Efficient and robust propensity-score-based methods for population inference using epidemiologic cohorts

Lingxiao Wang, Barry I. Graubard, Hormuzd A. Katki and Yan Li

1 The Joint Program in Survey Methodology, University of Maryland, College Park, Maryland, USA
2 Division of Cancer Epidemiology & Genetics, Biostatistics Branch, National Cancer Institute, Bethesda, Maryland, USA

Corresponding author: yli6@umd.edu

Summary

Most epidemiologic cohorts are composed of volunteers who do not represent the general population. To improve population inference from cohorts, propensity-score (PS)-based matching methods, such as PS-based kernel weighting (KW) method, utilise probability survey samples as external references to develop PSs for membership in the cohort versus survey. We identify a strong exchangeability assumption (SEA) that underlies existing PS-based matching methods whose failure invalidates inferences, even if the propensity model is correctly specified. Herein, we develop a framework of propensity estimation and relax the SEA to a weak exchangeability assumption (WEA) for matching methods. To recover efficiency, we propose a scaled KW (KW.S) matching method by scaling survey weights in propensity estimation. We prove consistency of KW.S estimators of means/prevalences under WEA and provide consistent finite population variance estimators.

In simulations, the KW.S estimators had smallest mean squared error (MSE). Our data example showed the KW estimates requiring the SEA had large bias, whereas the proposed KW.S estimates had the smallest MSE.

Key words: Nonprobability cohorts; finite population inference; propensity score weighting; weight scaling; Taylor series linearisation variance.

1 Introduction

Assembling epidemiologic cohorts using probability sampling substantially increases costs, especially if biospecimens are required (Duncan, 2008; LaVange et al., 2001). Thus, many cohort studies assemble samples using volunteer-based recruitment, but they generally cannot represent the target population. For example, the UK Biobank has half the mortality rate of the UK population (Collins, 2012). However, the recent survey sampling literature is burgeoning with propensity-score (PS)-based methods that use probability-based survey samples as external references to improve population inferences for non-probability samples (Chen et al., 2019; Elliott & Valliant, 2017; Lee & Valliant, 2009; Valliant & Dever, 2011). The PS-based methods can also...
be applied in cohort analysis for inference of outcome variables available in cohorts but not in surveys, such as cancer incidence and novel molecular or genetic risk factors. Intuitively, these methods use covariates measured in both cohort and survey sample to construct a PS model to estimate the participation propensity in the cohort (Chen et al., 2019; Elliott & Valliant, 2017; Valliant & Dever, 2011) or the propensity of participating in the cohort versus in the survey sample (Lee & Valliant, 2009; Wang et al., 2020). Then, the estimated PS are used to create pseudo-weights for the cohort participants so that the under-represented are upweighted, and the over-represented are downweighted, in order to improve the cohort’s population representativeness.

One popular type of PS-based method is PS-based weighting methods that use PS to estimate participation rates of the cohort units, for example, the inverse PS weighting proposed in Valliant and Dever (2011). A second approach is PS-based matching methods that use PS to measure similarity of the cohort and survey sample units, such as PS adjustment by subclassification (PSAS) (Lee & Valliant, 2009), Rivers’ matching method (Rivers, 2007) and kernel weighting (KW) (Wang et al., 2020). Compared with PS-based weighting methods, PS-based matching methods can be more robust to propensity model mis-specification because the PS is used to measure the similarity of the cohort units with the survey sample units, in terms of the distribution of the covariates that are used in the propensity model (Wang et al., 2020). In PS-based matching methods, the propensity model is fitted to the combined (cohort vs. unweighted survey) sample to avoid inefficiency in propensity estimation due to using the survey weights (Brick, 2015; Lee & Valliant, 2009; Rivers, 2007; Wang et al., 2020). The KW method, which uses kernel smoothing, ensures consistent pseudo-weighted estimation under regularity conditions (Wang et al., 2020). Empirical results show that the KW estimation of finite population means generally have smaller mean squared error (MSE) compared with the inverse of PS weighting (IPSW) or the PSAS estimation.

We demonstrate that all PS-based matching methods that fit the propensity model to the combined (cohort vs. unweighted survey) sample require a hidden, but critical, strong exchangeability assumption (SEA) for estimating finite population means. The SEA states that the expectation of the outcome variable given the estimated PS is the same in the cohort, the survey and the finite population. We prove that, without the SEA, current PS-based matching estimates are not consistent, even when the propensity model is correctly specified. We establish a unified framework of propensity estimation for both PS-based matching and weighting methods. For the matching methods, the matching scores are defined by the (functions of) PSs estimated from the propensity models fitted to the combined (cohort vs. weighted survey) sample. We relax the SEA to a weak exchangeability assumption (WEA) that is more realistic for data analysis.

