Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

Improving Survey Inference Using Administrative Records Without Releasing Individual-Level Continuous Data

Sharifa Z. Williams[1,2,3] | Jungang Zou[1] | Yutao Liu[1] | Yajuan Si[4] | Sandro Galea[5] | Qixuan Chen[1,3]

[1]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York, USA | [2]Edward J. Bloustein School of Planning and Public Policy, Rutgers University, New Brunswick, New Jersey, USA | [3]Center for Research on Cultural and Structural Equity, Nathan S. Kline Institute for Psychiatric Research, Orangeburg, New York, USA | [4]Institute for Social Research, University of Michigan, Ann Arbor, Michigan, USA | [5]School of Public Health, Boston University, Boston, Massachusetts, USA

Correspondence: Sharifa Z. Williams (sharifa.williams@rutgers.edu)

Received: 22 May 2023 | **Revised:** 21 August 2024 | **Accepted:** 19 October 2024

Funding: This study was supported by Mailman School of Public Health, Columbia University (Initiative for Maximizing Student Diversity), National Science Foundation (SES2217456) and National Institutes of Health (R01AG067149, R01ES035784, R01MD12248 and U01MD017867).

Keywords: Bayesian predictive inference | continuous auxiliary variables | generalized additive model | inclusion propensity | poststratification | Rstan

ABSTRACT

Probability surveys are challenged by increasing nonresponse rates, resulting in biased statistical inference. Auxiliary information about populations can be used to reduce bias in estimation. Often continuous auxiliary variables in administrative records are first discretized before releasing to the public to avoid confidentiality breaches. This may weaken the utility of the administrative records in improving survey estimates, particularly when there is a strong relationship between continuous auxiliary information and the survey outcome. In this paper, we propose a two-step strategy, where the confidential continuous auxiliary data in the population are first utilized to estimate the response propensity score of the survey sample by statistical agencies, which is then included in a modified population data for data users. In the second step, data users who do not have access to confidential continuous auxiliary data conduct predictive survey inference by including discretized continuous variables and the propensity score as predictors using splines in a Bayesian model. We show by simulation that the proposed method performs well, yielding more efficient estimates of population means with 95% credible intervals providing better coverage than alternative approaches. We illustrate the proposed method using the Ohio Army National Guard Mental Health Initiative (OHARNG-MHI). The methods developed in this work are readily available in the R package `AuxSurvey`.

1 | Introduction

Probability samples play an important role in survey research, facilitating inference about health measures in large finite populations using moderately sized samples. However, probability surveys have suffered increasingly high nonresponse rates in the past several decades [1–5], which makes these probability surveys often nonrepresentative, challenging the validity of survey inference. Auxiliary information about the target population can be used to improve survey inference [6]. Such data can be obtained from sources such as administrative records.

When only discrete auxiliary variables are available, poststratification or raking can be used to weight the sample to reduce bias in the survey estimation [4, 5, 7, 8]. Poststratification requires the joint population distributions of the auxiliary variables to be known, whereas raking utilizes the more commonly available marginal population distributions [7, 8]. Although

https://doi.org/10.1002/sim.10270

poststratification and raking are easy to implement, the resultant weights can be highly variable, and consequently, the weighted estimators can be unstable [9–13]. Predictive inference represents an alternative framework for correcting nonresponse bias, offering the advantage of improved efficiency in survey inference [9, 10, 12–15]. Multilevel regression and poststratification (MRP) is a widely used model-based alternative to the described weighted estimation approaches with important applications in social and political sciences [14, 16]. MRP models survey outcomes using a multilevel regression and can generate accurate survey estimates even from nonprobability samples [17]. Although MRP has an advantage in improving efficiency in survey estimation over poststratification and raking, it requires all the auxiliary variables to be discrete.

When administrative records contain individual-level continuous auxiliary variables, MRP models cannot be directly applied. Instead, prediction models that allow flexible associations between survey outcomes and these continuous auxiliary variables are desirable. Liu et al. proposed a regularized prediction approach using soft Bayesian Additive Regression Trees (BART) [6], which predicts survey outcomes in the population using detailed individual-level data from administrative records. This approach has been shown to effectively reduce bias and improve efficiency in survey estimates. However, individual-level continuous administrative records are often inaccessible due to confidentiality concerns [18, 19]. With discrete auxiliary variables, only frequency tables of the population data need to be released to the public to allow weighting or model-based adjustments. In contrast, continuous auxiliary variables require individual-level population information to facilitate prediction or weighting. When confidentiality concerns prevent the release of individual-level data, continuous auxiliary variables are often discretized using percentiles or meaningful cutoffs before making the population data available. Using discretized versions of continuous variables may lead to efficiency loss in predictive survey inference, particularly when there is a strong relationship between the survey outcome and the continuous variables. Therefore, we seek a data analysis strategy that makes the best use of continuous information while avoiding the need to release individual-level continuous variables.

In this paper, we propose a two-step strategy for predictive survey inference that maximizes the use of continuous data information in the population while controlling disclosure risk. In the first step, statistical agencies with access to confidential administrative records link survey data with administrative records and estimate inclusion propensities using both continuous and discrete auxiliary variables. They then create a modified population data set by replacing individual-level continuous data with continuous inclusion propensity scores and discretized versions of the continuous variables. In the second step, data users, who do not have access to the confidential records, perform predictive survey inference using the survey sample and the modified population data. This division of tasks ensures that statistical agencies handle the confidential data while data users conduct survey inference with the modified population data. We conduct simulation studies to evaluate the performance of this two-step strategy compared with alternative methods. We also illustrate this two-step strategy using the Ohio Army National Guard Mental Health Initiative (OHARNG-MHI) Survey, estimating the percentage of lifetime alcohol abuse among all service members in the Ohio National Guard in 2008. Our R package, AuxSurvey, is available on Github: https://github.com/zjg540066169/AuxSurvey and provides researchers with a user-friendly interface for conducting analyses we discuss in this paper.

2 | Methods

2.1 | Notation and Background

We consider a target population consisting of N units with survey outcome Y. Let $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$ denote p discrete auxiliary variables and $\mathbf{X} = (X_1, \ldots, X_q)^T$ be q continuous auxiliary variables, known for the population. For simplicity, we let $q = 1$. The continuous variable X is usually not publicly available due to confidentiality concerns. Instead, before releasing to the public, X is discretized, denoted with X^*, using percentiles or meaningful cutoffs. The population can be partitioned into J disjoint and exhaustive cells or poststrata defined by the joint distribution of $(\mathbf{Z}, X^*)$, with N_j units in cell j, $j = 1 \ldots J$ and $\sum_{j=1}^{J} N_j = N$ where $N_j > 0$. With $\overline{Y}_j = \sum_{i=1}^{N_j} y_i / N_j$, the population mean of Y within cell j, the overall population mean $\overline{Y}$ can be written as,

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

Let s denote a probability survey sample of size n selected from the population with survey outcome values, $y_1, \ldots, y_n$. We can then use $(\mathbf{Z}, X^*)$ to divide the sample into J cells; the corresponding sample size in cell j is n_j with $\sum_{j=1}^{J} n_j = n$ and the sample mean of Y in cell j is $\overline{y}_j = \sum_{i=1}^{n_j} y_i / n_j$. Assuming that $\overline{y}_j$ is an unbiased estimate of $\overline{Y}_j$, the poststratification estimator 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 sample unit i in cell j, $j = 1, \ldots, J$.

