Published in final edited form as:

Stat Med. 2020 October 30; 39(24): 3299–3312. doi:10.1002/sim.8666.

Multinomial Logistic Regression with Missing Outcome Data: An Application to Cancer Subtypes

Ching-Yun Wang, Li Hsu

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, P.O. Box 19024, Seattle, WA 98109-1024, U.S.A.

SUMMARY

Many diseases such as cancer and heart diseases are heterogeneous and it is of great interest to study the disease risk specific to the subtypes in relation to genetic and environmental risk factors. However, due to logistic and cost reasons, the subtype information for the disease is missing for some subjects. In this paper, we investigate methods for multinomial logistic regression with missing outcome data, including a bootstrap hot deck multiple imputation (BHMI), simple inverse probability weighted (SIPW), augmented inverse probability weighted (AIPW), and expected estimating equation (EEE) estimators. These methods are important approaches for missing data regression. The BHMI modifies the standard hot deck multiple imputation method such that it can provide valid confidence interval estimation. Under the situation when the covariates are discrete, the SIPW, AIPW and EEE estimators are numerically identical. When the covariates are continuous, nonparametric smoothers can be applied to estimate the selection probabilities and the estimating scores. These methods perform similarly. Extensive simulations show that all of these methods yield unbiased estimators while the complete-case analysis can be biased if the missingness depends on the observed data. Our simulations also demonstrate that these methods can gain substantial efficiency compared to the complete-case analysis. The methods are applied to a colorectal cancer study in which cancer subtype data are missing among some study individuals.

Keywords

Hot deck multiple imputation; Inverse probability weighting; Missing at random

1 Introduction

In biomedical studies, estimation of exposure or genetic effects on disease incidence is a frequent objective. However, due to practical consideration, some variables may not be available for all subjects from the study cohort. Incomplete data can be due to various reasons such as noncompliance with data acquisition requests or by design to reduce costs. Regression analysis may encounter a challenge due to missing data in some individuals of the study cohort. Methodology for missing data is often required to address these challenges.

We consider an example of missing outcome regression. In a case-control study of colorectal cancer, there is an interest in studying whether the effects of genetic or environmental risk factors on colorectal cancer differ by microsatellite instability (MSI) status in tumor.¹ This is because it is known that MSI is associated with survival outcomes and treatment response.^{2,3}

However, due to logistical and cost reasons, not all tumors have the information on MSI. The problem is multinomial logistic regression with missing outcomes. To address this problem, a simple approach is the complete-case (CC) analysis, which is to perform the usual analysis using the observations in a subsample with complete data, namely the *complete-case set*. However, this approach can give rise to bias and reduced statistical efficiency. In addition to the CC estimator, there is a rich body of literature on missing data regression; see Little and Rubin for a review.⁴ By using the terminology of Rubin, data are missing completely at random (MCAR) when the missing data process is independent of both the outcomes and covariates, and data are missing at random (MAR) when the missing data process is independent of the missing data given the observed covariates (and auxiliary variables if available).⁵ In contrast, a nonignorable missing data process may depend on the unobserved data given the observed data. Imputation and weighting are 2 important approaches in dealing with missing data problems.^{6,7} Wang et al. showed that when the covariates are MAR, under many situations inverse selection probability weighted estimators are numerically identical to imputation of estimating scores.⁸ The selection probability in the missing data context is defined as the probability of an individual to be in the complete-case set. The idea of the simple inverse probability weighted (SIPW) estimator is to adjust the CC analysis by using the inverse of the selection probability as the weight. We also investigate an augmented inverse probability weighted (AIPW) estimator. A general imputation approach is multiple imputation (MI) which imputes multiple sets of data for the missing variables.⁹ The multiple imputed data sets are then analyzed by using some standard procedures for the complete data and combining the results from these analyses to adjust for the uncertainty due to the missing data. MI has been widely applied in the research; for example, Wu and Wu (2002) proposed a three-step hierarchical multiple-imputation method, implemented by Gibbs sampling, for viral dynamic models in acquired immune deficiency syndrome research.¹⁰ In this paper, we investigate a bootstrap hot deck multiple imputation (BHMI) when the covariates are discrete, and the BHMI is modified to impute data from the nearest neighbor when the covariates are continuous. In addition to imputation and weighting, we extend the idea of expected estimating equation (EEE) which is the same as the maximum likelihood (ML) estimator when the covariates are discrete.⁸

There are other MI approaches for missing outcome regression. For example, an alternative to the standard implementation of MI for handling missing data in both outcome and exposure variables is the “multiple imputation, then deletion” approach (MID), where observations with imputed outcomes are excluded from the analysis.¹¹ However, MID is not advisable when the imputation model contains auxiliary variables for the outcome.¹² Another consideration is that in longitudinal studies in which patients may drop out due to many practical reasons, there could be only a small group of individuals who may have complete data.¹³ Hence, under longitudinal data, MI is usually a more efficient approach than the SIPW approach because MI may use the observed longitudinal data in the imputation process while SIPW does not. For longitudinal outcomes with missing data, Bayesian or non-Bayesian likelihood modeling and the use of sensitivity analysis have been widely applied to the analysis of missing longitudinal data.¹⁴ However, these methods often make use of the association of the longitudinal outcomes. In this paper we focus on missing outcomes in multinomial regression and there is no longitudinal data to establish an

additional association, or sensitivity analysis. Hence, the Bayesian or non-Bayesian modeling for missing longitudinal data problems may not be applicable to the regression setting in this paper.

In this paper, we investigate methods for multinomial logistic regression when the outcomes are MAR. Motivated by the colorectal cancer example mentioned above, our methods would treat the binary cancer indicator as an auxiliary variable for the multinomial outcome variable. We study the BHMI, SIPW, AIPW, and EEE estimators. We show that they are valid, SIPW, AIPW and EEE are numerically identical when the covariates are discrete, and they have similar performance when the covariates are continuous. Our simulations demonstrate that the four methods can gain efficiency by using auxiliary information compared to the CC analysis. Especially, the bias reduction can be substantial under many practical situations. In terms of computation, the BHMI and SIPW estimator are in general computationally friendly. The paper is outlined as the following. In Section 2, we describe the BHMI estimator and in Section 3, we review the SIPW and AIPW estimators. The EEE estimator is investigated in Section 4. In Section 5, the results from our simulation studies are presented. Some finite sample issues are noted. In Section 6, we apply the proposed methods to the colorectal cancer study described above. Some concluding remarks are given in Section 7. Sketched proofs are provided in the Appendix.

