On the Use of Auxiliary Variables in Multilevel Regression and Poststratification[*]

Yajuan Si[†]

Oct 28, 2020

Abstract

Multilevel regression and poststratification (MRP) has been a popular approach for selection bias adjustment and subgroup estimation, with successful and widespread applications from social sciences to health sciences. We demonstrate the capability of MRP to handle the methodological and computational issues in data integration and inferences of probability and nonprobability-based surveys, and the broad extensions in practical applications. Our development is motivated by the Adolescent Brain Cognitive Development (ABCD) Study that has collected children across 21 U.S. geographic locations for national representation but is subject to selection bias, a common problem of nonprobability samples. Though treated as the gold standard in public opinion research, MRP is a statistical technique that has assumptions and pitfalls, the validity of which prominently depends on the quality of available auxiliary information. In this paper, we develop the statistical foundation of how to incorporate auxiliary variables under MRP. We build up a systematic framework under MRP for statistical data integration and inferences. Our simulation studies indicate the statistical validity of MRP with a tradeoff between robustness and efficiency and present the improvement over alternative methods. We apply the approach to evaluate cognition performances of diverse groups of children in the ABCD study and find that the adjustment of auxiliary variables has a substantial effect on the inference results.

Key words: data integration; nonprobability sample; robust inference; model-based; design-adjusted

1. Introduction

Nonprobability samples are quickly emerging owing to the rapidly declining response rates and increasing costs of probability samples, and offer detailed outcomes of interest with large sample sizes that are not available in probability surveys. The selection mechanism may be voluntary or deterministic, and the inclusion probabilities are unknown. Our motivating application, the Adolescent Brain Cognitive Development (ABCD) study aims for national representation but is a nonprobability sample (ABCD, 2018). The 21 research sites across the U.S. geography are selected for convenience with operational constraints, and the sample enrollment is conditional on the school and parental consents. The ABCD study design can result in selection bias. The lack of randomization and sampling frames demolishes the inferential framework, leading the validity to rely on the quality of auxiliary information and underlying population model specification (Smith, 1983).

Current approaches for inferences with nonprobability surveys rely on calibration with a reference probability sample or population control information (Elliott and Valliant, 2017). The design-based approach assumes quasi-randomization and constructs pseudo-weights in the combination of the probability and nonprobability samples. Model-based approaches fit a model of the survey outcome based on the sample and predict the outcome for the nonsampled population

---

[†]University of Michigan; email: `yajuan@umich.edu`

units (Ghosh and Meeden, 1997). Doubly robust (DR) methods combine the two approaches and propose a weighted estimator that is also a function of predicted outcomes (Chen et al., 2019; Yang et al., 2020). DR estimators offer protection against the misspecification of the sample inclusion model or the outcome model. Kang and Schafer (2007) find that many DR methods perform better than simple inverse-probability weighting, however, none of the studied has improved upon the performance of simple regression-based prediction of the nonsampled values.

As a prediction approach to generating synthetic populations, multilevel regression and post-stratification (MRP, Gelman and Little (1997)), originally applied to estimate state-level public opinion from sociodemographic subgroups using sample surveys, has become increasingly popular. MRP is now a standard approach for adjusting selection bias and facilitating small area estimation (Pfeffermann, 2013; Rao and Molina, 2015; Ghosh, 2020) to extrapolate sample inferences to the target population. MRP has two key components: 1) multilevel regression for small area estimation by setting up a predictive model with a large number of covariates and regularizing with Bayesian prior specifications; and 2) poststratification to adjust for selection bias and correct for imbalances in the sample composition. The flexible modeling of survey outcomes can capture complex data structures conditional on poststratification cells, which are determined by the cross-tabulation of categorical variables that affect the sample inclusion (selection and response) and use population control information to balance the sample discrepancy (Holt and Smith, 1979; Gelman and Carlin, 2000).

Besides its widespread and successful applications in social sciences, especially in the U.S. and U.K. election forecasting (e.g., Lax and Phillips (2009a,b); Wang et al. (2015); Lauderdale et al. (2020); Zahorski (2020)), MRP also demonstrates its stability for small area estimation of disease prevalence and adjustment for nonresponse bias in public health (e.g., Zhang et al. (2015); Downes et al. (2018); Downes and Carlin (2020)). Notably, MRP has been actively used for the 2020 U.S. presidential election forecasting (The Economist, 2020; Yougov, 2020) and COVID-19 prevalence estimates (Gelman and Carpenter, 2020; Covello et al., 2020).

We aim to apply MRP to correct for the selection bias and generate valid and stabilized estimates of child groups with diverse socio-demographic characteristics in the ABCD study. The large-scale survey is designed to assess the target population of the U.S. 9- and 10-year old children with diverse biological, familial, social and environmental factors (Garavan et al., 2018). Population representativeness is crucial for external validity to generalize the results. Calibrating the ABCD study to eliminate the sample discrepancy, MRP has the potential to yield valid estimates, especially for minority groups with small sample sizes.

However, criticisms on treating MRP "as a widely used gold standard" emerge and argue with substantial variation in the performances (Buttice and Highton, 2013; Valliant, 2019). MRP is a statistical method of adjusting for selection/nonresponse bias and data sparsity to improve survey estimates for overall population and domain inferences, and could be subject to poorly predictive auxiliary information, model misspecification and invalid assumptions. However, there is no guideline but much confusion on the selection and modeling of auxiliary information in practical applications of MRP. This paper fills in the knowledge gap. We develop the statistical foundation on the use of auxiliary information in MRP.

MRP is deeply rooted in survey methodology. Survey inference is essentially a data integration process that calibrates the sample data with auxiliary information to achieve the population representativeness and valid inferences. Accounting for the inclusion mechanism in prediction modeling, MRP combines design-based and model-based approaches for survey inference, similar to the

design-adjusted, model-assisted estimation (Sarndal et al., 1992; Breidt and Opsomer, 2017). As post-adjustments after data collection, MRP can unify inferences for probability and nonprobability samples under a systematic data integration framework. MRP integrates the sample data with the auxiliary or population control information, from census records or large-scale survey data that have low variability, for example, the American Community Survey (ACS) or the Current Population Survey (CPS). The availability and quality of external information and the integrating method affect the inferential validity. Propagating all sources of uncertainty in the synthetic population generation, we generalize MRP as a unified framework for data integration and robust survey inferences with challenging data settings in practice.

The paper structure is organized as below. In Section 2 we develop the theoretical foundation of MRP in the aspects of poststratification and robust survey inferences and propose a systematic framework for data integration. We use simulation studies to illustrate the improvements of the MRP framework in Section 3 and demonstrate the application of inferences with the ABCD nonprobability survey in Section 4. We summarize existing challenges and potential extensions in Section 5.

2. Methodology

Rubin (1983) and Little (1983) point out that any model for survey outcomes should condition on all information that predicts inclusion probabilities. Suppose the outcome in the population is Y_i, the inclusion indicator is I_i, and the auxiliary variables of the population are denoted by X_i, for $i = 1, \ldots, N$, where N is the population size. We consider the inference framework (Smith, 1983)

$$f(Y_i, I_i | X_i) = f(Y_i | I_i, X_i) f(I_i | X_i),$$

the validity of which relies on the inclusion mechanism $f(I_i | X_i)$ and the outcome model $f(Y_i | I_i, X_i)$. When the outcome Y_i is correlated to the inclusion indicator I_i, the inclusion mechanism is informative and has to be accounted for in the analysis of the sampled data $f(Y_i | I_i = 1, X_i)$. Survey practice often assumes that the inclusion mechanism is ignorable, i.e., missing at random (MAR), given the auxiliary information, $f(Y_i | I_i, X_i) = f(Y_i | X_i)$ (Rubin, 1976). However, this demands correct model specification with rich, highly predictive information X_i from integrated data sources. By constructing poststratification cells with discretized auxiliary variables X_i from the target population, MRP is a post-collection adjustment method that can make inferences of both probability and nonprobability samples and account for the design and response mechanisms. In the paper, we also refer to the included samples as respondents.

We develop statistical guidelines on the use of such categorical auxiliary information in MRP to achieve inferential validity and balance estimation bias and variance. We first examine the properties of MRP in the aspects of poststratification and robust survey inferences, and then present the proposed systematic MRP framework for data integration and inferences. As popular quantities of interest in survey inference, we focus on the descriptive summaries of the population: the overall mean and subdomain inferences.