Alternatively, w_i in formula (2) can be created using the raking method when only population margins of $(\mathbf{Z}, X^*)$ are available. Raking weights are obtained via an iterative proportional fitting procedure that begins by adjusting design weights to the marginal distribution of the first auxiliary variable. These adjusted weights are then updated to conform to the marginal distribution of the second auxiliary variable. This process of updating the adjusted weights is carried out for each auxiliary variable and the first iteration ends when the weights are adjusted using the last auxiliary variable. Subsequent iterations are performed until the weights conform to the marginal distributions of all the auxiliary variables, that is, the algorithm converges [7]. The Newton–Raphson method can also be used to obtain the same set of weights [8].

When $(\mathbf{Z}, X^*)$ are related to survey outcome Y and the distributions of $(\mathbf{Z}, X^*)$ differ between sample and population, the poststratification and raking methods can effectively reduce bias in the estimation of θ, by assigning a weight w_i for each sample unit. However, the weights can be highly variable, especially when p+q is large and n_j is small for some j, $j = 1, \ldots, J$. Consequently, the weighted estimator can be unstable. Alternatively, model-based approaches that predict the nonsampled units in the population based on a prediction model of the sample units can be used to improve efficiency in the θ estimation.

The MRP model [14] has been widely used as a model-based alternative to poststratification. MRP uses a hierarchical regression model to model the survey outcome of interest given the auxiliary variables,

$$g(E(Y_i|Z_i, X_i^*)) = \alpha_0 + \sum_{k=1}^{p_1}\alpha_k Z_{ik} + \sum_{l=p_1+1}^{p}\alpha_{m[i]}^{Z_l} + \alpha_{m[i]}^{x^*} \tag{3}$$

where $g(\cdot)$ is a link function, such as using an identity link for continuous Y and a logit link for binary Y; α_0 is the intercept; α_k is the slope associated with the binary auxiliary variable Z_k, $k = 1, \ldots, p_1$ and $p_1 \leq p$; and $\alpha_{m[i]}^{Z_l}$ and $\alpha_{m[i]}^{x^*}$ correspond to varying coefficients associated with polytomous variables Z_l, $l = (p_1 + 1), \ldots, p$, and the discretized continuous variable X^*, respectively, where $m[i]$ indicates the category of the polytomous variable to which the ith unit belongs. The varying coefficients $\alpha_{m[i]}^{Z_l}$ and $\alpha_{m[i]}^{x^*}$ are given independent normal prior distributions, $\alpha_{m[i]}^{Z_l} \overset{iid}{\sim} N(0, \tau_l^2)$ and $\alpha_{m[i]}^{x^*} \overset{iid}{\sim} N(0, \tau_{x^*}^2)$, and the variance components $(\tau_{(p_1+1)}^2, \ldots, \tau_p^2, \tau_{x^*}^2)$ are assigned a hyper-prior distribution each using $\sim$ inv-$\chi^2(\nu, \sigma_0^2)$ with a weakly informative or noninformative prior for ν and σ_0. Using Bayesian simulations, a posterior draw of θ is then obtained by replacing $\overline{Y}_j$ in (1) with $\widehat{E}(Y_j|\mathbf{Z}, X^*)^{(d)}$, the dth draw from the posterior distribution of $E(Y|\mathbf{Z}, X^*)$ for units in cell j, with $d = 1, \ldots, D$,

$$\widehat{\theta}^{(d)} = \frac{\sum_{j=1}^{J} N_j \widehat{E}(Y_j|\mathbf{Z}, X^*)^{(d)}}{\sum_{j=1}^{J} N_j} \tag{4}$$

The posterior mean or median of $\widehat{\theta}^{(d)}$ serves as the estimate for θ.

The hierarchical structure in MRP partially pools estimates in poststrata formed by the auxiliary variables [17, 20, 21], and improves the estimation of $\overline{Y}_j$ in cells with sparse data by borrowing data information from other poststrata. It also improves efficiency in the estimation of population means in the presence of dispersed weights when models are well constructed [11, 13]. The model in (3) can also be extended to include two-way or higher order interactions between auxiliary variables [22, 23].

2.2 | Generalized Additive Model of Inclusion Propensity

When the administrative records contain continuous auxiliary variables, discretizing the continuous variables may result in the loss of important information, especially when there are strong smooth relationships between the continuous auxiliary variables and the survey outcome. In the missing data literature, Little and An [24] showed that a penalized spline of propensity prediction model can improve mean estimation when imputation models are misspecified, where the logit-transformed estimated response propensity is included in the model using a spline. We borrow this modeling idea. In addition to including $\mathbf{X}^*$ as covariates in the prediction model of the survey outcome, if the sample inclusion propensity is available, we can also include the logit-transformed inclusion propensity as a covariate in the prediction model. With the sample inclusion propensity in conjunction with discretized variables, we aim to best utilize the continuous auxiliary information when access to the individual-level continuous data is not feasible. Thus, we propose a two-step strategy. In the first step, statistical agencies who have access to the confidential administrative records link survey sample with administrative records and estimate the inclusion propensity for both sample and nonsampled units using the continuous and discrete auxiliary variables measured in the population. They then create a modified population data set by replacing the continuous variables $\mathbf{X}$ with $\mathbf{X}^*$ and the estimated inclusion propensity. In the second step, the data users, who do not have access to administrative records, utilize this modified population data together with the sample data for survey inference.

The sample inclusion propensity can be estimated using a propensity model [25–27]. Let the sample inclusion indicator I be coded as 1 for the units in the sample and 0 for the rest of the units in the population. The estimated inclusion propensity, denoted by $\hat{\pi}_i$ for unit i in the population, can be obtained via fitting a logistic or probit regression of I on the discrete auxiliary variables $\mathbf{Z}$ and the original continuous variables $\mathbf{X}$ by linking survey sample with administrative records. When the number of auxiliary variables is small, parametric or semi-parametric regression models can be used to estimate the propensity score. When there exist high-dimensional auxiliary data, machine learning methods such as binary BART are recommended [6].

The estimated sample inclusion propensity can be included in the modified population data together with the discrete auxiliary variables, and serve as a covariate in the prediction model for Y. Despite the advantages of MRP, the varying coefficients of $\alpha_{m[i]}^{x^*}$ in the MRP model in (3) assume an exchangeable prior distribution. Because X^* is discretized from the continuous X, the ordering of the categories of X^* may matter. When there is a smooth relationship between X^* and Y, an exchangeable prior distribution is not sufficient. Therefore, we extend the MRP model in (3) to allow a smooth relationship of Y with both X^* and logit$(\hat{\pi})$,

$$\begin{aligned} g(E(Y_i|\mathbf{Z}_i, X_i^*)) = \alpha_0 + \sum_{k=1}^{p_1}\alpha_k Z_{ik} + \sum_{l=p_1+1}^{p}\alpha_{m[i]}^{Z_l} \\ + s_1(x_i^*) + s_2(\text{logit}(\hat{\pi}_i)) \end{aligned} \tag{5}$$

where $s_1(x_i^*)$ and $s_2(\text{logit}(\hat{\pi}_i))$ are smooth functions of x^* and logit$(\hat{\pi}_i)$, respectively. These smooth functions allow flexible associations between the specified covariates and the survey outcome and thus protect against potential model misspecification.

