The Annals of Applied Statistics
2021, Vol. 15, No. 1, 126–147
https://doi.org/10.1214/20-AOAS1386

ROBUST INFERENCE WHEN COMBINING INVERSE-PROBABILITY WEIGHTING AND MULTIPLE IMPUTATION TO ADDRESS MISSING DATA WITH APPLICATION TO AN ELECTRONIC HEALTH RECORDS-BASED STUDY OF BARIATRIC SURGERY

BY TANAYOTT THAWEETHAI[1], DAVID E. ARTERBURN[2], KAREN J. COLEMAN[3] AND SEBASTIEN HANEUSE[4]

[1]*Biostatistics Center, Massachusetts General Hospital, tthaweethai@mgh.harvard.edu*
[2]*Health Research Institute, Kaiser Permanente Washington, David.E.Arterburn@kp.org*
[3]*Department of Research & Evaluation, Kaiser Permanente Southern California, Karen.J.Coleman@kp.org*
[4]*Department of Biostatistics, Harvard T. H. Chan School of Public Health, shaneuse@hsph.harvard.edu*

While electronic health records present a rich and promising data source for observational research, they are highly susceptible to missing data. For settings like these, Seaman et al. (*Biometrics* **68** (2012) 129–137) proposed a strategy wherein one handles missingness in some variables using inverse-probability weighting and others using multiple imputation. Seaman et al. (*Biometrics* **68** (2012) 129–137) show that Rubin's variance estimator for averaging results across datasets is asymptotically valid when the analysis and imputation models are correctly specified and the weights are either known or correctly specified. Modeled after the approach of Robins and Wang (*Biometrika* **87** (2000) 113–124), we propose a method for asymptotically valid inference that is robust to violation of these conditions. Following a simulation study in which we demonstrate that a proposed variance estimator can reduce bias due to model misspecification, we illustrate this approach in an electronic health records-based study investigating whether differences in long-term weight loss between bariatric surgery techniques are associated with chronic kidney disease at baseline. We observe that the weight loss advantage after five years of Roux-en-Y gastric bypass surgery, compared to vertical sleeve gastrectomy, is less pronounced among patients with chronic kidney disease at baseline compared to those without.

1. Introduction. Electronic health records (EHR) present an enormous opportunity for conducting observational research and include information on large populations over long periods of time. While EHR data are relatively inexpensive to obtain, they are generated for clinical and/or billing purposes, not research (Bruce Bayley et al. (2013), Haneuse and Shortreed (2017)). Upon defining a research question and the set of variables needed to perform an observational study answering that question using EHR data, it is rarely the case that all variables are routinely collected in clinical care, and such data is often not collected consistently across time. Indeed, missing data is extremely common in EHR-based studies, and, while it is possible to perform an intended analysis using only "complete cases" (i.e., those individuals with no missing data in the variables needed to perform the analysis), such an analysis may be subject to selection bias if the associations of interest are different in the included and excluded populations (Haneuse (2016), Seaman and White (2013)).

Consider a hypothetical study of the association between covariates X_1, X_2, X_3 and outcome Y, perhaps through the statistical model $Y = f(X_1, X_2, X_3; \theta)$, where θ is the parameter that indexes that model and is the quantity of interest. We refer to the statistical model

Received February 2020; revised August 2020.
Key words and phrases. Missing data, multiple imputation, inverse-probability weighting, model misspecification, electronic health records, bariatric surgery, obesity, chronic kidney disease.

TABLE 1
Missingness patterns for a hypothetical study population

Pattern	X_1	X_2	X_3	Y	Z_1	Z_2
1	✓	✓			✓	✓
2	✓	✓	✓		✓	✓
3	✓	✓		✓	✓	✓
4	✓	✓	✓	✓	✓	✓
5	✓				✓	✓
6	✓		✓		✓	✓
7	✓			✓	✓	
8	✓		✓	✓	✓	

For each missingness pattern, ✓ indicates that the variable is observed. A blank space indicates that the variable is missing.

being fit in this study as the *analysis model*. In practice, suppose we observe a population consisting of individuals with varying missingness patterns in the data as described in Table 1. Auxiliary variables (i.e., those not directly involved in the analysis model) Z_1 and Z_2 are also recorded. Analysts have at their disposal a wide range of tools for dealing with missingness in the analysis model variables, including inverse-probability weighting (IPW), multiple imputation (MI) and doubly-robust methods. In IPW, one initially constructs and fits a model for the probability of being a complete case for the analysis model variables (missingness pattern 4 in Table 1). Then, an analysis is performed using only the complete cases but with their contributions weighted by the inverse of their estimated probabilities of being a complete case (Seaman and White (2013)). In contrast, an MI analysis first constructs and fits a model for the joint distribution of the analysis model variables that exhibit missingness (X_2, X_3, Y), conditional on the analysis model variables that are fully observed (X_1) as well as any fully observed variables associated with missingness or the values of the missing variables (Z_1, potentially). One generates M "complete" datasets by using this distribution to impute any missing values. The analysis model is fit on all complete datasets, and the results are averaged over the completed datasets (Kenward and Carpenter (2007)). Doubly-robust methods augment standard IPW methods, in a sense, by specifying a second model for the distribution of the missing variable(s) in addition to the model for the probability of being a complete case. Doubly-robust methods, unlike IPW or MI alone, allow the analyst to perform unbiased estimation and inference that is "robust" to model misspecification, if either the model for the distribution of the missing data or the model for the probability of being a complete case is correctly specified (Bang and Robins (2005)).

When both auxiliary models are correctly specified, so that IPW alone, MI alone and the doubly-robust estimator are all consistent, MI is known to be most efficient. Seaman et al. (2012) proposed a novel strategy that combines IPW and MI in a way that is distinct from established doubly-robust methods. Rather than positing a model for the probability that all missing variables are observed as well as a model for the distribution of all missing variables (as in doubly-robust methods), the analyst begins by partitioning the set of variables with missingness into those handled using IPW and those handled using MI and then specifying separate models for each.

In the Seaman et al. (2012) approach, the analyst first specifies some rule to identify a subset of individuals whose missing values will be multiply imputed, say, those with X_2 observed (missingness patterns 1–4 in Table 1). Subjects who fail to meet the inclusion criteria are excluded from the main analysis. Then, any remaining missing variables (X_3 and/or Y in individuals with missingness patterns 1–3) have their values imputed. Following multiple

imputation, M "quasi-complete" datasets are generated, so called because data is complete only for individuals who meet the analyst-specified rule. Each quasi-complete dataset is then analyzed using IPW to account for exclusion of individuals who failed to meet the inclusion criteria and from which M point estimates, $\tilde{\theta}^{(1)}, \ldots, \tilde{\theta}^{(M)}$, are obtained. For each of these analyses, a sandwich-type estimator of the variance is used to account for estimation and use of IPW (Robins, Rotnitzky and Zhao (1994)), yielding $\tilde{V}^{(1)}, \ldots, \tilde{V}^{(M)}$. Finally, an overall estimate $\tilde{\theta} = M^{-1} \sum_{j=1}^{M} \tilde{\theta}^{(j)}$ is computed, with inference based on an estimate of the variance of $\tilde{\theta}$ that uses Rubin's rules (Rubin (2004)):

$$\tilde{V} = \frac{1}{M} \sum_{j=1}^{M} \tilde{V}^{(j)} + \frac{(1 + M^{-1})}{(M - 1)} \sum_{j=1}^{M} \big(\tilde{\theta}^{(j)} - \tilde{\theta}\big)\big(\tilde{\theta}^{(j)} - \tilde{\theta}\big)^{T}.$$

Theorem 1 of Seaman et al. (2012) establishes conditions for the consistency of $\tilde{\theta}$ as an estimator of θ, while Theorem 2 establishes consistency of $\tilde{V}$ as an estimator of the variance of $\tilde{\theta}$ when:

(c.1) the analysis model is a linear regression and is correctly specified,
(c.2) the outcome variable is properly imputed from its posterior predictive distribution using the regression imputation procedure of Schenker and Welsh (1988),
(c.3) all of the pairwise interactions between the inclusion criteria selection weights and the variables used in the weighting model are included in the imputation model as covariates,
(c.4) the imputation model is correctly specified, and
(c.5) the weights used for IPW are known.

Quartagno, Carpenter and Goldstein (2019) developed a strategy for incorporating weights when multiple imputation is used, proposing a method in which the weights form distinct strata and the covariance matrices in the imputation model follow a random distribution across strata such that condition (c.3) is met. Seaman et al. (2012) observed, via theoretical results and simulations, that $\tilde{V}$ is approximately unbiased in settings where covariates are imputed in addition to the outcome. Practically, they found that, when both weighting and imputation models were correctly specified, combining IPW with MI offered efficiency advantages compared to using IPW alone, whereas MI alone was most efficient. However, when the imputation model was misspecified (thus violating condition c.4), point estimates obtained using IPW and MI were less biased than those obtained using MI alone.

This paper is concerned with settings where researchers employ the framework of Seaman et al. (2012) but that the weighting, imputation and/or analysis models are potentially misspecified. If this is the case, then $\tilde{\theta}$ will converge to some value, θ^*, which may or may not equal θ. Regardless of whether θ^* equals θ, which may occur in some settings, valid inference remains of interest. Indeed, this is the central concern of the seminal paper by White (1982), which focuses on valid likelihood-based inference in the presence of model misspecification, as well as more recent work in this area (Abadie, Imbens and Zheng (2014), Imbens and Kolesár (2016), Stefanski and Boos (2002)). However, Rubin's rules, as presented above, are only guaranteed to generate asymptotically valid inference concerning θ^* when the weighting, imputation and analysis models are correctly specified. Since one or more of these models may be misspecified, we propose a novel analysis strategy for the Seaman et al. (2012) framework that enables the analyst to conduct inference that is robust to misspecification of the weighting, imputation and analysis models while remaining asymptotically valid. Toward this, we consider an alternative estimator of θ, denoted $\hat{\theta}$, that is based on an improper imputation procedure, as defined by Wang and Robins (1998), rather than proper imputation. Henceforth, robust inference in this paper refers to valid variance estimation for $\hat{\theta}$ in the presence of model misspecification, regardless of whether $\hat{\theta}$ converges to θ. That is, Wald intervals generated for $\hat{\theta}$ will, in large samples, achieve their nominal coverage about θ^*.

Robins and Wang (2000) considered misspecification in the analysis and imputation models in the setting where improper multiple imputation alone is used to handle missing data. They propose a variance estimator for that setting that is robust to misspecification of these models. In this paper we extend their approach and propose a variance estimator that incorporates inverse-probability weights while maintaining these robustness properties. Because we use a sandwich-type estimator to account for estimation and use of the weights, the proposed variance estimator also demonstrates robustness against misspecification of the weighting model.