2.1. Poststratification

We start by estimating the population mean of a single survey response: $\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$. Assume the population means in the respondents and nonrespondents are $\bar{Y}_R$ and $\bar{Y}_M$, respectively. The

3

population proportion of respondents is ψ. The overall population $\bar{Y}$ is given by $\bar{Y} = \psi \bar{Y}_R + (1 - \psi)\bar{Y}_M$. Suppose units in the population and the sample can be divided into J poststratification cells with population cell size N_j and sample cell size n_j for each cell $j = 1, \ldots, J$, with the population size $N = \sum_{j=1}^{J} N_j$ and the sample size $n = \sum_{j=1}^{J} n_j$. Let $\bar{Y}_j$ be the population mean and $\bar{y}_j$ be the sample mean within cell j. The overall mean in the population is $\bar{Y} = \sum_{j=1}^{J} \frac{N_j}{N} \bar{Y}_j$. For subdomain estimation, since the poststratification cells are constructed at the finest level, we will need to group the set of cell-wise estimates that belong to the subdomain. Let the poststratification cell means for respondents and nonrespondents be $\bar{Y}_{jR}$ and $\bar{Y}_{jM}$, respectively, and the population cell proportions of respondents be $\bar{\psi}_j$. The population mean can be expressed as

$$\bar{Y} = \sum_{j=1}^{J} \frac{N_j}{N} (\psi_j \bar{Y}_{jR} + (1 - \psi_j)\bar{Y}_{jM}).$$

To make inference about $\bar{Y}$ with the sample data, the unweighted (UnW) estimator is the average of the sample cell means

$$\bar{y}_s = \sum_{j=1}^{J} \frac{n_j}{n} \bar{y}_j. \tag{1}$$

The poststratification (PS) estimator accounts for the population cell sizes as a weighted average of the sample cell means,

$$\bar{y}_{ps} = \sum_{j=1}^{J} \frac{N_j}{N} \bar{y}_j. \tag{2}$$

Motivated by the model-based perspective of poststratification (Fay, 1979; Little, 1993), Gelman and Little (1997) propose MRP by modeling cell estimates and predicting nonresponded cases. Assuming the survey outcome follows a normal distribution with cell-specific mean and variance values,

$$y_{ij} \sim N(\theta_j, \sigma_j^2), \tag{3}$$

the proposed MRP estimator for the population mean can be expressed in weighted form,

$$\tilde{\theta}^{\text{mrp}} = \sum_{j=1}^{J} \frac{N_j}{N} \tilde{\theta}_j, \tag{4}$$

where $\tilde{\theta}_j$ is the model-based estimate of $\bar{Y}_j$ in cell j. MRP fits multilevel models to borrow information and smooth cell-wise estimates $\tilde{\theta}_j$, i.e., partial pooling (Gelman and Hill, 2007).

Under a Bayesian paradigm, given an exchangeable prior distribution, $\theta_j \sim N(\mu, \sigma_\theta^2)$, where the hyperparameters (μ, σ_θ^2) are assigned with noninformative prior distributions, the posterior mean estimate for MRP is

$$\tilde{\theta}^{\text{mrp}} = \sum_{j=1}^{J} \frac{N_j}{N} \frac{\bar{y}_j + \delta_j \frac{\sum_{j=1}^{J}(\bar{y}_j/(1+\delta_j))}{\sum_{j=1}^{J}(1/(1+\delta_j))}}{1 + \delta_j}, \text{ where } \delta_j = \frac{\sigma_j^2}{n_j \sigma_\theta^2}.$$

Given the fitted model, the ratio of sums $\frac{\sum_{j=1}^{J}(\bar{y}_j/(1+\delta_j))}{\sum_{j=1}^{J}(1/(1+\delta_j))}$ is a constant that does not depend on j, and can be approximated by $\sum \bar{y}_j * n_j/n = \bar{y}_s$. This is appropriate if the cell-wise inclusion mechanisms are independent of the group sizes N_j and the variance parameters $(\sigma_j^2, \sigma_\theta^2)$. Hence, the approximated MRP estimator is given by

$$\tilde{\theta}^{\mathrm{mrp}} \approx \sum_{j=1}^{J} \frac{N_j}{N} \frac{\bar{y}_j + \delta_j \bar{y}_s}{1 + \delta_j}, \text{ where } \delta_j = \frac{\sigma_j^2}{n_j \sigma_\theta^2}, \tag{5}$$

as a combined estimator of $\bar{y}_{ps}$ and $\bar{y}_s$ with the shrinkage factor δ_j. The shrinkage factor depends on the sample cell size n_j, within-cell variance σ_j^2 and between-cell variance σ_θ^2, the values of which affect the tendency of $\tilde{\theta}^{\mathrm{mrp}}$ toward the unweighted estimate $\bar{y}_s$, as illustrated by the empirical work in Buttice and Highton (2013). When $n_j \to 0$, $\delta_j \to \infty$, i.e., with small cell sizes, the estimate will be pooled toward the overall unweighted mean $\tilde{\theta}^{\mathrm{mrp}} \to \bar{y}_s$; when $n_j \to \infty$, $\delta_j \to 0$, i.e., with large cell sizes, the estimate will be lead toward the PS estimator with cell-wise direct estimates $\tilde{\theta}^{\mathrm{mrp}} \to \bar{y}_{ps}$.

Next, we examine the bias and variance in the comparison of the UnW, PS and MRP estimators. Assume that the sample respondents are a random sample of the population respondents: $E(\bar{y}_j) = \bar{Y}_{jR}$, similar to Kalton and Kasprzyk (1988), and we can calculate the bias

$$\mathrm{bias}(\bar{y}_s) = \sum_{j=1}^{J} \frac{N_j}{N} \bar{Y}_{jR} \frac{(\psi_j - \bar{\psi})}{\bar{\psi}} + \sum_{j=1}^{J} \frac{N_j}{N} (1 - \psi_j)(\bar{Y}_{jR} - \bar{Y}_{jM}) \doteq A + B$$

$$\mathrm{bias}(\bar{y}_{ps}) = \sum_{j=1}^{J} \frac{N_j}{N} (1 - \psi_j)(\bar{Y}_{jR} - \bar{Y}_{jM}) \doteq B,$$

where $\bar{\psi} = \sum_{j=1}^{J} \frac{N_j}{N} \psi_j$ denotes the overall proportion of respondents in the population. The term A captures the variation between cell-wise response propensities, and the term B is a weighted average of cell-wise covariances between the response propensities and outcomes. The general conditions that the absolute values $|\mathrm{bias}(\bar{y}_{ps})| < |\mathrm{bias}(\bar{y}_s)|$ are if: 1) A and B have the same sign or 2) $|A| > 2|B|$. The biases can be approximated in a stochastic model (Bethlehem, 2002)

$$bias(\bar{y}_s) = \frac{Cov(\psi, y)}{\bar{\psi}}, \qquad bias(\bar{y}_{ps}) = \sum_{j=1}^{J} \frac{N_j}{N} \frac{Cov_j(\psi, y)}{\bar{\psi}_j},$$

where $Cov(\psi, y)$ is the covariance between the response probabilities and the outcome values, $Cov_j(\psi, y)$ is the covariance between the response probabilities and the outcome values within cell j, and $\bar{\psi}_j$ is the mean of response probabilities in cell j. The expression $bias(\bar{y}_s)$ is similar to the data defect index discussed in Meng (2018).

The bias of the MRP estimator is given by

$$\mathrm{bias}(\tilde{\theta}^{\mathrm{mrp}}) = \sum_{j=1}^{J} \frac{N_j}{N} \frac{1}{1 + \delta_j} (1 - \psi_j)(\bar{Y}_{jR} - \bar{Y}_{jM}) + \sum_{j=1}^{J} \frac{\delta_j}{1 + \delta_j} (\sum_{j=1}^{J} \frac{N_j \psi_j}{N \bar{\psi}} \bar{Y}_{jR} - \psi_j \bar{Y}_{jR} - (1 - \psi_j)\bar{Y}_{jM}).$$

If the dataset is a simple random sample, that is, missing completely at random, $\psi_j \equiv \bar{\psi}$, then all three estimators are unbiased. When the cells are homogeneous with respect to either the

response probability or the outcome variable, equivalently, based on the model (3), $\bar{Y}_{jR} = \bar{Y}_{jM}$, $Cov_j(\psi, y) = 0$, and thus, $B = 0$, the PS estimator $\bar{y}_{ps}$ is unbiased. However, with the exchangeable prior distribution, the MRP estimator $\tilde{\theta}^{\mathrm{mrp}}$ will be biased, the second term of which is nonzero, though the bias will be smaller than bias$(\bar{y}_s)$.

Conditional on the sample cell sizes $\vec{n} = (n_1, \ldots, n_J)$, the variance estimates are

$$var(\bar{y}_s|\vec{n}) = \sum_{j=1}^{J} \frac{n_j}{n^2} s_j^2$$

$$var(\bar{y}_{ps}|\vec{n}) = \sum_{j=1}^{J} \frac{N_j^2}{N^2}(1 - n_j/N_j)\frac{s_j^2}{n_j},$$

where s_j^2 is the element variance inside cell j. The conditional variance estimate for $\tilde{\theta}^{\mathrm{mrp}}$ is approximated by

$$var(\tilde{\theta}^{\mathrm{mrp}}|\vec{n}, \vec{\delta}) = \sum \frac{N_j^2}{N^2}[(\frac{1}{1+\delta_j})^2 \frac{s_j^2}{n_j} + (\frac{\delta_j}{1+\delta_j})^2 \frac{n_j}{n^2} s_j^2 + \frac{2\delta_j}{(1+\delta_j)^2} \frac{s_j^2}{n}].$$

With small n_j's, the variance $var(\bar{y}_{ps}|\vec{n})$ could be large, and $var(\tilde{\theta}^{\mathrm{mrp}}|\vec{n}, \vec{\delta})$ reduces the variance with the shrinkage effect. Under a Bayesian paradigm, the variance estimate of $\tilde{\theta}^{\mathrm{mrp}}$ will propagate all sources of uncertainty in the posterior computation.