When the association between X and Y varies between different levels of a categorical variable, say Z_1 with H levels, model (5) can be extended to reflect the interaction effect by allowing different smooth functions of X^* and logit$(\hat{\pi})$ in each level of Z_1,

$$\begin{aligned} g(E(Y_i|\mathbf{Z}_i, X_i^*)) = \alpha_0 + \sum_{k=1}^{p_1}\alpha_k Z_{ik} + \sum_{l=p_1+1}^{p}\alpha_{m[i]}^{Z_l} \\ + \sum_{h=1}^{H} s_{1h}(x_i^*) + \sum_{h=1}^{H} s_{2h}(\text{logit}(\hat{\pi}_i)) \end{aligned} \tag{6}$$

where $s_{1h}(x_i^*)$ and $s_{2h}(\text{logit}(\hat{\pi}_i))$ are smooth functions for x^* and $\text{logit}(\hat{\pi}_i)$ in category h of Z_1.

The smooth functions can be modeled using spline or kernel functions. In this article, we use a smooth spline to model each of these associations. We execute the Bayesian generalized additive model (GAM) in (5) and (6) using the `stan_gamm4()` function in `rstanarm`, an R package that estimates models using RStan [28]. RStan is an `R` interface to Stan for obtaining Bayesian inference using the No-U-Turn Sampler, a variant of Hamiltonian Monte Carlo [29, 30]. The `stan_gamm4()` function fits the specified GAM by adding priors on the hyperparameters of smooth splines, which is different from performing (restricted) maximum likelihood estimation with the `lme4` package in `R`. Bayesian estimation provides better estimates for the uncertainty in the parameter estimates. We monitor the convergence of our parameter estimates using the convergence measure $\hat{R}$ that suggests the chains mix well if close to 1.

Models (5) and (6) are used to yield predictions for Y among the nonsampled units in the population [11, 13, 14]. Our model-based predictive estimator of θ is written as

$$\tilde{\theta}^m = N^{-1}\left(\sum_{i\in s} y_i + \sum_{i\notin s}\hat{y}_i\right) = N^{-1}\left(\sum_{i=1}^{N}\hat{y}_i + \sum_{i\in s}(y_i-\hat{y}_i)\right) \qquad (7)$$

where y_i denotes the observed Y in the sample, and $\hat{y}_i$ denotes the predicted Y based on the prediction model. The posterior distribution of θ is simulated by generating a large number of draws using the predictive estimator, with the median of these draws defining the generalized additive model of propensity (GAMP) estimator.

3 | Simulation Study

3.1 | Design

We simulate a population of size $N = 3{,}000$, and generate three independent binary variables $\mathbf{Z} = (Z_1, Z_2, Z_3)$ with the marginal probabilities of $(0.7, 0.5, 0.4)$ and one continuous variable X with a standard normal distribution. We consider two survey outcomes generated using additive nonlinear models. The first is a continuous outcome with $Y_1 \sim \text{N}(15 + 2.5Z_1 - Z_2 + Z_3 - 2X + 3.75X^2, 3)$. The second is a binary outcome Y_2 generated via $\text{logit}\left(Pr(Y_2 = 1)\right) = -2.5 + 0.75Z_1 - 2.5Z_2 + 1.5Z_3 - 0.25X + 1.5X^2$. Both models show a smooth association between X and survey outcomes. We then select a sample with approximate $n = 600$ cases from the population. We repeat this simulation process 500 times. We compare the performance of various estimators under three settings using absolute empirical bias, root mean squared error (RMSE), and average width and coverage rate of the 95% confidence or credible interval (CI). Credible intervals for Bayesian methods are calculated using the highest probability density method.

Let $\hat{\theta}_t$ be an estimate of the population mean θ_t in the t^{th} simulation, $t = 1, \dots, 500$. The absolute empirical bias and RMSE are defined as follows,

$$\text{Absolute bias} = \left|\frac{1}{500}\sum_{t=1}^{500}(\hat{\theta}_t - \theta_t)\right|,$$

$$\text{RMSE} = \sqrt{\frac{1}{500}\sum_{t=1}^{500}(\hat{\theta}_t - \theta_t)^2}$$

Estimators with smaller absolute bias, smaller RMSE, shorter 95% CI, and coverage rate closer to the nominal level are desired.

3.2 | Setting One: *X* is Associated With Both Outcome *Y* and Inclusion *I*

In this setting, we consider samples selected via an inclusion model, $\text{logit}(\pi) = -1.25 - Z_1 + 1.25Z_2 - 0.75Z_3 + 0.75X - 0.10X^2$, so that X is associated with both the survey outcomes Y_1 and Y_2 and the sample inclusion probability π. The propensity scores are estimated by fitting a binary BART model with predictors $\mathbf{Z}$ and X by linking the survey sample with the corresponding population data. BART is a sum-of-trees machine learning model [31]. It is less sensitive to model misspecification and allows for nonlinear effects and multi-way interactions between auxiliary variables and outcomes of interest. The estimated propensity scores $\hat{\pi}$ are the posterior mean of the predictive probability of inclusion.

We estimate population means for the overall population and for the subset where $Z_1 = 1$. We compare the performance of various estimators, including an unweighted estimator using sample mean, eight weighted estimators, and four prediction model-based estimators. The eight weighted estimators are

- IPW and eIPW: the inverse propensity weighted estimators. IPW uses true propensity scores π, while eIPW uses the estimated propensity scores $\hat{\pi}$ from BART.
- PostStrat_3 and ($\text{eIPW} + \text{PostStrat}_3$): poststratification using the population joint distributions of $\mathbf{Z}$ and X_3^* (the discretized X using population tertiles). ($\text{eIPW} + \text{PostStrat}_3$) applies poststratification on the top of the base weights constructed using the inverse of the estimated propensity score from BART.
- Raking_5 and ($\text{eIPW} + \text{Raking}_5$): raking using the population margins of $\mathbf{Z}$ and X_5^* (the discretized X using population quintiles). ($\text{eIPW} + \text{Raking}_5$) applies raking on the top of the base weights constructed using the inverse of the estimated propensity score from BART.
- Raking_{10} and ($\text{eIPW} + \text{Raking}_{10}$): raking using the population margins of $\mathbf{Z}$ and X_{10}^* (the discretized X using population deciles) and raking on the top of the base weights.

Poststratification using $\mathbf{Z}$ and X_5^* or X_{10}^* can lead to sparse or zero poststratification cells in the sample and thus is not considered here. Both the unweighted and weighted estimates are obtained using the `survey` package in `R`, in which the finite population correction is incorporated in variance estimation. We also consider four prediction models implemented using the `rstanarm` package in `R`:

TABLE 1 | Comparison between absolute bias (×100), root mean squared error (RMSE ×100), average interval width (×100), and coverage rate of 95% CI (×100) for the 13 estimators from setting one.

	Continuous outcome Y_1				**Binary outcome Y_2**			
Estimators	**Bias**	**RMSE**	**Width**	**Coverage**	**Bias**	**RMSE**	**Width**	**Coverage**
Sample mean	146.53	148.05	81.68	0	9.73	9.86	5.9	0
Weighting								
IPW	0.32	71.43	207.87	84.6	0.02	3.49	11.98	91.2
eIPW	34.95	50.56	158.41	78.0	0.89	3.01	11.91	94.6
PostStrat$_3$	52.89	60.46	90.38	38.6	1.79	3.29	8.02	78.4
eIPW + PostStrat$_3$	38.17	47.87	99.86	59.8	1.29	3.05	8.45	83.2
raking$_5$	34.69	43.11	91.16	61.6	1.32	2.76	8.54	89.8
eIPW + raking$_5$	25.41	37.82	110.89	77.6	0.61	2.6	9.56	94.4
raking$_{10}$	18.10	29.76	82.97	80.2	0.64	2.45	8.28	91.8
eIPW + raking$_{10}$	16.64	29.82	91.61	84.0	0.23	2.51	8.94	93.4
Predictive estimators								
True model	0.41	15.79	61.4	94.4	0.02	2.06	7.63	93.6
MRP$_{10}$	21.51	31.28	87.44	81.6	1.20	2.48	8.36	91.2
GAMP$_{10}$	8.02	38.35	96.4	80.8	0.02	2.25	8.07	93.0
eGAMP$_{10}$	20.56	31.60	97.0	86.4	0.28	2.28	8.27	93.6

- True Model: the true outcome model.
- MRP$_{10}$: MRP model in (3) using $\mathbf{Z}$ and X^*_{10}.
- GAMP$_{10}$ and eGAMP$_{10}$: GAMP model in (5) using $\mathbf{Z}$, X^*_{10}, and logit propensity scores. GAMP$_{10}$ uses logit(π) while eGAMP$_{10}$ uses logit($\hat{\pi}$).

