Journal of Data Science 0 (0), 1–21
August 2025
DOI: 10.6339/25-JDS1179
Statistical Data Science

Matched Mass Imputation for Survey Data Integration

Jeremy Flood and Sayed A. Mostafa*

Department of Mathematics & Statistics, North Carolina A&T State University, Greensboro, NC, USA

Abstract

Analysis of nonprobability survey samples has gained much attention in recent years due to their wide availability and the declining response rates within their costly probabilistic counterparts. Still, valid population inference cannot be deduced from nonprobability samples without additional information, which typically takes the form of a smaller survey sample with a shared set of covariates. In this paper, we propose the matched mass imputation (MMI) approach as a means for integrating data from probability and nonprobability samples when common covariates are present in both samples but the variable of interest is available only in the nonprobability sample. The proposed approach borrows strength from the ideas of statistical matching and mass imputation to provide robustness against potential nonignorable bias in the nonprobability sample. Specifically, MMI is a two-step approach: first, a novel application of statistical matching identifies a subset of the nonprobability sample that closely resembles the probability sample; second, mass imputation is performed using these matched units. Our empirical results, from simulations and a real data application, demonstrate the effectiveness of the MMI estimator under nearest-neighbor matching, which almost always outperformed other imputation estimators in the presence of nonignorable bias. We also explore the effectiveness of a bootstrap variance estimation procedure for the proposed MMI estimator.

Keywords *data integration; mass imputation; nonignorable missingness; nonprobability samples; statistical matching*

1 Introduction

For the past century, probability samples have been considered the gold standard by statisticians as they facilitate statistical inference about the target population. These samples are drawn using a probability sampling design, $p(.)$, which ensures that every unit in the population has a positive selection probability (Lohr, 2021; Särndal et al., 2003). They are not without challenges, however; recent literature explicitly highlights the impracticality of probability samples due to factors such as design costs, non-response rates, and other limitations (Yang and Kim, 2020; Wiśniowski et al., 2020; Wang et al., 2020; Kern et al., 2021; Li et al., 2023). The recent surge in high-dimensional convenience datasets has popularized the use of nonprobability samples to maximize available statistical information; nevertheless, without an explicit sampling design, results derived from these samples are likely to suffer from estimation bias (Yang et al., 2021; National Academies of Sciences, Engineering, and Medicine, 2018). An intuitive compromise involves *data integration*, an umbrella term for techniques that combine non-probability and probability data to leverage the strengths of both. Such techniques usually require a distinct

*Corresponding author. Email: sabdelmegeed@ncat.edu.

Received August 16, 2024; Accepted March 20, 2025

pattern of *monotone missingness*, where a set of shared covariates, $\boldsymbol{X}$, exist in both the probability and nonprobability sample, but the target variable, Y, is available only in the latter. Specifically, with $\mathcal{U}$ denoting the finite population of interest, let $\boldsymbol{A} \subseteq \mathcal{U}$ denote a probability sample of size n_{A} that contains sample inclusion probabilities, π_i, and covariates, $\boldsymbol{X}_i$, for all elements $i \in \boldsymbol{A}$. Furthermore, let $\boldsymbol{B} \subseteq \mathcal{U}$ denote a nonprobability sample of size n_{B} that is assumed to contain information on the response variable, Y_j, and covariates, $\boldsymbol{X}_j$, for all $j \in \boldsymbol{B}$. In these contexts, blending sample $\boldsymbol{A}$ with sample $\boldsymbol{B}$ is of expressed interest, but remains challenging due to varying information in each sample.

This data integration problem has received considerable attention recently, with most research focusing on estimating the finite population mean, $\mu_N = \frac{1}{N}\sum_{i\in\mathcal{U}} Y_i$, using $\boldsymbol{A}$ and $\boldsymbol{B}$ together. Note that, if Y_i was measured for all elements in $\boldsymbol{A}$, a design-unbiased estimate of μ could be obtained by use of Horvitz and Thompson (1952)'s (HT) mean estimator, defined as $\hat{\mu}_\pi = \frac{1}{N}\sum_{i\in\boldsymbol{A}} \pi_i^{-1} Y_i$. Of course, $\hat{\mu}_\pi$ cannot be calculated since the values of Y are not observed in sample $\boldsymbol{A}$, but it may be possible to fill in these missing Y values using predictions from a regression model trained on sample $\boldsymbol{B}$. This approach is suitably named *mass imputation* since it 'imputes' mass values of Y in sample $\boldsymbol{A}$ (Kim et al., 2021; Chen et al., 2022). In a similar vein, one may consider building a model on $\boldsymbol{C} = \boldsymbol{A} \cup \boldsymbol{B}$ to obtain predicted values of π_j for all $j \in \boldsymbol{B}$, and using these values to calculate a pseudo-HT mean estimator; these procedures fall under *propensity score estimation*, which estimates an individual's 'propensity' to belong in $\boldsymbol{B}$ (Beaumont and Rao, 2021; Chen et al., 2020). Furthermore, *doubly-robust* mean estimators combine estimated propensities and outcome predictions to ensure consistency, provided that either the propensity model or the outcome model is correctly specified (Yang et al., 2020).

A common assumption in the above work is the *ignorability* condition, which assumes that the response variable Y_j is independent of the sample membership indicators δ_j^B given $\boldsymbol{X}_j$ for all $j \in \boldsymbol{B}$. Similar to the *missing at random* (MAR) assumption in missing data problems, ignorability is untestable since verifying it requires access to unavailable data (Little, 1988). To address potential bias from nonignorable sampling in $\boldsymbol{B}$, *statistical matching* pairs nonprobability sample observations with their probabilistic counterparts based on a chosen distance measure (Dever, 2018; Kalay, 2021). While Rivers (2007) and Bethlehem (2016) showed that matching can reduce estimation bias, the effectiveness of their approach is limited to MAR scenarios. Moreover, Bethlehem (2016) focused only on categorical $\boldsymbol{X}$, restricting its broader applicability.

To address this gap, we propose the *matched mass imputation* (MMI) procedure, which combines statistical matching with mass imputation. The approach consists of two steps: (1) identifying the subset of $j \in \boldsymbol{B}$ most similar to elements in $\boldsymbol{A}$ and (2) using an outcome model trained on these matched units to impute the missing Y values in $\boldsymbol{A}$. By building the outcome model only on the matched subset, MMI reduces dependence on the full nonprobability sample and offers greater robustness to nonignorable selection mechanisms.

The remainder of the paper is organized as follows. Section 2 introduces the notation, data integration setup, and key definitions. Section 3 provides an overview of existing data integration procedures, highlighting their strengths and weaknesses. In Section 4, we introduce the proposed MMI procedure. Section 5 presents results from Monte Carlo simulations, while Section 6 summarizes findings from a real data application—both comparing the performance of our proposed estimators against various competitors. We conclude with a discussion of results in Section 7.

Table 1: Data structure for the probability ($\boldsymbol{A}$) and nonprobability ($\boldsymbol{B}$) samples.

Sample	d	X_1	X_2	$\cdots$	X_p	Y
$\boldsymbol{A}$	✓	✓	✓	✓	✓	×
$\boldsymbol{B}$	×	✓	✓	✓	✓	✓

2 Notation and Preliminaries

Let $\mathcal{U} = \{1, 2, \ldots, N\}$ denote an index set for the units in a finite population of size N. Also, let $\boldsymbol{A}$ denote a $n_{\text{A}} \times (p+1)$ probability sample drawn from $\mathcal{U}$ with measured variables $S_{\text{A}} = \{d, X_1, X_2, \ldots, X_p\}$, where d denotes the *design weights*, $d_i = 1/\pi_i$ for $i \in \{1, \ldots, n_{\text{A}}\}$, and $X_1, \ldots, X_p$ the auxiliary variables (or, *covariates*). Similarly, let $\boldsymbol{B}$ denote a $n_{\text{B}} \times (p+1)$ nonprobability sample from $\mathcal{U}$ with measured variables $S_{\text{B}} = \{Y, X_1, X_2, \ldots, X_p\}$, where Y denotes a response variable of interest. This data structure is illustrated in Table 1.

Given that Y and d are missing in $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, we may estimate μ by using a survey-weighted mean of imputed $\hat{Y}$ (for *mass imputation*), estimated $\hat{\pi}$ (for *propensity score estimation*), or some amalgamation of the two (for *doubly-robust estimation* and *statistical matching*). These methods require *positivity* and *transportability* assumptions, stated in Definitions 1 and 2, respectively.

Definition 1 (Positivity Assumption)**.** Let δ_j^B denote the sample $\boldsymbol{B}$ membership indicator which takes the value 1 if $j \in \boldsymbol{B}$. The assumption of positivity is satisfied if $\Pr(\delta_j^B = 1 | \boldsymbol{X} = \boldsymbol{x}) > 0$ for all $j \in \boldsymbol{B}$ and $\boldsymbol{x}$ in the support of $\boldsymbol{X}$. That is, for every value $\boldsymbol{x}$, there is a positive probability for the corresponding units to be selected in sample $\boldsymbol{B}$.

Definition 2 (Transportability Assumption)**.** Let $f(Y|\boldsymbol{X})$ denote the distribution of Y conditional on $\boldsymbol{X}$. The assumption of transportability is satisfied if $f(Y|\boldsymbol{X}, \delta_j^B = 1) = f(Y|\boldsymbol{X})$.

Transportability is a crucial assumption as it allows us to 'transport' a prediction model built on sample $\boldsymbol{B}$ to sample $\boldsymbol{A}$. As described by Kim et al. (2021), a sufficient condition for transportability is the *ignorability condition* in Definition 3, which is essentially Rubin (1976)'s *missing at random* (MAR) assumption.

Definition 3 (Ignorability Condition)**.** Let $\Pr(\delta_j^B = 1|\boldsymbol{X})$ denote the inclusion probability of unit $j \in \boldsymbol{B}$ conditional on $\boldsymbol{X}$. The assumption of ignorability is satisfied if $\Pr(\delta_j^B = 1|\boldsymbol{X}, Y) = \Pr(\delta_j^B = 1|\boldsymbol{X})$ for all $j \in \boldsymbol{B}$. Otherwise, the sample is said to possess a nonignorable, or *informative*, sampling mechanism.

Under the above setup, the following section provides an overview of some existing data integration methods, namely, mass imputation, propensity score estimation, and statistical matching.

3 Overview of Existing Methods

3.1 Mass Imputation

The mass imputation literature generally falls into two categories based on the assumed prediction model: (1) parametric models, which assume a known conditional mean function, $\mathbb{E}(Y|\boldsymbol{X}) =$

$m(\boldsymbol{X};\boldsymbol{\beta})$, parameterized by an unknown coefficient vector $\boldsymbol{\beta}$, and (2) nonparametric models, which directly estimate a nonparametric conditional mean function, $m(\boldsymbol{X})$. These two approaches are described below.