Holt and Smith (1979) show that poststratification based on cells that are homogeneous with respect to the target variable both reduces variance and bias. Poststratification based on cells that are homogeneous with respect to the response probabilities reduce the bias but not necessarily the variance (Little, 1986). Consistent with Little and Vartivarian (2005), we recommend incorporating auxiliary variables that are predictive of either the survey outcomes (primarily) or response propensities (secondly) in the construction of poststratification cells, and induce prior distributions to shrink the non-predictive terms to be zero. With data-driven shrinkage modeling to find a tradeoff between bias and variance in the comparison with $\bar{y}_s$ and $\bar{y}_{ps}$, we expect that $\tilde{\theta}^{\mathrm{mrp}}$ under a flexible model structure and prior distributions yields the smallest root mean squared error (RMSE) and robust inferences.

This highlights the incorporation of cell-wise covariates to introduce a conditionally exchangeable model for θ_j given the auxiliary variables X_j, for example

$$\theta_j = X_j\beta + \gamma_j, \ \gamma_j \sim N(0, \sigma_\theta^2).$$

Here we use the cell indices j in the notation of auxiliary variables X_j, rather than the unit indices i, because the poststratification cells are the resulting cross-tabulation of X_i. It is appropriate to assume a conditional exchangeable model after incorporating sufficient relevant information in X_j that the cells can be thought of as randomly assigned. A nonexchangeable model would be appropriate if information relevant to the outcome were conveyed in the unit indexes rather than by explanatory variables. With cell-wise covariates X_j, the MRP estimator, which is commonly used in practice, becomes

$$\tilde{\theta}^{\mathrm{mrp}} \approx \sum_{j=1}^{J} \frac{N_j}{N} \frac{\bar{y}_j + \delta_j X_j\beta}{1+\delta_j}, \text{ where } \delta_j = \frac{\sigma_j^2}{n_j\sigma_\theta^2}, \tag{6}$$

6

as a combined estimator of $\bar{y}_{ps}$ and $\sum_{j=1}^{J} \frac{N_j}{N} X_j \beta$. The MRP estimator is a population average of the composite estimators that are known in the small area estimation (Ghosh, 2020).

Lahiri and Mukherjee (2007) have examined the design-consistency property of a hierarchical Bayes estimator of a finite population stratum mean when the sample size is large. Specifically, they prove that with an exponential family distribution for the outcome and a normal prior distribution for θ_j, the posterior mean estimator $\tilde{\theta}_j \to \bar{y}_j$ as $n_j \to \infty$ and $n_j/N_j \to f_j$ for some $0 < f_j < 1$, and the corrected estimator $\tilde{\theta}_j - (\bar{y}_j - \bar{y}_{j,ds})$ is design-consistent, where $\bar{y}_{j,ds}$ is any design-consistent estimator of $\bar{Y}_j$ (Theorem 3.1, Lahiri and Mukherjee (2007)). Extending to MRP, we have the corollary below.

Corollary 1. *Assume the following regularity conditions.*
(R.1) The survey outcome y_{ij} follows an exponential family distribution with cell-specific mean θ_j and variance parameters σ_j^2.
(R.2) The prior distribution of θ_j is $N(\mu_j, \sigma_\theta^2)$.
(R.3) The cell-wise proportion $n_j/N_j \to f_j$ for some $0 < f_j < 1$.
(R.4) The poststratification cell structure fully accounts for the design information.
Then the MRP estimator

$$\tilde{\theta}^{\mathrm{mrp}} = \sum_{j=1}^{J} \frac{N_j}{N} \tilde{\theta}_j \to \sum_{j=1}^{J} \frac{N_j}{N} \bar{y}_j$$

as $n_j \to \infty$, is design-consistent.

Here one of the key regularity conditions is the availability of auxiliary information that constructs the poststratification cells and fully accounts for the design information and nonresponse mechanism. Design-consistency is a desirable property in the randomization approach to finite population sampling. Moreover, MRP improves estimation efficiency and balances bias and variance for robust inferences.

2.2. Robust inferences

Baker et al. (2013) emphasize that inference for any probability or nonprobability survey requires reliance on modeling assumptions, a coherent framework and an accompanying set of measures for evaluating the quality. MRP improves overall population inferences and small area estimation by adjusting for selection/nonresponse bias into hierarchical modeling that is robust against misspecification. The crucial guidelines for generalizable and valid inferences of MRP on the use of auxiliary information include 1) poststratification with auxiliary information, and 2) flexible modeling strategies.

Poststratification with auxiliary information: With a rich set of highly predictive covariates, Wang et al. (2015) have applied MRP to obtain estimates of voting behavior in the 2012 U.S. Presidential election based on a non-representative poll of 350,000 Xbox users and calibration with the auxiliary information from the exit poll data. The cross-tabulation of covariates about voting behavior—sex, race, age, education, state, party ID, political ideology, and reported 2008 vote—results in 176,256 poststratification cells, across which a multilevel logistic regression is fit. Next, the exit poll data with 101,638 respondents from the 2008 presidential election are used as the reference sample for poststratification to adjust for the discrepancies in the sample decompositions of the Xbox data to match the population control information. Their findings show that MRP estimates are in line with the forecasts from leading poll analysts, which were based on aggregating

hundreds of traditional polls conducted during the election cycle, and non-representative samples have the potential for population-based inferences. The success of MRP mostly comes from the adjustment of selection bias to which vote swings are mostly attributed (Gelman et al., 2016).

Current approaches for statistical inferences with nonprobability survey samples assume the relevant auxiliary information is available from a reference probability survey sample (Elliott and Valliant, 2017; Chen et al., 2019; Yang and Kim, 2020; Kim and Tam, 2020). Incorporating auxiliary information as covariates in response propensity modeling, inverse propensity score weighting (IPW) estimators are usually used but are sensitive to misspecification, especially with a wide range of values, and can result in highly variable estimates, for example, discussed by Tan (2007). Flexible predictive models are fit to construct stabilized pseudo-weights for nonprobability samples, such as Bayesian Additive Regression Trees (BART, Rafei et al. (2020)) and kernel weighting approaches (Wang et al., 2020). In contrast to survey weighting, MRP constructs poststratification cells based on the auxiliary information and groups individuals to reduce the variability that is also stabilized by the multilevel outcome modeling.

Flexible modeling strategies: DR estimators improve the efficiency and the robustness of IPW estimators by combining a prediction model for the survey outcome (Bang and Robins, 2005). Kang and Schafer (2007) show that the regression prediction estimator outperforms DR estimators with a predictive model. Moreover, most DR estimators do not apply to subgroup estimation. With a linear regression model, the well-known general regression (GREG) estimator improves estimation accuracy in the combination with poststratification adjustments (Deville and Särndal, 1992). MRP replaces the linear regression in GREG with a multilevel model with smoothing effects across poststratification cells.

Lauderdale et al. (2020) consider three pre-election polling applications and conclude that careful model specification is essential in applying MRP. Flexible models of survey outcomes are essential to yield robust model-based inferences. The election forecasting model applied to the Xbox survey is a Bayesian hierarchical model with only main effects. The existing examples of flexible models under MRP include hierarchical models with high-order interactions and global-local shrinkage prior specifications (Ghitza and Gelman, 2013; Si et al., 2020), Gaussian process (GP) regression models (Si et al., 2015), BART (Bisbee, 2019), and stacked regression (Ornstein, 2020). The flexible models under MRP improve small area estimation and facilitate subdomain and overall population inferences. In addition to a careful selection of predictive auxiliary variables and their high-order interaction terms, we recommend incorporation a flexible function form of inclusion probabilities into the outcome modeling. Next we develop a systematic data integration framework under MRP.

2.3. Systematic data integration

Even with numerous successful applications, MRP is faced with challenges in practical settings. Most MRP applications integrate the probability/nonprobability sample of interest with large probability samples (e.g., ACS, CPS, or census records) and ignore the sampling uncertainty of the latter. That is, MRP treats the population cell counts N_j's from external data as fixed. In practice, the reference probability sample may be small, shown in Figure 1. We have to estimate population information based on the reference probability sample and then use the estimates for calibration or imputation of outcome values for the nonsampled units. Often the joint population distributions of all auxiliary information are incomplete, e.g., with the joint distribution of only a subset of variables or only marginal distributions being available.