Table 1 presents the simulation results for this setting. The sample mean, which ignores the unequal probability of inclusion, performs poorly, exhibiting large bias and RMSE, with the 95% CI yielding a zero coverage rate. The eight weighted estimators demonstrate much smaller bias and RMSE compared with the sample mean. Among the eight weighted estimators, IPW yields the smallest bias but the largest RMSE. Using the estimated propensity score increases bias but reduces RMSE, which is expected due to the lower variation in the estimated propensity score compared with the true propensity score. Applying poststratification and raking on top of the base weights constructed with estimated propensity scores results in smaller bias and RMSE compared with poststratification and raking without considering the base weights for PostStrat$_3$ and raking$_5$, but the improvement is small for raking$_{10}$. Additionally, raking estimators using X^*_{10} perform better than those using X^*_5 and yield the smallest RMSE among all weighted estimators.

As expected, the predictive estimator using the true outcome model performs best, with the smallest bias, RMSE and average width, and coverage rate close to the nominal level. The MRP$_{10}$ performs similarly to raking$_{10}$. The proposed GAMP$_{10}$ yields smaller bias but a wider 95% CI compared with MRP$_{10}$ for Y_1 and achieves both smaller bias and a shorter 95% CI for Y_2. When replacing the true response propensity with the estimated one, the bias increases but the RMSE decreases and coverage rate of 95% CI improves for Y_1 while the results for Y_2 remain largely unchanged. The 95% CI coverage rates are significantly below the nominal level for all methods except the true model for Y_1. This occurs because the continuous covariate X has a strong smooth relationship with both Y_1 and inclusion I, but discretizing X into quintiles or deciles across methods leads to information loss.

The results of the subgroup analysis are presented in Supporting Information eTable 1. The conclusions are similar to the overall population analysis, except that the model-based estimators now show greater efficiency gains compared with the weighted estimators.

3.3 | Setting Two: X is Associated With Y but Not I

In the second setting, we simulate one additional independent continuous auxiliary variable from a standard normal distribution, $W \sim N(0, 1)$, in the population. Then, samples are selected using inclusion model, $\text{logit}(\pi) = -1.25 - Z_1 + 1.25Z_2 - 0.75Z_3 + 0.75W - 0.1W^2$. In this setting, predictors $\mathbf{Z}$, W, X are used to fit BART to estimate propensity scores $\hat{\pi}$. Note that W is associated with the sample inclusion probability π but not the survey outcomes Y_1 or Y_2, and X is associated with the survey outcomes but not the sample inclusion.

We compare the same thirteen estimators as setting one. All the weighted and predictive estimators, except for the true model estimator, use all available auxiliary information of ($\mathbf{Z}, X^*, W^*$). Specifically, PostStrat$_3$ uses ($\mathbf{Z}$, X^*_3, W^*_3), raking$_5$ uses ($\mathbf{Z}$, X^*_5, W^*_5), and raking$_{10}$ uses ($\mathbf{Z}$, X^*_{10}, W^*_{10}), where W^*_3, W^*_5 and W^*_{10} are the discretized W using population tertiles, quintiles, and deciles, respectively. The MRP$_{10}$ uses ($\mathbf{Z}$, W^*_{10}, X^*_{10}), whereas GAMP$_{10}$ and eGAMP$_{10}$ also include logit(π) and logit($\hat{\pi}$), respectively.

TABLE 2 | Comparison between absolute bias (×100), root mean squared error (RMSE ×100), average interval width (×100), and coverage rate of 95% CI (×100) for the 13 estimators from setting two.

	Continuous outcome Y_1				**Binary outcome Y_2**			
Estimators	**Bias**	**RMSE**	**Width**	**Coverage**	**Bias**	**RMSE**	**Width**	**Coverage**
Sample mean	70.2	73.82	94.08	16.8	9.28	9.42	5.95	0
Weighting								
IPW	0.94	40.72	138.26	92	0.04	3.26	11.11	92.6
eIPW	7.75	31.27	132.49	96.2	0.05	2.87	11.42	94.4
PostStrat$_3$	3.6	31.62	85.99	81.6	0.99	3.01	6.45	72
eIPW + PostStrat$_3$	8.63	29.43	83.82	84.6	1.12	3.03	6.42	72.2
raking$_5$	0.77	28.46	98.08	92	0.11	2.4	8.3	91
eIPW + raking$_5$	4.97	26.99	102.48	92.4	0.002	2.53	9.16	92.4
raking$_{10}$	0.76	24.33	84.45	91.2	0.17	2.38	8.07	90.6
eIPW + raking$_{10}$	3.07	23.67	87.04	92.6	0.01	2.55	8.74	90.6
Predictive estimators								
True model	0.07	13.4	52.76	95.4	0.01	1.8	6.75	95.8
MRP$_{10}$	0.79	19.56	82.81	97	0.3	1.94	7.48	94.8
GAMP$_{10}$	1.1	22.35	89.78	95.6	0.17	2.09	7.77	93.6
eGAMP$_{10}$	2.54	20.85	95.07	97.6	0.16	2.11	7.96	93.8

Table 2 presents the results for estimators on the overall population. The results for the subgroup analysis are shown in Supporting Information eTable 2. Similar to the first setting, the sample mean performs poorly. Among the eight weighted estimators, IPW yields the smallest bias but the largest RMSE. Using estimated response propensity, eIPW results in larger bias but smaller RMSE compared with IPW. The raking$_5$ shows a smaller bias and RMSE and a 95% CI coverage rate closer to the nominal level than PostStrat$_5$. However, unlike setting one, raking$_{10}$ does not lead to significant improvements than raking$_5$, with improvements only seen in RMSE for Y_1. Applying poststratification and raking on the top of the base weights constructed using the inverse of propensity score does not lead to clear improvements compared with those ignoring the base weights, resulting in reduced RMSE but larger bias for Y_1 and reduced bias but larger RMSE for Y_2.

All the four prediction model-based estimators have lower bias and reduced RMSE than the weighted estimators for both Y_1 and Y_2. The inclusion of W in this setting increases the variation in the sample weights and thus the uncertainty in the weighted estimates, but the impact of W is relatively small in the prediction model-based estimators. Because only W (and not X) is related to π, the estimated inclusion propensity $\hat{\pi}$ does not contain useful information about X. As such, the GAMP$_{10}$ and eGAMP$_{10}$ that incorporate the inclusion propensities does not offer improvements over MRP$_{10}$. All model-based estimators yield 95% CIs with coverage rate close to the nominal level.

