Reducing Information and Selection Bias in EHR-Linked Biobanks via Genetics-Informed Multiple Imputation and Sample Weighting

Maxwell Salvatore[1,2], Ritoban Kundu[2,3], Jiacong Du[2,3], Christopher R Friese[3,4,5], Alison M Mondul[1,4], David Hanauer[6], Haidong Lu[7,8], Celeste Leigh Pearce[1,4], Bhramar Mukherjee[9]

[1] Department of Epidemiology, University of Michigan, Ann Arbor, MI, USA
[2] Center for Precision Health Data Science, University of Michigan, Ann Arbor, MI, USA
[3] Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA
[3] Rogel Cancer Center, University of Michigan, Ann Arbor, MI, USA
[4] Department of Systems, Populations, and Leadership, School of Nursing, University of Michigan, Ann Arbor, MI, USA
[5] Department of Health Management and Policy, University of Michigan, Ann Arbor, MI, USA
[6] Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, USA
[7] Section of General Internal Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA
[8] Public Health Modeling Unit, Yale School of Public Health, New Haven, CT, USA
[9] Department of Biostatistics, Yale University, New Haven, CT, USA

Corresponding author:
Bhramar Mukherjee
60 College St
New Haven, CT 06510
bhramar.mukerhjee@yale.edu
(203) 737-8644

Keywords (up to 5): electronic health records, missing data, exposome, biobank, selection bias

ABSTRACT

Electronic health records (EHRs) are valuable for public health and clinical research but are prone to many sources of bias, including missing data and non-probability selection. Missing data in EHRs is complex due to potential non-recording, fragmentation, or clinically informative absences. This study explores whether polygenic risk score (PRS)-informed multiple imputation for missing traits, combined with sample weighting, can mitigate missing data and selection biases in estimating disease-exposure associations. Simulations were conducted for missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) conditions under different sampling mechanisms. PRS-informed multiple imputation showed generally lower bias, particularly when combined with sample weighting. For example, in biased samples of 10,000 with exposure and outcome MAR data, PRS-informed imputation had lower percent bias (3.8%) and better coverage rate (0.883) compared to PRS-uninformed (4.5%; 0.877) and complete case analyses (10.3%; 0.784) in covariate-adjusted, weighted, multiple imputation scenarios. In a case study using Michigan Genomics Initiative (n=50,026) data, PRS-informed imputation aligned more closely with a sample-weighted All of Us-derived benchmark than analyses ignoring missing data and selection bias. Researchers should consider leveraging genetic data and sample weighting to address biases from missing data and non-probability sampling in biobanks.

Electronic health records (EHRs) represent a rich, longitudinal resource that researchers increasingly use to address questions of public health and clinical significance. EHR-linked biobanks, which often contain genetic information linked to other data sources (e.g., administrative and insurance claims, neighborhood-level characteristics, and complementary survey data), are growing in both the number of participants (n) and the breadth of measured variables (p). However, EHR data has not been collected for research purposes, so researchers must carefully consider potential biases (i.e., systematic errors). Potential sources of bias include missing data[1–3] (including clinically informative visiting processes[2,4,5]), selection bias,[6–8] misclassification,[7,9,10] confounding,[11,12] immortal time bias,[13,14] and clinical practice and data collection and processing heterogeneity across EHRs.[15,16] Although the advent of large-scale secondary data (colloquially, "Big Data"[17]) effectively minimizes the threat of random error, systematic sources of bias are ever-present adversaries, unphased by ever-increasing sample sizes. In fact, large sample sizes amplify these biases relative to the very small variance, frequently making inference erroneous, a phenomenon commonly characterized as the Big Data Paradox.[18]

Missing data is ubiquitous in epidemiology[19–23] and almost universally encountered in health research.[24,25] Complete case analyses, which ignore observations with missing data for variables of interest (e.g., exposures, outcomes, or important covariates), are the most commonly employed approach in randomized clinical trials[25] and observational studies[24] in the presence of missing data. However, complete case analyses can lead to biased parameter estimation depending on the missing data mechanism, or the reason why the data are missing.[1,19,26]

Missing data mechanisms broadly fall into three classes: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).[27] Naïve complete-case analyses are expected to produce unbiased results when data are MCAR. However, there are several reasons that make the assumption that missing data are MCAR in EHR-linked biobanks less reasonable,[6,30–32] including non-random patient-provider interactions,[33–35] clinically informative observation processes,[2,36] and EHR fragmentation. For these reasons, MAR and MNAR assumptions are more plausible.[37] Among the existing missing data methods that improve precision and reduce bias (e.g., inverse probability weighting[38–40] and full-information maximum likelihood[41–44]), multiple imputation is a commonly used and frequently recommended approach for handling missing data in EHR-linked biobanks.[3,37,45–47] It is important to note that MNAR data cannot be empirically distinguished from MAR data while the MCAR assumption can be tested.[48,49]

A hallmark of major biobanks is availability of genetic data on a large fraction of participants and an active genetics research community producing polygenic risk scores (PRS) for a variety of traits.[50,51] It is an interesting question whether PRS observed on a large sample can improve imputation of the traits they are constructed for. The actual traits may be missing for a large number of participants, and PRS can serve as a weak proxy.[52,53]

Adding to the missing data challenge is the fact that EHR-linked biobanks often do not represent their source (or target) population, introducing potential selection bias.[54] Recruitment mechanisms like recruiting patients awaiting surgery (as in the Michigan Genomics Initiative (MGI)[55]) and oversampling groups historically

underrepresented in biomedical research (as in the NIH All of Us Research Program[56]) as well as participant-driven factors like healthy volunteer bias (as in the UK Biobank[57,58]) explain differences between the study cohorts and their underlying source and target populations.[57,59] Weighting-based methods like inverse probability weighting and poststratification weighting are often employed to reduce selection bias in parameter estimation when individual or summary data from an external non-probability sample are available. Recent papers have shown that weighted analyses reduce (but do not remove) bias due to selection in EHR-linked biobanks.[58,60,61]

In this study we considered, to the best of our knowledge, an unexplored question: can PRS-informed multiple imputation reduce bias due to missing exposure data in association estimation? We investigated (a) whether PRS-informed multiple imputation meaningfully reduces bias due to missing data in probability samples and (b) the joint impact of PRS-informed multiple imputation and sample weighting on exposure-outcome association estimation in biased samples (Figure 1). We calculated unweighted and weighted complete case- and multiple imputation-based estimates of the body mass index (BMI) coefficient for glucose in realistic simulations, followed by a case study stratified by non-Hispanic White (n=42,999) and non-Hispanic Black (n=2,297) status in MGI. First, our simulation studies explored the joint impacts of multiple imputation with and without exposure and outcome PRS for missing data in MCAR, MAR, and MNAR settings and weighting in random and biased samples. We considered sampling weights in biased samples. Our case study applied these methods to MGI data to estimate the BMI coefficient for glucose using the same missing data methods and stratum-specific selection weights (as described previously[60]) to

demonstrate differences in association estimates under different analytical strategies in real-world data relative to National Health Interview Survey-weighted All of Us-based benchmark.

METHODS

Simulation Design

Generating outcome, exposure, covariates, and polygenic risk scores jointly

We simulated 1,000 replicates of a pseudo-population with size 100,000 (Figure 2). To achieve this, we first generated an 8-dimensional multivariate normal distribution, $\mathbf{X} \sim \mathcal{N}_8(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, mimicking the joint distribution of age, sex, non-Hispanic White (NHW), smoking status (ever/never), BMI, glucose, BMI PRS, and glucose PRS, assuming mean standardized variables ($\boldsymbol{\mu} = 0$) and $\boldsymbol{\Sigma}$ as observed in MGI (see Eq.1 below).

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.998 & 0.209 & 0.254 & -0.177 & 0.060 & 0.092 & 0.084 & 0.092 \\ 0.209 & 0.997 & 0.051 & 0.020 & -0.006 & 0.022 & 0.301 & -0.005 \\ 0.254 & 0.051 & 0.923 & -0.126 & 0.160 & 0.165 & -0.052 & -0.004 \\ -0.177 & 0.020 & -0.126 & 1.001 & -0.038 & -0.115 & 0.022 & -0.008 \\ 0.060 & -0.006 & 0.160 & -0.038 & 0.907 & 0.078 & -0.033 & 0.023 \\ 0.092 & 0.0217 & 0.165 & -0.115 & 0.078 & 1.005 & 0.076 & 0.001 \\ 0.084 & 0.301 & -0.052 & 0.022 & -0.033 & 0.076 & 0.990 & 0.061 \\ 0.092 & -0.005 & -0.004 & -0.008 & 0.023 & 0.001 & 0.061 & 1.005 \end{pmatrix} \quad \textit{(Eq.1)}$$

Binary variables were recoded (sex, NHW, smoking status) from the generated continuous variables such that they preserved their correlation with age in observed MGI data.