The remainder of this paper is as follows. In Section 2 we describe motivating questions in assessing the efficacy of bariatric surgery among individuals with and without chronic kidney disease. We describe the methodology for estimation and inference in Section 3. A worked example is provided in Section 4, while Section 5 presents a simulation study that investigates small-sample properties, including the robustness of inference of the proposed methods relative to that proposed by Seaman et al. (2012). We then apply the proposed analytic framework to an EHR-based study comparing weight loss across different techniques for bariatric surgery among patients with and without impaired renal function in Section 6. Section 7 concludes the paper with a discussion.

2. Motivation. Bariatric surgery is a commonly performed surgical procedure that has been shown to achieve dramatic and lasting weight loss in obese patients (Li et al. (2019), Maciejewski et al. (2016)). There are multiple treatment options for patients undergoing bariatric surgery, the two most common of which are Roux-en-Y gastric bypass (RYGB) and vertical sleeve gastrectomy (VSG). RYGB is a surgical procedure in which a small pouch is created from the stomach and is connected to the small intestine, bypassing most of the stomach, and has been considered the "gold standard" for decades. However, VSG, in which part of the stomach is removed, has become increasingly popular, as it is believed to be less complex, less risky and less invasive than RYGB (Zellmer et al. (2014)). In a retrospective observational cohort study involving over 60,000 bariatric surgery procedures at 41 health systems, adults who underwent RYGB were observed to have lost more weight at one, three and five years than VSG but had higher 30-day rates of major adverse events (Arterburn et al. (2018)).

There is also growing evidence that impaired renal function (e.g., chronic kidney disease) is associated with increased risk of complications following bariatric surgery (Neff, Olbers and le Roux (2013), Nguyen et al. (2011), Turgeon et al. (2012)). RYGB has also been shown to be associated with more renal complications than VSG; in one observational study of 762 patients who underwent bariatric surgery, the risk of nephrolithiasis was higher for patients undergoing RYGB compared to VSG (Lieske et al. (2015)). Oxalate nephropathy, which generally results in rapid progression to end-stage renal disease, has also been reported in association with RYGB (Nasr et al. (2008)). Given the complex relationship between renal function and bariatric surgery, it is not well known whether the weight loss advantage of RYGB compared to VSG holds among those with presurgical chronic kidney disease and, if that advantage persists, whether that outweighs the potential increased risks associated with RYGB.

We implement the proposed framework in an analysis of long-term weight loss among patients in the DURABLE (DURAtion of Bariatric Long Term Effects) study, a large, ongoing, NIH-funded, multicenter retrospective cohort study investigating the health outcomes of patients who undergo bariatric surgery. Data in this study are derived from patient electronic health records (EHR) in three healthcare systems: Kaiser Permanente Northern California, Kaiser Permanente Southern California and Kaiser Permanente Washington. The amount of available data regarding BMI and comorbid conditions varies widely between patients. One

particular challenge is that the absence of an indicator for a condition, such as chronic kidney disease, does not necessarily mean that renal function is normal. One must, therefore, distinguish between the absence of a condition and missing data on whether the condition is present. We return to these motivating questions in Section 6.

3. Methods.

3.1. *Analyses based on complete data.* In a random sample of size N, we define D as the set of variables included in the analysis model (i.e., the model of substantive interest). We refer once more to the hypothetical study population described in Table 1. In this example, $D = (X_1, X_2, X_3, Y)$. In the presence of complete data, let $U(\theta; D_i)$ denote an individual's contribution to the (unweighted) complete-data estimating equations of the analysis model, where θ is the parameter of substantive interest.

3.2. *A modification of the Seaman et al. (2012) approach to missing data.* Let R_i denote the missingness pattern in D for the ith individual. We partition D_i into D_i^m, the missing values of D for the ith individual, and D_i^o, the observed values of D for the ith individual. In the hypothetical study, if the ith individual has missingness pattern 1, then $D_i^m = (X_{3,i}, Y_i)$ and $D_i^o = (X_{1,i}, X_{2,i})$. At this point we suppose that a decision is made by the analysis team that, effectively, identifies a subsample of the N individuals who will be "included" in the main analysis (i.e., directly contribute to estimation of θ) and those who will be "excluded." Seaman et al. (2012) conceptualize this in the form of a user-specified function of the missing data pattern, termed a "rule," $\mathcal{R}$. Here, we denote this rule, as applied to the ith individual, by $\mathcal{R}_i \equiv \mathcal{R}(R_i)$, with $\mathcal{R}_i = 1$ indicating that the individual is included and $\mathcal{R}_i = 0$ indicating that they are excluded.

Theorem 1 of Seaman et al. (2012) requires that the distribution of $\mathcal{R}$ depends only on variables that are observed for all subjects in order to achieve unbiased estimation. In the hypothetical example, suppose we define $\mathcal{R}(R_i)$ to be equal to 1 when X_2 is observed and 0 otherwise. Consequently, we include only individuals with missingness patterns 1, 2, 3 and 4 in the main analysis. The probability that $\mathcal{R} = 1$ can depend only on X_1 and Z_1, since those are the only fully observed variables for all subjects.

Among individuals with $\mathcal{R} = 1$, some will have complete data in the sense that D is fully observed and therefore $U(\theta; D_i)$ can be evaluated (i.e., missingness pattern 4). Others, however, may not, and Seaman et al. (2012) propose that any remaining missingness be resolved via multiple imputation. Restricting to the subpopulation of subjects with $\mathcal{R} = 1$, by Theorem 1 of Seaman et al. (2012), any missing variables must be missing at random (MAR); that is, the probability that data is missing can depend only on observed data (Rubin (1976)). In the hypothetical example, the remaining missing variables to be imputed are X_3 and Y, and missingness in those variables can depend only on variables that are fully observed among individuals with $\mathcal{R} = 1$: X_1, X_2, Z_1 and Z_2. Whereas the general framework being introduced in this paper matches that of Seaman et al. (2012), the most notable departure is the use of improper imputation, as opposed to proper imputation. The following subsections formalize the procedure.

3.3. *Estimation.* Let H be a vector of fully observed variables, some of which may belong to D, that predict the probability an individual has a missing pattern R_i such that $\mathcal{R}_i = 1$. In the hypothetical example in Table 1, only X_1 and Z_1 may be included in H. Specify a model for $\text{pr}(\mathcal{R} = 1|H; \alpha)$ indexed by α. Let $S_{\alpha,i}(\alpha) = S_\alpha(\alpha; H_i)$ denote an individual's contribution to the score equation for estimating $\text{pr}(\mathcal{R} = 1|H; \alpha)$. Then, $\hat{\alpha}$ is the solution to estimating equations $\sum_{i=1}^N S_{\alpha,i}(\alpha) = 0$, where $\hat{\alpha}$ converges in probability to a limit α^*. Define $W(\alpha; H) = \text{pr}(\mathcal{R} = 1 \mid H; \alpha)^{-1}$.

Based on the missingness patterns observed in the population of individuals with $\mathcal{R} = 1$, D is partitioned into $\{D^o_{\mathcal{R}}, D^m_{\mathcal{R}}\}$, where $D^o_{\mathcal{R}}$ denotes the variables in D that are fully observed among all individuals with $\mathcal{R} = 1$ and $D^m_{\mathcal{R}}$ denotes variables that exhibit any missingness among those individuals. In the hypothetical example, $D^o_{\mathcal{R}} = (X_1, X_2)$ and $D^m_{\mathcal{R}} = (X_3, Y)$. $D^o_{\mathcal{R},i}$ and $D^m_{\mathcal{R},i}$ denote the values of $D^o_{\mathcal{R}}$ and $D^m_{\mathcal{R}}$ for the ith individual, respectively.

There is a subtle, yet important, distinction between $D^m_{\mathcal{R},i}$ and D^m_i and between $D^o_{\mathcal{R},i}$ and D^o_i: $D^m_{\mathcal{R}}$ is defined by the missingness patterns of the entire sample of individuals with $\mathcal{R} = 1$, whereas D^m_i represents the values of variables that are missing for the ith individual alone. D^m_i is completely unobserved. A "complete case" would have $D^m_i = \varnothing$, but $D^m_{\mathcal{R},i}$ would be the ith individual's values of the variables for which any individuals with $\mathcal{R} = 1$ have missing data. D^o_i and $D^o_{\mathcal{R},i}$ are defined analogously. In the hypothetical example, if the ith individual has missingness pattern 3, $D^m_{\mathcal{R},i} = (X_{3,i}, Y_i)$ and $D^o_{\mathcal{R},i} = (X_{1,i}, X_{2,i})$, whereas $D^m_i = X_{3,i}$ and $D^o_i = (X_{1,i}, X_{2,i}, Y_i)$

Next, define $D^\dagger$ as the set of variables distinct from D that, conditional on $\mathcal{R} = 1$, are associated with missingness in D and/or will be used for imputing D. $D^\dagger$ must be fully observed among individuals with $\mathcal{R} = 1$. $D^\dagger$ is often needed to make the MAR assumption, as described at the end of Section 3.2, tenable. In the hypothetical example the only variables that may be included in $D^\dagger$ are Z_1 and Z_2. Although Z_2 is not always observed, it is observed for all individuals with $\mathcal{R} = 1$, and it is not already included in D.

Specify an imputation model $g(D^m_{\mathcal{R}} \mid D^o_{\mathcal{R}}, D^\dagger, \mathcal{R} = 1; \psi)$, where ψ indexes the model. A possible imputation model in the hypothetical example could, therefore, be $g(X_3, Y \mid X_1, X_2, Z_1, Z_2, \mathcal{R} = 1; \psi)$. Let $S_{\psi,i}(\psi)$ denote an individual's contribution to the score equation for $g(D^m_{\mathcal{R}} \mid D^o_{\mathcal{R}}, D^\dagger, \mathcal{R} = 1; \psi)$; that is, $S_{\psi,i}(\psi) = \partial \log g(D^m_{\mathcal{R},i} \mid D^o_{\mathcal{R},i}, D^\dagger_i, \mathcal{R}_i = 1; \psi)/\partial\psi$. The observed score for the ith individual is then defined as $S^{\text{obs}}_{\psi,i}(\psi) = E_{D^m_i}[S_{\psi,i}(\psi)|D^o_i]$. Then, $\hat{\psi}$ is the observed-data maximum likelihood estimator; that is, it is the solution to estimating equations $\sum_{i=1}^N S^{\text{obs}}_{\psi,i}(\psi) = 0$, where $\hat{\psi}$ converges in probability to a limit ψ^*.

Missing values D^m_i among individuals with $\mathcal{R} = 1$ are improperly imputed M times using the model $g(D^m_{\mathcal{R}} \mid D^o_{\mathcal{R}}, D^\dagger, \mathcal{R} = 1; \psi)$. Throughout when referenced, an improper imputation procedure will refer to the method defined in Wang and Robins (1998), wherein the preliminary estimate of the imputation model parameters (generally the maximum likelihood estimator) is used to generate all imputations. Then, the estimating equations for the analysis model are solved using data from all M complete datasets, simultaneously, to obtain the improper MI estimator. This is in contrast to a proper imputation procedure, in which imputation model parameters are generally drawn M times from their Bayesian predictive distribution such that Rubin's rules provide valid inference, and then the estimating equations for the analysis model are solved within each of the M datasets and the M parameter estimates averaged (Nielsen (2003)).

