Statistics
in Medicine WILEY

Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts

Lingxiao Wang[1] | **Richard Valliant**[1,2] | **Yan Li**[1]

[1]The Joint Program in Survey Methodology, University of Maryland, College Park, Maryland, USA

[2]Institute for Social Research, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Yan Li, The Joint Program in Survey Methodology, University of Maryland, 1218 Lefrak Hall, 7521 Preinkert Dr, College Park, MD 20742, USA.
Email: yli6@umd.edu

Many epidemiologic studies forgo probability sampling and turn to nonprobability volunteer-based samples because of cost, response burden, and invasiveness of biological samples. However, finite population (FP) inference is difficult to make from the nonprobability sample due to the lack of population representativeness. Aiming for making inferences at the population level using nonprobability samples, various inverse propensity score weighting methods have been studied with the propensity defined by the participation rate of population units in the nonprobability sample. In this article, we propose an adjusted logistic propensity weighting (ALP) method to estimate the participation rates for nonprobability sample units. The proposed ALP method is easy to implement by ready-to-use software while producing approximately unbiased estimators for population quantities regardless of the nonprobability sample rate. The efficiency of the ALP estimator can be further improved by scaling the survey sample weights in propensity estimation. Taylor linearization variance estimators are proposed for ALP estimators of FP means that account for all sources of variability. The proposed ALP methods are evaluated numerically via simulation studies and empirically using the naïve unweighted National Health and Nutrition Examination Survey III sample, while taking the 1997 National Health Interview Survey as the reference, to estimate the 15-year mortality rates.

KEYWORDS
finite population inference, nonprobability sample, propensity score weighting, survey sampling, variance estimation

1 | INTRODUCTION

In the big data era, assembling volunteer-based epidemiologic cohorts within integrated healthcare systems that have electronic health records and a large preexisting base of volunteers are increasingly popular due to their cost-and-time efficiency, such as the UK Biobank in the UK National Health Service.[1] However, samples of volunteer-based cohorts are not randomly selected from the underlying finite target population, and therefore cannot well represent the target population. As a result, the naïve sample estimates obtained from the cohort can be biased for the finite population (FP) quantities. For example, the estimated all-cause mortality rate in the UK Biobank was only half that of the UK population,[2] and the Biobank is not representative of the UK population with regard to many sociodemographic, physical, lifestyle, and health-related characteristics.

Aiming for making inferences at the population level using nonprobability samples, various propensity-score weighting, and matching methods have been proposed to improve the population representativeness of nonprobability samples, by using probability-based survey samples as external references in survey research.[3-6]

Inverse propensity score weighting (IPSW) methods have been studied with the propensity defined by the participation rate of population units in the nonprobability sample. We review two methods—both assume that the units in the nonprobability sample are observed according to some random, but unknown, mechanism. Because that mechanism is unknown, the inclusion probability of each unit must be estimated. As described in Section 2, all methods are based on estimating a pseudo log-likelihood, although the methods differ in their details. Valliant and Dever[7] estimated participation rates by fitting a logistic regression model to the combined nonprobability sample and a reference, probability sample. Sample weights for the probability sample were scaled by a constant so that the scaled probability sample was assumed to represent the complement of the nonprobability sample. Each unit in the nonprobability sample was assigned a weight of one. This results in the sum of the scaled weights in the combined probability plus nonprobability sample being an estimate of the population size. This method will be referred to as the rescaled design weight (RDW) method. The participation rate for each nonprobability sample unit was estimated by the inverse of the estimated inclusion (or participation) probability.

The RDW estimator is biased especially when the participation rate of the nonprobability sample is large, as noted by Chen et al.[4] As a remedy, Chen et al[4] estimated the participation rate by manipulating the log-likelihood estimating equation in a somewhat different way. The resulting estimator, denoted by CLW, is consistent and approximately unbiased regardless of the magnitude of participation rates. Compared with the CLW method, which requires special programming, the RDW method has the advantage of easy implementation by ready-to-use software such as R, Stata, or SAS. Survey practitioners can simply fit a logistic regression model with scaled survey weights in the probability sample to obtain the estimated participation rates.

In this article, we propose an adjusted logistic propensity weighting (ALP) method to estimate the participation rates for nonprobability sample units. Like the CLW, the proposed ALP method relaxes the assumptions required by the RDW method,[7,8] by formulating the method in an innovative way. As in the RDW method, the proposed ALP method retains the advantage of easy implementation by fitting a propensity model with survey weights in ready-to-use software. Taylor linearization (TL) variance estimators are proposed for ALP estimates that account for variability due to differential pseudo-weights in the nonprobability sample, complex survey design of the reference probability survey, as well as the estimation of the propensity scores. The variance of the proposed estimator has the order of the inverse of the nonprobability sample size (as shown in Appendix C). Moreover, under the logistic propensity model, the ALP method can flexibly scale the probability sample weights for propensity estimation to further improve efficiency. In summary, the contributions of the proposed ALP method include (1) easy implementation with ready-to-use software, (2) high efficiency, as well as (3) the justification of a set of pseudo-estimating equation (14) that underly the straightforward implementation in survey software.

2 | METHODS

2.1 | Basic setting

Let FP = $\{1, \cdots, N\}$ represent the FP with size N. We are interested in estimating the FP mean $\mu = N^{-1} \sum_{i \in FP} y_i$. Suppose a volunteer-based nonprobability sample s_c of size n_c is selected from FP by a self-selection mechanism, with $\delta_i^{(c)}$ ($= 1$ if $i \in s_c$; 0 otherwise) denoting the indicator of s_c inclusion. The underlying participation rate of nonprobability sample for a FP unit is defined as

$$\pi_i^{(c)} = P(i \in s_c \,|\, FP) = E_c\{\delta_i^{(c)} | y_i, \boldsymbol{x}_i\}, \quad i \in FP,$$

where the expectation E_c is with respect to the nonprobability sample selection, and $\boldsymbol{x}_i$ is a vector of self-selection variables, that is, covariates related to the probability of inclusion in s_c. The corresponding implicit nonprobability sample weight is $w_i = 1/\pi_i^{(c)}$ for $i \in FP$.

We consider the following assumptions for the nonprobability sample self-selection.

A1. The nonprobability sample selection is uncorrelated with the variable of interest given the covariates, that is, $\pi_i^{(c)} = E_c\{\delta_i^{(c)} | y_i, \boldsymbol{x}_i\} = E_c\{\delta_i^{(c)} | \boldsymbol{x}_i\}$ for $i \in FP$.

A2. All FP units have a positive participation rate, that is, $\pi_i^{(c)} > 0$ for $i \in FP$.

A3. The indicators of participation in the nonprobability cohort are uncorrelated with each other given the self-selection variables, that is, $\text{cov}(\delta_i^{(c)}, \delta_j^{(c)}|x_i, x_j) = 0$ for $i \neq j$.

An independent reference probability-based survey sample s_p of size n_p is randomly selected from FP. The sample inclusion indicator, selection probability, and the corresponding sample weights are defined by $\delta_i^{(p)}(=1$ if $i \in s_p$; 0 otherwise), $\pi_i^{(p)} = E_p(\delta_i^{(p)}|x_i)$, and $d_i = 1/\pi_i^{(p)}$, respectively, where E_p is with respect to the survey sample selection.

2.2 | Existing logistic propensity weighting method

In this section, we first briefly introduce the existing RDW and CLW methods and discuss their pros and cons.