2 Regression Model and Bootstrap Hot deck MI

Assume that the total sample size of the study cohort is n . The regression model of interest in this paper is multinomial logistic regression. Let Y_i be the response variable, X_i be the vector of covariates for individual i , $i = 1, \dots, n$. In the colorectal cancer example described in the introduction, $Y_i = 0, 1$, or 2 for control, MSI stable, and MSI unstable, respectively. We assume that the vector of covariates X_i is discrete that may include e.g., BMI (obese, overweight, normal, and underweight) and gender (men and women) among other risk factors. In many applications, there exists an auxiliary variable, denoted by S_i , for the outcome variable. For example, in the aforementioned colorectal cancer study, the binary case-control status could serve as an auxiliary variable. That is, $S_i = 0$ if $Y_i = 0$, and $S_i = 1$ if $Y_i = 1$ or $Y_i = 2$. Assume that Y_i has $k + 1$ possible categories $Y_i = 0, 1, \dots, k$. The main interest is to estimate the vector of regression coefficients β in the following regression model

$$\text{pr}(Y_i = j | X_i) = \frac{\exp(\beta_{0,j} + \beta'_{1,j} X_i)}{1 + \sum_{r=1}^k \exp(\beta_{0,r} + \beta'_{1,r} X_i)}$$

for $j = 1, \dots, k$. Let R_i indicate whether Y_i is observed ($R_i = 1$) or not ($R_i = 0$). The complete-case data set ($R_i = 1$) consists of (Y_i, X_i, S_i) , and the non-complete-case data set ($R_i = 0$) consists of (S_i, X_i) . The selection probability π_i ($i = 1, \dots, n$) is defined as $\text{pr}(R_i = 1 | Y_i, X_i, S_i)$. When Y_i is MCAR, π_i is a constant, while Y_i is MAR if π_i depends on only the observed data (X_i, S_i) . If Y_i is nonignorable missing, then π_i may depend on Y_i and the observed data. In this paper, our methodology development is under the assumption that Y is MAR.

To adjust for bias and increase efficiency due to missing Y for some individuals, a popular approach is hot deck multiple imputation (HMI). Briefly speaking, HMI replaces a missing value with an observed outcome from an individual with the same characteristics.^{4,15} In our problem, the characteristics are determined by X_i and S_i . The HMI can be implemented by creating M (say 10) complete datasets by the following procedure: (i) For the m th imputed set ($m = 1, \dots, M$), $Y_{i,m}$ is randomly drawn among the individuals $j = 1, \dots, n$ with $X_j = X_i$ and $S_j = S_i$. That is, the imputed $Y_{i,m}$ is drawn among the individuals with the same characteristics. (ii) Create M imputed datasets. (iii) Let $\hat{\beta}^{(m)}$ be the estimate from the m th imputed set, $m = 1, \dots, M$, with V_m the variance of $\hat{\beta}^{(m)}$. (iv) Set $\hat{\beta}_{MI} = \sum_{m=1}^M \hat{\beta}^{(m)} / M$, with variance $\bar{V} + (1 + M^{-1})B$, where $\bar{V} = \sum_{m=1}^M V_m / M$, $B = \sum_{m=1}^M (\hat{\beta}^{(m)} - \hat{\beta}_{MI})(\hat{\beta}^{(m)} - \hat{\beta}_{MI})' / (M - 1)$. However, HMI is a simple random imputation which does not take into account the conditional posterior distribution of the parameters given the observed data.¹⁶ Hence it would generally produce confidence intervals of β that are too narrow. Rubin and Schenker proposed a Bayesian bootstrap imputation and an approximate Bayesian bootstrap imputation to address the under-estimation problem.¹⁶ Our idea to correct for interval estimation is similar to the bootstrap method described in Little and Rubin (Chapter 5); we would modify the HMI by resampling the observed data using bootstrap before performing step (i) of HMI.⁴ The BHMI follows the following steps:

- i. For $m = 1, \dots, M$, we resample the observed data with replacement before performing imputation. Let the variables of the bootstrap sample be denoted by $(Y_i^*, X_i^*, S_i^*, R_i^*)$, in which Y_i^* is missing if $R_i^* = 0$.
- ii. For the m th bootstrap sample, we perform imputation by randomly drawing $Y_{i,m}^*$ if i th subject is missing, i.e., $R_i^* = 0$, among the individuals $j = 1, \dots, n$ with $X_j^* = X_i^*$, $S_j^* = S_i^*$ and $R_j^* = 1$.
- iii. Let $\hat{\beta}^{(m)}$ be the estimate from the m th bootstrap imputed dataset, with V_m the variance of $\hat{\beta}^{(m)}$.
- iv. Set $\hat{\beta}_{BHMI} = \sum_{m=1}^M \hat{\beta}^{(m)} / M$, with variance $V_{BHMI} = \sum_{m=1}^M (\hat{\beta}^{(m)} - \hat{\beta}_{BHMI})(\hat{\beta}^{(m)} - \hat{\beta}_{BHMI})' / (M - 1)$.

Briefly speaking, BHMI takes a bootstrap sample with replacement of the observed data in step (i) of HMI prior to the m th imputation. Then the variance of the BHMI estimator is the sample variance of the M regression coefficient estimates, rather than the conventional MI variance estimator that sums the average of the M variances and the between-sample variance. In our simulation study, the coverage probabilities based on $M = 30$ have satisfactory performance. The number of bootstrap imputed data sets ($M = 30$) needed for BHMI variance estimator to be stable is in general slightly larger than the number of imputations (say, $M = 10$) for most MI procedures. An important advantage of BHMI is that it can provide valid confidence interval estimation at the expense of slightly more computation time, while the SEs from the typical formula of a simple HMI would likely be too small. An additional note is that if we apply the typical variance for MI to the BHMI,

then the confidence interval estimation would be too wide. The reason for the difference is due to the bootstrap resampling in BHMI.