Sample selection

For each pseudo-population, we performed sampling under two scenarios: random and biased/covariate-informed. Covariate-informed sampling probabilities depended on observed age, glucose (the outcome), and BMI (the exposure) (e.g., $logit\big(P(S = 1|age, glucose, BMI)\big) = \gamma_0 + \gamma_{age} age + \gamma_{glucose} glucose + \gamma_{BMI} BMI$, where

S is an indicator variable for selection into the sample; $\gamma_{age}, \gamma_{glucose}, \gamma_{BMI} = 1$). For each scenario, the intercept, γ_0, was selected to draw a sample of approximately 1,000, 2,500, 5,000, and 10,000 (γ_0 = -6.92, -5.83, -4.94, -3.93, respectively) from the pseudo-population where individual i had selection probability $P\left(S_i = 1 | age_i, glucose_i, BMI_i\right)\big)$ dependent on the exposure (BMI) and the outcome (glucose) as well as the covariate (age).

Missingness generation

We simulated (a) exposure only (i.e., BMI) and (b) exposure and outcome (i.e., BMI and glucose) missingness under MCAR, MAR, and MNAR mechanisms for each selected sample size and mechanism. Approximately 25% missingness was generated for each variable. Under MCAR, the probability of missingness (e.g., $P(R_{BMI} = 1)$ where R_{BMI} is an indicator for whether BMI is missing) was 25% for all observations. Under MAR, exposure missingness depended on the outcome (glucose) and covariates (age, sex, race/ethnicity, smoking status) while outcome (glucose) missingness depended only on covariates. Under MNAR, exposure and outcome missingness was dependent on the whole set of exposure, outcome, and covariates. In all settings, all non-intercept regression coefficients were set equal to 1 and only the intercept was tuned to attain the desired sample size. Supplementary Table 1 shows the intercept coefficient values by missingness mechanism to (approximately) achieve the desired sample sizes.

Analytic Methods

For each scenario, we performed unweighted and weighted analyses (for simple random sampling, these are equivalent). The weights were proportional to the inverse of the *known* covariate-informed sampling probabilities for each individual i ($\omega_i \propto$

$P(S_i = 1|age_i, glucose_i, BMI_i)^{-1}$), and weighted analyses were carried out using the survey R package (version 4.4-2).[62] In addition to complete case analysis, we performed multiple imputation to address missing data. For each sample, multiple imputation methods using age, sex, NHW, smoking status, BMI, glucose (without PRS; woPRS-imputed) and additionally exposure (BMI) and outcome (glucose) PRS (PRS-imputed) were carried out using the R package mice (version 3.16.0; m = 5 imputations).[63,64] Beta coefficients across imputations were pooled using Rubin's rule, with confidence intervals calculated from pooled standard errors based on within and between imputation variances.[65,66] For multiply imputed analyses of biased samples, weighted analyses were conducted on each imputed dataset before pooling.

Our target quantity was the true coefficient of BMI in a linear regression model for glucose (β_{BMI}) adjusted for age, sex, NHW race/ethnicity, and smoking status (ever/never) (Eq. 2).

$$Glucose_i = \beta_0 + \beta_{BMI} BMI_i + \beta_{age} age_i + \beta_{sex} sex_i + \beta_{NHW} NHW_i + \beta_{smoke} smoke_i + \epsilon_i \quad \text{(Eq.2)}$$

$$\text{where } \epsilon_i \sim N(0, \sigma^2)$$

For each replicate, the true β_{BMI} was obtained from the pseudo-population of size 100,000 and the sample estimates were obtained in the selected samples of sizes 1,000, 2,500, 5,000, and 10,000. In each sample, we conducted unweighted and weighted complete case, woPRS-imputed, and PRS-imputed analyses, extracting the coefficient estimate of BMI for glucose ($\hat{\beta}_{BMI}$). We evaluated association estimation properties using percent bias, coverage rate, average 95% confidence interval width, and root mean square error (RMSE), averaged over the 1,000 replicates.

Case study: Michigan Genomics Initiative

Description of the study cohort

MGI is an EHR-linked biobank that began in 2012, initially recruiting adult patients through pre-/peri-operative appointments requiring anesthesia from the University of Michigan Health System. As of September 2023, ~100,000 consented participants have provided access to their EHR and a biospecimen for genotyping, with a recent follow-up effort collecting complementary survey data.[67] This paper included 42,999 (25,520 complete cases) non-Hispanic White and 2,297 (1,240 complete cases) non-Hispanic Black participants aged 40 or older without a diabetes diagnosis and with demographic, health measurement, laboratory, and polygenic risk score data. MGI protocols were reviewed and approved by the University of Michigan Medical School Institutional Review Board (IRB ID HUM00099605 and HUM00155849).

Outcome, exposure, covariates, and polygenic risk score

The outcome and exposure of interest were glucose (mg/dL; logical observation identifiers names and codes (LOINC) code: 2345-7) and BMI (kg/m^2), respectively. The longitudinal data in the EHR was reduced to the participant's median value after removing extreme values (values outside 1.5x the interquartile range) for the corresponding variable. Age was considered the participant's age at the time of data pull (March 23, 2022). Sex (indicator for female) and race/ethnicity were obtained from EHR data. Multiple measurements of self-reported smoking status were recorded and recoded into a binary ever/never indicator variable.

Ma and colleagues previously calculated several PRS for 27 exposures in MGI participants.[51] In this paper, we selected the Lassosum PRS for BMI and the

deterministic Bayesian sparse linear mixed model PRS for glucose because they had the highest R^2 value for their respective traits in the published paper.[51] These PRSs relied on publicly available GWAS summary statistics of UK Biobank data (Neale lab[68,69]). Both PRSs were predictive in MGI, with BMI PRS being much stronger (Pearson correlation between BMI and BMI PRS: 0.30; glucose and glucose PRS: 0.09; Supplementary Figure 1).

Estimated regression coefficient corresponding to BMI with glucose as the outcome

The target estimand of interest was the regression coefficient corresponding to BMI with glucose as the outcome. We conducted analyses among individuals 40 and older without a diabetes diagnosis in non-Hispanic White and non-Hispanic Black strata as well as in the full (i.e., unstratified) cohort. Unlike in the simulations, the selection weights in MGI were *not known*. Salvatore and colleagues estimated inverse probability selection weights to make MGI more representative of the US adult population using National Health Interview Survey data.[60] Using the same methods to calculate stratum-specific weights, we conducted weighted versions of each regression analysis. Using the non-Hispanic White and non-Hispanic Black samples with missing data (n=42,999 and 2,297, respectively), we performed multiple imputation with and without PRS (adjusting for age, sex, and smoking status). We also conducted a PRS-informed multiple imputation analysis where observations were restricted to only those with observed PRS (PRS-imputed (subset): n=25,520 and 1,240 for non-Hispanic Whites and non-Hispanic Blacks, respectively). We reported the estimated beta coefficients and 95% confidence intervals.

Software

Analyses were conducted using R version 4.3.3. The code used to conduct analyses in this paper is available at https://github.com/maxsal/exprs_imputation.

RESULTS

Simulation study

In **random sampling with exposure-only missingness**, when BMI data was MCAR (Figure 3), all analyses successfully maintained the nominal 95% coverage rates and exhibited no bias in the estimated BMI coefficients for glucose. However, when the data was MAR, complete case analysis showed a decline in coverage rates as sample size increased, and consistently exhibited bias (e.g., 8.86% for n=1,000; 7.80% for n=10,000). woPRS-imputed analyses, in contrast, provided more stable coverage rates (e.g., 0.931 for n=1,000; 0.924 for n=10,000) and reduced bias (e.g., 2.89% for n=1,000; 0.10% for n=10,000) as sample sizes grew. PRS-imputed analyses outperformed both, fully retaining the nominal coverage rate across all sample sizes and achieving the least bias (e.g., 1.5% for n=1,000; 0.04% for n=10,000). Under MNAR conditions, none of the analyses could maintain the nominal coverage rate, and all exhibited substantial bias exceeding 30%. However, PRS-imputed analyses performed slightly better than woPRS-imputed analyses, achieving marginally higher coverage rates (e.g., for n=1,000: 0.637 for PRS-imputed; 0.561 for woPRS-imputed) and lower bias (e.g., for n=1,000: 31.95% for PRS-imputed; 36.12% for woPRS-imputed).

In **biased sampling scenarios where only exposure data were missing**, MCAR conditions led to significant bias and a failure to retain nominal coverage across all unweighted analyses, with the bias worsening as sample sizes increased (e.g.,

>29%; Figure 4). When the missingness was MAR or MNAR, unweighted complete case analyses exhibited less bias than multiple imputation methods, likely due to the amplification of biases by multiple imputation when it does not account for sampling weights. However, when analyses were weighted, multiple imputation approaches, particularly PRS-imputed, showed substantial improvements in coverage rates and bias reduction, consistently outperforming complete case analyses. For instance, in a 10,000-observation sample with MAR missingness, the coverage rate was 0.784 for complete case analysis, 0.877 for woPRS-imputed, and 0.883 for PRS-imputed. When data was MNAR, coverage decreased, and bias increased across all methods, but PRS-imputed analyses performed slightly better, with reduced bias as sample size increased.

