Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

to appear in Statistics in Medicine

RESEARCH ARTICLE

Multilevel Regression and Poststratification Using Margins of Poststratifiers: Improving Inference for HIV Health Outcomes During the COVID-19 Pandemic

Amy J. Pitts[1] | **Maiko Yomogida**[2] | **Angela Aidala**[2] | **Andrew Gelman**[3] | **Qixuan Chen**[1,4]

[1]Department of Biostatistics, Columbia University, New York, USA

[2]Department of Sociomedical Sciences, Columbia University, New York, USA

[3]Department of Statistics, Columbia University, New York, USA

[4]Division of Social Solutions and Services Research, Nathan S. Kline Institute for Psychiatric Research, New York, USA

Correspondence
Corresponding author Amy Pitts
Email: ajp2257@cumc.columbia.edu

Funding Information
This research was supported by the T32DA031099 for Pitts, by Health Resources and Services Administration grant H89HA00015 for Yomogida and Aidala through a subcontract with the New York City Department of Health and Mental Hygiene, and by National Institutes of Health grants R01AG067149 and R01ES035784 for Gelman and Chen.

Abstract

Multilevel regression and poststratification (MRP) has surged in popularity for population inference using survey samples. The method consists of two stages, first, fitting a regularized model that regresses the outcome on poststratification variables; second, predicting the outcome using the regularized model and aggregating the predictions to make population inferences. Existing MRP methods mostly focus on settings where the joint distribution of the poststratifiers is known in the population. However, in practice, such data information is often not available; instead, we are provided with the margins of the poststratifiers. Motivated by this challenge, we propose an adapted MRP in which we model both the survey outcome that we would like to estimate in the population and the population sizes of subgroups formed by the poststratifiers. We consider Poisson and negative binomial models for the population sizes of subgroups when the number of poststratifiers is small and Bayesian additive regression trees when there are many poststratifying variables. We apply the adapted MRP to estimate the proportion of viral load suppression and means of mental and physical health scales among persons with HIV in New York City using the 2018–2021 wave of the Community Health Advisory and Information Network survey, in which sampling and in-person data collection was disrupted by COVID-19 pandemic.

KEYWORDS
BART, Bayesian inference, MRP, Raking, Sampling Bias

1 | INTRODUCTION

Mitigating survey-related sampling bias is crucial for ensuring the reliability of survey results. Recent global events, notably the COVID-19 pandemic, have severely constrained the capacity to recruit and interview participants due to the heightened risk of disease transmission. Response rates to national and state surveys have decreased since the onset of the pandemic, with the decline being particularly pronounced in studies involving participants from lower-income and less educated backgrounds.[1] The socially patterned reduction in response rates poses a risk of sampling bias, which challenges the generalizability of results to a target population.

The Community Health Advisory and Information Network (CHAIN) project is an ongoing prospective community cohort study of people with HIV in New York City (NYC) and the tri-county regions of Westchester, Rockland, and Putnam Counties.[2,3,4] CHAIN provides comprehensive data from the perspective of individuals with HIV. It covers their healthcare and social services need, interactions with the entire spectrum of HIV-related services, and overall physical, mental, and social well being.[2,3,4] HIV-positive adults were recruited using a two-stage sampling strategy. The second cohort was recruited in 2002, followed by a refresher cohort in 2009–2010, and a third cohort (2015–2020) specifically targeting HIV-positive adults under 40. For each cohort, 30–40 service sites were randomly sampled from a compiled listing of HIV service agencies with at least 20 adult cases,

excluding private physicians or group practices, grouped by agency type (medical or social service) and geography (borough). Staff in selected agencies assisted with recruitment of a random sample of clients, either drawn from agency rosters or using sequential enrollment procedures.[5] This paper analyzes data from the 2018–2021 CHAIN survey, comprising 504 participants from the refresher and third cohorts. Due to the intentional sampling of younger HIV-positive adults in the third cohort, the survey sample skews younger than the target population. Additionally, like other in-person surveys, CHAIN has encountered recruitment and follow-up challenges attributed to the disruptions caused by the COVID-19 pandemic in its 2018-2021 wave of survey. Consequently, concerns have arisen regarding the representativeness of the 2018-2021 wave of sample in reflecting the population of interest due to potential nonresponse bias and selection bias. These biases stems from overrepresentation of younger adults, pandemic-related restrictions that halted data collection for several months, as well as the subsequent challenges of locating and conducting in-person interviews after restrictions were lifted.[2,4,6,7] To generalize the finding from this sample to people in NYC with HIV, it is imperative that steps are taken to adequately adjust for the sampling bias using population demographic and geographic information.

Known differences between sample and population in terms of demographic and geographic characteristics can be adjusted using poststratification, which uses the joint distribution of discrete auxiliary variables from the population to weight the sample and reduce bias in survey estimation.[8,9] However, access to the joint distribution of the population characteristics can be challenging. An alternative technique is raking, which only requires marginal distributions of population characteristics. Raking, also known as iterative proportional fitting, adjusts unit weights until the sample weighted distributions align with the population's marginal distributions for these discrete auxiliary variables.[10] Poststratification and raking are prone to instability, especially in cases with small subgroup samples.[11,12,13,14]

To obtain more efficient estimates, one increasingly popular technique is multilevel regression and poststratification (MRP)[15]. MRP is a two-step approach, first fitting an individual response model given the discrete auxiliary variables, then obtaining survey estimates using a poststratification table obtained from the joint distribution of the population characteristics.[15,16,17] MRP requires the complete joint distribution of the population characteristics, not merely the marginal distribution of each population characteristic, which is all that is used in classical raking adjustments. For the CHAIN project, due to considerations of patient security and confidentiality, the joint distribution of this population is not publicly accessible. As a result, MPR requires estimation of joint distribution of the population characteristics using the marginal distributions of specific auxiliary variables.

This paper introduces a novel adaptation of the MRP that uses just the data on population margins rather than the complete joint distribution of the population characteristics. Section 2 explains the method. Section 3 reports simulation studies examining the performance of the MRP adaptation compared to alternative methods in the settings when the auxiliary variables associated with survey outcomes and sample inclusion have perfect, partial, and no overlap. Section 4 applies the proposed MRP adaptation to the CHAIN cohort, which faced challenges to both selection bias and survey nonresponse. The paper concludes in Section 5 with recommendations for selecting various models when constructing MRP adaptations.

2 | METHODS

2.1 | Notation

Let y be a single outcome measure, taking either binary or continuous values. The main parameter of interest is the average of y in the population of interest, denoted $\overline{Y} = \sum_{i=1}^{N} Y_i/N$, with N being the population size. For every individual i, demographic and geographic information $\mathbf{X} = (X_1, \ldots, X_{Q+K})^T$ is gathered and grouped into either $q = 1, \ldots, Q$ where Q represents the total number of binary predictors, or $k = 1, \ldots, K$ where K represents the total number of polytomous variables, i.e., categorical variables with more than 2 levels. In the latter case, these categorical variables are further differentiated by $l = 1, \ldots L^{(k)}$ where $L^{(k)}$ signifies the total number of categories for the k-th variable. The same demographic and geographic variables, featuring identical categorizations, are assumed to be measured for both the survey sample and the population data.