3.4 | Setting Three: Nonadditive Association With Both Y and I

In the third simulation setting, we consider two survey outcomes generated from models with nonadditive association between X and Y. The first is a continuous outcome $Y_1 \sim \mathrm{N}(15 + 2.5Z_1 - Z_2 + Z_3 - 2X + X^2 + Z_1 * X - 2.5Z_1 * X^2, 2)$. The second is a binary outcome generated via $\mathrm{logit}\big(Pr(Y_2 = 1)\big) = -1.75 + 0.75Z_1 - 1.5Z_2 + 1.5Z_3 - 1.5X + X^2 + Z_1 * X - 2.5Z_1 * X^2$. Then, samples are selected using inclusion model $\mathrm{logit}(\pi) = -0.9 - 0.5Z_1 + 0.75Z_2 - Z_3 + 0.5X - 0.05X^2 + 0.5Z_1 * X - 0.75Z_1 * X^2$, which also includes nonadditive association between X and π.

We consider the similar weighted and predictive estimators as before. For raking, we use one-variable margins for Z_2 and Z_3 and two-variable margins for Z_1 and X^*_5 to model the interaction between Z_1 and X. The raking estimators using X^*_{10} are not included due to the potential for sparse samples in the two-variable margins for Z_1 and X^*_{10}. For MRP$_{10}$, we include interactions between Z_1 and X^*_{10}. For GAMP$_{10}$, the interaction between Z_1 and $\mathrm{logit}(\pi)$ is also included. Similarly, interaction between Z_1 and $\mathrm{logit}(\hat{\pi})$ is specified for eGAMP$_{10}$. To further assess whether the predictive estimators are sensitive to misspecification in the prediction models, we also compare eGAMP$_{10}$ to two alternative eGAMP models that either omit the interaction between Z_1 and X^*_{10} or omit both interactions and compare MRP$_{10}$ to the alternative that ignores the interaction between Z_1 and X^*_{10} for the continuous outcome Y_1.

Table 3 presents the results for this setting. The results for the subgroup analysis are available in Supporting Information eTable 3. The results on the sample mean and weighted estimators are similar to the first setting. However, unlike the first setting, the predictive estimators show significant improvements over the weighted estimators, with much smaller bias, lower RMSE, and 95% CI coverage rates closer to the nominal level. Compared with MRP$_{10}$, GAMP$_{10}$ and eGAMP$_{10}$ result in lower bias, improved efficiency, and closer to nominal level coverage, especially for the continuous outcome Y_1.

TABLE 3 | Comparison between absolute bias (×100), root mean squared error (RMSE ×100), average interval width (×100), and coverage rate of 95% CI (×100) for the 11 estimators from setting three.

	Continuous outcome Y_1				Binary outcome Y_2			
Estimators	**Bias**	**RMSE**	**Width**	**Coverage**	**Bias**	**RMSE**	**Width**	**Coverage**
Sample mean	39.29	40.65	39.12	3.2	4.63	4.85	5.33	11.8
Weighting								
IPW	4.87	38.52	89.67	85.2	0.41	2.95	9.7	93.4
eIPW	26.71	30.26	60.41	59	2	3.13	9.41	87.2
PostStrat_3	33.96	36.47	40.58	12.6	1.42	2.91	7.27	75.2
PostStrat_3 eIPW	28.65	31.93	43.99	29.6	1.45	2.99	7.05	73.8
raking_5	26.9	30.25	47.18	38.8	0.85	2.5	8.16	91.4
eIPW + raking_5	22.23	27.13	53.67	57.8	0.91	2.73	8.77	91.4
Predictive estimators								
True model	1.13	14.11	53.69	94.6	0.23	1.87	7.61	95.8
MRP_{10}	17.05	21.06	58.73	83.4	0.77	1.99	8.49	97.2
GAMP_{10}	8.22	18.32	75.84	95.6	0.49	2.05	9.29	97
eGAMP_{10}	1.99	14.19	64.96	96.2	0.37	2.09	9.19	96.6

TABLE 4 | Comparison between absolute bias (×100), root mean squared error (RMSE ×100), average interval width (×100), and coverage rate of 95% CI (×100) to assess whether the GAMP estimator is sensitive to misspecification in the predictive model and compare to MRP and raking that omit the interaction effect of Z_1 and X^* using the data in simulation setting three.

	Continuous outcome Y_1			
Estimators	**Bias**	**RMSE**	**Width**	**Coverage**
raking_5 (omit interaction $Z_1 \times X_5^*$)	51.06	52.87	55.21	6.0
raking_5 (include interaction $Z_1 \times X_5^*$)	26.90	30.25	47.18	38.8
MRP_{10} (omit interaction $Z_1 \times X_{10}^*$)	69.03	70.6	49.02	0
MRP_{10} (include interaction $Z_1 \times X_{10}^*$)	17.05	21.06	58.73	83.4
eGAMP_{10} (omit interaction $Z_1 \times X_{10}^*$)	14.60	18.20	42.10	72.0
eGAMP_{10} (omit both interactions)	38.2	42.23	68.58	39.6
eGAMP_{10} (include both interactions)	1.99	14.19	64.96	96.2

Table 4 shows the sensitivity analysis for raking_5, MRP_{10}, and eGAMP_{10}. Omitting the interaction between Z_1 and X^* results in much large bias and RMSE for all three estimators. However, the impact on the eGAMP_{10} (omitting interaction $Z_1 \times X_{10}^*$) is smallest among the three estimators, with smaller bias and RMSE than the other estimators. Omitting both the $Z_1 \times X_{10}^*$ and $Z_1 \times \text{logit}(\hat{\pi}_i)$ interactions leads to poor performance of eGAMP_{10}.

4 | Application to the Ohio Army National Guard Study

The Ohio Army National Guard Mental Health Initiative (OHARNG-MHI) study provides information about the prevalence and risk factors of mental health-related outcomes among National Guard service members with the aim of identifying areas of intervention that can be modified during deployment to improve the psychological well-being of soldiers [32, 33]. We apply the methods described in this paper to the OHARNG-MHI study to estimate the percentage of lifetime alcohol abuse among all the service members as a data illustration.

The target population for the OHARNG-MHI study included all active members of the OHARNG between June 2008 and February 2009. All members with address information listed with the Guard were notified of the study via mailed letter and opt-out card. Although some members chose to opt out of the study, others refused participation when contacted or were not contacted before the cohort closed [34]. Furthermore, service members with no or incorrect telephone numbers could not be contacted to complete the 60 min structured computer-assisted telephone interview [34]. As such, the statistical analysis of the OHARNG-MHI study data is complicated by nonsampling errors due to survey nonresponse and sampling frame undercoverage. Information on age group (17–24 years, 25–34 years, 35–44 years, 45 years or older), gender (male, female), race (White, Black, Other), rank (enlisted, officer), marital status (single, married, other), and number of years in service for the target population ($N = 10{,}994$) was available in OHARNG administrative files. A total of $n = 2{,}600$ service members completed the survey. Table 5 shows the distribution of the auxiliary information in the population and

TABLE 5 | Distribution of auxiliary variables in the OHARNG population and OHARNG-MHI survey sample.