With a large number of auxiliary variables, the number of poststratification cells J could be

large, for example, $J = 2 \times 4 \times 4 \times 5 \times 50 = 8000$, in the adjustment of sex (2 levels), race/ethnicity (4 levels), education (4 levels), age (5 levels) and state (50 levels). The resulting sample poststratification cells could be sparse and even empty, as the computational bottleneck for MRP. Little (1993) recommends collapsing small cells to reduce the variance, with a payoff of increased bias. Either 1) when the inclusion probabilities of the individuals are the same within cells; or 2) when the included individuals are similar to those excluded within cells, i.e., conditional MAR inside cells, the poststratification and MRP estimators will be unbiased. Combing cells may violate such assumptions, however, the data sparsity requires shrinkage to stabilize inferences. The Bayesian paradigm of MRP allows the data to determine the pooling effects.
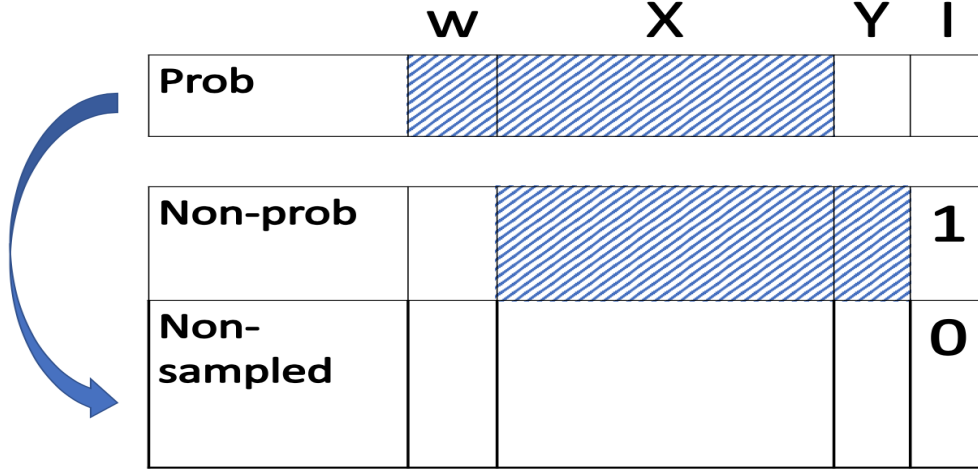

Figure 1: *Illustrative systematic data integration framework of combing a small probability (prob) sample and a nonprobability (Non-prob) sample, where they share a common set of covariates X and the survey outcome Y is only collected in Non-prob samples. We create an inclusion indicator I for the nonprobability samples. The probability sample will be used to estimate the population information for the nonsampled units.*

We would like to generate synthetic population information if unknown and fit MRP with a flexbile model across cells and Bayesian shrinkage prior distributions, as a systematic data integration framework in Figure 1 on the use of auxiliary information. The general MRP estimator for the population mean would be

$$\tilde{\theta}^{\mathrm{mrp}} = \sum_{j=1}^{J} \frac{\hat{N}_j}{\sum \hat{N}_j} \tilde{\theta}_j, \tag{7}$$

where both $\hat{N}_j$ and $\tilde{\theta}_j$ are model-based estimates.

The MRP under the systematic data integration framework combines the quasi-randomization approach and the superpopulation modeling approach by constructing three main models.

1. The cell sizes: assume that the respondents within poststratification cell j are independently and quasi-randomly sampled from the population cell cases,

$$(n_1, \ldots, n_J | \vec{\psi}) \sim \mathrm{Multinomial}((cN_1\psi_1, \ldots, cN_J\psi_J), n), \tag{8}$$

where c is a normalizing constant, $c = 1/\sum_{j=1}^{J} N_j \psi_j$. This can be approximated by J Poisson distributions when J is sufficiently large or the inclusion probability ψ_j is sufficiently small.

Here the population cell counts N_j's are estimated from the reference sample. Often the relevant auxiliary information is partially complete in the reference sample and requires modeling to generate the synthetic population distribution (Reilly et al., 2001; Kastellec et al., 2015). Ghitza and Gelman (2020) turn to large-scale voter registration databases to accommodate the census or exit polls for a more accurate picture of the target population with powerful and stable political covariates. A variety of nonparametric and parametric approaches are available to estimate N_j. With survey weights available from the reference probability sample, weighted Bayesian bootstrap approaches have been developed to obtain the population distribution of X_j (Zangeneh and Little, 2012; Dong et al., 2014; Zhou et al., 2016; Makela et al., 2018). Si et al. (2015) have used survey weights in probability samples by assuming $\psi_j = 1/w_j$, where w_j is the unique weight value in cell j and model (8) to yield smoothed estimates of N_j's with sparse cells even only containing one unit.

Leemann and Wasserfallen (2017) combine MRP with raking when only population margins of the auxiliary variables are known. The raking model assumes that the auxiliary variables have independent effects on the inclusion, equivalent to the prediction model of inclusion probabilities below but with only the main effects of auxiliary models (Little and Wu, 1991; Si and Zhou, 2020).

2. The inclusion probabilities: we can assume that they are concentrated around the average inclusion probability of the sample, for example, a Beta distribution with a mean of n/N. Alternatively, with covariates Z_j that affect inclusion propensities and a link function $g(\cdot)$ (e.g., *logit*), we model the inclusion probabilities,

$$g(\psi_j) = Z_j \gamma, \tag{9}$$

where Z_j's denote cell-wise covariates and γ represents the model parameters.

3. The outcome: within poststratification cells the units are included with the same probability and independently distributed. We assume that the outcome depends on the cell inclusion probabilities and follows a normal distribution with cell-specific mean and variance values,

$$y_{ij} \sim N(f(\psi_j) + X_j\beta, \sigma_j^2). \tag{10}$$

where $f(\psi_j)$ is a function of ψ_j, and X_j denotes the cell-wise covariates that predict the outcome variable, may include both main effects and high-order interaction terms, and overlap with Z_j. For notational simplicity, we ignore individual-level covariates, which can be predictive, even continuous covariates, and added in the outcome model, though used in the poststratification.

Flexible prior distributions can be introduced on $f(\psi_j)$, such as penalized spline functions and nonparametric Bayesian distributions, to account for potential dependency structure and smoothing effects across cells. Si et al. (2015) induce a GP prior distribution to the mean function, such as $f(\psi_j) \sim GP(\log(\psi_j)\beta, \Sigma(\vec{\psi}))$, where the mean $\log(\psi_j)\beta$ is a linear function of the logarithm of the inclusion probability, and the covariance matrix $\Sigma(\vec{\psi})$ depends on the differences between ψ_j's, e.g., $Cov(\psi_j, \psi_{j'}) = (1 - g)\tau^2 exp(-\frac{(\psi_j - \psi_{j'})^2}{l^2}) + g\tau^2 I_{j=j'}$ with weakly informative or noninformative prior specifications for hyperparameters (g, τ, l). The smoothing function $f(\psi_j)$ can be pre-specified or approximate with basis expansion functions if the number of cells is large. With rich information X_j and correct specification of the mean structure $X_j\beta$, the role of the function $f(\psi_j)$ could be minimal. We recommend to include the flexible specification of $f(\psi_j)$ that offers protection against model misspecification, similar to the doubly robust penalized spline of propensity prediction methods (Little and An, 2004; Breidt et al., 2005; Zhang and Little, 2009).

The computation for the models specified in (8), (9), and (10) can be implemented in a fully Bayesian procedure that integrates all sources of uncertainty. Stan makes the computation of MRP generally accessible. As a state-of-the-art platform for statistical modeling and high-performance statistical computation, Stan can perform fully Bayesian analyses with Markov chain Monte Carlo (MCMC) computation. The implementation of MRP is straightforward with the publicly available R packages such as *Rstan* (Stan Development Team, 2020), and *Rstanarm* (Goodrich and Gabry, 2020). For example, in *Rstanarm*, the function *stan_glmer* fits a multilevel model and the function *posterior_predict* imputes the outcome for all nonsampled units in the population. The imputation step draws posterior predictive samples and takes into account all sources of uncertainty, which generates multiple synthetic populations and is implicitly poststratification. We can obtain subgroup estimates by extracting the imputed values for the corresponding domains. We implement the proposed systematic data integration framework in Stan, which has not been available in the public R packages yet.

3. Simulation studies

We conduct simulations to first evaluate the bias and variance trade-off of MRP with varying associations of the poststratification cell structure with the survey outcome and inclusion mechanism. Then we implement the systematic data integration framework with cell-wise covariates and illustrate the improvement of MRP on robust inferences. We compare the UnW estimator $\bar{y}_s$, the PS estimator $\bar{y}_{ps}$ and the MRP estimator $\tilde{\theta}^{\mathrm{mrp}}$ in inferences for the overall finite population and subdomain mean estimates.

3.1. Associations of cell structure with outcome and inclusion

To evaluate the randomness properties, we perform repeated sampling of 100 samples to assess the bias, standard error (SE), RMSE and nominal coverage rates (CR) of 95% confidence intervals. We simulate a population of 1000000 units with $J = 10$ cells of equal sizes. We are interested in the overall population and five subgroup mean estimates, where the gth subgroup composites $g + 1$ number of cells randomly selected from the total 10 cells, for $g = 1, \ldots, 5$, the goal of which is to generate subgroups with various sample sizes. These six quantities of interest are denoted as *All, G1, G2, G3, G4, G5* in the outputs shown in Figure 2. For each repetition, we randomly draw $n = 400$ samples based on the pre-specified outcome modeling and inclusion mechanisms in Table 1. For the MRP specification we assume $\theta_j = \beta_0 + \beta_1 j$ with $\beta_1 \sim Cauchy(0, 3)$, a Cauchy prior distribution, and assign noninformative prior distributions to other parameters. This facilitates computation performances under repeated sampling, and can be replaced by the flexible outcome model in (10).

Table 1: Outcome models and inclusion mechanisms in the simulation studies.