2.2 | Poststratification and Raking

In order to capture the population characteristics, a summarization of $\mathbf{X}$ is consolidated into a poststratification table. This table encompasses $j = 1, \ldots, J$ distinct cells that correspond to different demographic and geographic subgroups where N_j denotes the count of individuals within the j-th population subgroup. With $\overline{Y}_j = \sum_{i=1}^{N_j} y_i/N_j$ being the population mean of y within cell j, the

overall population mean $\overline{Y}$ can be written as,

$$\theta = \overline{Y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{\sum_{j=1}^{J} N_j \overline{Y}_j}{\sum_{j=1}^{J} N_j}. \tag{1}$$

Using the sample data, the goal is to estimate the population mean. Let n_j denote the sample size in the j-th cell where $\sum_{j=1}^{J} n_j = n$. In addition, the sample mean in the j-th cell is $\overline{y}_j = \sum_{i=1}^{n_j} y_i/n_j$ and assumed to be an unbiased estimate of $\overline{Y}_j$. The poststratification estimate of θ can be written as

$$\hat{\theta} = \frac{\sum_{j=1}^{J} N_j \overline{y}_j}{\sum_{j=1}^{J} N_j} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}, \tag{2}$$

where $w_i = N_j/n_j$ for each sample unit i in cell j. Poststratification weights the sample based on the weight w_i created from the poststratification table. Variance estimation for $\hat{\theta}$ is performed using linearization or resampling methods.[18]

However, N_j is not always known. Raking, another weighting technique, can also be used to create the w_i weights. Different from poststratification, raking only requires the marginal distributions of the auxiliary variables in the population. The raking procedure takes one auxiliary variables at a time and weights the sample to match the marginal distribution of the population for the variables. The procedure is iterated until the weighted distributions of the auxiliary variables in the sample conform to the marginal distributions for each of the auxiliary variables in the population.[8,9]

2.3 | Multilevel Regression and Poststratification (MRP)

The MRP method for estimating the population mean consists of two steps. The first step is to fit an individual response model using demographic and geographic variables as predictors.[15,16,17] Here we follow the usual choices of logistic regression for binary outcomes and normal regression for continuous outcomes. The formulation of the multilevel regression can then be expressed as follows:

$$g\left(\mathrm{E}(y_i|\mathbf{x})\right) = \alpha_0 + \sum_{q=1}^{Q} \alpha_q x_{iq} + \sum_{k=1}^{K} \alpha_{\ell[i]}^{(k)}, \tag{3}$$

$$\alpha_{\ell[i]}^{(k)} \mid \sigma^{(k)} \sim \text{normal}\left(0, \sigma^{(k)}\right), \tag{4}$$

where $g(\cdot)$ is a link function using an identity link for continuous y and a logit link for binary y. In this context, α_0 denotes the intercept, α_q represents the regression coefficient associated with the qth binary predictor, and $\alpha_{l[i]}^{(k)}$ encapsulates the variable coefficient associated with the lth category of the kth polytomous variable. In the second level of the model in (4), $\alpha_{l[i]}^{(k)}$ are modeled using a normal distribution with mean at zero and the variable-specific standard deviation $\sigma^{(k)}$. Bayesian inference can be applied. Prior information is specified for α_0, α_q, and $\sigma^{(k)}$, such as a uniform prior for the α and α_q coefficients and a truncated normal distribution for $\sigma^{(k)}$. The above model can be expanded to allow interactions between the predictors. Both the main effects and interaction terms associated with the polytomous variables are specified as multilevel random effects to achieve partial pooling during model fitting. Partial pooling stabilizes estimates for small groups by borrowing information from larger groups, preventing overfitting in small sample cells while preserving meaningful variation across categories.[19] The choice of the number of categories in a polytomous variable (e.g., using three large age groups vs. seven more detailed age groups) can influence MRP estimates. More categories allow for greater flexibility in capturing heterogeneity, but when sample sizes in some groups are small, estimates may become unstable. Thus small cells may be merged with adjacent categories that share similar outcome distributions to improve stability.

In the second step, the multilevel regression model in (3) and (4) is used to obtain the estimated mean of y in the j-th population subgroup, denoted using $\hat{\theta}_j$. When N_j is known, the overall population mean in (1) can be estimated using

$$\hat{\theta}^{\text{MRP}} = \frac{\sum_{j=1}^{J} N_j \hat{\theta}_j}{\sum_{j=1}^{J} N_j}. \tag{5}$$

This estimation can also be expanded to encompass the estimation of any specific sub-population denoted as s:

$$\hat{\theta}_s^{\text{MRP}} = \frac{\sum_{j \in J_s} N_j \hat{\theta}_j}{\sum_{j \in J_s} N_j}, \tag{6}$$

where J_s is the subset of all the cells in the poststratification table that are part of s.

The MRP model was implemented using Hamiltonian Monte Carlo (HMC) via the `stan_glmer` function in the `rstanarm` package in R.[20,21] This function estimates Bayesian regression models through RStan, which provides an R interface to the Stan C++ library for Bayesian estimation.[21] The point estimate of θ is obtained by averaging posterior draws of (5) or (6) from their posterior distributions using the `posterior_epred()` function. The credibility interval is derived by computing the quantiles of these posterior draws.

MRP improves efficiency in the population mean estimation than poststratification and raking, especially when data are sparse in some cells. However, the good performance of MRP relies on the inclusion of important predictors of y. In addition, MRP requires N_j to be known for all poststratification cells, but in practice, they are often unknown. Motivated by this challenge, in the next section, we present an adaptation of MRP that allows for the case where N_j is unknown, but the margins of the auxiliary variables are known for the population.

2.4 | Adaptation of MRP When N_j is Unknown

To address situations where obtaining the joint distribution of the auxiliary variables in the population is not possible, we propose an adaptation of the MRP that estimates N_j. We consider three ways of estimating N_j, including raking, modeling using parametric Poisson or negative binomial regression, and a machine learning approach using Bayesian additive regression trees (BART). In all three approaches, we assume population margins of the auxiliary variables are known for the population.

2.4.1 | MRP Rake Count

One strength of raking is it only requires the marginal distribution of the population to adjust for the estimation of the sample data. This property allows us to estimate the joint distribution of the auxiliary variables in the population by raking on the joint distribution of these auxiliary variables in the sample. Let $\hat{N}_j^{Rake}$ denote the estimated N_j using raking. To obtain an estimate of θ, we then use

$$\hat{\theta}^{\text{MRP rake}} = \frac{\sum_{j=1}^{J} \hat{N}_j^{\text{Rake}} \hat{\theta}_j}{\sum_{j=1}^{J} \hat{N}_j^{\text{Rake}}}. \tag{7}$$

Similarly, we can replace N_j in equation (6) with $\hat{N}_j^{Rake}$ to obtain estimation for specific subgroups of the population. Although using raking to estimate N_j is straightforward and easy to implement, simply treating $\hat{N}_j^{Rake}$ as if they were the true N_j is problematic. It ignores the uncertainty associated with the estimated N_j and thus could lead to probability intervals that are too narrow and may miss the true population parameters.

2.4.2 | MRP Poisson and Negative Binomial

To account for the uncertainty in estimating N_j we extend the raking approach to a two-step approach. We first model the sample count c_j for cells in the poststratification table formed by the auxiliary variables using a Bayesian model. We then apply raking to the estimated c_j drew from their posterior predictive distributions based on the fitted models. With c_j being count data, we consider a Bayesian Poisson loglinear model or Bayesian negative binomial (NB) model with the latter being used if there is hypothesised overdispersion in the sample count data.

For the Poisson loglinear model, we first assume $c_j \sim \text{Poisson}(\lambda_j)$. Then we set up a multilevel regression model as

$$\log(\lambda_j) = \beta_0 + \sum_{q=1}^{Q} \beta_q x_{jq} + \sum_{k=1}^{K} \beta_{\ell[j]}^{(k)}. \tag{8}$$

The β coefficients in equation (8) are defined similarly to the α coefficients in equation (3), but they represent the associations between the auxiliary variables and the sample counts. Prior distributions are specified independently for each β coefficient.

For the varying coefficients $\beta_{\ell[j]}^{(k)}$, independent normal distributions are specified as $\beta_{\ell[j]}^{(k)} \mid \tau^{(k)} \sim \text{normal}\left(0, \tau^{(k)}\right)$ where $\tau^{(k)}$ follows a positive truncated normal distribution. If there is hypothesised overdispersion, a similar approach can be taken with the assumption that c_j follows a negative binomial distribution instead of a Poisson distribution. Since c_j is only used in estimating model (8), it is acceptable for some values to be small.

