HHS Public Access

Author manuscript

J Off Stat. Author manuscript; available in PMC 2023 September 01.

Published in final edited form as:

J Off Stat. 2022 September ; 38(3): 875–900. doi:10.2478/jos-2022-0038.

Variable inclusion strategies through directed acyclic graphs to adjust health surveys subject to selection bias for producing national estimates

Yan Li[1,*], **Katherine E. Irimata**[2], **Yulei He**[2], **Jennifer Parker**[2]

[1]Joint Program in Survey Methodology, Department of Epidemiology and Biostatistics, 1218 Lefrak Hall, University of Maryland College Park, College Park, MD 20742

[2]Division of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD 20782

Abstract

Along with the rapid emergence of web surveys to address time-sensitive priority topics, various propensity score (PS)-based adjustment methods have been developed to improve population representativeness for nonprobability- or probability-sampled web surveys subject to selection bias. Conventional PS-based methods construct pseudo-weights for web samples using a higher-quality reference probability sample. The bias reduction, however, depends on the outcome and variables collected in both web and reference samples. A central issue is identifying variables for inclusion in PS-adjustment. In this paper, directed acyclic graph (DAG), a common graphical tool for causal studies but largely under-utilized in survey research, is used to examine and elucidate how different types of variables in the causal pathways impact the performance of PS-adjustment. While past literature generally recommends including all variables, our research demonstrates that only certain types of variables are needed in PS-adjustment. Our research is illustrated by NCHS' Research and Development Survey, a probability-sampled web survey with potential selection bias, PS-adjusted to the National Health Interview Survey, to estimate U.S. asthma prevalence. Findings in this paper can be used by National Statistics Offices to design questionnaires with variables that improve web-samples' population representativeness and to release more timely and accurate estimates for priority topics.

Keywords

kernel weighting; logistic regression; propensity score model; survey inference

1. Introduction

Producing timely data is a priority of National Statistics Offices (NSOs). However, some of the more timely data collections, including web-based surveys, may be subject to biases relative to large nationally representative surveys conducted by NSOs due to lower coverage

* yli6@umd.edu .

and response rates. Adjusting these timelier sources with less timely but higher quality reference surveys may decrease their biases.

Selection bias has been acknowledged in different areas (Hernán 2004) and is becoming more critical in the big data era with the rapid emergence of various web surveys to address time-sensitive priority topics, referred to here as target samples. Data collected in target samples, such as web panels, can result in attrition and response rates are often found to be 10% or lower (Baker et al., 2013). Although low response is not necessarily indicative of response bias (Groves and Peytcheva, 2008; Brick and Tourangeau, 2017), selection bias has been of great concern because the composition of web panels often differs from that of the underlying population. Panel members tend to be more educated and to have higher socioeconomic status than non-panel-members (Craig et al., 2013). Epidemiologic target samples often recruit “healthy volunteers” with lower estimates of disease incidence and mortality that a general population (Pinsky et al., 2007). To reduce the selection bias of the target samples, various propensity score (PS)-based adjustment methods have been developed which use an existing high-quality probability sample (e.g., national representative surveys) as a reference, where high quality refers to probability-sampled surveys with relatively low sampling and non-sampling errors that lead to confidence in their ability to produce representative estimates (Groves, 1989). Recent PS adjustment methods include, but are not limited to, PS weighting (Valliant, 2020; Chen et al., 2019) and PS matching (Kern et al., 2020) methods.

The amount of bias reduction, however, varies depending on the outcome and variables that are collected in both the target and reference data sources. Wang et al. (2020a) studied the bias reduction of different PS adjustment methods using the non-representative U.S. National Institutes of Health–American Association of Retired Persons cohort (NIH-AARP, 2006), with the National Health Interview Survey (NHIS) as the reference survey. Among the ten selected diseases examined, they found the amount of relative bias reduction ranged from 8% to 30% using their proposed PS-based kernel weighting (KW) method. There is still a large amount of bias that is not removable by PS adjustment methods alone due to the uncollected information in the reference probability sample. The effectiveness of PS adjustment methods depends on the identification of the proper set of covariates, their availability, and their quality (Baker et al., 2013). Some references (e.g., Mercer et al. 2018) have even argued that choosing the correct variables can be more important than choosing the correct adjustment models, including PS methods.

High-quality probability samples surveyed through well-designed questionnaires are in great demand as reference surveys for at least two reasons: 1) Different PS adjustment methods, including PS-based weighting and matching methods, require a high-quality probability sample as the reference in order to create a set of pseudo-weights for the target sample to better represent the underlying target population; 2) Different target samples may use a common high-quality probability sample as the reference for cost efficiency by using the same questions with exact wordings to avoid potential reporting/measurement error. Given a high-quality population representative reference survey, we are interested in identifying the types of variables that are critical for collection in the target sample to improve its external

validity in estimating population quantities. The findings can be used in turn to plan for future surveys.

The target sample motivating this research is collected through the National Center for Health Statistics' (NCHS) Research and Development Survey (RANDS), a probability-based panel survey that has been conducted using online and phone administration (https://www.cdc.gov/nchs/rands). Although the RANDS data are more structured than nonprobability samples, RANDS is subject to potential selection bias as RANDS has lower response rates, as well as potential measurement and coverage errors compared to traditional interviewer-administered national population health surveys (Parker et al., 2020). On the other hand, probability survey panels such as RANDS have the potential to produce more timely information than national population surveys. To reduce the potential selection bias in RANDS estimates, NCHS' NHIS has been used as a reference sample to construct pseudo-weights using PS-based weighting methods (Parker et al., 2020; Irimata et al., 2020) and raking. These adjustments have been applied to the estimation of several population health outcomes, including diagnosed asthma, diagnosed hypertension, diagnosed diabetes, health insurance, as well as for health outcomes related to the coronavirus disease 2019 (COVID-19) pandemic such as access to health care. In these studies, adjustment to the NHIS (PS-based weighting or raking) has typically been performed using the main effects for all common covariates between RANDS and the NHIS, including sociodemographic, health, and internet use variables; the adjustment for RANDS during COVID-19 used a limited subset of variables for the public release of COVID-19-related estimates (https://www.cdc.gov/nchs/covid19/rands.htm). While PS weighting and raking adjustments have been shown to improve RANDS estimates relative to those without any adjustment to the NHIS (Parker et al., 2020; Irimata et al., 2020), stability of the estimated propensity scores and how the inclusion of different variables in the propensity model or calibration affects bias and efficiency of the estimated population mean for various outcomes have been a major concern.

Propensity model variable inclusion has been widely studied in different areas, including clinical trial or medical research and survey research. In clinical trial research, participants are included for clinical and experimental purposes (mainly for treatment effect estimation) and are not necessarily representative of the U.S. population. Simulations (Brookhart et al., 2006; Leyrat et al., 2013) were performed to examine the effect of the choice of variables that are included in a propensity model has on the bias, variance, and mean squared error of estimated treatment effects. It was concluded that omitting confounding factors increases bias and the inclusion of variables that are independent of the exposure but related to the outcome in the propensity model gains efficiency without increasing bias of estimated treatment effects. However, covariate inclusion for propensity score models in clinical trial research has been limited. Ali et al. (2015) provided a systematic review of covariate inclusion in the PS model for medical studies and concluded that the quality of reporting variable inclusion is far from optimal in the medical literature. Similarly, Grose et al. (2020) found 90% out of 303 systematically reviewed studies did not provide justification for covariates included in their PS models.

In survey research, propensity analyses have been conducted to estimate response propensity (Groves, 2006; Iannacchione et al., 1991) and to adjust sampling weights in representative surveys to reduce the estimation bias due to unit nonresponse. The best auxiliary variables to be included for nonresponse adjustment are those simultaneously correlated with response propensity and the key survey outcomes (Lessler and Kalsbeek, 1992). Little and Vartivarian (2005) further suggested that most important feature of variables for inclusion is that the variables are predictive of survey outcomes; prediction of response propensity is a secondary, though useful, goal.

This paper, in contrast to the interest of estimating treatment effects in clinical research, aims to estimate population quantities such as the population mean. We are interested in identifying key auxiliary information in a reference probability survey to improve the external validity of inferences from a target dataset. This is an important obligation for survey designers because the choice and inclusion of these variables has a tremendous effect on both the bias and the precision of the estimates of population quantities. This differs from the goal of nonresponse adjustment which uses chosen covariates for predicting response propensities as, in nonresponse adjustment research, respondents are nested within the sampled units, and respondents and nonrespondents share common sampling design variables. As a result, unweighted analysis of response propensity can be performed conditional on the design and response predictive variables (Little and Vartivarian, 2003). However, this is not true for estimating the propensity of target sample inclusion because the reference survey and the target samples are often independent without sharing design variables (Wang et al., 2021). Variable choice for the propensity model used to predict the target sample inclusion propensity should be performed with additional care.

This paper aims to examine how different types of variables included in a propensity model impact the performance of population mean estimation using target samples through the directed acyclic graph (DAG), a common graphical tool in causal studies but largely under-utilized in survey research. The DAG is used to identify certain types of variables in the causal pathway to be included in the PS model which results in the lowest bias and highest precision under various scenarios. Estimated population means and their variances are evaluated analytically and numerically under various mis-specified propensity models, including with and without interactive effects. Different levels of variable correlations in the finite population are considered to mimic real data scenarios. The findings are applied to RANDS, with NHIS as the reference, to estimate the prevalence of asthma in the U.S. The RANDS evaluation demonstrates the advantage of this approach compared to the approach when the propensity model includes all available variables.

The results from this research provide insight for data analysts on propensity model construction to improve the population representativeness of target samples. It also provides insight for questionnaire designers on the critical auxiliary information to collect from the reference survey. NSOs, using the paper results, can design the questionnaires for both the target and reference surveys and release accurate estimates for priority topics from more timely data sources.

2. Methods

We first introduce some notation. Suppose Y is a binary outcome of interest (e.g., for estimating the prevalence of a disease or health condition: Y=1 if event and 0 otherwise). In the context of survey sampling, suppose A is the binary selection indicator variable (i.e., A=1 if a population unit participates in the target sample and 0 otherwise). Note A indicates the target sample participation with value of one representing population units who are recruited and respond to the survey.

