RESEARCH ARTICLE

Accounting for nonmonotone missing data using inverse probability weighting

Rachael K. Ross[1] | **Stephen R. Cole[1]** | **Jessie K. Edwards[1]** | **Daniel Westreich[1]** | **Julie L. Daniels[1]** | **Jeffrey S.A. Stringer[1,2]**

[1]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, USA

[2]Department of Obstetrics and Gynecology, School of Medicine, University of North Carolina, Chapel Hill, North Carolina, USA

Correspondence
Rachael K. Ross, Department of Epidemiology, Gillings School of Global Public Health, UNC Campus Box 7435m Chapel Hill, NC 27599-6435, USA.
Email: rkross@unc.edu

Inverse probability weighting can be used to correct for missing data. New estimators for the weights in the nonmonotone setting were introduced in 2018. These estimators are the unconstrained maximum likelihood estimator (UMLE) and the constrained Bayesian estimator (CBE), an alternative if UMLE fails to converge. In this work we describe and illustrate these estimators, and examine performance in simulation and in an applied example estimating the effect of anemia on spontaneous preterm birth in the Zambia Preterm Birth Prevention Study. We compare performance with multiple imputation (MI) and focus on the setting of an observational study where inverse probability of treatment weights are used to address confounding. In simulation, weighting was less statistically efficient at the smallest sample size and lowest exposure prevalence examined (n = 1500, 15% respectively) but in other scenarios statistical performance of weighting and MI was similar. Weighting had improved computational efficiency taking, on average, 0.4 and 0.05 times the time for MI in R and SAS, respectively. UMLE was easy to implement in commonly used software and convergence failure occurred just twice in >200 000 simulated cohorts making implementation of CBE unnecessary. In conclusion, weighting is an alternative to MI for nonmonotone missingness, though MI performed as well as or better in terms of bias and statistical efficiency. Weighting's superior computational efficiency may be preferred with large sample sizes or when using resampling algorithms. As validity of weighting and MI rely on correct specification of different models, both approaches could be implemented to check agreement of results.

KEYWORDS
imputation, missing data, nonmonotone, simulation, weighting

1 | INTRODUCTION

Missing data plague research. Reviews of the epidemiologic and clinical literature show that missing data are often inadequately reported and that complete case analysis, where records with missing data are excluded, remains the most frequently implemented approach to handle missing data.[1-6] Complete case analyses can be statistically inefficient and are valid only under certain assumptions, generally that data are missing completely at random, or that missingness is

independent of the outcome and effect modifiers.[7-10] Weighting is an alternative approach to handle missing data that is valid under different assumptions, principally that the data are missing at random.[11-13]

It is generally straightforward to estimate weights to account for missing data when missing data follow a uniform pattern (ie, for each individual, the variables with missingness are either all observed or all missing) or a monotone pattern (ie, there is an ordering in which a variable is observed only if the previous variable is observed, such as missing data after lost to follow-up) (see illustration in Appendix A).[12] However, until recently, weighting approaches for nonmonotone missing data (ie, when missingness is neither uniform nor monotone) have been challenging to implement.[14,15] In 2018, Sun and Tchetgen Tchetgen published two estimators for weights in the setting of nonmonotone missingness.[16,17] Unlike prior approaches, their estimators can be readily implemented in commonly used software.

In this paper, we describe and illustrate the estimators from Sun and Tchetgen Tchetgen, and examine their performance in simulation and an applied example estimating the effect of anemia on spontaneous preterm birth in the Zambia Preterm Birth Prevention Study (ZAPPS).[18,19] We compare performance with multiple imputation (MI), a commonly used alternative. We specifically examine the setting of an observational study where inverse probability of treatment weights are used to address confounding. In Section 2, we introduce our motivating application. In Section 3, we detail our parameter of interest, a sufficient set of identification assumptions, and weighted estimators. In Section 4, we describe the weighted estimators from Sun and Tchetgen Tchetgen using a simple example to aid illustration. In Section 5, we present results from a limited simulation study to compare performance with MI in finite samples. Section 6 presents results from the motivating application. Finally, in Section 7, we discuss the findings and consider the choice between weighting and MI.

2 | MOTIVATING APPLICATION

Our objective was to estimate the effect of maternal anemia on the risk of spontaneous preterm birth among people seeking prenatal care in Lusaka, Zambia. Some research has suggested an association between maternal anemia, particularly when diagnosed early in pregnancy, and poor pregnancy outcomes.[20-22] However, this finding has not been consistently observed.[23,24] To estimate this effect, we used data from ZAPPS,[18,19] an observational prospective cohort of 1450 people recruited at prenatal care initiation in Lusaka, Zambia between 2015 and 2017. A person was eligible if they were ≥ 18 years old, had a viable intrauterine single or twin pregnancy, presented to prenatal care prior to 20 weeks of gestation if HIV-seropositive or 24 weeks if HIV-seronegative, and resided within Lusaka with no plans to relocate during follow-up. Anemia was diagnosed at enrollment if the capillary hemoglobin concentration was <10.5 g/dL (HemoCue Hb 201).[25] Spontaneous preterm birth was defined as delivery occurring after spontaneous labor or membrane rupture prior to 37 weeks of gestation. Additional covariates collected at enrollment and used in this analysis included gestational age at enrollment, maternal age, maternal HIV serostatus, and previous pregnancy and birth history. Three people experienced a miscarriage and were excluded from the analysis, resulting in 1447 people.

Table 1 shows cohort characteristics and occurrence of the outcome, overall and stratified by anemia diagnosis. ZAPPS is typical of many prospective cohorts. Despite rigorous study procedures and active efforts to maximize study retention, some data are missing. In particular, our exposure and outcome have notable missingness; 425 people (29%) did not have a hemoglobin measurement and 239 people (17%) were lost to follow-up. Ignoring missing data, 13.5% of the cohort was anemic and 9.9% had a spontaneous preterm birth. The risk of spontaneous preterm birth was higher among anemic people (12.4%) compared to people without anemia (9.5%) or people with missing anemia status (9.7%). There was also missingness in some covariates: maternal age (n = 41, 3%), maternal HIV serostatus (n = 3, <1%), and prior stillbirth (n = 85, 6%). Table 2 shows the 16 missing data patterns for the 5 variables with missingness. Just over half (781, 54%) of the cohort were complete cases. Among the 666 people with some missing data, 4 patterns accounted for 88%. There were nine patterns with <1% of the cohort.

3 | PARAMETER, IDENTIFICATION, AND WEIGHTED ESTIMATORS

3.1 | Parameter

Our parameter is the sample average causal effect of a time-fixed binary exposure on the outcome risk, quantified by the risk difference, $\theta = E\left(Y^{x=1}\right) - E\left(Y^{x=0}\right)$, where Y^x is the potential outcome when exposure X is set to x. This parameter

TABLE 1 Cohort characteristics overall and stratified by anemia status at enrollment

	N (%)			
		Anemia status		
	Overall	**Positive**	**Negative**	**Unknown**
Characteristic	**(n = 1447)**	**(n = 138)**	**(n = 884)**	**(n = 425)**
Spontaneous preterm birth	120 (9.9)	15 (12.4)	70 (9.6)	35 (9.7)
Missing	239	17	157	65
At enrollment				
Gestational age (weeks)				
Median (IQR)	16.1 (13.3, 18.3)	16.9 (14.7, 18.9)	15.9 (12.9, 18.1)	16.6 (13.6, 18.6)
First trimester	432 (29.9)	29 (21)	286 (32.4)	117 (27.5)
Missing	0	0	0	0
Maternal age				
Median (IQR)	27 (23, 32)	27 (22, 33)	27 (23, 32)	27 (22, 31)
< 20 years	111 (7.9)	11 (8.3)	59 (6.9)	41 (9.8)
20-34 years	1113 (79.2)	102 (76.7)	677 (79.2)	334 (79.9)
≥ 35 years	182 (12.9)	20 (15)	119 (13.9)	43 (10.3)
Missing	41	5	29	7
HIV+	349 (24.2)	58 (42)	195 (22.1)	96 (22.7)
Missing	3	0	0	3
Nulliparous	457 (31.6)	40 (29)	286 (32.4)	131 (30.8)
Missing	0	0	0	0
Prior preterm birth	410 (28.3)	40 (29)	258 (29.2)	112 (26.4)
Missing	0	0	0	0
Prior stillbirth	126 (9.3)	15 (11.6)	80 (9.6)	31 (7.8)
Missing	85	9	48	28

Abbreviation: IQR, interquartile range.

requires identification of two sample average risks, $E(Y^x)$, one under exposure ($x = 1$) and one under no exposure ($x = 0$). We focus on identification of the risk in the observational setting in which there are common causes of the exposure and outcome, Z, that produce confounding bias such that the risk is not identified by the crude conditional risk, that is, $E[Y^x] \neq E[Y|X = x]$, even in the absence of missing data.