Each missingness pattern found in the data corresponds to a particular conditional distribution for D^m_i induced by the original imputation model that is fit. Then, D^m_i is imputed using that derived conditional model. If, in the hypothetical example, we fit the imputation model $g(X_3, Y \mid X_1, X_2, Z_1, Z_2, \mathcal{R} = 1; \psi)$, then, in order to impute data for individuals with missingness pattern 2, we would derive the imputation model $g(Y \mid X_1, X_2, X_3, Z_1, Z_2, \mathcal{R} = 1; \psi)$ and impute missing values of Y with that induced model. As such, only one joint imputation model needs to be fit from the data; every conditional model is induced by that initial fit. Let $D^{m,(j)}_i$ denote the jth imputed value of D^m_i for $j = 1, \ldots, M$, and let $D^{(j)}_i = \{D^{m,(j)}_i, D^o_i\}$.

Following multiple imputation, denote the contribution to the weighted estimating equations of the analysis model by individual i in the jth imputed dataset as

$$S_{\theta,i}^{(j)}(\theta,\alpha,\psi)=S_\theta\big(\theta,\alpha,\psi;D_i^{(j)},H_i\big)=\mathcal{R}_i W_i(\alpha;H_i)U\big(\theta;D_i^{(j)}\big),$$

where $U(\theta;D_i^{(j)})$, the ith individual's contribution to the complete-data estimating equation of the analysis model in the jth imputed dataset, is implicitly a function of ψ because ψ has been used to generate the imputed data $D_i^{(j)}$. Let $\hat\theta$ be the solution to weighted estimating equations

$$\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}S_{\theta,i}^{(j)}(\theta,\hat\alpha,\hat\psi)=\frac{1}{N}\sum_{i=1}^{N}\bar S_{\theta,i}(\theta,\hat\alpha,\hat\psi)=0,$$

where $\bar S_{\theta,i}(\theta,\alpha,\psi)=M^{-1}\sum_{j=1}^{M}S_{\theta,i}^{(j)}(\theta,\alpha,\psi)$, $\hat\alpha$ is used to calculate the inverse-probability weights and $\hat\psi$ has been used for imputation. Under suitable regularity conditions, $\hat\theta$ converges in probability to a limit θ^*. By Theorem 1 of Seaman et al. (2012), if:

(i) model $\mathrm{pr}(\mathcal{R}=1|H;\alpha)$ is correctly specified,
(ii) model $g(D_{\bar{\mathcal{R}}}^m \mid D_{\mathcal{R}}^o, D^\dagger, \mathcal{R}=1;\psi)$ is correctly specified,
(iii) $\mathrm{pr}(\mathcal{R}=1\mid D, D^\dagger, H)=\mathrm{pr}(\mathcal{R}=1\mid H)$,
(iv) $p(R\mid D, D^\dagger, W, \mathcal{R}=1)=p(R\mid D_{\mathcal{R}}^o, D^\dagger, W, \mathcal{R}=1)$ and
(v) $D_{\bar{\mathcal{R}}}^m \perp\!\!\!\perp W \mid D_{\mathcal{R}}^o, D^\dagger, \mathcal{R}=1$,

then, when $M=\infty$, $\hat\theta$ is a consistent estimator of θ. (Note that above, p denotes a law, or distribution, and pr denotes a probability.) Since H can be high dimensional and the true weights $W(\alpha;H)$ are typically unknown, by Corollary 1 of Seaman et al. (2012), an alternative to satisfying Condition (v) is to include the estimated weights $W(\hat\alpha;H)$ in $D^\dagger$.

3.4. *Inference*. Whereas correct specification of the weighting and imputation models are two of five conditions described in Section 3.3 that are sufficient for consistency of $\hat\theta$, correct specification of these models is not required to ensure valid inference (at least with respect to the limit of $\hat\theta$, θ^*). The asymptotic variance of $\hat\theta$ is provided in Theorem 3.1, followed by a consistent estimator for the variance.

THEOREM 3.1. *If missing values are improperly imputed M times from a fully parametric imputation model $g(D_{\bar{\mathcal{R}}}^m|D_{\mathcal{R}}^o, D^\dagger, \mathcal{R}=1;\hat\psi)$ among individuals with $\mathcal{R}=1$, and $\hat\theta$ is estimated, as described in Section 3.3, then $N^{1/2}(\hat\theta-\theta^*)$ is asymptotically normal with mean zero and variance $\Sigma=\tau^{-1}\Omega\tau^{-\top}$, where*

$$\tau=-E\big[\partial S_\theta^{(j)}(\theta,\alpha^*,\psi^*)/\partial\theta^\top\big]_{\theta=\theta^*},\qquad \Omega=E\big[\nu(\theta^*,\alpha^*,\psi^*)^{\otimes 2}\big],$$

$$\nu(\theta^*,\alpha^*,\psi^*)=\bar S_\theta(\theta^*,\alpha^*,\psi^*)-\delta I_\alpha^{-1}S_\alpha(\alpha^*)-\kappa I_\psi^{-1}S_\psi^{\mathrm{obs}}(\psi^*),$$

$$\delta=-E\big[\partial\bar S_\theta(\theta^*,\alpha,\psi^*)/\partial\alpha^\top\big]_{\alpha=\alpha^*},\qquad I_\alpha=-E\big[\partial S_\alpha(\alpha)/\partial\alpha^\top\big]_{\alpha=\alpha^*},$$

$$\kappa=-E\big[S_\theta^{(j)}(\theta^*,\alpha^*,\psi^*)S_\psi^{\mathrm{mis},(j)}(\psi^*)^\top\big],$$

$$I_\psi=-E\big[\partial S_\psi^{\mathrm{obs}}(\psi)/\partial\psi^\top\big]_{\psi=\psi^*},$$

$$S_\psi^{\mathrm{mis},(j)}(\psi^*)=\partial\log g\big(D^{m,(j)}|D^o,D^\dagger,\mathcal{R}=1;\psi\big)/\partial\psi\,|_{\psi=\psi^*}.$$

Note $A^{\otimes 2}\equiv AA^\top$ for any matrix A. The expectation is taken with respect to the density $\{\prod_{j=1}^M g(D_{\bar{\mathcal{R}}}^{m,(j)}|D_{\mathcal{R}}^o,D^\dagger,\mathcal{R}=1;\psi)\}g(D_{\mathcal{R}}^o,D^\dagger|\mathcal{R}=1;\psi)$. It is assumed that the derivatives above exist, and all inverses can be taken as described.

The proof of Theorem 3.1 can be found in the Supplementary Material (Section 1). A consistent estimator of Σ is $\hat{\Sigma} = \hat{\tau}^{-1}\hat{\Omega}\hat{\tau}^{-\top}$, where

$$\hat{\tau} = -N^{-1}\sum_{i=1}^{N} \partial \bar{S}_{\theta,i}(\theta, \hat{\alpha}, \hat{\psi})/\partial\theta^{\top}\,|_{\theta=\hat{\theta}}, \qquad \hat{\Omega} = N^{-1}\sum_{i=1}^{N} \hat{\nu}_i(\hat{\theta}, \hat{\alpha}, \hat{\psi})^{\otimes 2},$$

$$\hat{\nu}_i(\hat{\theta}, \hat{\alpha}, \hat{\psi}) = \bar{S}_{\theta,i}(\hat{\theta}, \hat{\alpha}, \hat{\psi}) - \hat{\delta}\hat{I}_{\alpha}^{-1} S_{\alpha,i}(\hat{\alpha}) - \hat{\kappa}\hat{I}_{\psi}^{-1} S_{\psi,i}^{\text{obs}}(\hat{\psi}),$$

$$\hat{\delta} = -N^{-1}\sum_{i=1}^{N} \partial \bar{S}_{\theta,i}(\hat{\theta}, \alpha, \hat{\psi})/\partial\alpha^{\top}\,|_{\alpha=\hat{\alpha}},$$

$$\hat{I}_{\alpha} = -N^{-1}\sum_{i=1}^{N} \partial S_{\alpha,i}(\alpha)/\partial\alpha^{\top}\,|_{\alpha=\hat{\alpha}},$$

$$\hat{\kappa} = -(NM)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M} S_{\theta,i}^{(j)}(\hat{\theta}, \hat{\alpha}, \hat{\psi}) S_{\psi,i}^{\text{mis},(j)}(\hat{\psi})^{\top},$$

$$\hat{I}_{\psi} = -N^{-1}\sum_{i=1}^{N} \partial S_{\psi,i}^{\text{obs}}(\psi)/\partial\psi^{\top}\,|_{\psi=\hat{\psi}},$$

$$S_{\psi,i}^{\text{mis},(j)}(\hat{\psi}) = \partial \log g\big(D_i^{m,(j)} | D_i^{o}, D_i^{\dagger}, \mathcal{R}_i = 1; \psi\big)/\partial\psi\,|_{\psi=\hat{\psi}}.$$

Consistency follows from Slutsky's theorem.

3.5. *Code*. We provide code in `R` for implementing the proposed methods. The code makes use of the `jacobian` function in the `numDeriv` package for calculating the derivatives needed for $\hat{\delta}$, $\hat{\kappa}$, $\hat{I}_{\alpha}$ and $\hat{I}_{\psi}$. The code can accommodate any weighting model as long as the user can specify the score equation $S_{\alpha}(\alpha, H_i)$ and function for the weights $W(\hat{\alpha}; H_i)$. The code can modified by the end user to accommodate any parametric imputation model. Code and a corresponding tutorial can be found in the Supplementary Material Thaweethai et al. (2021) and online at https://github.com/tthaweethai/robustipwmi.

4. A worked example. We consider a detailed, worked example where we are interested in estimating the coefficients of a linear regression of outcome Y on a set of covariates X. This implies that $D = \{X, Y\}$. There is missingness in both X and Y, and so we choose to define $\mathcal{R}(R) = 1$ when X is fully observed and 0 otherwise and then use a logistic regression with predictors H to estimate the probability of fully observing X, where H is fully observed. We then use the inverse of those probabilities as weights to account for exclusion of individuals with any missingness in X. The only missingness that remains in D, once we restrict to $\mathcal{R} = 1$ is in Y which implies that $D_{\mathcal{R}}^{m} = Y$ and $D_{\mathcal{R}}^{o} = X$. We identify an additional set of variables that are fully observed among individuals with $\mathcal{R} = 1$ to include in the imputation model, which we define as $D^{\dagger}$. Collectively, we refer to $\{D_{\mathcal{R}}^{o}, D^{\dagger}\}$ as Z, and so our imputation model is a linear regression model with outcome Y, predictors Z and homoskedastic errors. We use this model to impute missing Y among individuals with X fully observed.