When the covariate vector contains a component that is continuous, the BHMI needs to be modified to impute data from the nearest neighbor, as in Chapter 4 of Little and Rubin.⁴ For simplicity in presenting the idea, we consider the situation when X is a scalar in this paragraph. The main question is how to impute a missing Y_i based on the observed S_j and X_j . Step (ii) of the BHMI algorithm would be modified to (ii): For the m th bootstrap sample, we perform imputation by randomly drawing $Y_{i,m}^*$ if i th subject is missing, i.e., $R_i^* = 0$, among the individuals $j = 1, \dots, n$ with $S_j^* = S_i^*$, $R_j^* = 1$, and that $|X_j^* - X_i^*| < h$, where h is the width of the nearest neighbor. Bias will occur if the width is too wide, but large uncertainty may occur if the width is too narrow. We are not aware of theory about the width of the nearest neighbor. To address this issue, we consider the idea of bandwidth selection when estimating the selection probability by a nonparametric smoother, and suggest that we use the rate of $n^{-1/3}$ for the nearest neighbor.¹⁷ Additional discussion on the bandwidth selection for estimating the selection probability nonparametrically is provided in the next section. In our simulation study, we use the bandwidth of $h = 2\sigma_{X,S,C} n_{S,C}^{-1/3}$, in which $\sigma_{X,S,C}$ is the standard deviation of X given $S = s$ in the CC set, and $n_{S,C}$ is the sample size of the subcohort. If in case for a specific X there is no data in the nearest neighbor with this h , then we double the bandwidth until there are data in the neighbor to be included in the imputation procedure.

3 Inverse Probability Weighted Method

Another useful approach for missing data is an SIPW estimator, which takes the inverse of the selection probabilities as weights. The SIPW estimator under the regression model is to solve the following weighted estimating equation:

$$\sum_{i=1}^n \frac{R_i}{\pi_i} \phi(Y_i, X_i, \beta) = 0, \quad (1)$$

where $\phi(Y_i, X_i, \beta)$ is the estimating equation for β when $Y_i, i = 1, \dots, n$ are all observed. The selection probabilities $\pi_i, i = 1, \dots, n$, may be known in some situations (such as a two-stage design study), but this may not be the case in general. In addition, it is usually more efficient to use estimated π values even if the true π 's are known.¹⁷ Because X_j and S_j are discrete, π_j can be consistently estimated by a nonparametric estimator:

$$\hat{\pi}(X_i, S_i) = \frac{\sum_{j=1}^n R_j I[X_j = X_i, S_j = S_i]}{\sum_{j=1}^n I[X_j = X_i, S_j = S_i]}, \quad (2)$$

where $I(\cdot)$ is an indicator function. The SIPW estimator applies the inverse probability weights to the estimating function for subjects from the complete-case set. However, it does not directly apply the expected estimating function for subjects from the non-complete-case set. Therefore, it is natural to include an augmented term and the augmented inverse probability weighted (AIPW) estimator is to solve the following estimating equation:

$$\sum_{i=1}^n \left[\frac{R_i}{\pi_i} \phi(Y_i, X_i, \beta) + \{1 - (R_i/\pi_i)\} E\{\phi(Y_i, X_i, \beta) | S_i, X_i\} \right] = 0, \quad (3)$$

When the covariates are discrete, the estimating score in the augmentation term can be consistently estimated by

$$\hat{E}\{\phi(Y_i, X_i, \beta) | X_i, S_i\} = \frac{\sum_{j=1}^n R_j \phi(Y_j, X_j, \beta) I[X_j = X_i, S_j = S_i]}{\sum_{j=1}^n R_j I[X_j = X_i, S_j = S_i]}. \quad (4)$$

The AIPW estimator has a double robustness property in the sense that it only requires the correct specification of either the selection probability or the conditional distribution of the missing data given the observed data.^{6,17} However, when the covariates are discrete both the selection probabilities and the augmentation term can be estimated by their empirical averages and hence there is no issue of misspecification. Wang et al. showed that under this situation the estimating equation with the augmented term is the same as without the augmented term.⁸

Estimation of the standard errors of the SIPW estimator for β would require additional efforts due to the estimation of the selection probabilities. The asymptotic variance can be obtained by a sandwich variance estimator where the vector of the estimating equations is obtained by stacking the estimating equations for β and the nuisance parameters involved in the selection probabilities. However, if there are many covariates in the modeling of the selection probabilities, then it will be computationally easier to use bootstrap variance estimation to obtain the standard errors.

When the covariate vector contains a component that is continuous, the selection probability estimator in (2) will need to be modified. For simplicity of presentation, we consider the situation when X is a scalar in this paragraph. The Nadaraya-Watson kernel smoother for the selection probability can be presented as:

$$\hat{\pi}(X_i, S_i) = \frac{\sum_{j=1}^n R_j I[S_j = S_i] K_h(X_i - X_j)}{\sum_{j=1}^n I[S_j = S_i] K_h(X_i - X_j)},$$

where $K_h(\cdot) = K(\cdot/h)$, h is the bandwidth and $K(\cdot)$ is a kernel function such that $\int K(u) du = 1$ and $\int u K(u) du = 0$. In our simulation study, we use the bandwidth of $h = 4\sigma_{X,S} n_s^{-1/3}$, in which $\sigma_{X,S}$ is the standard deviation of X given $S = s$, and n_s is the sample size of the subcohort with $S = s$.¹⁷ If in case for a specific X there is no data within the bandwidth, then we double the bandwidth until there are data within the bandwidth in the selection probability smoother above. The estimating score in the augmentation term can be consistently estimated by

$$\hat{E}\{\phi(Y_i, X_i, \beta) | X_i, S_i\} = \frac{\sum_{j=1}^n R_j \phi(Y_j, X_j, \beta) I[S_j = S_i] K_h(X_i - X_j)}{\sum_{j=1}^n R_j I[S_j = S_i] K_h(X_i - X_j)} \quad (5)$$

In (5), in our simulation we use bandwidth $h = 4\sigma_{x,s,c}n_{s,c}^{-1/3}$, based on the CC set with the same S . If there are more than 2 components of $\mathbf{X}$ that are continuous, then there may be curse of dimensionality when directly applying kernel smoothers. In this case, a parametric selection probability model is likely more practical.