2.2.1 | RDW method

Valliant and Dever[6,7] assumed a logistic regression model for the participation rates $\pi_i^{(c)}(\gamma)$

$$\log\left\{\frac{\pi_i^{(c)}(\gamma)}{1 - \pi_i^{(c)}(\gamma)}\right\} = \gamma^T x_i, \text{ for } i \in \text{FP}, \tag{1}$$

where γ is a vector of unknown parameters, and x_i is a vector of covariates for $i \in \text{FP}$. To simplify the notation, we use $\pi_i^{(c)}$ below. They considered (implicitly) the population likelihood function of $\pi_i^{(c)}$ as

$$L(\gamma) = \prod_{i \in \text{FP}} \{\pi_i^{(c)}\}^{\delta_i^{(c)}} \{1 - \pi_i^{(c)}\}^{1 - \delta_i^{(c)}}. \tag{2}$$

Then, the log-likelihood function can be written as

$$l(\gamma) = \sum_{i \in \text{FP}} [\delta_i^{(c)} \log \pi_i^{(c)} + \{1 - \delta_i^{(c)}\} \log\{1 - \pi_i^{(c)}\}] = \sum_{i \in s_c} \log \pi_i^{(c)} + \sum_{i \in \text{FP} - s_c} \log\{1 - \pi_i^{(c)}\}, \tag{3}$$

where the set $\text{FP} - s_c$ represents the FP units that are not self-selected into the nonprobability sample. Since $\text{FP} - s_c$ is not available in practice, the pseudo-loglikelihood function was constructed to estimate $l(\gamma)$ by

$$\tilde{l}^{\text{RDW}}(\gamma) = \sum_{i \in s_c} w_i^* \log \pi_i^{(c)} + \sum_{i \in s_p} w_i^* \log\{1 - \pi_i^{(c)}\}, \tag{4}$$

where $w_i^* = \begin{cases} 1, & \text{for } i \in s_c \\ d_i(\hat{N}_p - n_c)/\hat{N}_p, & \text{for } i \in s_p \end{cases}$, with $\hat{N}_p = \sum_{i \in s_p} d_i$ being the survey estimate of the target FP size N. This leads to the total of the scaled weights across the probability sample units being $\sum_{i \in s_p} w_i^* = \hat{N}_p - n_c$. The rationale for rescaling is to weight the survey sample to represent the complement of s_c in the FP, that is, the set $\text{FP} - s_c$. Under the logistic regression model, the nonprobability sample participation rate $\pi_i^{(c)}$ for $i \in s_c$ can be estimated by fitting Model (1) to the combined sample of s_c and *scaled-weighted* s_p with scaled weights w_i^*, leading to the RDW estimates.

The RDW method has been shown to effectively reduce the bias of the naïve nonprobability sample estimates. However, the summand $\sum_{i \in \text{FP} - s_c} \log\{1 - \pi_i^{(c)}\}$ in (3) is not a fixed FP total because units in the nonprobability sample s_c are treated as being randomly observed. This leads to a bias as shown below.

Comparing the expectation of the population log-likelihood function $l(\gamma)$ in (3) and the expectation of the pseudo log-likelihood $\tilde{l}^{\text{RDW}}(\gamma)$ in (4), and letting $E(\cdot) = E_c E_p(\cdot)$ we have

$$E\{l(\gamma)\} = \sum_{i \in \text{FP}} \pi_i^{(c)} \log \pi_i^{(c)} + \sum_{i \in \text{FP}} \{1 - \pi_i^{(c)}\} \log\{1 - \pi_i^{(c)}\}, \text{ and}$$

$$E\{\tilde{l}^{\text{RDW}}(\gamma)\} = E_c E_p\{\tilde{l}^{\text{RDW}}(\gamma)\} = E_c\left[\sum_{i \in \text{FP}} \delta_i^{(c)} \log \pi_i^{(c)}\right] + E_p\left[\sum_{i \in \text{FP}} \delta_i^{(p)} \cdot \frac{\hat{N}_p - n_c}{\hat{N}_p} d_i \log\{1 - \pi_i^{(c)}\}\right]$$

$$\doteq \sum_{i \in \text{FP}} \pi_i^{(c)} \log \pi_i^{(c)} + \sum_{i \in \text{FP}} \left\{1 - \frac{n_c}{N}\right\} \log\{1 - \pi_i^{(c)}\}$$

by assuming $E_p(\widehat{N}_p) = N$. The difference of the two expectations, denoted by Δ_{RDW}, can be written as

$$\Delta_{RDW} = E\{\widetilde{l}^{RDW}(\gamma)\} - E\{l(\gamma)\} = \sum_{i \in FP} \left\{ \frac{n_c}{N} - \pi_i^{(c)} \right\} \log\{1 - \pi_i^{(c)}\}$$

which, in general, is nonzero. Accordingly, the nonprobability sample participation rates estimated by solving for γ in $\partial \widetilde{l}^{RDW}(\gamma)/\partial \gamma = 0$ under Model (1) can be biased, unless either (i) the nonprobability sample units have small participation rates, that is, both n_c/N and $\pi_i^{(c)}$ are close to 0 for all $i \in FP$, in which case $\log\{1 - \pi_i^{(c)}\} \approx 0$, or (ii) all population units are equally likely to participate in the nonprobability sample, that is, $\pi_i^{(c)} \equiv n_c/N$. In many practical applications, (i) will hold. For example, suppose that $n_c = 1000$ and the US population age 18 and over is the target population. The population size is approximately 210 million, so that $n_c/N \doteq 5 \times 10^{-6}$. If, instead, the population is for a small state like Wyoming where the 18+ population size is about 365 000, then $n_c/N \doteq 0.0027$. In both examples, with such small sampling fractions, all $\{\pi_i^{(c)}, i \in s_c\}$ should be near zero also.

2.2.2 | CLW method

Chen et al[4] proposed another IPSW method using the same likelihood function $L(\gamma)$ in (2), but rewriting the population log-likelihood as

$$l(\gamma) = \sum_{i \in s_c} \log \frac{\pi_i^{(c)}}{1 - \pi_i^{(c)}} + \sum_{i \in FP} \log\{1 - \pi_i^{(c)}\}. \tag{5}$$

In contrast to the RDW method, CLW estimated the population total of $\log\{1 - \pi_i^{(c)}\}$ by a weighted reference sample total and constructed the pseudo log-likelihood as

$$\widetilde{l}^{CLW}(\gamma) = \sum_{i \in s_c} \log \frac{\pi_i^{(c)}}{1 - \pi_i^{(c)}} + \sum_{i \in s_p} d_i \log\{1 - \pi_i^{(c)}\}. \tag{6}$$

Under the same logistic regression model (1), the participation rate $\pi_i^{(c)}$ was estimated by solving the pseudo estimation equation

$$\widetilde{S}(\gamma) = \frac{1}{N} \left\{ \sum_{i \in s_c} x_i - \sum_{i \in s_p} d_i \pi_i^{(c)} x_i \right\} = 0, \tag{7}$$

derived from the pseudo log-likelihood (6). The resulting CLW weights are calculated as $\{w_i^{CLW} = 1 + \exp^{-1}(\widehat{\gamma}^T x_i), i \in s_c\}$. Chen et al[4] proved that the CLW estimator of the FP mean, $\widehat{\mu}^{CLW} = \left(\sum_{i \in s_c} w_i^{CLW} \right)^{-1} \sum_{i \in s_c} w_i^{CLW} y_i$, was design consistent when model (1) for the participation rates was correct.

In contrast to the RDW method, CLW does not require condition (i) or (ii) in the RDW method for unbiased estimation of participation rates $\{\pi_i^{(c)}, i \in s_c\}$. In the following section, we propose an ALP method, which corrects the bias in the RDW method. The proposed ALP method provides consistent estimators of FP means and is as easy to implement as the RDW method.