An individual's contribution to the estimating equation for estimating α is given by the score for the weighting model:

$$S_{\alpha}(\alpha; H_i) = S_{\alpha,i}(\alpha) = H_i\big[\mathcal{R}_i - \exp(\alpha^{\top}H_i)\{1 + \exp(\alpha^{\top}H_i)\}^{-1}\big].$$

We solve the estimating equations $\sum_{i=1}^{N} S_{\alpha,i}(\alpha) = 0$ to obtain $\hat{\alpha}$, from which we can generate the probabilities used for weighting: $\text{pr}(\mathcal{R}_i = 1 | H_i; \hat{\alpha}) = \exp(\hat{\alpha}^{\top}H_i)[1 + \exp(\hat{\alpha}^{\top}H_i)]^{-1}$.

We define $R_{Y,i}$ to be an indicator of whether the ith individual has Y observed in addition to X. Therefore, if $R_{Y,i} = 1$, then it is necessarily true that $\mathcal{R}_i = 1$. We fit the following imputation model among individuals with $R_{Y,i} = 1$, indexed by parameter ψ, which we partition into $\psi = (\beta^\top, \sigma)^\top$:

$$Y_i = \beta^\top Z_i + \epsilon_i, \quad \text{where } \epsilon_i \sim N(0, \sigma).$$

The contribution of individual i to the estimating equation for the imputation model is

$$\begin{aligned} S_\psi^{\text{obs}}(\psi; Y_i, Z_i) = S_{\psi,i}^{\text{obs}}(\psi) &= \begin{bmatrix} S_\beta^{\text{obs}}(\psi; Y_i, Z_i) \\ S_\sigma^{\text{obs}}(\psi; Y_i, Z_i) \end{bmatrix} \\ &= \begin{bmatrix} R_{Y,i}\sigma^{-2} Z_i (Y_i - \beta^\top Z_i) \\ R_{Y,i}\{-\sigma^{-1} + \sigma^{-3}(Y_i - \beta^\top Z_i)^2\} \end{bmatrix}. \end{aligned}$$

We solve the estimating equations $\sum_{i=1}^N S_{\psi,i}^{\text{obs}}(\psi) = 0$ to obtain $\hat{\psi} = (\hat{\beta}^\top, \hat{\sigma})^\top$. For each of M imputations, following the improper imputation, we impute $Y_i^{(j)}$ for individuals with $\mathcal{R}_i = 1$ but $R_{Y,i} = 0$ as follows:

$$Y_i^{(j)} = \hat{\beta}^\top Z_i + \tilde{\epsilon}_i^{(j)},$$

where $\tilde{\epsilon}_i^{(j)}$ is a single draw from a normal distribution with mean 0 and standard deviation $\hat{\sigma}$. For individuals with $\mathcal{R}_i = R_{Y,i} = 1$, $Y_i^{(j)} = Y_i$ (i.e., their observed value). Finally, the contribution of individual i in the jth imputed dataset to the weighted estimating equations is

$$S_\theta(\theta, \alpha, \psi; D_i^{(j)}, H_i) = S_{\theta,i}^{(j)}(\theta, \alpha, \psi) = \mathcal{R}_i \text{pr}(\mathcal{R}_i = 1 | H_i; \hat{\alpha})^{-1} X_i (Y_i^{(j)} - \theta^\top X_i).$$

We solve the estimating equations $\sum_{i=1}^N \sum_{j=1}^M S_{\theta,i}^{(j)}(\theta, \hat{\alpha}, \hat{\psi}) = 0$ to obtain $\hat{\theta}$. To estimate $\text{var}(\hat{\theta})$, we use the proposed estimator in Section 3.4, the inputs of which evaluate as follows:

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^N \mathcal{R}_i \text{pr}(\mathcal{R}_i = 1 | H_i; \hat{\alpha})^{-1} X_i X_i^\top,$$

$$\begin{aligned} \hat{I}_\psi &= -\frac{1}{N} \sum_{i=1}^N \begin{bmatrix} \partial S_{\beta,i}^{\text{obs}}(\psi)/\partial\beta^\top & \partial S_{\beta,i}^{\text{obs}}(\psi)/\partial\sigma \\ \partial S_{\sigma,i}^{\text{obs}}(\psi)/\partial\beta^\top & \partial S_{\sigma,i}^{\text{obs}}(\psi)/\partial\sigma \end{bmatrix}_{\psi=\hat{\psi}} \\ &= \frac{1}{N} \sum_{i=1}^N R_{Y,i} \begin{bmatrix} \hat{\sigma}^{-2} Z_i Z_i^\top & 2\hat{\sigma}^{-3} Z_i (Y_i - \hat{\beta}^\top Z_i) \\ 2\hat{\sigma}^{-3}(Y_i - \hat{\beta}^\top Z_i) Z_i^\top & -\hat{\sigma}^{-2} + 3\hat{\sigma}^{-4}(Y_i - \hat{\beta}^\top Z_i)^2 \end{bmatrix}, \end{aligned}$$

$$\hat{\delta} = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M \mathcal{R}_i \exp(\hat{\alpha}^\top H_i)^{-1} X_i (Y_i^{(j)} - \hat{\theta}^\top X_i) H_i^\top,$$

$$\hat{I}_\alpha = \frac{1}{N} \sum_{i=1}^N H_i \exp(\hat{\alpha}^\top H_i) [1 + \exp(\hat{\alpha}^\top H_i)]^{-2} H_i^\top,$$

$$\begin{aligned} S_\psi^{\text{mis},(j)}(\psi; Y_i^{(j)}, Z_i) &= S_{\psi,i}^{\text{mis},(j)}(\psi) \\ &= \begin{bmatrix} (1 - R_{Y,i})\sigma^{-2} Z_i (Y_i^{(j)} - \beta^\top Z_i) \\ (1 - R_{Y,i})(-\sigma^{-1} + \sigma^{-3}(Y_i^{(j)} - \beta^\top Z_i)^2) \end{bmatrix} \end{aligned}$$

and the remaining inputs are sums of products of scores or other inputs defined above.

5. Simulations.

5.1. *Set-up.* We conduct a simulation study modeled after the simulation study described in Section 4 of Seaman et al. (2012) but introduce an additional level of model misspecification to demonstrate the robustness properties of the proposed variance estimator. We consider two main simulation settings: one with homoskedastic errors in the outcome variable, precisely mirroring Seaman et al. (2012), and one with heteroskedastic errors in the outcome variable which was introduced in Robins and Wang (2000) as a setting under which Rubin's rules are biased for variance estimation.

For $N = 1000$ individuals, we generated the following covariates: X_1 is 1 with probability 0.5 and 0 otherwise, (X_2, X_3, X_4) are independent and identically distributed normal random variables with mean 0 and standard deviation 1 and X_5 is normally distributed with mean $X_2 \times X_3$ and standard deviation 1. Under homoskedastic errors, response Y is generated from a normal distribution with mean $-3 + X_1X_2 + X_1X_3 + 0.5X_2X_3 + X_4 + 0.5X_5$ and standard deviation 1. Under heteroskedastic errors, Y is generated from a normal distribution with the same mean but with standard deviation 1 if $X_1 = 0$ and 2 if $X_1 = 1$.

X_1 is observed for all individuals, but (X_2, X_3, X_4, X_5) is only observed with probability $0.8 - 0.6X_1$ and is otherwise missing. Given that (X_2, X_3, X_4, X_5) is observed, Y is observed with probability $[1 + \exp(-1.5 + 0.6X_2X_4)]^{-1}$ and is otherwise missing. Under this missingness mechanism, approximately 50% of individuals have (X_2, X_3, X_4, X_5) observed, and, approximately, 40% have (X_2, X_3, X_4, X_5, Y) observed.

The analysis model is a linear regression assuming homoskedastic errors: $Y = \theta_0 + \theta_2X_2 + \theta_3X_3 + \theta_{23}X_2X_3 + e$, where $E(e|X_2, X_3) = 0$. As in Seaman et al. (2012), upon integrating over the distribution of X_1, X_4 and X_5, it can be shown that the analysis model is correctly specified in the absence of missing data and the true value of the parameters induced by this model is $(\theta_0, \theta_2, \theta_3, \theta_{23})=(-3, 0.5, 0.5, 1)$. In the heteroskedastic errors setting, however, the analysis model is misspecified, since it assumes homoskedasticity.

5.2. *Analysis approaches.* We compare three strategies for estimating θ:

(i) use only complete cases,

(ii) combine IPW and MI using a correctly specified weighting model and an imputation model that contains all interaction terms such that it is correctly specified in the homoskedastic error setting and

(iii) combine IPW and MI as in strategy (ii), but using an imputation model that omits certain interactions such that it is misspecified under both homoskedastic and heteroskedastic errors.

To combine inverse probability weighting and multiple imputation, we mirror the baseline approach of Seaman et al. (2012) and define the criteria $\mathcal{R}(R) = 1$ when (X_2, X_3, X_4, X_5) is observed and impute missing Y among individuals with $\mathcal{R}(R) = 1$. We define $H = (1, X_1)^\top$ to predict whether $\mathcal{R}_i = 1$ and use a correctly specified linear probability model for the weights, $\text{pr}(\mathcal{R} = 1|H) = \alpha_0 + \alpha_1X_1$, which implies $S_\alpha(\alpha; H_i) = H_i(\mathcal{R}_i - \alpha^\top H_i)$ and $\text{pr}(\mathcal{R}_i = 1|H_i; \hat{\alpha}) = \hat{\alpha}^\top H_i$. While using a linear probability model is an unusual choice, there is no risk of estimating negative weights because the only predictor, X_1, is binary. In Section 4 we demonstrate how a logistic regression can be used for the weighting model and provide full expressions for the particular combination of models used in this simulation study in the Supplementary Material, Section 2.2. For strategy (ii), we use an imputation model that assumes homoskedastic errors: $Y = \beta^\top Z + \epsilon$, where $\epsilon \sim N(0, \sigma)$ and $Z = (1, X_1, X_2, X_3, X_4, X_5, X_1X_2, X_1X_3, X_2X_3, X_1X_2X_3)^\top$. $S^{\text{obs}}_{\psi,i}(\psi)$ is defined as in Section 4. For strategy (iii), we repeat strategy (ii) but use $Z^\dagger$ in the place of Z, where $Z^\dagger = Z \setminus \{X_2X_3, X_1X_2X_3, X_5\}$

TABLE 2
Model specification under different analysis strategies

	Model predictors		Homoskedastic errors		Heteroskedastic errors	
Strategy	Imputation	Analysis	Imputation	Analysis	Imputation	Analysis
(i) CC	N/A	X^*	N/A	✓	N/A	×
(ii) IPW/MI	Z	X^*	✓	✓	×	×
(iii) IPW/MI	$Z^\dagger$	X^*	×	✓	×	×

The analysis approaches defined in Section 5 correspond to different combinations of predictors for the imputation and analysis models. Under different error structures (homoskedastic vs. heteroskedastic), each model is either correctly specified (✓) or misspecified (×). For IPW/MI, the weighting model is always correctly specified

$$Z = (1, X_1, X_2, X_3, X_4, X_5, X_1X_2, X_1X_3, X_2X_3, X_1X_2X_3)^\top,$$
$$Z^\dagger = (1, X_1, X_2, X_3, X_4, X_1X_2, X_1X_3)^\top, \qquad X^* = (1, X_2, X_3, X_2X_3)^\top.$$

which, as explained in Seaman et al. (2012), causes θ_{23} to be underestimated by approximately 60%. We summarize the consequences of each choice of model in each strategy on model misspecification in Table 2.