We adapt the framework of Brookhart et al. (2006) of employing a directed acyclic graph (DAG) to study potential selection bias induced by three types of covariates (see Figure 1a):

1. variables related to both the outcome Y and the selection indictor A of the target sample — confounders (X_1);
2. variables related to Y but not related to A – outcome predictors (X_2);
3. variables related to A but not related to Y – selection variables (X_3).

We now present some background about PS adjustment methods. For estimation of the finite population (FP) mean of a binary outcome $E(Y) = p(Y = 1)$, the naïve unweighted estimate using the selected target sample ($A = 1$) has bias relative to the FP, given by $Bias = p(Y = 1|A = 1) - p(Y = 1) = (R - 1)p(Y = 1)$ with $R = \frac{p(A = 1 \mid Y = 1)}{p(A = 1)}$. In order to remove the bias, it requires the conditional distribution of A given Y is the same as the distribution of A, denoted by $A \perp Y$ and adjustment methods based on PS are often employed (Valliant 2020).

More specifically in PS-based adjustment methods, the population mean μ of the outcome Y, is estimated by

$$\hat{\mu}(x) = \frac{1}{\sum_{j \in S_c} w_j(x)} \sum_{j \in S_c} w_j(x) y_j, \tag{1}$$

where S_c denotes the set of sample units in the target sample of size n_c; y_j for $j \in S_c$ and x are, respectively, realized values of the outcome Y and X; the pseudo-weights $w_j(x)$ for $j \in S_c$ is constructed to balance the distribution in covariates between the target sample and the reference survey. Note X can be a single covariate or a vector of covariates from X_1-X_3, and are available in both the target sample and the reference survey, while the outcome variable Y is available in the target sample S_c only.

Various PS-based adjustment methods, including PS weighting and PS matching methods, have been developed under the following assumptions. First, the reference survey sample (in our real data example, the NHIS), through weighting, properly represents the target population of interest. Second, all finite population units have a positive participation rate (i.e., each individual in the population has a positive propensity to volunteer to participate in RANDS panel). Third, conditional exchangeability holds with no unmeasured confounders, that is, the probability for each individual in the FP to participate in the target sample is not related to his/her outcome, after adjusting for all measured variables. It is a common practice

that the variables in the target sample are measured using same question wordings as in the reference survey to avoid potential reporting or measurement error.

While PS weighting and PS matching methods have similar assumptions, PS *weighting* methods construct the pseudo-weight by the inverse of the inclusion probability conditional on x, i.e., $w_j = \frac{1}{e(x_j)}$ for $j \in S_c$, with $e(x) = p(A = 1|x)$, the target sample inclusion probability conditional on x. It can also be verified that $A \perp x|e(x)$. In contrast, PS *matching* methods distribute the survey sample weights to target sample units that have similar predicted propensity scores. For example, the KW method (Wang et al. 2021) first assigns the sample weight of each survey unit, say unit i, to cohort members proportionally according to kernel distances, defined by propensity scores $K\left(\frac{e(x_i) - e(x_j)}{h}\right)$ for $j \in S_c$, where $K(.)$ a kernel function such as the standard normal density function, and h is the bandwidth selected by Silverman's rule of thumb method (Silverman, 1986). As such, most of the sample weight for survey unit i is assigned to those cohort members with similar propensity scores. The assigned portions from survey members to cohort member j are then summed up to form the pseudo-weight w_j.

In sections 2.1-2.2, we assume that X_1, X_2, and X_3 are mutually independent in the FP and study how the PS-based adjustment methods reduce the bias and variance through the incorporation of different types of variables in the propensity models. We further consider real situations in section 2.3 when different types of variables are correlated in the FP using DAG. Although DAG is a graphical tool developed for causal interpretation, we used it to rule out possible confounding and identify conditioning covariate set for $Y \perp A$. The actual causation is not important in this context. Section 2.4 summarizes some practical guidelines for identifying the variable types in real data and choosing between PS-based methods to construct pseudo-weights when covariates interactively affect the target sample participation and the outcome.

2.1 Bias of $\hat{\mu}(x)$ by various types of covariates

It is readily shown in Figure 1a that the confounders X_1 induce the bias when we use the simple sample mean to estimate the population mean $p(Y = 1)$. Intuitively, the information can be exchanged between the two nodes of A and Y through X_1, but not X_2 or X_3. This result is consistent with the bias calculation below. For selection variables ($X = X_3$) or predictors ($X = X_2$), we have $R = \frac{\sum_x [p(Y = 1 \mid X = x)p(A = 1 \mid X = x)p(X = x)]}{p(Y = 1)\sum_x [p(A = 1 \mid X = x)p(X = x)]} = 1$ and hence $Bias = 0$. For confounders, however, $Bias \neq 0$ To correct for the bias induced by X_1, PS-based adjustment methods create pseudo-weights and reweight the target sample such that the weighted sample distribution of the confounder X_1 is same as that in the FP, i.e., $X_1 \perp A$ as shown in Figure 1b. The dotted line denotes the path X_1-A is blocked (i.e., there is no information exchange between the two nodes) by reweighting the target sample and hence $A \perp Y$.

As a result, the estimator $\hat{\mu}(x_1)$ with the set of pseudo-weights $w(x_1)$, where x_1 is the realized value of the confounder X_1, is approximately unbiased. Analogously, it is readily shown that

the estimator $\hat{\mu}$ with pseudo-weights defined by the inverse of sample inclusion probabilities that balance the x_1 distribution between the target sample and the FP, including $e(x_1, x_2)$, $e(x_1, x_3)$, or $e(x_1, x_2, x_3)$, is also unbiased. Note that the three sets of pseudo-weights of $w(x_1, x_2)$, $w(x_1, x_3)$, or $w(x_1, x_2, x_3)$ balance the x_1 distribution and also the distribution of x_2, x_3, or x_2 and x_3, respectively, between the target sample S_c and the FP (Rosenbaum and Rubin, 1983).

In contrast, pseudo-weights of $w(x_2)$, $w(x_3)$ or $w(x_2, x_3)$ do not balance the X_1 distribution and therefore the corresponding weighted estimators in (1) are biased.

2.2 Variance of $\hat{\mu}(x)$ by various types of covariates

Among the four unbiased estimators based on $e(x_1)$, $e(x_1, x_2)$, $e(x_1, x_3)$, and $e(x_1, x_2, x_3)$, we compare their efficiencies. We first compare the variance of $\hat{\mu}(x)$ with x the realized value of $X = X_1$ versus $X = (X_1, X_3)$, denoted by $V(\hat{\mu}(x_1))$ and $V(\hat{\mu}(x_1, x_3))$, respectively. We write

$$V(\hat{\mu}(x_1)) = V\left(\frac{1}{\sum_{j \in S_c} w_j(x_1)} \sum_{j \in S_c} w_j(x_1) y_j\right), \text{and}$$

$$V(\hat{\mu}(x_1, x_3)) = V\left(\frac{1}{\sum_{j \in S_c} w_j(x_1, x_3)} \sum_{j \in S_c} w_j(x_1, x_3) y_j\right)$$

Note the selection variable is independent of the outcome and thus the pseudo-weights based on x_3 are non-informative of the outcome Y. The corresponding pseudo-weighted mean, although adding no bias, loses efficiency due to the differential non-informative pseudo-weights. Taking the adjusted logistic propensity pseudo-weights (denoted by ALP in Wang et al. 2021) as an example, under the logistic regression propensity model

$$\log \frac{p_j}{1 - p_j} = \log \pi_j = \beta_x^T x_j, \tag{2}$$

where $p_j = p(j \in S_c^* \mid FP^*)$ and $FP^* = S_c^* \cup FP$ denotes the pseudopopulation by combining S_c^* (i.e., a copy of the target sample S_c) and the FP, and β_x is the regression coefficient associated with x. The ALP pseudo-weight $w_j(x)$ is constructed as $w_j(x) = \exp^{-1}\left(\beta_x^T x_j\right)$ for $j \in S_c$.

For simple illustration, assume model (2) includes main effects of covariates, so $w_j(x_1, x_3) = w_j(x_1) w_j(x_3)$ and $w_j(x_3)$ are noninformative weights since $x_3 \perp y$. Under the assumption that the variance of the observations is approximately constant (Kish, 1992), the proportional increase in variance from weighting, denoted by L, is approximated to be

$$L = \frac{V(\hat{\mu}(x_1, x_3))}{V(\hat{\mu}(x_1))} - 1 = CV^2\left(w_j(x_3)\right) > 0,$$

where CV is the coefficient of variation of the $w_j(x_3)$ weights. Thus, $V(\hat{\mu}(x_1, x_3)) > V(\hat{\mu}(x_1))$.

Note that the model parameter $\beta_x = 0$ in (2) if x is an outcome predictor, which does not predict the target sample membership A, such as x_2 in Figure 1a. As a result, $w(x_1, x_2) = \exp^{-1}\left(\beta_{x_1}^T x_1 + 0\right) = w(x_1)$ and thus

$$V\left(\hat{\mu}(x_1, x_2) = \frac{1}{\sum_{j \in S_c} w_j(x_1, x_2)} \sum j \in S_c w_j(x_1, x_2) y_j\right) = V\left(\hat{\mu}(x_1) = \frac{1}{\sum_{j \in S_c} w_j(x_1)} \sum j \in S_c w_j(x_1) y_j\right).$$

Along the same line of justification, $V(\hat{\mu}(x_1, x_3)) = V(\hat{\mu}(x_1, x_2, x_3))$. In conclusion,

$$V(\hat{\mu}(x_1)) = V(\hat{\mu}(x_1, x_2)) < V(\hat{\mu}(x_1, x_3)) = V(\hat{\mu}(x_1, x_2, x_3)).$$

In summary, to achieve unbiasedness and efficiency of pseudo-weighted mean estimators, the propensity model that considers confounders (X_1) alone, or together with outcome predictors (X_2), should be used to construct the pseudo-weights in eq(1). The resulting mean estimates are unbiased and most efficient. The inclusion of selection variables in the propensity model, in addition to all confounders, adds no more bias, however, will inflate the variances of the estimates. In contrast, the inclusion of outcome predictors does not inflate the variance while retaining the unbiasedness of the FP mean estimates. A short version of the recommendation for PS-based pseudo-weights construction: include all confounders but avoid selection variables in the propensity model.