4 Expected Estimating Equation Estimator

Maximum likelihood (ML) is an alternative approach when the outcome is MAR; see for example, Magder and Hughes for logistic regression with misclassification in the outcomes.¹⁸ We now take a different view point linking the ML estimation and the conditional expectation of the *full data estimating equation*, namely the estimating equation when there is no missing data. Let $\mathcal{L}(Y_i|\mathbf{X}_i)$ be the likelihood function of Y_i given $\mathbf{X}_i$. If all the Y_i , $i = 1, \dots, n$ are available, then the estimating equation for $\boldsymbol{\beta}$ can be expressed as $\sum_{i=1}^n \phi(Y_i, \mathbf{X}_i, \boldsymbol{\beta}) = 0$, in which $\phi(Y_i, \mathbf{X}_i, \boldsymbol{\beta})$ is the derivative of $\log\{\mathcal{L}(Y_i|\mathbf{X}_i)\}$ with respect to $\boldsymbol{\beta}$, that is the full data estimating equation. Because the true Y_i is not observed in the non-complete-case set, the full data estimating equation can not be directly applied to the data. In the non-complete-case set, the estimating score will be from the likelihood of S_i given $\mathbf{X}_i$, denoted by $\mathcal{L}(S_i|\mathbf{X}_i)$. By some calculations (Appendix A), it can be shown that the estimating score for an individual in the non-complete-case set can be expressed as $E\{\phi(Y_i, \mathbf{X}_i, \boldsymbol{\beta})|\mathbf{X}_i, S_i\}$. That is, the observed data estimating score is the conditional expectation of the full data estimating score given the observed data. This approach for the ML estimation follows the idea of expected estimating equations.¹⁹ Therefore, the ML estimator can be obtained by solving the following estimating equation:

$$\sum_{i=1}^n R_i \phi(Y_i, \mathbf{X}_i, \boldsymbol{\beta}) + (1 - R_i) E\left\{ \phi(Y_i, \mathbf{X}_i, \boldsymbol{\beta}) \middle| \mathbf{X}_i, S_i \right\} = 0. \quad (6)$$

When $\mathbf{X}_i$ and S_i are discrete, the estimating score with missing Y_i can be consistently estimated by (4) given in the previous section. The EEE estimator is the same as the ML estimator when data are discrete, and the asymptotic variance of the EEE estimator for $\boldsymbol{\beta}$ can be obtained by a sandwich variance estimator where the vector of the estimating equations is obtained by stacking the estimating equations for $\boldsymbol{\beta}$ and the nuisance parameters involved in the conditional distribution of Y_i given $(\mathbf{X}_i, S_i)$. However, if there are quite a few covariates in the regression model, then bootstrap variance estimation will be more computationally practical for the standard errors.

When some covariates are continuous, the estimating score in the augmentation term can be consistently estimated by nonparametrical smoother (5). Estimating the conditional expectation of the full data score given the observed data nonparametrically could avoid potential misspecification of the distribution of the covariates. The bandwidth selection of the nonparametric smoother is similar to the approach that we discussed in the last section. In our simulation study for continuous covariates, the SE estimates for EEE are from bootstrap resampling of 50 replicates.

5 Simulation Study

We conducted a simulation study for multinomial logistic regression when the outcome variables (e.g., subtypes for diseased) were missing among some individuals, while an auxiliary variable (e.g., disease status) for the outcome variable was available for all individuals of the study cohort. In Tables 1 and 2, we study the situation when the disease status $S = \mathbb{I}[Y > 0]$ is available for all the study individuals. This is similar to the colorectal cancer study described earlier when some subjects may be missing information on tumor subtypes. We first generated binary X such that 50% individuals had $X = 0$ and 50% had $X = 1$. We assume that the log(odds) of $Y = 1$ versus $Y = 0$ is $\beta_{0,1} + \beta_{1,1}X$, and the log(odds) of $Y = 2$ versus $Y = 0$ is $\beta_{0,2} + \beta_{1,2}X$. The parameters of the regression model are provided in the tables. We compare the CC analysis, BHMI, SIPW, AIPW, and EEE estimators. In the tables, “bias” is calculated by taking the average of $\hat{\beta} - \beta$, “SD” denotes the sample standard deviation of the estimates, “ASE” denotes the average of the estimated standard errors of the estimates over the replicates. We also calculated the 95% confidence interval coverage probabilities (CP). To implement the SIPW estimator, we modeled the selection probabilities as a logistic function of the discrete covariates X and auxiliary outcome S . For the EEE estimator, we solved estimating equation (6) but with the conditional score given the observed data being estimated by (4). In Tables 1–3, the standard errors (SEs) of the SIPW and EEE estimates are obtained from sandwich variance estimates, while the BHMI estimates are from bootstrap. In Table 4 for continuous covariates, the SEs of the SIPW, AIPW and EEE estimates are obtained from bootstrap. In Table 1, the CC set is a random sample, and the sample size is 50% of the whole cohort. In addition to the regression parameters, we also study estimation of the difference between the covariate effects on $Y = 2$ versus $Y = 1$ (for example, MSI stable vs. MSI unstable); $\beta_{1,2} - \beta_{1,1}$. The BHMI, SIPW, AIPW and EEE estimators are more efficient than the CC estimator for the regression parameters. The efficiency gain is due to the use of the auxiliary variable S in BHMI, SIPW, AIPW and EEE. The BHMI estimator is slightly less efficient than the SIPW, AIPW, and EEE estimators. We also note that the SIPW, AIPW and EEE estimates are numerically identical. To understand this finding, in Appendix B we provide a brief proof to show that the SIPW, AIPW and EEE estimators are numerically identical when the covariates are discrete. However, for the estimation of the difference in odds ratio between $Y = 2$ and $Y = 1$ ($\beta_{1,2} - \beta_{1,1}$), there is no efficiency gain from BHMI, SIPW, AIPW or EEE over CC. The main reason for no improvement is because the auxiliary variable S does not provide any information to distinguish between $Y = 2$ and $Y = 1$.

In Table 2, we also investigate a scenario similar to Table 1, but the selection probability is $\{1 + \exp(X + S)\}^{-1}$; under the selection model the complete-case sample size is about 28% of the size of the whole cohort. This is a situation when Y is missing at random. One major difference between the results of Table 1 and Table 2 is that the CC analysis is consistent in the former, but not in the latter. The BHMI, SIPW, AIPW, and EEE estimators are all valid in estimating the regression parameters. It is worth noting that even though the CC analysis is not valid for the regression coefficients, but is valid for the estimation of $\beta_{1,2} - \beta_{1,1}$. This is because the biased sampling causes the same bias in both slopes $\hat{\beta}_{1,2}$ and $\hat{\beta}_{1,1}$ and the difference essentially cancels the biases.