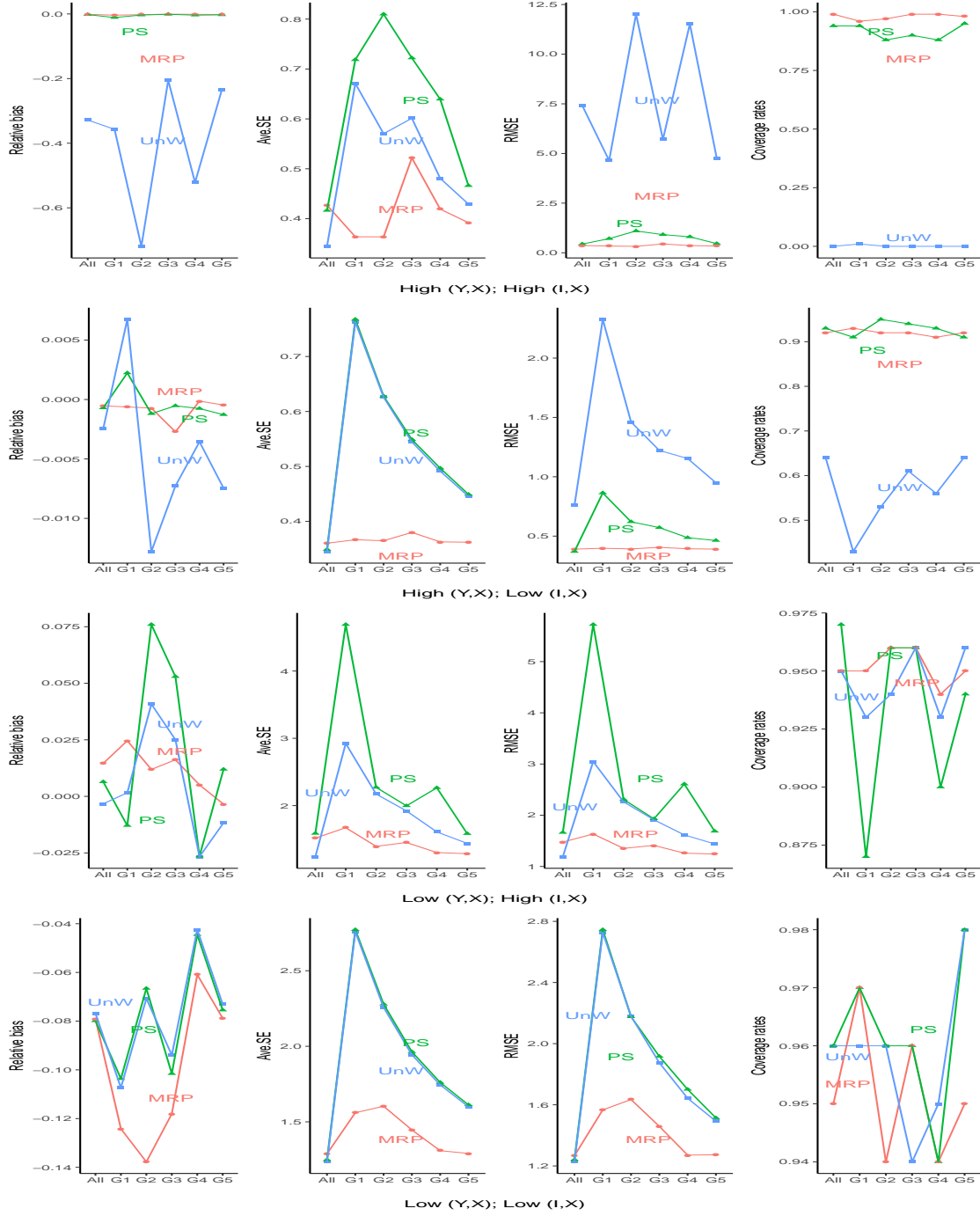| Association	Outcome	Inclusion
High (Y, X); high (I, X)	$Y \sim N(-5 + 5j, 7^2)$	$\psi_j = 0.1 + 0.8(j - 1)/9$
High (Y, X); low (I, X)	$Y \sim N(-5 + 5j, 7^2)$	$\psi_j = 0.5$
Low (Y, X); high (I, X)	$Y \sim N(-5, 25^2)$	$\psi_j = 0.1 + 0.8(j - 1)/9$
Low (Y, X); low (I, X)	$Y \sim N(-5, 25^2)$	$\psi_j = 0.5$

Figure 2: *Simulation outputs comparing the unweighted (UnW) estimator, the poststratification (PS) estimator and the MRP estimator on finite population and subdomain inferences. The panels from top to bottom represent four cases varying the strength of associations (Y, X) and (I, X) from high to low.*

We consider four different simulation scenarios by varying the strength of association of the auxiliary variables with inclusion probabilities (I, X) and outcome (Y, X) from high to low. The four cases are described in Table 1. Generally, across all cases, MRP is the most accurate approach with the smallest SEs and RMSEs, especially for the subgroups with small sample sizes. The top two panels in Figure 2 show that when the association between the outcome and the auxiliary variables used for the poststratification cell construction is high, the unweighted estimator is biased and has low coverage rates, and the MRP and PS estimators yield valid inferences. The findings hold the same when we change both of the associations to be medium.

When the outcome is weakly related to the cell structure, as the lower two panels in Figure 2, the PS estimator has large variability and potential low coverage, while the MRP estimator yields competitive coverage rates with small SEs. When both associations are low, among three estimators, MRP yields the smallest SE and RMSE with comparable bias values. Interestingly, the UnW and PS estimators have similar performances when the cell structure is weakly related with both the outcome and inclusion mechanism. This may be due to substantial group sizes. In all, with a weakly informative prior distribution, the MRP estimator outperforms the unweighted and poststratification estimators with valid inferences and a bias-variance tradeoff, especially for small subgroups.

3.2. Systematic data integration with cell-wise covariates

We use the ACS 2011-2015 data as auxiliary covariates and simulate the outcome variable to conduct the simulation study. Consider three ACS variables: sex (2 levels: female, male), race/ethnicity (race, 8 levels: white, black, Hispanic, Asian, AIAN, NHPI, Other, Multiple), and family income (inc, 6 levels: $< \$25K$, $\$25K - 49K$, $\$50K - 74K$, $\$75K - 99K$, $\$100K - 199K$, $\$200K+$). We simulate the outcome $Y_i = \beta_0 + \beta_{sex}sex_i + \beta_{race}race_i + \beta_{inc}inc_i + \epsilon_i$, $\epsilon_i \sim N(0, 2^2)$, and the inclusion indicator $I_i \sim \text{Bernoulli}(p_i)$ with $p_i = logit^{-1}(\alpha_0 + \alpha_{sex}sex_i + \alpha_{race}race_i + \alpha_{inc}inc_i)$, where the coefficients β_{var}'s and α_{var}'s are vectors if the covariate var has more than 2 levels, β_{var}'s are random draws between 1 and 10, and α_{var}'s are randomly selected from $\{-1, 0, 1\}$, $var \in \{sex, race, inc\}$. The simulated outcome depends on main effects of the three covariates, and the inclusion mechanism depends on only partial main effects, fewer than those in the outcome model. The resulting inclusion probabilities range from 0.27 to 0.98 with a mean value of 0.75.

For inference of each simulated sample, we implement the systematic data integration framework given by models (8), (9), and (10). We assume $f(\psi_j) \sim GP(0, \Sigma(\vec{\psi}))$, where the covariance matrix $\Sigma(\vec{\psi})$ has a pre-specified kernel function $Cov(\psi_j, \psi_{j'}) = (1-g)exp(-(\psi_j - \psi_{j'})^2) + gI_{j=j'}$ and a hyperparameter $g \in (0,1)$ that is given a noninformative prior distribution. The hierarchical prior specification of the coefficients is: $\beta_{var} \sim N(0, \sigma^2_{y:var})$, $\sigma_{y:var} \sim N^+(0, 3^2)$, $\alpha_{var} \sim N(0, \sigma^2_{i:var})$, $\sigma_{i:var} \sim N^+(0, 1^2)$, where $N^+(\cdot)$ denotes a half-normal distribution restricted to the positive values.

We perform repeated sampling 100 times and set the average sample size as 10000. Even with this large sample size, not all 96 cells will be available in the sample. MRP predicts cell-wise estimates for all the population cells. As inferences with large subgroups, Table 2 presents the results of the systematic MRP integration framework for the overall population and five subgroup mean estimates, where the subgroups compose different numbers of randomly selected cells with varying sizes. Comparing the three estimators, MRP and poststratification generate unbiased and valid estimates, while the unweighted estimators are biased with low nominal coverage rates. Since the outcome depends on all three variables, the PS estimator is appropriate, and the MRP framework induces a flexible model that captures the true structure with slightly increased variability and

Table 2: Inference results of the systematic MRP integration framework.

