needed to address the structural missing data problems encountered due to linkage ineligibility and due to program characteristics, including longitudinal changes in program components and participation that are observed in the administrative data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Abayomi K, Gelman A, and Levy M (2008), “Diagnostics for Multivariate Imputations,” Journal of the Royal Statistical Society: Series C, 57, 273–291.

Andridge RH, and Little RJ, (2010), “A Review of Hot Deck Imputation for Survey Nonresponse,” International Statistical Review, 78(1), 40–64. [PubMed: 21743766]

Barnard J, and Rubin DB (1999), “Small-Sample Degrees of Freedom with Multiple Imputation,” Biometrika, 86, 948–955.

Belsley DA, Kuh E, and Welsch RE (1980), Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, New York: John Wiley and Sons.

Braithwaite D, Zhu W, Hubbard RA, O’Meara ES, Miglioretti DL, Geller B, Dittus K, Moore D, Wernli KJ, Mandelblatt J, and Kerlikowske K (2013), “Screening Outcomes in Older US Women Undergoing Multiple Mammograms in Community Practice: Does Interval, Age or Comorbidity Score Affect Tumor Characteristics or False Positive Rates?,” Journal of the National Cancer Institute, 105, 334–341. [PubMed: 23385442]

Cheema JR (2014), “A Review of Missing Data Handling Methods in Education Research,” Review of Educational Research, 84, 487–508.

Chen J, Normand S-L, Wang Y, and Krumholz HM (2011), “National and Regional Trends in Heart Failure Hospitalization and Mortality Rates for Medicare Beneficiaries, 1998–2008,” Journal of the American Medical Association, 306(15), 1669–1678. [PubMed: 22009099]

Christen P (2012), Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection, Berlin: Springer.

Fleiss JL (1981), Statistical Methods for Rates and Proportions (2nd ed.), New York: John Wiley & Sons.

Gerhardt G, Yemane A, Apostle K, Oelschlaeger A, Rollins E, and Brennan N (2014), "Evaluating Whether Changes in Utilization of Hospital Outpatient Services Contributed to Lower Medicare Readmission Rate," Medicare & Medicaid Research and Review, 23, 4.

Gutman R, Afendulis CC, and Zaslavsky AM (2013), "A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs," Journal of American Statistical Association, 108, 34–47.

Harron K, Goldstein H, and Dibben C (2015), Methodological Developments in Data Linkage, New York: John Wiley & Sons.

He Y (2010), "Missing Data Analysis Using Multiple Imputation: Getting to the Heart of the Matter," Circulation: Cardiovascular Quality and Outcomes, 3, 98–105. [PubMed: 20123676]

Horton NJ, and Kleinman KP (2007), "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models," American Statistician, 61, 79–90. [PubMed: 17401454]

Ibrahim J, and Molenberghs G (2009), "Missing Data Methods in Longitudinal Studies: A Review," TEST: *An Official Journal of the Spanish Society of Statistics and Operations Research*, 18, 1–43.

Kinney S, and Reiter J (2009), "Inferences for Two Stage Multiple Imputation for Nonresponse," Journal of Statistical Theory and Practice, 3, 307–318.

Little RJA (2008), "Selection and Pattern-Mixture Models," *in* Advances in Longitudinal Data Analysis, eds. Fitzmaurice G, Davidian M, Verbeke G, and Molenberghs G, pp. 409–431 (London: CRC Press).

Little RJA, and Rubin DB (2002), Statistical Analysis with Missing Data, New York: Wiley.

Miller DM, Gindi RM, and Parker JD (2011), "Trends in Record-Linkage Refusal Rates: Characteristics of National Health Interview Survey Participants Who Refuse Record-Linkage," Paper presented at the Joint Statistical Meetings, Miami Beach, FL.

Miller EA, Miller DM, Judson DH, He Y, Day HR, Zevallos K, Parker JD, MacKinnon JA, Hernandez MN, Wohler B, Sherman R, Fernandez CA, McClure LA, LeBlanc WG, Tannenbaum SL, Zheng DD, Lee DJ, and Christ SL. (2014), "Linkage of 1986–2009 National Health Interview Survey with 1981–2010 Florida Cancer Data System," Vital Health Statistics, 2.