In **random sampling scenarios where both exposure and outcome data were missing,** MCAR conditions allowed all analyses to maintain the nominal coverage rates and remain unbiased (Figure 3). For example, the coverage rate for complete case analysis was 0.938 for n=1,000 and remained above 0.950 for larger sample sizes. However, under MAR conditions, coverage rates decreased, and bias remained stable with larger sample sizes (e.g., coverage rate dropped from 0.925 for n=1,000 to 0.784 for n=10,000) in complete case analyses. woPRS-imputed and PRS-imputed analyses effectively returned to nominal coverage rates in MAR data (e.g., 0.946 and 0.947, respectively, for n=10,000) and exhibited little to no bias (e.g., 1.32% for woPRS-imputed; 0.97% for PRS-imputed). Under MNAR conditions, no analysis method could maintain nominal coverage, and all showed significant bias. Nonetheless, PRS-imputed analyses slightly outperformed others, achieving better coverage rates (e.g., 0.299 for

PRS-imputed versus 0.000 for complete case in n=1,000) and lower bias (e.g., 53.89% for PRS-imputed versus 62.27% for complete case in n=1,000). Plots describing average 95% CI width and RMSE are shown in Supplementary Figure 2.

In **biased sampling with both exposure and outcome missingness**, unweighted PRS-imputed analyses failed to recover nominal coverage and exhibited significant bias across all missingness mechanisms, with MCAR data showing the least bias (Figure 4). When sampling weights were applied, PRS-imputed multiple imputation analyses performed better than complete case analyses, with improved coverage rates, particularly in MAR data. For example, PRS-imputed coverage rates for MAR data exceeded those observed for MCAR as sample sizes increased (e.g., 0.840 for n=1,000 to 0.906 for n=10,000). Despite these improvements, MNAR data analyses remained problematic across all methods, with PRS-imputed analyses showing slightly better performance but still exhibiting considerable bias and suboptimal coverage (e.g., percent bias of 32.40% for n=1,000 and 23.19% for n=10,000; plots depicting average 95% CI width and RMSE are shown in Supplementary Figures 3 and 4).

Analysis in the Michigan Genomics Initiative (MGI)

Descriptive characteristics of the study population

We analyzed a cohort of 50,026 MGI participants aged 40 or older without diabetes, of which 54.5% were female and 86.0% non-Hispanic White. The mean age was 62.9 years (SD: 12.5), BMI was 29.1 (6.0), and glucose was 99.0 mg/dL (14.1) (Supplementary Table 8). Due to suspected racial/ethnic heterogeneity,[70,71] we stratified the analysis into non-Hispanic White and non-Hispanic Black groups.

Among the 42,999 non-Hispanic White participants, 53.8% were female, with a mean age of 63.5 years (12.5), BMI of 29.1 (6.0), and glucose of 99.5 mg/dL (14.2) (Table 1). Participants with missing data were generally younger (mean age 62.6 vs. 64.1 years), more likely to be female (54.6% vs. 53.3%), less likely to have smoked (46.5% vs. 50.7%), and had slightly lower glucose levels (99.47 vs. 99.55 mg/dL), with no differences in BMI PRS (p=0.6) or glucose PRS (p=0.4) compared to those with complete data.

Among the 2,297 non-Hispanic Black participants, 63.3% were female, with a mean age of 57.8 years (11.4), BMI of 30.8 (6.4), and glucose of 95.1 mg/dL (12.5) (Table 2). Those with missing data were less likely to have smoked (39.2% vs. 44.7%) and had lower glucose levels (94.0 vs. 95.8 mg/dL), with no differences in BMI PRS (p=0.8) or glucose PRS (p=0.061) compared to complete cases.

Subsets of 30,492 non-Hispanic Whites and 1,437 non-Hispanic Blacks had complete PRS data. Smoking status and glucose showed moderate missingness (14% and 11% in the non-Hispanic White sample and 13% and 7% in the non-Hispanic Black, respectively), while BMI was rarely missing in both groups (0.5% in non-Hispanic Whites and 1.1% in non-Hispanic Blacks).

Estimation of the coefficient for BMI with glucose as the outcome

Among non-Hispanic White individuals aged 40 years or older without diabetes in the MGI cohort, the unweighted, covariate-adjusted, complete case coefficient estimate was 0.288 (0.264, 0.312) (Figure 5), which differed from the benchmark range of 0.376 to 0.423, derived from NHIS-weighted All of Us data. Using sampling weights, the estimate improved to 0.324 (0.283, 0.365), aligning more closely with the benchmark.

Multiple imputation alone also improved the estimates, with woPRS-imputed and PRS-imputed estimates rising to 0.300 (0.277, 0.323) and 0.302 (0.280, 0.324), respectively. When these imputation methods were combined with sample weighting, they approached the benchmark even more closely, with final estimates of 0.331 (0.292, 0.371) for woPRS-imputed and 0.338 (0.280, 0.324) for PRS-imputed analyses. *Overall, sample weighting and multiple imputation consistently brought estimates closer to the benchmark*. Similar trends were seen in the corresponding unadjusted analyses in Figure 5A.

In non-Hispanic Black individuals aged 40 years or older without diabetes, the unweighted complete case estimate was 0.178 (0.097, 0.261) (Figure 5), also differing from the benchmark range of 0.196 to 0.297. However, applying sampling weights brought the estimate within the benchmark range at 0.202 (0.086, 0.317). Multiple imputation alone saw nominal increases in estimates, with woPRS-imputed and PRS-imputed estimates at 0.204 (0.119, 0.288) and 0.203 (0.116, 0.290), respectively. The weighted analyses produced similar results, with estimates of 0.200 (0.082, 0.318) for woPRS-imputed and 0.214 (0.096, 0.332) for PRS-imputed analyses. Unlike the non-Hispanic White group, weighting had a smaller impact because the estimates were already within the benchmark range.

In the unstratified results for the entire MGI cohort aged 40 years or older without diabetes (Supplementary Figure 5), which was predominantly non-Hispanic White (86%), the findings mirrored those of the non-Hispanic White stratum. For example, the weighted PRS-imputed estimate (0.312 (0.274, 0.349)) was closer to the benchmark

range of 0.346 to 0.386 than the unweighted complete case estimate (0.277 (0.255, 0.299)).

DISCUSSION

We investigated the combined impact of missing data and selection bias on association estimates using simulation studies and real-world EHR data. Our study shows that biobanks with genetic data can reduce these biases by incorporating genetic summaries of exposures and outcomes. Building on previous research,[3,51] we assessed the effectiveness of PRS-informed multiple imputation in improving association estimates, and to examine the interaction between multiple imputation and sample weighting using simulations and a case study. To our knowledge, this is the first paper to explore the joint impacts of genetic-informed multiple imputation and sample weighting methods in EHR-linked biobank data.

Our simulations revealed that PRS-informed multiple imputation generally outperformed standard methods, particularly for MAR and MNAR data, by offering smaller confidence intervals (Supplementary Figure 3) and reduced bias. PRS preserve correlation between underlying traits despite often being weakly predictive of the trait itself.[51] Because PRS are observed on a large fraction of the sample, they may help with selection and non-response biases. However, while PRS-imputed analyses improved coverage rates and reduced bias compared to complete case analyses, as expected, [26,65] they did not fully recover the nominal coverage rate, especially under MNAR conditions. Notably, PRS-imputed analyses also demonstrated the lowest RMSE in MAR scenarios, suggesting better estimation accuracy.

Missingness in EHR-linked biobank data often deviate from the MCAR assumption due to factors such as patient health status, healthcare access, and EHR fragmentation, leading to biased observation processes.[2,4,6,35,47,72–77] PRS-informed imputation performed best but struggled to achieve nominal coverage or bias reduction when data were MNAR. While correlations between exposures and their PRS are weak (Supplementary Figure 1), stronger correlates would likely improve multiple imputation.

EHR-linked biobank data are subject to selection bias, including due to healthy volunteer bias (as in the UK Biobank[57]) or non-random recruitment strategies (as in MGI[55] and the NIH All of Us Research Program[56]). When simulating selection bias by oversampling by age, BMI, and glucose, all methods showed substantial bias in unweighted analyses (≥21%). Weighting improved PRS-imputed analyses' performance, especially for MAR data, significantly reducing bias and nearly restoring nominal coverage rates (e.g., woPRS-imputed vs. PRS-imputed coverage rate for n=1,000: 0.842 vs. 0.840; n=10,000: 0.884 vs. 0.906). PRS-imputed methods only slightly improved bias and RMSE for MNAR data (e.g., n=1,000: 33.67% vs. 32.40%; n=10,000: 25.46% vs. 23.19%).