From this fitted model we take 1000 draws of the sample counts from their posterior predictive distributions. For each set of draws, we apply the raking method using the known marginal distribution of the population data and obtain $\hat{N}_{j,d}^{\text{Poisson}}$ (or $\hat{N}_{j,d}^{\text{NB}}$) for $d = 1, \ldots, 1000$ draws. Similarly, we obtain $d = 1, \ldots, 1000$ draws of $\hat{\theta}_{j,d}$ based on model (3). For each draw of $\hat{N}_{j,d}$ and $\hat{\theta}_{j,d}$, we obtain a draw of

$$\hat{\theta}_d^{\text{MRP Poisson}} = \frac{\sum_{j=1}^{J} \hat{N}_{j,d}^{\text{Poisson}} \hat{\theta}_{j,d}}{\sum_{j=1}^{J} \hat{N}_{j,d}^{\text{Poisson}}} \qquad \text{or} \qquad \hat{\theta}_d^{\text{MRP NB}} = \frac{\sum_{j=1}^{J} \hat{N}_{j,d}^{\text{NB}} \hat{\theta}_{j,d}}{\sum_{j=1}^{J} \hat{N}_{j,d}^{\text{NB}}}. \tag{9}$$

These draws are averaged to obtain the point estimate of θ. The 95% probability interval is formed by taking the 2.5th and 97.5th percentile of the posterior distributions of θ. To account for the uncertainty in multiplying $\hat{N}_{j,d}$ and $\hat{\theta}_{j,d}$, we introduced variability by randomly shuffling the 1000 $\hat{N}_{j,d}$ values before performing the multiplication.

Using a model-based approach to estimate sample counts $c_j, j = 1, \ldots, J$, propagates the uncertainty in estimating N_j and thus improves the coverage of probability interval associated with θ. However, poor model specification for c_j can lead to poor estimation of N_j and thus biased estimate of θ. Therefore, the model for the survey sample count c_j should include all the important predictors of c_j. When the number of auxiliary variables is not small, deciding which variables to include and whether to include interaction terms is not a simple task and requires expert knowledge in the field or exploratory variable selection methods. This motivated the development of the next approach using Bayesian machine learning methods.

2.4.3 | MRP BART

Bayesian additive regression trees (BART) is a sum-of-trees model for nonparametric function estimation. Over the past decade, it has gained significant attention due to its ensemble-based methodology, which provides flexibility through the use of decision trees while preventing overfitting by incorporating a regularization prior on the model parameters.[22] This regularization prior restricts each binary tree to be explained by a subset of available relationships between variables. BART offers a Bayesian, non-parametric approach to modeling sample counts.

In this approach, we follow similar steps as the MRP Poisson and MRP negative binomial methods but use BART to model the sample count data c_j. Current software for BART supports a range of outcome types including continuous, binary, categorical, and time-to-event data.[23] While there exists literature describing a BART procedure that can accommodate count outcomes, dedicated software package for this purpose is currently unavailable.[24] As a workaround, we adopt a common practice of transforming the data by taking the square root of the counts. This transformation mitigates the skewness inherent in count data and permits analysis on a continuous scale.

To establish the sum-of-trees model for the square root of the counts, let M be number of trees, T_m be the mth tree, $m = 1, \ldots, M$, and $\mu_m = \{\mu_1, \ldots, \mu_{b_m}\}$ contain a list of parameter values associated with each of the b_m terminal nodes of T_m. Let $g(\mathbf{x}_j; T_m, \mu_m)$ denote the function that assigns μ_m according to $\mathbf{x}_j$. The model takes the form

$$c_j^* = \sqrt{c_j} = \sum_{m=1}^{M} g(\mathbf{x}_j, T_m, \mu_m) + \epsilon_j, \quad \epsilon_j \overset{\text{iid}}{\sim} \text{normal}(0, \sigma). \tag{10}$$

Regularization priors are specified for $p(T_m)$, $p(\mu_m|T_m)$, and $p(\sigma)$ such that each single tree T_m is a weak learner. The $p(T_m)$ prior has three aspects which control whether the tree branches grow or are pruned. First, the probability that a node at depth d is nonterminal is given by $\alpha/(1+d)^\gamma$ where $\alpha \in (0, 1)$ and $\gamma \in [0, \infty)$. The default values recommended are $\alpha = 0.95$ and $\gamma = 2$.[25] Second, a uniform prior is specified for the splitting variables assignment at each interior node. Third, conditional on the splitting variables a uniform prior is specified for the splitting rule assignment in each interior node. For $p(\mu_m|T_m)$, a normal prior distribution is used with mean 0 and variance σ_μ^2 where $\sigma_\mu = 0.5/k\sqrt{m}$, with the suggested value of k between 1 and 3.[22] This prior limits the effects of individual tree components by shrinking the tree parameter towards zero. The prior for σ is specified as the scaled inverse chi-square distribution with degrees of freedom ν and scale λ. The recommended default values are $\nu = 3$. The λ values can be user specified but is defaulted to take an estimate of $\hat{\sigma}$ which can be either user specified or determined via linear regression or the sample standard deviation. The posterior inference is completed using a Markov chain Monte Carlo and back fitting technique[22,23,26] and implemented using the `BART` package in R.[25]

Based on model (10), we obtain 1000 draws of c_j^* and take square to obtain 1000 draws of c_j. We then follow the same procedure as the MRP Poisson to obtain the posterior distribution for θ. Specifically, for each set of draws of c_j, we apply the raking method using the known marginal distributions of $\mathbf{x}$ in the population to obtain $\hat{N}_{j,d}^{\text{BART}}$ for $d = 1, \ldots, 1000$ draws. We then replace $\hat{N}_{j,d}^{\text{Poisson}}$ with $\hat{N}_{j,d}^{\text{BART}}$ in (9) and obtain the posterior distribution of θ.

3 | SIMULATION STUDY

3.1 | Data Generation

The objective of the data generation process is to create a synthetic population dataset that mimics the real data application but allows us to know the truth. To maintain simplicity, we create a population dataset that contains four categorical covariates, $X_1, \ldots, X_4$, and two outcome variables, one binary and the other continuous. Each covariates is generated with 4 categories and coded using numbers 1–4, thus a poststratification table with up to 256 cells can be created to denote each possible combination of the 4 covariates. To simulate the population data with certain counts in each of the poststratification cells, a Poisson distribution is used with $\log(\lambda) = 10\,(X_1 + X_2 + X_3 + X_4)$ using the numerical values of the X variables.

To generate the outcomes we use a multilevel regression model approach. We let

$$\eta_i = a_0 + Z_i^{(X_1)}\gamma^{(X_1)} + Z_i^{(X_2)}\gamma^{(X_2)} + Z_i^{(X_1 \times X_2)}\gamma^{(X_1 \times X_2)}, \tag{11}$$

where $\gamma^{(\cdot)}$ is a column vector of random draws from a normal distribution, with $\gamma^{(X_1)}$ and $\gamma^{(X_2)}$ having 4 elements and $\gamma^{(X_1 \times X_2)}$ having 16 elements; and the $Z_i^{(X_\cdot)}$ is a row vector with one in the entry corresponding to the value of $X_\cdot$ for unit i and 0 otherwise. For the binary outcome, we let $a_0 = -3$, $\gamma \overset{iid}{\sim} \text{normal}(0, 0.5)$, and generate y from a Bernoulli distribution with $\Pr(y_i = 1) = \text{logit}^{-1}(\eta_i)$. For the continuous outcome, we let $a_0 = 0$, $\gamma \overset{iid}{\sim} \text{normal}(0, 1)$, and generate y_i from a normal distribution with $y_i \overset{ind}{\sim} \text{normal}(\eta_i, 1)$.