All simulations were performed in `R`. For strategies (ii) and (iii) we compare three imputation techniques: improper imputation, proper imputation and Multiple Imputation by Chained Equations, or MICE (van Buuren and Groothuis-Oudshoorn (2011)). Proper imputation was performed using the method of Schenker and Welsh (1988). MICE was performed with the `MICE` package in `R`, using the `formulas` argument to specify interactions in the imputation model. This method was included as a comparison, as it is one of the most frequently used tools for performing MI in practice.

To estimate standard errors, we use Rubin's rules for all three imputation techniques but use the proposed variance estimator of Section 3.4 for only improper and proper imputation. Unlike improper imputation, the proper imputation estimator does not use the single preliminary estimate $\hat{\psi}$ obtained from the observed data to generate the M imputed datasets. Instead, it uses a different quantity $\hat{\psi}^{(j)}$ for each imputed dataset, where each $\hat{\psi}^{(j)}$ is drawn from the posterior density of ψ given the observed data. Since the proposed estimator in Section 3.4 assumes an improper imputation procedure, we explain in the Supplementary Material (Section 2.1) how the proposed estimator was adapted for proper imputation. We cannot use the proposed variance estimation procedure, when MICE is used, because we do not know the value of $\hat{\psi}^{(j)}$ that was used to generate each imputed dataset.

We generated 10,000 datasets and calculated $\hat{\theta}$ under each estimation strategy, the empirical standard error of $\hat{\theta}$ (defined as the standard deviation of the parameter estimates) under each strategy and the average estimated standard error using each variance estimation method. We then calculated the bias of the parameter estimates $\hat{\theta}$ compared to the truth and the bias of each standard error estimation method compared to the empirical standard error. $M = 10$ imputations were performed. For every analysis (besides the complete case analysis), we estimated θ two ways: first, by combining all M quasi-complete datasets into a single dataset and fitting the analysis model once, and, second, by fitting the analysis model in each of M imputed datasets and averaging the point estimates.

5.3. *Results*. Across all simulations, the point estimates obtained by combining all M datasets into a single dataset and fitting the analysis model once were practically identical to the point estimates obtained by fitting the analysis model once for each imputed dataset and averaging the point estimates. The largest absolute difference between these two quantities for any single dataset was less than 1.0×10^{-13}.

TABLE 3
Percent bias in point estimates for θ, $N = 1000$

Strategy	Imputation method	Homoskedastic errors				Heteroskedastic errors			
		θ_0	θ_2	θ_3	θ_{23}	θ_0	θ_2	θ_3	θ_{23}
(i) CC	N/A	0.0	−82.7	−60.1	0.0	0.0	−82.6	−60.1	0.1
(ii) IPW/MI	Improper	0.0	−1.4	−1.4	0.3	0.1	−1.3	−1.4	0.3
	Proper	0.0	−1.4	−1.4	0.3	0.1	−1.3	−1.5	0.4
	MICE	0.2	−4.6	−3.1	−2.5	0.2	−4.7	−3.2	−2.5
(iii) IPW/MI	Improper	0.0	−1.6	−1.2	−22.2	0.0	−1.5	−1.3	−22.1
	Proper	0.0	−1.6	−1.3	−22.2	0.1	−1.6	−1.4	−22.1
	MICE	−0.3	−5.3	−3.9	−21.0	−0.3	−5.3	−4.0	−21.0

Percent bias in parameter estimate is calculated by $(\hat{\theta} - \theta)/\theta \times 100$, where θ is the true value under the induced marginal model. The strategies are introduced in Section 5.2 and summarized in Table 2.

Table 3 provides the percent bias in the analysis model point estimates for θ, compared to the true value of θ. $\hat{\theta}_2$ and $\hat{\theta}_3$ were approximately −83% and −60% biased, respectively, using a complete case analysis (strategy i) under both error structures.

When improper and proper imputation were performed using an imputation model that included all key interactions (strategy ii), combining IPW and MI corrected the bias in these estimates (between −1.5% and 0.4% bias) under both error structures. MICE was slightly biased (between −4.7% and 0.2% bias). Notably, even though this imputation model is misspecified under heteroskedastic errors, it was still able to generate unbiased estimates of θ. When key interactions in the imputation model were excluded, combining inverse probability weighting with multiple imputation resulted in substantial bias in $\hat{\theta}_{23}$ (approximately 0.1% bias to between −21.0% and −22.2% bias) under both error structures.

Figure 1 shows how the proposed variance estimator compares to Rubin's rules for estimating the empirical standard error. Full numerical results and calculated percent biases can be found in the Supplementary Material (Table 1). Under homoskedastic errors (panels a and c, Figure 1), both inferential methods approximate the empirical standard error roughly equally well, even when the imputation model omits some interactions (strategy ii; between −7.3% and 2.8% bias for Rubin's rules and between −7.2% and −1.6% bias for the proposed estimator). As shown in panels (b) and (d) of Figure 1, under heteroskedastic errors, for parameters (θ_0, θ_2, θ_3), Rubin's rules substantially underestimates the empirical standard error (between −11.5% and −5.8% bias), while the proposed variance estimation procedure is considerably less biased (between −3.7% and −2.4% bias). As shown in panel (b), in the presence of heteroskedastic errors, the proposed variance estimation procedure for θ_{23} reduced the bias when all key interactions were included in the imputation model (−14.2% to −7.5% and −12.3% to −7.4% bias for improper and proper imputation, respectively). When the imputation model was missing some interactions, as shown in panel (d), Rubin's rules performed slightly better than the proposed variance estimation procedure at estimating the standard error for θ_{23} (−6.8% bias for the proposed method vs. between 1.2% and 2.0% for Rubin's rules). Despite containing all key interactions, the imputation model used in panel (b) was still misspecified due to the error structure which explains why the proposed variance estimator was less biased than Rubin's rules for all four parameters.

MICE was sometimes slightly more efficient (i.e., had smaller empirical standard errors) than improper or proper imputation which were approximately equally efficient. However,

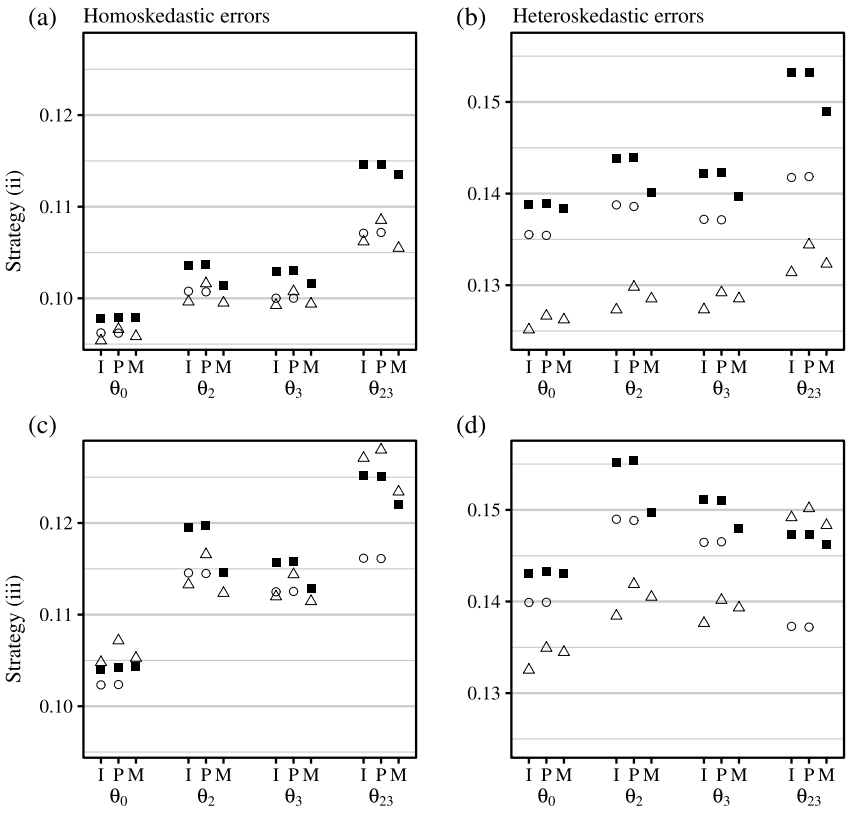

FIG. 1. *Each panel represents the combination of a particular analysis strategy* (*rows*) *and simulation error structure* (*columns*), *as defined in Section* 5. *In each panel and for each component of* θ *obtained following combining inverse probability weighting with multiple imputation*, *we consider three different imputation methods*: *improper* (*I*), *proper* (*P*) *and MICE* (*M*). *Empirical standard errors* (*black squares*) *are plotted vertically against average estimated standard errors using two methods*: *Rubin's rules* (*white triangles*) *and the proposed variance estimation procedure* (*white circles*). *Only Rubin's rules were used to estimate standard errors when MICE was used. Results are based on* 10,000 *simulations with sample size* 1000. *While the vertical axes are different in the two rows of panels*, *the scales are the same.*

Rubin's rules still generally underestimated the standard error when MICE was used under heteroskedastic errors.

When the sample size was increased from 1000 to 10,000, the proposed variance estimation method was uniformly better at estimating the empirical standard error compared to Rubin's rules (Supplementary Material, Figure 1), thus resolving the observed phenomenon that Rubin's rules were less biased than the proposed variance estimator for $\hat{\theta}_{23}$ in panel (d). That result is largely due to the small sample size in the original simulation study, as this was not observed when the sample size was increased to 10,000. This is likely due to the fact that the proposed variance estimator was derived under asymptotic conditions. Full simulation results, including coverage probabilities, can be found in the Section 3 of the Supplementary Material. We performed an additional simulation study where we considered misspecification of the weighting model in addition to the imputation and analysis models, the results of which can be found in Supplementary Material Section 5. The variance estimation procedure was observed to be robust to the weighting, imputation and analysis models, though the additional robustness to the weighting model was expected, since a sandwich-type estimator was used to account for the weighting.