3.1.1 Parametric Mass Imputation

In the parametric context, the finite population is considered as a realization from a superpopulation model, ξ_{P}, given by

$$Y = m(\boldsymbol{X};\boldsymbol{\beta}) + \nu(\boldsymbol{X})\epsilon, \tag{1}$$

where $m(\boldsymbol{X};\boldsymbol{\beta}) = \mathbb{E}_{\xi_{\mathrm{P}}}(Y|\boldsymbol{X})$ is a *known* function of $\boldsymbol{X}$ parameterized by an unknown parameter vector $\boldsymbol{\beta}$, $\nu(\cdot)$ is a known, strictly positive variance function, and ϵ is an error term satisfying $\mathbb{E}_{\xi_{\mathrm{P}}}(\epsilon|\boldsymbol{X}) = 0$ and $\mathbb{E}_{\xi_{\mathrm{P}}}\left(\epsilon^2|\boldsymbol{X}\right) = \sigma_\epsilon^2$. If $\mathcal{U}$ was fully observed, a natural estimator of $\boldsymbol{\beta}$ could be chosen to solve the finite population score function

$$U(\boldsymbol{\beta}) = \frac{1}{N}\sum_{u\in\mathcal{U}}\Big(Y_u - m(\boldsymbol{X}_u;\boldsymbol{\beta})\Big)\boldsymbol{W}(\boldsymbol{X}_u;\boldsymbol{\beta}) = 0$$

for some p-dimensional function $\boldsymbol{W}$; this estimator is henceforth denoted as $\boldsymbol{\beta}_{\mathrm{N}}$ since it requires information for all N population units. Since Y is measured in sample $\boldsymbol{B}$ only, $\boldsymbol{\beta}$ can be estimated by solving a sample-based score function,

$$\widehat{U}(\boldsymbol{\beta}) = \frac{1}{n_{\mathrm{B}}}\sum_{j\in\boldsymbol{B}}\Big(Y_j - m(\boldsymbol{X}_j;\boldsymbol{\beta})\Big)\boldsymbol{W}\left(\boldsymbol{X}_j;\boldsymbol{\beta}\right) = 0,$$

whose solution is henceforth denoted as $\widehat{\boldsymbol{\beta}}$.

Under this framework, Kim et al. (2021) proposed a *parametric mass imputation estimator*, henceforth abbreviated as PMIE, that uses predictions from a semiparametric regression model to estimate the finite population mean μ_N. Specifically, let $\hat{Y}_{i,\mathrm{par}} = m(\boldsymbol{X}_i;\widehat{\boldsymbol{\beta}})$ denote the predicted value of $Y_i, i\in\boldsymbol{A}$. Then, the PMIE is defined as

$$\hat{\mu}_{\mathrm{PMIE}} = \frac{1}{N}\sum_{i\in\boldsymbol{A}}\pi_i^{-1}\hat{Y}_{i,\mathrm{par}},$$

where π_i^{-1} denotes the first-order sampling weight associated with unit i. Under some regulatory conditions, Kim et al. (2021) showed that $\hat{\mu}_{\mathrm{PMIE}} = \tilde{\mu}_{\mathrm{PMIE}} + o_p(n_B^{-1/2})$, where

$$\tilde{\mu}_{\mathrm{PMIE}} = \frac{1}{N}\sum_{i\in\boldsymbol{A}}\pi_i^{-1}Y_i^* + \frac{1}{n_{\mathrm{B}}}\sum_{j\in\boldsymbol{B}}\left(Y_j - Y_j^*\right)\boldsymbol{W}(\boldsymbol{X}_j;\boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{c},$$

with $Y_i^* = m(\boldsymbol{X}_i;\boldsymbol{\beta}^*)$, $\boldsymbol{\beta}^* = \operatorname{plim}\widehat{\boldsymbol{\beta}}$, and

$$\boldsymbol{c} = \left[n_{\mathrm{B}}^{-1}\sum_{j\in\boldsymbol{B}}\frac{\partial m(\boldsymbol{X}_j;\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}^*}\boldsymbol{W}(\boldsymbol{X}_j;\boldsymbol{\beta}^*)^{\mathrm{T}}\right]^{-1}\frac{1}{N}\sum_{u\in\mathcal{U}}\frac{\partial m(\boldsymbol{X}_u;\boldsymbol{\beta}^*)}{\partial\boldsymbol{\beta}^*}.$$

From here, they show that

$$\mathbb{E}\left(\tilde{\mu}_{\mathrm{PMIE}} - \mu_N\right) = -\mathbb{E}\left(N^{-1}\sum_{u\in\mathcal{U}}\left(Y_u - m(\boldsymbol{X}_u;\boldsymbol{\beta}^*)\right)\right),$$

which implies that $\hat{\mu}_{\text{PMIE}}$ is asymptotically unbiased if the model is correct and ignorability holds (i.e., $\boldsymbol{\beta}^* = \boldsymbol{\beta}$). If ignorability is not satisfied, the asymptotic bias of the PMIE will still be smaller than that of the naive estimator (mean of sample $\boldsymbol{B}$) if $\text{Var}\left(Y_u - m(\boldsymbol{X}_u; \hat{\boldsymbol{\beta}})\right) < \text{Var}(Y_u)$ for $u \in \mathcal{U}$, which is likely true for reasonably specified models.

3.1.2 Nonparametric Mass Imputation

Nonparametric regression techniques can be used to provide robustness against misspecification of the functional form of the conditional mean function, $m(\boldsymbol{X})$. Assume that $\mathcal{U}$ is a realization from the following nonparametric superpopulation model, ξ_{NP}:

$$Y = m(\boldsymbol{X}) + \nu(\boldsymbol{X})\epsilon. \tag{2}$$

To obtain mass imputation predictions for the Y values in sample $\boldsymbol{A}$, the conditional mean function $m(\boldsymbol{X})$ can be estimated by training a model on sample $\boldsymbol{B}$ using nonparametric regression techniques such as k-nearest neighbors (Rivers, 2007), kernel regression (Chen et al., 2022), or generalized additive models (GAMs, Chen et al., 2022). However, when the covariates $\boldsymbol{X}$ are high-dimensional, k-nearest neighbors and kernel regression may suffer from the curse of dimensionality. Therefore, we focus on mass imputation with GAM as described by Chen et al. (2022).

Assume that, conditional on $\boldsymbol{X}$, Y follows an exponential family distribution with

$$g^{-1}\left(\mathbb{E}_{\xi_{\text{NP}}}(Y|\boldsymbol{X})\right) \coloneqq g^{-1}\left(m(\boldsymbol{X})\right) = \sum_{k=1}^{p} \rho_k(X_k), \tag{3}$$

where $g^{-1}(\cdot)$ denotes an inverse link function and $\{\rho_k\}_{k=1}^{p}$ a sequence of unknown smooth functions. Under this framework, Chen et al. (2022) considered using regression splines to estimate ρ_k, which approximate $g^{-1}\left(\mathbb{E}_{\xi_{\text{NP}}}(Y|\boldsymbol{X})\right)$ with linear combinations of $(m+1)$-order basis functions, $B_m(\cdot)$ for $m = \{1, 2, \ldots, M\}$, and spline coefficients, $\boldsymbol{\gamma}_m^k$, such that

$$g^{-1}\left(\mathbb{E}_{\xi_{\text{NP}}}(Y|\boldsymbol{X}_i)\right) \approx \sum_{k=1}^{p}\sum_{m=1}^{M} \boldsymbol{\gamma}_m^k B_m\left(X_{k,i}\right)$$

(Wood, 2017; James et al., 2013). To offset the risk of overfitting, the coefficient vector $\boldsymbol{\gamma}$ was chosen to minimize the penalized likelihood $-2\ln L(\boldsymbol{\gamma}) + \sum_{k=1}^{p} \lambda_k \boldsymbol{\gamma}_k^{\mathrm{T}} \boldsymbol{S}_k \boldsymbol{\gamma}_k$, where λ_k denotes the smoothing hyperparameter of covariate X_k, $L(\boldsymbol{\gamma}) = \prod_{j \in \boldsymbol{B}} f\left(Y_j|\boldsymbol{X}_j; \boldsymbol{\gamma}\right)$ is the likelihood function, $\boldsymbol{\gamma}_k^{\mathrm{T}} = \begin{bmatrix}\boldsymbol{\gamma}_1^k & \boldsymbol{\gamma}_2^k & \cdots & \boldsymbol{\gamma}_M^k\end{bmatrix}$, and

$$\boldsymbol{S}_k = \begin{bmatrix} \int B_1''(X_k)B_1'(X_k)\,dX_k & \int B_1''(X_k)B_2'(X_k)\,dX_k & \cdots & \int B_1''(X_k)B_M'(X_k)\,dX_k \\ \vdots & \vdots & \vdots & \vdots \\ \int B_M''(X_k)B_1'(X_k)\,dX_k & \cdots & \cdots & \int B_M''(X_k)B_M'(X_k)\,dX_k \end{bmatrix}.$$

Then, Chen et al. (2022)'s nonparametric mass imputation estimator is defined as follows:

$$\begin{aligned}\hat{\mu}_{\text{MIgam}} &\coloneqq \frac{1}{N}\sum_{i \in \boldsymbol{A}} \pi_i^{-1} g\left(\sum_{k=1}^{p}\sum_{m=1}^{M} \hat{\boldsymbol{\gamma}}_m^k B_m\left(X_{k,i}\right)\right) \\ &= \frac{1}{N}\sum_{i \in \boldsymbol{A}} \pi_i^{-1} \hat{Y}_{i,\text{gam}}.\end{aligned} \tag{4}$$

3.2 Propensity Score Estimation

An alternative to mass imputation is the propensity score approach, which estimates the missing sample inclusion probabilities π_j^B for units in sample $\boldsymbol{B}$ using a propensity score model built from combined data from $\boldsymbol{A}$ and $\boldsymbol{B}$. The estimated π_j^B values are then used to apply the Horvitz-Thompson estimator to the Y values in sample $\boldsymbol{B}$. Similar to Chen et al. (2020), suppose that π_j^B can be modeled as $\pi_j^B = \Pr\left(j \in \boldsymbol{B} | \boldsymbol{X}_j\right) = \pi(\boldsymbol{X}_j; \boldsymbol{\zeta})$, where $\boldsymbol{\zeta}$ denotes a set of unknown nuisance parameters. We may choose $\hat{\boldsymbol{\zeta}}$ to maximize $\ell^*(\boldsymbol{\zeta})$, a pseudo log-likelihood function given by

$$\ell^*(\boldsymbol{\zeta}) = \sum_{j \in \boldsymbol{B}} \log\left\{\frac{\pi(\boldsymbol{X}_j; \boldsymbol{\zeta})}{1 - \pi(\boldsymbol{X}_j; \boldsymbol{\zeta})}\right\} + \sum_{i \in \boldsymbol{A}} \frac{1}{\pi_i^A} \log\{1 - \pi(\boldsymbol{X}_i; \boldsymbol{\zeta})\}. \tag{5}$$