The above justification assumes the logistic regression model (2) with main effects is true. More rigorous justification is needed when different types of covariates are correlated; propensity models are mis-specified (that is, the logistic regression model (2) is not the true model); and the pseudo-weights (i.e., propensity model coefficients) are unknown and have to be estimated.

2.3 Correlation between covariates

We now consider more realistic scenarios in which the confounders, the outcome predictors, and the selection variables can be correlated to each other. Figure 2 shows cases where correlation exists between the pairs X_1 and X_3, X_1 and X_2, and X_2 and X_3, respectively. In addition, any two or all three pairs can be correlated simultaneously in the FP.

For unbiased estimation of the FP mean of Y using the target sample (A=1), we need to block all paths connecting A and Y such that $Y \perp A$. We focus on paths that point to A since in the propensity model we construct weights for the target sample units (with A=1) so that the weighted target sample and the FP have same distributions in certain covariates X, i.e. A⊥Xle(X).

As shown by the dotted lines in Figure 2, two paths in Figure 2a and 2c, and one path in Figure 2b are identified and need to be blocked, i.e., prevent information flow between A and Y, in order to achieve $Y \perp A$. The backdoor criteria (Pearl, 2009) is a way to rule out confounding via conditioning, and allows identifying the causal effect from A to Y (equal to

zero in Figure 2, i.e., $Y \perp A$) after conditioning a set of covariates that block the backdoor paths between A and Y. Here the identified paths in Figure 2 are backdoor paths because the arrows point into A (not the opposite direction if arrows point from A towards X_1-X_3). By the backdoor criteria, X_1 blocks the identified paths in Figure 2a-b. As follows, we construct PS-based pseudo-weights $w(x_1)$ so that the X_1 distribution in the weighted target sample is same as that in the FP in Figure 2a-b when X_1 and X_3 or X_1 and X_2 are correlated (i.e., $\rho_{x_1x_3} \neq 0$ or $\rho_{x_1x_2} \neq 0$). Thus, the $w(x_1)$-weighted target sample mean of Y is an unbiased estimator of the FP mean. In Figure 2c, X_2 or X_3, in addition to X_1, block the two identified paths (Pearl, 2009). Following the same logic, pseudo-weights that balance the distributions in X_2 or X_3, in addition to X_1, denoted by $w(x_1, x_2)$ or $w(x_1, x_3)$, should be constructed for the target sample units when the pair of X_2 and X_3 are correlated in Figure 2c ($\rho_{x_2x_3} \neq 0$). This result also applies to cases when any two pairs or all three pairs of covariates are simultaneously correlated in the FP, and the $w(x_1, x_2)$- or $w(x_1, x_3)$-pseudo-weighted target sample means are approximately unbiased.

In summary, similar to the scenario shown in Figure 2, in order to block the dotted paths when covariates interactively affect the outcome, PS-based adjustment methods can be applied to construct pseudo-weights that balance the distributions of X_1 (Figure 2a-b) and (X_1, X_2) or (X_1, X_3) (Figure 2c) in the pseudo-weighted sample and the FP.

2.4 Practical guidelines

In practice, the variable types (confounder, predictor, selection variable) need to be identified for propensity model construction. Since we are not concerned about model interpretation, parametric models with complex functional forms or nonparametric models can be fitted. In our RANDS example (Section 4), both the outcome and propensity models were selected by automatic backward selection methods. Starting from the full model containing all factors and their pairwise interaction terms, we removed the interaction term with the largest p-value and re-fit the model. We continued the iterations until all p-values of the interaction terms were less than 0.05. For each interaction, complex survey designs were accounted for in the logistic regression analysis using the svyglm() function in the R survey package (Lumley, 2020). The main effects with p-values greater than 0.05 were removed only if they were not involved in any of significant interaction terms. As results, each type of variables is identified: confounders are common terms in both the selected propensity and the outcome models; the selection variables (or predictors) are those selected in the propensity (or outcome) model only. Note that each type of variables may contain multiple variables as well as their nonlinear or nonadditive functions (e.g., pairwise interactions) in the final outcome and the propensity models.

Alternative model selection criteria can be employed, such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) (Lumley and Scott, 2015) to identify variable types. Yang et al. (2020) and Chen, Valliant, and Elliott (2018) have also proposed variable selection methods, including penalized estimating equations or LASSO regression, which can be used to identify variable types for inclusion in the PS model. In the case where the outcome of interest is not available in the reference probability sample or the outcome has not yet been collected in the target sample, subject matter literature and knowledge

may have to be used to assign the covariate types. Variable type identification is critical in practice and comparing different model selection methods to create the final models is of future research interest.

The true propensity model of the underlying selection mechanism of the target sample (A=1) is often unknown but complicated, which may involve covariate terms of higher orders of nonlinearity and/or nonadditivity. For example, X_1 and X_2 (or X_1 and X_3) can interactively affect the outcome Y (or selection indicator A), the scenario considered in simulation study 3 (to be shown in section 3). In order to estimate the propensity scores accurately to achieve the covariate balance so that the condition of $Y \perp A$ holds, data analysts need to be careful in choosing the PS-based adjustment methods among PS weighting or matching methods based on parametric models such as logistic regression and nonparametric methods such as machine learning.

For example, PS weighting methods (such as the ALP) can be sensitive to model misspecification (Wang et al., 2020a). ALP-weighted target sample distributions match the FP distributions when the assumed propensity model is true. For instance, ALP pseudo-weights that are constructed based on the propensity model of A on X_1, X_3, and their interaction $X_1 * X_3$, produce unbiased estimators. The estimators, however, are biased if the model is misspecified, for example, the interaction term is omitted from the propensity model.

In contrast, PS *matching* methods construct weights by matching target sample units with reference sample units using the estimated propensity scores, followed by distributing reference sample weights to target sample units with similar propensities. It has been well recognized that PS matching, compared to PS weighting methods, is more robust to model misspecification (Wang et al., 2020a). In our limited simulation studies, it is shown that the balance in covariate distributions between the KW-weighted sample and the FP can be achieved as long as the blocking variables, that is, X_1 in Figure 2a-b; X_1 and X_3 or X_1 and X_2 in Figure 2c, are included (with or without $X_1 * X_3$ interaction) in the propensity model (as shown in Figure 3). For complicated propensity models with higher orders of interactions and/or nonlinearity, including only the main effects of the blocking variables in the propensity model by KW methods might not be sufficient. Nonparametric modeling such as machine learning methods may be promising (Kern et al., 2020) in identifying nonlinear or nonadditive relationships of covariates with the target sample selection.

3. Simulation

Simulation studies were conducted to evaluate the performance of the mean estimator from (1) with the pseudo-weights constructed by the ALP and the KW methods based on propensity models that consider different types of covariates.

Population Generation

We generate a finite population $FP = \{X_{1i}, X_{2i}, X_{3i}, Y_i \text{ for } i = 1, \ldots N\}$ with population size $N = 20{,}000$. Three covariates X_1, X_2, and X_3 follow standard trivariate normal distributions

with pairwise correlations $\rho_{x_1x_2}$, $\rho_{x_1x_3}$, $\rho_{x_2x_3}$. A binary outcome Y is generated following the Bernoulli distribution with a mean of

$$p(Y=1)=\frac{\exp(\alpha_0+\alpha_1X_1+\alpha_2X_2+\alpha_{12}X_1X_2)}{1+\exp(\alpha_0+\alpha_1X_1+\alpha_2X_2+\alpha_{12}X_1X_2)}.$$

We specify $(\alpha_0, \alpha_1, \alpha_2) = (-1, .5, .5)$ so that x_1 and x_2 are associated with Y as in Figure 1, but vary $\alpha_{12} = 0.5$ or 0 with and without the interaction term. As a result, the FP mean $\bar{Y} \approx 0.29$.

Selection of the target sample (with A=1)

A sample of size $n_c = 1{,}000$, denoted by S_c, is selected from the FP, using the design of probability proportional to size (PPS) sampling with measure of size (mos): $mos = \exp(\beta_0 + \beta_1X_1 + \beta_3X_3 + \beta_{13}X_1X_3)$ so that the inclusion probability is

$$p(j\in S_c \mid FP)=\frac{n_c\times mos_j}{\sum_{i\in FP}(mos)_i}\propto \exp(\beta_1X_{1j}+\beta_3X_{3j}+\beta_{13}X_{1j}X_{3j}).$$

We specify $\beta = (\beta_0, \beta_1, \beta_3) = (-1, .5, .5)$ so that x_1 and x_3 are associated with A as in Figure 1. In addition, we vary $\beta_{13} = .5$ or 0 with or without the interactive effect in the propensity model. We have the target sample participation rate of $E(A) = .05$.

The inclusion probabilities (i.e., sample weights) are masked in the analysis and treated as unknown (i.e., equal sample weights of 1 used). Note that the target sample without weights is not representative of the population.

Selection of a probability sample

An independent probability sample of size $n_s = 500$, denoted by S_s, is selected using the same sampling design as the target sample selection. The selected probability sample has known selection probabilities. The weighted probability sample is used as the reference survey, representing the underlying FP in the propensity analysis.

Pseudo-weighted means, i.e., (1), with estimated pseudo-weights constructed under different propensity models, including the confounders (X_1), outcome predictors (X_2), the selection variables (X_3), and/or their interactions, were compared. Three simulation studies are conducted with results presented in Tables 1-2 and Figure 3. Simulation 1 considers a simple scenario of independent covariates in the FP (with $\rho_{x_1x_2} = \rho_{x_1x_3} = \rho_{x_2x_3} = 0$) without interaction effects of covariates on the outcome or the target sample inclusion (i.e., $\alpha_{12} = \beta_{13} = 0$). Simulation 2 varies the covariate correlation in the FP by ($\rho_{x_1x_2}$, $\rho_{x_1x_3}$, $\rho_{x_2x_3}$) = (.6,0,0), (0,.6,0), (0,0,.6), (.6,.6,0), (.6,0,.6), (0,.6,.6), or (.6,.6,.6) , while keeping $\alpha_{12} = \beta_{13} = 0$. Simulation 3 further complicates the underlying outcome model and the propensity model by including the interaction terms with $\alpha_{12} = \beta_{13} = 0.5$.