The auxiliary variable for the outcome variable in Tables 1 and 2 is the disease status $S = \mathbb{I}[Y > 0]$, which is available for all the study cohort. In Table 3, we investigated a setting that was almost the same as Table 2 except that 10% of disease outcomes may be misclassified as disease-free, such that $S = \mathbb{I}[Y > 0] \times \mathbb{I}[U > 0.1]$ is available where U is an independent uniform $[0,1]$ variable. Based on the selection probability model $\{1 + \exp(X + S)\}^{-1}$, the CC sample size is about 29% of the size of the whole cohort. This setting is like an example in a tumor study, an actual case may be missed by a screening test. The simulation result of Table 3 is similar to that of Table 2. The missing data mechanism in this table is still MAR, and hence the BHMI, SIPW, AIPW and EEE estimators are still unbiased for the regression parameters. The BHMI estimator is slightly less efficient than the SIPW, AIPW, and EEE estimators, and the SIPW, AIPW, and EEE estimates are numerically identical.

In Table 4, we study the performance of the methods when X is continuous. We generated X from a uniform distribution with mean 0 and variance 1, and the outcomes were generated similarly to those in Tables 1–3. The auxiliary variable $S = \mathbb{I}[Y > 0] \times \mathbb{I}[U > 0.3] + \mathbb{I}[Y = 0] \times \mathbb{I}[U > 0.8]$ is available for all individuals in which U is from an independent uniform $[0,1]$ distribution. This is like a situation when a control could be misclassified as a case based on the S , while a case could be misclassified as a control. Based on the selection probability model $\{1 + \exp(X + S)\}^{-1}$, the CC sample size is about 42% of the size of the whole cohort. Except for CC, the SE estimates in this table for continuous X are based on 50 bootstrap samples. The BHMI, SIPW, EEE and AIPW estimates have similar performance in most cases. They are all unbiased while CC is substantially biased. The AIPW estimator has about the same efficiency as the SIPW estimator. This finding is similar to a missing covariate problem in which Wang and Wang showed that for continuous covariates and auxiliary variables, the SIPW, AIPW and EEE are asymptotically equivalent.²⁰ Hence, from this table, the finding suggests that for continuous X there is no need to calculate a more complicated AIPW estimator since SIPW has the same performance. The EEE estimator appears to be slightly better than SIPW or AIPW. The reason for the finite sample performance could be likely because the selection probability could be very close to 0 with a large X . When the sample size is large enough, the differences between BHMI, SIPW, AIPW and EEE essentially disappear. Nevertheless, this is based on nonparametric estimation of the selection probability and the conditional score. If the selection probability and the conditional score are estimated parametrically, the AIPW estimator would have the double robustness property while SIPW or EEE would not.

6 DATA ANALYSIS

We applied the methods to a case-control study of colorectal cancer as described in Introduction.¹ The study includes 1,794 cases and 2,684 controls drawn from the Colon Cancer Family Registry including 6 recruitment centers. Approximately 10% - 20% of colorectal tumors display MSI, defined as the expansion or contraction of small repeated sequences in the DNA of tumor tissue relative to nearby normal tissue. The goal of the analysis is to evaluate the association between overweight or obesity and colorectal cancer risk by tumor MSI status. We created a subset of data based on Table 3 of Campbell et al., including individuals with microsatellite stable (MSS), or with high microsatellite instability (MSI-H), but excluding individuals with low microsatellite instability.¹ The data set includes

1,053 cases and 1,595 controls. Among 1,053 the cases, 871 cases are MSS, and 182 are MSI-H. The body mass index (BMI) is categorized into two groups: normal (BMI between 18.5 and 24.99) and overweight/obese (BMI larger than 25.00). To illustrate the proposed methods, we conducted a sensitivity analysis to randomly select about 33% of individuals as the CC subset ($n = 866$). We do not intend to interpret our findings as the Colon Cancer Family Registry results. In this data application, the response variable is the colorectal cancer status, with $Y = 0$ for control, $Y = 1$ for case with MSS, and $Y = 2$ for case with MSI-H. Covariate $X = 0$ for normal BMI, while $X = 1$ for overweight or obese individuals. The auxiliary variable $S_i = 0$ if $Y_i = 0$, and $S_i = 1$ if $Y_i = 1$ or $Y_i = 2$.

The analysis results are given in Table 5. Based on the CC analysis, perhaps due to the limited sample size, the effect of overweight/obesity on either MSS or MSI-H tumors is not significant at level 0.05. Using our proposed methods BHMI, SIPW, AIPW or EEE, we are able to establish the association between overweight/obesity and MSS colorectal cancer, while the CC estimator could not. There is no association between overweight/obesity and MSI-H colorectal cancer by any of the methods given above. These findings are consistent with those in Campbell et al. (2010).¹ They also cautioned that one possible reason for not being able to establish the association of overweight/obesity and MSI-H colorectal cancer was due to insufficient sample size in MSI-H. For the comparison of the effects of overweight/obesity between MSS and MSI-H tumors ($\beta_{1,2} - \beta_{1,1}$), there is no difference, which is likely due to the limited sample size in this data analysis.

7 Discussion

There are advantages and disadvantages between CC and MI in general, and the comparisons can be extended to SIPW, AIPW and EEE as well. White and Carlin compared CC and MI in terms of bias and efficiency under a variety of missing data mechanisms.²¹ They remarked that for missing outcome regression, CC and MI are essentially the same or approximately the same. Under this situation, the parameters to be used in the imputation model are essentially the same as the parameters estimated from the CC analysis. It is no longer the case if there is an auxiliary variable (S) for the outcome. With the additional information from S , the imputation model is conditional on X and S , which is different from the CC model that is conditional on X only. Therefore, the BHMI, SIPW, AIPW, and EEE estimators can improve over the CC estimator. Especially, the bias reduction can be substantial under many practical situations. For discrete covariates, SIPW, EEE and AIPW estimates are identical and BHMI estimates are very close to them. For continuous covariates, BHMI, SIPW, AIPW and EEE estimates are very close from our simulations. In terms of computation, the BHMI and SIPW estimators can be implemented with less efforts. It will be more practical to use bootstrap than the sandwich estimator to estimate the SEs of both methods in real data analysis. From our simulation experience, 50 bootstrap replicates are in general enough.