TABLE 4
Distribution of baseline covariates by surgical type

Variable	Roux-en-Y gastric bypass (RYGB)	Vertical sleeve gastrectomy (VSG)
Overall count (%)	7213 (76.6%)	2198 (23.4%)
Healthcare system		
A	653 (9.1%)	10 (0.5%)
B	4046 (56.1%)	1757 (79.9%)
C	2514 (34.9%)	431 (19.6%)
Sex		
Female	5882 (81.5%)	1776 (80.8%)
Male	1331 (18.5%)	422 (19.2%)
Race/Ethnicity		
Black	951 (13.2%)	452 (20.6%)
Hispanic	2056 (28.5%)	691 (31.4%)
White	3753 (52.0%)	977 (44.4%)
Other	453 (6.3%)	78 (3.5%)
Year of surgery		
2008	2380 (33.0%)	189 (8.6%)
2009	2359 (32.7%)	563 (25.6%)
2010	2474 (34.3%)	1446 (65.8%)
Age at time of surgery		
39 or less	2248 (31.2%)	703 (32.0%)
40 to 50	2399 (33.3%)	703 (32.0%)
51 or more	2566 (35.6%)	792 (36.0%)

6. Application of proposed framework in a study of bariatric surgery and chronic kidney disease.

6.1. *Study population.* We consider 9411 patients in the DURABLE study cohort, described in Section 2, who underwent either RYGB (76.6%) or VSG (23.4%) between 2008 and 2010. For the purposes of minimizing patient reidentification, we henceforth mask the names of the three healthcare systems in which patients received care randomly as A, B and C. Demographic information for the patients included in the study are found in Table 4.

6.2. *Analysis model and missing data.* For this analysis we are interested in studying whether the difference in weight loss between different surgical techniques is the same when considering whether an individual has chronic kidney disease before surgery. To study this, we consider the outcome of percent total weight change (PTWC) after bariatric surgery, defined as BMI at follow-up minus BMI at baseline divided by BMI at baseline. We use this outcome because PTWC has been found to be less confounded by baseline BMI than other commonly reported measures, such as percent excess weight loss (Hatoum and Kaplan (2012), Maciejewski et al. (2016)). BMI at baseline is defined as the most recent BMI measurement recorded in the 30 days leading up to and including the day of surgery. BMI at follow-up is considered to be observed if there is at least one BMI measurement less than six months before or after five years after the day of surgery. If there are multiple BMI measures in that time window, we fit a linear regression to those BMI measurements as a function of time and use that to obtain a "fitted" BMI at five years. Otherwise, we used the BMI measure closest to the five-year time point. Usually, this has very little impact compared to taking the BMI measure closest to five years, but there are very occasionally erratic BMI measurements very different from surrounding measures but happen to be the closest to the five-year time point.

Further details on this procedure can be found in the Supplementary Material, Section 7. Using these definitions, 10.8% and 32.1% of patients in the study were missing baseline and follow-up BMI, respectively.

The substantive model of interest is a linear regression with outcome PTWC and the following predictors, with the reference categories listed last: surgery type (RYGB/VSG), site (A/C/B), sex (male/female), age category at time of surgery (39 or less/51 or more/40 to 50), year of surgery (2008/2009/2010), race/ethnicity (Hispanic/White/Other—including Hawai'ian or Pacific Islander, Multiple, Asian, Native American or Alaskan Native, Other, and Unknown/Black), baseline BMI category in kg/m^2 (39 or less/45 to 49/50 to 54/55 to 59/60 or greater/40 to 44), baseline Charlson–Elixhauser combined comorbidity score, as defined by Gagne et al. (2011) (less than 0/greater than 0/0), baseline chronic kidney disease (1 if present/0 if absent) and the interaction between surgical type and baseline chronic kidney disease.

The Charlson–Elixhauser combined comorbidity score is calculated for a given time interval by determining the presence or absence of 20 conditions, defined by specific ICD-9 codes. The presence of each condition (counted only once) is associated with a weight, and the weights for the present conditions are summed to obtain an overall score. While the absence of evidence in EHR for any comorbidities may be due to the true absence of any comorbidities (and, therefore, a score of 0), it may be due to missing data. A proxy method we propose for determining whether the combined comorbidity score can be calculated for the six months leading up to surgery will be by examining whether BMI or any comorbid conditions are recorded in three time windows: zero to two, two to four and four to six months before surgery. The implicit assumption is that, if there is at least some BMI or comorbid condition recorded in the EHR during those periods, any other comorbid conditions in that six month period would also be reliably captured; otherwise, it is missing. Using this definition, 27.5% of patients are missing baseline combined comorbidity scores.

One condition included in the combined comorbidity score is a diagnosis of chronic kidney disease, which includes the following ICD-9 codes: 403.11, 403.91, 404.12, 404.92, 585.x, 586.x, V42.0, V45.1, V56.0 or V56.8. If any of these ICD-9 codes is present in the six months before surgery, the patient is considered to have baseline chronic kidney disease. If there is no evidence of chronic kidney disease and the combined comorbidity score is not missing as defined previously, then the patient categorized as not having chronic kidney disease. If there is no evidence of chronic kidney disease and the combined comorbidity score is missing, then we say that chronic kidney disease status at baseline is missing. In other words, if someone is missing their combined comorbidity score, their chronic kidney disease status is also missing, unless there is evidence of chronic kidney disease at any point in the six months before surgery. Using this definition, 26.6% of patients are missing chronic kidney disease status at baseline.

6.3. *Approach for addressing missingness.* The analysis model can only be fit among patients with complete data for BMI at baseline, follow-up, combined comorbidity score, chronic kidney disease status, surgical type and all potential confounders (site, sex, age at time of surgery, year of surgery and race/ethnicity category). We define this set of variables as D, using the notation of Section 3. There is missingness in baseline and follow-up BMI, combined comorbidity score and chronic kidney disease status.

We combine IPW and MI in this setting by first defining the inclusion rule $\mathcal{R}$ as equal to 1 when baseline BMI, combined comorbidity score and chronic kidney disease status are all measured. We fit a logistic regression model with outcome $\mathcal{R}$ and predictors H which includes the potential confounders listed previously and surgical type. We then multiply impute missing values of follow-up BMI (i.e., $D^m_{\mathcal{R}}$) using a linear regression imputation model

TABLE 5

Distribution of baseline covariates and BMI by missingness pattern. $\mathcal{R}$ is an indicator of whether baseline BMI, combined comorbidity score and chronic kidney disease status are all observed

Variable	$\mathcal{R}=0$	$\mathcal{R}=1$, follow-up BMI missing	$\mathcal{R}=1$, follow-up BMI measured
Overall count (%)	3140 (33.4%)	1888 (20.1%)	4383 (46.6%)
	Counts (%), by missingness pattern		
Surgery type			
RYGB	2530 (80.6%)	1471 (77.9%)	3212 (73.3%)
VSG	610 (19.4%)	417 (22.1%)	1171 (26.7%)
Healthcare system			
A	296 (9.4%)	130 (6.9%)	237 (5.4%)
B	1810 (57.6%)	1123 (59.5%)	2870 (65.5%)
C	1034 (32.9%)	635 (33.6%)	1276 (29.1%)
Sex			
Female	2501 (79.6%)	1522 (80.6%)	3635 (82.9%)
Male	639 (20.4%)	366 (19.4%)	748 (17.1%)
Race/Ethnicity			
Black	468 (14.9%)	228 (12.1%)	707 (16.1%)
Hispanic	905 (28.8%)	536 (28.4%)	1306 (29.8%)
White	1530 (48.7%)	984 (52.1%)	2216 (50.6%)
Other	237 (7.5%)	140 (7.4%)	154 (3.5%)
Year of surgery			
2008	930 (29.6%)	516 (27.3%)	1123 (25.6%)
2009	952 (30.3%)	507 (26.9%)	1463 (33.4%)
2010	1258 (40.1%)	865 (45.8%)	1797 (41.0%)
Age at surgery			
39 or less	1024 (32.6%)	712 (37.7%)	1215 (27.7%)
40 to 50	1063 (33.9%)	602 (31.9%)	1437 (32.8%)
51 or more	1053 (33.5%)	574 (30.4%)	1731 (39.5%)
Combined comorbidity score			
≤ -1	–	388 (20.6%)	895 (20.4%)
0	–	797 (42.2%)	1757 (40.1%)
≥ 1	–	703 (37.2%)	1731 (39.5%)
Chronic kidney disease status			
Yes	–	91 (4.8%)	277 (6.3%)
No	–	1797 (95.2%)	4106 (93.7%)
	Means (standard deviation), by missingness pattern		
Baseline BMI (kg/m^2)	–	45.2 (7.6)	44.4 (7.3)
Follow-up BMI (kg/m^2)	–	–	34.6 (7.1)

that conditions on variables in D that are fully observed for individuals with $\mathcal{R}=1$ ($D^o_{\mathcal{R}}$), which includes H as well as baseline BMI, combined comorbidity score and chronic kidney disease status. Missing follow-up BMI among patients with $\mathcal{R}=1$ is improperly imputed $M=50$ times using this model. We compare this approach to a complete case analysis and then compare the proposed estimation approach for the standard errors with Rubin's rules. The analysis cohort, stratified by missingness pattern, is described in Table 5.

6.4. *Results*. Table 6 includes the results of the complete case analysis compared to the proposed analytic approach.

TABLE 6

Regression coefficients and associated standard errors for a linear regression on percent total weight change between baseline and five years under different missing data strategies

	Complete cases		IPW/MI			
	Coef.	Std. err.	Coef.	Std. err. (proposed)	Std. err. (Rubin)	Pct. diff.
(Intercept)	−15.86	0.64	−15.78	0.65	0.69	−5.28
Surgery type: RYGB	−6.64	0.42	−6.74	0.43	0.44	−1.66
Chronic kidney disease	−3.03	1.56	−3.48	1.62	1.52	6.16
RYGB × CKD	3.02	1.78	3.02	1.83	1.74	4.84
Health System: A	−2.30	0.74	−1.83	0.78	0.76	2.32
Health System: C	1.40	0.42	1.45	0.43	0.45	−4.91
Sex: Male	3.00	0.46	3.07	0.47	0.47	−1.09
Age: ≤39	−1.29	0.44	−1.30	0.45	0.47	−4.37
Age: ≥51	0.59	0.40	0.68	0.41	0.45	−8.50
Race: White	−1.81	0.50	−2.06	0.52	0.53	−2.80
Race: Hispanic	−1.32	0.50	−1.62	0.52	0.54	−3.84
Race: Other	−2.40	1.00	−2.62	1.02	0.96	5.96
Year of surgery: 2008	0.35	0.45	0.48	0.46	0.45	1.65
Year of surgery: 2009	−0.47	0.40	−0.37	0.40	0.42	−3.54
Baseline BMI: <40	1.26	0.45	1.38	0.46	0.50	−8.60
Baseline BMI: 45–49	−1.24	0.46	−1.22	0.46	0.47	−2.35
Baseline BMI: ≥50	−1.58	0.50	−1.24	0.52	0.47	9.32
Comorbidity score: ≤ −1	0.24	0.43	0.18	0.45	0.46	−2.14
Comorbidity score: ≥1	−0.24	0.40	−0.05	0.41	0.40	0.84

Coef.: Regression coefficient, Std. err.: standard error, Pct. diff: [Std. err. (proposed)—Std. err. (Rubin)]/Std. err. (Rubin) ×100. Reference categories: Surgery type (VSG), Chronic kidney disease status (0), health system (B), sex (Female), age category (40–50), race (Black), year of surgery (2010), baseline BMI category (40–44), and combined comorbidity score (0). CKD refers to chronic kidney disease.