The second term on the right-hand side of (5) incorporates information from sample $\boldsymbol{A}$ to estimate the unknown population sum appearing in the corresponding term of the following log-likelihood function for sample $\boldsymbol{B}$:

$$\begin{aligned}\ell(\boldsymbol{\zeta}) &= \sum_{u \in U} \left\{\delta_u^B \log \pi_u^B + (1 - \delta_u^B) \log(1 - \pi_u^B)\right\} \\ &= \sum_{j \in \boldsymbol{B}} \log\left\{\frac{\pi(\boldsymbol{X}_j; \boldsymbol{\zeta})}{1 - \pi(\boldsymbol{X}_j; \boldsymbol{\zeta})}\right\} + \sum_{u \in U} \log\{1 - \pi(\boldsymbol{X}_u; \boldsymbol{\zeta})\}.\end{aligned}$$

Then, the maximum likelihood estimator (MLE) of π_j is defined as $\hat{\pi}_j = \pi(\boldsymbol{X}_j; \hat{\boldsymbol{\zeta}})$. The propensity score model, at its best, produces $\hat{\pi}_j \approx \pi_j$ for all $j \in \boldsymbol{B}$; however, if the model is misspecified, $\hat{\pi}_j$ will be substantially biased, which makes this approach unfavorable in practice (Beaumont and Rao, 2021; Wang et al., 2020; Yang et al., 2020; Lee et al., 2011). As a compromise, Chen et al. (2020) proposed a doubly-robust approach that integrates propensity scores, $\hat{\pi}$, and mass imputations, $\hat{Y}$, in one estimator:

$$\hat{\mu}_{\mathrm{DR}} = \frac{1}{N}\left(\sum_{j \in \boldsymbol{B}} \frac{Y_j - \hat{Y}_j}{\hat{\pi}_j} + \sum_{i \in \boldsymbol{A}} \frac{\hat{Y}_i}{\pi_i}\right). \tag{6}$$

The attractiveness of the DR estimator stems from its flexibility, in that it is consistent for μ_N if either the propensity score model or the outcome model is correctly specified. We refer the reader to Chen et al. (2020) for the details regarding the asymptotic properties of $\hat{\mu}_{\mathrm{DR}}$.

3.3 Statistical Matching

We conclude this section with a discussion on statistical matching, a procedure that pairs elements in sample $\boldsymbol{A}$ with closely related counterparts in sample $\boldsymbol{B}$. The goal is to create a subset of $\boldsymbol{B}$, denoted as $\boldsymbol{M}$, that is less biased than $\boldsymbol{B}$ itself.

For continuous $\boldsymbol{X}$, Rivers (2007) facilitated cold-deck matching by pairing each unit $i \in \boldsymbol{A}$ with a donor $m \in \boldsymbol{B}$ that satisfies $|\boldsymbol{X}_m - \boldsymbol{X}_i| \leqslant \left|\boldsymbol{X}_j - \boldsymbol{X}_i\right|$ across all $j \in \boldsymbol{B}$. Letting $\tilde{Y}_{i,\mathrm{R}} \coloneqq Y_m$ denote unit m's corresponding Y, Rivers (2007) proposed the following mean estimator:

$$\hat{\mu}_{\mathrm{SM:R}} = \frac{1}{N} \sum_{i \in \boldsymbol{A}} \pi_i^{-1} \tilde{Y}_{i,\mathrm{R}}. \tag{7}$$

Rivers (2007) reported that, under some regularity conditions (including ignorability), $\hat{\mu}_{\text{SM:R}}$ is asymptotically unbiased for μ_N.

Bethlehem (2016) extended this work to categorical $\boldsymbol{X}$ by considering (random) cold-deck matching, which facilitates imputation by randomly selecting units in $\boldsymbol{B}$ with exact profiles in $\boldsymbol{A}$. To describe their procedure, suppose that there exists a set of categorical covariates $\boldsymbol{X}$ shared between $\boldsymbol{A}$ and $\boldsymbol{B}$ that, when crossed, produce H disjoint groups (or, *strata*). Let n_h and m_h denote the number of observations in $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, that lie in stratum h. Assuming that $\boldsymbol{B}$ is sufficiently large (i.e., $n_h \leqslant m_h$ for all $h \in H$) and the design of $\boldsymbol{A}$, $p(\boldsymbol{A})$, is a simple random sample without replacement (SRS), Bethlehem (2016) draws n_h individuals from m_h as a SRS, and uses their naive mean to estimate μ. This yields the following cold-deck estimator:

$$\hat{\mu}_{\text{SM:B}} \coloneqq \frac{1}{n_{\text{A}}} \sum_{i \in \boldsymbol{A}} \tilde{Y}_{i,\text{B}} = \frac{1}{n_{\text{A}}} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \tilde{Y}_{i_h}. \tag{8}$$

Similar to Rivers (2007)'s $\hat{\mu}_{\text{SM:R}}$, Bethlehem (2016)'s $\hat{\mu}_{\text{SM:B}}$ is also asymptotically unbiased for μ_N if $p(\boldsymbol{B})$ is ignorable. Their simulation results supported this claim, as the bias of $\hat{\mu}_{\text{SM:B}}$ was indeed similar to $\hat{\mu}_{\pi}$ when the missingness in sample $\boldsymbol{B}$ was missing at random; for nonignorable missingness, though, the bias of $\hat{\mu}_{\text{SM:B}}$ was quite noticeable and tended to match that of the naive mean of $\boldsymbol{B}$.

Taken together, the works of Rivers (2007) and Bethlehem (2016) demonstrate the impact of cold-deck matching on reducing estimation bias. However, neither approach proves effective when $p(\boldsymbol{B})$ is nonignorable. In the next section, we propose the *matched mass imputation* procedure as a compromise between cold-deck matching and mass imputation, offering robustness to nonignorable selection in the design of the nonprobability sample.

4 Matched Mass Imputation (MMI)

The aforementioned literature assumes *transportability* (Definition 2) or *ignorability* (Definition 3) which, given the lack of design information, are both incompatible with convenience samples. These assumptions are nonetheless crucial components of any imputation engine, and cannot be omitted; thus, it is better to replace some haphazard $\boldsymbol{B}$ with $\boldsymbol{M} \subset \boldsymbol{B}$, a set of statistical matches designed to be as close to $\boldsymbol{A}$ as possible. This makes transportability (or, sufficiently, ignorability) more plausible, as it only needs to hold within a less-biased $\boldsymbol{M}$. For this reason, we consider replacing $\boldsymbol{B}$ in the mass imputation framework with $\boldsymbol{M}$, a subset of $\boldsymbol{B}$ that possesses maximal similarity to $\boldsymbol{A}$. Specifically, our approach uses statistical matching (detailed below) to identify the subset $\boldsymbol{M}$. Then, it fits a regression model on $\boldsymbol{M}$ to obtain $\hat{Y}^{M}_{A,i}$ for all $i \in \boldsymbol{A}$ and uses the survey-weighted mean of the resulting predictions as an estimator for μ_N. The resulting matched mass imputation (MMI) estimator, $\hat{\mu}_{\text{MMI}}$, is defined as

$$\begin{aligned} \hat{\mu}_{\text{MMI}} &= \left(\sum_{i \in \boldsymbol{A}} \frac{1}{\pi_i} \right)^{-1} \sum_{i \in \boldsymbol{A}} \frac{1}{\pi_i} \hat{Y}^{M}_{A,i} \\ &= \hat{N}^{-1} \Pi^{\top} \hat{Y}^{M}_{A}, \end{aligned} \tag{9}$$

where $\Pi^{\top} \coloneqq \begin{bmatrix} \frac{1}{\pi_1} & \frac{1}{\pi_2} & \cdots & \frac{1}{\pi_1} \end{bmatrix}$ is a $(1 \times n_{\text{A}})$ row vector of design weights and $\hat{Y}^{M}_{\text{A}}$ is a $(n_{\text{A}} \times 1)$ vector of predictions. The use of $\hat{N} \coloneqq \sum_{i \in \boldsymbol{A}} {\pi_i}^{-1}$ to estimate N makes the mean estimator a

Algorithm 1 Matched mass imputation.

1. Use some statistical matching model to identify and isolate observations in $\boldsymbol{B}$ with the highest similarity to those in $\boldsymbol{A}$. These units now belong to $\boldsymbol{M}$, a set of statistical matches.
2. Build a regression model on $\boldsymbol{M}$, say $\hat{m}_{\mathrm{M}}(\boldsymbol{X})$, and use it with the covariates $\boldsymbol{X}_A$ to obtain predictions for the missing Y values in sample $\boldsymbol{A}$, $\hat{Y}_A^M$.
3. Calculate $\hat{\mu}_{\mathrm{MMI}} = \hat{N}^{-1}\Pi^\top \hat{Y}_A^M$.

Hájek-type estimator, which is known to be particularly efficient for unequal probability sampling designs (Hájek, 1964). The general MMI procedure is summarized in Algorithm 1, with the remainder of this section detailing the identification of the set of matches $\boldsymbol{M}$.

4.1 Statistical Matching Models

In this subsection, we introduce three popular statistical matching models that can be used for the proposed MMI estimator: exact matching, coarsened exact matching, and nearest-neighbor (NN) matching.

4.1.1 Exact Matching

The exact matching model is straightforward, as it simply requires finding the intersection of samples $\boldsymbol{A}$ and $\boldsymbol{B}$. The simplicity of this approach allows us to establish several useful remarks.

Let $\boldsymbol{M} \coloneqq \boldsymbol{A} \cap \boldsymbol{B}$, and let $\Pr(j \in \boldsymbol{A} \cap \boldsymbol{B}) = \Pr(j \in \boldsymbol{A}) \Pr(j \in \boldsymbol{B})$ (that is, $p(\boldsymbol{A}) \perp p(\boldsymbol{B})$). Under the assumption that $\boldsymbol{B} \subseteq \mathcal{U}$ and that the selection mechanism of $\boldsymbol{B}$ involves independent draws, we establish the following set of remarks.