2.3 | ALP method

The ALP method also aims to estimate the cohort sample participation rates $\{\pi_i^{(c)}, i \in s_c\}$ and use the inverse of estimated $\pi_i^{(c)}$ as the pseudo-weight for $i \in s_c$. As a computational device, we construct a pseudo-population of $s_c^* \cup FP$, where s_c^* is a copy of s_c that has the same joint distributions of covariates x and outcome y with the original s_c. The number of units in

$s_c^* \cup FP$ is $n_c + N$. In the union of $s_c^* \cup FP$, s_c^* and s_c are treated as two different sets. We use R_i to indicate the membership of s_c^* in $s_c^* \cup FP$ (=1 if $i \in s_c^*$; 0 if $i \in FP$), and $p_i = P(R_i = 1) = P(i \in s_c^* | s_c^* \cup FP)$. Instead of directly modeling $\pi_i^{(c)}$ as in the RDW and CLW methods, we model p_i as a function of $\pi_i^{(c)}$:

$$p_i = \frac{\pi_i^{(c)}}{1 + \pi_i^{(c)}}, \quad \text{or equivalently}, \quad \pi_i^{(c)} = \frac{p_i}{1 - p_i}. \tag{8}$$

The relationship between p_i and $\pi_i^{(c)}$ follows because

$$\frac{p_i}{1 - p_i} = \frac{P(i \in s_c^* | s_c^* \cup FP)}{P(i \in FP | s_c^* \cup FP)} = \frac{P(i \in s_c)}{P(i \in FP)} = P(i \in s_c | FP) = \pi_i^{(c)} \tag{9}$$

since s_c^* is a copy of s_c and $P(i \in s_c^* | s_c^* \cup FP) = P(i \in s_c | s_c^* \cup FP)$. Notice that, derived from Formula (8),

$$p_i \leq \frac{1}{2},$$

since $\pi_i^{(c)} \leq 1$ and the equality holds only if $\pi_i^{(c)} = 1$, that is, the FP unit i participates in the cohort with certainty. As illustrated by the examples at the end of section 2.2.1, requiring $p_i \leq 1/2$ is not unrealistic in typical applications because $\pi_i^{(c)}$ is generally quite small.

Suppose that p_i can be modeled parametrically by $p_i = p(\mathbf{x}_i; \boldsymbol{\beta}) = \text{expit}(\boldsymbol{\beta}^T \mathbf{x}_i)$, where $\boldsymbol{\beta}$ is a vector of unknown model parameters. That is,

$$\log \left\{ \frac{p_i}{1 - p_i} \right\} = \boldsymbol{\beta}^T \mathbf{x}_i, \text{ for } i \in s_c^* \cup FP \tag{10}$$

Notice that $\boldsymbol{\beta}$, the coefficients in Model (10), differ from the coefficients $\boldsymbol{\gamma}$ in Model (1) because the two logistic regression models have different dependent variables. Based on (8), expression (10) implies that $\pi_i^{(c)}$ is being modeled as $\exp(\boldsymbol{\beta}^T \mathbf{x}_i)$, which differs from the RDW/CLW model in (1) where $\pi_i^{(c)} = \exp(\boldsymbol{\gamma}^T \mathbf{x}_i)/\{1 + \exp(\boldsymbol{\gamma}^T \mathbf{x}_i)\}$. The corresponding "likelihood" function can be written as

$$L^*(\boldsymbol{\beta}) = \prod_{i \in s_c^* \cup FP} p_i^{R_i} (1 - p_i)^{(1 - R_i)}, \tag{11}$$

where R_i indicates the membership of s_c^* in $s_c^* \cup FP$ (=1 if $i \in s_c^*$; 0 if $i \in FP$). We put "likelihood" in quotes because $L^*(\boldsymbol{\beta})$ varies depending on which set of units is selected for s_c^*. This contrasts with the population likelihood in (2) which applies regardless of which sample is selected. Note that $L^*(\boldsymbol{\beta})$ in (11) is written as if the units are independent when they are not. This is a standard procedure in pseudo-MLE estimation, and the resulting parameter estimators remain design-consistent even when some units may be correlated due to, for example, clustering.[9,10] The quantity $L^*(\boldsymbol{\beta})$ should be viewed as motivation for developing the estimating equations given below in (13). The log-likelihood generated from $L^*(\boldsymbol{\beta})$ is

$$l^*(\boldsymbol{\beta}) = \sum_{i \in s_c^* \cup FP} \{R_i \cdot \log p_i + (1 - R_i) \log(1 - p_i)\} = \sum_{i \in FP} \delta_i^{(c)} \log p_i + \sum_{i \in FP} \log(1 - p_i). \tag{12}$$

Notice that the randomness of $L^*(\boldsymbol{\beta})$ and $l^*(\boldsymbol{\beta})$ comes from the cohort selection, that is, $\delta_i^{(c)}$ in the first summand in the last line of (12). In reality, since the unit level information of FP is unknown, we replace the second summand in (12) by a survey sample estimate, $\sum_{i \in s_p} d_i \log(1 - p_i)$, and obtain the maximum pseudo-likelihood estimator $\widehat{\boldsymbol{\beta}}$ by solving the pseudo-estimating equation

$$\widetilde{S}^*(\boldsymbol{\beta}) = \frac{1}{N + n_c} \left\{ \sum_{i \in s_c} (1 - p_i) \mathbf{x}_i - \sum_{i \in s_p} d_i p_i \mathbf{x}_i \right\} = \mathbf{0}. \tag{13}$$

Assuming that p_i is bounded by 0 and 1/2 implies that $\pi_i^{(c)}$ is automatically bounded by 0 and 1. Note that (13) falls in a general class of estimating equations that ensure unique solutions of parameters[4,11,12] (eg, equation (6) in CLW[4] if their function $h(\boldsymbol{x}_i, \boldsymbol{\theta})$ is set equal to $\boldsymbol{x}_i\{1 + \pi_i^{(c)}\}^{-1}$).

The ALP estimator of μ is

$$\hat{\mu}^{\text{ALP}} = \frac{\sum_{i \in s_c} w_i^{\text{ALP}} y_i}{\sum_{i \in s_c} w_i^{\text{ALP}}}, \tag{14}$$

where $w_i^{\text{ALP}} = 1/\pi_i^{(c)}(\hat{\boldsymbol{\beta}})$ for $i \in s_c$. Although $L^*(\boldsymbol{\beta})$ is not a standard likelihood, $\hat{\mu}^{\text{ALP}}$ is a consistent estimator of the population mean as shown in the theorem below.

We consider the following limiting process for the theoretical development.[4,13] Suppose there is a sequence of FPs FP_k of size N_k, for $k = 1, 2, \cdots$. Cohort $s_{c,k}$ of size $n_{c,k}$ and survey sample $s_{p,k}$ of size $n_{p,k}$ are sampled from FP_k. The sequences of the FP, the cohort and the survey sample have their sizes satisfy $\lim_{k \to \infty} n_{t,k}/N_k \to f_t$, where $t = c$ or p and $0 < f_t \leq 1$ (regularity condition C1 in Appendix A). In the following the index k is suppressed for simplicity.

Theorem. *Consistency of ALP estimator of FP mean (see Appendix B).*

Under the regularity conditions A1 to A3, and C1 to C5 in Appendix A, and assuming logistic regression model (10) for p_i, the ALP estimate $\hat{\mu}^{\text{ALP}}$ is design consistent for μ, in particular $\hat{\mu}^{\text{ALP}} - \mu = O_p(n_c^{-1/2})$, with the FP variance

$$\text{Var}(\hat{\mu}^{\text{ALP}}) \doteq N^{-2} \sum_{i \in \text{FP}} p_i(1 - 2p_i)\left\{\frac{(y_i - \mu)}{p_i} - \boldsymbol{b}^T \boldsymbol{x}_i\right\}^2 + \boldsymbol{b}^T \boldsymbol{D} \boldsymbol{b}, \tag{15}$$

where $p_i = \text{expit}(\boldsymbol{\beta}^T \boldsymbol{x}_i)$, $\boldsymbol{b}^T = \left\{\sum_{i \in \text{FP}}(y_i - \mu)\boldsymbol{x}_i^T\right\}\left\{\sum_{i \in \text{FP}} p_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right\}^{-1}$, and $\boldsymbol{D} = N^{-2} V_p\left(\sum_{i \in s_p} d_i p_i \boldsymbol{x}_i\right)$ is the design-based variance-covariance matrix under the probability sampling design for s_p.

In practice, the ALP estimator of a FP mean can be obtained by three steps:

Step 1   Search for covariates $\boldsymbol{x}$ available in both the cohort (s_c) and the reference survey sample (s_p) and combine the two samples. Assign $R_i = 1$ for $i \in s_c$ and $R_i = 0$ for $i \in s_p$ in the combined sample.

Step 2   Fit a logistic regression model for $p_i = P(R_i = 1)$ in the combined s_c and weighted s_p, with the survey sample weights $\{d_i, i \in s_p\}$, and obtain the estimate $\hat{p}_i$ for $i \in s_c$.