However, fitting the propensity model to the weighted sample (when compared with the unweighted sample) increases variability in propensity estimation that can greatly increase the variance of pseudo-weighted estimation. To recover efficiency, we propose scaling survey weights by their mean in propensity estimation, which is motivated by the method of scaling weights in population-based case–control studies where the sample weights are highly variable among the cases and the controls (Landsman & Graubard, 2013; Li et al., 2010; Scott & Wild, 1986, 2001). We demonstrate that this simple scaling greatly reduces variance while retaining the consistency of the PS-based pseudo-weighted estimators.

Finally, we derive Taylor linearisation (TL) variances (i.e. analytical variances) for pseudo-weighted estimators of finite population means that take propensity estimation into account. We apply our methods to an example where we use the naïve (not sample-weighted) US National Health and Nutrition Examination (NHANES) III as the ‘cohort’ and the sample-weighted 1997 US National Health Interview Survey (NHIS) as the reference survey. The use of naïve NHANES allows for clear assessment of the bias reduction obtained by the proposed
PS-based methods without being obscured by differences in the population coverage and measurement errors between the cohort and the reference survey interfering to any large extent.

2 Basic Setting

Let the target finite population (FP) consist of \( N \) individuals indexed by \( i \in \{1, \ldots, N\} \), where each individual \( i \) has a value for the outcome variable of interest \( y_i \) and for the vector of covariates \( x_i \). We focus on estimating the FP mean of \( y \), that is, \( \mu = N^{-1} \sum_{i \in FP} y_i \). Let \( s_c \subset FP \) denote a cohort with \( n_c \) individuals. Following notation in Chen et al. (2019), we define a random indicator variable \( \delta^{(c)}(i) = 1 \) if \( i \in s_c \); 0 otherwise) that specifies which individuals in FP participate in \( s_c \). Note that FP and \( s_c \) are also used to denote sets of indices for the target finite population and the cohort, respectively. The underlying cohort participation rate for each \( i \in s_c \) is defined by

\[
\pi^{(c)}_i \equiv P(i \in s_c | FP) = E_c(\delta^{(c)}_i | FP),
\]

where the expectation \( E_c \) is with respect to the unknown random cohort sample participation process from FP. The corresponding cohort implicit sample weight is \( w_i = 1/\pi^{(c)}_i \) for \( i \in s_c \). We require the following standard assumptions for cohort participation (Chen et al., 2019; Elliott & Valliant, 2017):

**A1** The cohort participation indicator \( \delta^{(c)} \) is independent of the outcome variable \( y \) given the covariates \( x \), that is, \( Pr(\delta^{(c)} = 1 | y, x) = Pr(\delta^{(c)} = 1 | x) \).

**A2** All finite population units have a positive participation rate, that is, \( \pi^{(c)}_i > 0 \) for \( i \in FP \).

In addition, a reference survey sample \( s_s \) with \( n_s \) individuals is randomly selected from the FP. The sample inclusion indicator, inclusion probability and the corresponding sample weights are defined by \( \delta^{(s)}_i (= 1 \) if \( i \in s_s \); 0 otherwise), \( \pi^{(s)}_i = E_s(\delta^{(s)}_i | FP) \) and \( d_i = 1/\pi^{(s)}_i \), respectively, where \( E_s \) is the expectation with respect to the survey sample selection and \( s_s \) also denotes the subset of indices for individuals in the survey sample from the FP. In practice, we assume the inclusion probabilities and sample weights are adjusted for non-response and calibrated for FP quantities. Similar conditions of **A1** and **A2** are also assumed for survey sampling, that is, \( \pi^{(s)}_i > 0 \) for \( i \in FP \) and \( Pr(\delta^{(s)} = 1 | y, x) = Pr(\delta^{(s)} = 1 | x) \).

3 Existing PS-Based Matching Methods under SEA

The PS-based matching methods use a matching score (a function of PS) to measure the similarity of the cohort and survey units in terms of the covariate distributions. Hence, they do not require that the matching scores accurately estimate the participation rates \( \pi^{(c)}_i \) for the cohort units. Many existing PS-based matching methods use the propensity of participating in the cohort \( (s_c) \) versus being selected in the survey sample \( (s_s) \) as the matching score to avoid low efficiency of pseudo-weighted estimates due to high variability of the estimated pseudo-weights. However, we found this commonly used matching score requires an SEA for unbiased estimated of \( \mu \), which can be invalid even the propensity is correctly estimated. In this section, we identify this research gap with a simple illustration example.
The PS adjustments by the subclassification (PSAS) method (Lee & Valliant, 2009), the Rivers matching method (Rivers, 2007) and the KW method (Wang et al., 2020) use the matching score defined by $\tilde{p}_i = P\{i \in s_c| i \in s_c \cup^* s_s, FP\}$, where the union $s_c \cup^* s_s$ allows for duplication of individuals in both $s_c$ and $s_s$ and the duplicate individuals are included twice such that the union has $n_c + n_s$ units. In practice, the set of duplicates is usually small, and it is usually not possible to identify the duplicate individuals (Lee & Valliant, 2009). Also, we assume that the relationship between $\tilde{p}_i$ and $x$ follows a logistic regression model:

$$\log\left(\frac{\tilde{p}_i}{1 - \tilde{p}_i}\right) = \beta^T x_i, \ i \in \{s_c \cup^* s_s\}, \ (3.1)$$

where $\hat{\beta}$ is a vector of unknown regression coefficients and can be estimated by fitting model (3.1) to the combined ($s_c$ vs. unweighted $s_s$) sample. Please refer Wang et al. (2020) for details.