	Distribution, *n*(%)	
Variable	**Population**	**MHI Sample**
All	10,944 (100.0)	2,600 (100.0)
Age group		
17–24 years	4,077 (37.2)	877 (33.7)
25–34 years	3,788 (34.6)	844 (32.5)
35–44 years	2,196 (20.1)	631 (24.3)
45 years or older	883 (8.1)	248 (9.5)
Gender		
Male	9,398 (85.9)	2,213 (85.1)
Female	1,546 (14.1)	387 (14.9)
Race		
White	9,593 (87.6)	2,284 (87.8)
Black	1,107 (10.1)	194 (7.5)
Other	244 (2.2)	122 (4.7)
Rank		
Enlisted	9,857 (90.1)	2,260 (86.9)
Officer	1,087 (9.9)	340 (13.1)
Marital status		
Single	6,008 (54.9)	1,129 (43.4)
Married	4,069 (37.2)	1,222 (47.0)
Other	867 (7.9)	249 (9.6)
Number of years in service, mean (min–max)	9.7 (0–42)	10.1 (0–40)

survey sample. It indicates that older, other race, officer rank, and married service members were overrepresented in the sample, whereas younger, Black race, enlisted rank, and single service members were underrepresented. The outcome of interest, presence of lifetime alcohol abuse, was only measured among survey participants.

Figure 1a shows the association between number of years in service and the logit-transformed proportion of lifetime alcohol abuse in the sample. Here, the proportion of lifetime alcohol abuse is computed as the number of participants with lifetime alcohol abuse divided by the total number of participants in each year of service. The plot of the logit-transformed proportion by years of service is overlaid with the fitted loess curve. The proportion increases with the number of years in service. Figure 1b is created similarly and shows a smooth association between the logit-transformed sample inclusion propensity and years in service. The logit-transformed sample inclusion probability has a linear association with the number of years in service, increasing from 0 to about 20 years, and then plateaus.

The data analysis includes three main steps. First, statisticians who have access to the confidential OHARNG administrative files, which contain the individual-level continuous years of service variable, estimate the sample inclusion propensities. For this analysis, a binary BART machine learning model is utilized, incorporating all available auxiliary variables from the administrative files as covariates, including both discrete and continuous variables listed in Table 5. In the second step, a modified population data set is created for sharing with data users. The continuous years of service variable in the administrative files is replaced with its discretized version, categorized using deciles of the population values, along with the estimated inclusion propensities, $\hat{\pi}$. Finally, data users who do not have access to the confidential administrative files use the survey sample and the modified population data to conduct survey inference. The proportion of lifetime alcohol abuse in the OHARNG population is estimated using several methods: unweighted sample mean, eIPW, raking, eIPW + raking, MRP, and eGAMP. For raking, all discrete covariates from Table 5 and the discretized version of years in service using deciles are used for weighting. For eGAMP, a GAM is fitted for the binary lifetime alcohol abuse variable, including all discrete covariates from Table 5, a spline of the deciles of the years of service variable, and a spline of the estimated inclusion propensities. For comparison, we also provide the GAM(x) estimator by fitting a GAM model regressing on a spline of the original continuous years in service variable and the other discrete auxiliary variables.

Figure 2 shows estimates and 95% CIs for the proportion of lifetime alcohol abuse. The unweighted sample mean estimates a higher proportion of lifetime alcohol abuse compared with the weighted and predictive estimators. This is expected, given that service members with more years of service were more likely to be included in the sample and also had a higher proportion of lifetime alcohol abuse (Figure 1). All the weighted and predictive estimators yield similar point estimates for the proportion of lifetime alcohol abuse. However, eGAMP produces a wider 95% CI than the others. Due to the limited variation in the estimated inclusion propensities among the sample units in this application, we do not observe the typically wide confidence intervals associated with the weighted estimators.

5 | Discussion

We consider finite population inference from a nonrepresentative sample where a number of auxiliary variables, both continuous and discrete, are measured in both the sample and the population via administrative records. This auxiliary information can be used to improve survey inference of population quantities through weighting or predictive models. However, individual-level continuous administrative records are often inaccessible due to confidentiality concerns. The common practice is to discretize the continuous auxiliary variables using percentiles or other meaningful cutoffs, which can lead to a loss of information. Motivated by this challenge, we develop a method for predictive survey inference that makes the best use of continuous auxiliary information in the administrative records without requiring the release of individual-level continuous variables for the entire population.

We propose a two-step strategy. In the first step, statistical agencies with access to confidential population data estimate probabilities of inclusion in the sample for all population units using both continuous and discrete auxiliary variables. They then

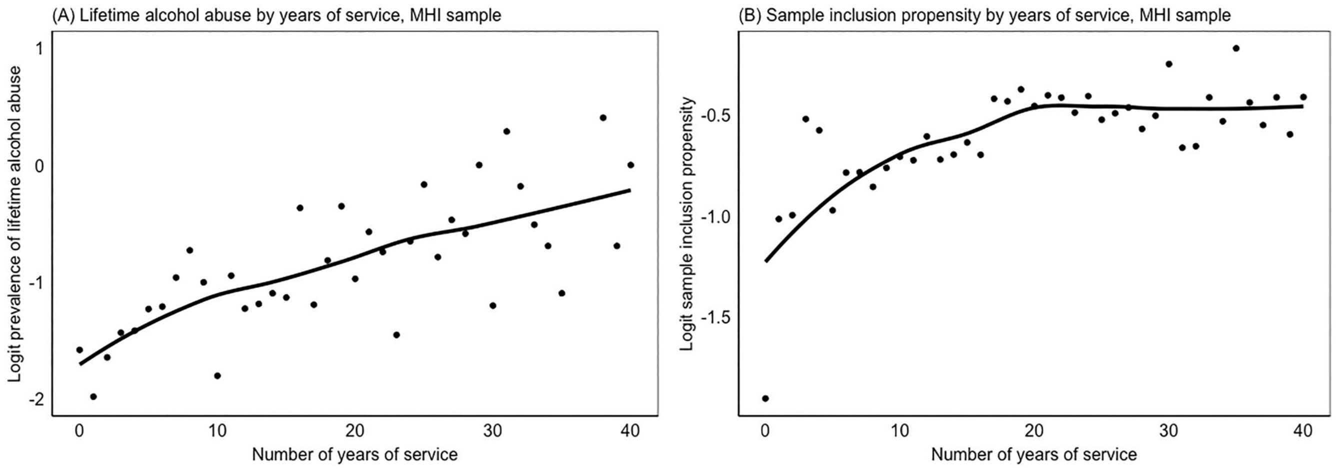

FIGURE 1 | Associations of lifetime alcohol abuse and response propensity with years in service, Ohio National Guard Mental Health Initiative Study, 2008-2009.

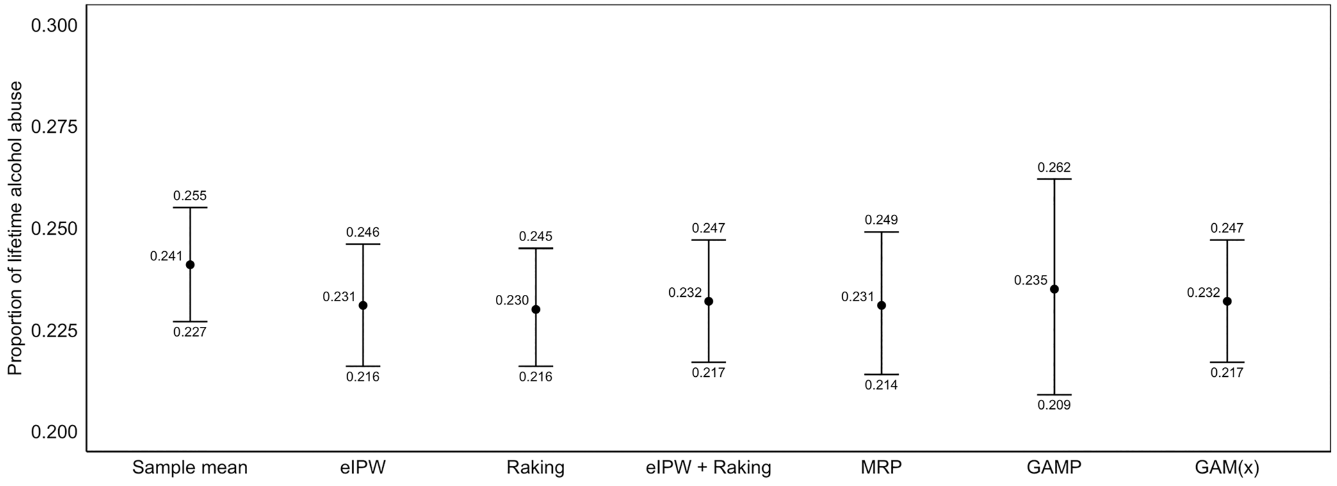

FIGURE 2 | Estimated proportion of lifetime alcohol abuse among all service members of the OHARNG between June 2008 and February 2009 using the Ohio National Guard Mental Health Initiative Study.