Tables 1-2 show the bias, empirical variance (EmpVar), and MSE of the KW estimate, over B=500 iterations, from simulations 1-2, respectively, and

$$bias = \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}^{(b)} = \bar{Y}; EmpVar = (B-1)^{-1}\sum_{b=1}^{B}\left\{\hat{\mu}^{(b)} - B^{-1}\sum_{b=1}^{B}\hat{\mu}^{(b)}\right\}^2; \text{ and}$$
$$\text{MSE} = \frac{1}{B}\sum_{b=1}^{B}\left\{\hat{\mu}^{(b)} - \bar{Y}\right\}^2,$$

where $\hat{\mu}^{(b)}$ is the KW estimate of the population mean using the bth simulated target sample under various analytical propensity models. The w(x_1), w(x_{12}), and w(x_{13}) denote the propensity models including main effects of, respectively, x_1, x_1 and x_2, x_1 and x_3. Models including x_2 only, x_3 only, and x_2 and x_3 are denoted as w(x_2), w(x_3), and w(x_{23}), respectively.

Three observations are made in Table 1. *Firstly*, consistent with our expectations, all propensity models that include the confounder x_1, i.e. w(x_1), w(x_{12}), w(x_{13}), produce approximately unbiased estimates of the FP mean of Y; the estimates are badly biased under the propensity models which include x_2 only, x_3 only, or x_2 and x_3. *Secondly*, the propensity model w(x_3) yields inflated variance estimates compared to w(x_1) or w(x_2), and w(x_2) has the smallest empirical variances. *Thirdly*, among the three approximately unbiased estimators, w(x_1) yields the most efficient estimates relative to w(x_{12}) or w(x_{13}).

Table 2 presents results from simulation 2 with varying covariate correlations. Three observations are made. *Firstly*, pseudo-weights that balance the distributions in x_2 or x_3, in addition to x_1, produced approximately unbiased estimates across various correlations; see the shaded two columns of w(x_{12}) and w(x_{13}). *Secondly*, among the two, w(x_{12}) and w(x_{13}), the empirical variance estimates and MSEs under w(x_{12}) tend to be smaller than those under w(x_{13}). *Thirdly*, the inclusion of only the confounder x_1 in the propensity model, i.e., $w(x_1)$, although efficient, may induce bias, especially when correlation exists between x_2 and x_3.

Simulation 3 compares biases of estimated population means by the KW matching method and the ALP weighting method when the underlying outcome and propensity models include the interaction terms, i.e., $\alpha_{12} = \beta_{13} = 0.5$ (see Figure 3). Four analytic propensity models, including X_2 or X_3 in addition to the confounder x_1, are considered and they are 1) w(x_{12}), X_1 and X_2 main effects only, 2) w(x_{13}), X_1 and X_3 main effects only, 3) w(x_1 * x_2), X_1 and X_2 main effects and their interaction, and 4) w(x_1 * x_3), including X_1 and X_3 main effects and their interaction. Recall KW is a type of PS matching method and expected to be more robust to model misspecification compared to the ALP method. As expected, the KW method consistently yields approximately unbiased estimates across four propensity models with or without interaction terms. In the contrast, the ALP approach directly uses the inverse of the participation rates estimated from the assumed propensity model as pseudo-weights, and the ALP estimates are approximately unbiased only under the true propensity model w(x_1 * x_3). Furthermore, it can be observed that biases of the ALP estimates are consistently closer to zero than the KW under the true model. Results with covariate correlations ($\rho_{x_1x_2}$, $\rho_{x_1x_3}$, $\rho_{x_2x_3}$) = (.6,0,0), (0,0,.6), (.6,0,.6) and (0,.6,.6) showed a similar pattern and hence are not shown.

4. Real data analysis

RANDS, a series of web-based probability panel surveys conducted at NCHS (https://www.cdc.gov/nchs/rands), has been used for methodological research and, more recently, for providing early experimental estimates on the COVID-19 pandemic. RANDS has the capability to collect data quickly and is less costly than traditional national household surveys, but is subject to potential selection bias due to low response rates. Adjustment methods to construct pseudo-weights, including propensity-score based methods, are applied to balance the covariate distributions in the target sample and the FP, and are an important component of the RANDS program. We consider the simulation findings from this paper for selecting PS-model covariates to estimate the national prevalence of asthma compared to NCHS' NHIS.

Data from the third round of RANDS (RANDS 3) is evaluated. RANDS 3 was collected in 2019 using NORC's AmeriSpeak® Panel (https://amerispeak.norc.org) and included responses from 2,646 panelists aged 18 years and older. RANDS 3 panelists were surveyed via web and were asked questions related to general and mental health, medical conditions, opioid use, and pain. The RANDS 3 cumulative response rate was 18.1%. The RANDS 3 original panel weights were developed by the inverse of the probability of inclusion in the AmeriSpeak® Panel, subject to nonresponse adjustment and poststratification adjustment to external population totals of age, sex, education, race/ethnicity, housing tenure, telephone status, and Census Division (National Center for Health Statistics, 2020). The original panel-weighted estimate of diagnosed asthma (ever been told you had asthma) in RANDS 3 was 16.86% (standard error = 0.98%). For comparison, the unweighted estimate of diagnosed asthma in RANDS 3 was 16.40% (standard error = 0.72%). The 2019 NHIS (n=31,997) is evaluated as the gold standard. The NHIS (https://www.cdc.gov/nchs/nhis) is a cross-sectional household interview survey that collects information on a broad range of health topics, primarily through face-to-face interviews. The NHIS sample adult file, which is a collection of responses from one randomly selected adult per selected household, was used to evaluate the prevalence of ever having asthma among adults. The percentage of adults who ever had asthma based on the 2019 NHIS (n=31,997) was 13.46% (standard error = 0.25%).

Common covariates available in RANDS 3 and the 2019 NHIS that were potentially related to diagnosed asthma or the selection indicator were considered (see Table 3). All percent estimates in Table 3 (when expressed as proportions) meet the NCHS data presentation standards for proportions (Parker et al., 2017). As observed, demographic variables of age, sex and race/ethnicity have similar weighted distributions in the RANDS and NHIS. This result is as expected, since these variables are poststratification variables used to construct the sample (or original panel) weights in both NHIS and RANDS. As observed, persons with higher levels of education, with selected health conditions (i.e., diagnosed high cholesterol, diagnosed chronic obstructive pulmonary disease (COPD), emphysema, or chronic bronchitis, diagnosed diabetes, and diagnosed hypertension), who are current or former smokers, or who are not married (with the exception of those who are widowed) participated in RANDS at a higher rate compared to the NHIS. Since the percent of missing values across all considered variables was relatively low for both data sources, ranging

from 0%-0.68% for RANDS and 0%-2.64% for NHIS (unweighted), missing values were excluded for evaluation.

To check for correlation between covariates, bivariate correlations were assessed on the weighted NHIS data. Bivariate correlations for all selected covariates were statistically significant. Prior to evaluating the propensity models, the survey weights for both data sets were normalized to their respective sample sizes (n=2,646 for RANDS, n=31,997 for NHIS) as suggested by Li et al. (2011) and Wang et al. (2021). The KW method was implemented for demonstration to construct RANDS pseudo-weights that adjust for potential selection bias due to differential non-response and under-coverage of some groups on the sample frame using the NHIS data as the reference dataset.

A full propensity model (denoted by model.all) that includes all covariates and their pairwise interactions was used to create pseudo-weights. Due to the large number of parameters in the full model, estimated propensity scores can be unstable. As a result, some form of stepwise propensity model selection methods have been conducted in different studies (Weitzen et al., 2004; Austin 2008; Wang et al., 2020a), using the combined target sample and the reference sample to identify significant terms out of the full propensity model. In the framework of our paper, the combination of the confounders and selection predictors (i.e., model.x13 which contains X_1 and X_3), which can be main effects of covariates or their nonlinear/nonadditive combinations such as pairwise interactions, are recommended as terms for inclusion. Based on the simulation results, we expect that the pseudo-weighted mean under model.x13 would be unbiased but with higher variability when compared with the estimates under model.x12 that includes the confounders and outcome predictors.

Accordingly, we conducted the outcome model selection using backward selection on the reference survey (e.g., the 2019 NHIS), to identify terms which were confounders or outcome predictors. We defined the selected model as model.x12.n (contains X_1 and X_2) with “n” indicating that the outcome predictors were identified using the NHIS. However, it is often the case that the reference probability surveys have no collected information on the outcome variable. In this case, we have only the target sample (e.g., RANDS) available for outcome model selection. With the assumption of the conditional noninformative sampling of the target sample, it is expected the unweighted regression of the outcome would produce unbiased estimates of regression coefficients (Korn and Graubard, 1999). As follows, outcome model variable selection was conducted based on the unweighted outcome regression of the RANDS data, and the selected model included both confounders and outcome predictors, denoted by model.x12.r (contains X_1 and X_2) indicating that the outcome predictors were identified using RANDS. The common terms in model.x13 and model.x12 (denoted by either x12.n or x12.r based on the information available) are confounders, and the corresponding propensity model is denoted by model.x1. The identified covariate types under each model are reported in the Appendix. Due to the correlation between x1, x2 and x3, we expect estimates under model.x1, albeit efficient, may not remove as much bias as under the model.x12 or model.x13.

The outcome models utilized the observations in the NHIS or the RANDS, whereas the propensity model utilized the observations in the combined NHIS and RANDS data, from

which the estimated propensities were obtained and used for construction of the KW pseudo-weight for each individual in RANDS. Note that RANDS has panel weights, which were computed as an overall sampling weight for the selection of each panel member from the sampling frame and the selection of the panel member into RANDS. We considered two scenarios of 1) panel weights or 2) no panel weights for the propensity analysis.