Remark 1. $\Pr(j \in \boldsymbol{A} \cap \boldsymbol{B}) > 0$ for all $j \in \boldsymbol{B}$.

Proof. By definition of a probability sample, the inclusion probabilities for each observation in the population must be known and strictly positive (Särndal et al., 2003). Since $\boldsymbol{B} \subseteq \mathcal{U}$, every unit in $\boldsymbol{B}$ has a strictly positive chance of being included in $\boldsymbol{A}$, even if $p(\boldsymbol{B})$ is non-ignorable. □

Remark 2. Let n_{M} denote the number of units in $\boldsymbol{M}$ and $n_{B_{\mathrm{Uq.}}}$ the number of unique units in $\boldsymbol{B}$. Then, $n_{B_{\mathrm{Uq.}}} \to N \implies n_{\mathrm{M}} \to n_{\mathrm{A}}$.

Proof. Note that $\max\left(n_{B_{\mathrm{Uq.}}}\right) = N$, as no sample from $\mathcal{U}$ can contain more information than $\mathcal{U}$ itself; therefore, $n_{B_{\mathrm{Uq.}}} = N$ implies that $\boldsymbol{B} = \mathcal{U}$, which further implies that $\boldsymbol{M} \coloneqq \boldsymbol{A} \cap \boldsymbol{B} = \boldsymbol{A}$. Therefore, as the number of unique elements in $\boldsymbol{B}$ increases to N, the number of observations in $\boldsymbol{M}$ also increases and converges to n_{A}. □

Remark 3. Assuming that a selection mechanism of $\boldsymbol{B}$ exists and is independent for all j,

$$n_{\mathrm{M}} \sim \mathrm{PoissonBinomial}\left(\lambda = \sum_{u \in \mathcal{U}} \Pr(u \in \boldsymbol{A}) \times \Pr(u \in \boldsymbol{B})\right)$$

and

$$\mathbb{E}_{\mathcal{D}}\left(\sum_{j \in \boldsymbol{B}} \Pr(j \in \boldsymbol{A}) \Pr(j \in \boldsymbol{B}) - n_{\mathrm{M}}\right) = 0,$$

where $\mathbb{E}_{\mathcal{D}}(\cdot)$ denotes the design-based expectation.

Proof. Let δ_j^A and δ_j^B denote the sample membership indicators for samples $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. Assuming that the sampling design of $\boldsymbol{B}$ involves independent draws and that $p(\boldsymbol{A}) \perp p(\boldsymbol{B})$, it follows that

$$\delta_j^A \delta_j^B \sim \text{Bernouli}\Big(\Pr(j \in \boldsymbol{A}) \times \Pr(j \in \boldsymbol{B}) \Big)$$

and

$$n_{\text{M}} = \sum_{j \in \boldsymbol{B}} \delta_j^A \delta_j^B.$$

The sum of independent Bernoulli trials with varying probabilities of success follows a Poisson Binomial distribution, which implies that $\boldsymbol{M}$ is a Poisson random sample. That is, for $k = \{1, 2, \ldots, N\}$,

$$p(\boldsymbol{M}) = \prod_{k \in \boldsymbol{M}} \{\Pr(k \in \boldsymbol{A}) \Pr(k \in \boldsymbol{B})\} \prod_{k \in \mathcal{U}/\boldsymbol{M}} \{1 - \Pr(k \in \boldsymbol{A}) \Pr(k \in \boldsymbol{B})\},$$

$$\mathbb{E}_{\mathcal{D}}(n_{\text{M}}) = \mathbb{E}_{\mathcal{D}} \left(\sum_{j \in \boldsymbol{B}} \delta_j^A \delta_j^B \right) = \sum_{u \in \mathcal{U}} \Pr(u \in \boldsymbol{A}) \Pr(u \in \boldsymbol{B}),$$

and

$$\text{Var}_{\mathcal{D}}(n_{\text{M}}) = \sum_{u \in \mathcal{U}} \Pr(u \in \boldsymbol{A}) \Pr(u \in \boldsymbol{B}) - \Big(\Pr(u \in \boldsymbol{A}) \Pr(u \in \boldsymbol{B}) \Big)^2,$$

where $\boldsymbol{M}$ belongs to the set of all 2^N subsets of $\mathcal{U}$ (Wang, 1993; Särndal et al., 2003). □

Therefore, for sufficiently large n_{B}, there will always exist an $\boldsymbol{M} \subseteq \boldsymbol{B}$ such that all units in $\boldsymbol{M}$ are also *exactly* in $\boldsymbol{A}$; and since $\pi_m \coloneqq \Pr(m \in \boldsymbol{M}) = \Pr(m \in \boldsymbol{B}) \Pr(m \in \boldsymbol{A})$, we would intuitively expect significant reductions in estimation bias as $n_{B_{\text{Uq.}}} \to N$ (Rivers, 2007).

4.1.2 Coarsened Exact Matching (CEM)

One limitation of exact matching is its tendency to discard too many observations, particularly when dealing with continuous covariates. This issue can be mitigated through *coarsened exact matching* (CEM), which performs exact matching after binning (or "coarsening") the data into discrete groups (Stuart et al., 2011). The concept is akin to that of Bethlehem (2016), as it involves (1) binning each $X_p \in \boldsymbol{X}$ into h_p strata, (2) crossing the binned variables into H groups, and (3) identifying the (binned) intersection between $\boldsymbol{A}$ and $\boldsymbol{B}$.

Although CEM tends to discard fewer data points than traditional exact matching, this depends heavily on the choice of h_p; setting h_p too low results in spurious estimates, while setting it too high replicates the exact matching procedure. In practice, h_p is either set to an arbitrary constant, typically between 3 and 5, or determined by Sturges (1926)'s rule, which for a dataset of size n is defined as $(1 + \log_2(n))$ (Scott, 2009).

4.1.3 Nearest-Neighbor (NN) Matching

Despite the overall simplicity of the (coarsened) exact matching procedure, several limitations remain worth discussing. For one, if $n_{\text{B}} \to N$, then $\hat{\mu}_B = n_{\text{B}}^{-1} \sum_{j \in \boldsymbol{B}} Y_j \to \mu_N$, which challenges the necessity of matching in the first place. We also question the construct validity of exact matching on covariates alone. A preferable alternative is *nearest-neighbor* (NN) matching, which

uses a distance measure, say $d(\cdot)$, to pair each observation in $\boldsymbol{A}$ with its closest neighbor in $\boldsymbol{B}$. The procedure is straightforward: for each $i \in \boldsymbol{A}$, find the $j \in \boldsymbol{B}$ whose distance d_{ij} is the smallest in the sample. These matching methods are considered "greedy" because they do not optimize an overall objective; nevertheless, they are widely used in the matching literature due to their simplicity and tendency to create well-matched groups (Stuart et al., 2011; Stuart, 2010).

Recall the NN matching algorithm proposed by Rivers (2007), which pairs each $i \in \boldsymbol{A}$ with an $m \in \boldsymbol{B}$ such that $|\boldsymbol{X}_m - \boldsymbol{X}_i| \leqslant |\boldsymbol{X}_j - \boldsymbol{X}_i|$ for all $j \in \boldsymbol{B}$. This procedure is theorized to bring the joint distribution of $\boldsymbol{X}$ in $\boldsymbol{M}$ closer to that in $\boldsymbol{A}$, assuming that closely matched covariates yield, on average, similar Y values, i.e., $\mathbb{E}_{\xi}(Y|\boldsymbol{X})$ is almost surely Lipschitz continuous on the support of $\boldsymbol{X}$. However, since observations are matched based only on $\boldsymbol{X}$, any information linking $\boldsymbol{X}$ to Y is effectively discarded.

To efficiently utilize all available information in $\boldsymbol{A}$ and $\boldsymbol{B}$, we propose two NN matching algorithms: one for continuous (or, at least, unique) Y_B values and another that accommodates general cases including noncontinuous Y_B with ties. These algorithms are presented below.

4.1.4 NN Matching: Unique Y_{B}

If $\mathbb{E}(Y|\boldsymbol{X})$ were known a priori, one could simply pair each unit $i \in \boldsymbol{A}$ with a donor $m \in \boldsymbol{B}$ that satisfies $|\mathbb{E}(Y_i|\boldsymbol{X}_i) - Y_m| \leqslant \left|\mathbb{E}(Y_i|\boldsymbol{X}_i) - Y_j\right|$ for all $j \in \boldsymbol{B}$, assuming that there exists no $j, k \in \boldsymbol{B}$ such that $Y_j = Y_k$. Since $\mathbb{E}(Y|\boldsymbol{X})$ is nonetheless unknown, we consider replacing it with $\widehat{\mathbb{E}}_B(Y_i|\boldsymbol{X}_i) \coloneqq \hat{Y}^B_{A,i}$, the estimated value of $Y_{A,i}$ derived from a model built on $\boldsymbol{B}$ and applied to the covariates in $\boldsymbol{A}$. The corresponding MMI estimator is presented in Algorithm 2.

Algorithm 2 Matched mass imputation under NN matching (unique Y_{B}).

1. Obtain $\hat{Y}^B_A$ using some outcome model trained on $\boldsymbol{B}$.
2. Match, with replacement, each unit $i \in \boldsymbol{A}$ with a unit $m \in \boldsymbol{B}$ such that

$$\left|\hat{Y}^B_{A,i} - Y_m\right| \leqslant \left|\hat{Y}^B_{A,i} - Y_j\right| \quad \forall j \in \boldsymbol{B}.$$

   These units now belong to $\boldsymbol{M}_U$, a set of nearest-neighbor statistical matches from $\boldsymbol{B}$ with unique (highlighted by U) Y values.
3. Build a regression model on $\boldsymbol{M}_U$ to obtain $\hat{Y}^M_{A:U}$.
4. Calculate $\hat{\mu}_{\mathrm{MMIunn}} = \hat{N}^{-1}\Pi^{\top}\hat{Y}^M_{A:U} = \hat{N}^{-1}\sum_{i\in\boldsymbol{A}}\pi_i^{-1}\hat{Y}^M_{A,i:U}$.

Using $\left|\hat{Y}^B_{A,i} - Y_j\right|$ in place of $\left|\boldsymbol{X}_i - \boldsymbol{X}_j\right|$ allows each covariate to be weighted by its estimated relationship with Y, effectively accounting for covariates that may be discrete, nonlinear, or of varying importance.

4.1.5 NN Matching: Tied Y_{B}

Note that, if the variance of $\hat{Y}^M_{A:U}$ is too small, or if there exists repeated Y in $\boldsymbol{B}$, the risk of some $j \in \boldsymbol{B}$ being sampled too often under Algorithm 2 becomes substantial. While the former issue cannot be addressed in practice, the latter can be avoided by assuming that Y in $\boldsymbol{B}$ is unique. This assumption is rather strong and may not hold in practice (see Section 6 for an example).