3.2 | No missing data

We can point-identify the risk under the assumptions of conditional exchangeability with positivity, causal consistency, and no measurement error. Conditional exchangeability means that, conditional on a set of measured confounders, the potential outcomes are independent of the observed exposure, $Y^x \coprod X \mid Z$, such that $E[Y^x|Z] = E[Y^x \mid X = x, Z]$.[26,27] Positivity means that every person has a nonzero probability of having each level of exposure across the distribution of Z.[26-28] Causal consistency means that the potential outcome Y^x is the observed outcome Y for people with observed exposure x.[27,29] With these conditions, $E(Y^x)$ is identified by a weighted risk where the weight is the inverse of the confounder-conditional probability of exposure (ie, inverse probability of treatment weight, hereinafter *treatment weight*),

TABLE 2 Missing data patterns in the motivating example data from the ZAPPS cohort, n = 1447

Pattern	Anemia	Spont. PTB	Maternal Age	HIV serostatus	Prior stillbirth	N	%
1	O	O	O	O	O	781	54.0
2	M	O	O	O	O	330	22.8
3	O	M	O	O	O	151	10.4
4	M	M	O	O	O	58	4.0
5	O	O	O	O	M	45	3.1
6	M	O	O	O	M	26	1.8
7	O	O	M	O	O	21	1.5
8	O	M	M	O	O	12	0.8
9	O	M	O	O	M	11	0.8
10	M	M	M	O	O	5	0.4
11	M	O	O	M	O	2	0.1
12	O	O	M	O	M	1	0.1
13	M	O	M	O	O	1	0.1
14	M	O	M	O	M	1	0.1
15	M	M	O	O	M	1	0.1
16	M	M	O	M	O	1	0.1

Note: "O" indicates variable is observed and "M" indicates variable is missing. Variables not included here (gestational age at enrollment, maternal HIV serostatus, nulliparity, and prior preterm birth) did not have missingness.

Abbreviations: Spont. PTB, spontaneous preterm birth.

formally

$$E[Y^x] = E\left[\frac{YI(X = x)}{\Pr(X = x|Z = z)}\right],$$

where $I(a)$ is an indicator that takes the value 1 when a is true and 0 otherwise and $\Pr(X = x|Z = z)^{-1}$ is the treatment weight. An estimator of this weighted risk is $\frac{1}{n}\sum_i \frac{Y_i I(X_i = x)}{\widehat{\Pr}(X_i = x|Z_i)}$ where i indexes the independent and identically distributed n people included in the sample. $\Pr(X_i = x|Z_i)$ can be estimated nonparametrically or using a parametric model, commonly a logistic regression, called a propensity score model (proof[30] in Appendix B).

3.3 | With missing data

When some data are missing, the identification conditions described above are not sufficient because the weighted risk above is no longer expressed in terms of fully observed data. Let $R = 1$ for complete cases. We can point-identify the risk among the complete cases by incorporating a second weight, formally

$$E\left[\frac{YI(X = x)}{\Pr(X = x|Z = Z)}\right] = E\left[\frac{\Pr(R = 1|Z = z, X = x, Y = y)}{\Pr(R = 1|Z = z, X = x, Y = y)}\frac{YI(X = x)}{\Pr(X = x|Z = Z)}\right]$$

$$= E\left[\frac{E[I(R = 1) \mid Z = z, X = x, Y = y]YI(X = x)}{\Pr(R = 1|Z = z, X = x, Y = y)\Pr(X = x|Z = z)}\right]$$

$$= E\left[\frac{YI(X = x)I(R = 1)}{\Pr(R = 1|Z = z, X = x, Y = y)\Pr(X = x|Z = z)}\right].$$

The first equality is multiplication by 1, the second is the equivalence between probability and expectation of an indicator function, and the third is the law of total probability. The additional weight, $\Pr(R = 1|Z = z, X = x, Y = y)^{-1}$, is the inverse of the conditional-probability of missingness, hereinafter the missingness weight. We require additional conditions to obtain $\Pr(R = 1|Z = z, X = x, Y = y)$ because some data on $Z, X,$ or Y are not observed for people with $R \neq 1$. These conditions are that the data are missing at random (MAR) and there is positivity. We reserve full explanation of MAR and estimation of $\Pr(R = 1|Z = z, X = x, Y = y)$ for Section 4. In the missing data setting, positivity means that everyone has a nonzero probability of being a complete case.[16,17] Once we obtain the missingness weight, we subsequently obtain the treatment weight among the weighted complete cases (proof in Appendix C).[31] An estimator of the weighted risk is

$$\frac{1}{n_{wt}} \sum_i \frac{Y_i I(X_i = x, R_i = 1)}{\widehat{\Pr}(R_i = 1|Z_i = z, X_i = x, Y_i = y)\widehat{\Pr}(X_i = x|Z_i = z)},$$

where $n_{wt} = \sum_i \frac{I(X_i = x, R_i = 1)}{\widehat{\Pr}(R_i = 1|Z_i = z, X_i = x, Y_i = y)\widehat{\Pr}(X_i = x|Z_i = z)}$.

4 | SUN AND TCHETGEN TCHETGEN ESTIMATORS FOR THE MISSINGNESS WEIGHT

To aide illustration, we introduce a simple example with exposure X, outcome Y and confounder Z. There are non-monotone missing data with four patterns (Table 3). Let R denote the pattern to which a person belongs where $R = 1$ is reserved for complete cases. For the missingness weight, we require the conditional probability of being a complete case, $P(R = 1 | Z = z, X = x, Y = y)$, which is the complement of the sum of the probabilities of the other patterns,

$$P(R = 1|Z, X, Y) = 1 - P(R = 2|Z, X, Y) - P(R = 3|Z, X, Y) - P(R = 4|Z, X, Y).$$

The MAR condition is that missingness is independent of the missing data, conditional on the observed data. For example, Y is missing in pattern $R = 2$. Under MAR, $I(R = 2) \coprod Y \mid Z, X$, such that $P(R = 2|Z, X, Y) = P(R = 2 \mid Z, X)$. Applying this condition to each pattern, we get

$$\Pr(R = 1|Z, X, Y) = 1 - \Pr(R = 2|Z, X) - \Pr(R = 3|X) - \Pr(R = 4|X, Y).$$

When missingness is not independent from the missing data, neither marginally nor conditional on observed data, then data are missing not at random (MNAR). Data are missing completely at random (MCAR) when missingness is marginally independent of the observed and missing data, a stronger condition than MAR.

Sun and Tchetgen Tchetgen developed two estimators for the conditional probability of being a complete case under MAR.[16,17] Let π_r be the probability for each pattern. We specify logistic models for each pattern $R > 1$, formally

$$\Pr(R = 2|Z, X) = \pi_2(\gamma_2) = 1/(1 + \exp(-(\gamma_{20} + \gamma_{21} X + \gamma_{22} Z))),$$

$$\Pr(R = 3|X) = \pi_3(\gamma_3) = 1/(1 + \exp(-(\gamma_{30} + \gamma_{31}X))),$$

$$\Pr(R = 4|X, Y) = \pi_4(\gamma_4) = 1/(1 + \exp(-(\gamma_{40} + \gamma_{41}X + \gamma_{42}Y))).$$

TABLE 3 Missing data patterns in the simple illustrative example

Pattern (R)	Z	X	Y
1	O	O	O
2	O	O	M
3	M	O	M
4	M	O	O

Note: "O" indicates variable is observed and "M" indicates variable is missing.