Step 3   Estimate the FP mean by Formula (14) with the ALP pseudo weight $w_i^{\text{ALP}} = \hat{p}_i/(1 - \hat{p}_i)$ for $i \in s_c$.

Notice that Step 2 can be accomplished by any existing survey software, such as svyglm in survey package of R, svy:logit in Stata, and PROC SURVEYLOGISTIC in SAS. In addition to being easy to implement, the ALP estimator from (14) does not require conditions (i) or (ii), unlike RDW. Moreover, we prove that in large samples, $\text{Var}(\hat{\mu}^{\text{ALP}}) = O(n_c^{-1})$ is as or more efficient compared with $\text{Var}(\hat{\mu}^{\text{CLW}}) = O\{\min{(n_p, n_c)}^{-1}\}$ under their correct propensity model, respectively, which depends on both the nonprobability and probability sample sizes (see Appendix C).

An alternative method would be to omit the odds transformation, which uses p_i to approximate the participation rate $\pi_i^{(c)}$. Denote this method by FDW for full design weight, which contrasts to the scaling of the survey sample weights in the RDW method. Comparing the expectation of the population log-likelihood function $l(\boldsymbol{\gamma})$ in (3) and the expectation of the pseudo log-likelihood $\tilde{l}^*(\boldsymbol{\beta})$ in (12) with $\pi_i^{(c)}$ replacing p_i by the FDW method, that is, $\tilde{l}^*(\boldsymbol{\gamma}) = \sum_{i \in s_c} \log \pi_i^{(c)} + \sum_{i \in s_p} d_i \log(1 - \pi_i^{(c)})$, we have their difference, denoted by Δ_{FDW}, written as

$$\begin{aligned}
\Delta_{\text{FDW}} = E\{\tilde{l}^*(\boldsymbol{\gamma})\} - E\{l(\boldsymbol{\gamma})\} &= \sum_{i \in \text{FP}} \pi_i^{(c)} \log \pi_i^{(c)} + \sum_{i \in \text{FP}} \log\{1 - \pi_i^{(c)}\} \\
&- \sum_{i \in \text{FP}} \pi_i^{(c)} \log \pi_i^{(c)} - \sum_{i \in \text{FP}}(1 - \pi_i^{(c)}) \log\{1 - \pi_i^{(c)}\} = \sum_{i \in \text{FP}} \pi_i^{(c)} \log\{1 - \pi_i^{(c)}\}.
\end{aligned}$$

The bias is zero only if $\pi_i^{(c)}$ for $i \in \text{FP}$ are all close to zero. Thus, the odds transformation step in ALP could be skipped if all nonprobability participation rates are extremely small; but, in general, that step is essential for unbiased estimation.

2.4 | Variance estimation

Using the FP variance formula (15), the first summand can be consistently estimated by

$$\{\widehat{N}^{(c)}\}^{-2} \sum_{i \in s_c} (1 - \widehat{p}_i)(1 - 2\widehat{p}_i) \left\{ \frac{(y_i - \widehat{\mu}^{\text{ALP}})}{\widehat{p}_i} - \widehat{\boldsymbol{b}}^T \boldsymbol{x}_i \right\}^2, \tag{16}$$

where $\widehat{p}_i$ is the prediction for $i \in s_c$, $\widehat{N}^{(c)} = \sum_{i \in s_c} w_i^{\text{ALP}}$, and $\widehat{\boldsymbol{b}}^T = \left\{ \sum_{i \in s_c} (y_i - \widehat{\mu}^{\text{ALP}}) \boldsymbol{x}_i^T \right\} \left\{ \sum_{i \in s_c} \widehat{p}_i \boldsymbol{x}_i \boldsymbol{x}_i^T \right\}^{-1}$. The second summand $\boldsymbol{b}^T \boldsymbol{D} \boldsymbol{b}$ is estimated by $\widehat{\boldsymbol{b}}^T \widehat{\boldsymbol{D}} \widehat{\boldsymbol{b}}$, where $\widehat{\boldsymbol{D}}$ is the survey design consistent variance estimator of $\boldsymbol{D}$. For example, under stratified multistage cluster sampling with H strata and a_h primary sampling units (PSUs) in stratum h selected with replacement,

$$\widehat{\boldsymbol{D}} = \{\widehat{N}^{(p)}\}^{-2} \cdot \sum_{h=1}^{H} \frac{a_h}{a_h - 1} \sum_{l=1}^{a_h} (\boldsymbol{z}_l - \overline{\boldsymbol{z}})(\boldsymbol{z}_l - \overline{\boldsymbol{z}})^T, \tag{17}$$

where $\widehat{N}^{(p)} = \sum_{i \in s_p} d_i$, $\boldsymbol{z}_l = \sum_{i \in s_{p(hl)}} d_i \widehat{p}_i \boldsymbol{x}_i$ is the weighted PSU total for cluster l in stratum h, $s_{p(hl)}$ is the set of sample elements stratum h and cluster l, and $\overline{\boldsymbol{z}} = a_h^{-1} \sum_l^{a_h} \boldsymbol{z}_l$ is the mean of the PSU totals in stratum h.

2.5 | Scaling survey weights in the likelihood for the ALP method

The proposed ALP can flexibly scale the survey weights in estimating equation (13) to improve efficiency. For case-control studies, Scott and Wild[14] and Li et al[15] previously used the technique we propose below to reduce variances of estimates of relative risks when weights for cases and controls are substantially different. We multiply the second summand in $\widetilde{S}^*(\boldsymbol{\beta})$ by a constant λ, say $\lambda = n_c / \left(\sum_{i \in s_p} d_i \right)$, so that the sum of the scaled survey weights (λd_i) is n_c. Accordingly, the score function becomes

$$\widetilde{S}_\lambda^*(\boldsymbol{\beta}) = \sum_{i \in s_c} (1 - p_i) \boldsymbol{x}_i - \lambda \sum_{i \in s_p} d_i p_i \boldsymbol{x}_i. \tag{18}$$

Solving $\widetilde{S}_\lambda^*(\boldsymbol{\beta}) = 0$ for $\boldsymbol{\beta}$, and the resulting vector of estimates is denoted by $\widehat{\boldsymbol{\beta}}_\lambda = (\widehat{\beta}_{0,\lambda}, \widehat{\boldsymbol{\beta}}_{1,\lambda})$, where $\widehat{\beta}_{0,\lambda}$ is estimate of the intercept. Similar derivations to those in Chambers and Skinner[10] and Beaumont[11] can be used to prove that $\widehat{\boldsymbol{\beta}}_{1,\lambda}$ is design-consistent with various efficiency gains, depending on the variability of survey weights vs the nonprobability sample weights (with implicit common value of 1). However, the estimate of the intercept $\widehat{\beta}_{0,\lambda}$ can be badly biased with scaled weights. As a result, the estimate of participation rate $\exp(\widehat{\boldsymbol{\beta}}_\lambda^T \boldsymbol{x}_i)$ including $\widehat{\beta}_{0,\lambda}$ would also be biased. The bias of $\widehat{\beta}_{0,\lambda}$, however, would not affect the estimate of population mean because the scaled ALP-weighted mean, $\widehat{\mu}^{\text{ALP.S}}$,

$$\widehat{\mu}^{\text{ALP.S}} = \frac{\sum_{i \in s_c} w_i^{\text{ALP.S}} y_i}{\sum_{i \in s_c} w_i^{\text{ALP.S}}} = \frac{\sum_{i \in s_c} \exp^{-1}(\widehat{\boldsymbol{\beta}}_{1,\lambda}^T \boldsymbol{x}_i) y_i}{\sum_{i \in s_c} \exp^{-1}(\widehat{\boldsymbol{\beta}}_{1,\lambda}^T \boldsymbol{x}_i)}, \tag{19}$$

depends on $\widehat{\boldsymbol{\beta}}_{1,\lambda}$, but not $\widehat{\beta}_{0,\lambda}$, where $w_i^{\text{ALP.S}} = \exp^{-1}(\widehat{\boldsymbol{\beta}}_{1,\lambda}^T \boldsymbol{x}_i)$ is the scaled ALP pseudo weight.