In Tables 1 and 2 of our simulation study, the disease status $S = \mathbb{I}[Y > 0]$ is available for all study individuals, while the multinomial $Y(0, 1, 2)$ was available only in a sub-sample. We found that the BHMI, SIPW, AIPW and EEE estimators were more efficient than the CC estimator for the regression parameters, but for the estimation of the difference of log-odds

ratios between $Y = 2$ and $Y = 1$ ($\beta_{1,2} - \beta_{1,1}$), there is no improvement. This is because that the auxiliary variable (disease status $S = \mathbb{I}[Y > 0]$) does not provide any information to distinguish between $Y = 2$ and $Y = 1$. To gain efficiency in comparing the covariate effects between $Y = 2$ and $Y = 1$ ($\beta_{1,2} - \beta_{1,1}$), our study findings indicate that a strategy is to find an auxiliary variable S that is associated with the outcome variable Y , such as the mean value of S for $Y = 1$ is different from the mean value of S for $Y = 2$.

In this paper, the methods are developed under multivariate logistic regression. When the covariates are discrete, we show that the the SIPW/AIPW estimator and the EEE estimator are numerically identical. When the covariates are continuous, our simulation results show that the BHMI, SIPW, AIPW and EEE estimators have very close performance. We note that there are some other types of MI procedures other than the BHMI that we investigate in the paper.^{11,12} We investigate bandwidth selection for the nearest neighbor BHMI, which has good performance for continuous covariates. The findings in our paper may not be applicable to longitudinal missing data in which individuals may drop out during follow-up and there could a small CC set.¹³ For longitudinal missing data problems, in addition to imputation and weighting, Bayesian modeling or likelihood inference based on the available data have been widely applied. In practice, sensitivity analysis based on some modeling assumptions could be useful in the explanation of potential research findings.¹⁴

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments and Data Availability Statement

This research was partially supported by US National Cancer Institute grants CA189532, CA235122 (Wang and Hsu), CA86368, CA239168, National Heart, Lung, and Blood Institute grant HL130483 (Wang), and a travel award from the Mathematics Research Promotion Center of National Science Council of Taiwan (Wang). The data that support the findings of this study are available from Table 3 of Campbell et al.¹

References

1. Campbell PT, Jacobs ET, Ulrich CM, Figueiredo JC, Poynter JN, McLaughlin JR, Haile RW, Jacobs EJ, Newcomb PA, Potter JD and Le Marchand L. Case-control study of overweight, obesity, and colorectal cancer risk, overall and by tumor microsatellite instability status. *Journal of the National Cancer Institute*. 2010; 102: 391–400. [PubMed: 20208017]
2. Gryfe R, Kim H, Hsieh ET, et al. Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. *N Engl J Med*. 2000; 342: 69–77. [PubMed: 10631274]
3. Ribic CM, Sargent DJ, Moore MJ, et al. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. *N Engl J Med*. 2003; 349: 247–257. [PubMed: 12867608]
4. Little RJA and Rubin DB. *Statistical Analysis with Missing Data*, John Wiley & Sons, 2nd Edition New York; 2002.
5. Rubin DB. Inference and missing data. *Biometrika*. 1976; 63: 581–592.
6. Robins JM, Rotnitzky A and Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89: 846–866.
7. Qi L, Wang YF, He Y. A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates. *Statistics in Medicine*. 2010; 29: 2592–604. [PubMed: 20806403]

8. Wang CY, Lee SM and Chao EC. Numerical equivalence of imputing scores and weighted estimators in regression analysis with missing covariates. *Biostatistics*. 2007; 8: 468–473. [PubMed: 16971377]
9. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons; 1987.
10. Wu H and Wu L. Identification of significant host factors for HIV dynamics models by nonlinear mixed-effect models. *Statistics in Medicine*. 2002; 21: 753–771. [PubMed: 11870815]
11. von Hippel PT. Regression with missing Ys: an improved strategy for analyzing multiply imputed data. *Social Methodol*. 2007;37(1):83117.
12. Sullivan TR, Salter AB, Ryan P, Lee KJ. Bias and precision of the “multiple imputation, then deletion” method for dealing with missing outcome data. *Am J Epidemiol*. 2015;182(6):52834.
13. Lavori PW, Dawson R, Shera D. A multiple imputation strategy for clinical trials with truncation of patient data. *Statistics in Medicine*, 1995;14:191325.
14. McElreath R (2016) *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: Chapman & Hall/CRC.
15. Andridge RR and Little RJA. A review of hot deck imputation for survey non-response. *International Statistical Review*. 2010; 78: 40–64. [PubMed: 21743766]
16. Rubin DB and Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*. 1986; 81: 366–374.
17. Qi L, Wang CY and Prentice RL. Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*. 2005; 100: 1250–1263.
18. Magder LS and Hughes JP. Logistic regression when outcome is measured with uncertainty. *American Journal of Epidemiology*. 1997; 246: 195–203.
19. Wang CY, Huang Y, Chao EC, and Jeffcoat MK. Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorably missing data. *Biometrics*. 2008; 64: 85–95. [PubMed: 17608787]
20. Wang S and Wang CY. Asymptotic comparisons of kernel assisted estimators in missing covariate regression. *Statistics and Probability Letters*. 2001; 55: 439–449.
21. White IR and Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*. 2010; 29: 2920–2931. [PubMed: 20842622]

Table 1:Simulation study in multinomial logistic regression when Y is MCAR while $S = \mathbb{I}[Y > 0]$ is available

	CC	BHMI	SIPW	AIPW	EEE	CC	BHMI	SIPW	AIPW	EEE
	<hr/>					<hr/>				
	<i>n</i> = 500					<i>n</i> = 1000				
	<hr/>									
$\beta_{0,1} = -\ln(2)$										
Bias	0.005	-0.007	0.005	0.005	0.005	-0.006	-0.012	-0.008	-0.008	-0.008
SD	0.230	0.199	0.191	0.191	0.191	0.157	0.133	0.129	0.129	0.129
ASE	0.221	0.193	0.180	0.180	0.180	0.155	0.134	0.127	0.127	0.127
CP	0.944	0.936	0.928	0.928	0.928	0.956	0.954	0.946	0.946	0.946
$\beta_{1,1} = \ln(2)$										
Bias	0.013	0.013	0.007	0.007	0.007	-0.002	0.003	0.001	0.001	0.001
SD	0.317	0.259	0.249	0.249	0.249	0.237	0.193	0.190	0.190	0.190
ASE	0.313	0.267	0.251	0.251	0.251	0.220	0.188	0.177	0.177	0.177
CP	0.956	0.952	0.954	0.954	0.954	0.930	0.950	0.932	0.932	0.932
$\beta_{0,2} = -\ln(2)$										
Bias	0.006	-0.007	0.005	0.005	0.005	-0.001	-0.006	-0.003	-0.003	-0.003
SD	0.232	0.196	0.186	0.186	0.186	0.149	0.127	0.124	0.124	0.124
ASE	0.221	0.194	0.180	0.180	0.180	0.155	0.135	0.127	0.127	0.127
CP	0.944	0.944	0.954	0.954	0.954	0.962	0.954	0.956	0.956	0.956
$\beta_{1,2} = \ln(2)$										
Bias	0.003	0.002	-0.004	-0.004	-0.004	-0.008	-0.005	-0.006	-0.006	-0.006
SD	0.323	0.268	0.260	0.260	0.260	0.226	0.186	0.183	0.183	0.183
ASE	0.314	0.268	0.251	0.251	0.251	0.220	0.188	0.177	0.177	0.177
CP	0.960	0.946	0.950	0.950	0.950	0.944	0.936	0.942	0.942	0.942
$\beta_{1,2} - \beta_{1,1}$										
Bias	-0.011	-0.011	-0.007	-0.011	-0.011	-0.007	-0.008	-0.007	-0.007	-0.007
SD	0.345	0.359	0.345	0.345	0.345	0.250	0.257	0.250	0.250	0.250
ASE	0.338	0.382	0.338	0.338	0.338	0.238	0.270	0.238	0.238	0.238
CP	0.942	0.960	0.942	0.942	0.942	0.930	0.958	0.930	0.930	0.930