The first estimator is the unconstrained maximum likelihood estimator (UMLE). We maximize the joint log-likelihood of the models using the observed data,

$$\ln \mathcal{L}(\gamma) = \sum_{i=1}^{n} \left\{ \left[\sum_{r=2}^{4} I(R_i = r) \ln \pi_r(\gamma_r) \right] + I(R_i = 1) \ln \left[1 - \sum_{r=2}^{4} \pi_r(\gamma_r) \right] \right\}.$$

Each person contributes a term to the likelihood that corresponds to the pattern to which they belong. The log-likelihood can be maximized in standard software and we provide code in SAS (using the NLMIXED procedure) and R (using NLM in the Stats package)[32] (available at https://github.com/rachael-k-ross/Wgt-NonmonotoneMiss). The UMLE does not naturally impose the constraint that $\pi_1 > 0$, so the log-likelihood may fail to converge if there is a fitted π_1 for a complete case close to zero. Therefore, Sun and Tchetgen Tchetgen also proposed the constrained Bayesian estimator (CBE).

The CBE bounds the probability of being a complete case away from zero by discarding draws from the posterior that do not meet a user-specified constraint. Let c be a user-specified constraint that is a small positive number. CBE produces a posterior distribution that is proportional to the combination of the likelihood, constraint, and prior distributions $f(\gamma)$,

$$f(\gamma|X, Z, Y) \propto \mathcal{L}(\gamma) I(R_i = 1) I(\pi_2(\gamma_2) + \pi_3(\gamma_3) + \pi_4(\gamma_4) < 1 - c) f(\gamma).$$

Sun and Tchetgen Tchetgen used independent diffuse priors for each γ parameter, $\gamma \sim N(0,100)$ (ie, multivariate normal with an identity covariance matrix), $c = 10^{-8}$ and adaptive Gibbs sampling to sample from the posterior.[33] They used the median of the posterior samples to estimate $\hat{\gamma}$. Sun and Tchetgen Tchetgen provided OpenBugs code for implementation.[34] On GitHub (https://github.com/rachael-k-ross/Wgt-NonmonotoneMiss), we provide R code using R2jags package[35] which calls Just Another Gibbs Sampler (JAGS).[36] The OpenBugs and R2jags code are opaque, so we provide code in SAS and R for a more transparent (though less efficient) manually coded Metropolis-Hastings algorithm with rejection sampling that imposes the user-specified constraint by rejecting draws that violate it.[37] Once $\hat{\gamma}$ are estimated by UMLE or CBE, the missingness weight can be estimated for each complete case.

4.1 | Inference

The naïve SE estimate from the weighted analyses described above (incorporating treatment weights or missingness weights, alone or together) is not consistent and resulting Wald-type confidence intervals (CI) may have poor coverage.[38] Two options for estimating the SE and obtaining appropriate CIs are (1) nonparametric bootstrap and (2) the sandwich variance estimator. The sandwich variance estimator can incorporate estimation of all parameters (including the nuisance parameters γ) or only the parameters of interest, treating the nuisance parameters as known (ie, not estimated). This latter approach is commonly called the "robust" (Huber-White) sandwich estimator and is expected to produce conservative estimates leading to over-coverage in CIs.[39,40] In contrast, nonparametric bootstrapping will produce nominal CIs though it can be computationally intensive.[41,42] Here we use the robust sandwich estimator.

5 | SIMULATION

We conducted a limited simulation study to assess finite-sample performance of weighting for nonmonotone missingness. The design was guided by the motivating application. We used both SAS and R.

5.1 | Data generation

We simulated 5000 studies each with n independent individuals, a binary exposure X, a binary outcome Y, and three correlated confounders Z (one continuous and two binary). Under no exposure ($X = 0$), the marginal incidence of the outcome was 10%. See Appendix D for data generation details and causal diagram. We varied $n = (1500, 5000)$, marginal prevalence of exposure $p_x = (15\%, 50\%)$, and the true risk difference, in percentage points, $\theta = (0, 5)$.

TABLE 4 Missing data patterns in the simulation study

Pattern (R)	X	Y	Z_1	Z_2	Z_3	% in each pattern					
						Primary	Secondary scenarios				
1	O	O	O	O	O	50	65	35	50	65	35
2	M	O	O	O	O	15	10	15	10	5	15
3	O	M	O	O	O	15	10	15	10	5	10
4	M	M	O	O	O	10	5	15	10	5	10
5	O	O	O	O	M	5	5	10	5	5	10
6	M	O	O	O	M	5	5	10	5	5	10
7	O	O	M	O	O				5	5	5
8	O	M	M	O	O				5	5	5

Note: "O" indicates variable is observed and "M" indicates variable is missing.

After generating the full data, we induced missingness guided by our motivating example. In the primary scenario, we generated missing data with 6 patterns and 50% complete cases (Table 4). For each pattern where $R > 1$, we specified a logistic model

$$\Pr(R = r|Z, X, Y) = 1/(1 + \exp(-(\gamma_{r0} + \gamma_{r1}X + \gamma_{r2}Y + \gamma_{r3}Z_1 + \gamma_{r4}Z_2 + \gamma_{r5}Z_3)))$$

to obtain individual probabilities of being in that pattern. The probability of being a complete case was the complement of the sum of the other probabilities. The observed missing data pattern, R, was generated from a multinomial distribution and then missing data were imposed according to the observed pattern (ie, value of any variable missing under that pattern was set to missing). We varied the γ coefficients to produce missing data that were MAR (coefficients for variables missing in that pattern were zero), MCAR (all coefficients were zero), and MNAR (some coefficients for missing variables were nonzero). The γ intercepts were set by numerical approximation to achieve the desired prevalence of each pattern.[43] We also explored how the percent of complete cases and number of patterns affected results when data are MAR (secondary scenarios, Table 4).

5.2 | Analysis

In each simulated cohort, we implemented three approaches to address missing data. Regardless of missing data approach used, we used inverse probability of treatment weighting to address confounding and fit a weighted linear-binomial outcome model conditional on exposure ($E(Y) = \lambda + \theta X$) using generalized estimating equations with an independence covariance structure to estimate the risk difference, θ, and robust SE for Wald-type 95% CIs.

First, we conducted an analysis restricted to complete cases (ie, conditional on $R = 1$). We fit the treatment propensity score model among the complete cases and estimated the treatment weight as $\widehat{\Pr}(X = x|Z = z, R = 1)^{-1}$.

Second, we implemented weighting for missingness. We implemented UMLE and, if UMLE failed to converge, we estimated $\widehat{\gamma}$ as the posterior median obtained by CBE implemented by adaptive Gibbs sampling with a single chain of 10 000 iterations with 5000 burn-in samples discarded. We used diffuse priors $\gamma \sim N(0,100)$ and set $c = 10^{-8}$. For each missingness model, all variables observed in that pattern were included. Models were correctly specified except in MNAR scenarios. Using $\widehat{\gamma}$, we estimated the missingness weight, $\widehat{\Pr}(R = 1|Z = z, X = x, Y = y)^{-1}$. We subsequently fit the missingness-weighted propensity score model among the complete cases to obtain the treatment weight, $\widehat{\Pr}(X = x|Z = z)^{-1}$. The final weight was the product of the missingness weight and treatment weight.

Third, we implemented MI by chained equations.[44,45] Using logistic regression for binary variables and linear regression for continuous variables, we imputed missing data 20 times (mice package in R[46] and MI procedure in SAS). All variables were included in each imputation model; a linear and quadratic term was included for the continuous confounder. In each imputed dataset, we fit the treatment propensity score model and estimated the treatment weight, $\widehat{\Pr}(X = x|Z = z)^{-1}$. From the treatment-weighted outcome models, the 20 estimates were combined by Rubin's rule,[47] ($\overline{\theta} = \frac{1}{20}\sum_{k=1}^{20}\widehat{\theta}_k$ where $\widehat{\theta}_k$

TABLE 5 Failures in 5000 simulated datasets when exposure prevalence was 15% and sample size was 1500 for primary missing data scenario (6 patterns, 50% complete cases)

	Risk difference 0%		Risk difference 5%	
Missing data approach[a]	R	SAS	R	SAS
MAR				
CC	12 (0.2%)	25 (0.5%)	0 (0.0%)	1 (<0.1%)
MI	18 (0.4%)	25 (0.5%)	0 (0.0%)	1 (<0.1%)
Weighting[b]	12 (0.2%)	25 (0.5%)	0 (0.0%)	1 (<0.1%)
MNAR				
CC	165 (3.3%)	202 (4.0%)	32 (0.6%)	52 (1.0%)
MI	163 (3.3%)	168 (3.4%)	22 (0.4%)	38 (0.8%)
Weighting[b]	162 (3.2%)	202 (4.0%)	34 (0.7%)	52 (1.0%)

Abbreviations: CC, complete case analysis; MAR, missing at random; MI, multiple imputation; MNAR, missing not at random.