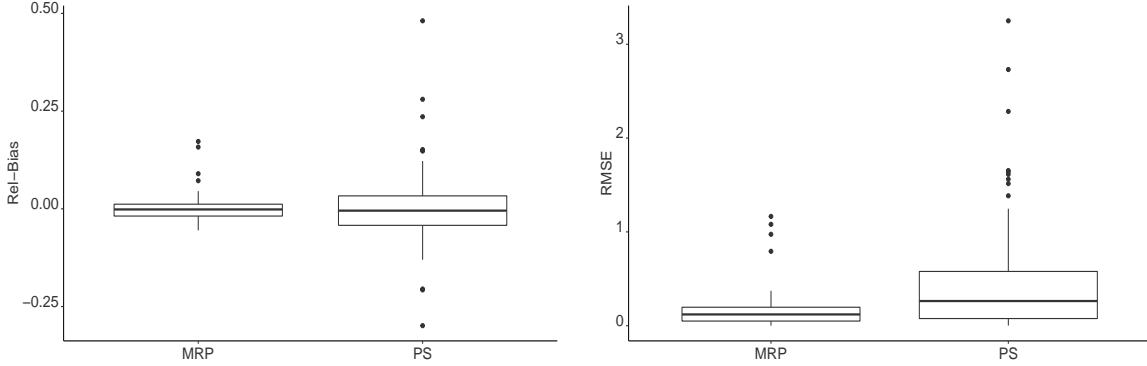|	MRP				UnW				PS			
	Bias	SE	RMSE	CR	Bias	SE	RMSE	CR	Bias	SE	RMSE	CR
G1	-0.00	0.06	0.04	1.00	-0.02	0.03	0.48	0.00	-0.00	0.03	0.03	0.98
G2	-0.00	0.06	0.03	1.00	-0.06	0.03	1.27	0.00	-0.00	0.04	0.04	0.98
G3	-0.00	0.07	0.06	0.98	-0.05	0.04	1.04	0.00	-0.00	0.05	0.04	0.97
G4	0.00	0.06	0.04	0.98	-0.05	0.04	0.90	0.00	0.00	0.05	0.05	0.96
G5	0.00	0.06	0.04	0.98	-0.00	0.04	0.14	0.38	0.00	0.05	0.04	0.94
All	0.00	0.05	0.03	1.00	-0.04	0.02	0.78	0.00	-0.00	0.02	0.02	0.95

Figure 3: *Simulation outputs comparing the poststratification (PS) estimator and the MRP estimator on small domain estimates when the outcome depends on main effects of sex and race/ethnicity.*

conservative confidence coverages, mainly due to the nonparametric Bayesian GP prior specification.

In the assessment of the cell-wise mean estimates, we randomly draw one sample, and the selected sample covers 80 cells, 11 of which only have 1 unit, and 23 cells have fewer than 5 cases. MRP generates smaller bias and RMSE values than the PS estimator. Figure 4 in the Appendix summarizes the relative bias and RMSE of the sampled 80 cell mean estimates.

To evaluate the performances when the auxiliary information becomes redundant with a sparse cell structure, we consider another simulation scenario with the ACS auxiliary variables. Suppose that the outcome only depends on sex and race: $Y_i = \beta_0 + \beta_{sex}sex_i + \beta_{race}race_i + \epsilon_i$, $\epsilon_i \sim N(0, 2^2)$, and the inclusion indicator depends on three variables with probabilities $p_i = logit^{-1}(\alpha_0 + \alpha_{sex}sex_i + \alpha_{race}race_i + \alpha_{inc}inc_i)$, where the coefficients β_{var}'s are random draws between 1 and 10, and α_{var}'s are randomly selected from the range $(-2.5, 1.5)$ with a break of 0.5. The resulting inclusion probabilities are generally low, ranging from 0 to 0.38 with a mean value of 0.04. The randomly drawn sample has 9939 cases and 72 cells, 19 of which have fewer than 5 units. We implement MRP with the same setup as described above. Figure 3 depicts the output for the 72 cell mean estimates under this sparse setting. MRP has substantially smaller biases and RMSEs than the PS estimator. This provides further supporting evidence that MRP outperforms the PS estimator and improves small area estimation with efficiency and accuracy, especially with a sparse data structure.

Table 3: Sociodemographic distribution (%) comparison between the ABCD baseline cohort (March 2019, $n = 11875$) and the ACS (2011–2015, $N = 8211605$, adjusted by the ACS weights).

	ABCD	ACS
Age		
9	56.6	49.6
10	43.4	50.4
Sex		
Male	52.1	51.2
Female	47.9	48.8
Race/Ethnicity		
White	52.1	52.4
Black	15.0	13.4
Hispanic	20.3	24.0
Asian	2.1	5.9
Other	10.5	4.2
Family Type		
Married couple	73.1	66.1
Other	26.9	33.9
Household Size		
2-3 Persons	17.5	18.5
4 Persons	33.7	33.5
5 Persons	25.2	25.4
6 Persons	14.3	12.5
7 or more Persons	9.3	10.1
Household Income		
$< 25K$	16.0	21.5
$25K–$49K	15.2	21.7
$50K–$74K	13.7	17.0
$75K–$99K	14.2	12.5
$100K–$199K	29.9	20.5
$200K+	11.0	6.8
Family and Labor Force (LF) Status		
Married, both in LF	49.9	40.8
Married, 0/1 in LF	23.2	25.6
Single parent, in LF	21.2	26.5
Single parent, not in LF	5.7	7.0

4. Application

We apply MRP to make inferences of our motivating nonprobability survey, the ABCD study. The ABCD study is a prospective cohort study and has collected social, health, imaging and genetics measures of 11,875 children aged 9-10 for environmental exposure, neuroimaging and substance use analysis from 21 U.S. research sites between 2016 and 2018. The ABCD sampling and recruitment process is designated to yield an overall sample that closely approximates national sociodemographics of the targeted U.S. children aged 9–10 (Garavan et al., 2018). However, the 21 research sites are not randomly chosen based on reasons of convenience such as neuroimaging resource allocation and accessibility, a form of nonprobability-based selection, and the child enrollment is conditional on the school and parental consents, resulting in potential selection and nonresponse bias.

Comparing with the ACS 2011-2015 data, Table 3 shows that the ABCD study oversamples 9-year old children, males, high-income families, and certain race/ethnicity groups, with slightly more representation of children from families with married couples and employment. By design, the race/ethnicity composition for major classes (White, Black, Hispanic) matches the ACS targets fairly closely with children of Asian ancestry being underrepresented and children with self-reported other (AIAN, NHPI, Multiple) race /ethnicity being overrepresented relative to the U.S. population of 9- and 10-year olds. This illustrates sociodemographic discrepancies from the U.S. population and potential selection bias analyzing data (Compton et al., 2019).

Heeringa and Berglund (2019) have constructed weights to match the ABCD sample to the ACS population control information by predicting pseudo-probabilities of sample inclusion and performing raking adjustments. Specifically, a multiple logistic regression model was fit to the concatenated

ACS and ABCD data with covariates: age, sex, race/ethnicity, family income, family type, household size, patients' labor force status, and census region, and predict the pseudo-probabilities of inclusion in the ABCD sample, the inverse of which are treated as the initial weights. The initial weights are trimmed at the 2% and 98% quantiles of the distribution and calibrated with raking adjustments of age, sex and race/ethnicity, whose marginal distributions are matched to the ACS. However, the final weights are still widely spread and can result in unstable variances.

In contrast to weighting, we apply MRP to adjust for the sample discrepancies in the estimation of average cognition test scores for the overall U.S. population of 9/10-year olds and diverse sociodemographic groups of interest. The score is a total composite score of cognition based on the NIH Toolbox (ABCD, 2019). We use seven auxiliary variables in Table 3 to construct poststratification cells. The cross-tabulation results in 4800 ($= 2*2*5*2*5*6*4$) cells, 1517 of which are available in the ABCD data and 3128 cells are available in the ACS data. There are 3 cells in the ABCD that are not available in the ACS. In the ABCD data, 962 cells have fewer than or equal to 5 units, and 333 cells only include 1 unit.

We integrate these two datasets by fitting a model to the ABCD and then predicting the potential outcomes of the ACS dataset (in the adjustment of the ACS weights), similar to Figure 1, but with a large probability sample of the ACS. The outcome model incorporates the main effects of seven auxiliary variables

$$y_{ij} = \alpha_0 + \alpha_{age}I(age_{j[i]} = 10) + \alpha_{fem}fem_{j[i]} + \alpha_{mar}mar_{j[i]} + \alpha_{j[i]}^{race} + \alpha_{j[i]}^{inc} + \alpha_{j[i]}^{hhsize} + \alpha_{j[i]}^{lf} + \epsilon_i.$$

Here we use dummy indicators for age (9, 10), sex (fem: female, male), and family type (mar: married, other), and multiple terms indicating levels for race/ethnicity, family income (inc), household size (hhsize), and family labor force status (lf), which are assigned with hierarchical prior distributions: $\alpha_j^{var} \sim N(0, \sigma_{var}^2)$, $\sigma_{var} \sim N^+(0, 1^2)$, $var \in \{race, inc, hhsize, lf\}$. Assume the error term inside cells follows a Cauchy distribution $\epsilon_i \sim$ Cauchy$(0, 1)$. Our analysis results are not sensitive to the hyperparameter specification and generate the same findings when we increase the values (e.g., $N^+(0, 3^2)$, Cauchy$(0, 3)$).

We perform the MCMC computation in Stan with two chains of 2000 iterations each. The diagnostic measures $\hat{R}$ have values around 1 and indicate convergence. The covariates age, sex, race/ethnicity, family income, household size and labor force status are significant predictors of the test score, while the marital status is not significant ($\hat{\alpha}_{married} = -0.85$, with 95% CI (-1.96, 0.67)). The 10-year olds tend to have a higher score than the 9-year olds ($\hat{\alpha}_{age} = 4.61$, with 95% CI (4.54, 4.67)), and the girls have higher scores than boys ($\hat{\alpha}_{fem} = 0.96$, with 95% CI (0.91, 1.01)). The posterior mean estimates of the variance parameters are significantly different from 0, $\hat{\sigma}_{race} = 2.27$ (95% CI: 1.57, 3.33), $\hat{\sigma}_{inc} = 2.16$ (95% CI: 1.50, 3.17), $\hat{\sigma}_{hhsize} = 0.78$ (95% CI: 0.45, 1.87), and $\hat{\sigma}_{lf} = 0.69$ (95% CI: 0.31, 1.80), showing the covariates are predictive for the cognition outcome.