Various propensity models that included different types of covariates were evaluated by the coefficient of variation (CV) of the KW pseudo-weights ($CV = sd(KW)/mean(KW)$ with sd denoting standard deviation), relative bias (relBias $= \frac{\hat{\mu}_{RANDS} - \hat{\mu}_{NHIS}}{\hat{\mu}_{NHIS}} \times 100\,\%$), standard error (se), and mean squared error (MSE $= (\hat{\mu}_{RANDS} - \hat{\mu}_{NHIS})^2 + se^2(\hat{\mu}_{RANDS})$). The relative bias was calculated as the estimated asthma prevalence in RANDS relative to the NHIS estimate where the RANDS estimate, $\hat{\mu}_{RANDS}$, was calculated using the KW pseudo-weights produced from the various propensity models. The standard error $se(\hat{\mu}_{RANDS})$ considered the variability due to estimating the propensity scores, sampling, kernel weighting, as well as differential pseudo-weights by the Taylor linearization method (Wang et al., 2020b). For comparison purposes, we also report the relative bias, standard error, and MSE of the original panel-weighted and unweighted estimates of asthma prevalence in RANDS 3. Results are presented in Table 4.

Four observations can be made from Table 3. *Firstly*, all 12 (panel-weighted or unweighted) propensity-adjusted estimates perform better, with a smaller MSE (or relative bias), compared to the original panel-weighted RANDS estimate of asthma prevalence without PS adjustment. When the RANDS panel weights are considered in the propensity analysis, the standard errors tend to be inflated, relative to those in the lower pane, due to more variable KW pseudo-weights with their CVs ranging from 1.07-1.13 (see the upper pane) *vs.* 0.69-0.83 (see the lower pane). Accordingly, observations 2-4 focus on the results in the lower panel when the panel weights are not used to construct KW pseudo-weights. *Secondly*, consistent with our expectations, the propensity models that contain confounders and selection variables, i.e. Model.x13, produce larger estimated variances compared with Model.x12 irrespective of the outcome predictors being selected from RANDS or the NHIS (e.g., se=0.97 vs. 0.81-0.84). *Thirdly*, comparing the estimates under the propensity models containing the confounders and outcome predictors that are selected using NHIS data (i.e. Model.x12.n) vs. the RANDS data (i.e. Model.x12.r), similar relative bias, se and MSE are observed (relBias = 11.38 vs. 11.37; se=0.81 vs. 0.84, MSE=3.01 vs. 3.04). *Lastly*, the relative biases under Model.x1 are somewhat larger than that under Model.x12 (relBias = 13.67 vs. 11.38 or 13.44 vs. 11.37). This result could be due to the existing correlation between the confounders and the outcome predictors. Adjusting for confounders only may not be sufficient for maximum bias reduction.

In brief, for the evaluation of diagnosed asthma using the RANDS data, we would recommend the pseudo-weights constructed under Model.x12.n with the confounders and predictors selected from the reference survey (e.g., NHIS). In situations where outcome variables are not collected in the reference survey but available only in the target sample

(e.g., RANDS), Model.x12.r can be the alternative model to construct the KW pseudo-weights, assuming conditional noninformative sampling holds for the target sample.

5. Discussion

Identifying and collecting the best information on more timely target sample and on higher quality reference surveys can increase the ability of NSOs to produce timely estimates with lower bias from target samples. This paper examined how different types of variables that are included in a propensity model impact the performance of PS-based pseudo-weighted estimators for population mean estimation from a target sample. Means and variances of estimated population means under various mis-specified propensity models, including different types of variables with and without interactive effects, were evaluated analytically and numerically. Different levels of variable correlations in the finite population were also considered to reflect real data scenarios. We have the following findings: 1) confounders, the variables related to both the selection indicator and the outcome of interest, are important variables to include in the propensity model; 2) pseudo-weights that balance the distributions in the outcome predictor x_2 or the selection variable x_3, in addition to the confounder x_1, denoted by $w(x_1, x_2)$ or $w(x_1, x_3)$, should be constructed for the target sample units so that the corresponding pseudo-weighted target sample mean is approximately unbiased; 3) compared to $w(x_1, x_3)$, the pseudo-weights $w(x_1, x_2)$ gain more efficiency in estimating population means. In contrast, the inclusion of selection variables, compared to the outcome predictors, in the propensity model tended to inflate the estimated variances. Intuitively, the outcome predictor is related to the outcome variable; including outcome predictors in the propensity score model distinguishes differences between the outcome in the reference and target samples, which results in weights related to outcome and therefore yields estimates with smaller variance estimates. Finally, findings are applied to real target data from RANDS, a survey that uses commercial probability panels, which has potential selection bias. Under the model with confounders and outcome predictors (Model.x12) or model with confounders and selection variables (Model.x13), the RANDS estimate of U.S. asthma prevalence had the greatest bias reduction (relative bias ranging from 11.37%-13.51% compared to the NHIS) when the panel weights are not used to construct KW pseudo-weights, compared to the original panel-weighted RANDS estimates (relative bias of 25.31%).

Results from this paper have several important applications in practice for NSOs that collect data from both target surveys and high-quality reference surveys. First, this study provides a principled approach to select covariates for the PS model. Rather than including all variables or selecting certain demographic variables, covariates are assessed based on their variable type (confounder, outcome predictor, selection variable) to be included in the PS model for population mean estimation. Second, guidance on how to design the questionnaire for a target survey with specific research questions (e.g., SARS-CoV2 seropositivity web survey by Kalish et al., 2021) is provided to survey practitioners. The attributes that are most effective in reducing bias/variances of estimates can be collected and used to reduce potential selection bias for the purpose of timely data collection and minimum response burden. Third, the findings from this study can be used for future development of a high-quality probability survey, including the planning of covariates to collect through paradata

or the survey questionnaire with minimized measurement/reporting error, to be used as a high-quality reference survey by various nonprobability or web-based probability surveys with selection bias.

The proposed variable inclusion strategies have limitations that can be of interest for future research. First, the strategy is developed for single-outcome studies with research questions related to one outcome of interest, e.g., SARS-CoV2 seropositivity study (Kalish et al., 2021). The target sample was collected in a web survey with questions related to COVID-19 infection only. For studies with multiple key outcome variables, it would be of interest to study how the correlation among outcome variables, the overlap for each variable type across outcomes, and their interplays affect population mean estimation by different variable inclusion strategies. Second, in our simulation, we demonstrated the use of a PS matching method (KW) and a PS weighting method numerically. It showed that KW produced approximately unbiased estimates when the analytic propensity model is slightly mis-specified (without the interaction term) while the PS weighting methods require the true propensity model to obtain unbiased estimates. In practice, the underlying selection mechanism of the target sample is often complicated, involving higher orders of nonlinearity and/or nonadditivity. For complicated propensity models, only including main or interactive effects of blocking variables in the logistic model by KW methods may not be sufficient. Nonparametric modeling such as machine learning methods may be promising (Kern et al., 2020). Third, the focus of this paper was on evaluating the bias and variance reduction of Horvitz-Thompson estimators of FP mean by the types of covariates in the propensity model and thus we did not study how the pseudo-weights, when combined with different estimators, affect the FP mean estimation. Alternative analysis methods, such as doubly robust estimators (Chen, Li and Wu, 2018) or augmented estimation equations in the missing data imputation context (Robins, Rotnitzky and Zhao, 1994) can be employed, after identifying the appropriate type of variables to include in the propensity score model. Fourth, selection bias in target samples, compared to more rigorous probability samples, can be induced by low response, different question wording/ordering, topic salience for different types of questions (for example health and health conditions can have large selection bias as shown in Table 2). It would be interesting to study how the proposed variable inclusion strategy can be adapted to reduce the selection bias in target samples with different response rates, question order/wording and salience effects. Lastly, in our data example the backward selection is employed for identifying the type of variables. It would be interesting to employ and compare alternative variable selection methods such as AIC or BIC (Lumley and Scott, 2015) that incorporate complex sample designs for the model selection.

Disclaimer:

The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

Appendix: Real Data Analysis Covariate Types

Covariate types (X_1, confounder; X_2, predictor; X_3, selection indicator) reported for each model covariate used in the real data analysis (Section 4). Covariate interactions are denoted by *. Eleven predictors (age group; sex; race/ethnicity; marital status; education level;

smoking status; diagnosed with high cholesterol; diagnosed with COPD, emphysema, or chronic bronchitis; diagnosed with diabetes; diagnosed with hypertension; and employment status) and their pairwise interactions were included in the initial propensity score models for all adjustments. The covariate types are reported by model set up including the inclusion of RANDS panel weights (panel weights) and exclusion of RANDS panel weights (no weights). Model.n indicates that the outcome predictors were identified using the NHIS; Model.r indicates that the outcome predictors were identified using RANDS.