Miller JW, Betancourt B, Zaidi A, Wallach H, and Steorts RC (2015), "Microclustering: When the Cluster Sizes Grow Sublinearly with the Size of the Data Set," Advances in Neural Information Processing Systems (NIPS), Bayesian Nonparametrics: The Next Generation Workshop.

Moriarity C, and Scheuren F (2001), "Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure," Journal of Official Statistics, 17, 407–422.

National Center for Health Statistics (2005), "2004 National Health Interview Survey (NHIS) Public Use Data Release: NHIS Survey Description."

National Center for Health Statistics (2006), "2005 National Health Interview Survey (NHIS) Public Use Data Release: NHIS Survey Description."

National Center for Health Statistics (2011), "Linkages between Survey Data from the National Center for Health Statistics and Medicare Program Data from the Centers for Medicare and Medicaid Services."

Raghunathan TE, Lebkowski JM, VanHoewyk J, and Solenberger P (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," Survey Methodology, 27, 85–95.

Rodgers WL (1984), "An Evaluation of Statistical Matching," Journal of Business and Economic Statistics, 2, 91–102.

Rubin DB (1978), "Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse," Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 20–34.

Rubin DB (1987), Multiple Imputation for Non-Response in Surveys, New York: John Wiley.

Rubin DB (1996), "Multiple Imputation after 18+ Years," Journal of American Statistical Association, 91, 473–489.

Rubin DB, and Schenker N (1986), "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse," Journal of the American Statistical Association, 81, 366–374.

Sadinle M (2014), "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach," Annals of Applied Statistics, 8, 2404–2434.

Sadinle M, and Fienberg S (2013), "A Generalized Fellegi-Sunter Framework for Multiple Record Linkage with Application to Homicide Record-Systems," Journal of the American Statistical Association, 108, 385–397.

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

Schenker N, Raghunathan TE, Chiu P-L, Makuc DM, Zhang G, and Cohen AJ (2006), "Multiple Imputation of Missing Income Data in the National Health Interview Survey," Journal of the American Statistical Association, 101, 924–933.

Simon AE, and Schoendorf KC (2014), "Medicaid Enrollment Gap Length and Number of Medicaid Enrollment Periods among US Children," American Journal of Public Health, 104, e55–e61.

Steorts RC (2015), "Entity Resolution with Empirically Motivated Priors," Bayesian Analysis, 10, 849–875.

Steorts RC, Ventura S, Sadinle M, and Fienberg S (2014), "A Comparison of Blocking Methods for Record Linkage," *in* Privacy in Statistical Databases, pp. 253–268, Springer.

Steorts RC, Hall R, and Fienberg SE (2015), "A Bayesian Approach to Graphical Record Linkage and De-duplication," Journal of the American Statistical Association, DOI: 10.1080/01621459.2015.1105807.

Townsend-Rocchiccioli J, and Steele S (2002), "Reimbursement for Screening Mammography- The Medicare Disparity: A Policy Perspective," Policy, Politics, & Nursing Practice, 3, 240–247.

Van Buuren S (2007), "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification," Statistical Methods in Medical Research, 16, 219–242. [PubMed: 17621469]

Van Buuren S (2012), Flexible Imputation of Missing Data, Boca Raton, FL: Chapman & Hall/CRC.

Van Buuren S, and Karin G (2011), "Mice: Multivariate Imputation by Chained Equations in R," Journal of Statistical Software, 45(3).

Wang Y, Schrag D, Brooks G, and Dominici F (2014), "National Trends in Pancreatic Cancer Outcomes and Pattern of Care among Medicare Beneficiaries, 2000 through 2010," Cancer, 120, 1050–1058. [PubMed: 24382787]

White IR, Royston P, and Wood AM (2011), "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice," Statistics in Medicine, 30(4), 377–399. [PubMed: 21225900]

Wu Y (1995), "Random Shuffling: A New Approach to Matching Problem," Proceedings of the Statistical Computing Section of the American Statistical Association, pp. 69–74.