Sample data are drawn from the population using an inclusion mechanism that resembles the hypothesized sampling bias in real data scenarios. We allow the sample inclusion indicator r_i to be correlated with different subsets of Xs:

$$\Pr(r_i = 1) = \text{logit}^{-1}(-2.5 + 1.75X_{1i} + 1.75X_{2i} - 0.25X_{1i} \times X_{2i}) \quad \text{(perfect overlap)}$$
$$\Pr(r_i = 1) = \text{logit}^{-1}(-1.75 + 1.5X_{1i} + 1.5X_{3i} - 0.25X_{1i} \times X_{3i}) \quad \text{(semi overlap)}$$
$$\Pr(r_i = 1) = \text{logit}^{-1}(-1 + 1.25X_{3i} + 1.25X_{4i} - 0.25X_{3i} \times X_{4i}) \quad \text{(no overlap)}$$

The X variables that are associated with r are also the predictors of y in the "perfect overlap" scenario but are not related to y in the "no overlap" scenario. The "semi overlap" scenario is in the middle, with one common predictor in the two models but each model also has its own predictors. The semi-overlap scenario is the more common setting with the perfect overlap and no overlap representing the two extreme cases. To determine what scenario applies to a specific data application with regards to specific auxiliary variables the association with the inclusion and the outcome can be examined.

We conducted simulations with 500 replicates under each sample data generation mechanism. We compared the performance of the four adapted MRP methods, including MRP rake, MRP Poisson, MRP NB, and MRP BART, to the original MRP method assuming N_j is known, the weighted method using raking, and the unadjusted estimate that simply averages the sample data and calculates the 95% confidence interval using the normal distribution. We considered the same multilevel regression model of y on X_1, X_2, X_3, X_4 and all their two-way interactions in both the original MRP and adapted MRP methods. In using the MRP Poisson, MRP NB, and MRP BART, we included all 4 covariates in the models predicting c_j, and all the 2-way interactions in the Poisson and NB models. Raking and MRP adaptations were conducted using the population margins of all the 4 covariates.

Evaluation is based on the absolute bias, root mean square error (RMSE), coverage rate of the 95% credible or confidence interval (CI), and average width of the 95% CI.

3.2 | Simulation Results

We generated a population comprising 25,606 observations, wherein the binary outcome variable y has a true proportion of 0.42, while the continuous outcome variable y has a true mean of 5.10. These values serve as the benchmarks for evaluating the

performance of each method employed in our simulation study. For each sample-generating mechanism, the average sample size is 800 observations. Exploring the distribution of the 256 cell counts formed by $X_1 - X_4$, as depicted in Figure 1A, we observe that the histogram approximately adheres to a normal distribution, with its center located at 100. To mimic the increasing trend of physical wellness scores across age groups in our real data application, we ordered the simulated values of γs in (11). Figure 1B and 1C exhibit the distributions of the two outcome variables (binary and continuous) by the two influential covariates, X_1 and X_2.

Table 1 provides a comprehensive assessment of each method's performance in recovering the true population values across various sample-generating mechanisms. Looking at the perfect-overlap and semi-overlap data generation mechanism it becomes apparent that the unadjusted estimates exhibit significant biases and RMSE, along with a striking absence of coverage (0%) for both binary and continuous outcomes. These findings imply that the unadjusted estimates deviate substantially from the population truth, and the 95% confidence interval never encompasses the true values. This pronounced lack of fit highlights the inadequacy of the unadjusted approach in these scenarios. In the no-overlap data generating case, the unadjusted method's bias and RMSE are only slightly above that of the raking, MRP, and MRP adaptations. Moreover, the coverage rate is 93% for the binary outcome and 78% for the continuous outcome. These results align with expectations, as the no-overlap sampling mechanism uses the covariates that have no influence on the outcome, thus not accounting for the sampling bias does not result in extremely biased results. However, accounting for covariate information in the analysis does improve results by reducing bias and RMSE and increasing the coverage of 95% intervals.

The raking method has the smallest bias across all scenarios, except for the continuous y under the perfect-overlap and no-overlap cases. However, raking has a larger RMSE than MRP and its adaptations and also produces the widest 95% confidence intervals.

When examining all sample data mechanisms for both binary and continuous outcomes, we find that MRP and its four adaptations produce similar results, with MRP showing slightly lower bias and RMSE. This consistency across various scenarios is encouraging, suggesting that in this data-generating context, adaptations relying solely on the marginal population distribution perform just as effectively as the MRP approach, which uses the entire joint population distribution.

When comparing the four MRP adaptations, we observed similar performance across various outcome types and sample data-generating mechanisms. However, the MRP rake method has the largest RMSE compared to the other three MRP adaptations in the perfect-overlap and semi-overlap scenarios. Accounting for uncertainty in the N_j estimation using either parametric or machine learning models improves coverage rate for the continuous outcome. MRP BART performs comparably to MRP Poisson and MRP NB, making it a promising alternative when dealing with a large number of covariates, where traditional parametric models may become less feasible.

4 | APPLICATION TO THE CHAIN DATA

4.1 | Data Structure and Model Specification

The 2018–2021 sample of the CHAIN study includes 504 individuals with HIV in NYC. The primary interest is the viral load suppression, recorded as a binary endpoint. Specifically, data were collected as self-reported most recent HIV viral load as an actual numerical value or reported medical provider designation as "undetectable." Viral load of < 200 copies per milliliter (mL) or provider report as "undetectable" were coded as "suppressed viral load" and otherwise as "unsuppressed viral load." In addition, secondary endpoints such as mental and physical functioning scales provide insight into the overall health, mental health functioning, and quality of life, as shaped by social determinants of health. The Mental Component Summary (MCS) and Physical Component Summary (PCS) scores are derived from the Medical Outcome Study Measures of Quality of Life Short Form Survey Version 2 (MOS SF-12v2), which uses norm-based scoring ranging from 0 to 100, with a mean of 50 and a standard deviation of 10, based on US general population studies. Low mental health functioning is indicated by MCS score lower than 42, associated with clinically significant mental health symptoms; Low physical health functioning is indicated by PCS score lower than 50, associated with some degree of physical limitation or impairment.[27,28] Missing data is a minor concern in this dataset. Missingness in the covariates only makes up about 8.73% of the dataset. A hot deck imputation was employed to impute the missing covariates.[29,30,31] There are also some missing data in the outcomes, with 51 (10%) observations missing the viral suppression and 3 (0.6%) observations missing the MCS and PCS variables. For each outcome, only the data with complete outcome measures were used and thus our total sample size varies by outcomes with $n = 453$ for the viral load suppression and $n = 501$ for the MCS and PCS.

Covariate selection is an important component of the MRP method and its adaptations. Auxiliary variables that are included as covariates in the models need to be both measured in the CHAIN data and available as summary statistics in the population. We obtained the population summary statistics from three data sources, including the 2020 NYC HIV/AIDS annual surveillance statistics, the 2013–2014 NYC Medical Monitoring Project data, and the 2019 Ryan White HIV/AIDS Program annual client-level data report.[32,33,34] Among the binary covariates considered are gender, men who have sex with men (MSM), individuals with a history of injection drug use (EverIDU), and those who acquired HIV perinatally versus behaviorally. Transgender individuals are categorized within the women gender group due to the presence of only a limited number of transgender participants (14 individuals) in the dataset, all of whom identify as women. Additionally, our analysis incorporates polytomous variables, including age groups (20–29, 30–39, 40–49, 50–59, $\geq$ 60 years), race/ethnicity (Black, Latino, White, Other), NYC borough of residence (Bronx, Brooklyn, Manhattan, Queens, Staten Island), education levels (less than high school, high school, more than high school, including but not limited to some college), housing status (Stable, Temporary, Unstable), and poverty status (0–100%, 101–138%, 139–250%, 251–400%, > 400% federal poverty level).

We estimated the prevalence of HIV viral suppression and the mean MCS and PCS scores using the 4 MRP adaptation methods, including MRP rake, MRP Poisson, MRP NB, and MRP BART, which were then compared to the raking approach and the unadjusted estimate using sample mean.