		Covariate Type			
		Panel weights		No weights	
	Variable	**Model.n**	**Model.r**	**Model.n**	**Model.r**
1	Age group (years)	X_1	X_1	X_1	X_1
2	Sex	X_1	X_1	X_1	X_1
3	Race/Ethnicity	X_1	X_1	X_1	X_1
4	Marital status	X_1	X_1	X_1	X_1
5	Education level	X_1	X_1	X_1	X_1
6	Smoking status	X_1	X_1	X_1	X_1
7	Diagnosed with high cholesterol	X_1	X_1	X_1	X_1
8	Diagnosed with COPD, emphysema, or chronic bronchitis	X_1	X_1	X_1	X_1
9	Diagnosed with diabetes	X_1	X_1	X_1	X_1
10	Diagnosed with hypertension	X_1	X_1	X_1	X_1
11	Employment status	X_1	X_1	X_1	X_1
12	Age group (years) * Sex				
13	Age group (years) * Race/Ethnicity	X_1	X_1	X_1	X_1
14	Age group (years) * Marital status	X_3	X_1	X_3	X_1
15	Age group (years) * Education level	X_3	X_3	X_3	X_3
16	Age group (years) * Smoking status	X_2	X_2	X_2	
17	Age group (years) * Diagnosed with high cholesterol				
18	Age group (years) * Diagnosed with COPD, emphysema, or chronic bronchitis	X_3	X_3	X_3	
19	Age group (years) * Diagnosed with diabetes	X_3	X_1	X_3	
20	Age group (years) * Diagnosed with hypertension			X_3	X_3
21	Age group (years) * Employment status			X_3	X_3
22	Sex * Race/Ethnicity				
23	Sex * Marital status	X_2	X_2	X_2	X_2
24	Sex * Education level	X_1	X_3	X_1	X_3
25	Sex * Smoking status	X_2		X_1	X_3
26	Sex * Diagnosed with high cholesterol	X_2	X_2	X_2	X_2
27	Sex * Diagnosed with COPD, emphysema, or chronic bronchitis		X_2		
28	Sex * Diagnosed with diabetes				X_2
29	Sex * Diagnosed with hypertension	X_3	X_3		

		Covariate Type			
		Panel weights		No weights	
	Variable	Model.n	Model.r	Model.n	Model.r
30	Sex * Employment status			X_3	X_3
31	Race/Ethnicity * Marital status	X_2	X_2	X_1	X_1
32	Race/Ethnicity * Education level	X_3	X_3	X_3	X_3
33	Race/Ethnicity * Smoking status	X_3	X_3	X_3	X_3
34	Race/Ethnicity * Diagnosed with high cholesterol				
35	Race/Ethnicity * Diagnosed with COPD, emphysema, or chronic bronchitis				
36	Race/Ethnicity * Diagnosed with diabetes				
37	Race/Ethnicity * Diagnosed with hypertension	X_2		X_2	
38	Race/Ethnicity * Employment status				
39	Marital status * Education level		X_2		X_2
40	Marital status * Smoking status		X_2		X_2
41	Marital status * Diagnosed with high cholesterol		X_2		X_2
42	Marital status * Diagnosed with COPD, emphysema, or chronic bronchitis		X_2		X_2
43	Marital status * Diagnosed with diabetes	X_2	X_2	X_2	X_2
44	Marital status * Diagnosed with hypertension		X_2		X_2
45	Marital status * Employment status		X_2		X_2
46	Education level * Smoking status				X_2
47	Education level * Diagnosed with high cholesterol	X_3	X_3	X_3	X_3
48	Education level * Diagnosed with COPD, emphysema, or chronic bronchitis				
49	Education level * Diagnosed with diabetes				
50	Education level * Diagnosed with hypertension				
51	Education level * Employment status				
52	Smoking status * Diagnosed with high cholesterol				
53	Smoking status * Diagnosed with COPD, emphysema, or chronic bronchitis	X_2		X_2	
54	Smoking status * Diagnosed with diabetes		X_2		X_2
55	Smoking status * Diagnosed with hypertension				
56	Smoking status * Employment status				
57	Diagnosed with high cholesterol * Diagnosed with COPD, emphysema, or chronic bronchitis	X_2		X_2	
58	Diagnosed with high cholesterol * Diagnosed with diabetes			X_2	
59	Diagnosed with high cholesterol * Diagnosed with hypertension	X_3	X_3	X_3	X_3
60	Diagnosed with high cholesterol * Employment status			X_3	X_3
61	Diagnosed with COPD, emphysema, or chronic bronchitis * Diagnosed with diabetes	X_2			X_2
62	Diagnosed with COPD, emphysema, or chronic bronchitis * Diagnosed with hypertension				

		Covariate Type			
		Panel weights		No weights	
	Variable	Model.n	Model.r	Model.n	Model.r
63	Diagnosed with COPD, emphysema, or chronic bronchitis * Employment status				
64	Diagnosed with diabetes * Diagnosed with hypertension				
65	Diagnosed with diabetes * Employment status				
66	Diagnosed with hypertension * Employment status				

References

Ali MS, Groenwold RHH, Belister S, Pestman WR, Hoes AW, Roes KCB, et al. 2015. "Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review." Journal of Clinical Epidemiology, 68, 122–31. 10.1016/j.jclinepi.2014.08.011. [PubMed: 25579639]

Austin P 2008. "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003." Statistics in Medicine, 27, 2037–2049. 10.1002/sim.3150. [PubMed: 18038446]

Baker R, Brick JM, Bates NA, Battaglia M, Couper MP, Dever JA, et al. 2013. "Summary report on the APPOR task force on non-probability sampling." Journal of Survey Statistics and Methodology, 1, 90–143. 10.1093/jssam/smt008

Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J and Stürmer T 2006. "Variable selection for propensity score models." Am. J. of Epi, 163, 1149–1156. 10.1093/aje/kwj149

Brick JM, Tourangeau R 2017. "Responsive Survey Designs for Reducing." Journal of Official Statistics, 33:735–752. 10.1515/jos-2017-0034

Chen Y, Li P and Wu C 2019. "Doubly Robust Inference with Nonprobability Survey Samples." Journal of the American Statistical Association: 1–11. 10.1080/01621459.2019.1677241. [PubMed: 34012183]

Chen JKT, Valliant R and Elliott MR 2018. "Model-assisted calibration of non-probability sample survey data using adaptive LASSO." Survey Methodology, 44, 117–144. http://link.gale.com/apps/doc/A663832408/AONE?u=anon~9db3f809&sid=googleScholar&xid=411f8f5e . Accessed 25 Mar. 2022.

Craig BM, Hays RD, Pickard AS, Cella D, Revicki DA, Reeve BB. Comparison of US panel vendors for online surveys. Journal of the Medical Internet Research. 2013;15 (11):e260. 10.2196/jmir.2903. [PubMed: 24292159]

Grose E, Wilson S, Barkun J, Bertens K, Martel G, Balaa F, et al. 2020. "Use of Propensity Score Methodology in Contemporary High-Impact Surgical Literature." Journal of the American College of Surgeons, 230, 101–112.e2. 10.1016/j.jamcollsurg.2019.10.003. [PubMed: 31672675]

Groves RM 1989. Survey Errors and Survey Costs. New York: Wiley.

Groves RM 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." Public Opinion Quarterly, 70, 646–675. 10.1093/poq/nfl033

Groves RM, Peytcheva E 2008. "The impact of nonresponse rates on nonresponse bias." Public Opinion Quarterly 72:167–189. 10.1093/poq/nfn011

Hernán MA 2004. "A definition of causal effect for epidemiological research." Journal of Epidemiology & Community Health, 58, 265:271. 10.1136/jech.2002.006361 [PubMed: 15026432]

Iannacchione VG, Milne JG and Folsom RE 1991. "Response probability weight adjustments using logistic regression." Proceedings of the section on survey research methods, American Statistical Association, 637–42. http://www.asasrms.org/Proceedings/papers/1991_109.pdf, accessed on 03/25/2022

Irimata KE, He Y, Cai B, Shin H-C, Parsons VL and Parker JD 2020. "Comparison of Quarterly and Yearly Calibration Data for Propensity Score Adjusted Web Survey Estimates." Survey Methods: Insights from the Field, Special issue "Advancements in Online and Mobile Survey Methods." 10.13094/SMIF-2020-00018

Kalish H, Klumpp-Thomas C, Hunsberger S, Baus HA, Fay MP, Siripong N, et al. 2021. "Mapping a Pandemic: SARS-CoV-2 Seropositivity in the United States." https://www.medrxiv.org/content/10.1101/2021.01.27.21250570v1

Kern C, Li Y and Wang L 2020. "Boosted Kernel Weighting – Using Statistical Learning to Improve Inference from Nonprobability Samples." Journal of Survey Statistics and Methodology, 10.1093/jssam/smaa028

Kish L 1992. "Weighting for unequal Pi." Journal of Official Statistics, 8, 183–200. https://www.scb.se/contentassets/f6bcee6f397c4fd68db6452fc9643e68/weighting-for-unequal-empemsubemiemsub.pdf, accessed on 03/25/2022

Korn E and Graubard BI 1999. Analysis of Health Surveys. John Wiley & Sons.

Lessler JT and Kalsbeek WD. 1992. Nonsampling Errors in Surveys. New York: John Wiley.

Leyrat C, Caille A, Donner A, and Giraudeau B 2013. "Propensity scores used for analysis of cluster randomized trials with selection bias: a simulation study." Statistics in Medicine, 32, 3357–3372. 10.1002/sim.5795. [PubMed: 23553813]

Li Y, Graubard B, and DiGaetano R 2011. "Weighting methods for population-based case-control study." Journal of Royal Statistical Society C, 60, 165–185. 10.1111/j.1467-9876.2010.00731.x.

Little RJ and Vartivarian S 2003. "On weighting the rates in non-response weights." Statistics in Medicine, 22, 1589–1599. 10.1002/sim.1513. [PubMed: 12704617]

Little RJ and Vartivarian S 2005. "Does weighting for nonresponse increase the variance of survey means?" Survey Methodology, 31, 161–168. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9046-eng.pdf?st=VelcJijp. Accessed on 3/25/2022.

Lumley T and Scott A 2015. "AIC and BIC for modeling with complex survey data." Journal of Survey Statistics and Methodology, 3, 1–18. 10.1093/jssam/smu021.

Lumley T 2020. Survey: analysis of complex survey samples. R package version 4.4.0.

NIH-AARP (National Institutes of Health and AARP Diet and Health Study) Data Dictionary. August 2006. Available: http://dietandhealth.cancer.gov/docs/DataDictionary_Aug2006.pdf.

Mercer Andrew, Kennedy Courtney, Klein Hannah. 2018. "For weighting online opt-in samples, what matters most." Pew Research Center. https://www.pewresearch.org/methods/wp-content/uploads/sites/10/2018/01/Weighting-Online-Opt-In-Samples.pdf. Accessed on 3/25/22.

National Center for Health Statistics. 2020. RANDS 3 Technical Documentation. Hyattsville, Maryland.