NOTE: The selection probability is a constant with 50% individuals in the complete-case sample. CC is the complete-case analysis. The weighted estimator is similar to the CC but weighted by the inverse selection probabilities that are from an empirical function of (X, S) . The BHMI estimates are obtained from 30 bootstrap imputed data sets. The results are from 500 simulation replicates.

Table 2:Simulation study in multinomial logistic regression when Y is MAR while $S = \mathbb{I}(Y > 0)$ is available

	CC	BHMI	SIPW	AIPW	EEE	CC	BHMI	SIPW	AIPW	EEE
	$n = 1000$					$n = 2000$				
$\beta_{0,1} = -\ln(2)$										
Bias	-0.635	-0.023	-0.015	-0.015	-0.015	-0.625	-0.011	-0.007	-0.007	-0.007
SD	0.199	0.162	0.158	0.158	0.158	0.134	0.108	0.105	0.105	0.105
ASE	0.196	0.164	0.153	0.153	0.153	0.138	0.115	0.104	0.104	0.104
CP	0.074	0.956	0.952	0.952	0.952	0.006	0.950	0.940	0.940	0.940
$\beta_{1,1} = \ln(2)$										
Bias	-0.187	-0.007	0.001	0.001	0.001	-0.182	0.006	0.008	0.008	0.008
SD	0.357	0.273	0.264	0.264	0.264	0.221	0.172	0.169	0.169	0.169
ASE	0.337	0.262	0.243	0.243	0.243	0.236	0.178	0.170	0.170	0.170
CP	0.906	0.948	0.934	0.934	0.934	0.896	0.934	0.944	0.944	0.944
$\beta_{0,2} = -\ln(2)$										
Bias	-0.625	-0.011	-0.004	-0.004	-0.004	-0.610	0.003	0.008	0.008	0.008
SD	0.182	0.154	0.148	0.148	0.148	0.132	0.105	0.101	0.101	0.101
ASE	0.195	0.163	0.152	0.152	0.152	0.137	0.114	0.107	0.107	0.107
CP	0.052	0.972	0.958	0.958	0.958	0.002	0.966	0.968	0.968	0.966
$\beta_{1,2} = \ln(2)$										
Bias	-0.209	-0.027	-0.021	-0.021	-0.021	-0.212	-0.022	-0.022	-0.022	-0.022
SD	0.341	0.255	0.246	0.246	0.246	0.227	0.172	0.168	0.168	0.168
ASE	0.337	0.266	0.243	0.243	0.243	0.236	0.179	0.170	0.170	0.170
CP	0.910	0.958	0.946	0.946	0.946	0.860	0.946	0.944	0.944	0.944
$\beta_{1,2} - \beta_{1,1}$										
Bias	-0.022	-0.021	-0.022	-0.022	-0.022	-0.029	-0.028	-0.029	-0.029	-0.029
SD	0.425	0.445	0.425	0.425	0.425	0.282	0.288	0.282	0.282	0.282
ASE	0.408	0.452	0.408	0.408	0.408	0.284	0.304	0.285	0.285	0.285
CP	0.942	0.950	0.942	0.942	0.942	0.934	0.954	0.934	0.934	0.934

NOTE: The selection probability is $\{1 + \exp(X + S)\}^{-1}$, and the complete-case sample size is about 28% of the size of the whole cohort. CC is the complete-case analysis. The weighted estimator is similar to the CC but weighted by the inverse selection probabilities that are from an empirical function of (X, S) . The BHMI estimates are obtained from 30 bootstrap imputed data sets. The results are from 500 simulation replicates.

Table 3:

Simulation study in multinomial logistic regression when Y is MAR while $S = \mathbb{I}[Y > 0] \times \mathbb{I}[U > 0.9]$ is available in which U is uniform $[0,1]$

	CC	BHMI	SIPW	AIPW	EEE	CC	BHMI	SIPW	AIPW	EEE
	$n = 1000$					$n = 2000$				
$\beta_{0,1} = -\ln(2)$										
Bias	-0.540	-0.017	-0.005	-0.005	-0.005	-0.543	-0.009	-0.004	-0.004	-0.004
SD	0.187	0.165	0.159	0.159	0.159	0.136	0.108	0.106	0.106	0.106
ASE	0.189	0.170	0.155	0.155	0.155	0.136	0.117	0.109	0.109	0.106
CP	0.148	0.960	0.946	0.946	0.946	0.016	0.944	0.956	0.956	0.956
$\beta_{1,1} = \ln(2)$										
Bias	-0.179	-0.008	-0.009	-0.009	-0.009	-0.166	-0.006	-0.007	-0.007	-0.007
SD	0.312	0.275	0.261	0.261	0.261	0.230	0.183	0.178	0.178	0.178
ASE	0.324	0.282	0.254	0.254	0.254	0.227	0.190	0.178	0.178	0.178
CP	0.942	0.960	0.948	0.948	0.948	0.890	0.946	0.958	0.958	0.958
$\beta_{0,2} = -\ln(2)$										
Bias	-0.526	-0.003	0.006	0.006	0.006	-0.539	-0.004	0.000	0.000	0.000
SD	0.185	0.157	0.151	0.151	0.151	0.142	0.116	0.114	0.114	0.114
ASE	0.188	0.170	0.154	0.154	0.154	0.133	0.118	0.109	0.109	0.109
CP	0.166	0.962	0.958	0.958	0.958	0.018	0.938	0.936	0.936	0.936
$\beta_{1,2} = \ln(2)$										
Bias	-0.195	-0.031	-0.024	-0.024	-0.024	-0.158	-0.001	-0.002	-0.002	-0.002
SD	0.313	0.265	0.251	0.251	0.251	0.224	0.177	0.169	0.169	0.169
ASE	0.324	0.277	0.254	0.254	0.254	0.227	0.188	0.177	0.177	0.177
CP	0.916	0.968	0.958	0.958	0.958	0.894	0.956	0.964	0.964	0.964
$\beta_{1,2} - \beta_{1,1}$										
Bias	-0.015	-0.023	-0.015	-0.015	-0.015	0.007	0.005	0.005	0.005	0.005
SD	0.387	0.428	0.403	0.403	0.403	0.258	0.271	0.262	0.262	0.262
ASE	0.387	0.443	0.397	0.397	0.397	0.270	0.297	0.277	0.277	0.277
CP	0.952	0.956	0.948	0.948	0.948	0.966	0.958	0.964	0.964	0.964