Since these covariates have different distributions between ABCD and ACS, we expect that the poststratification adjustment to the ACS will generate different results from the unadjusted ABCD sample analysis. Table 4 presents the mean estimates of cognition test scores and 95% CIs for the overall population and four sociodemographic subgroups of U.S. 9/10-year-old children. We can predict the outcome values for the 3128 cells in the ACS and also all the 4800 cells, and we have the population counts for the ACS cells. For the overall mean score estimates, the MRP estimates based on the 3128 cells are 86 (95% CI: 85.94, 86.07), which are different from the MRP estimate based on the 1517 cells: 85.32 (95% CI: 85.23, 85.4). For comparison, we present the MRP estimates based on the cells that are also available in the ABCD in Table 4.

Table 4: Finite population inferences of average cognition test scores by groups (95% confidence intervals in parenthesis).

	MRP	Weighted	Sample
Overall (n=11875)	85.32 (85.23, 85.4)	85.73 (85.55, 85.92)	86.2 (86.03, 86.36)
Girls in married families (n=479)	86.94 (86.84, 87.03)	87.23 (86.93, 87.53)	87.8 (87.54, 88.06)
Black (n=1782)	79.46 (79.3, 79.63)	78.65 (78.23, 79.07)	79.05 (78.63, 79.47)
Low-income ($< 25K$), white (n=413)	83.44 (83.25, 83.62)	85.03 (84.4, 85.66)	85.02 (84.41, 85.62)
Married, both in LF, HH size > 3 (n=411)	86.44 (86.34, 86.57)	87.67 (87.42, 87.93)	87.92 (87.7, 88.15)

We also apply the pseudo-weights of the ABCD baseline survey and compute the weighted estimates. Based on Table 4, for the overall mean score estimates, the MRP estimate (85.32) is different from the weighted estimate (85.73), and both are significantly lower than the sample estimate without adjustments (86.2). MRP generates lower scores for girls from married families, low-income white children, and those whose parents are married and employed with larger households, but slightly higher scores for black children, compared to both the weighted and sample estimates. The variance of MRP estimators is lower than that of the weighted and unweighted sample estimators, an illustration of efficiency gains.

We have shown that cognitive performances vary across child groups with diverse sociodemographic and familial characteristics, and the adjustment of auxiliary variables with the ACS study yields substantially different results. MRP uses the highly predictive auxiliary information of the cognitive assessments and adjusts for the differential response propensities. The MRP estimates compose model-based predictions weighted by population cell counts and have efficiency gains. The weighted estimates use the raw values that are subject to measurement error. The poststratification of MRP borrows the joint distribution of the auxiliary variables from ACS, however, the weighting adjustment only uses their marginal distributions. It is possible that the joint distributions of the auxiliary variables are substantially different between the population and the sample, and the high-order interaction terms affect the sample inclusion propensities, leading to different estimates. Here, we incorporate auxiliary information into MRP and stress the potentially substantial impacts on inferences from the sample non-representation. The external validity assessment of different findings requires additional information and substantive knowledge, which will be discussed below.

5. Discussion

Large-scale nonprobability surveys quickly emerge and demand qualified auxiliary information and robust statistical adjustments to achieve representative inferences. Integrating nonprobability samples with probability surveys, MRP adjusts for selection/nonresponse bias and data sparsity to improve survey estimates for population inferences, especially small area estimation. MRP constructs poststratification cells, fits flexible hierarchical models, and pools cell estimates weighted by the population cell counts as weights. The inference validity relies on the poststratification structure and model specification. The key to success is the availability of highly predictive auxiliary variables.

MRP assumes quasi-randomization given the auxiliary variables, where the inclusion probabilities are treated as equal inside cells. MRP fits a superpopulation model to predict outcome values of nonsampled units. This article develops guidelines on the use of auxiliary information in MRP and a systematic framework for data integration and inference. The systematic framework propagates all sources of uncertainty, including the estimation of unknown population control information. We have demonstrated that MRP can achieve a balance between robustness and efficiency.

This comprehensive study of MRP opens up several interesting future extensions. First, a sensitivity analysis framework will offer insights into potential bias if the sample inclusion is non-ignorable, still correlated with the survey outcomes conditional on the available auxiliary information. Little et al. (2019) have proposed a measure of the degree of departure from ignorable sample selection. Depending on the magnitude of the correlation structure, a sensitivity index can be developed to reflect the range of potential bias values. Second, integrating multiple datasets, beyond two samples, can supplement the list of predictive auxiliary variables that can be sequentially imputed, and also help access the inferential validity. Data integration techniques learn about the common target population structure while accounting for the heterogeneity of different studies. Existing organic databases and administrative records could provide unprecedented information. Linking auxiliary information across multiple data sources has become a research priority for most statistical agencies. Third, MRP focuses on the modeling of a single survey outcome and can extend to multivariate outcomes. The selection of auxiliary variables will become the union of those predictive for any outcome. Weighting cells can be constructed based on predictive mean matching and propensity score subgrouping, to achieve double robustness. Fourth, MRP implicitly generates multiple synthetic populations according to the specified model and brings in methodological and practical challenges in model evaluation. The synthetic population prediction can utilize state-of-the-art machine learning and deep learning algorithms to improve accuracy. The evaluation should include goodness of fit in the sample data and achieved representation of the target population. Model evaluation in terms of generalizability is challenging in practical application studies (Keiding and Louis, 2016). Election forecasting can be checked with actual results, however, a gold standard does not exist in most areas. This is an area that needs further developments and collaborations across multidisciplines.

Qualified research emphasizes both sound design and analysis approaches. While soliciting the predictive auxiliary information for analysis, improving the data collection process is essential. MRP is a useful remedy after data collection. Building an integrated database combining multiple data sources to provide big data for a systematic design and analysis process and promote open science, is another promising direction.

In this application, we focus on the cognition assessment at baseline in the comparison of socio-demographic groups in the ABCD study and have demonstrated the use of auxiliary information in MRP to make representative inferences. The ABCD study is a longitudinal study and advocates "population neuroscience" for child development studies. Our future work also includes nonresponse bias adjustment due to attrition and population-based inferences with the brain and neurological outcome measures and methodology developments combining MRP with image modeling.

References

ABCD (2018). The Adolescent Brain Cognitive Development Study. https://abcdstudy.org.

ABCD (2019). The ABCD data repository. https://nda.nih.gov/abcd.

Baker, R., J. M. Brick, N. A. Bates, M. Battaglia, M. P. Couper, J. A. Dever, K. J. Gile, and R. Tourangeau (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology 1*(2), 90–143.

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*, 962–972.

Bethlehem, J. G. (2002). Weighting nonresponse adjustments based on auxiliary information. In R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little (Eds.), *Survey Nonresponse*. Wiley.

Bisbee, J. (2019). BARP: Improving Mister P using Bayesian Additive Regression Trees. *American Political Science Review 113*(4), 1060–1065.

Breidt, F. and J. Opsomer (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science 32*, 190–205.

Breidt, F. J., G. Claeskens, and J. D. Opsomer (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika 92*, 831–846.

Buttice, M. K. and B. Highton (2013). How does multilevel regression and poststratification perform with conventional national surveys? *Political Analysis 21*(4), 449–467.

Chen, Y., P. Li, and C. Wu (2019). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 1–11.

Compton, W. M., G. J. Dowling, and H. Garavan (2019). Ensuring the best use of data: The Adolescent Brain Cognitive Development Study. *JAMA Pediatrics 173*(9), 809–810.

Covello, L., A. Gelman, Y. Si, and S. Wang (2020). A design for hospital-based coronavirus tracking. Technical Report.

Deville, J. C. and C. E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association 87*, 376–382.

Dong, Q., M. R. Elliott, and T. E. Raghunathan (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey Methodology 40*(1), 29–46.

Downes, M. and J. Carlin (2020). Multilevel regression and poststratification versus survey sample weighting for estimating population quantities in large population health studies. *American Journal of Epidemiology 189*(7), 717–725.

Downes, M., L. C. Gurrin, D. R. English, J. Pirkis, D. Currier, M. J. Spittal, and J. B. Carlin (2018). Multilevel regression and poststratification: A modeling approach to estimating population quantities from highly selected survey samples. *American Journal of Epidemiology 187*(8), 1780–1790.

Elliott, M. R. and R. Valliant (2017). Inference for nonprobability samples. *Statistical Science 32*(2), 249–264.

Fay, R. E. R. A. H. (1979). Estimates of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association 366*, 269–277.

Garavan, H., H. Bartsch, K. Conway, A. Decastro, R. Goldstein, S. Heeringa, T. Jernigan, A. Potter, W. Thompson, and D. Zahs (2018). Recruiting the ABCD sample: Design considerations and procedures. *Developmental Cognitive Neuroscience 32*, 16–22.

Gelman, A. and J. B. Carlin (2000). Poststratification and weighting adjustments. In *Survey Nonresponse*. Wiley.

Gelman, A. and B. Carpenter (2020). Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society Series C (Applied Statistics) forthcoming*.