In the case study using MGI data, we estimated the BMI coefficient for glucose and found small differences between complete case and imputed estimates, likely due to low levels of missingness (non-Hispanic Whites and Blacks: glucose: 14% and 13%; BMI: 0.5% and 1.1%). However, accounting for selection bias resulted in more substantial changes, underscoring its greater impact than missing data.

Our findings suggest that while PRS-informed multiple imputation can enhance the accuracy of association estimates, particularly in MAR scenarios, it does not fully

address challenges when data are MNAR. Sensitivity analyses, alongside expert knowledge[78,79] and tools like m-graphs or m-DAGs,[80–83] are recommended and methods like Heckman imputation[84,85] and pattern-mixture models[86,87] can be explored. For most regression models, complete case analyses can give unbiased results when the probability of being a complete case is independent of the outcome after taking covariates into account, regardless of the missingness mechanism (Supplementary Table 9).[88–90] Combining PRS-informed multiple imputation with sampling weights can reduce bias and improve coverage, but careful consideration of underlying data-generating mechanisms is essential.

Strengths and limitations

This study emphasizes the need to address missing data and selection bias in EHR-linked biobanks and suggests actions for researchers. Our simulations highlight the effectiveness of multiple imputation combined with weighting methods. However, our study has limitations. First, our simulations considered a single level of missingness (~25%) in two scenarios: exposure alone and exposure and outcome, whereas in practice, multiple patterns and levels of missingness can affect exposures, outcomes, and covariates simultaneously. Future studies should explore a wider range of missingness scenarios. Second, selection bias varies across EHR-linked biobanks due to differing recruitment strategies. For instance, MGI has notable selection biases relative to the US adult population, which may not be as pronounced in population-based biobanks like the NIH All of Us Research Program or the UK Biobank. Third, our case study examined a single association parameter with a relatively small level of missingness and without a gold standard estimate. Future research should investigate

associations where gold standard estimates are available. Fourth, the study focused on glucose levels, which can be collected without fasting conditions and managed with medication, complicating interpretation. Codes specifying fasting conditions were rarely used and thus not considered in our analyses, and we did not consider other factors that might impact glucose levels (e.g., surgery, metformin use in people with pre-diabetes). Fifth, we examined the association between two continuous variables after collapsing longitudinal data, whereas many studies utilize binary outcomes and longitudinal data. Future work should address these data types. Lastly, clinically informative visiting processes in EHR data increase the likelihood of MNAR data. Although PRS-imputed analyses showed some improvements for MNAR data, future research should incorporate methods that specifically model these processes.[91–94]

Conclusion

Missing data is a critical issue in EHR-linked biobank data. We leveraged non-missing genetic data – a key feature of biobanks – to assess if PRS-informed multiple imputation could reduce bias in association estimation. Our simulations demonstrated a substantial reduction in bias for MAR data when incorporating genetic information. Using real-world MGI data, selection bias was relatively more impactful than missing data. Our findings call for exploring additional missingness patterns and levels across associations. Biobanks should provide PRS for common exposures available as proxies and sampling weights to address selection bias. This approach will help researchers better mitigate multiple biases in EHR-linked biobank association analyses, enhancing the reliability and validity of their findings.

ACKNOWLEDGMENTS

Michigan Genomics Initiative: The authors acknowledge the Michigan Genomics Initiative participants, Precision Health at the University of Michigan, the University of Michigan Medical School Central Biorepository, the University of Michigan Medical School Data Office for Clinical and Translational Research, and the University of Michigan Advanced Genomics Core for providing data and specimen storage, management, processing, and distribution services, and the Center for Statistical Genetics in the Department of Biostatistics at the School of Public Health for genotype data curation, imputation, and management in support of the research reported in this publication/grant application/presentation.

Funding

This work was funded by National Cancer Institute grant P30CA046592 and the Training, Education, and Career Development Graduate Student Scholarship of the University of Michigan Rogel Cancer Center.

Data-availability statement

Patient confidentiality prevents the sharing of data publicly. However, the data underlying the study's results are available from the Michigan Genomics Initiative at https://precisionhealth.umich.edu/ourresearch/michigangenomics/ for researchers who meet the criteria for confidential data access. The code used to conduct analyses in this paper is publicly available at https://github.com/maxsal/exprs_imputation.

Conflict of interest

The authors have no financial or non-financial conflicts of interest related to this research.

Institutional approval

The institutional review board of the University of Michigan Medical School gave ethical approval for this work (HUM00155849).

REFERENCES

1. Pedersen A, Mikkelsen E, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol*. 2017;Volume 9:157-166. doi:10.2147/CLEP.S129785

2. Tan ALM, Getzen EJ, Hutch MR, et al. Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record? *J Biomed Inform*. 2023;139:104306. doi:10.1016/j.jbi.2023.104306

3. Li R, Chen Y, Moore JH. Integration of genetic and clinical information to improve imputation of data missing from electronic health records. *J Am Med Inform Assoc*. 2019;26(10):1056-1063. doi:10.1093/jamia/ocz041

4. Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. *Diagn Progn Res*. 2020;4(1):8. doi:10.1186/s41512-020-00077-0

5. Goldstein BA, Phelan M, Pagidipati NJ, Peskoe SB. How and when informative visit processes can bias inference when using electronic health records data for clinical research. *J Am Med Inform Assoc*. 2019;26(12):1609-1617. doi:10.1093/jamia/ocz148

6. Haneuse S, Daniels M. A General Framework for Considering Selection Bias in EHR-Based Studies: What Data are Observed and Why? *EGEMs Gener Evid Methods Improve Patient Outcomes*. 2016;4(1):16. doi:10.13063/2327-9214.1203

7. Beesley LJ, Mukherjee B. Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics*. Published online December 3, 2020:biom.13400. doi:10.1111/biom.13400

8. Lu H, Howe CJ, Zivich PN, Gonsalves GS, Westreich D. The Evolution of Selection Bias in the Recent Epidemiologic Literature—A Selective Overview. *Am J Epidemiol*. Published online August 12, 2024:kwae282. doi:10.1093/aje/kwae282

9. Young JC, Conover MM, Jonsson Funk M. Measurement Error and Misclassification in Electronic Medical Records: Methods to Mitigate Bias. *Curr Epidemiol Rep*. 2018;5(4):343-356. doi:10.1007/s40471-018-0164-x

10. Tong J, Huang J, Chubak J, et al. An augmented estimation procedure for EHR-based association studies accounting for differential misclassification. *J Am Med Inform Assoc*. 2020;27(2):244-253. doi:10.1093/jamia/ocz180

11. Lu CY. Observational studies: a review of study designs, challenges and strategies to reduce confounding. *Int J Clin Pract*. 2009;63(5):691-697. doi:10.1111/j.1742-1241.2009.02056.x

12. Streeter AJ, Lin NX, Crathorne L, et al. Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *J Clin Epidemiol*. 2017;87:23-34. doi:10.1016/j.jclinepi.2017.04.022

13. Tyrer F, Bhaskaran K, Rutherford MJ. Immortal time bias for life-long conditions in retrospective observational studies using electronic health records. *BMC Med Res Methodol*. 2022;22(1):86. doi:10.1186/s12874-022-01581-1

14. Prada-Ramallal G, Takkouche B, Figueiras A. Bias in pharmacoepidemiologic studies using secondary health care databases: a scoping review. *BMC Med Res Methodol*. 2019;19(1):53. doi:10.1186/s12874-019-0695-y

15. Fu S, Leung LY, Raulli AO, et al. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. *BMC Med Inform Decis Mak*. 2020;20(1):60. doi:10.1186/s12911-020-1072-9

16. Glynn EF, Hoffman MA. Heterogeneity introduced by EHR system implementation in a de-identified data resource from 100 non-affiliated organizations. *JAMIA Open*. 2019;2(4):554-561. doi:10.1093/jamiaopen/ooz035

17. Japec L, Kreuter F, Berg M, et al. Big Data in Survey Research: AAPOR Task Force Report. *Public Opin Q*. 2015;79(4):839-880. doi:10.1093/poq/nfv039

18. Meng XL. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann Appl Stat*. 2018;12(2). doi:10.1214/18-AOAS1161SF

19. Cole SR, Zivich PN, Edwards JK, et al. Missing Outcome Data in Epidemiologic Studies. *Am J Epidemiol*. 2023;192(1):6-10. doi:10.1093/aje/kwac179

20. Howe CJ, Cain LE, Hogan JW. Are All Biases Missing Data Problems? *Curr Epidemiol Rep*. 2015;2(3):162-171. doi:10.1007/s40471-015-0050-8

21. Edwards JK, Cole SR, Westreich D. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. *Int J Epidemiol*. 2015;44(4):1452-1459. doi:10.1093/ije/dyu272

22. 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

23. Little RJ, D'Agostino R, Cohen ML, et al. The Prevention and Treatment of Missing Data in Clinical Trials. *N Engl J Med*. 2012;367(14):1355-1360. doi:10.1056/NEJMsr1203730