Adjusting for the variables included in the substantive model, patients who underwent RYGB had greater percent total weight loss after five years than patients who underwent VSG. Generally, it appears that the main effect of baseline chronic kidney disease is equal in magnitude but opposite in direction to the interaction of RYGB with baseline chronic kidney disease. We can interpret this to mean that, among patients undergoing VSG, those with chronic kidney disease experienced greater percent weight loss than those without, but among those undergoing RYGB, those with and without chronic kidney disease experienced similar percent weight loss. An alternative interpretation is that the difference in percent weight loss comparing RYGB to VSG was smaller among those with chronic kidney disease at baseline compared to those without. Compared to the complete case analysis, IPW and MI did not dramatically shift the conclusions one might draw regarding the associations of interest in this study, as the standard errors were generally large compared to the differences in the point estimates.

6.5. *Robust variance estimation.* When combining IPW with MI, we consider two methods for estimating standard errors: Rubin's rules and the proposed robust variance estimator. Standard error estimates were generally larger for IPW/MI compared to the complete case analysis. Standard error estimates obtained from the proposed method, compared to Rubin's rules, were not uniformly larger or smaller, though for the effects of interest (main effect of chronic kidney disease and the interaction of chronic kidney disease with surgery type) the proposed standard error estimate was slightly larger.

7. Discussion. Combining IPW and MI using the Seaman et al. (2012) approach requires specification of three models: a weighting model used for IPW, an imputation model used for MI and an analysis model. Asymptotically unbiased point estimation and variance estimation via Rubin's rules are guaranteed when these three models are correctly specified. While we do not guarantee that point estimates obtained using the proposed framework are asymptotically unbiased when one or more of these models are misspecified, we provide an alternative procedure based on improper imputation that permits asymptotically valid variance estimation that is robust to misspecification of these models, based on Robins and Wang (2000).

Although not explored in this paper, the proposed method for performing inference is also robust to uncongeniality between the imputation and analysis models (Meng (1994)). Such a situation can arise when the amount of information used to fit the imputation model and generate imputations differs greatly from the analysis model, particularly in the presence of interactions in one but not both models. The validity of Rubin's rules in this setting has been the topic of much debate (Kim et al. (2006), Meng and Romero (2003), Robins and Wang (2000), Tang (2017)). While the Robins and Wang (2000) approach offers protection against uncongeniality, multiple authors have pointed to the mathematical complexity of this approach, exacerbated by the lack of available software for implementation (Hughes, Sterne and Tilling (2016), Seaman, White and Leacy (2014)). We provide software in `R` in Section 3.5 to overcome this hurdle and assist analysts in implementing the proposed method. If the analyst does not use the provided software, we note that improper imputation is computationally more straightforward to perform compared to proper imputation and that improper imputation was not found to be any less efficient than proper imputation in our simulations.

During the review process, one reviewer queried the utility of the point estimate, $\hat{\theta}$, if it is not guaranteed to converge to θ; why would one care about inference for settings where the analysis model is misspecified? While a reasonable query, we take the pragmatic position that analysis models are likely almost always misspecified and, indeed, sometimes purposely so. Moreover, the analysis model is generally specified in relation to the motivating research question, where interest can lie in estimating marginal effects rather than modeling the precise data generating mechanism. For example, in the application described in this paper we have chosen to focus on an interaction effect with respect to one particular comorbidity (chronic kidney disease) while ignoring other potential interactions with other comorbidities included in the Charlson–Elixhauser comorbidity score. A model that adjusts for all potential interactions in the interest of creating a model that is most likely to adhere to the true data generating mechanism would likely result in standard errors that are too large to be useful. There are also many patient characteristics and behaviors that are not routinely recorded in EHR that are likely to impact the magnitude of weight loss among patients who undergo bariatric surgery, including diet and exercise behaviors (Arterburn et al. (2013), Courcoulas et al. (2015)). Omission of these factors would contribute to model misspecification, despite our best efforts. Alternatively, one might be uninterested in modeling or unable to model the precise functional form of covariate associations in a model, choosing to use first-order associations rather than those that are higher order. Even though such a model might be misspecified, it can still be worthwhile to obtain appropriate confidence intervals for point estimates that are obtained.

Further, in the simulation study in Section 5 we observed that asymptotically unbiased point estimates were obtained despite misspecification of both the imputation and analysis models (i.e., $\hat{\theta}$ did, in fact, converge to θ). When an imputation model with all interactions (strategy ii) was used in the heteroskedastic errors setting, the resulting point estimates were unbiased even though the imputation and analysis models were misspecified, as they both assumed homoskedastic errors. The use of Rubin's rules resulted in biased variance estimation

whereas the proposed variance estimation procedure reduced that bias. When $N = 10,000$, that bias was eliminated entirely under the proposed estimation procedure, as shown in the Supplementary Material Figure 1.

Practically, if the analyst is very concerned that the weighting and/or imputation models are so severely misspecified such that the point estimates would not be useful, the analyst can consider other strategies for handling missing data, such as doubly-robust methods, or semiparametric or nonparametric imputation procedures, such as Bayesian additive regression trees (Xu, Daniels and Winterstein (2016)). However, the validity of doubly-robust methods has been shown to be highly sensitive to minor misspecification of both models, so caution must be taken (Kang and Schafer (2007), Seaman and Vansteelandt (2018)). If the analyst decides to proceed with the general framework Seaman et al. (2012) introduced and combine IPW with MI, an analyst can never definitively know whether their imputation or weighting models are correctly specified and will, of course, never know if their estimate of θ is asymptotically unbiased for its true value. Our suggestion for this setting is to consider calculating standard errors using both Rubin's rules and the proposed variance estimation approach.

One consideration when implementing the combined IPW/MI estimator is how imputation is performed when $\mathcal{R}$ is defined such that multiple variables are to be imputed. This occurs, for example, in the hypothetical example described in Section 3.3 where missingness in X_2 is resolved using IPW and missingness in X_3 and Y is resolved using MI. One way forward is to model the missing variables using a multivariate normal distribution akin to the multivariate normal imputation (MVNI) strategy first implemented by Schafer (1997). While the assumption that the missing data follow a multivariate normal distribution will frequently not hold, particularly when some of the missing variables are categorical, there is some evidence that MVNI performs reasonably well when an adaptive rounding procedure is used for imputing categorical variables (Bernaards, Belin and Schafer (2007), Lee and Carlin (2010)). The robustness of this approach to the assumption of multivariate normality has been the subject of much debate; for further discussion of the performance of MVNI in the presence of nonnormal data, see Horton and Kleinman (2007), Huque et al. (2018), Kropko et al. (2014), Lee and Carlin (2017), von Hippel (2013), Xia and Yang (2016), Yucel, He and Zaslavsky (2011).

We also illustrated how the proposed analytic framework can be used in a study of weight change following bariatric surgery using weight measures derived from EHR. While the standard errors were moderately large, the results of the data analysis suggest that, regardless of which inferential approach was used, RYGB resulted in greater percent total weight loss than VSG but that the weight loss difference was smaller in magnitude when considering individuals with chronic kidney disease at baseline. RYGB is already generally considered to be a higher-risk procedure compared to VSG, both in general and with regards to renal function. That the weight loss advantage of RYGB compared to VSG is slightly attenuated among surgical candidates with chronic kidney disease provides further evidence that baseline chronic kidney disease should be taken into consideration when making surgical treatment recommendations. The impact of these results is, of course, limited by the dichotomization of chronic kidney disease status, since chronic kidney disease is a progressive condition.

The consideration of missing data for EHR presented in this paper is just one possible approach of many. For instance, we could have specified different "rules" for inclusion in the imputation stage (i.e., redefining what it means for $\mathcal{R} = 1$). Even within the same rule, there are infinitely many ways one might have defined missingness: one could have considered a wider window of time for eligible follow-up BMI measurements, or one could have defined missingness in baseline Charlson–Elixhauser combined comorbidity score and chronic kidney disease status in a way that could have resulted in more or less stringent requirements for completeness. In developing these definitions, one must consider both the clinical meaningfulness of the data being captured and the process by which data appears in the EHR, a

process sometimes referred to as the data provenance (Haneuse and Daniels (2016)). One future avenue of research is assessment of sensitivity to EHR data that is believed to be missing not at random (i.e., whether a value is missing is related to the value itself or other missing data), which is well acknowledged as a concern when handling EHR data, as patients tend to interact more with the healthcare system when they are sick compared to when they are well (Pivovarov et al. (2014)). A key consideration is how inference can be performed in a robust fashion in that setting.

Acknowledgments. We would like to thank the Editor, Associate Editor and referees for their insightful feedback regarding this manuscript which resulted in meaningful improvements in both the methodology and the data application. We would also like to thank Lisa J. Herrinton, a principal investigator of the DURABLE study, for providing data from Kaiser Permanente Northern California. T.T. is also affiliated with the Department of Medicine at Harvard Medical School and was supported by NIH/NIDDK Award Number F-31 DK118817. D.E.A., K.J.C. and S.H. were supported by NIH/NIDDK Award Number R-01 DK105960. S.H. was supported by NIH/NCI Award Number P-50 CA244433-01.

SUPPLEMENTARY MATERIAL

Supplement to "Robust inference when combining inverse-probability weighting and multiple imputation to address missing data with application to an electronic health records-based study of bariatric surgery". (DOI: 10.1214/20-AOAS1386SUPPA; .pdf). We provide the proof for Theorem 3.1, full results from the simulation study, and results when increasing the sample size to $N = 10{,}000$. We also describe a second simulation study that considers misspecification of the weighting model. We also provide additional simulations describing the asymptotic equivalence of certain aspects of improper compared to proper imputation. We conclude with information about how potential data entry errors were corrected for when defining BMI measured at follow-up in the data application of Section 6.

R code for "Robust inference when combining inverse-probability weighting and multiple imputation to address missing data with application to an electronic health records-based study of bariatric surgery". (DOI: 10.1214/20-AOAS1386SUPPB; .zip). R code and tutorial for implementing the robust variance estimator presented in this paper.

REFERENCES

ABADIE, A., IMBENS, G. W. and ZHENG, F. (2014). Inference for misspecified models with fixed regressors. *J. Amer. Statist. Assoc.* **109** 1601–1614. MR3293613 https://doi.org/10.1080/01621459.2014.928218

ARTERBURN, D., LIVINGSTON, E. H., OLSEN, M. K., SMITH, V. A., KAVEE, A. L., KAHWATI, L. C., HENDERSON, W. G. and MACIEJEWSKI, M. L. (2013). Predictors of initial weight loss after gastric bypass surgery in twelve veterans affairs medical centers. *Obes. Res. Clin. Pract.* **7** e367–e376.