To address this issue, in Algorithm 3, we consider matching unit $i \in \boldsymbol{A}$ to satisfy $|\|\mathbf{a}_i\| - \|\mathbf{b}_m\|| \leqslant |\|\mathbf{a}_i\| - \|\mathbf{b}_j\||$ for all $j \in \boldsymbol{B}$, where $\|\cdot\|$ denotes the L_2 norm, $\mathbf{a}_i \coloneqq \begin{bmatrix}\hat{Y}^B_{A,i} & \boldsymbol{X}_i\end{bmatrix}$, and

Algorithm 3 Matched mass imputation under NN matching (with ties).

1. Obtain $\hat{Y}_A^B$ using some outcome model trained on $\boldsymbol{B}$.
2. Calculate
$$\mathcal{N}_A = \begin{bmatrix} \|\mathbf{a}_1\| \\ \|\mathbf{a}_2\| \\ \cdots \\ \|\mathbf{a}_{n_\mathrm{A}}\| \end{bmatrix} \quad \text{and} \quad \mathcal{N}_B = \begin{bmatrix} \|\mathbf{b}_1\| \\ \|\mathbf{b}_2\| \\ \cdots \\ \|\mathbf{b}_{n_\mathrm{B}}\| \end{bmatrix},$$
where $\mathbf{a}_i \coloneqq \begin{bmatrix} \hat{Y}_{A,i}^B & \boldsymbol{X}_i \end{bmatrix}$, $\mathbf{b}_j \coloneqq \begin{bmatrix} Y_{B,j} & \boldsymbol{X}_j \end{bmatrix}$, and $\|\cdot\|$ denotes the L_2 norm.
3. Match, with replacement, each unit $i \in \boldsymbol{A}$ with a $m \in \boldsymbol{B}$ such that
$$\left|\mathcal{N}_{A,i} - \mathcal{N}_{B,m}\right| \leqslant \left|\mathcal{N}_{A,i} - \mathcal{N}_{B,j}\right| \quad \forall j \in \boldsymbol{B}.$$
These units now belong to $\boldsymbol{M}_R$, a set of nearest-neighbor statistical matches from $\boldsymbol{B}$ with possibly repeated (highlighted by R) Y.
4. Build a regression model on $\boldsymbol{M}_R$ to obtain $\hat{Y}_{A:R}^M$.
5. Calculate $\hat{\mu}_{\mathrm{MMIrnn}} = \hat{N}^{-1}\Pi^\top \hat{Y}_{A:R}^M = \hat{N}^{-1}\sum_{i\in\boldsymbol{A}} \pi_i^{-1}\hat{Y}_{A,i:R}^M$.

$\mathbf{b}_j \coloneqq \begin{bmatrix} Y_{B,j} & \boldsymbol{X}_j \end{bmatrix}$. Assuming that $\boldsymbol{X}$ is numeric and unique, as in Rivers (2007), the requirement for unique Y_B becomes moot. Another potential benefit of explicitly including $\boldsymbol{X}$ in the matching procedure is that it may reduce reliance on $\widehat{\mathbb{E}}_B(Y_i|\boldsymbol{X}_i)$, thus protecting against potential bias in sample $\boldsymbol{B}$. However, a key drawback is that noisy covariates in $\boldsymbol{X}$ can degrade the matching quality. Additionally, the computational burden of calculating the L_2 norm increases when $\boldsymbol{X}$ is high-dimensional.

Despite their differences, both Algorithms 2 and 3 seek to identify an $\boldsymbol{M} \subset \boldsymbol{B}$ with maximal similarity to $\boldsymbol{A}$, in an attempt to align the distributions of $\delta^B \mid Y, \boldsymbol{X}$ and $\delta^A \mid Y, \boldsymbol{X}$, thus providing robustness to nonignorability in the design of $\boldsymbol{B}$, and consequently, protecting against possible violations of the transportability assumption. Given this rationale, MMI estimators are expected to outperform mass imputation estimators when the assumption of ignorability is violated. However, under ignorability, mass imputation estimators may be preferable, as MMI may needlessly discard observations from sample $\boldsymbol{B}$. Since the ignorability assumption is untestable, practitioners should use their judgment about the nonprobability sample's potential deviation from ignorability when choosing between MMI estimators and standard mass imputation. The performance of the MMI estimators from Algorithms 2 and 3 is thoroughly assessed through Monte Carlo simulations in Section 5 and a real data application in Section 6.

4.2 Variance Estimation

We conclude this section by describing a bootstrap variance estimator for the sampling variance of $\hat{\mu}_{\mathrm{MMI}}$. Following Kim et al. (2021), we first generate L sets of replicate weights for sample $\boldsymbol{A}$ and draw L bootstrap samples (of size n_B) with replacement from sample $\boldsymbol{B}$. We then use each weight-resample pair to compute the MMI estimator using (coarsened) exact or NN matching (say, $\hat{\mu}_I^l$, $l = 1, 2, \ldots, L$). Using both the replicate estimates ($\hat{\mu}_I^l$) and the MMI estimator from

the original $\boldsymbol{A}$ and $\boldsymbol{B}$ sets (say, $\hat{\mu}_I$), we calculate the following bootstrap variance estimator:

$$\widehat{\text{Var}}_{\text{Boot}}\left(\hat{\mu}_I\right) = \frac{1}{L}\sum_{l=1}^{L}\left(\hat{\mu}_I^l - \hat{\mu}_I\right)^2. \tag{10}$$

The estimator in (10) can be used to construct approximately $(1-\alpha)\times 100\%$ confidence intervals for μ_N as follows:

$$\hat{\mu}_I \pm z_{\alpha/2} \times \sqrt{\widehat{\text{Var}}_{\text{Boot}}\left(\hat{\mu}_I\right)},$$

where $z_{\alpha/2}$ is the z-score corresponding to the $(1-\alpha/2)$ percentile under the standard normal distribution. We explore the behavior of this bootstrap variance estimator and the corresponding confidence intervals in the following section.

5 Simulation Study

We conducted a Monte Carlo study to contrast the performance of our MMI estimators with various competitors. We also examined the performance of the bootstrap variance estimator described above.

5.1 Simulation Settings

To thoroughly assess the performance of the proposed MMI estimators, we generated finite populations of size $N = 100{,}000$ from five superpopulation models:

- Model ξ_1:

$$Y = \sum_{a=1}^{2} 4X_a + \sum_{b=3}^{4} 2X_b + \epsilon,$$

  where $X_1, X_2 \sim \text{N}\left(\mu = 2, \sigma = 1\right)$, $X_3, X_4 \sim \text{N}\left(\mu = 4, \sigma = 1\right)$, and $\epsilon \sim \text{N}\left(\mu = 0, \sigma = 3\right)$.
- Model ξ_2:

$$Y = \sum_{a=1}^{2} 4X_a + \sum_{b=3}^{4} 2X_b + \sum_{c\in\{1,3\}} \left(X_c + X_{c+1}\right)^2 + \epsilon,$$

  where $X_1, X_2 \sim \text{Uniform}\left(0, 4\right)$, $X_3, X_4 \sim \text{Uniform}\left(4, 8\right)$, and $\epsilon \sim \text{N}\left(\mu = 0, \sigma = 17.5\right)$.
- Model ξ_3 (Maia et al., 2021):

$$\begin{aligned} Y = {} & X_1 + 2.121X_2^2 + 2\mathbb{1}\left(|X_3| > \bar{X}_3\right) + 2.619\log\left(|X_1|\right)|X_3| \\ & + 2.682X_2X_4 + 6\mathbb{1}\left(|X_5| > \bar{X}_5\right) + 1.392\text{e}^{X_6} + \epsilon, \end{aligned}$$

  where $\mathbb{1}\left(\cdot\right)$ denotes the usual indicator function, $X_1, \ldots, X_6 \sim \text{Uniform}(-5, 5)$, and $\epsilon \sim \text{N}(\mu = 0, \sigma = \sqrt{1.8})$.
- Model ξ_4:

$$\begin{aligned} Y = {} & 4X_1^2 + 2X_2^2X_3\text{e}^{-|X_4|} + 1.5X_6\log\left(|X_5|\right) - 12X_5\log\left(|X_6|\right) + \sin X_3^2\log(X_3^2) \\ & + \sin\left(X_3X_4 - X_5\right) - \frac{1}{\text{e}^{1/X_4}} + X_2^2X_3^2 - X_3^2X_1^2\log\left(|X_2 - X_5|\right) + \sin\left(X_4X_5^2\right) + \epsilon, \end{aligned}$$

  where $X_1, X_3, X_5 \sim \text{Uniform}(-5, 5)$, $X_2, X_4, X_6 \sim \text{Uniform}(5, 10)$, and $\epsilon \sim \text{N}(\mu = 300, \sigma = 20\sqrt{30})$.

- Model ξ_5: Same as model ξ_4, except $\epsilon \sim \mathrm{N}(\mu = 750, \sigma = 50\sqrt{30})$.

From each finite population, a simple random sample without replacement of size $n_\mathrm{A} = 1{,}000$ was used to obtain sample $\boldsymbol{A}$. For sample $\boldsymbol{B}$ with $\boldsymbol{n}_\mathrm{B} = [n_\mathrm{A} \quad 10n_\mathrm{A} \quad 20n_\mathrm{A}]$, stratified simple random sampling was used as follows: first, all units $u \in \mathcal{U}$ with $Y_u < \mathrm{Median}\,(Y)$ were placed in stratum I, and the remaining units were placed in stratum II. Then, to allow an informative $p(\boldsymbol{B})$, we selected $\boldsymbol{r} \times \boldsymbol{n}_\mathrm{B}$ and $(1 - \boldsymbol{r}) \times \boldsymbol{n}_\mathrm{B}$ observations from each respective stratum, where $\boldsymbol{r} = \big(0.50\ 0.30\ 0.15\ 0.05\ 0.01\big)$. Note that when $r = 0.5$, both strata are sampled at equal rates, resulting in a noninformative design for $\boldsymbol{B}$. As r deviates from 0.50, the design of $\boldsymbol{B}$ becomes increasingly informative, making violations of the ignorability condition more likely. Following these steps, samples $\boldsymbol{A}$ and $\boldsymbol{B}$ were drawn $n_\mathrm{sim} = 1000$ times and used to calculate the following sets of estimators:

- Existing estimators
  - $\hat{\mu}_\mathrm{B}$: The naive mean of $\boldsymbol{B}$.
  - $\hat{\mu}_\mathrm{MIgam}$: The MIgam with natural cubic splines (see Eq. 4).
  - $\hat{\mu}_\mathrm{DRgam}$: The doubly-robust estimator using logistic propensity scores and GAMs (GAM was fitted using `R`'s `lm()` and `bs()` base functions with 10 degrees of freedom) with natural cubic splines (see Eq. 6).
  - $\hat{\mu}_\mathrm{SMexact}$, $\hat{\mu}_\mathrm{SMcem}$, $\hat{\mu}_\mathrm{SMunn}$, and $\hat{\mu}_\mathrm{SMrnn}$: The Hájek estimator on exact, coarsened exact (using Sturges (1926)'s rule), and nearest-neighbor matches (using Algorithms 2 and 3, respectively; see Eq. 7 and 8). Exact and CEM models were built using Stuart et al. (2011)'s `MatchIt` package in `R`.
- Proposed estimators
  - $\hat{\mu}_\mathrm{MMIexact}$, $\hat{\mu}_\mathrm{MMIcem}$, $\hat{\mu}_\mathrm{MMIunn}$, and $\hat{\mu}_\mathrm{MMIrnn}$: Matched mass imputation estimators using GAMs trained on exact, coarsened exact, and nearest-neighbor matches (see Eq. 9).