[a]All approaches addressed confounding using inverse probability of treatment weights.

[b]UMLE used to estimate the missingness weights.

is the estimated risk difference from imputation k; $V(\overline{\theta}) = \frac{1}{20}\sum_{k=1}^{20}\widehat{V}(\widehat{\theta}_k) + \left(1 + \frac{1}{20}\right)\left(\frac{1}{20-1}\right)\sum_{k=1}^{20}\left(\widehat{\theta}_k - \overline{\theta}\right)^2$ where $\widehat{V}(\widehat{\theta}_k)$ is the estimated robust variance for $\widehat{\theta}_k$ in imputation k).

Finally, we analyzed the full data without inducing missingness. While not available in practice, the full data provide a reference against which we compare the three approaches. We fit the treatment propensity score model in the full data and estimated the treatment weight as $\widehat{\Pr}(X = x|Z = z)^{-1}$.

To compare estimator performance, we estimated (1) number of times the estimator failed to produce results, (2) bias (mean of the risk difference estimates minus the true risk difference), (3) empirical SE (SD of the estimates), (4) root mean squared error (square root of the average of the squared differences between each risk difference estimate and the true risk difference), (5) average model SE (square root of the average of the variance estimates), and 6) CI coverage (proportion of estimated 95% CIs that included the true risk difference).[48] Failures did not contribute to the estimation of the other performance measures. We also captured the run time as the average time in seconds for a single cohort (each scenario was run separately on a single 2.5 GHz processor with up to 15 GB of memory allocated; versions R 4.1.0 and SAS 9.4).

5.3 | Results

5.3.1 | Failure to produce results

Failures were rare and occurred with similar frequency across the missing data approaches (Table 5). In the primary scenario, failures only occurred when exposure prevalence was 15%, sample size was 1500, and data were missing MAR (≤0.5% failures) or MNAR (≤4.0%). In the primary scenario, there were no failures at the larger sample size, 50% exposure prevalence, or when data were MCAR. In general, failures for weighting and complete analysis occurred when the weighted outcome model restricted to the complete cases did not converge due to too few exposed individuals or too few outcomes. Weighting can also fail if UMLE fails to converge and thus missingness weights are not estimated, however this UMLE non-convergence never occurred in SAS and occurred just twice in R (across 5000 cohorts in 44 scenarios). In those two instances, CBE successfully estimated the weights. For the remainder of the simulation results, only weighting results using UMLE are presented.

5.3.2 | Statistical performance

Results from R and SAS were approximately equivalent; we present R only. For the primary scenario (6 patterns, 50% complete cases), results are presented in Table 6 and Figure 1 (MAR), Appendix Table 1 and Appendix Figure 1 (MCAR),

TABLE 6 Bias, empirical SE, root mean squared error, average model SE, and 95% confidence interval coverage of the risk difference (in percentage points) for primary missing data scenario (6 patterns, 50% complete cases) when data are missing at random (MAR)

	Exposure prevalence 15%					Exposure prevalence 50%				
Missing data approach[a]	Bias	ESE	RMSE	Avg. ModSE	95% CI coverage	Bias	ESE	RMSE	Avg. ModSE	95% CI Coverage
Risk difference 0										
n = 1500										
Full	−0.1	2.2	2.2	2.2	94%	0.0	1.6	1.6	1.6	96%
CC[b]	−1.6	2.8	3.2	2.8	81%	−1.1	1.9	2.2	1.9	91%
MI[b]	0.4	4.0	4.1	4.3	94%	0.0	2.6	2.6	2.6	94%
Weighting[b,c]	−0.5	4.4	4.5	4.5	90%	0.0	2.7	2.7	2.7	95%
n = 5000										
Full	0.0	1.2	1.2	1.2	95%	0.0	0.9	0.9	0.9	95%
CC	−1.6	1.5	2.2	1.5	77%	−1.0	1.0	1.4	1.0	84%
MI	0.2	2.3	2.3	2.4	94%	0.1	1.4	1.4	1.4	95%
Weighting[c]	−0.2	2.4	2.4	2.5	94%	0.1	1.4	1.4	1.5	96%
Risk difference 0.05										
n = 1500										
Full	−0.1	2.6	2.6	2.6	95%	0.0	1.7	1.7	1.8	95%
CC	−3.4	3.5	4.9	3.5	74%	−2.5	2.1	3.3	2.2	78%
MI	0.0	4.7	4.7	4.8	94%	0.0	2.8	2.8	2.8	94%
Weighting[c]	−0.6	5.2	5.2	5.3	91%	0.0	2.8	2.8	3.0	96%
n = 5000										
Full	0.0	1.4	1.4	1.4	95%	0.0	1.0	1.0	1.0	95%
CC	−3.4	1.9	3.8	1.9	55%	−2.5	1.2	2.7	1.2	45%
MI	−0.1	2.6	2.6	2.6	94%	0.1	1.5	1.5	1.5	95%
	−0.1	2.8	2.8	2.9	95%	0.1	1.5	1.5	1.6	96%

Abbreviations: avg. ModSE, average model SE; CC, complete case analysis; CI, confidence interval; ESE, empirical SE; MI, multiple imputation; RMSE, root mean squared error.

[a] All approaches addressed confounding using inverse probability of treatment weights.

[b] Results from 5000 simulated cohorts except at 15% prevalence for approaches marked (see Table 5 for number of failures).

[c] UMLE used to estimate the missingness weights.

and Appendix Table 2 and Appendix Figure 2 (MNAR). When data were MAR, complete case analysis was notably biased and thus had poor coverage. At 50% exposure prevalence, MI and weighting performed similarly with negligible bias and nearly the same RMSE and coverage. At 15% exposure prevalence with n = 1500, weighting performance declined with a small increase in bias and reduced precision and coverage. Comparably, MI had greater precision and thus lower RMSE. When data were MCAR, bias was negligible for all approaches. Weighting had the same precision as complete case analysis. Again, improved precision of MI over weighting was apparent only at 15% exposure prevalence with n = 1500. When data were MNAR, all missing data approaches were biased. At 15% exposure prevalence, weighting had more bias than MI though, at 50% exposure prevalence, bias was similar.