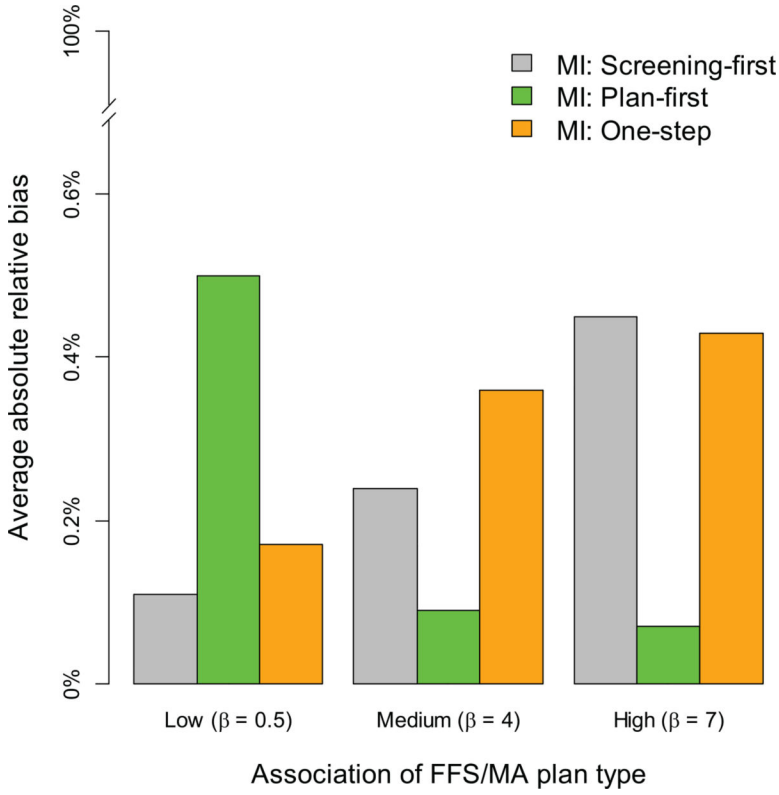

Figure 1. Average Absolute Relative Bias (AARB) of Estimated Percentages of FFS Medicare Participants with Mammography Screening by Association of FFS/MA Plan Type Across Years —Results of Simulation 1.

NOTE. The AARBs for the available-case analysis were 13.11 percent, 12.92 percent, and 12.46 percent when β=0.5, 4, and 7.

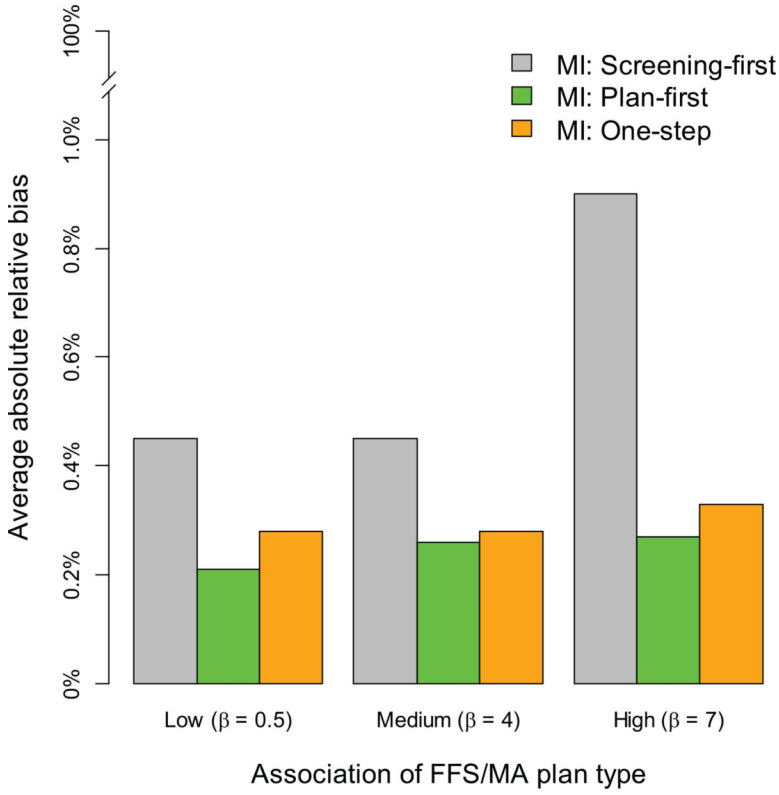

Figure 2. Average Absolute Relative Bias (AARB) of Estimated Percentages of FFS Medicare Participants with Mammography Screening by Association of FFS/MA Plan Type Across Years —Results of simulation 2.