For the sake of space, the evaluation of the variance estimator and corresponding confidence intervals was restricted to model ξ_1 with $n_\mathrm{A} = 1{,}000$, $n_\mathrm{B} = 20n_\mathrm{A}$, and $\boldsymbol{r} = (.50\ .30\ .15)$. $L = 500$ sets of replicate weights were generated from sample $\boldsymbol{A}$, and corresponding samples with replacement (of size n_B) were generated from sample $\boldsymbol{B}$; these were jointly used to calculate each of the aforementioned $\hat{\mu}$, whose values were subsequently used to calculate $\widehat{\mathrm{Var}}_\mathrm{Boot}$ using Eq. (10). This process was repeated $n_\mathrm{sim} = 500$ times to calculate the following performance metrics.

5.2 Performance Metrics

We evaluate the performance of the various mean estimators described above by the root mean squared error ratio (RMSER):

$$\mathrm{RMSER}\left(\hat{\mu}\right) = \frac{\mathrm{RMSE}\left(\hat{\mu}\right)}{\mathrm{RMSE}\left(\hat{\mu}_\pi\right)}, \tag{11}$$

and the absolute bias ratio (ABR):

$$\mathrm{ABR}\left(\hat{\mu}\right) = \left|\frac{\mathrm{Bias}\left(\hat{\mu}\right)}{\mathrm{Bias}\left(\hat{\mu}_\pi\right)}\right|,$$

where $\hat{\mu}_\pi$ denotes the 'gold-standard' Horvitz-Thompson (HT) estimator from sample $\boldsymbol{A}$. The performance of the bootstrap variance estimator is assessed using the following statistics:

- CL: The percentage of instances for which μ_N was captured in the 90% confidence interval.
- LL (UL): The mean of the confidence intervals' lower (upper) limit.
- $\bar{\hat{\mu}}_I$: The Monte Carlo mean of $\hat{\mu}_I$.
- $\bar{d}$: The average width of the confidence intervals.
- $\bar{\widehat{\text{Var}}}_{\text{Boot}}\left(\hat{\mu}_I\right)$: The Monte Carlo mean of $\widehat{\text{Var}}_{\text{Boot}}\left(\hat{\mu}_I\right)$.
- RB: The percent relative bias of $\bar{\widehat{\text{Var}}}_{\text{Boot}}$ relative to its Monte Carlo equivalent,

$$\widehat{\text{Var}}_{\text{MC}}(\hat{\mu}_I) = \frac{1}{n_{\text{sim}} - 1} \sum_{m=1}^{n_{\text{sim}}} \left(\hat{\mu}_{I,m} - \frac{1}{n_{\text{sim}}} \sum_{m=1}^{n_{\text{sim}}} \hat{\mu}_{I,m} \right)^2, \tag{12}$$

defined as

$$\text{RB} = \frac{\widehat{\text{Var}}_{\text{MC}} - \bar{\widehat{\text{Var}}}_{\text{Boot}}}{\widehat{\text{Var}}_{\text{MC}}} \times 100.$$

5.3 Mean Estimation Results

5.3.1 RMSER

The RMSER results are presented in Figure 1 below and in Tables 1(a), 2(a), and 3(a) in the Supplementary Material. We begin by addressing the instability of $\hat{\mu}_{\text{MMIexact}}$ and $\hat{\mu}_{\text{MMIcem}}$ when $n_{\text{B}} = n_{\text{A}}$. In this setting, the expected number of exact matches shrinks to 10, which is too small to support stable GAM smoothing. While coarsened exact matches provided some robustness—particularly for models with fewer covariates (ξ_1 and ξ_2)—the only true remedy was increasing n_{B}. Even then, the RMSERs for these estimators were either marginally better or noticeably worse than those of $\hat{\mu}_{\text{MIgam}}$.

On the other hand, both $\hat{\mu}_{\text{MMIunn}}$ and $\hat{\mu}_{\text{MMIrnn}}$ effectively mitigated $\boldsymbol{B}$'s selection bias. Except at $r = 0.50$, i.e., when $\boldsymbol{B}$ is drawn by sampling at equal rates from the two strata, the effect of nearest-neighbor matching on mass imputation is substantial, yielding consistent efficiency gains across varying degrees of bias, from slight ($r = 0.30$) to moderate ($r = 0.15$), large ($r = 0.05$), and severe ($r = 0.01$). The superiority of these MMI estimators was evident across four of the five models, with the only exception occurring under model ξ_4 when $r = 0.30$, where they exhibited slightly higher RMSER values than the standard mass imputation estimator $\hat{\mu}_{\text{MIgam}}$. These results suggest that both $\hat{\mu}_{\text{MMIunn}}$ and $\hat{\mu}_{\text{MMIrnn}}$ provide considerable robustness to nonignorable bias in the design of sample $\boldsymbol{B}$. Although the differences between the two estimators were minor, some trends emerged: $\hat{\mu}_{\text{MMIunn}}$ performed better for well-specified $\widehat{\mathbb{E}}_B(Y|\boldsymbol{X})$ (i.e., the linear ξ_1 and quadratic ξ_2), whereas $\hat{\mu}_{\text{MMIrnn}}$ was superior under slight to moderate model misspecification (ξ_3 and ξ_4).

The RMSERs for the cold-deck estimators were less impressive: $\hat{\mu}_{\text{SMexact}}$ and $\hat{\mu}_{\text{SMcem}}$ closely mirrored $\hat{\mu}_{\text{B}}$, while $\hat{\mu}_{\text{SMrnn}}$ and $\hat{\mu}_{\text{SMunn}}$ resembled $\hat{\mu}_{\text{MIgam}}$. Evaluations of alternative cold-deck estimators may yield different findings and are of interest for future research. Nevertheless, the current results overwhelmingly support matched mass imputation over its cold-deck counterparts, particularly in the context of nonignorable sample design for $\boldsymbol{B}$.

5.3.2 ABR

The ABR results are displayed in Figure 2 below and in Tables 1(b), 2(b), and 3(b) in the Supplementary Material. Performance trends closely mirrored those observed under RMSER, though

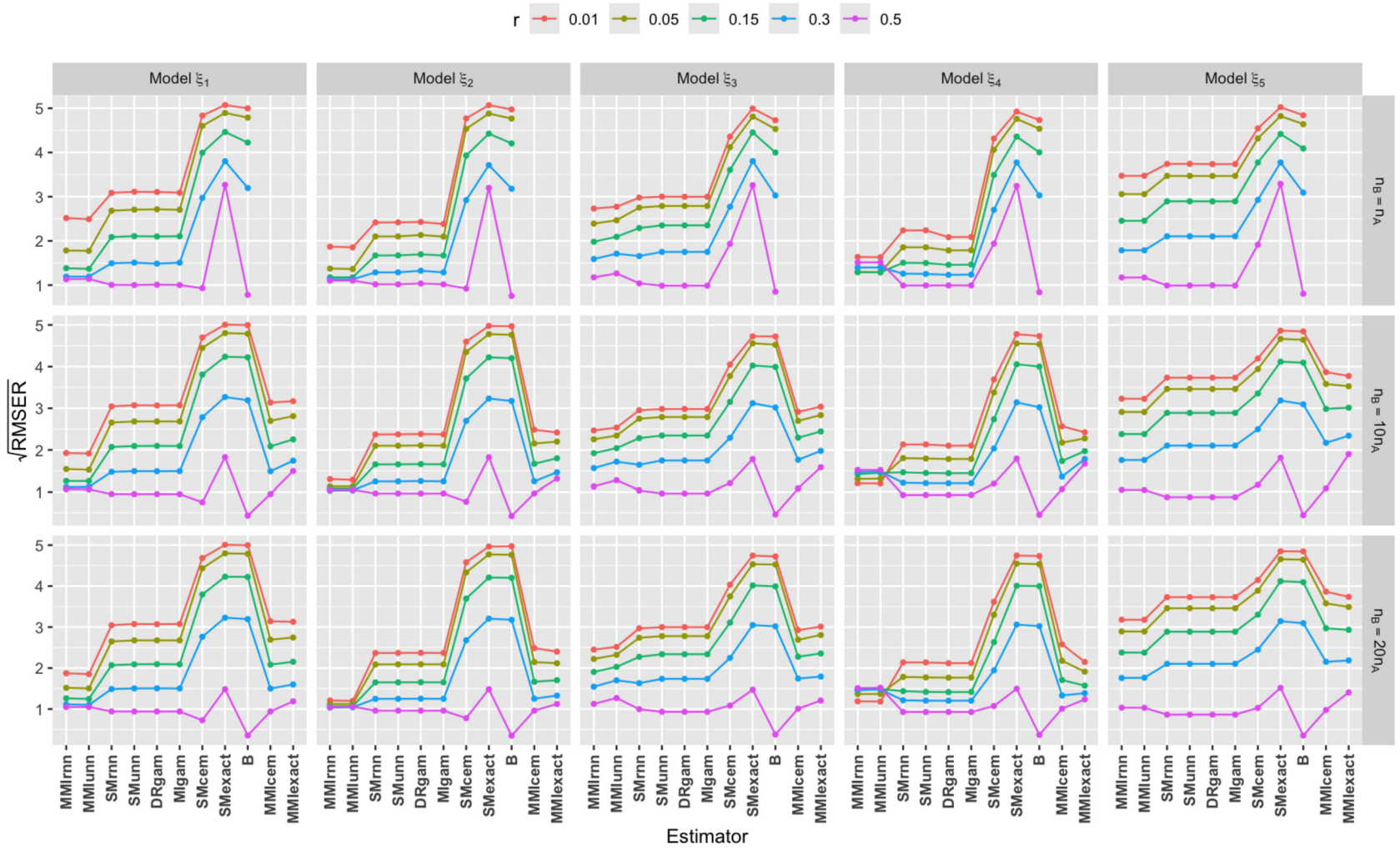

Figure 1: Square root of RMSER for eleven mean estimators. To improve readability, $\hat{\mu}_{\text{MMIcem}}$ and $\hat{\mu}_{\text{MMIexact}}$ are omitted when $n_{\text{B}} = n_{\text{A}}$ due to their extreme values.