Results from scenarios varying percent of complete cases and number of missing data patterns are presented in Appendix Tables 3 and 4. Data were MAR with 50% exposure prevalence. Bias was negligible for MI and weighting so plots of results (Figure 2 and Appendix Figure 3) focus on RMSE and coverage. At n = 1500, MI had a lower RMSE than weighting when there were 35% complete cases. A small difference persisted at 50% complete cases only when there were eight patterns and the difference nearly disappeared at 65% complete cases. A similar pattern was present when n = 5000 though the differences between MI and weighting were smaller. For both sample sizes, there was decreasing coverage for
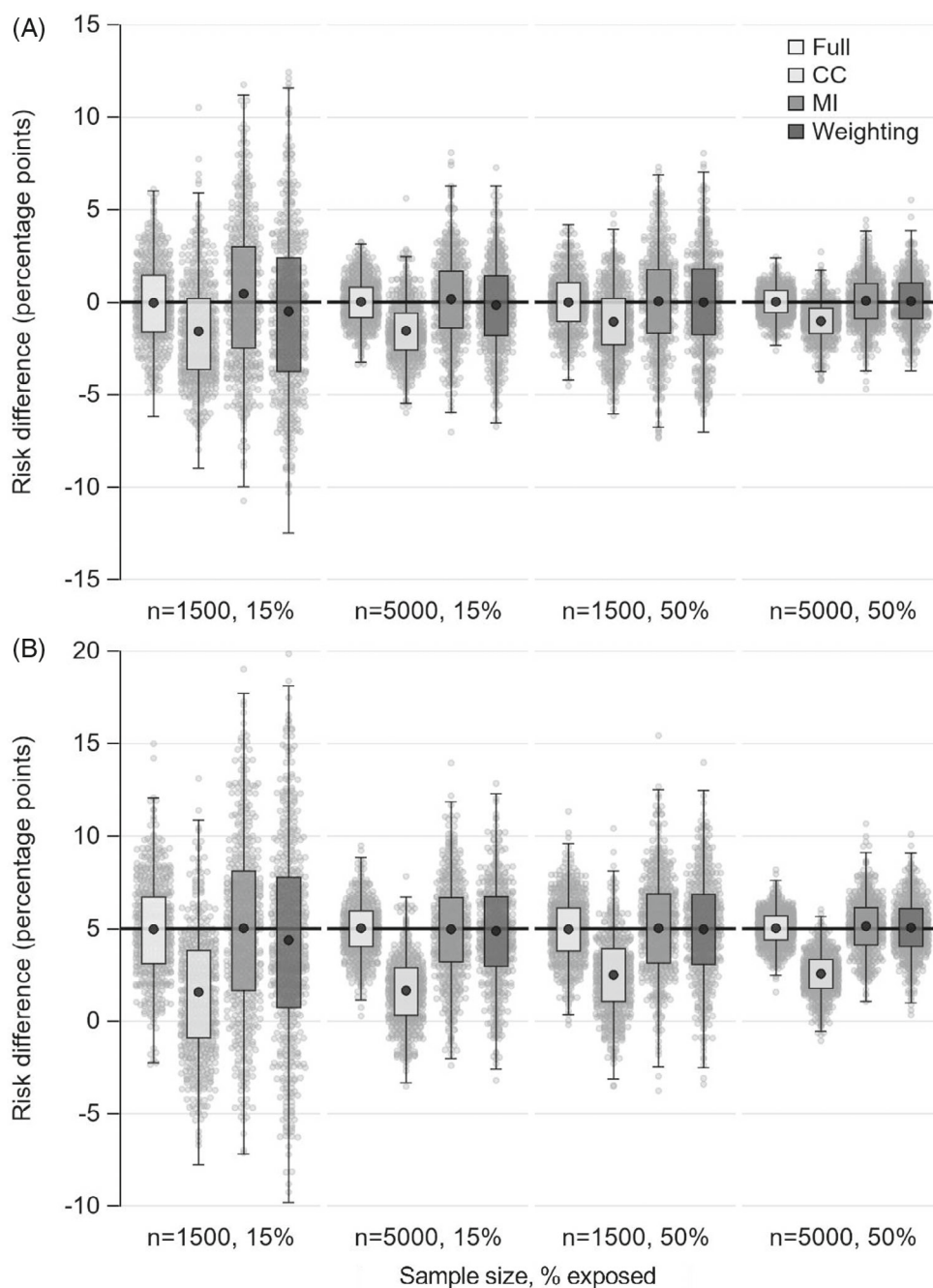
FIGURE 1 Boxplots of risk difference estimates for primary missing data scenario (6 patterns, 50% complete cases) when data are missing at random (MAR). Panel (A) true risk difference is 0%; panel (B) true risk difference is 5%. Horizontal black line marks true risk difference; black dot marks mean; small gray dots are a 10% random sample of estimates. CC, complete case analysis; MI, multiple imputation

MI (dropping <95%) and increasing coverage for weighting (>95%) as the percent of complete cases declines. There were little differences in results when there were six or eight patterns.

5.3.3 | Computational efficiency

Figure 3 plots the time in seconds for MI and weighting for all 22 scenarios by sample size (Appendix Table 5 for numeric results). On average, MI took 2.7 times as long as weighting in R (13.0 vs 4.9 s) and 18.3 times as long in SAS (19.2 vs 1.0 s). Weighting was faster in SAS than in R (average 1.0 vs 4.9); MI was slower in SAS than in R (average 19.2 vs 13.0).
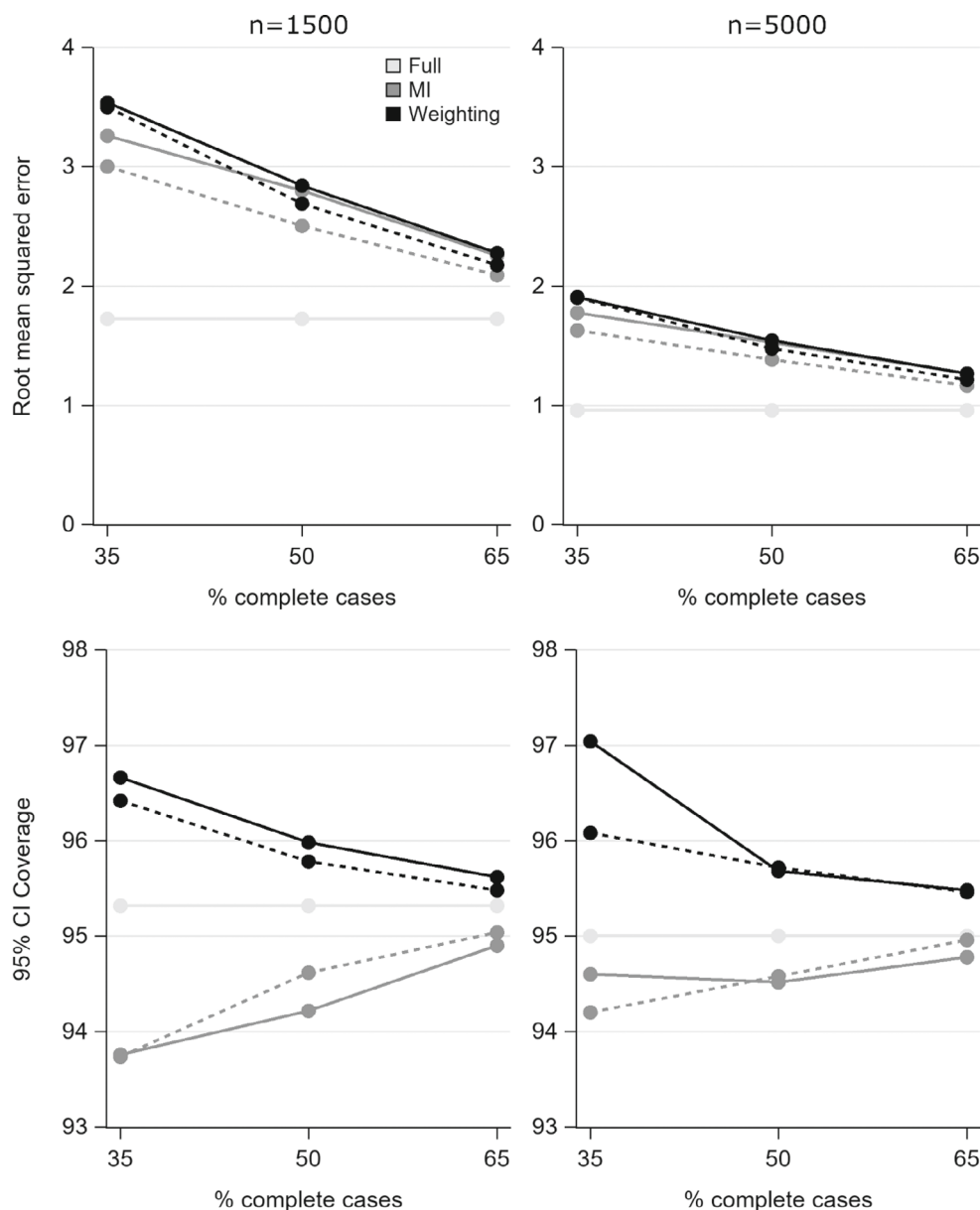
FIGURE 2 Root mean squared error and confidence interval coverage as percent complete cases and number of patterns varies when the true risk difference is 5% and data are missing at random (MAR). Solid line indicates 6 patterns and dashed line indicates 8 patterns (for Full, the two lines are exactly overlaid). MI, multiple imputation; CI, confidence interval

6 | APPLICATION

6.1 | Methods

As outlined above, we aimed to estimate the effect of maternal anemia in pregnancy on the risk of spontaneous preterm birth. We implemented three approaches to address missing data that mirror the simulation study. In all, we used treatments weights to address confounding by gestational age at enrollment, maternal age, maternal HIV serostatus, and birth history including parity, prior preterm birth, and prior stillbirth. We fit a weighted linear-binomial model to estimate the risk difference and obtained Wald-type 95% CIs using the robust SE. Gestational age and maternal age were modeled with restricted quadratic splines with four knots at the fifth, 35th, 65th, and 95th percentiles.[49]