24. Eekhout I, De Boer RM, 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. doi:10.1097/EDE.0b013e3182576cdb

25. 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):118. doi:10.1186/1471-2288-14-118

26. 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

27. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592. doi:10.1093/biomet/63.3.581

28. Westreich D. Berkson's bias, selection bias, and missing data. *Epidemiol Camb Mass*. 2012;23(1):159-164. doi:10.1097/EDE.0b013e31823b6296

29. 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

30. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*. 2014;21(2):221-230. doi:10.1136/amiajnl-2013-001935

31. McClatchey KD, ed. *Clinical Laboratory Medicine*. 2nd ed. Lippincott Wiliams & Wilkins; 2002.

32. Banerjee D, Chung S, Wong EC, Wang EJ, Stafford RS, Palaniappan LP. Underdiagnosis of Hypertension Using Electronic Health Records. *Am J Hypertens*. 2012;25(1):97-102. doi:10.1038/ajh.2011.179

33. Goldstein BA, Bhavsar NA, Phelan M, Pencina MJ. Controlling for Informed Presence Bias Due to the Number of Health Encounters in an Electronic Health Record. *Am J Epidemiol*. 2016;184(11):847-855. doi:10.1093/aje/kww112

34. Phelan M, Bhavsar N, Goldstein BA. Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference. *EGEMs Gener Evid Methods Improve Patient Outcomes*. 2017;5(1):22. doi:10.5334/egems.243

35. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. *AMIA Annu Symp Proc AMIA Symp*. 2013;2013:1472-1477.

36. Groenwold RHH, White IR, Donders ART, Carpenter JR, Altman DG, Moons KGM. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *Can Med Assoc J*. 2012;184(11):1265-1269. doi:10.1503/cmaj.110977

37. Wells BJ, Nowacki AS, Chagin K, Kattan MW. Strategies for Handling Missing Data in Electronic Health Record Derived Data. *EGEMs Gener Evid Methods Improve Patient Outcomes*. 2013;1(3):7. doi:10.13063/2327-9214.1035

38. 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

39. Seaman SR, White IR, Copas AJ, Li L. Combining Multiple Imputation and Inverse-Probability Weighting. *Biometrics*. 2012;68(1):129-137. doi:10.1111/j.1541-0420.2011.01666.x

40. Little RJ, Carpenter JR, Lee KJ. A Comparison of Three Popular Methods for Handling Missing Data: Complete-Case Analysis, Inverse Probability Weighting, and Multiple Imputation. *Sociol Methods Res*. Published online August 5, 2022:004912412211138. doi:10.1177/00491241221113873

41. Enders CK. A Primer on Maximum Likelihood Algorithms Available for Use With Missing Data. *Struct Equ Model Multidiscip J*. 2001;8(1):128-141. doi:10.1207/S15328007SEM0801_7

42. Enders CK. The Performance of the Full Information Maximum Likelihood Estimator in Multiple Regression Models with Missing Data. *Educ Psychol Meas*. 2001;61(5):713-740. doi:10.1177/0013164401615001

43. Lee T, Shi D. A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. *Psychol Methods*. 2021;26(4):466-485. doi:10.1037/met0000381

44. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol Methods*. 2001;6(4):330-351.

45. Beesley LJ, Salvatore M, Fritsche LG, et al. The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. *Stat Med*. 2020;39(6):773-800. doi:10.1002/sim.8445

46. Li J, Yan XS, Chaudhary D, et al. Imputation of missing values for electronic health record laboratory data. *Npj Digit Med*. 2021;4(1):147. doi:10.1038/s41746-021-00518-0

47. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and Managing Missing Structured Data in Electronic Health Records: Data Analysis. *JMIR Med Inform*. 2018;6(1):e11. doi:10.2196/medinform.8960

48. Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. *J Am Stat Assoc*. 1988;83(404):1198-1202. doi:10.1080/01621459.1988.10478722

49. Berrett TB, Samworth RJ. Optimal nonparametric testing of Missing Completely At Random and its connections to compatibility. *Ann Stat*. 2023;51(5). doi:10.1214/23-AOS2326

50. Lambert SA, Gil L, Jupp S, et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. *Nat Genet*. 2021;53(4):420-425. doi:10.1038/s41588-021-00783-5

51. Ma Y, Patil S, Zhou X, Mukherjee B, Fritsche LG. ExPRSweb: An online repository with polygenic risk scores for common health-related exposures. *Am J Hum Genet*. 2022;109(10):1742-1760. doi:10.1016/j.ajhg.2022.09.001

52. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet*. 2014;23(R1):R89-R98. doi:10.1093/hmg/ddu328

53. Burgess S, Butterworth A, Thompson SG. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genet Epidemiol*. 2013;37(7):658-665. doi:10.1002/gepi.21758

54. Lu H, Cole SR, Howe CJ, Westreich D. Toward a Clearer Definition of Selection Bias When Estimating Causal Effects. *Epidemiology*. 2022;33(5):699-706. doi:10.1097/EDE.0000000000001516

55. Zawistowski M, Fritsche LG, Pandit A, et al. The Michigan Genomics Initiative: A biobank linking genotypes and electronic clinical records in Michigan Medicine patients. *Cell Genomics*. Published online January 2023:100257. doi:10.1016/j.xgen.2023.100257

56. All of Us Research Program Investigators, Denny JC, Rutter JL, et al. The “All of Us” Research Program. *N Engl J Med*. 2019;381(7):668-676. doi:10.1056/NEJMsr1809937

57. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol*. 2017;186(9):1026-1034. doi:10.1093/aje/kwx246

58. Van Alten S, Domingue BW, Faul J, Galama T, Marees AT. Reweighting UK Biobank corrects for pervasive selection bias due to volunteering. *Int J Epidemiol*. 2024;53(3):dyae054. doi:10.1093/ije/dyae054

59. Zeng C, Schlueter DJ, Tran TC, et al. Comparison of phenomic profiles in the All of Us Research Program against the US general population and the UK Biobank. *J Am Med Inform Assoc*. 2024;31(4):846-854. doi:10.1093/jamia/ocad260

60. Salvatore M, Kundu R, Shi X, et al. To weight or not to weight? The effect of selection bias in 3 large electronic health record-linked biobanks and recommendations for practice. *J Am Med Inform Assoc*. Published online May 14, 2024:ocae098. doi:10.1093/jamia/ocae098

61. Schoeler T, Speed D, Porcu E, Pirastu N, Pingault JB, Kutalik Z. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat Hum Behav*. Published online April 27, 2023. doi:10.1038/s41562-023-01579-9

62. Lumley T. CRAN - Package survey. 2023. Accessed August 11, 2023. https://cran.r-project.org/web/packages/survey/index.html

63. Buuren SV, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw*. 2011;45(3). doi:10.18637/jss.v045.i03

64. Buuren S van, Groothuis-Oudshoorn K, Vink G, et al. mice: Multivariate Imputation by Chained Equations. Published online November 24, 2021. Accessed July 6, 2022. https://CRAN.R-project.org/package=mice

65. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience; 2004.

66. Marshall A, Altman DG, Holder RL, Royston P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med Res Methodol*. 2009;9(1):57. doi:10.1186/1471-2288-9-57

67. Salvatore M, Clark-Boucher D, Fritsche LG, et al. Cohort profile: Epidemiologic Questionnaire (EPI-Q) – a scalable, app-based health survey linked to electronic health record and genotype data. *Epidemiol Health*. Published online August 8, 2023:e2023074. doi:10.4178/epih.e2023074

68. Howrigan DP, Abbott L, rkwalters, Palmer D, Francioli L, Hammerbacher J. Nealelab/UK_Biobank_GWAS: v2. Published online June 6, 2023. doi:10.5281/zenodo.8011558

69. UK Biobank. Neale lab. Accessed April 24, 2024. http://www.nealelab.is/uk-biobank

70. Carson AP, Muntner P, Selvin E, et al. Do glycemic marker levels vary by race? Differing results from a cross-sectional analysis of individuals with and without diagnosed diabetes. *BMJ Open Diabetes Res Care*. 2016;4(1):e000213. doi:10.1136/bmjdrc-2016-000213

71. Zhu Y, Sidell MA, Arterburn D, et al. Racial/Ethnic Disparities in the Prevalence of Diabetes and Prediabetes by BMI: Patient Outcomes Research To Advance Learning (PORTAL) Multisite Cohort of Adults in the U.S. *Diabetes Care*. 2019;42(12):2211-2219. doi:10.2337/dc19-0532

72. Haneuse S, Arterburn D, Daniels MJ. Assessing Missing Data Assumptions in EHR-Based Studies: A Complex and Underappreciated Task. *JAMA Netw Open*. 2021;4(2):e210184. doi:10.1001/jamanetworkopen.2021.0184

73. Hersh WR, Weiner MG, Embi PJ, et al. Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. *Med Care*. 2013;51(Supplement 8Suppl 3):S30-S37. doi:10.1097/MLR.0b013e31829b1dbd

74. Bourgeois FC. Patients Treated at Multiple Acute Health Care Facilities: Quantifying Information Fragmentation. *Arch Intern Med*. 2010;170(22):1989. doi:10.1001/archinternmed.2010.439

75. Wei WQ, Leibson CL, Ransom JE, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc*. 2012;19(2):219-224. doi:10.1136/amiajnl-2011-000597

76. Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS. Challenges in Using Electronic Health Record Data for CER: Experience of 4 Learning Organizations and Solutions Applied. *Med Care*. 2013;51:S80-S86.