create a modified population data for data users by replacing the continuous variables with their discretized versions, along with the continuous estimated inclusion propensities. In the second step, data users, who do not have access to the protected population data, utilize the modified population dataset and the survey sample for statistical inference of population quantities using the proposed GAMP model. The GAMP model extends MRP by using a hierarchical structure of discrete variables and spline functions of continuous variables, incorporating discretized continuous variables and estimated inclusion propensities. This approach is inspired by Little and An [24], who included spline functions of the logit-transformed response propensity score in the prediction models, which yields robust estimates of sample means even when the model is misspecified. Similarly, including the logit-transformed estimated propensities in GAMP's predictive models can improve survey inference of population quantities when discretized continuous variables do not fully capture the associations between the original continuous variables and the survey outcome of interest. We provide an R package, `AuxSurvey`, for conducting analyses using our proposed GAMP estimator.

We assess the performance of the proposed GAMP method and compare it to existing weighting and prediction model-based approaches through a simulation study. Our simulations confirm established findings on the importance of appropriate statistical analyses to adjust for nonsampling errors such as nonresponse and undercoverage [4, 5, 35]. Both weighting and model-based approaches yield more accurate population estimates than unadjusted estimates across all simulation settings. Importantly, our proposed GAMP estimator outperforms the weighting approaches using IPW, poststratification, or raking, providing less biased and more efficient estimates with a 95% CI coverage rate closer to the nominal level. GAMP also outperforms MRP when continuous auxiliary variables have more complicated associations with both the outcome of interest and sample inclusion. GAMP has a smaller bias and improved efficiency compared with modeling the discretized continuous variables alone in MRP (e.g., simulation setting three). When interactions exist between continuous variables and a categorical variable in the outcome model, omitting these interaction effects can lead to large bias and RMSE in raking and MRP estimators. In contrast, the bias and RMSE of GAMP are relatively small as long as the interactions between categorical variables and the logit-transformed propensity scores are included, even if the interactions between categorical variables and discretized continuous variables are omitted. In cases where continuous auxiliary variables are not associated with the outcome of interest but are associated with sample inclusion (e.g., simulation setting two), the GAMP model, including a spline of the estimated inclusion propensities, does not offer additional benefits and incurs a slight increase in bias and RMSE compared with MRP. Nevertheless, the GAMP estimator still outperforms the weighted estimators in this setting.

In practice, the number of auxiliary variables available for the entire finite population is often limited, so flexible parametric models are usually preferred. In this paper, we consider a Bayesian GAM implemented using the `stan_gamm4()` function in `rstanarm`. Using the hierarchical structure similar to MRP, our GAM model partially pools estimates in multiple categories of discrete auxiliary covariates. Including spline functions of the discretized continuous variables and the estimated inclusion propensities, the model is flexible enough to catch their potential nonlinear associations with the survey outcome of interest. When there are interactions between the smooth splines and discrete auxiliary variables, the model can also be easily modified to reflect these interaction effects using stratified splines. Furthermore, the Bayesian framework is straightforward for quantification of uncertainty. In the settings where there exists a high dimension of auxiliary variables, machine learning techniques, such as BART, can be used instead.

It is crucial to balance data utility and confidentiality protection. The two-step strategy we propose aims to ensure that continuous auxiliary data in administrative records remain useful for survey inference while protecting confidentiality. Although we focus on continuous auxiliary variables, the proposed method can also be applied to settings with sensitive categorical covariates, such as fine-level geographic identifiers. In such cases, the fine-level geographic identifier is used to estimate the inclusion propensity, which is then released along with coarse-level geographic information for predictive survey inference. To avoid assigning the exact same estimated propensity score to units with identical categorical predictors, a random draw from the posterior distribution of the propensity score—rather than the posterior mean—can be obtained from the binary BART model for inclusion propensity.

The proposed GAMP method has several limitations. First, like other predictive estimators, its effectiveness depends on the inclusion of auxiliary variables that are predictive of the outcomes. Additionally, these auxiliary variables in the sample should have common support to those in the target population. If important predictors of outcomes are not available in the administrative records, or if the ranges of the auxiliary variables in the sample are narrower than those in the population, the model predictions may not perform well. Second, the sample inclusion propensity often requires estimation. Our simulation shows that the performance of GAMP using the true versus BART-estimated propensity scores is comparable, although the eGAMP estimator (using the BART-estimated inclusion propensity) results in larger bias but smaller RMSE compared with GAMP using the true inclusion propensity. In the eGAMP estimator, we ignore the uncertainty associated with π estimation. One way to address the uncertainty from estimating π is obtaining multiple sets of estimated propensity scores from their posterior distributions and repeating the predictive models multiple times with these sets of estimated π values. Third, estimating inclusion propensity often involves linking the survey sample with administrative records, which may be subject to record linkage errors and need further correction.

In conclusion, our study advocates for a prediction model-based approach that leverages continuous auxiliary information in administrative records to improve survey inference while controlling disclosure risk, thus eliminating the need to release individual-level continuous auxiliary variables. When survey inference is conducted by those with access to confidential records, a predictive model using splines on the continuous variables can be directly applied. However, for data users without such access, our two-step strategy offers an effective solution. Although the use of predictive inferences, Bayesian hierarchical models, and splines for propensity scores is not new, our major contribution lies in combining these components to leverage record-level continuous variables for improving survey inference in nonrepresentative probability samples or nonprobability samples while maintaining confidentiality. This paper focuses on model-based predictive inference that includes the propensity score as a covariate. An alternative approach that combines weighting and prediction is model-assisted estimation [36]. This research also opens up several exciting directions for future investigation. For example, modeling two separate propensities—one for undercoverage (noncontact) and one for response (responded out of those contacted)—and including them as smoothed terms in the prediction model could be explored. Another potential extension is to model the unknown population distribution of auxiliary variables and incorporate the uncertainty in estimating this distribution for survey inference [37].

Author Contributions

Sharifa Z. Williams and Qixuan Chen conceived and designed the research and wrote the paper with contributions from Yajuan Si, Jungang Zou, and Sandro Galea. Sharifa Z. Williams, Jungang Zou, and Yutao Liu conducted the simulation study. Sharifa Z. Williams performed the real data analysis. Jungang Zou wrote the `R`package. Sharifa Z. Williams, Qixuan Chen, and Yajuan Si interpreted the results. Sandro Galea contributed the application data from the Ohio Army National Guard Mental Health Initiative study.

Acknowledgments

The authors would like to acknowledge and thank Gregory Cohen for preparing the dataset used in the data application, and the Ohio Army National Guard Mental Health Initiative study team including Dr. Joseph Calabrese, Dr. Mario Tamburrino, and Dr. Israel Liberzon for access to the 2008-2009 data. We also would like to thank the Editor, the Associate Editor, and the referees for their thoughtful and constructive comments which greatly improved the paper.