First, we conducted a complete case analysis restricted to the 781 people with complete data. Second, we implemented weighting for missingness. There were a number of missing data patterns with few people so we combined rare patterns
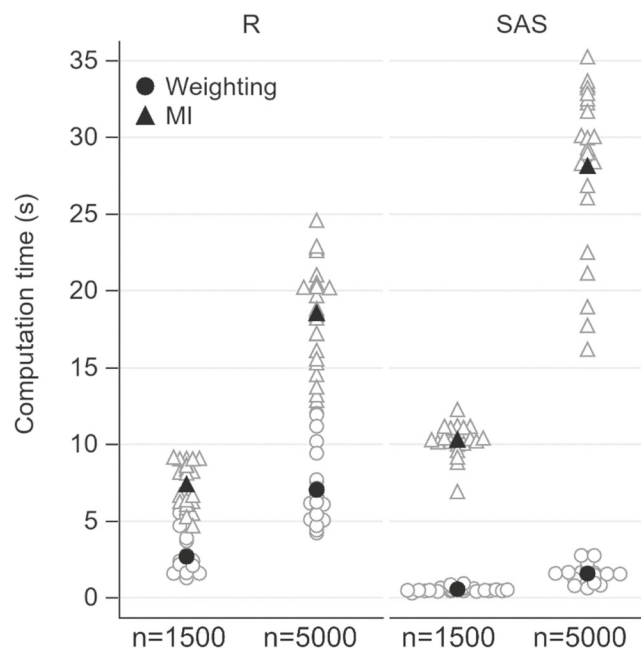
FIGURE 3 Computational time in seconds to implement MI (triangle) and weighting using UMLE (circle) from 22 scenarios by sample size and computer program. Filled in black symbol marks the mean. MI, multiple imputation; UMLE, unconstrained maximum likelihood estimator

as suggested by Sun and Tchetgen Tchetgen.[16] We combined patterns in two ways: (1) patterns 8 through 16 (each <1%, total 2.4%) resulting in eight total patterns and (2) patterns 6 through 16 (each <2%, total 5.7%) resulting in six total patterns. All observed variables were included in the missingness models. For the combined pattern, the model included variables observed across all patterns in the set: gestational age, nulliparity, and prior preterm birth. We implemented both UMLE and CBE (with diffuse priors $\gamma \sim N(0,100)$ and $c = 10^{-8}$) to estimate the weights. For CBE, we used Gelman-Rubin statistic (Rhat) to assess convergence[50] and estimated $\hat{\gamma}$ by the median across 3 chains of 120 000 iterations with the first half of samples discarded. Using $\hat{\gamma}$, we estimated the missingness weights for the complete cases and then fit the missingness-weighted treatment propensity score model to estimate the treatment weights. The final weight was the product of the two weights.

Third, we implemented MI by chained equations. We implemented MI four ways varying (1) whether spline terms were created before imputation (active imputation) or after imputation (passive imputation) and (2) imputation model flexibility. The imputation models included all variables from the analysis and the exposure-outcome interaction. To increase flexibility, we included interactions between the outcome and strong outcome predictors: prior preterm birth, maternal age, and HIV serostatus. We imputed missing data 20 times. In each imputed dataset, we fit the treatment propensity score model, estimated the treatment weights and then estimated the risk difference and robust variance from the treatment-weighted linear-binomial model. The 20 estimates were combined by Rubin's rule (as detailed above).

6.2 | Results

In the complete case analysis, the estimate of the effect of anemia on spontaneous preterm birth was 2.7 percentage points (95% CI—5.7, 11.0) (Table 7). Although anemia appeared to elevate the risk of preterm birth, the confidence interval was wide and encompassed effects ranging from strongly protective to strongly causative. While addressing missingness resulted in small changes in the point estimate, the confidence intervals remained wide so that these changes were not meaningful. Across the weighting approaches, the point estimates varied from 2.1 to 2.3. There was modest improvement in statistical efficiency with weighting compared to complete case analysis (SE: complete case 4.3, UMLE 4.0, CBE 4.1). Across the MI specifications, the point estimates varied more widely, from 1.9 to 3.0. The SEs ranged from 3.8 to 4.1.

TABLE 7 Risk difference estimates (in percentage points) and uncertainty from application examining effect of anemia on spontaneous preterm birth

Missing data approach[a]	RD	95% CI[b]	SE[b]
Complete case analysis	2.7	−5.7, 11.0	4.3
Weighting, UMLE, 6 patterns[c]	2.1	−5.8, 10.0	4.0
Weighting, UMLE, 8 patterns[d]	2.0	−5.9, 9.9	4.0
Weighting, CBE, 6 patterns[c]	2.3	−5.8, 10.3	4.1
Weighting, CBE, 8 patterns[d]	2.2	−5.8, 10.2	4.1
MI, transform/impute[e]	1.9	−5.7, 9.6	3.8
MI, transform/impute, more flexible[e,f]	3.0	−5.3, 11.2	4.1
MI, impute/transform[g]	2.4	−5.6, 10.5	4.0
MI, impute/transform, more flexible[g,f]	3.0	−5.0, 11.0	4.0

Abbreviations: CBE, constrained Bayesian estimator; CI, confidence interval; MI, multiple imputation; RD, risk difference in percentage points; SE, standard error; UMLE, unconstrainted maximum likelihood estimator.

[a] All approaches addressed confounding using inverse probability of treatment weights.

[b] Robust SE estimated by "robust" (Huber-White) sandwich estimator.

[c] Combined patterns 6 through 16 (each <2%) resulting in 6 total patterns.

[d] Combined patterns 8 through 16 (each <1%) resulting in 8 total patterns.

[e] Spline terms were created first and then missing data were imputed.

[f] More flexible imputation models included interactions between outcome and strong predictors of the outcome: prior preterm birth, maternal age, and HIV serostatus.

[g] Missing data were imputed first and then spline terms were created.

7 | DISCUSSION

In this work, we illustrated implementation of a novel weighting approach for nonmonotone missing data using data from a cohort (including code) and examined the approach's comparative performance to the most used approach, multiple imputation. This work provides an important demonstration of the utility of this novel approach and provides tools for its uptake. Our work supports the finding of Sun and Tchetgen Tchetgen that weighting is an alternative to MI to account for nonmonotone missingness. The UMLE estimator for missingness weights is easy to implement in commonly used software and intuitive, in contrast to prior approaches for weighting for nonmonotone missingness.[14-17] We provide code in R and SAS to support uptake of this approach.

Although theoretically UMLE may fail to converge if there are complete cases with a small probability of being a complete case, this was extremely rare in our simulation study (twice in 220 000 simulated cohorts). CBE is an alternative that can be used in these settings, though implementation is more challenging. Future work could examine characteristics of the datasets in which UMLE failed in an effort to anticipate failures. Even when UMLE converges, the weighted outcome model may fail, though, this was also rare in our simulation study (≤0.5% under MAR). Weighting failed when the complete case analysis also failed as both approaches fit the outcome model restricted to the complete cases. The complete case analysis could be used to assess whether weighting is likely to succeed. In our context, failure could also occur, for weighting or imputation, if the treatment propensity score model fails to converge or cannot be estimated due to chance non-positivity. Imputation may fail if the imputation models do not converge. Alternative approaches such as robust logistic regression or penalization could be used to address non-convergence of the propensity score or imputation models.[51] UMLE non-convergence was previously examined in a simple simulation by Sun and Tchetgen Tchetgen in which the parameters of interest were the parameters of a logistic model and only missingness weights were used.[16] They reported that UMLE failed in approximately 1% of simulation replicates. However, the authors did not say specifically whether these failures were due to UMLE non-convergence or non-convergence of the weighted outcome model. They also do not report failures of the other approaches.