We first fit the outcome models using multilevel linear regression for the MCS and PCS scores and multilevel logistic regression for the viral suppression. To account for potential correlations among individuals recruited from the same medical or social service agency, the agency is modeled as a polytomous variable with random coefficients. We regressed each outcome on all the auxiliary variables except for perinatally acquired HIV. The decision to exclude "perinatally acquired HIV" from this initial modeling step was attributed to its notably low prevalence (2% in the population). Further, instead of including all the 2-way interactions, we only included in the models the interactions of key predictors of each outcome. To decide the importance of each covariate in predicting the outcomes, we ran a BART model on each outcome and obtained the posterior variable importance scores.[22] For the MCS outcome, important predictors were identified as ethnicity, education level, and NYC borough. In contrast, for the PCS outcome, key predictors encompassed age category, gender, and NYC borough. Lastly for the viral suppression, important variables included poverty category, NYC borough, and gender. These variables were subsequently chosen for inclusion as interaction terms in the model.

We then applied the raking technique to estimate population counts formed by all these 3 binary and 6 polytomous auxiliary variables, and estimated the prevalence of viral suppression and mean MCS and PCS scores by combining the draws of cell-specific means from the outcome models and the estimated population counts using the four MRP adaptations. To account for the uncertainty in the estimates of population counts, in MRP Poisson and MRP NB, we fit a multilevel Poisson or negative binomial regression model on the sample counts, including all these 9 covariates in a marginal manner. In MRP BART, we fit a BART model for normal data on the square root transformed sample counts which allows variable selection and interaction detection.

4.2 | Results

Table 2 compares the distribution of the auxiliary variables in the CHAIN sample and in the population and shows the percentage of viral suppression and means of the MCS and PCS scores in the sample and stratified by each auxiliary variable. Young people with HIV were over-represented in our sample, which can be problematic because outcome values varied across age groups. For example, looking at the age group of 20 to 29 years there was about 5.8% of the population in this subgroup but in our sample we had 17.5% falling in this category. In contrast, for the age group of 60 years or older there was 29.7% of the population in this group but the sample only had 10.5% of the participants in this group. The issue becomes apparent when looking at the physical health scale category. For the 20–29 age group the mean value was 53.0 whereas in the 60 years or older age group the mean value was 42.6. This difference in age distribution could bias the estimate of mean PCS score of the population if the sample data were used without adjustment. A similar but opposite trend can be seen in the mental health scale where younger age groups had a lower score compared to older age groups. A similar story can be told for the majority of demographic and geographic variables listed in Table 2.

Figure 2 displays the results from the unadjusted estimate using sample mean and the adjusted methods using raking and the four MRP adaptations for each outcome of interest. The unadjusted method estimated that 94% (95% CI: 91%–96%) of individuals with HIV in NYC were viral suppressed with a mean MCS score of 40.3 (95% CI: 39.6–41.1) and a mean PCS score of 49.7 (95% CI: 48.8–50.6). The four MRP adaptations yielded similar estimates for all three outcomes. Compared with the unadjusted estimates, the MRP adaptations estimated a similar viral suppression rate (93%), a higher mean MCS score

(41.1), and a lower mean PCS score (47.3). The changes in the MCS and PCS estimates from the unadjusted method to the MRP adaptations were expected. In our exploratory analysis in Table 2 we see that older adults were under-represented in our sample and they had higher MCS scores but lower PCS scores than young adults. The MRP adaptations corrected the differences in the distributions of age and other auxiliary variables between the sample and the population and thus reduced bias in the estimates of population quantities. Finally, the estimates using the raking approach differed from the other methods for all three outcomes, with the lowest estimate of viral suppression rate, the lowest mean estimate of PCS score, but the highest mean estimate of MCS score among all the methods. Further, the raking approach led to a much wider confidence interval than the other methods.

We also conducted a subgroup analysis of the three primary outcome measures across distinct age groups (categorized as age < 40, age 40–49, and age > 50), as illustrated in Figure 2. The raking approach produced even more unstable estimates in subgroup analysis compared to the overall analysis. This instability is evident in the considerably wider 95% CI and the substantial disparities in the point estimates across all three outcome measures. Compared to the unadjusted estimates, the MRP-adapted methods yielded similar estimates of all three outcomes among individuals younger than 40 years; had lower viral suppression rate, lower mean MCS score, and slightly higher mean PCS score among those between 40 and 49 years of age; and had slightly lower viral suppression and much lower MCS score among those older than 50 years. The MRP-adapted method did not inflate the variability in the estimates, yielding 95% credible intervals with similar lengths compared to the unadjusted method even in the subgroup analyses. Unlike the overall analysis, the difference in mean PCS scores between MRP adaptations and the unadjusted estimate is subtler in the subgroup analysis. Age, the key factor influencing the physical health scale, reduces selection bias when conditioned upon. In contrast, MRP adaptations yielded lower mean MCS scores than the unadjusted analysis in the subgroup analysis, particularly among HIV patients over 50. This suggests additional factors, beyond age, contribute to biased mental health scale estimates.

5 | DISCUSSION

Multilevel regression and poststratification (MRP) has surged in popularity for population inference using survey samples. Traditional MRP approaches typically assume that the joint distribution of poststratifiers is known in the population, but in practice, such data are often unavailable. While some studies have explored cases where the joint distribution is available for certain auxiliary variables but not for other important outcome predictors,[35,36] few have examined scenarios where only the marginal distributions of the poststratifiers are known. To address this challenge, we propose MRP adaptations in which we model both the survey outcome of interest and the population sizes of subgroups formed by the poststratifiers. We consider multiple ways to estimate the joint distribution of auxiliary variables in the population using survey data and population marginal distributions, including raking, parametric modeling with Poisson or negative binomial families, and a machine-learning approach based on BART model.

To examine the finite sample performance of the approaches, a simulation study was conducted to compare the MRP adaptations against an unadjusted estimate and the weighting method using raking. They are also compared against the original MRP approach where the true joint distribution of the population is used. Three different sampling mechanisms were used to explore the approaches on differing degrees of sampling bias. Consistent with previous research, MRP yields smaller RMSE, narrower 95% CI, and better coverage than raking. In addition, the unadjusted estimate is biased, and the degree of bias decreases as the overlap between covariates associated with survey outcomes and sampling is reduced. The MRP adaptations perform similarly to MRP, although MRP exhibits slightly smaller bias and RMSE by using true population counts in subgroups instead of estimated counts. Finally, incorporating the uncertainty in estimating population counts in subgroups using either parametric or machine learning models improves the estimation, resulting in slightly smaller bias and RMSE, and enhanced coverage rates, especially in the no-overlap case for the continuous outcome.

We apply the adapted MRP to estimate the proportion of viral load suppression and means of mental and physical health scales among persons with HIV in New York City using the 2018–2021 sample of the Community Health Advisory & Information Network survey, in which the survey sample was overrepresented by young adults and the in-person data collection was disrupted by COVID-19 pandemic. Our results show that relying on the unadjusted estimate would have overestimated the physical health, leading to the misinterpretation that people with HIV had better physical health than they actually did. This finding is expected, given that 2018–2021 sample is younger compared to the broader population of individuals with HIV in New York City. The age subgroup analysis further highlights that age is the primary driver of sampling imbalance and bias in the physical health scale outcome, as shown by the unadjusted estimate aligning closely with the four MRP adaptations. In contrast, for mental health outcome, the unadjusted estimate deviates from the four adaptations in the age-stratified analysis, suggesting that other variables

also contributed to this discrepancy. Consistent with the simulation study, the raking approach produced unstable estimates with wide confidence intervals, especially in the age subgroup analysis.

To assess the impact of each auxiliary variable in reducing sampling bias, a comparative table, similar to Table 2, showing the distribution of auxiliary variables in both the sample and target population, as well as outcome distributions across different auxiliary variable categories in the sample, can be highly informative. This comparison helps identify which auxiliary variables are linked to sampling bias and which are associated with outcomes. The degree of overlap between these two sets of variables defines the three scenarios considered in our simulation. Such a table can also aid in determining which variables should be included when fitting the outcome model (3).