It can be proved that $\widehat{\mu}^{\text{ALP.S}}$ is a consistent estimator of the FP mean, μ. The TL variance estimator of $\widehat{\mu}^{\text{ALP.S}}$ can be obtained by substituting w_i^{ALP}, $\widehat{\mu}^{\text{ALP}}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{p}_i$, and d_i by $w_i^{\text{ALP.S}}$, $\widehat{\mu}^{\text{ALP.S}}$, $\widehat{\boldsymbol{\beta}}_\lambda$, $\widehat{p}_{i,\lambda} = \exp(\widehat{\boldsymbol{\beta}}_\lambda^T \boldsymbol{x}_i)$, and λd_i, respectively, in Formulae (16) and (17). Details on the variance and the consistency of ALP.S are discussed in the dissertation by Wang.[16]

3 | SIMULATIONS

3.1 | FP generation and sample selection

We applied simulation setups similar to those in Chen et al.[4] In the FP of size $N = 500\,000$, a vector of covariates $\boldsymbol{x}_i = (x_{1i}, x_{2i}, x_{3i}, x_{4i})^T$ was generated for $i \in$ FP where $x_{1i} = v_{1i}$, $x_{2i} = v_{2i} + 0.3x_{1i}$, $x_{3i} = v_{3i} + 0.2(x_{1i} + x_{2i})$, $x_{4i} = v_{4i} + 0.1$

$(x_{1i} + x_{2i} + x_{3i})$, with $v_{1i} \sim \text{Bernoulli}(0.5)$, $v_{2i} \sim \text{Uniform}(0, 2)$, $v_{3i} \sim \text{Exponential}(1)$, and $v_{4i} \sim \chi^2(4)$. The variable of interest $y_i \sim \text{Normal}(\mu_i, 1)$, where $\mu_i = -x_{1i} - x_{2i} + x_{3i} + x_{4i}$ for $i \in$ FP. The parameter of interest was the FP mean $\mu = N^{-1} \sum_{i \in \text{FP}} y_i = 3.97$.

The probability-based survey sample s_p with the target sample size $n_p = 12\,500$ (sampling fraction $f_p = 2.5\%$) was selected by Poisson sampling, with inclusion probability $\pi_i^{(p)} = (n_p \cdot q_i)/ \sum_{i \in \text{FP}} q_i$ for $i \in$ FP, where $q_i = \text{const} + x_{3i} + 0.03y_i$ to control for the variation of the survey weights, $d_i = 1/\pi_i^{(p)}$. We set $\text{const} = -0.26$ so that $\max q_i / \min q_i = 20$.

As noted in Section 2, the ALP and CLW methods do assume somewhat different models for the participation rate. Thus, it is interesting to check their performances both when their underlying models are correct and when the assumed participation rate models fail. The volunteer-based nonprobability sample s_c (with a target sample size n_c) was also selected by Poisson sampling but with different inclusion probabilities $\pi_i^{(c)}$ for $i \in$ FP. We considered two scenarios with different functional forms of $\pi_i^{(c)}$ so that the ALP (and FDW) or the CLW method had the true linear logistic regression propensity model in one scenario but not in the other. In Scenario 1, $\pi_i^{(c)} = \exp(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i)$ was the specified participation rate for the ith population unit to be included into the nonprobability sample. The underlying true propensity model for ALP (and FDW) methods, shown in (10), was $\text{logit}(p_i) = \log\{\pi_i^{(c)}\} = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i$, which implies $\text{logit}\{\pi_i^{(c)}\} = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i - \log\{1 - \pi_i^{(c)}\}$. This model differs from the underlying linear model (1) assumed by the CLW method by the addition of the term $\log\{1 - \pi_i^{(c)}\}$. In Scenario 2, $\pi_i^{(c)} = \text{expit}(\gamma_0 + \boldsymbol{\gamma}^T \mathbf{x}_i)$ was specified so that $\text{logit}\{\pi_i^{(c)}\} = \gamma_0 + \boldsymbol{\gamma}^T \mathbf{x}_i$, which was the model (1) assumed by the CLW method. This model, however, implied that $\text{logit}(p_i) = \log\{\pi_i^{(c)}\} = \gamma_0 + \boldsymbol{\gamma}^T \mathbf{x}_i + \log\{1 - \pi_i^{(c)}\}$, which was different from the model assumed by the ALP and the FDW method (by the extra term $\log\{1 - \pi_i^{(c)}\}$). Hence, ALP and CLW estimates of the population mean are expected to be unbiased in one scenario but not the other since both methods assume a linear logistic propensity model. The biases of the FDW and RDW estimates, as measured by Δ_{FDW} and Δ_{RDW}, depend on $\pi_i^{(c)}$, and go to 0 as $\pi_i^{(c)}$ approaches 0. The biases become larger as $\pi_i^{(c)}$ increases in either scenario.

In both scenarios, the coefficients were set to be $\boldsymbol{\beta} = \boldsymbol{\gamma} = (0.18, 0.18, -0.27, -0.27)^T$. The parameters were chosen so that $0 < \pi_i^{(c)} < 1$ for all units $i \in$ FP. The intercepts β_0 and γ_0 were also controlled so that the expected number of nonprobability sample units $E_c(n_c) = \sum_{\text{FP}} \pi_i^{(c)}$ was varied from 1250, 2500, 5000, to 10 000 with the corresponding overall participation rate $f_c = E_c(n_c)/N$ being 0.5%, 5%, 10%, or 20%.

3.2 | Evaluation criteria

We examined the performance of five IPSW estimators of FP mean μ: (1) to (2) $\widehat{\mu}^{\text{ALP}}$ and $\widehat{\mu}^{\text{ALP.S}}$ described in Sections 2.2 to 2.5; (3) $\widehat{\mu}^{\text{FDW}}$ using weights from the ALP method omitting the odds transformation; (4) $\widehat{\mu}^{\text{CLW}}$ proposed by Chen et al[4]; and (5) $\widehat{\mu}^{\text{RDW}}$ proposed by Wang et al,[6] compared with the naïve nonprobability sample mean ($\widehat{\mu}^{\text{Naïve}}$) that did not use weights, and the weighted nonprobability sample mean, $\widehat{\mu}^{\text{TW}}$, with weights equal to the inverse of the true nonprobability sample inclusion probabilities. Note that $\widehat{\mu}^{\text{TW}}$ is unavailable in practice because the true nonprobability sample inclusion probabilities are unknown. Relative bias (%RB), empirical variance (V), mean squared error (MSE) of the point estimates were used to evaluate the performance of the four IPSW point estimates, calculated by

$$\%\text{RB} = \frac{1}{B} \sum_{b=1}^{B} \frac{\widehat{\mu}^{(b)} - \mu}{\mu} \times 100, V = \frac{1}{B-1} \sum_{b=1}^{B} \left\{ \widehat{\mu}^{(b)} - \frac{1}{B} \sum_{b=1}^{B} \widehat{\mu}^{(b)} \right\}^2, \text{MSE} = \frac{1}{B} \sum_{b=1}^{B} \{\widehat{\mu}^{(b)} - \mu\}^2,$$

where $B = 4000$ is the number of simulation runs, $\widehat{\mu}^{(b)}$ is one of the point estimates obtained from the bth simulated sample, and μ is the true FP mean.

We also evaluated the variance estimates using the variance ratio (VR) and 95% confidence interval coverage probability (CP), which were calculated as

$$\text{VR} = \frac{\frac{1}{B} \sum_{b=1}^{B} \widehat{v}^{(b)}}{V} \times 100, \text{ and CP} = \frac{1}{B} \sum_{b=1}^{B} I(\mu \in \text{CI}^{(b)}),$$

where $\widehat{v}^{(b)}$ is the proposed analytical variance estimate in simulated sample b, and $\text{CI}^{(b)} = (\widehat{\mu}^{(b)} - 1.96\sqrt{\widehat{v}^{(b)}}, \widehat{\mu}^{(b)} + 1.96\sqrt{\widehat{v}^{(b)}})$ is the 95% confidence interval from the bth simulated sample.