Generally weighting is a semi-parametric approach; in contrast, the more widely-used MI is a fully parametric approach. The stronger parametric assumption improves precision of MI,[12] however, in our simulation, this improved efficiency was only notable at the smaller sample size with 15% exposure prevalence. This observed relationship with exposure prevalence may be because there was a lot of missingness in the exposure in our simulation (30% in primary

scenario marginally). In the other scenarios, there was little practical difference in efficiency between weighting and MI. Increasing the number of imputations may improve efficiency of MI.[52] To improve efficiency of weighting, augmented weighted estimators (also derived by Sun and Tchetgen Tchetgen) could be used.[16] We chose to use the non-augmented estimators in this work because they are easier to implement and thus more likely to be used in practice.

There were two instances where performance was notably different between weighting and MI. First, when data were MNAR, both MI and weighting were biased, however MI was less biased than weighting. It is not clear whether this is a characteristic of the estimators themselves or a product of our data generating mechanism. In Sun and Tchetgen Tchetgen's simulation, which examined different parameters of interest than examined in our simulation, there was not a consistent pattern of bias in the estimators across the parameters of interest under MNAR.[16] Second, weighting was computationally faster than MI. At the small sample sizes examined here, the difference was trivial, however with large sample sizes or when estimating the variance using resampling algorithms, the difference could be meaningful. Under linear extrapolation of our results, 1000 bootstrap resamples for $n = 5000$ is expected to take 7.8 h using MI and 0.4 h using weighting in SAS. In R, MI is expected to take 5.1 h and weighting 1.9 h.

In our empirical example, there was little change in the estimates across the approaches and results were impre-cise. The application highlights that implementing weighting and MI each require a number of analytic decisions. For weighting, some missing data patterns were rare, so we combined rare patterns. The model for this combined set could only include variables that were observed across all patterns in the set and therefore induced an assumption stronger than MAR. Combining patterns may induce bias, but such bias is likely small given that the combined patterns include few observations.[16] After estimation, it is good practice to examine the distribution of the weights as extreme weights can inflate the variance. Truncation of extreme weights can improve precision though potentially at the cost of bias.[53] Arguably, MI demands a greater number of analytic decisions.[12,54] Some of these decisions include: the iterative proce-dure (Markov chain Monte Carlo vs chained equations); transforming skewed variables; passive vs active imputation; number of iterations and imputed datasets; and specifying prior distributions.[11,54-64]

Both weighting and MI require correct model specification. In MI by chained equations, we specify a model for each variable with missingness. In weighting, we specify a model for each pattern. The number of potential patterns grows exponentially with the number of variables with missingness so weighting will often require specification of more mod-els than MI. The missingness models in weighting all have binomial dependent variables whereas the distribution of the dependent variables in the imputation models may take different distributions. Some imputation approaches assume a normal distribution and if this assumption is grossly violated, performance can be poor.[60] It may be necessary to trans-form variables before imputation. Alternatively, to reduce the strength of the correct model specification assumption, nonparametric approaches for imputation, for example, predictive mean matching,[57] can be used. Although it has been argued that it is easier to correctly specify the missingness models than the imputation models,[12] accurate specification of models depends on context-specific knowledge of what variables cause missingness and the functional form of those variables with missingness itself (for missingness models) or with the unobserved data (for imputation models). Previous work suggests that imputation may still perform well when imputation models are misspecified.[65] Future work is needed to investigate performance of weighting under model misspecification. For both MI and weighting, we can include aux-iliary variables to increase plausibility of the MAR assumption and potentially improve precision.[12,66] If the auxiliary variables also have missingness, then the complexity of the imputation model will increase and the number of missing-ness patterns/models for weighting will likely increase. Finally, an important limitation of MI is that specification of the imputation models can impose constraints on the model of interest (ie, the analyst must ensure the imputation models are *congenial* with the analysis model).[57] Issues of congeniality can arise by omitting the outcome or important interac-tions from imputation models, or using different functional forms of variables in imputation models and analysis models. Departures from congeniality can be avoided by specifying highly flexible imputation models or modifying the standard chained equations approach.[67,68] Congeniality is not a concern with weighting where specification of the missingness models is independent of the full data and therefore does not impose restrictions on the analysis model.

Our work has limitations. Our simulation design was relatively simple compared to most data settings and we did not vary all elements of the data generating mechanism. We mimicked the patterns observed in the application where missing exposure and outcome accounted for most of the missingness and results may differ as these patterns change. Other than the MCAR and MNAR scenarios, we did not vary the strength of relationships between observed data and missingness. We generated missingness in such a way that allowed us to control whether data were MAR. It is arguably more realistic to generate missingness by variable (as opposed to pattern) however such missingness would be MNAR. Future work should examine the performance of weighting under alternative generation mechanisms. Finally, we specifically focused on estimation of marginal causal effects where confounding is addressed via inverse probability of treatment weighting.

Comparative performance of weighting and MI may differ for other estimators (eg, g-computation). A strength of our work is that our simulation employed increased complexity over the only other simulation comparing Sun and Tchetgen Tchetgen's estimators with MI for nonmonotone missingness.[16] Additionally, in contrast to that previous simulation, we focused on estimating causal effects. We also provide code for the estimators in two commonly used software programs.

8 | CONCLUSION

Weighting can be used to address confounding and missing data simultaneously.[31] When missingness is nonmonotone, Sun and Tchetgen Tchetgen's UMLE is an easy to implement estimator for the missingness weights. Though less statistically efficient, weighting is an alternative to MI. Weighting's superior computational efficiency may be preferred with large sample sizes or when using resampling algorithms to estimate variance. As validity of weighting and MI rely on correct specification of different models, both approaches could be implemented to check agreement of results.

CONFLICT OF INTEREST STATEMENT
Authors have no conflict of interest relevant to this article.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are openly available in osf at https://osf.io/vzqdf/. Code is available on Github at https://github.com/rachael-k-ross/Wgt-NonmonotoneMiss. [Correction added on 15 September 2023, after first online publication: The data links in the Data Availability Statement have been updated in this version.]

ORCID
Rachael K. Ross [ORCID] https://orcid.org/0000-0002-2049-6918
Jessie K. Edwards [ORCID] https://orcid.org/0000-0002-1741-335X

REFERENCES
1. Eekhout I, de Boer MR, Twisk JWR, de Vet HCW, Heymans MW. Missing data: a systematic review of how they are reported and handled. *Epidemiology*. 2012;23(5):729-732.
2. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *Br J Cancer*. 2004;91(1):4-8. doi:10.1038/sj.bjc.6601907
3. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials*. 2004;1(4):368-376. doi:10.1191/1740774504cn032oa
4. Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; a review of the top medical journals. *BMC Med Res Methodol*. 2014;14(1):1-8. doi:10.1186/1471-2288-14-118
5. Sullivan TR, Yelland LN, Lee KJ, Ryan P, Salter AB. Treatment of missing data in follow-up studies of randomised controlled trials: a systematic review of the literature. *Clin Trials*. 2017;14(4):387-395. doi:10.1177/1740774517703319
6. Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. *Am J Epidemiol*. 2008;168(4):355-357. doi:10.1093/aje/kwn071
7. Westreich D. Berksons bias, selection bias, and missing data. *Epidemiology*. 2012;23(1):159-164. doi:10.1097/EDE.0b013e31823b6296
8. Ross RK, Breskin A, Westreich D. When is a complete-case approach to missing data valid? The importance of effect-measure modification. *Am J Epidemiol*. 2020;189(12):1583-1589. doi:10.1093/aje/kwaa124
9. Bartlett JW, Harel O, Carpenter JR. Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *Am J Epidemiol*. 2014;182(8):730-736. doi:10.1093/aje/kwv114
10. Choi J, Dekkers OM, le Cessie S. A comparison of different methods to handle missing data in the context of propensity score analysis. *Eur J Epidemiol*. 2019;34(1):23-36. doi:10.1007/s10654-019-00553-y

11. Perkins NJ, Cole SR, Harel O, et al. Principled approaches to missing data in epidemiologic studies. *Am J Epidemiol*. 2018;187(3):568-575. doi:10.1093/aje/kwx348

12. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res*. 2013;22(3):278-295. doi:10.1177/0962280210395740

13. Mansournia MA, Altman DG. Inverse probability weighting. *BMJ*. 2016;352:1-2. doi:10.1136/bmj.i189

14. Robins JM, Gill RD. Non-response models for the analysis of non-monotone non-ignorable missing data. *Stat Med*. 1997;16(1–3):21-37. doi:10.1002/(SICI)1097-0258(19970115)16:1<21::AID-SIM470>3.0.CO;2-F

15. Li L, Shen C, Li X, Robins JM. On weighting approaches for missing data. *Stat Methods Med Res*. 2013;22(1):14-30. doi:10.1177/0962280211403597

16. Sun BL, Tchetgen Tchetgen EJ. On inverse probability weighting for nonmonotone missing at random data. *J Am Stat Assoc*. 2018;113(521):369-379. doi:10.1080/01621459.2016.1256814

17. Sun BL, Perkins NJ, Cole SR, et al. Inverse-probability-weighted estimation for monotone and nonmonotone missing data. *Am J Epidemiol*. 2018;187(3):585-591. doi:10.1093/aje/kwx350

18. Castillo MC, Fuseini NM, Rittenhouse KJ, et al. Zambian preterm birth prevention study (ZAPPS): cohort characteristics at enrollment. *Gates Open Res*. 2019;2(25):1-18.