Fitting the predictive model for n_j can be challenging when many auxiliary variables are included. Our simulation study and real data application have shown that BART is an attractive alternative to parametric models for estimating n_j. BART allows for the inclusion of a large number of covariates without needing to specify which covariates to include or their functional forms. We recommend using Poisson or negative binomial models when the number of poststratifiers is small and the BART approach when there are a large number of poststratifiers.

Ideally, incorporating a larger number of auxiliary variables with finer categories can help correct sampling bias. However, when the number of auxiliary variables and categories becomes too large, it can lead to zero poststratification cells, which in turn may cause slow or failed convergence.[37,38] Similarly, while finer categories enhance flexibility in capturing heterogeneity in outcomes, small sample sizes in certain cells can lead to unstable estimates in the multilevel model. One solution is to combine variables and collapse categories to ensure sufficient sample sizes in each collapsed cell, balancing granularity and stability in estimation. Alternatively, variable selection can be applied to include only auxiliary variables that are strongly associated with both outcomes and sampling bias in the adapted MRP methods. This approach promotes a more parsimonious and stable model while also improving convergence in raking.

In this paper, we assume correct model specification for survey outcomes in the first step of the MRP framework. In practice, identifying the true model structure is often difficult, especially in settings with a large number of auxiliary variables and complex relationships. In such data-rich environments, BART offers a nonparametric alternative for estimating population quantities from nonrandom samples.[39] One possible extension of our MRP adaptations is to replace the parametric multilevel model for survey outcomes with BART, resulting in a "double BART" approach in which BART is used to model both the survey outcomes and the sample counts across subgroups.

So far, we have considered surveys without complex survey designs. When a survey involves stratification and clustering, these design features can be incorporated into the MRP model by including strata and cluster indicators as polytomous covariates. The coefficients corresponding to each level of the strata and cluster indicators are modeled hierarchically using a normal distribution with mean zero and a standard deviation estimated from the data. The models can be easily fit using widely available statistical software, such as `rstanarm` used in this paper, even when the number of strata and cluster groups is large. If the survey includes sample weights, reflecting selection probabilities and nonresponse adjustments, these weights should also be integrated into the MRP model. Expanding our methods to incorporate survey weights is an important area for future research. One possible approach is to model the joint distribution between outcome and survey weights.[40]

Modeling population counts in subgroups formed by the poststratifiers requires access to auxiliary variable information at the population level. In our setting, we assume that marginal population counts for the auxiliary variables are available. However, in practice, such information may be unavailable, or only available for a subset of the auxiliary variables. In these situations, well-designed and properly executed probability surveys targeting the same population can serve as valuable alternative data sources for estimating auxiliary variable distributions. Crucially, the uncertainty associated with these estimated population counts must be appropriately incorporated into the MRP framework. Extending our current methods to model poststratification cell counts using data from probability surveys represents another important direction for future research.

ACKNOWLEDGMENTS

We thank the Associate Editor and the two referees for their thoughtful and constructive comments which greatly improved the paper, and we thank the U.S. National Science Foundation, National Institutes of Health, and Office of Naval Research for partial support of this work.

REFERENCES

1. Krieger N, LeBlanc M, Waterman PD, Reisner SL, Testa C, Chen JT. Decreasing survey response rates in the time of COVID-19: Implications for analyses of population health and health inequities. *American Journal of Public Health.* 2023;113:667–670.
2. Aidala A, Yomogida M. Housing need, housing assistance, and engagement with HIV medical care. *Community Health Advisory & Information Network (CHAIN). HIV Health & Human Services Planning Council of New York.* 2019.

3. HIV Health & Human Services Planning Council of New York. HIV Data Reports & Resources. https://nyhiv.org/tdb_templates/chain-reports-template-12/; Updated June 25, 2024.
4. Aidala A, Yomogida M. Housing need, housing assistance, and engagement with HIV medical care: Program and policy implications. *APHA 2019 Annual Meeting and Expo.* 2019.
5. Abramson D, Messeri P, Aidala AA, Healton C, Jessop D, Jetter D. Recruiting Rare and Hard-to-reach Populations: A sampling strategy for surveying NYC residents living with HIV/AIDS. *Journal of the American Statistical Association.* 1995.
6. Messeri P, Aidala A, Abramson D, Healton C, Jones-Jessop D, Jetter D. Recruiting rare & hard to reach populations: A sampling strategy for surveying NYC residents living with HIV/AIDS using agency recruiters. *American Statistical Association Proceedings of the Section on Survey Research Methods.* 1995;2:1064–1068.
7. Messeri P, McAllister-Hollod L, Irvine M. NYC CHAIN 2012-8 report - 5/30/13 - Validating self-reported HIV test results using surveillance registry data. *Columbia University Mailman School of Public Health.* 2013.
8. Deming WE, Stephan FF. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics.* 1940;11(4):427–444.
9. Deville JC, Särndal CE, Sautory O. Generalized raking procedures in survey sampling. *Journal of the American Statistical Association.* 1993;88(423):1013–1020.
10. Mercer A, Lau A, Kennedy C. For weighting online opt-in samples, what matters most?. *Pew Research Center.* 2018.
11. Little RJA. Post-stratification: A modeler's perspective. *Journal of the American Statistical Association.* 1993;88(423):1001–1012.
12. Elliott MR, Little RJA. Model-based alternatives to trimming survey weights. *Journal of Official Statistics.* 2000;16(3):191–210.
13. Little RJ. To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association.* 2004;99(466):546–556.
14. Battaglia MP, Hoaglin DC, Frankel MR. Practical considerations in raking survey data. *Survey Practice.* 2009;2(5):2953.
15. Gelman A. Poststratification into many categories using hierarchical logistic regression. *Survey Methodology.* 1997;23:127.
16. Lopez-Martin J, Phillips JH, Gelman A. Multilevel regression and poststratification case studies. https://bookdown.org/jl5522/MRP-case-studies/; 2021.
17. Kuh S, Kennedy L, Chen Q, Gelman A. Using leave-one-out cross-validation (LOO) in a multilevel regression and poststratification (MRP) workflow: A cautionary tale. *Statistics in Medicine.* 2024;43:953–982.
18. Dever JA, Valliant R. A comparison of variance estimators for poststratification to estimated control totals. *Survey Methodology.* 2010;36(1):45–56.
19. Gelman A. *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2007.
20. Brilleman SL, Elci EM, Novik JB, Wolfe R. Bayesian survival analysis using the rstanarm R package. *arXiv:2002.09633.* 2020.
21. Goodrich B, Gabry J, Ali I, Brilleman S. rstanarm: Bayesian applied regression modeling via Stan. https://mc-stan.org/rstanarm/; 2023. R package version 2.21.4.
22. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *Annals of Applied Statistics.* 2010;4:266–298.
23. Sparapani R, Spanbauer C, McCulloch RE. Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software.* 2021;97(1):1–66. doi: 10.18637/jss.v097.i01
24. Murray JS. Log-linear Bayesian additive regression trees for multinomial logistic and count regression models. *Journal of the American Statistical Association.* 2021;116(534):756–769.
25. Sparapani R, Spanbauer C, McCulloch R. Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R Package. *Journal of Statistical Software.* 2021;97(1):1–66. doi: 10.18637/jss.v097.i01
26. Hill J, Linero A, Murray J. Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application.* 2020;7:251–278.
27. Ware J, Kosinski M, Turner-Bowker D, Gandek B. How to score: version 2 of the SF-12v2 Health Survey. *Boston: Health Assessment Lab.* 2002;22(32.9):32–35.
28. Kosinski M, Ware JE, Turner-Bowker DM, Gandek B. *User's manual for the SF-12v2 health survey: with a supplement documenting the SF-12® Health Survey*, 2007.
29. Cranmer SJ, Gill J. We have to be discrete about this: A non-parametric imputation technique for missing categorical data. *British Journal of Political Science.* 2013;43(2):425–449.
30. Gill J, Cranmer SJ, Jackson N, Murr A, Armstrong D, Heuberger S. *hot.deck: Multiple hot deck imputation*. CRAN; https://CRAN.R-project.org/package=hot.deck: 2021. R package version 1.2.
31. Andridge RR, Little RJA. A review of hot deck imputation for survey non-response. *International Statistical Review.* 2010;78(1):40–64.
32. NYC Department of Health & Mental Hygiene. HIV in NYC: Statistics and Reports. HIV Surveillance Annual Report. https://www.nyc.gov/assets/doh/downloads/pdf/dires/hiv-surveillance-annualreport-2020.pdf; 2020.
33. NYC Department of Health & Mental Hygiene. Medical Monitoring Project (MMP). Results from the MMP survey of people receiving HIV care in 2013-2014 in NYC. https://www.nyc.gov/assets/doh/downloads/pdf/dires/mmp-report-2013-2014.pdf; .
34. Health Resources and Services Administration (HRSA) . Ryan White HIV/AIDS program annual client-level data report. https://ryanwhite.hrsa.gov/sites/default/files/ryanwhite/data/rwhap-annual-client-level-data-report-2019.pdf; 2019.
35. Zhang X, Holt JB, Yun S, et al. Multilevel small-area estimation of health behaviors: An extension of multilevel regression and poststratification (MRP) approach via bootstrapping. *JSM Proceedings, Survey Research Methods Section.* 2017.
36. Li K, Si Y. Embedded multilevel regression and poststratification: Model-based inference with incomplete auxiliary information. *Statistics in Medicine.* 2023;43(2):256—278.
37. Bishop Y, Fienberg S, Holland P. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA, 1972.
38. Brick JM, Montaquila J, Roth S. Identifying problems with raking estimators. *In JSM Proceedings, Survey Research Methods Section.* 2003.
39. Liu Y, Gelman A, Chen Q. Inference from nonrandom samples using Bayesian machine learning. *Journal of Survey Statistics and Methodology.* 2023;11(2):433–455.
40. Gelman A, Si Y, West BT. MRPW: Regression, poststratification, and small-area estimation with sampling weights. https://sites.stat.columbia.edu/gelman/research/unpublished/weight_regression.pdf; 2024.