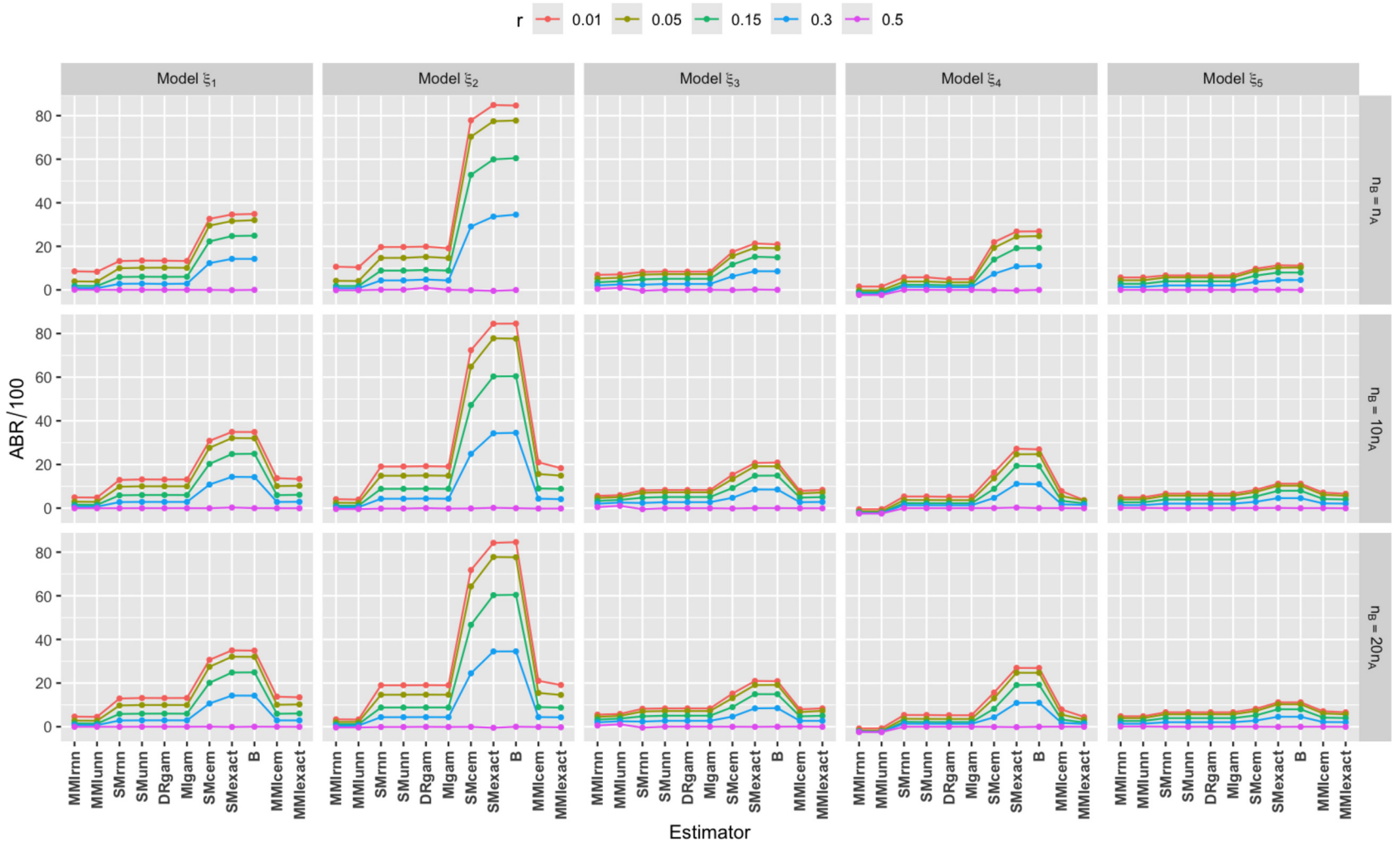

Figure 2: ABR divided by 100 for eleven mean estimators. To improve readability, $\hat{\mu}_{\text{MMIcem}}$ and $\hat{\mu}_{\text{MMIexact}}$ are omitted when $n_{\text{B}} = n_{\text{A}}$ due to their extreme values.

some additional observations warrant discussion. First, the ABRs for $\hat{\mu}_{\text{MMIunn}}$ and $\hat{\mu}_{\text{MMIrnn}}$ under $r \neq 0.5$ were noticeably smaller overall, suggesting that both estimators can effectively accommodate informative nonprobability samples. However, results for $r = 0.5$ highlight the consequence of needlessly discarding useful information. When $\boldsymbol{B}$ is truly representative of $\mathcal{U}$ (as is the case when $r = 0.5$), it is optimal to retain as many observations as possible. In such cases, discarding data based on Y induces missingness not at random (MNAR) in a sample that would otherwise be at least missing at random (MAR).

Additionally, we observed that both ABR and RMSER for $\hat{\mu}_{\text{DRgam}}$ were nearly identical to those of $\hat{\mu}_{\text{MIgam}}$. Upon further inspection, it appears that the propensity score model yielded highly inaccurate inclusion probability estimates. The double-robust property of $\hat{\mu}_{\text{DRgam}}$ mitigated this issue by effectively eliminating propensity-based influence, leaving predictions that were primarily model-driven through GAM.

5.4 Variance Estimation Results

The variance estimation results are presented in Table 2. We begin by noting a severely biased $\widehat{\text{Var}}$, which, for most mean estimators, was significantly smaller than the Monte Carlo variance $\widehat{\text{Var}}_{\text{MC}}$. These underestimated variances led to confidence intervals that were too narrow, frequently excluding μ_N at disproportionate rates. Consequently, coverage rates were significantly lower under $r = 0.30$ and $r = 0.15$, even for estimators that fared quite well in mean estimation, such as $\hat{\mu}_{\text{MMIunn}}$ and $\hat{\mu}_{\text{MMIrnn}}$. The only exception occurred at $r = 0.50$, where most mean estimators achieved nominal coverage or close to nominal. However, $\hat{\mu}_{\text{B}}$, $\hat{\mu}_{\text{SMcem}}$, $\hat{\mu}_{\text{SMunn}}$, and $\hat{\mu}_{\text{SMrnn}}$ deviated from this trend: $\hat{\mu}_{\text{B}}$ produced overly conservative intervals due to large variance estimates, while the others were too liberal.

Overall, these results are unsatisfactory, as they indicate a strong reliance on MAR missingness for valid bootstrap variance estimates. If this assumption fails, the procedure proposed by Kim et al. (2021) will severely underestimate the true sampling variance. Alternative variance estimation methods, such as linearization estimators, could be explored to improve the robustness of variance estimation for MMI. We shall investigate this further in future research.

6 Real Data Application

In this section, we evaluate the performance of our proposed estimators using the National Health and Nutrition Examination Survey (NHANES) dataset published by the Centers for Disease Control and Prevention (CDC) (2015–2020). NHANES is a multistage, stratified random sample designed to be nationally representative of non-institutionalized individuals in the United States.

In this study, we estimate total cholesterol (Y, in mg/dL) in the U.S. adult population using five covariates: age (X_1), glycohemoglobin (X_2, in %), triglycerides (X_3, in mg/dL), direct high-density lipoprotein cholesterol (X_4, in mg/dL), and body mass index (X_5, in kg/m^2). The probability sample ($\boldsymbol{A}$, $n_{\text{A}} = 2{,}701$) consists of individuals with available data from the 2015–2016 NHANES round, while the nonprobability sample ($\boldsymbol{B}$, $n_{\text{B}} = 4{,}558$) comprises individuals with available data from the 2017–2020 rounds. Notably, the CDC classified observations from the 2019–2020 round as 'nonrepresentative' due to confidentiality concerns, requiring their merger with the 2016–2017 round. Readers interested in accessing the 2019–2020 dataset must submit a formal request through the CDC website.

Table 2: Performance statistics for the bootstrap variance estimator in Eq. (10) and corresponding 90% confidence intervals.

	Performance	$\hat{\mu}_{\text{B}}$	$\hat{\mu}_{\text{MIgam}}$	$\hat{\mu}_{\text{DRgam}}$	$\hat{\mu}_{\text{SMexact}}$	$\hat{\mu}_{\text{SMcem}}$	$\hat{\mu}_{\text{SMunn}}$	$\hat{\mu}_{\text{SMrnn}}$	$\hat{\mu}_{\text{MMIexact}}$	$\hat{\mu}_{\text{MMIcem}}$	$\hat{\mu}_{\text{MMIunn}}$	$\hat{\mu}_{\text{MMIrnn}}$
	CL	99.80	90.20	90.20	89.00	61.20	14.20	14.60	92.00	90.20	86.00	86.20
	LL	31.90	31.67	31.67	31.16	31.89	31.96	31.96	31.42	31.66	31.63	31.62
	$\bar{\hat{\mu}}_I$	31.98	32.00	32.00	31.98	31.99	32.00	31.99	31.98	31.99	32.00	32.00
$r = .50$	UL	32.07	32.33	32.33	32.79	32.09	32.03	32.03	32.55	32.33	32.37	32.37
	$\bar{d}$	0.16	0.66	0.66	1.63	0.20	0.07	0.07	1.12	0.66	0.75	0.76
	$\widehat{\bar{\text{Var}}}_{\text{Boot}}$	0.00	0.04	0.04	0.25	0.00	0.00	0.00	0.12	0.04	0.05	0.05
	RB	−270.40	−1.90	−1.80	3.10	74.00	98.80	98.70	−23.8	−1.30	17.20	18.60
	CL	0.00	30.60	30.60	0.00	0.00	1.80	1.40	63.40	31.40	81.60	82.60
	LL	34.14	32.10	32.10	33.44	33.56	32.39	32.37	31.85	32.10	31.70	31.71
	$\bar{\hat{\mu}}_I$	34.22	32.42	32.43	34.21	33.65	32.42	32.41	32.43	32.42	32.08	32.08
$r = .30$	UL	34.29	32.75	32.75	34.98	33.75	32.46	32.45	33.01	32.75	32.46	32.45
	$\bar{d}$	0.15	0.65	0.65	1.54	0.19	0.07	0.08	1.16	0.65	0.75	0.74
	$\widehat{\bar{\text{Var}}}_{\text{Boot}}$	0.00	0.04	0.04	0.22	0.00	0.00	0.00	0.12	0.04	0.05	0.05
	RB	−217.10	7.70	7.70	1.60	74.00	98.80	98.70	−9.40	7.60	23.70	20.30
	CL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	20.40	0.00	68.60	69.00
	LL	35.82	32.62	32.63	35.24	35.06	32.89	32.87	32.29	32.61	31.85	31.86
	$\bar{\hat{\mu}}_I$	35.89	32.93	32.94	35.91	35.14	32.93	32.91	32.95	32.93	32.22	32.23
$r = .15$	UL	35.96	33.24	33.24	36.59	35.22	32.97	32.95	33.60	33.24	32.59	32.59
	$\bar{d}$	0.14	0.62	0.62	1.35	0.16	0.08	0.09	1.31	0.62	0.74	0.73
	$\widehat{\bar{\text{Var}}}_{\text{Boot}}$	0.00	0.04	0.04	0.17	0.00	0.00	0.00	0.16	0.04	0.05	0.05
	RB	−155.40	−4.30	−4.50	7.30	70.80	98.00	98.00	−16.40	−3.60	14.50	14.90