77. Samal L, Dykes PC, Greenberg JO, et al. Care coordination gaps due to lack of interoperability in the United States: a qualitative study and literature review. *BMC Health Serv Res*. 2016;16(1):143. doi:10.1186/s12913-016-1373-y

78. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychol Methods*. 2002;7(2):147-177.

79. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338(jun29 1):b2393-b2393. doi:10.1136/bmj.b2393

80. Lee KJ, Carlin JB, Simpson JA, Moreno-Betancur M. Assumptions and analysis planning in studies with missing data in multiple variables: moving beyond the MCAR/MAR/MNAR classification. *Int J Epidemiol*. 2023;52(4):1268-1275. doi:10.1093/ije/dyad008

81. Mohan K, Pearl J, Tian J. Graphical models for inference with Missing data. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'13. Curran Associates Inc.; 2013:1277-1285.

82. Mohan K, Pearl J. Graphical Models for Processing Missing Data. *J Am Stat Assoc*. 2021;116(534):1023-1037. doi:10.1080/01621459.2021.1874961

83. Thoemmes F, Mohan K. Graphical Representation of Missing Data Problems. *Struct Equ Model Multidiscip J*. 2015;22(4):631-642. doi:10.1080/10705511.2014.937378

84. Galimard J, Chevret S, Protopopescu C, Resche-Rigon M. A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Stat Med*. 2016;35(17):2907-2920. doi:10.1002/sim.6902

85. Galimard JE, Chevret S, Curis E, Resche-Rigon M. Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. *BMC Med Res Methodol*. 2018;18(1):90. doi:10.1186/s12874-018-0547-1

86. Shen C, Weissfeld L. Application of pattern-mixture models to outcomes that are potentially missing not at random using pseudo maximum likelihood estimation. *Biostatistics*. 2005;6(2):333-347. doi:10.1093/biostatistics/kxi013

87. Fiero MH, Hsu C, Bell ML. A pattern-mixture model approach for handling missing continuous outcome data in longitudinal cluster randomized trials. *Stat Med*. 2017;36(26):4094-4105. doi:10.1002/sim.7418

88. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int J Epidemiol*. 2019;48(4):1294-1304. doi:10.1093/ije/dyz032

89. Bartlett JW, Harel O, Carpenter JR. Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression. *Am J Epidemiol*. 2015;182(8):730-736. doi:10.1093/aje/kwv114

90. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 1st ed. Wiley; 2002. doi:10.1002/9781119013563

91. Liang Y, Lu W, Ying Z. Joint Modeling and Analysis of Longitudinal Data with Informative Observation Times. *Biometrics*. 2009;65(2):377-384. doi:10.1111/j.1541-0420.2008.01104.x

92. Lin DY, Ying Z. Semiparametric and Nonparametric Regression Analysis of Longitudinal Data. *J Am Stat Assoc*. 2001;96(453):103-126. doi:10.1198/016214501750333018

93. Sun J, Park DH, Sun L, Zhao X. Semiparametric Regression Analysis of Longitudinal Data With Informative Observation Times. *J Am Stat Assoc*. 2005;100(471):882-889. doi:10.1198/016214505000000060

94. Bůržková P, Lumley T. Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables. *Can J Stat*. 2007;35(4):485-500. doi:10.1002/cjs.5550350402

Tables

Table 1 Comparison of demographic, health measurements, and polygenic risk score values overall and among non-Hispanic Whites 40 or older without diabetes, with and without any missing values in the Michigan Genomics Initiative.

Characteristic	Overall N = 42,999[a]	Incomplete observations N = 17,479[a]	Complete observations N = 25,520[a]	p-value[b]	Non-missing PRS N = 30,942[a]
Age	63.5 (12.5)	62.6 (12.4)	64.1 (12.5)	<0.001	63.7 (12.5)
Female	53.8 (23,145)	54.6 (9,547)	53.3 (13,598)	0.006	53.4 (16,509)
Smoking status (ever)	49.4 (18,187)	46.5 (5,255)	50.7 (12,932)	<0.001	50.1 (14,320)
Missing	*6,178*	*6,178*	*0*		*2348*
BMI	29.1 (6.0)	29.1 (6.0)	29.0 (5.9)	0.3	29.0 (5.9)
Missing	*236*	*236*	*0*		*138*
Glucose	99.5 (14.2)	99.5 (14.7)	99.5 (14.0)	0.023	99.8 (14.3)
Missing	*4,836*	*4,836*	*0*		*3560*
BMI PRS[c]	0.000 (1.000)	0.004 (1.021)	-0.001 (0.996)	0.6	0.000 (1.000)
Missing	*12,057*	*12,057*	*0*		*0*
Glucose PRS[c]	0.000 (1.000)	0.013 (0.989)	-0.003 (1.002)	0.4	0.000 (1.000)
Missing	*12,057*	*12,057*	*0*		*0*

[a] continuous: mean (SD); dichotomous: % (n)
[b] Wilcoxon rank sum test; Pearson's Chi-squared test
[c] PRS were mean standardized
Abbreviations: BMI, body mass index; PRS, polygenic risk score

Table 2 Comparison of demographic, health measurements, and polygenic risk score values overall and among non-Hispanic Blacks 40 or older without diabetes, with and without any missing values in the Michigan Genomics Initiative.

Characteristic	Overall N = 2,297[a]	Incomplete observations N = 1,057[a]	Complete observations N = 1,240[a]	p-value[b]	Non-missing PRS N = 1,437[a]
Age	57.8 (11.4)	57.3 (11.2)	58.2 (11.6)	0.069	57.7 (11.6)
Female	63.3 (1,454)	63.2 (668)	63.4 (786)	>0.9	63.3 (910)
Smoking status (ever)	42.6 (847)	39.2 (293)	44.7 (554)	0.017	44.3 (597)
Missing	*310*	*310*	*0*		*88*
BMI	30.8 (6.4)	30.6 (6.3)	31.0 (6.5)	0.2	31.0 (6.5)
Missing	*26*	*26*	*0*		*8*
Glucose	95.1 (12.5)	94.0 (12.6)	95.8 (12.4)	<0.001	95.8 (12.4)
Missing	*160*	*160*	*0*		*116*
BMI PRS[c]	0.000 (1.000)	-0.006 (0.977)	0.001 (1.004)	0.8	0.000 (1.000)
Missing	*860*	*860*	*0*		*0*
Glucose PRS[c]	0.000 (1.000)	0.158 (1.064)	-0.025 (0.988)	0.061	0.000 (1.000)
Missing	*860*	*860*	*0*		*0*

[a] continuous: mean (SD); dichotomous: % (n)
[b] Wilcoxon rank sum test; Pearson's Chi-squared test
[c] PRS were mean standardized
Abbreviations: BMI, body mass index; PRS, polygenic risk score