TABLE 1 Results from 4000 simulated survey samples and nonprobability samples with low to high participation rates under various propensity score models

	Scenario 1 True propensity model for ALP					Scenario 2 True propensity model for CLW				
	%RB	V (×10⁵)	VR	MSE (×10⁵)	CP⁵	%RB	V (×10⁵)	VR	MSE (×10⁵)	CP
$f_c = 0.5\%$										
$\hat{\mu}^{\text{Naive}}$	−42.76	0.22	0.99			−42.61	0.22	1.00		
$\hat{\mu}^{\text{TW}}$	−0.13	4.38	0.93	4.39	0.90	−0.12	4.38	0.93	4.38	0.90
$\hat{\mu}^{\text{RDW}}$	−0.29	3.73	0.93	3.75	0.87	−0.40	3.63	0.93	3.66	0.87
$\hat{\mu}^{\text{FDW}}$	−0.28	3.73	0.93	3.75	0.87	−0.40	3.63	0.93	3.66	0.87
$\hat{\mu}^{\text{ALP}}$	−0.07	3.70	0.93	3.77	0.88	−0.19	3.66	0.93	3.67	0.88
$\hat{\mu}^{\text{CLW}}$	0.05	3.87	0.93	3.87	0.89	−0.07	3.76	0.93	3.76	0.88
$\hat{\mu}^{\text{ALP}.S}$	−0.11	3.54	0.92	3.54	0.87	−0.21	3.45	0.92	3.45	0.87
$f_c = 5\%$										
$\hat{\mu}^{\text{Naive}}$	−42.74	0.02	0.99			−41.21	0.02	1.01		
$\hat{\mu}^{\text{TW}}$	−0.04	0.50	0.98	0.50	0.92	−0.02	0.46	1.00	0.47	0.93
$\hat{\mu}^{\text{RDW}}$	−2.15	0.56	1.00	1.29	0.66	−3.05	0.43	1.01	1.89	0.45
$\hat{\mu}^{\text{FDW}}$	−2.05	0.57	1.00	1.23	0.68	−2.95	0.43	1.01	1.81	0.47
$\hat{\mu}^{\text{ALP}}$	−0.01	0.62	1.00	0.62	0.94	−1.03	0.47	1.01	0.64	0.85
$\hat{\mu}^{\text{CLW}}$	1.29	0.84	1.00	1.10	0.95	0.01	0.61	1.01	0.61	0.94
$\hat{\mu}^{\text{ALP}.S}$	−0.05	0.45	1.00	0.45	0.92	−0.63	0.35	1.02	0.41	0.86
$f_c = 10\%$										
$\hat{\mu}^{\text{Naive}}$	−42.74	0.01	1.11			−39.65	0.01	1.11		
$\hat{\mu}^{\text{TW}}$	−0.01	0.25	1.02	0.25	0.94	−0.01	0.22	1.01	0.22	0.94
$\hat{\mu}^{\text{RDW}}$	−4.25	0.34	1.00	3.20	0.17	−5.62	0.22	0.99	5.20	0.02
$\hat{\mu}^{\text{FDW}}$	−3.87	0.35	1.00	2.71	0.24	−5.28	0.22	0.99	4.62	0.03
$\hat{\mu}^{\text{ALP}}$	0.01	0.42	1.00	0.42	0.95	−1.81	0.27	0.99	0.79	0.65
$\hat{\mu}^{\text{CLW}}$	2.94	0.80	1.00	2.16	0.76	0.03	0.42	0.99	0.42	0.95
$\hat{\mu}^{\text{ALP}.S}$	−0.03	0.27	1.02	0.27	0.94	−0.86	0.18	1.01	0.29	0.81
$f_c = 20\%$										
$\hat{\mu}^{\text{Naive}}$	−42.75	0.00	1.26			−36.50	0.01	1.21		
$\hat{\mu}^{\text{TW}}$	−0.02	0.15	0.93	0.15	0.93	−0.02	0.11	0.96	0.11	0.93
$\hat{\mu}^{\text{RDW}}$	−8.58	0.21	0.95	11.83	0.00	−9.59	0.10	0.97	14.60	0.00
$\hat{\mu}^{\text{FDW}}$	−7.15	0.23	0.96	8.29	0.01	−8.51	0.11	0.98	11.53	0.00
$\hat{\mu}^{\text{ALP}}$	0.00	0.32	0.96	0.32	0.95	−2.85	0.16	0.98	1.44	0.19
$\hat{\mu}^{\text{CLW}}$	7.80	1.67	0.92	11.27	0.06	0.01	0.33	0.98	0.33	0.95
$\hat{\mu}^{\text{ALP}.S}$	−0.03	0.19	0.96	0.19	0.94	−1.02	0.10	1.00	0.27	0.73

Abbreviations: ALP, adjusted logistic propensity; CP, coverage probability; MSE, mean squared error; VR, variance ratio.

3.3 | Results

Table 1 presents simulation results for the seven nonprobability sample estimators of the FP mean. The naïve estimator $\hat{\mu}^{\text{Naive}}$ that ignored the underlying sampling scheme had relative biases ranging from −36.5% to −42.8% while the true weighted nonprobability sample estimator, $\hat{\mu}^{\text{TW}}$, was approximately unbiased in all scenarios. The variance of $\hat{\mu}^{\text{Naive}}$ was much smaller than that of the other estimators, but its bias caused the MSE to be extremely high (not reported).

Consistent with the bias theory in Section 2, the RDW point estimator $\widehat{\mu}^{\text{RDW}}$ and the FDW point estimator $\widehat{\mu}^{\text{FDW}}$ were approximately unbiased when $\pi_i^{(c)}$ was small for all $i \in \text{FP}$ and the overall participation rate $f_c = N^{-1} \sum_{i \in \text{FP}} \pi_i^{(c)}$ was low, but more biased as f_c increased. The CPs decreased correspondingly.

As expected, the ALP estimators $\widehat{\mu}^{\text{ALP}}$ and $\widehat{\mu}^{\text{ALP.S}}$ (or the CLW estimator $\widehat{\mu}^{\text{CLW}}$) consistently provided unbiased point estimators in the scenarios where they were expected to be unbiased, that is, scenario 1 for $\widehat{\mu}^{\text{ALP}}$ and $\widehat{\mu}^{\text{ALP.S}}$, and scenario 2 for $\widehat{\mu}^{\text{CLW}}$. When the underlying model was incorrect for an estimator, biases occurred. For example, the relative biases of $\widehat{\mu}^{\text{CLW}}$ in scenario 1 were 0.05%, 1.29%, 2.94%, and 7.80% as f_c increased from 0.5%, 5%, 10%, to 20%, respectively. In scenario 2, the corresponding relative biases for $\widehat{\mu}^{\text{ALP}}$ are −0.19%, −1.03%, −1.81%, and −2.85%.

Consistent with the theory in Section 2, the ALP estimator $\widehat{\mu}^{\text{ALP}}$ was more efficient than $\widehat{\mu}^{\text{CLW}}$ with consistently smaller empirical variances in all scenarios, especially when the nonprobability cohort size was much larger than the probability sample size. Among all considered methods, $\widehat{\mu}^{\text{ALP.S}}$ was approximately unbiased with the smallest variance under Scenario 1 of the correct model. Under Scenario 2 of a misspecified model, $\widehat{\mu}^{\text{ALP.S}}$ was biased but most efficient, and therefore achieved smallest MSE.

The variance estimators for $\widehat{\mu}^{\text{ALP}}$, $\widehat{\mu}^{\text{ALP.S}}$, and $\widehat{\mu}^{\text{CLW}}$ performed very well (with VR's near 1), providing CPs close to the nominal level under the correct propensity models when f_c was large. The lower coverage of the nominal level (about 88%) when $f_c = 0.5\%$ was due to the small sample bias with skewed distributions of underlying sampling weights in the selected nonprobability sample.

4 | REAL DATA EXAMPLE

We use the same data example as Wang et al[6] for illustration purposes. We estimated prospective 15-year all-cause, all-cancer, and heart disease mortality rates for adults in the US using the adult household interview part of The Third U.S. National Health and Nutrition Examination Survey (NHANES III) III conducted in 1988 to 1994, with sample size $n_c = 20\,050$. We ignored all complex design features of NHANES III and treated it as a nonprobability sample. The coefficient of variation of sample weights is 125%, indicating highly variable selection probabilities, and thus low representativeness of the unweighted sample. For estimating mortality rates, we approximated that the entire sample of NHANES III was randomly selected in 1991 (the midpoint of the data collection time period).