Parker J, Miller K, He Y, Scanlon P, Cai B, Shin H-C, et al. 2020. "Overview and Initial Results of the National Center for Health Statistics' Research and Development Survey." Statistical Journal of the International Association for Official Statistics, 36, 1199–1211. 10.3233/SJI-200678. [PubMed: 35923778]

Parker JD, Talih M, Malec DJ, Beresovsky V, Carroll M, Gonzalez JF, et al. 2017. "National Center for Health Statistics Data Presentation Standards for Proportions. National Center for Health Statistics." Vital Health Stat, 2(175).

Pearl J 2009. Causality: Models, Reasoning, and Inference. Cambridge, England: Cambridge University Press, 2nd edn.

Pinsky PF, Miller A, Kramer BS, Church T, Reding D, Prorok P, Gelmann E, Schoen RE, Buys S, Hayes RB and Berg CD 2007. "Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial." Am. J. Epidem, 165, 874–881. 10.1093/aje/kwk075.

Robins J, Rotnitzky A, and Zhao L 1994. "Estimation of regression coefficients when some regressors are not always observed." Journal of the American Statistical Association,89, 846–86. 10.1080/01621459.1994.10476818.

Rosenbaum PR and Rubin DB 1983. "The central role of the propensity score in observational studies for causal effects." Biometrika, 70, 41–55. 10.1093/biomet/70.1.41.

Silverman BW 1986. Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.

Valliant R 2020. “Comparing alternatives for estimation from nonprobability samples.” Journal of Survey Statistics and Methodology, 8, 231–263. 10.1093/jssam/smz003.

Wang L, Graubard BI, Hormuzd AK and Li Y 2020a. “Improving External Validity of Epidemiologic Cohort Analyses: a Kernel Weighting Approach.” Journal of the Royal Statistical Society Series A, 183, 1293–1311. 10.1111/rssa.12564.

Wang L, Graubard BI, Hormuzd AK and Li Y 2020b. “Efficient and Robust Propensity-Score-Based Methods for Population Inference using Epidemiologic Cohorts.” 10.1111/insr.12470.

Wang L, Valliant R and Li Y 2021. “Adjusted Logistic Propensity Weighting Methods for Population Inference using Nonprobability Volunteer-Based Epidemiologic Cohorts.” Statistics in Medicine, 10.1002/sim.9122.

Weitzen S, Lapane KL, Toledano AY, Hume AL and Mor V 2004. “Principles for modeling propensity scores in medical research: a systematic literature review.” Phamacoepidemiology and Drug Safety, 13, 841–53. 10.1002/PDS.969.

Yang S, Kim JK, and Song R 2020. “Doubly robust inference when combining probability and non-probability samples with high dimensional data.” Journal of the Royal Statistical Society Series B, 82, 445–465. 10.1111/rssb.12354

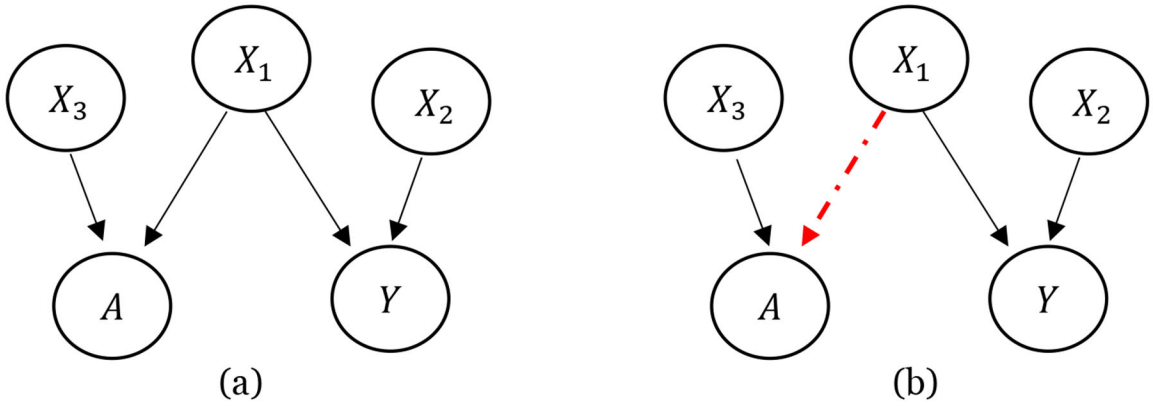

Figure 1.
DAG for three types of covariates with the selection indicator (A) and the outcome (Y)

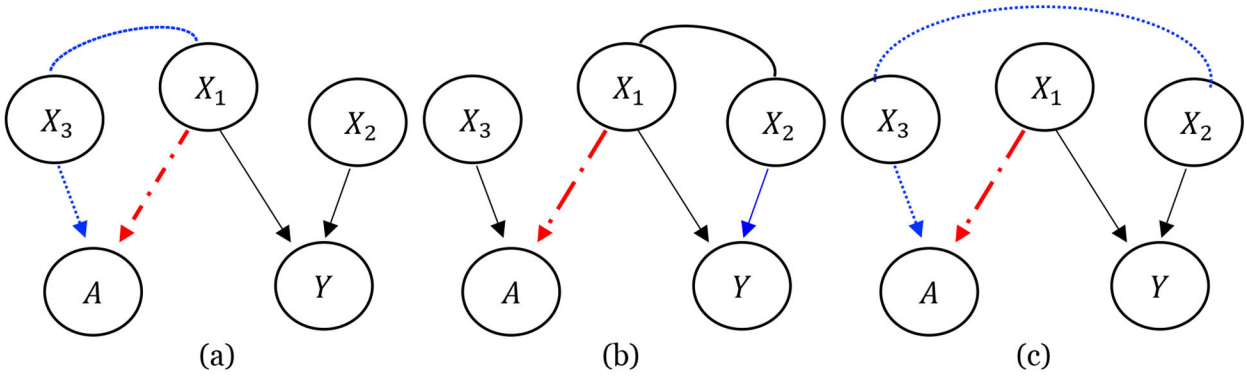

Figure 2.
DAG for $Y \perp A$ when correlation exists between X_1-X_3(a), X_1-X_2(b) and X_2-X_3(c). Blocking dotted path(s) to have $Y \perp A$.

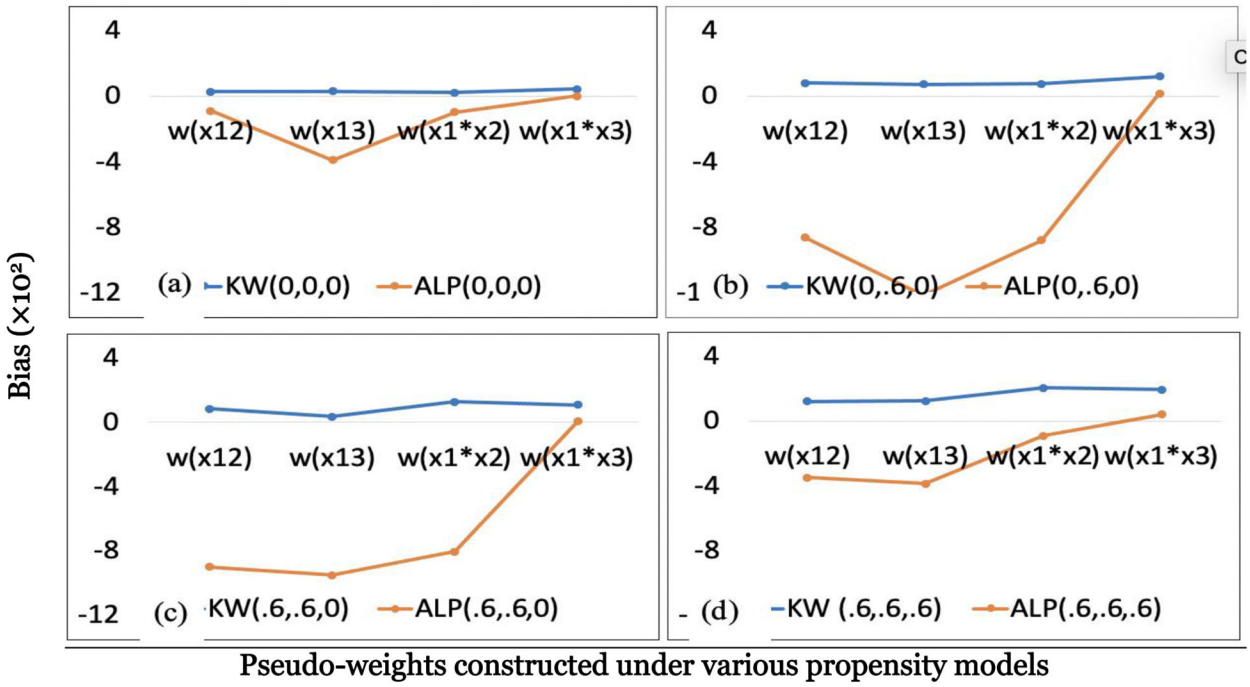

Figure 3.

Bias of Kernel Weighting (KW) vs. Adjusted Logistic Propensity (ALP) Estimated under Various Propensity Models with w(x_{12}), w(x_{13}), w(x_1 * x_2), and w(x_1 * x_3) including, respectively, main effects of X_1 and X_2, main effects of X_1 and X_3, main and interactive effects of X_1 and X_2, and main and interactive effects of X_1 and X_3, with ($\rho_{x_1x_2}$, $\rho_{x_1x_3}$, $\rho_{x_2x_3}$) = (0,0,0) (a), (0,.6,0) (b), (.6,.6,0) (c), and (.6,.6,.6) (d), to cover 0, 1, 2, and 3 pair(s) of covariate correlations in the FP, and interactive effects $\alpha_{12} = \beta_{13} = 0.5$. Propensity model with w(x_1 * x_3) is the true model.

Table 1.

Results from population mean estimation[1] under various propensity score models[2] with covariate correlations $(\rho_{x_1x_2}, \rho_{x_1x_3}, \rho_{x_2x_3}) = (0, 0, 0)$ and interaction effects $\alpha_{12} = \beta_{13} = 0$.