Figures

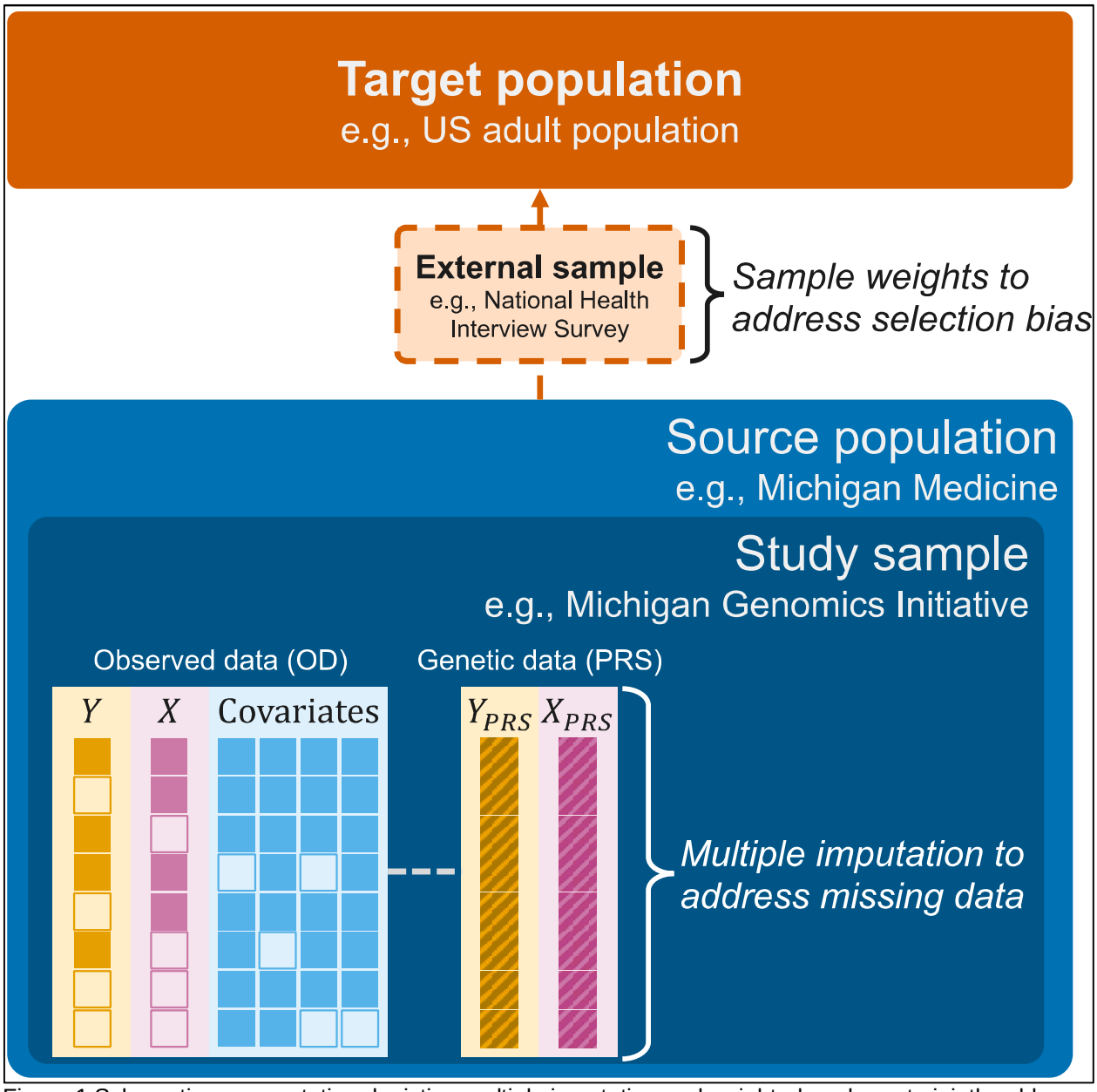

Figure 1 Schematic representation depicting multiple imputation and weighted analyses to jointly address missing data and selection bias. Y represents the outcome (e.g., glucose), X represents the exposure (e.g., body mass index) and covariates could include age, sex, race/ethnicity, and smoking status. The empty boxes represent missing data. Y_{PRS} and X_{PRS} are the polygenic risk scores (PRS) corresponding to the outcome and exposure, respectively.

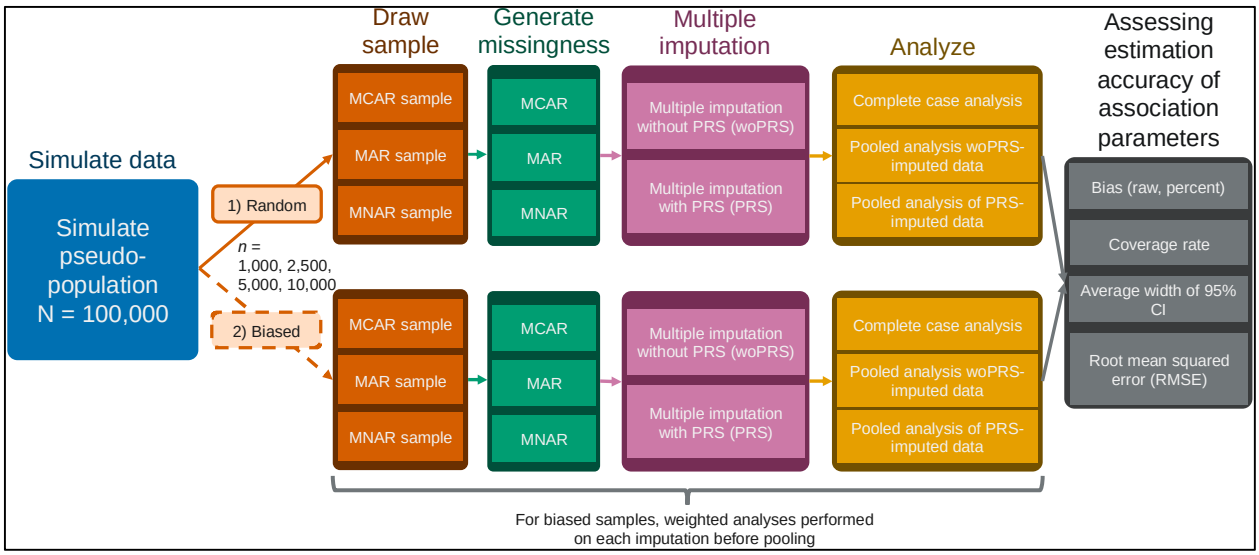

Figure 2 Schematic representation of random and biased sampling simulation analyses. Abbreviations: CI, confidence interval; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; PRS, polygenic risk score; RMSE, root mean square error; woPRS, without polygenic risk score

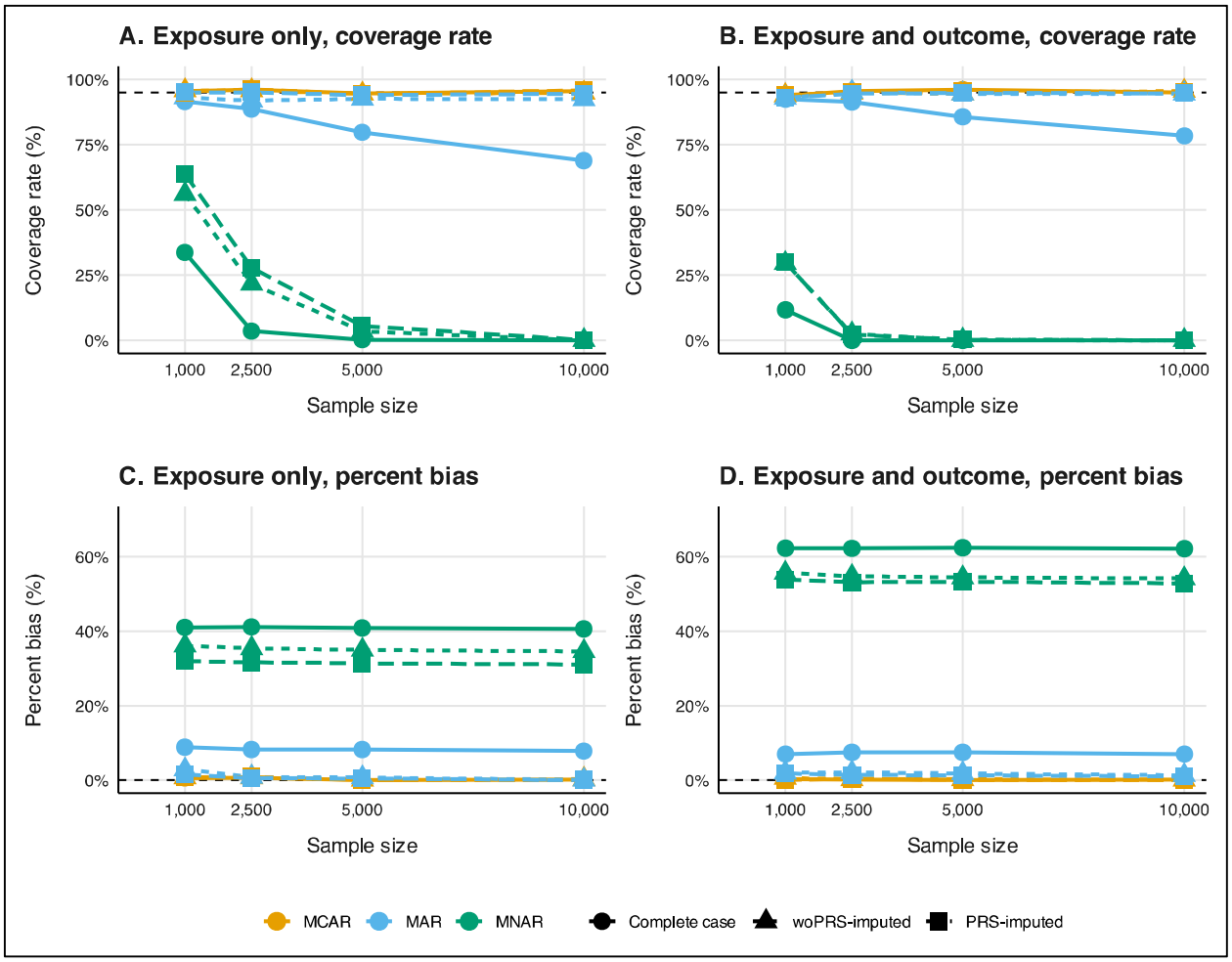

Figure 3 Coverage rate (panels A and B) and percent bias (panels C and D) diagnostics for exposure only (panels A and C) and exposure and outcome missingness (panels B and D) BMI coefficient for glucose by missing data mechanism and method and sample size under random sampling in a 1,000-iteration simulation. Analyses were adjusted for age, sex, non-Hispanic White, and smoking status (ever/never). Corresponding coverage rate, percent bias, average confidence interval width, and root mean squared error diagnostics are reported in Supplementary Table 2 and Supplementary Table 3. Abbreviations: MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; PRS-imputed, polygenic risk score-informed multiple imputation; woPRS-imputed, multiple imputation without exposure and outcome PRS.