Gelman, A., S. Goel, D. Rivers, and D. Rothschild (2016). The mythical swing voter. *Quarterly Journal of Political Science 11*(1), 103–130.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A. and T. C. Little (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology 23*, 127–135.

Ghitza, Y. and A. Gelman (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science 57*(3), 762–776.

Ghitza, Y. and A. Gelman (2020). Voter registration databases and MRP: Toward the use of large-scale databases in public opinion research. *Political Analysis forthcoming*.

Ghosh, M. (2020). Small area estimation: its evolution in five decades. *Statistics in Transition 21*(4), 1–22.

Ghosh, M. and G. Meeden (1997). *Bayesian Methods for finite population sampling*. Chapman Hall / CRC Press.

Goodrich, B. and J. S. Gabry (2020). rstanarm: Bayesian applied regression modeling via Stan. https://cran.r-project.org/web/packages/rstanarm/.

Heeringa, S. G. and P. A. Berglund (2019, June). A guide for population-based analysis of the Adolescent Brain Cognitive Development (ABCD) study baseline data. https://doi.org/10.1101/2020.02.10.942011.

Holt, D. and T. M. F. Smith (1979). Post stratification. *Journal of the Royal Statistical Society Series A 142*(1), 33–46.

Kalton, G. and D. Kasprzyk (1988). The treatment of missing survey data. *Survey Methodology 12*(1), 1–16.

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

Kastellec, J. P., J. R. Lax, M. Malecki, and J. H. Phillips (2015). Polarizing the electoral connection: Partisan representation in supreme court confirmation politics. *The Journal of Politics 77*(3), 787–804.

Keiding, N. and T. A. Louis (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society Series A 179*(2), 319–376.

Kim, J. K. and S.-M. Tam (2020, June). Data integration by combining big data and survey sample data for finite population inference. https://arxiv.org/abs/2003.12156.

Lahiri, P. and K. Mukherjee (2007). On the design-consistency property of hierarchical bayes estimators in finite population sampling. *The Annals of Statistics 35*(2), 724–737.

Lauderdale, B. E., D. Bailey, J. Blumenau, and D. Rivers (2020). Model-based pre-election polling for national and sub-national outcomes in the US and UK. *International Journal of Forecasting 36*, 399–413.

Lax, J. and J. Phillips (2009a). Gay rights in the states: Public opinion and policy responsiveness. *American Political Science Review 103*, 367–386.

Lax, J. and J. Phillips (2009b). How should we estimate public opinion in the states? *American Journal of Political Science 53*, 107–121.

Leemann, L. and F. Wasserfallen (2017). Extending the use and prediction precision of subnational public opinion estimation. *American Journal of Political Science 61*(4), 1003–1022.

Little, R. (1983). Comment on "An evaluation of model-dependent and probability-sampling inferences in sample surveys", by M. H. Hansen, W. G. Madow and B. J. Tepping. *Journal of the American Statistical Association 78*, 797–799.

Little, R. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review 54*(2), 139–157.

Little, R. (1993). Post-stratification: A modeler's perspective. *Journal of the American Statistical Association 88*, 1001–1012.

Little, R. and H. An (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica 14*, 949–968.

Little, R. and S. Vartivarian (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology 31*(2), 161–168.

Little, R. and M. Wu (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American Statistical Association 86*, 87–95.

Little, R. J. A., B. T. West, P. S. Boonstra, and J. Hu (2019). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology smz023*.

Makela, S., Y. Si, and A. Gelman (2018). Bayesian inference under cluster sampling with probability proportional to size. *Statistics in Medicine 37*(26), 3849–3868.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Ann. Appl. Stat. 12*(2), 685–726.

Ornstein, J. T. (2020). Stacked regression and poststratification. *Political Analysis 28*(2), 293–301.

Pfeffermann, D. (2013, Feb). New important developments in small area estimation. *Statistical Science 28*(1), 40–68.

Rafei, A., C. A. C. Flannagan, and M. R. Elliott (2020). Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees. *Journal of Survey Statistics and Methodology 8*(1), 148–180.

Rao, J. and I. Molina (2015). *Small Area Estimation.* John Wiley & Sons, Inc.

Reilly, C., A. Gelman, and J. Katz (2001). Poststratication without population level information on the poststratifying variable, with application to political polling. *Journal of the American Statistical Association 96*, 1–11.

Rubin, D. B. (1976). Inference and missing data (with discussion). *Biometrika 63*, 581–592.

Rubin, D. B. (1983). Comment on "An evaluation of model-dependent and probability-sampling inferences in sample surveys," by M. H. Hansen, W. G. Madow and B. J. Tepping. *Journal of the American Statistical Association 78*, 803–805.

Sarndal, C., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Si, Y., N. S. Pillai, and A. Gelman (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis 10*(3), 605–625.

Si, Y., R. Trangucci, J. S. Gabry, and A. Gelman (2020). Bayesian hierarchical weighting adjustment and survey inference. Survey Methodology (accepted); https://arxiv.org/abs/1707.08220.

Si, Y. and P. Zhou (2020). Bayes-raking: Bayesian finite population inference with known margins. *Journal of Survey Statistics and Methodology smaa008*.

Smith, T. M. F. (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society. Series A (General) 146*(4), 394.

Stan Development Team (2020). RStan: the R interface to Stan. https://CRAN.R-project.org/package=rstan.

Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science 22*(4), 560–568.

The Economist (2020). Forecasting the US elections. https://projects.economist.com/us-2020-forecast/president.

Valliant, R. (2019). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology forthcoming.*

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

Wang, W., D. Rothschild, S. Goel, and A. Gelman (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting 31*(3), 980–991.

Yang, S. and J. K. Kim (2020, Jan). Statistical data integration in survey sampling: A review. https://arxiv.org/pdf/2001.03259.pdf.

Yang, S., J. K. Kim, and R. Song (2020). Doubly robust inference when combining probability and non-probability samples with high-dimensional data. *Journal of the Royal Statistical Society Series B Stat. Methodol. 82*, 445–465.

Yougov (2020). 2020 presidential election model. https://today.yougov.com/2020-presidential-election.

Zahorski, A. (2020). Multilevel regression with poststratification for the national level viber/street poll on the 2020 presidential election in Belarus. https://arxiv.org/abs/2009.06615.

Zangeneh, S. Z. and R. J. Little (2012). Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample. *Proceedings of the Section on Survey Research Methods.*

Zhang, G. and R. J. A. Little (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics 65*(3), 911–918.

Zhang, X., J. B. Holt, S. Yun, H. Lu, K. J. Greenlund, and J. B. Croft (2015). Validation of multi-level regression and poststratification methodology for small area estimation of health indicators from the behavioral risk factor surveillance system. *American Journal of Epidemiology 182*(2), 127–137.

Zhou, H., M. Elliott, and T. Raghunathan (2016). A two-step semiparametric method to accommodate sampling weights in multiple imputation. *Biometrics 72*(1), 242–252.

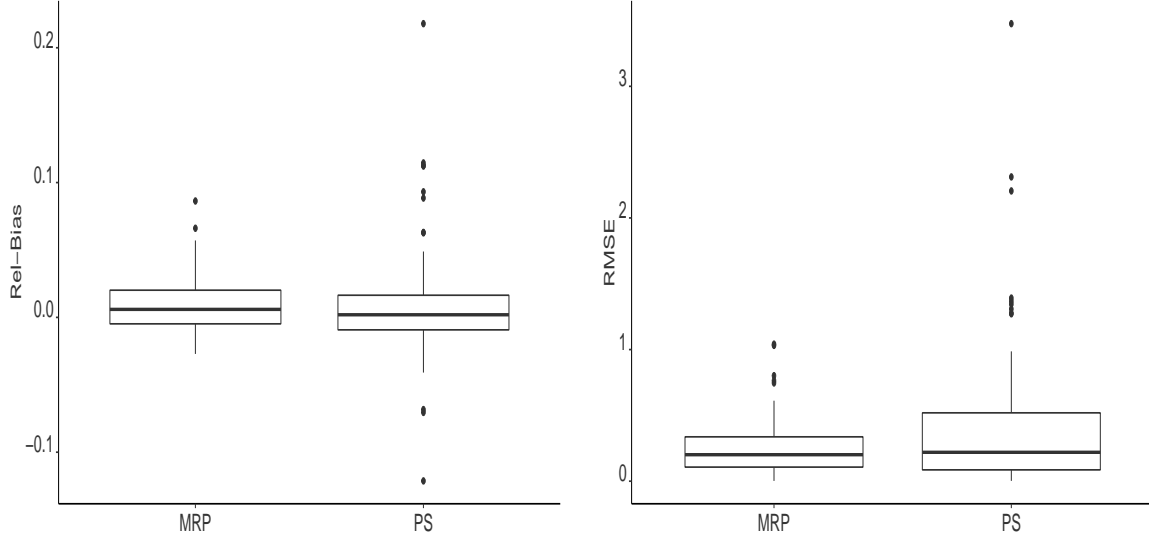**Title:** Figure supplement of outputs for Section 3.2

Figure 4: *Supplemental simulation outputs for Section 3.2 comparing the poststratification (PS) estimator and the MRP estimator on small domain estimates when the outcome and inclusion mechanism depend on main effects of sex, race/ethnicity and income.*