TABLE 1 Results from the simulation study. Methods included for comparison are unadjusted, raking, MRP, and the 4 MRP adaptations. The results displayed are absolute value of bias $\times 100$, RMSE $\times 100$, and the coverage and average width of the 95% interval.

	Binary Outcome				Continuous Outcome			
Method	|Bias|	RMSE	Coverage and avg width of 95% interval		|Bias|	RMSE	Coverage and avg width of 95% interval	
Perfect-Overlap								
Unadjusted	25.03	25.03	0.0	5.13	306.06	306.06	0.0	32.96
Raking	0.31	3.31	92.2	15.32	3.23	7.66	91.8	34.65
MRP	2.68	3.59	86.4	13.77	0.40	6.51	94.0	31.77
MRP rake count	2.71	3.61	88.2	13.87	0.88	6.54	93.8	32.18
MRP Poisson	2.68	3.58	86.8	13.63	0.95	6.50	94.0	31.54
MRP NB	2.68	3.58	86.8	13.63	0.96	6.50	93.8	31.56
MRP BART	2.67	3.58	86.6	13.79	1.03	6.53	93.8	31.96
Semi-Overlap								
Unadjusted	13.30	13.30	0.0	6.27	133.26	133.26	0.0	30.67
Raking	0.02	3.34	92.2	15.26	0.61	7.42	94.0	34.38
MRP	1.54	2.72	93.6	12.31	0.94	5.68	95.2	28.25
MRP rake count	1.57	2.76	95.4	12.51	1.28	5.86	93.8	28.71
MRP Poisson	1.62	2.74	94.4	12.28	1.17	5.70	94.2	28.11
MRP NB	1.61	2.74	94.8	12.29	1.22	5.70	94.6	28.11
MRP BART	1.69	2.78	94.0	12.38	0.78	5.70	94.6	28.52
No-Overlap								
Unadjusted	1.04	1.48	93.0	6.47	8.63	9.46	78.6	29.80
Raking	0.07	1.91	93.6	9.15	0.35	4.57	93.2	20.96
MRP	0.08	1.45	96.6	7.83	0.19	3.58	96.0	18.20
MRP rake count	0.53	1.46	96.6	7.79	2.52	4.17	92.8	18.77
MRP Poisson	0.43	1.46	96.6	7.79	2.18	3.98	93.8	18.66
MRP NB	0.40	1.46	96.4	7.80	2.05	3.93	94.4	18.67
MRP BART	0.63	1.49	95.8	7.86	2.89	4.35	93.4	19.31

TABLE 2 Distribution of demographic and geographic factors in the target population and among individuals with HIV in the CHAIN data in New York City from 2018–21. The table also displays unadjusted descriptive statistics for the 3 outcome measures of interest. The population data are obtained from the NYC HIV/AIDS annual surveillance Statistics 2020, Medical Monitoring Project (MMP) NYC 2013–2014 data, and the Ryan White HIV/AIDS Program annual client-level data report 2019.

	Participant Source				**Unadjusted Outcome Measure**					
	Population		**CHAIN Sample**		**Viral Suppression**		**Mental Health (MCS)**		**Physical Health (PCS)**	
Poststratification Factor	N	%	N	%	N	%	Mean	SD	Mean	SD
Total	129061	–	504	–	424	93.6	40.3	8.8	49.7	10.3
Gender[1]										
Women	35320	27.4	158	31.3	128	91.4	40.5	9.1	46.7	10.6
Men	93741	72.6	346	68.7	296	94.6	40.2	8.7	51.1	9.8
Age Group										
[20, 30)	7535	5.8	88	17.5	70	93.3	37.8	7.9	53.0	7.6
[30, 40)	20840	16.1	185	36.7	153	92.7	39.3	8.6	52.4	9.6
[40, 50)	23049	17.9	84	16.7	71	94.7	40.2	9.8	49.2	10.7
[50, 60)	39359	30.5	94	18.7	85	93.4	42.6	7.7	46.0	9.9
[60, 100)	38278	29.7	53	10.5	45	95.7	44.1	9.3	42.6	11.2
Ethnicity[2]										
Black	55790	43.2	250	49.6	210	93.3	41.8	7.9	49.5	10.1
Latino	42740	33.1	199	39.5	170	94.4	38.6	9.5	50.1	10.6
White	26165	20.3	21	4.2	18	94.7	39.2	10.5	49.7	11.0
Other	4366	3.3	32	6.3	24	88.9	40.0	7.5	49.6	9.8
NYC Boroughs										
Bronx	33823	26.2	192	38.1	162	95.9	39.5	8.6	50.5	9.5
Brooklyn	33235	25.7	144	28.6	121	93.8	40.9	9.5	49.0	11.1
Manhattan	35383	27.4	91	18.1	78	89.7	41.4	8.2	47.6	11.2
Queens	21481	16.6	49	9.7	42	93.3	40.3	8.5	51.2	9.6
Staten Island	5139	4.0	28	5.6	21	91.3	39.6	8.6	52.4	8.4
Type of Exposure										
MSM[3]	58656	45.5	249	49.4	214	93.9	39.4	8.7	52.3	9.0
EverIDU[2]	17099	13.3	71	14.1	61	95.3	39.5	8.1	48.7	11.1
Perinatally acqiared	2552	2.0	5	1.0	4	80.0	37.2	4.9	50.0	8.1
Education[2]										
Less than HS	45042	34.9	170	33.7	139	94.6	40.8	9.2	47.9	11.1
High school	34588	26.8	238	47.2	198	93.4	40.4	8.5	50.1	9.9
More than HS	49431	38.3	94	18.7	85	92.4	39.2	8.7	52.1	9.4
Housing Status[4]										
Stable	101313	78.5	379	75.2	325	93.4	40.8	8.9	49.5	10.3
Temporary	13810	10.7	54	10.7	42	91.3	38.8	7.9	51.9	8.2
Unstable	13938	10.8	60	11.9	48	96.0	38.7	8.6	49.7	11.5
Poverty Status[5]										
0–100% FPL	91763	71.1	288	57.1	242	92.4	39.9	8.7	49.5	10.1
101–138% FPL	17165	13.3	75	14.9	66	98.5	41.2	8.7	48.3	10.5
139–250% FPL	13293	10.3	58	11.5	50	90.9	42.9	9.4	47.8	11.6
251–400% FPL	4646	3.6	41	8.1	38	100	38.8	7.4	54.7	7.6
> 400% FPL	2194	1.7	15	3.0	11	78.6	38.4	8.3	55.5	6.5

Abbreviations: FPL: Federal Poverty Level; IDU: Injection Drug Use; MSM: Men Who Have Sex with Men; SD:Standard Deviation.