NOTE: The selection probability is $\{1 + \exp(X + S)\}^{-1}$, and the complete-case sample size is about 29% of the size of the whole cohort. CC is the complete-case analysis. The weighted estimator is similar to the CC but weighted by the inverse selection probabilities that are from an empirical function of (X, S) . The BHMI estimates are obtained from 30 bootstrap imputed data sets. The results are from 500 simulation replicates.

Table 4:

Simulation study with a continuous covariate when Y is MAR. Auxiliary variable $S = \mathbb{I}[Y > 0] \times \mathbb{I}[U > 0.3] + \mathbb{I}[Y = 0] \times \mathbb{I}[U > 0.8]$ is available where U is uniform $[0,1]$

	CC	BHMI	SIPW	AIPW	EEE	CC	BHMI	SIPW	AIPW	EEE
	$n = 1000$					$n = 2000$				
$\beta_{0,1} = -\ln(2)$										
Bias	-0.293	-0.030	-0.029	-0.019	-0.009	-0.277	-0.001	0.005	0.006	0.014
SD	0.154	0.154	0.154	0.154	0.154	0.101	0.108	0.106	0.107	0.106
ASE	0.145	0.160	0.151	0.154	0.148	0.102	0.110	0.104	0.104	0.103
CP	0.488	0.946	0.932	0.938	0.930	0.220	0.954	0.936	0.942	0.932
$\beta_{1,1} = \ln(2)$										
Bias	-0.091	-0.006	-0.002	-0.015	-0.002	-0.090	0.001	0.003	0.007	-0.005
SD	0.146	0.172	0.176	0.179	0.164	0.107	0.127	0.127	0.128	0.123
ASE	0.152	0.183	0.176	0.177	0.166	0.107	0.126	0.122	0.124	0.116
CP	0.930	0.952	0.934	0.940	0.950	0.864	0.946	0.936	0.938	0.932
$\beta_{0,2} = -\ln(2)$										
Bias	-0.298	-0.034	-0.023	-0.023	-0.012	-0.284	-0.010	-0.004	-0.005	0.003
SD	0.142	0.149	0.150	0.149	0.151	0.105	0.106	0.106	0.107	0.105
ASE	0.146	0.159	0.151	0.153	0.148	0.102	0.111	0.105	0.106	0.103
CP	0.468	0.962	0.952	0.952	0.946	0.210	0.956	0.946	0.946	0.954
$\beta_{1,2} = \ln(2)$										
Bias	-0.094	-0.004	0.000	0.003	-0.012	-0.095	-0.004	-0.002	0.003	-0.009
SD	0.150	0.182	0.181	0.182	0.172	0.109	0.122	0.122	0.123	0.118
ASE	0.152	0.183	0.178	0.180	0.167	0.107	0.128	0.124	0.125	0.118
CP	0.910	0.940	0.936	0.938	0.948	0.860	0.954	0.948	0.950	0.934
$\beta_{1,2} - \beta_{1,1}$										
Bias	-0.003	0.001	0.002	0.001	0.003	-0.005	-0.005	-0.005	-0.004	-0.004
SD	0.183	0.239	0.241	0.245	0.223	0.124	0.155	0.160	0.158	0.152
ASE	0.180	0.235	0.227	0.231	0.210	0.126	0.162	0.158	0.161	0.148
CP	0.938	0.926	0.908	0.910	0.916	0.956	0.948	0.932	0.944	0.938

NOTE: The selection probability is $\{1 + \exp(X + S)\}^{-1}$, and the complete-case sample size is about 42% of the size of the whole cohort. CC is the complete-case analysis. The weighted estimator is similar to the CC but weighted by the inverse selection probabilities that are from a smoothing function of (X, S) . The BHMI estimates are obtained from 50 bootstrap imputed data sets. The results are from 500 simulation replicates.

Table 5:

Analysis results of a colorectal cancer study

		CC	BHMI	SIPW	AIPW	EEE
Intercept-1	$\beta_{0,1}$	-0.773	-0.768	-0.789	-0.789	-0.789
	SE	0.123	0.075	0.080	0.080	0.080
BMI on MSS	$\beta_{1,1}$	0.145	0.245	0.280	0.280	0.280
	SE	0.155	0.088	0.099	0.099	0.099
Intercept-2	$\beta_{0,2}$	-2.119	-2.190	-2.134	-2.134	-2.134
	SE	0.211	0.205	0.189	0.189	0.189
BMI on MSI-H	$\beta_{1,2}$	-0.148	0.040	-0.012	-0.012	-0.012
	SE	0.277	0.225	0.250	0.250	0.250
	$\beta_{1,2} - \beta_{1,1}$	-0.293	-0.206	-0.293	-0.293	-0.293
	SE	0.291	0.249	0.291	0.291	0.291

Note: CC is the complete-case analysis. The SIPW and AIPW estimators are the simple and augmented inverse probability weighted estimators. The BHMI estimates are obtained from 30 bootstrap imputed data sets, the EEE estimator is the same as the ML estimator given that X and S are discrete. The analysis is based on MS stable and MSI-high cases and controls ($n = 2648$), and the CC subset is a random sample with size $n=866$.