Disclosure

Yutao Liu is now an employee of AstraZeneca and may or may not own stock options.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The survey data of the Ohio Army National Guard Mental Health Initiative study are not publicly available due to privacy or ethical restrictions.

References

1. Y. Si, R. J. Little, Y. Mo, and N. Sedransk, “A Case Study of Nonresponse Bias Analysis in Educational Assessment Surveys,” *Journal of Educational and Behavioral Statistics* 48, no. 3 (2022): 271–295.

2. J. Bacher, J. Lemcke, A. Quatember, and P. Schmich, “Probability and Nonprobability Sampling: Representative Surveys of Hard-To-Reach and Hard-To-Ask Populations. Current Surveys Between the Poles of Theory and Practice,” *Survey Insights: Methods From the Field* (2019), https://doi.org/10.13094/SMIF-2019-00018.

3. A. Gelman, “Struggles With Survey Weighting and Regression Modeling,” *Statistical Science* 22, no. 2 (2007): 153–164.

4. J. M. Brick and G. Kalton, “Handling Missing Data in Survey Research,” *Statistical Methods in Medical Research* 5 (1996): 215–238.

5. D. Holt and T. Smith, “Post Stratification,” *Journal of the Royal Statistical Society Series A* 142, no. 1 (1979): 33–46.

6. Y. Liu, A. Gelman, and Q. Chen, “Inference From Non-random Samples Using Bayesian Machine Learning,” *Journal of Survey Statistics and Methodology* 11 (2023): 433–455.

7. W. E. Deming and F. F. Stephan, “On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Margin Totals Are Known,” *Annals of Mathematics and Statistics* 11 (1940): 427–444.

8. J. Deville, C. Sarndal, and O. Sautory, “Generalized Raking Procedures in Survey Sampling,” *Journal of the American Statistical Association* 88, no. 423 (1993): 1013–1020.

9. D. Pfeffermann, “New Important Developments in Small Area Estimation,” *Statistical Science* 28, no. 1 (2013): 40–68.

10. R. Little, “To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling,” *Journal of the American Statistical Association* 99, no. 466 (2004): 546–556.

11. M. R. Elliott and R. J. A. Little, “Model-Based Alternatives to Trimming Survey Weights,” *Journal of Official Statistics* 16, no. 3 (2000): 191–209.

12. M. Ghosh and J. N. K. Rao, “Small Area Estimation: An Appraisal,” *Statistical Science* 9, no. 1 (1994): 55–76.

13. R. Little, “Post-Stratification: A modeler’s Perspective,” *Journal of the American Statistical Association* 88 (1993): 1001–1012.

14. A. Gelman and T. Little, “Postratification Into Many Categories Using Hierarchical Logistic Regression,” *Survey Methodology* 23, no. 2 (1998): 127–135.

15. J. Rao and I. Molina, *Small Area Estimation*, 2nd ed. (Hoboken: John Wiley and Sons, Inc., 2015).

16. D. K. Park, A. Gelman, and J. Bafumi, “Bayesian Multilevel Estimation With Poststratification: State-Level Estimates From National Polls,” *Political Analysis* 12 (2004): 375–385.

17. W. Wang, D. Rothschild, S. Goel, and A. Gelman, “Forecasting Elections With Non-Representative Polls,” *International Journal of Forecasting* 31 (2015): 980–991, https://doi.org/10.1016/j.ijforecast.2014.06.001.

18. F. Ritchie and J. Smith, *Confidentiality and Linked Data* (London: Government Statistical Service Methodology Advisory Committee, 2018), https://arxiv.org/ftp/arxiv/papers/1907/1907.06465.pdf.

19. K. Harron, C. Dibben, J. Boyd, et al., “Challenges in Administrative Data Linkage for Research,” *Big Data & Society* 4, no. 2 (2017), https://doi.org/10.1177/2053951717745678.

20. Y. Si, “On the use of Auxiliary Variables in Multilevel Regression and Poststratification,” Under Review (2023), https://arxiv.org/abs/2011.00360.

21. A. Gelman and J. Little, *Data Analysis Using Regression and Multilevel/Hierarchical Models* (New York: Cambridge, 2007).

22. Y. Si, R. Trangucci, J. S. Gabry, and A. Gelman, “Bayesian Hierarchical Weighting Adjustment and Survey Inference,” *Survey Methodology* 46, no. 2 (2020): 181–214.

23. Y. Ghitza and A. Gelman, “Deep Interactions With MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups,” *American Journal of Political Science* 57 (2013): 762–776.

24. R. Little and H. An, “Robust Likelihood-Based Analysis of Multivariate Data With Missing Values,” *Statistica Sinica* 14 (2004): 949–968.

25. B. K. Lee, J. Lessler, and E. A. Stuart, “Weight Trimming and Propensity Score Weighting,” *PLoS One* 6, no. 3 (2011): e18174, https://doi.org/10.1371/journal.pone.0018174.

26. M. Schonlau, v A. Soest, A. Kapteyn, and M. Couper, “Selection Bias in Web Surveys and the Use of Propensity Scores,” *Sociological Methods & Research* 37, no. 3 (2009): 291–318.

27. P. R. Rosenbaum and D. B. Rubin, “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika* 70, no. 1 (1983): 41–55.

28. The Stan Development Team, “RStan: The R Interface to Stan,” Version 2.5.0 (2014), http://mc-stan.org/rstan.html.

29. M. Hoffman and A. Gelman, “The No-u-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo,” *Journal of Machine Learning Research* 25 (2013): 1351–1381.

30. R. Neal, “MCMC Using Hamiltonian Dynamics,” in *Handbook of Markov Chain Monte Carlo*, eds. S. Brooks, A. Gelman, G. Jones, and X. L. Meng (BocaRaton, FL: Chapman and Hall/CRC Press, 2011), 113–162.

31. H. A. Chipman, E. I. George, and R. E. McCulloch, “BART: Bayesian Additive Regression Trees,” *Annals of Applied Statistics* 4, no. 1 (2010): 266–298.

32. E. Goldmann, J. Calabrese, M. Prescott, et al., “Potentially Modifiable Pre-, Peri-, and Post-Deployment Characteristics Associated With Deployment-Related Posttraumatic Stress Disorder Among Ohio Army National Guard Soldiers,” *Annals of Epidemiologys* 22 (2011): 71–78.

33. M. B. Tamburrino, P. Chan, M. Prescott, et al., “Baseline Prevalence of Axis I Diagnosis in the Ohio Army National Guard,” *Psychiatry Research* 226 (2015): 142–148.

34. J. Calabrese, M. Prescott, M. B. Tamburrino, et al., “PTSD Comorbidity and Suicidal Ideation Associated With PTSD Within the Ohio Army National Guard,” *Journal of Clinical Psychiatry* 72, no. 8 (2011): 1072–1078.

35. Q. Chen, M. Elliott, D. Haziza, et al., “Approaches to Improving Survey-Weighted Estimates,” *Statistical Science* 32, no. 2 (2017): 227–248.

36. J. Breidt and J. Opsomer, “Model-Assisted Survey Estimation With Modern Prediction Techniques,” *Statistical Science* 32, no. 2 (2017): 190–205.

37. J. Dever and R. Valliant, “General Regression Estimation Adjusted for Undercoverage and Estimated Control Totals,” *Journal of Survey Statistics and Methodology* 4, no. 3 (2016): 289–318.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.