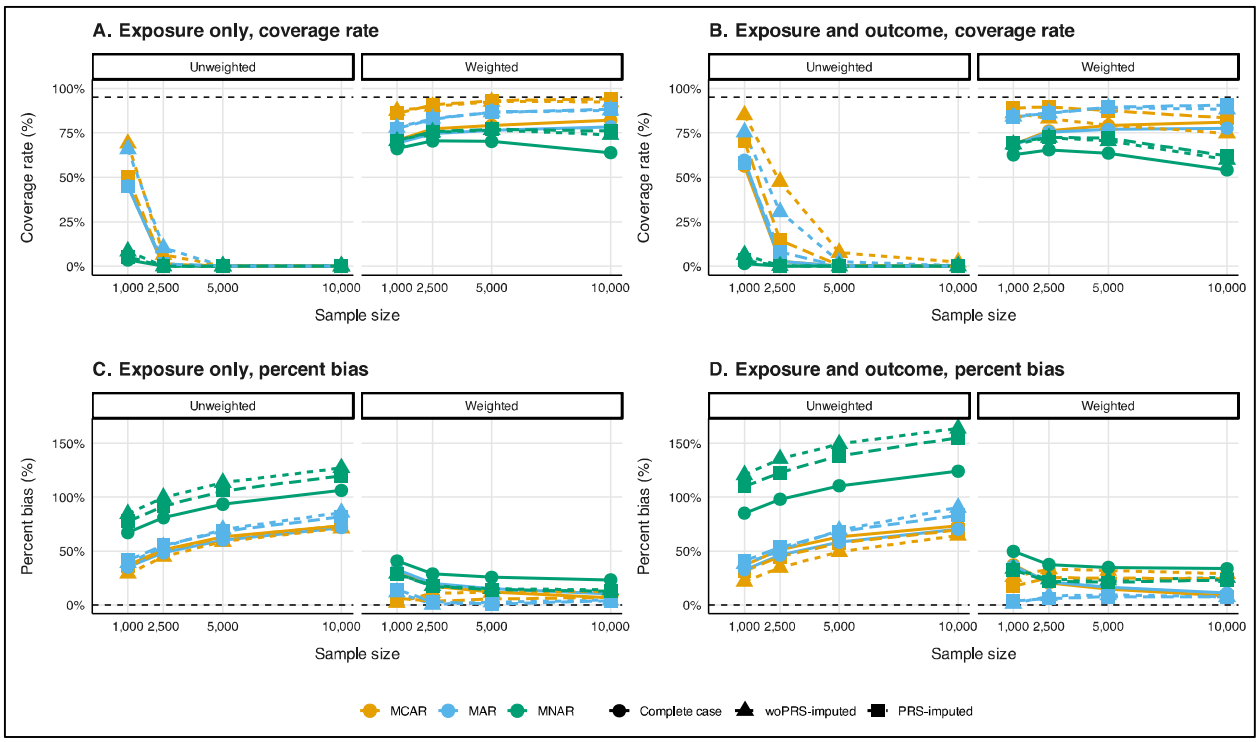

Figure 4 Coverage rate (panels A and B) and percent bias (panels C and D) diagnostics for unweighted (left) and weighted (right) BMI coefficient for glucose estimation by missing data mechanism and method and sample size under biased sampling and exposure only (panels A and C) and exposure and outcome missingness (panels B and D) in a 1,000-iteration simulation. For biased sampling simulations, unweighted and weighted diagnostics are reported in Supplementary Tables 4, 5, 6, and 7, respectively. Analyses were adjusted for age, sex, non-Hispanic White, and smoking status (ever/never).

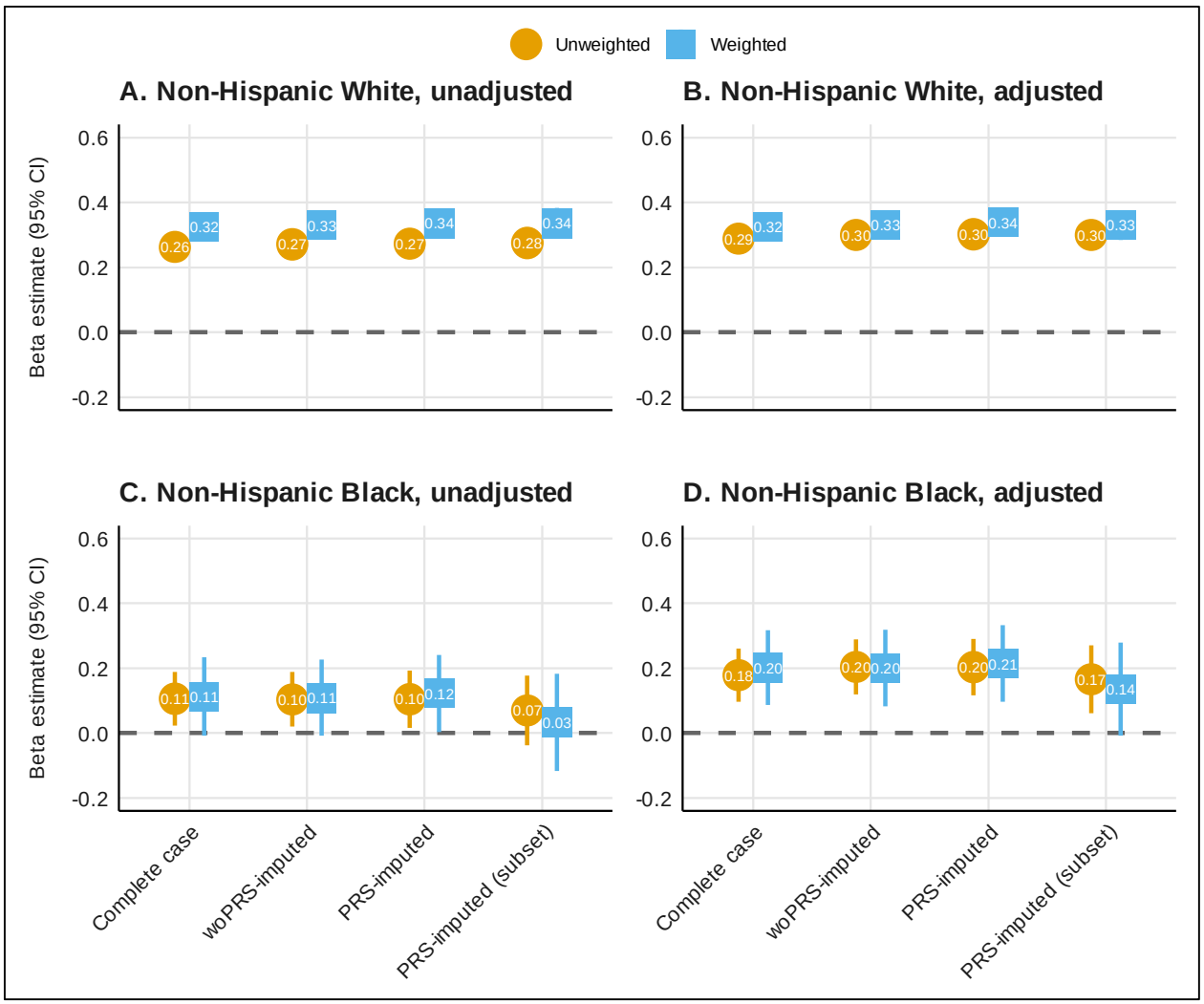

Figure 5 Estimation of the coefficient for BMI with glucose as the outcome by missing data method and weighting approach among non-Hispanic Whites (n=42,999; panels A and B) and non-Hispanic Blacks (n=2,297; panels C and D) in all MGI adults age 40 or older without diabetes. The PRS-imputed subset sample (n=30,942 for non-Hispanic Whites; n=1,437 for non-Hispanic Blacks) was restricted to individuals with non-missing genotype data before multiple imputation. Analyses were adjusted for age, sex, and smoking status (ever/never). Gray shaded regions represent corresponding 95% confidence interval from National Health Interview Survey-weighted All of Us data where weights are calculated separately for non-Hispanic Whites and non-Hispanic Blacks to make All of Us data for each of these groups more representative of their corresponding US population (target population). Results for the full, unstratified cohort are shown in Supplementary Figure 5. Abbreviations: PRS, polygenic risk score.