[1]The Trans individuals are included in the Women group; [2] There are 2 unknown; [3] There are 12 unknown; [4] There are 11 unknown; [5] There are 27 unknown.

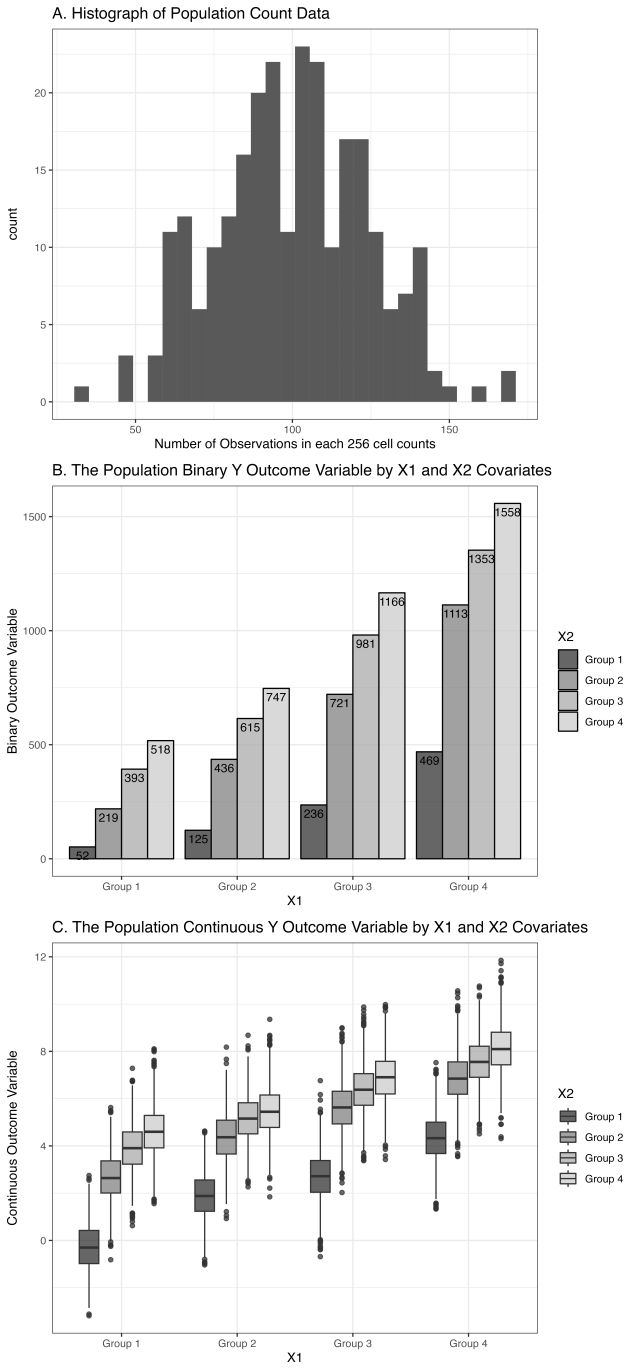

FIGURE 1 Population data in the simulation study. Plot A is the histogram of the population sizes N_j across the 256 subgroups formed by $X_1, \ldots, X_4$. Plot B shows the frequencies of the binary outcome y by covariates X_1 and X_2. Plot C is the box-plot of the continuous y outcome by covariates X_1 and X_2.

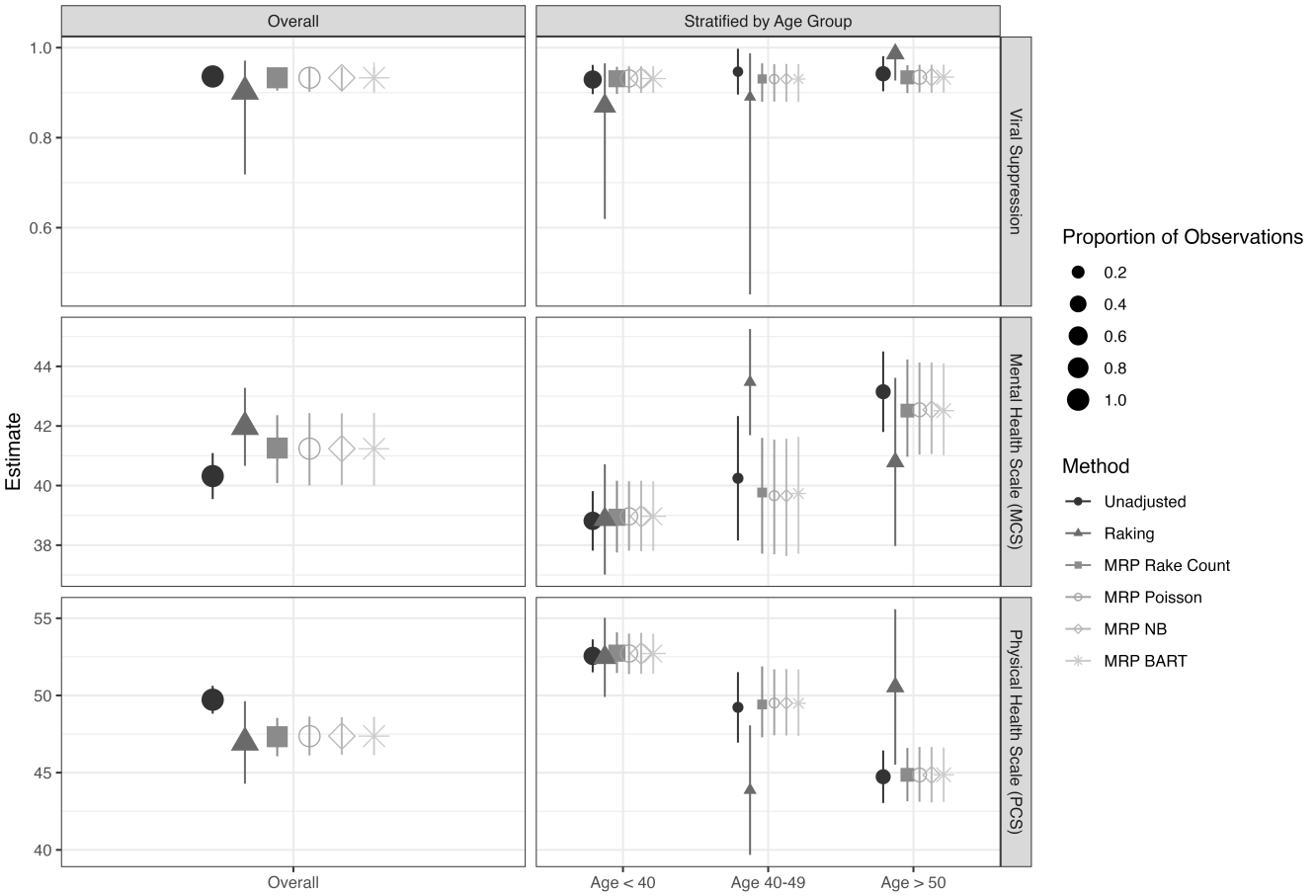

FIGURE 2 The point estimates and 95% intervals for population proportion of viral suppression and population means of mental and physical health scales, both overall and by age groups, in the CHAIN application. The methods used are the unadjusted estimate, raking, and the four MRP adaptations. The size of the point estimates represents the relative proportion of observations in each group.