For the reference survey, we used 1994 U.S. National Health Interview Survey (NHIS) respondents to the supplement for monitoring achievement of the Healthy People Year 2000 objectives. Adults aged 18 and older are included (sample size $n_p = 19\,738$). The 1994 NHIS used a multistage stratified cluster sample design with 125 strata and 248 pseudo-PSUs.[17,18] We collapsed strata with only one PSU with the next nearest stratum for variance estimation purposes.[19] Both samples of NHANES III and NHIS were linked to National Death Index (NDI) for mortality, allowing us to quantify the relative bias of unweighted NHANES estimates, assuming the NHIS estimates as the gold standard. Notice that the mortality information was obtained by statistical linkage between the survey sample and NDI,[20] but not responses from the questionnaires. The all-cancer and heart-disease mortality were classified according to National Center for Health Statistics death code.[21,22]

The usage of NHANES III as the "nonprobability cohort" has several advantages for illuminating the performance of the propensity weighting methods. The "nonprobability sample" and the reference survey sample have approximately the same target population, data collection mode, and similar questionnaires. This ensures that the pseudo-weighted "nonprobability sample" could potentially represent the target population, and thus enables us to characterize the performance of the propensity weighting methods in real data.

The distributions of selected common covariates and variables of interests in the two samples are presented in Table 2. As expected, the variables in the weighted samples of NHANES and 1994 NHIS have very close distributions because both weighted samples represent approximately the same FP. By contrast, covariates distribute quite differently in the *unweighted* NHANES from the weighted samples, especially for design variables such as age, race/ethnicity, poverty, and region, which leads to large biases in mortality rates estimated from the *unweighted* NHANES.

The propensity model included main effects of common demographic characteristics (age, sex race/ethnicity, region, and marital status), socioeconomic status (education level, poverty, and household income), tobacco usage (smoking

TABLE 2 Distribution of selected common variables in NIH-AARP and NHIS

		NHIS 1994		NHANES III	
		$n_p = 19\,738$	$\widehat{N}_p = 189\,608\,549$	$n_c = 20\,050$	$\widehat{N}_c = 187\,647\,206$
Variable	Total count	%	Weighted %	%	Weighted %
Age Group	18-24 years	10.5	13.3	15.8	15.8
	25-44 years	42.9	43.7	35.4	43.7
	45-64 years	26.1	26.6	22.6	24.6
	65 years and older	20.5	16.4	26.2	16.0
Race	NH-White	76.1	75.9	42.3	76.0
	NH-Black	12.6	11.2	27.4	11.2
	Hispanic	8.0	9.0	28.9	9.3
	NH-other	3.3	4.0	1.5	3.5
Region	Northeast	20.7	20.5	14.6	20.8
	Midwest	26.1	25.1	19.2	24.1
	South	31.5	32.5	42.7	34.3
	West	21.6	21.9	23.5	20.9
Poverty	No	79.1	82.3	67.9	80.3
	Yes	13.1	10.6	21.4	12.1
	Unknown	7.8	7.0	10.7	7.6
Education	Lower than high school	20.1	19.1	42.5	26.6
	High school/Some college	58.7	59.6	45.9	54.1
	College or higher	21.2	21.3	11.6	19.3
Health status	Excellent/Very good	60.5	62.0	39.0	51.6
	Good	25.7	25.7	35.9	32.7
	Fair/Poor	13.8	12.3	25.1	15.7
Mortality	All-cause	20.8	17.6	26.7	17.1
	Heart-disease	9.43	5.69	4.95	4.04
	All-cancer	5.57	4.11	5.10	4.47

status, and chewing tobacco), health variables (body mass index and self-reported health status), and a quadratic term for age. Appendix D shows the final propensity models for the five considered methods.

To evaluate the performance of the five PS-based methods, we used relative difference from the NHIS estimate %RD = $(\widehat{\mu} - \widehat{\mu}^{\text{NHIS}})/\widehat{\mu}^{\text{NHIS}} \times 100$, TL variance estimate ($V$), and estimated MSE = $(\widehat{\mu} - \widehat{\mu}^{\text{NHIS}})^2 + V$, which treated the NHIS estimate as truth. Table 3 shows that the naïve NHANES III estimate of overall mortality was ~52% biased from the NHIS estimate because older people who have higher mortalities were oversampled (Table 2). All five IPSW methods substantially reduced the bias from the naïve estimate. Consistent with the simulation results, the ALP, FDW, RDW, and CLW method yielded close estimates when the sample fraction of the nonprobability sample was small ($\widehat{f}_c = n_c/\widehat{N}_p = 1.06 \times 10^{-4}$ calculated from Table 2). The ALP.S method, by scaling the NHIS sample weights in propensity estimation, reduced more bias than the other methods, and was more efficient. Therefore, the ALP.S estimate had the smallest MSE. The results for inference of all-cancer mortality had the similar pattern as the results for all-cause mortality. All pseudo-weighting methods removed most bias of the naïve NHANES estimate (with %RD = $-3.21\% \sim 2.07\%$ reduced from 24.68%). By contrast, for heart-disease mortality, all pseudo-weighting methods were substantially less biased than the naïve estimate, with %RD = $42.58\% \sim 57.78\%$ reduced from 133.66%, but the alternative estimators still had undesirably large biases themselves. The bias reduction is not as much as that for all-cancer or all-cause mortality, and this may be due to the omission of important predictors of having heart disease and of being observed in the nonprobability sample in the propensity model.

TABLE 3 Relative difference (%RD) of all-cause 15-year mortality estimates from the NHIS estimate with estimated variance (V) and mean squared error (MSE)

Mortality	Method	Estimate (%)	%RD	V (×10^5)	MSE (×10^5)
All cause	NHIS	17.6			
	Naïve	26.7	52.16		
	ALP	18.6	6.08	1.87	13.27
	FDW	18.6	6.08	1.87	13.28
	RDW	18.6	6.08	1.87	13.28
	CLW	18.6	6.07	1.87	13.24
	ALP.S	17.2	−2.05	1.08	2.37
All cancer	NHIS	4.5			
	Naïve	5.6	24.68		
	ALP	4.6	2.07	0.38	0.46
	FDW	4.6	2.07	0.38	0.46
	RDW	4.6	2.07	0.38	0.46
	CLW	4.6	2.06	0.37	0.46
	ALP.S	4.3	−3.21	0.32	0.53
Heart disease	NHIS	4.0			
	Naïve	9.4	133.66		
	ALP	6.4	57.78	0.50	54.88
	FDW	6.4	57.78	0.50	54.90
	RDW	6.4	57.78	0.50	54.90
	CLW	6.4	57.77	0.50	54.87
	ALP.S	5.8	42.58	0.33	29.86

Abbreviations: ALP, adjusted logistic propensity; RDW, rescaled design weight.

5 | DISCUSSION

This article proposed ALP weighting methods for population inference using nonprobability samples. The proposed ALP method corrects the bias in the RDW method[7] by formulating the problem in an innovative way. As does the RDW method, the proposed ALP method retains the advantage of easy implementation by fitting a propensity model with survey weights in ready-to-use software. The proposed ALP estimators are design consistent if the assumed model for participation rate is correct. TL variance estimators for ALP estimates are derived. Consistency of the ALP FP mean estimators was proved theoretically and evaluated numerically.

A primary competitor to ALP is the CLW estimator developed by Chen et al.[4] If the nonprobability cohort is a small fraction of the population, ALP and CLW are very similar, although ALP does have computational advantages regardless of the size of the sampling fraction. As the sampling fraction increases, ALP and CLW become more distinct.