NOTE. The AARBs for the available-case analysis were 7.01 percent, 7.02 percent, and 7.16 percent when $\beta = 0.5$, 4, and 7.

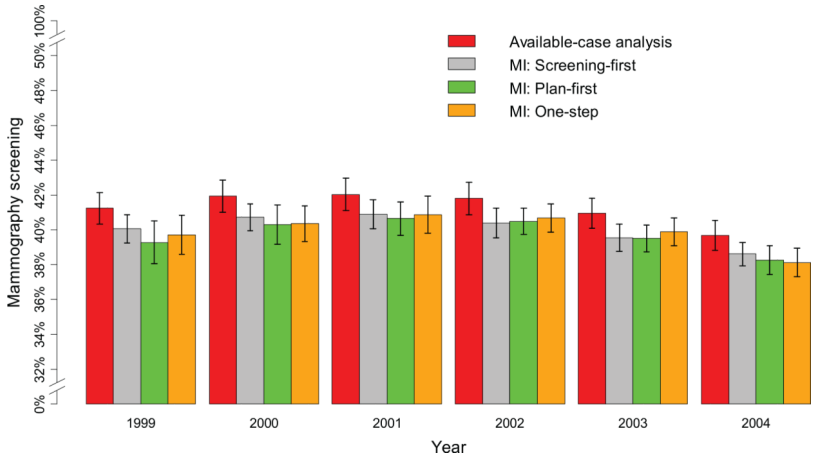

Figure 3. Estimated Percentages and Standard Errors of FFS Participants with Mammography Screening by Year, for Available-Case Analysis and Multiple Imputation Approaches.

2004–2005 NHIS Linked to 1999–2004 Medicare Denominator File.

NOTE. The error bars show one standard error above and one standard error below the estimated percentages.

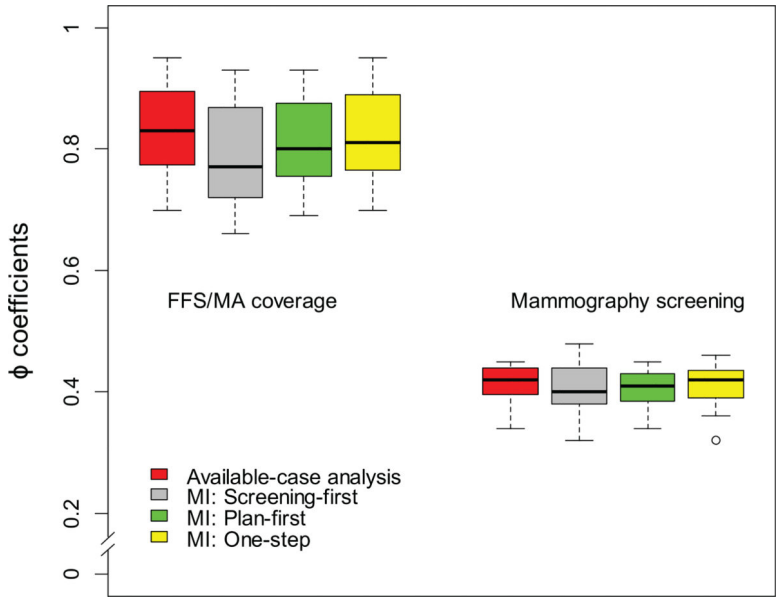

Figure 4.

Boxplots of ϕ Coefficients of FFS/MA Coverage and Mammography Screening Among FFS Participants—Results for the Linked NHIS-Medicare Data.