Table 3: Percent relative bias (%RB) of eleven mean estimators relative to $\hat{\mu}_\pi = 188.05$ obtained from the 2015–2016 NHANES dataset (sample $\boldsymbol{A}$).

Estimator	$\hat{\mu}_{\text{MMIrnn}}$	$\hat{\mu}_{\text{SMrnn}}$	$\hat{\mu}_{\text{MMIcem}}$	$\hat{\mu}_{\text{MIgam}}$	$\hat{\mu}_{\text{SMcem}}$	$\hat{\mu}_{\text{B}}$	$\hat{\mu}_{\text{DRgam}}$	$\hat{\mu}_{\text{SMunn}}$	$\hat{\mu}_{\text{MMIunn}}$	$\hat{\mu}_{\text{SMexact}}$	$\hat{\mu}_{\text{MMIexact}}$
$\hat{\mu}$	186.50	186.11	185.56	185.44	184.27	179.45	167.02	.	.	.	.
%RB	0.82	1.03	1.32	1.39	2.01	4.57	11.18	.	.	.	.

Using the two samples, we computed the eleven estimators described in the previous section and their percent relative bias (%RB) defined as

$$\%\text{RB} = \frac{\hat{\mu}_\pi - \hat{\mu}}{\hat{\mu}_\pi} \times 100.$$

The results are presented in Table 3.

We begin by addressing the absence of $\hat{\mu}_{\text{MMIexact}}$, $\hat{\mu}_{\text{SMexact}}$, $\hat{\mu}_{\text{MMIunn}}$, and $\hat{\mu}_{\text{SMunn}}$. The exact match procedure failed to produce any matches, leading to the omission of its associated mean estimators from the table. Additionally, only 5.4% of $\boldsymbol{Y}_{\text{B}}$ was unique, making one-to-one matching on $\hat{\boldsymbol{Y}}_A^B$ and $\boldsymbol{Y}_B$ infeasible. Specifically, each unit $i \in \boldsymbol{A}$ is matched with the first $m \in \boldsymbol{B}$ that satisfies the matching criteria, rendering $\hat{\mu}_{\text{MMI}}$ sensitive to the arbitrary ordering of $\boldsymbol{B}$.

Nevertheless, we observe that $\hat{\mu}_{\text{MMIrnn}}$ had the lowest percent relative bias, which was substantially lower than that of $\hat{\mu}_{\text{MIgam}}$ (0.82 vs. 1.39) and even more so compared to $\hat{\mu}_{\text{B}}$ (0.82 vs. 4.57). These results, along with those presented in Section 5, further confirm that our nearest-neighbor MMI estimators provide substantial robustness to nonignorable bias in $\boldsymbol{B}$. We also note that both $\hat{\mu}_{\text{SMrnn}}$ and $\hat{\mu}_{\text{MMIcem}}$ outperformed $\hat{\mu}_{\text{MIgam}}$. The bias of $\hat{\mu}_{\text{DRgam}}$ was notably higher than that of $\hat{\mu}_{\text{MIgam}}$, primarily due to severe misspecification of the propensity score model.

7 Discussion

In this paper, we addressed the problem of combining data from a probability sample ($\boldsymbol{A}$) and a nonprobability sample ($\boldsymbol{B}$) to estimate the finite population mean. We introduced the *matched mass imputation* (MMI) procedure, which replaces $\boldsymbol{B}$ in the mass imputation framework with a set of statistical matches ($\boldsymbol{M} \subset \boldsymbol{B}$). This approach shifts the burden of the ignorability assumption away from $\boldsymbol{B}$ and toward a less biased subset, thereby improving robustness.

The empirical simulations and real data analysis strongly support this approach, with our proposed nearest-neighbor mean estimators ($\hat{\mu}_{\text{MMIrnn}}$ and $\hat{\mu}_{\text{MMIunn}}$) demonstrating the best overall performance. We also explored variance estimation using the bootstrapping procedure proposed by Kim et al. (2021). However, our results indicated a severe underestimation of sampling variance when the ignorability assumption was violated. This finding highlights the need for future research on alternative variance estimation techniques to enhance the reliability of MMI-based inference. Future studies may also explore a formal investigation of the asymptotic properties of the MMI estimators, particularly $\hat{\mu}_{\text{MMIrnn}}$ and $\hat{\mu}_{\text{MMIunn}}$. This aspect was not pursued in the current work, as our primary objective was to introduce the MMI approach in its general form without restricting the discussion to a specific member of the MMI class.

Supplementary Material

The supplementary material includes the following: (1) additional simulation results showing the RMSER and ABR results in tabular format to complement the visualizations in Figures 1 and 2, (2) R code, and (3) README: a brief explanation of how to run the code.

Funding

This work of Jeremy Flood was funded by the North Carolina A&T State University Chancellor's Distinguished Fellowship, a Title III HBGI grant from the U.S. Department of Education.

References

Beaumont JF, Rao J (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, 83: 11–22.

Bethlehem J (2016). Solving the nonresponse problem with sample matching? *Social Science Computer Review*, 34(1): 59–77. https://doi.org/10.1177/0894439315573926

Centers for Disease Control and Prevention (CDC) (2015–2020). NHANES - National Health and Nutrition Examination Survey. https://www.cdc.gov/nchs/nhanes/index.htm (visited: 2023-10-11).

Chen S, Yang S, Kim JK (2022). Nonparametric mass imputation for data integration. *Journal of Survey Statistics and Methodology*, 10(1): 1–24. https://doi.org/10.1093/jssam/smaa036

Chen Y, Li P, Wu C (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532): 2011–2021. https://doi.org/10.1080/01621459.2019.1677241

Dever J (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. In: *Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference*, 1–15.

Hájek J (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4): 1491–1523.

Horvitz DG, Thompson DJ (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260): 663–685. https://doi.org/10.1080/01621459.1952.10483446

James G, Witten D, Hastie T, Tibshirani R, et al. (2013). *An Introduction to Statistical Learning*, volume 112. Springer.

Kalay AF (2021). Double Robust Mass-Imputation with Matching Estimators. arXiv preprint: https://arxiv.org/abs/2110.09275.

Kern C, Li Y, Wang L (2021). Boosted kernel weighting–using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*, 9(5): 1088–1113. https://doi.org/10.1093/jssam/smaa028

Kim JK, Park S, Chen Y, Wu C (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society. Series A. Statistics in Society*, 184(3): 941–963. https://doi.org/10.1111/rssa.12696

Lee BK, Lessler J, Stuart EA (2011). Weight trimming and propensity score weighting. *PLoS ONE*, 6(3): e18174. https://doi.org/10.1371/journal.pone.0018174

Li Y, Fay M, Hunsberger S, Graubard BI (2023). Variable inclusion strategies for effective quota sampling and propensity modeling: An application to sars-cov-2 infection prevalence estimation. *Journal of Survey Statistics and Methodology*, 11(5): 1204–1228. https://doi.org/10.1093/jssam/smad026

Little RJ (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404): 1198–1202.

Lohr SL (2021). *Sampling: Design and Analysis*. Chapman and Hall/CRC.

Maia M, Azevedo AR, Ara A (2021). Predictive comparison between random machines and random forests. *Journal of Data Science*, 19(4): 593–614. https://doi.org/10.6339/21-JDS1025

National Academies of Sciences, Engineering, and Medicine (2018). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. National Academies Press.

Rivers D (2007). Sampling for web surveys. American Statistical Association, Alexandria, VA, 1–26.

Rubin DB (1976). Inference and missing data. *Biometrika*, 63(3): 581–592. Publisher: Oxford University Press. https://doi.org/10.1093/biomet/63.3.581

Särndal CE, Swensson B, Wretman J (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Scott DW (2009). Sturges' rule. *Wiley Interdisciplinary Reviews. Computational Statistics*, 1(3): 303–306. https://doi.org/10.1002/wics.35

Stuart EA (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1): 1. https://doi.org/10.1214/09-STS313

Stuart EA, King G, Imai K, Ho D (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8): 1–28. https://doi.org/10.18637/jss.v042.i08

Sturges HA (1926). The choice of a class interval. *Journal of the American Statistical Association*, 21(153): 65–66. https://doi.org/10.1080/01621459.1926.10502161

Wang L, Graubard BI, Katki HA, Li Y (2020). Improving external validity of epidemiologic cohort analyses: A kernel weighting approach. *Journal of the Royal Statistical Society. Series A. Statistics in Society*, 183(3): 1293–1311.

Wang YH (1993). On the number of successes in independent trials. *Statistica Sinica*, 3(2): 295–312.

Wiśniowski A, Sakshaug JW, Perez Ruiz DA, Blom AG (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8(1): 120–147. https://doi.org/10.1093/jssam/smz051

Wood SN (2017). *Generalized Additive Models: An Introduction with R*. CRC Press.

Yang S, Kim JK (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3: 625–650. https://doi.org/10.1007/s42081-020-00093-w

Yang S, Kim JK, Hwang Y (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, 47(1): 29–58.

Yang S, Kim JK, Song R (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 82(2): 445–465. https://doi.org/10.1111/rssb.12354