	Sample[3]	w(x_1)	w(x_2)	w(x_3)	w(x_{12})	w(x_{13})	w(x_{23})
Bias ($\times 10^2$)	4.61	0.26	4.50	4.83	0.26	0.41	4.77
EmpVar ($\times 10^4$)	2.20	2.68	2.62	2.96	2.92	3.43	3.32
MSE ($\times 10^4$)	23.48	2.75	22.85	26.31	2.99	3.60	26.04

[1] Kernel weighting estimator (Wang et al., 2020b) is applied for population mean estimation.

[2] w(x_1), w(x_2), w(x_3), w(x_{12}), w(x_{13}), and w(x_{23}) denote pseudo-weighted means with pseudo-weights constructed under the propensity model with main effect(s) of x_1, x_2, x_3, x_1 and x_2, x_1 and x_3, and x_2 and x_3, respectively.

[3] sample denotes the unweighted mean

Table 2.

Results from population mean estimation[1] under various propensity score models[2] by covariate correlations with interaction effects $\alpha_{12} = \beta_{13} = 0$.

	Sample[3]	w(x_1)	w(x_2)	w(x_3)	w(x_{12})	w(x_{13})	w(x_{23})
		($\rho_{x_1x_2}$, $\rho_{x_1x_3}$, $\rho_{x_2x_3}$) = (.6, 0, 0)					
Bias (×10^2)	7.35	0.37	2.98	7.60	0.37	0.59	3.25
EmpVar (×10^4)	2.15	2.59	2.64	2.77	2.66	2.88	2.84
MSE (×10^4)	56.14	2.72	11.52	60.57	2.79	3.23	13.42
		($\rho_{x_1x_2}$, $\rho_{x_1x_3}$, $\rho_{x_2x_3}$) = (0, .6, 0)					
Bias (×10^2)	7.27	0.32	7.16	3.21	0.30	0.41	3.12
EmpVar (×10^4)	2.17	3.60	2.39	3.68	3.53	4.05	3.66
MSE (×10^4)	54.98	3.70	53.67	13.97	3.62	4.22	13.39
		($\rho_{x_1x_2}$, $\rho_{x_1x_3}$, $\rho_{x_2x_3}$) = (0, 0, .6)					
Bias (×10^2)	7.55	2.98	4.65	4.87	0.26	0.37	4.83
EmpVar (×10^4)	2.01	2.52	2.38	2.57	2.66	2.75	2.66
MSE (×10^4)	59.00	11.38	24.00	26.30	2.73	2.89	26.03
		($\rho_{x_1x_2}$, $\rho_{x_1x_3}$, $\rho_{x_2x_3}$) = (.6, .6, 0)					
Bias (×10^2)	9.81	−1.04	5.36	6.09	0.54	0.54	1.70
EmpVar (×10^4)	2.16	3.45	2.94	3.79	3.94	3.98	4.19
MSE (×10^4)	98.39	4.52	31.61	40.93	4.23	4.27	7.09
		($\rho_{x_1x_2}$, $\rho_{x_1x_3}$, $\rho_{x_2x_3}$) = (.6, 0, .6)					
Bias (×10^2)	10.27	3.11	1.67	7.60	0.50	0.60	2.36
EmpVar (×10^4)	2.33	2.84	2.96	2.94	2.93	3.11	3.00
MSE (×10^4)	107.86	12.51	5.76	60.72	3.18	3.46	8.58
		($\rho_{x_1x_2}$, $\rho_{x_1x_3}$, $\rho_{x_2x_3}$) = (0, .6, .6)					
Bias (×10^2)	10.09	3.11	7.24	1.56	0.33	0.44	2.34
EmpVar (×10^4)	1.98	3.65	2.63	3.28	3.45	3.61	3.48
MSE (×10^4)	103.83	13.33	54.98	5.71	3.55	3.80	8.97
		($\rho_{x_1x_2}$, $\rho_{x_1x_3}$, $\rho_{x_2x_3}$) = (.6, .6, .6)					
Bias (×10^2)	13.28	1.73	4.52	4.81	0.77	0.89	3.30
EmpVar (×10^4)	1.98	3.93	3.39	3.88	3.91	4.29	3.97
MSE (×10^4)	178.29	6.91	23.83	26.99	4.50	5.07	14.88

[1] Kernel weighting estimator (Wang et al., 2020b) is applied for population mean estimation.

[2] w(x_1), w(x_2), w(x_3), w(x_{12}), w(x_{13}), and w(x_{23}) denote pseudo-weighted means with pseudo-weights constructed under the propensity model with main effect(s) of x_1, x_2, x_3, x_1 and x_2, x_1 and x_3, and x_2 and x_3, respectively.

[3] sample denotes the unweighted mean.

Table 3.

Distribution of selected covariates and asthma in the Research and Development Survey (RANDS) 3 and the 2019 National Health Interview Survey (NHIS)

Variable	Subgroup	RANDS (n=2,646)			NHIS (n=31,997)	
		N	%	Wt %	n	Wt %
Outcome						
Ever diagnosed with asthma	Yes	431	16.4	16.9	4,229	13.5
	No	2,197	83.6	83.1	27,718	86.5
Covariates						
Age group (years)	18-34	721	27.2	29.9	7,058	29.7
	35-49	652	24.6	24.1	7,250	24.3
	50-64	687	26.0	25.1	8,313	24.9
	65+	586	22.1	20.9	9,376	21.1
Sex	Male	1,318	49.8	48.3	14,733	48.3
	Female	1,328	50.2	51.7	17,261	51.7
Race/Ethnicity	Non-Hispanic white	1,729	65.3	63.1	21,915	63.2
	Non-Hispanic black	273	10.3	11.9	3,483	11.8
	Non-Hispanic other	227	8.6	8.5	2,447	8.5
	Hispanic	417	15.8	16.5	4,152	16.5
Marital status	Married	1,282	48.5	47.7	14,759	52.4
	Widowed	134	5.1	4.5	3,115	6.0
	Divorced	350	13.2	12.4	4,317	9.0
	Separated	50	1.9	1.8	456	1.2
	Never married	618	23.4	24.3	6,368	22.5
	Living with partner	212	8.0	9.3	2,136	8.9
Education level	High school diploma or less	577	21.8	38.8	11,155	39.9
	Some college	1,222	46.2	27.7	9,386	31.1
	Bachelor's degree or more	847	32.0	33.5	11,277	29.0
Smoking status[1]	Current	409	15.5	17.2	4,296	14.0
	Former	811	30.8	28.9	7,973	22.5
	Never	1,411	53.6	53.9	18,931	63.5
Diagnosed with high cholesterol	Yes	976	37.1	36.4	9,179	24.9
	No	1,657	62.9	63.6	22,697	75.1
Diagnosed with COPD, emphysema, or chronic bronchitis	Yes	213	8.1	8.4	1,787	4.6
	No	2,420	91.9	91.6	30,158	95.4
Diagnosed with diabetes[2]	Yes	279	10.6	10.5	3,355	9.3
	No	2,352	89.4	89.5	28,594	90.7
Diagnosed with hypertension	Yes	989	37.5	37.0	11,480	31.7
	No	1,648	62.5	63.0	20,458	68.3
Employment status	Paid employee	1,630	61.6	58.6	18,810	64.6
	Looking for work	166	6.3	7.2	485	2.0

Variable	Subgroup	RANDS (n=2,646)			NHIS (n=31,997)	
		N	%	Wt %	n	Wt %
	Not looking for work	850	32.1	34.2	11,919	33.4

Notes: n=unweighted sample size, %=unweighted percent, Wt % = weighted percent

[1]Smoking status: Current smoker is defined as someone who has smoked at least 100 cigarettes in their lifetime and now smokes every day or some days. Former smoker is defined as someone who has smoked at least 100 cigarettes in their lifetime and now does not smoke. Never smokers are defined as persons who have smoked less than 100 cigarettes in their lifetime.

[2]Diagnosed diabetes excludes pre-diabetes and gestational diabetes.

Table 4.

Analysis Results for estimation of the prevalence of diagnosed asthma for adults from RANDS 3 under various propensity models and RANDS 3 weights

Propensity model[1]	CV[2](KW)	relBias[3] (%)	se[4] (× 10^2)	MSE[5] (× 10^4)
Original panel-weighted	0.91	25.31	0.98	12.56
Unweighted	0	21.89	0.72	9.19
	panel weights			
Model.all	1.13	17.55	1.21	7.04
Model.x13	1.10	13.35	1.04	4.31
Model.x12.n	1.07	11.41	0.93	3.23
Model.x1.n	1.07	12.38	0.95	3.67
Model.x12.r	1.08	12.85	0.97	3.94
Model.x1.r	1.08	12.85	0.97	3.93
	no weights			
Model.all	0.83	14.02	1.07	4.70
Model.x13	0.80	13.51	0.97	4.24
Model.x12.n	0.70	11.38	0.81	3.01
Model.x1.n	0.69	13.67	0.82	4.06
Model.x12.r	0.73	11.37	0.84	3.04
Model.x1.r	0.71	13.44	0.84	3.98

[1] Original panel-weighted denotes the RANDS 3 estimate using the original panel weights without PS adjustment; unweighted denotes the RANDS 3 estimate using weight = 1 without PS adjustment; model.all: the full propensity model with all main and pairwise interaction terms; Model.x13: the propensity model including selected terms of the confounders and selection predictors; Model.x12.n: propensity model including terms of the confounders and outcome predictors selected using the National Health Interview Survey (NHIS); Model.x12.r: propensity model including terms of the confounders and outcome predictors selected using the Research and Development Survey (RANDS). Panel weights indicates that the RANDS 3 original panel weights were used as the base weight for the PS adjustment. No weights indicates that the RANDS 3 original panel weights were not included in the PS adjustment.

[2] $CV = \frac{sd(KW)}{mean(KW)}$, for standard deviation sd

[3] $\text{relBias}\,(\%) = \frac{\hat{\mu}_{RANDS} - \hat{\mu}_{NHIS}}{\hat{\mu}_{NHIS}} \times 100\,\%$

[4] se=standard error of estimated mean

[5] $\text{MSE} = (\hat{\mu}_{RANDS} - \hat{\mu}_{NHIS})^2 + se^2(\hat{\mu}_{RANDS})$