ARTERBURN, D., WELLMAN, R., EMILIANO, A., SMITH, S. R., ODEGAARD, A. O., MURALI, S., WILLIAMS, N., COLEMAN, K. J., COURCOULAS, A. et al. (2018). Comparative effectiveness and safety of bariatric procedures for weight loss: A PCORnet cohort study. *Ann. Intern. Med.* **169** 741–750. https://doi.org/10.7326/M17-2786

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189 https://doi.org/10.1111/j.1541-0420.2005.00377.x

BRUCE BAYLEY, K., BELNAP, T., SAVITZ, L., MASICA, A., SHAH, N. and FLEMING, N. S. (2013). Challenges in using electronic health record data for CER: Experience of 4 learning organizations and solutions applied. *Med. Care* **51** S80–S86.

BERNAARDS, C. A., BELIN, T. R. and SCHAFER, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat. Med.* **26** 1368–1382. MR2345726 https://doi.org/10.1002/sim.2619

COURCOULAS, A. P., CHRISTIAN, N. J., O'ROURKE, R. W., DAKIN, G., PATCHEN DELLINGER, E., FLUM, D. R., MELISSA KALARCHIAN, P. D., MITCHELL, J. E., PATTERSON, E. et al. (2015). Preoperative factors and 3-year weight change in the Longitudinal Assessment of Bariatric Surgery (LABS) consortium. *Surg. Obes. Relat. Dis.* **11** 1109–1118.

GAGNE, J. J., GLYNN, R. J., AVORN, J., LEVIN, R. and SCHNEEWEISS, S. (2011). A combined comorbidity score predicted mortality in elderly patients better than existing scores. *J. Clin. Epidemiol.* **64** 749–759.

HANEUSE, S. (2016). Distinguishing selection bias and confounding bias in comparative effectiveness research. *Med. Care* **54** e23–e29. https://doi.org/10.1097/MLR.0000000000000011

HANEUSE, S. and DANIELS, M. (2016). A general framework for considering selection bias in EHR-based studies: What data are observed and why? *EGEMs* **4** 1–17.

HANEUSE, S. J. A. and SHORTREED, S. M. (2017). On the use of electronic health records. In *Methods in Comparative Effectiveness Research* 469–502. CRC Press, Boca Raton.

HATOUM, I. J. and KAPLAN, L. M. (2012). Advantages of percent weight loss as a method of reporting weight loss after roux-en-y gastric bypass. *Obesity* (*Silver Spring, Md.*) **21**.

HORTON, N. J. and KLEINMAN, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Amer. Statist.* **61** 79–90. MR2339152 https://doi.org/10.1198/000313007X172556

HUGHES, R. A., STERNE, J. A. C. and TILLING, K. (2016). Comparison of imputation variance estimators. *Stat. Methods Med. Res.* **25** 2541–2557. MR3572869 https://doi.org/10.1177/0962280214526216

HUQUE, M. H., CARLIN, J. B., SIMPSON, J. A. and LEE, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Med. Res. Methodol.* **18** 168–16.

IMBENS, G. W. and KOLESÁR, M. (2016). Robust standard errors in small samples: Some practical advice. *Rev. Econ. Stat.* **98** 701–712.

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458 https://doi.org/10.1214/07-STS227

KENWARD, M. G. and CARPENTER, J. (2007). Multiple imputation: Current perspectives. *Stat. Methods Med. Res.* **16** 199–218. MR2371006 https://doi.org/10.1177/0962280206075304

KIM, J. K., BRICK, J. M., FULLER, W. A. and KALTON, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 509–521. MR2278338 https://doi.org/10.1111/j.1467-9868.2006.00546.x

KROPKO, J., GOODRICH, B., GELMAN, A. and HILL, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Polit. Anal.* **22** 497–519.

LEE, K. J. and CARLIN, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *Am. J. Epidemiol.* **171** 624–632.

LEE, K. J. and CARLIN, J. B. (2017). Multiple imputation in the presence of non-normal data. *Stat. Med.* **36** 606–617. MR3594613 https://doi.org/10.1002/sim.7173

LI, R. A., LIU, L., ARTERBURN, D., COLEMAN, K. J., COURCOULAS, A. P., FISHER, D., HANEUSE, S., JOHNSON, E., THEIS, M. K. et al. (2019). Five-year longitudinal cohort study of reinterventions after sleeve gastrectomy and roux-en-y gastric bypass. *Ann. Surg.*. https://doi.org/10.1097/sla.0000000000003401

LIESKE, J. C., MEHTA, R. A., MILLINER, D. S., RULE, A. D., BERGSTRALH, E. J. and SARR, M. G. (2015). Kidney stones are common after bariatric surgery. *Kidney Inter., Suppl.* **87** 839–845. https://doi.org/10.1038/ki.2014.352

MACIEJEWSKI, M. L., ARTERBURN, D. E., SCOYOC, L. V., SMITH, V. A., YANCY, W. S., WEIDENBACHER, H. J., LIVINGSTON, E. H. and OLSEN, M. K. (2016). Bariatric surgery and long-term durability of weight loss. *JAMA Surg.* **151** 1046–1055. https://doi.org/10.1001/jamasurg.2016.2317

MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* **9** 538–558. https://doi.org/10.1214/ss/1177010269

MENG, X.-L. and ROMERO, M. (2003). Discussion: Efficiency and self-efficiency with multiple imputation inference. *Int. Stat. Rev.* **71** 607–618.

NASR, S. H., D'AGATI, V. D., SAID, S. M., STOKES, M. B., LARGOZA, M. V., RADHAKRISHNAN, J. and MARKOWITZ, G. S. (2008). Oxalate nephropathy complicating roux-en-y gastric bypass: An underrecognized cause of irreversible renal failure. *Clin. J. Am. Soc. Nephrol.* **3** 1676–1683. https://doi.org/10.2215/CJN.02940608

NEFF, K. J., OLBERS, T. and LE ROUX, C. W. (2013). Bariatric surgery: The challenges with candidate selection, individualizing treatment and clinical outcomes. *BMC Med.* **11** 8. https://doi.org/10.1186/1741-7015-11-8

NGUYEN, N. T., MASOOMI, H., LAUGENOUR, K., SANAIHA, Y., REAVIS, K. M., MILLS, S. D. and STAMOS, M. J. (2011). Predictive factors of mortality in bariatric surgery: Data from the nationwide inpatient sample. *Surgery* **150** 347–351.

NIELSEN, S. F. (2003). Proper and improper multiple imputation. *Int. Stat. Rev.* **71** 593–607.

PIVOVAROV, R., ALBERS, D. J., SEPULVEDA, J. L. and ELHADAD, N. (2014). Identifying and mitigating biases in EHR laboratory tests. *J. Biomed. Inform.* **51** 24–34. https://doi.org/10.1016/j.jbi.2014.03.016

QUARTAGNO, M., CARPENTER, J. R. and GOLDSTEIN, H. (2019). Multiple imputation with survey weights: A multilevel approach. *J. Surv. Stat. Methodology*. smz036. https://doi.org/10.1093/jssam/smz036

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

ROBINS, J. M. and WANG, N. (2000). Inference for imputation estimators. *Biometrika* **87** 113–124. MR1766832 https://doi.org/10.1093/biomet/87.1.113

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Classics Library*. Wiley Interscience, Hoboken, NJ. MR2117498

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. *Monographs on Statistics and Applied Probability* **72**. CRC Press, London. MR1692799 https://doi.org/10.1201/9781439821862

SCHENKER, N. and WELSH, A. H. (1988). Asymptotic results for multiple imputation. *Ann. Statist.* **16** 1550–1566. MR0964938 https://doi.org/10.1214/aos/1176351053

SEAMAN, S. R. and VANSTEELANDT, S. (2018). Introduction to double robust methods for incomplete data. *Statist. Sci.* **33** 184–197. MR3797709 https://doi.org/10.1214/18-STS647

SEAMAN, S. R. and WHITE, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Stat. Methods Med. Res.* **22** 278–295. MR3190658 https://doi.org/10.1177/0962280210395740

SEAMAN, S. R., WHITE, I. R. and LEACY, F. P. (2014). Comment on "Analysis of longitudinal trials with protocol deviations: A framework for relevant, accessible assumptions, and inference via multiple imputation," by Carpenter, Roger, and Kenward [MR3196115]. *J. Biopharm. Statist.* **24** 1358–1362. MR3273793 https://doi.org/10.1080/10543406.2014.928306

SEAMAN, S. R., WHITE, I. R., COPAS, A. J. and LI, L. (2012). Combining multiple imputation and inverse-probability weighting. *Biometrics* **68** 129–137. MR2909861 https://doi.org/10.1111/j.1541-0420.2011.01666.x

STEFANSKI, L. A. and BOOS, D. D. (2002). The calculus of M-estimation. *Amer. Statist.* **56** 29–38. MR1939394 https://doi.org/10.1198/000313002753631330

TANG, Y. (2017). On the multiple imputation variance estimator for control-based and delta-adjusted pattern mixture models. *Biometrics* **73** 1379–1387. MR3744550 https://doi.org/10.1111/biom.12702

THAWEETHAI, T., ARTERBURN, D. E., COLEMAN, K. J. and HANEUSE, S. (2021). Supplement to "Robust inference when combining inverse-probability weighting and multiple imputation to address missing data with application to an electronic health records-based study of bariatric surgery." https://doi.org/10.1214/20-AOAS1386SUPPA, https://doi.org/10.1214/20-AOAS1386SUPPB

TURGEON, N. A., PEREZ, S., MONDESTIN, M., DAVIS, S. S., LIN, E., TATA, S., KIRK, A. D., LARSEN, C. P., PEARSON, T. C. et al. (2012). The impact of renal function on outcomes of bariatric surgery. *J. Am. Soc. Nephrol.* **23** 885–894.

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). Mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** 1–67.

VON HIPPEL, P. T. (2013). Should a normal imputation model be modified to impute skewed variables? *Sociol. Methods Res.* **42** 105–138. MR3190726 https://doi.org/10.1177/0049124112464866

WANG, N. and ROBINS, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85** 935–948. MR1666715 https://doi.org/10.1093/biomet/85.4.935

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163 https://doi.org/10.2307/1912526

XIA, Y. and YANG, Y. (2016). Bias introduced by rounding in multiple imputation for ordered categorical variables. *Amer. Statist.* **70** 358–364. MR3574788 https://doi.org/10.1080/00031305.2016.1200486

XU, D., DANIELS, M. J. and WINTERSTEIN, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics* **17** 589–602. MR3603956 https://doi.org/10.1093/biostatistics/kxw009

YUCEL, R. M., HE, Y. and ZASLAVSKY, A. M. (2011). Gaussian-based routines to impute categorical variables in health surveys. *Stat. Med.* **30** 3447–3460. MR2861625 https://doi.org/10.1002/sim.4355

ZELLMER, J. D., MATHIASON, M. A., KALLIES, K. J. and KOTHARI, S. N. (2014). Is laparoscopic sleeve gastrectomy a lower risk bariatric procedure compared with laparoscopic roux-en-y gastric bypass? A meta-analysis. *Am. J. Surg.* **208** 903–910.