Both ALP and CLW methods fit a propensity model to the combined nonprobability sample and a weighted survey sample. Highly variable weights in the combined sample can lead to low efficiency of the estimated propensity model coefficients. Therefore, the variances of the ALP and the CLW estimators of the FP means can be large in some applications. However, the proposed ALP is proved analytically and numerically to have a variance that is less than or equal to that of the CLW method regardless of whether the propensity model underlying ALP is correct. It worth noting that ALP and CLW methods assume different logistic regression models for propensity score estimation. Propensity is defined as $p_i = P(i \in s_c^* | s_c^* \cup \text{FP})$ by ALP in (9) and $\pi_i = P(i \in s_c | \text{FP})$ by CLW in (2). Model diagnostics should be developed to select which propensity model is more appropriate for a given dataset and will be the focus of our future research.

An alternative ALP with scaled survey weights in the logistic regression propensity model produces consistent propensity estimates and further improves efficiency as shown in the simulation and the real data example. The scaled ALP had the smallest MSE in every scenario in our simulation study regardless of underlying model and had the smallest MSE for two of the three mortality causes in our real data application. The CLW estimator with the scaled survey weights, albeit more efficient than the unscaled CLW, is biased (simulation results not shown). The extension of scaling technique to the CLW method and other binary regression propensity models requires further investigation.

The theory for the ALP method implies that $p_i > 1/2$ since p_i defines the probability of the nonprobability sample inclusion among the combined FP units and the nonprobability sample, that is, $P(i \in s_c^* | s_c^* \cup \text{FP})$. In estimation, however, $\hat{p}_i > 1/2$ can happen, especially in the unusual case where the nonprobability sample s_c^* is a large proportion of the population. Fortunately, this will not be a concern for the estimation of population mean or regression coefficients. By scaling survey weights in the propensity model for the ALP method, we can control for all ALP weights $w_i^{\text{ALP.S}} = 1/\hat{\pi}_i^{\text{ALP.S}}$ to be greater than one. As proved in Section 2.5, using scaled survey weights, $w_i^{\text{ALP.S}} = \lambda \cdot \exp^{-1}(\hat{\boldsymbol{\beta}}_\lambda^T \boldsymbol{x}_i) = \exp^{-1}(\hat{\beta}_{0\lambda}^* + \hat{\boldsymbol{\beta}}_{1\lambda}^T \boldsymbol{x}_i)$, would not bias the mean estimates. For population total estimation, an option is to estimate the population mean first and then multiply by a known or estimated population size from an independent source. Another approach to avoid estimated $p_i > 1/2$ is to solve the pseudo-estimating equations in (14) using a constrained optimization algorithm that requires $\hat{p}_i \leq 1/2$ for all units. This would, of course, negate the computation advantage of the ALP estimator.

Both ALP and CLW are inverse-propensity-score-weighting methods that directly use (functions of) the propensity score to estimate the cohort participation rate. They can be sensitive to propensity model misspecification (eg, missing interaction terms in the fitted propensity model) due to inaccurate estimates of participation rates. Furthermore, extreme pseudo-weights can occur if the estimate of the participation rate is close to 0. By contrast, propensity-score-based matching methods (not included in this study) may be more robust to the model misspecification and less likely to produce extreme pseudo-weights, because they use propensity scores to measure the similarity between survey and cohort sample units and distribute survey sample weights to the cohort based on their similarity. Examples of matching methods are propensity-score adjustment by subclassification,[23] propensity-score-based kernel weighting methods,[5,6,16] and River's matching method.[24]

There are a number of shortcomings associated with the estimation of propensity scores using logistic regression. First, the logistic model is susceptible to model misspecification, requiring assumptions regarding correct variable selection and functional form, including the choice of polynomial terms and multiple-way interactions. If any of these assumptions are incorrect, propensity score estimates can be biased, and balance may not be achieved when conditioning on the estimated PS. Second, implementing a search routine for model specification, such as repeatedly fitting logistic regression models while in/excluding predictor variables, interactions, or transformations of variables can be computationally infeasible or suboptimal. In this context, parametric regression can be limiting in terms of possible model structures that can be searched over, particularly when many potential predictors are present (high-dimensional data). Various machine learning methods for estimating the propensity score that incorporate survey weights will also be our future research interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author.

ORCID

Lingxiao Wang https://orcid.org/0000-0002-3463-615X
Richard Valliant https://orcid.org/0000-0002-9176-2961
Yan Li https://orcid.org/0000-0001-8241-7464

REFERENCES

1. Collins R. What makes UK Biobank special. *Lancet*. 2012;379(9822):1173-1174.
2. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol*. 2017;186(9):1026-1034.
3. Elliott MR, Valliant R. Inference for nonprobability samples. *Stat Sci*. 2017;32(2):249-264.
4. Chen Y, Li P, Wu C. Doubly robust inference with nonprobability survey samples. *J Am Stat Assoc*. 2019;115(532):2011-2021.
5. Wang L, Graubard BI, Katki H, Li Y. Improving external validity of epidemiologic cohort analyses: a kernel weighting approach. *J Royal Stat Soc Ser A*. 2020; 183(3):1293–1311. https://doi.org/10.1111/rssa.12564.

6. Wang L, Graubard BI, Katki HA, Li Y. Efficient and robust propensity-score-based methods for population inference using epidemiologic cohorts; 2021. arXiv preprint arXiv:2011.14850.

7. Valliant R, Dever JA. Estimating propensity adjustments for volunteer web surveys. *Sociol Methods Res*. 2011;40(1):105-137.

8. Valliant R. Comparing alternatives for estimation from nonprobability samples. *J SurvStat Methodol*. 2020;8(2):231-263.

9. Binder DA. On the variances of asymptotically normal estimators from complex surveys. *Int Stat Rev*. 1983;51(3):279-292.

10. Chambers RA, Skinner CJ. *Sec. 6.3 Analysis of Survey Data*. New York, NY: Wiley; 2003.

11. Beaumont JF. Calibrated imputation in surveys under a quasi-model-assisted approach. *J Royal Stat Soc Ser B*. 2005;67(3):445-458.

12. Kim JK, Kim JJ. Nonresponse weighting adjustment using estimated probability. *Canadian J Stat*. 2007;35(4):501-514.

13. Krewski D, Rao JN. Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Ann Stat*. 1981;9(5):1010-1019.

14. Scott AJ, Wild CJ. Fitting logistic models under case-control or choice based sampling. *J Royal Stat Soc Ser B*. 1986;48(2):170-182.

15. Li Y, Graubard BI, DiGaetano R. Weighting methods for population-based case–control studies with complex sampling. *J Royal Stat Soc Ser C*. 2011;60(2):165-185.

16. Wang L. *Improving External Validity of Epidemiologic Analyses by Incorporating Data from Population-Based Surveys* [Doctoral dissertation]. University of Maryland, College Park; 2020. https://drum.lib.umd.edu/handle/1903/26125.

17. Massey JT. Design and estimation for the national health interview survey, 1985–94. US Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics; 1989.

18. Ezzati TM, Massey JT, Waksberg J, Chu A, Maurer KR. Sample design: third National Health and nutrition examination survey. *Vital Health Stat Ser 2 Data Eval Methods Res*. 1992;113:1-35.

19. Hartley HO, Rao JN, Kiefer G. Variance estimation with one unit per stratum. *J Am Stat Assoc*. 1969;64(327):841-851.

20. National Center for Health Statistics. *National Death Index User's Guide*. Hyattsville, MD: National Center for Health Statistics; 2013. https://www.cdc.gov/nchs/data/ndi/ndi_users_guide.pdf. Accessed April 2021.

21. National Center for Health Statistics. *Data Linkage Public-Use Linked Mortality File Data Dictionary*. Hyattsville, MD: National Center for Health Statistics; 2015. https://www.cdc.gov/nchs/data/datalinkage/public-use-2015-linked-mortality-files-data-dictionary.pdf. Accessed April 2021.

22. National Center for Health Statistic. *Data Linkage Underlying and Multiple Cause of Death Codes*. Hyattsville, MD: National Center for Health Statistics; 2018. https://www.cdc.gov/nchs/data/datalinkage/underlying_and_multiple_cause_of_death_codes.pdf. Accessed April 2021.

23. Lee S, Valliant R. Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociol Methods Res*. 2009;37(3):319-343.

24. Rivers D. Sampling for web surveys. Paper presented at: Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods; 2007; Salt Lake City, Utah.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.