19. Price JT, Vwalika B, Rittenhouse KJ, et al. Adverse birth outcomes and their clinical phenotypes in an urban Zambian cohort. *Gates Open Res*. 2020;3(1533):1-26.

20. Scholl TO, Hediger ML. Anemia and iron-deficiency anemia: compilation of data on pregnancy outcome. *Am J Clin Nutr*. 1994;59(suppl):492-501. doi:10.1093/ajcn/59.2.492S

21. Klebanoff MA, Shiono PH, Selby JV, Trachtenberg AI, Graubard BI. Anemia and spontaneous preterm birth. *Am J Obstet Gynecol*. 1991;164:59-63. doi:10.1016/0002-9378(91)90626-3

22. March of Dimes, PMNCH, Save the Children, World Health Organization. Born Too Soon: The Global Action Report on Preterm Birth; 2012. doi:10.2307/3965140

23. Goldenberg RL, Rouse DJ. Prevention of premature birth. *N Engl J Med*. 1998;338(5):313-320.

24. Klebanoff MA, Shiono PH, Berendes HW, Rhoads GG. Facts and artifacts about anemia and preterm delivery. *Jama*. 1989;262(4):511-515. doi:10.1001/jama.1990.03440090030018

25. American college of obstetricians and gynecologists. Anemia in pregnancy: ACOG practice bulletin, number 233. *Obstetrics and Gynecology*. 2021;138(2):e55-e63.

26. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006;60:578-586. doi:10.1136/jech.2004.029496

27. Westreich D. *Epidemiology by Design: A Causal Approach to the Health Sciences*. New York, NY: Oxford University Press; 2020.

28. Westreich D, Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol*. 2010;171(6):674-677. doi:10.1093/aje/kwp436

29. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*. 2009;20(1):3-5. doi:10.1097/EDE.0b013e31818ef366

30. Robins JM, Hernán MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G, eds. *Longitudinal Data Analysis*. Boca Raton, Fl: Chapman & Hall/CRC; 2008:553-507.

31. Ross RK, Breskin A, Breger TL, Westreich D. Reflection on modern methods: combining weights for confounding and missing data. *Int J Epidemiol*. 2021;51:679-684. doi:10.1093/ije/dyab205

32. R Core Team. R: A Language and Environment for Statistical Computing; 2022. http://www.r-project.org/

33. Gelfand AE, Smith AFM, Lee T-M. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J Am Stat Assoc*. 1992;87(418):523. doi:10.2307/2290286

34. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. *Stat Med*. 2009;28:3049-3067. doi:10.1002/sim

35. Su Y-S, Yajima M. R2jags; 2021. https://cran.r-project.org/package=R2jags

36. Plummer M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing; 2003. https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

37. Chib S, Greenberg E. Understanding the Metropolis-Hastings algorithm. *Am Stat*. 1995;49(4):327-335.

38. Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat Med*. 2016;35(30):5642-5655. doi:10.1002/sim.7084

39. Hernán MA, Robins JM. *Causal Inference: What if*. Chapman & Hall/CRC; 2020. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

40. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89(427):846-866. doi:10.1080/01621459.1994.10476818

41. Efron B, Tibshirani R. Boostrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci*. 1986;1(1):54-77. https://projecteuclid.org/download/pdf_1/euclid.ss/1177013437

42. Johnson RW. An introduction to the bootstrap. *Teach Stat*. 2001;23(2):49-54. doi:10.1111/1467-9639.00050

43. Robertson SE, Dahabreh IJ, Steingrimsson JA. Using numerical methods to design simulations: revisiting the balancing intercept. *Am J Epidemiol*. 2021;191(7):1283-1289.

44. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res*. 2011;20(1):40-49. doi:10.1002/mpr

45. Raghunathan TE, Lepkowski JM, Van HJ, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol*. 2001;27(1):85-95.

46. van Buuren S. Mice: Multivariate Imputation by Chained Equations; 2021. https://cran.r-project.org/package=mice

47. Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J Am Stat Assoc*. 1986;81(394):366-374. doi:10.1080/01621459.1986.10478280

48. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102. doi:10.1002/sim.8086

49. Howe CJ, Cole SR, Westreich D, Greenland S, Napravnik S, Eron JJ. Splines for trend analysis and continuous confounder control. *Epidemiology*. 2011;22(6):874-875. doi:10.1097/EDE.0b013e31823029dd

50. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci*. 1992;7(4):457-511.

51. Greenland S. Bayesian perspectives for epidemiological research II regression analysis. *Int J Epidemiol*. 2007;36(1):195-202. doi:10.1093/ije/dyl289

52. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev Sci*. 2007;8(3):206-213. doi:10.1007/s11121-007-0070-9

53. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656-664. doi:10.1093/aje/kwn164

54. Allison PD. Handling missing data by maximum likelihood. SAS Glob Forum; 2012:1-21.

55. Von Hippel PT. How to impute interactions, squares, and other transformed variables. *Sociol Methodol*. 2009;39(1):265-291. doi:10.1111/j.1467-9531.2009.01215.x

56. Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Med Res Methodol*. 2012;12:1-13. doi:10.1186/1471-2288-12-46

57. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med*. 2011;30(4):377-399. doi:10.1002/sim.4067

58. Harel O, Mitchell EM, Perkins NJ, et al. Multiple imputation for incomplete data in epidemiologic studies. *Am J Epidemiol*. 2018;187(3):576-584. doi:10.1093/aje/kwx349

59. Allison PD. Imputation of categorical variables with PROC MI. Proceedings of the 30th Meeting of SAS Users Group International, vol. 113; 2005:1-14.

60. Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *Am J Epidemiol*. 2010;171(5):624-632. doi:10.1093/aje/kwp425

61. Horton NJ, Lipsitz SR, Parzen M. A potential for bias when rounding in multiple imputation. *Am Stat*. 2003;57(4):229-232. doi:10.1198/0003130032314

62. Graham JW. Missing data analysis: making it work in the real world. *Annu Rev Psychol*. 2009;60:549-576. doi:10.1146/annurev.psych.58.110405.085530

63. Bernaards CA, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat Med*. 2007;26:1368-1382. doi:10.1002/sim

64. Stuart EA, Azur MJ, Frangakis C, Leaf P. Multiple imputation with large data sets: a case study of the children's mental health initiative. *Am J Epidemiol*. 2009;169(9):1133-1139. doi:10.1093/aje/kwp026

65. Murray JS. Multiple imputation: a review of practical and theoretical findings. *Stat Sci*. 2018;33(2):142-159. doi:10.1214/18-STS644

66. Zalla LC, Yang JY, Edwards JK, Cole SR. Leveraging auxiliary data to improve precision in inverse probability-weighted analyses. *Ann Epidemiol*. 2022;74:75-83. doi:10.1016/j.annepidem.2022.07.011

67. Bartlett JW, Seaman SR. Multiple imputation of covariates by fully conditional specification : accommodating the substantive model. *Stat Methods Med Res*. 2015;24(4):462-487. doi:10.1177/0962280214521348

68. Beesley LJ, Bondarenko I, Elliot MR, Kurian AW, Katz SJ, Taylor JMG. Multiple imputation with missing data indicators. *Stat Methods Med Res*. 2021;30(12):2685-2700. doi:10.1177/09622802211047346

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.