The PSAS method sorts the combined cohort and survey sample, $s_c \cup^* s_s$ by the PS $\{\tilde{p}_i, i \in s_c \cup^* s_s\}$ and partitions $s_c \cup^* s_s$ into subclasses. Then, the survey sample weights are evenly distributed among the cohort units within each subclass, which assumes that all cohort units within a subclass represent the same number of population units.

In contrast to the PSAS method, the KW methodfractionally distributes survey sample weights to cohort units with the KW pseudo-weight for cohort unit $i$ defined by

$$w_i^{KW} = \frac{\sum e_{j_1} \left\{ K\left\{ \frac{\tilde{p}_i^{(s)} - \tilde{p}_j^{(s)}}{h}\right\}\cdot d_j \right\} }{\sum e_{j_1} K\left\{ \frac{\tilde{p}_i^{(s)} - \tilde{p}_j^{(s)}}{h}\right\} }, \ i \in s_c, \ (3.2)$$

where $K(\cdot)$ is a zero-centred kernel function (Epanechnikov, 1969) (e.g. standard normal, or triangular density) and $h$ is the bandwidth decided by the distribution of $\{\tilde{p}_i^{(s)}, i \in s_c\}$ and the selected kernel function. As proved in Wang et al. (2020), using $\tilde{p}$ as the matching score requires the SEA

$$E(y|\tilde{p}, s_c) = E(y|\tilde{p}, s_s) = E(y|\tilde{p}, FP) \ (3.3)$$

for consistency of the KW estimators of finite population mean, where $E(\cdot)$ is the expectation with respect to the distribution of $y$ in FP. The SEA requires that two equalities hold among the sets $s_c$, $s_s$, and $FP$. This assumption is strong and can be violated even when $\tilde{p}$ is estimated under the correct propensity model fitted to the combined ($s_c$ vs. unweighted $s_s$) sample. This is because only the first equality of the SEA automatically holds under Model (3.1) (Rosenbaum & Rubin, 1983), but the second equality may not necessarily hold, resulting in biased pseudo-weighted estimation. We use the following simple examples to illustrate the case of unbiased mean estimation when the SEA is valid and the case of biased mean estimation when the SEA is violated even if the propensity model is correctly specified.

### 3.1 Simple Examples

Suppose the covariates $x$ include two binary variables age (=0 for young; =1 for old), and sex (=0 for male; =1 for female). The distribution of $y$ depends on age and sex with the expectation $\mu_{jk} = E(y|\text{age} = j, \text{sex} = k)$ for $j,k = 0,1$. We assume $\mu_{jk}$ differs by the four categories of age by sex. We assume $s_c$ and $s_s$ are two independent stratified simple random samples with the four strata defined by the age and sex groups. The implicit cohort sample weights are the same within the four strata, but different across the strata.
1. SEA Valid Case

In the SEA valid case, the values of $\tilde{p}$ are different in the four strata defined by age and sex. The SEA is satisfied because $E(y \mid \tilde{p}) = E(y \mid \text{age} = j, \text{sex} = k) = \mu_{jk}$ in $s_{sc}$, $s_{ss}$, and $FP$. The PS-based matching methods evenly distribute survey weights to the cohort units within each of the four matching groups (strata) defined by $\tilde{p}$ to construct pseudo-weights. As a result, the pseudo-weighted estimator of the finite population mean is unbiased.

2. SEA Invalid Case

SEA can be invalid if the value of $\tilde{p}$ cannot differentiate the four age-by-sex strata. This can happen even when the true propensity model (including both age and sex) is fitted to the combined ($s_c$ vs. unweighted $s_s$) sample. For example, if $s_c$ and $s_s$ have the same distribution of sex, the effect of sex is expected to be 0 in the fitted propensity model. As a result, the value of $\tilde{p}$ is the same within age group, regardless of sex. The second equality of SEA is violated, that is, $E(y \mid \tilde{p}, s_c) \neq E(y \mid \tilde{p}, FP)$, because $E(y \mid \text{age}, s_c) \neq E(y \mid \text{age}, FP)$. Applying a PS-based matching method, the survey sample weights would be distributed to the cohort units only according to age, resulting in common pseudo-weights within age categories, and thus yield biased pseudo-weighted estimator of the finite population mean.

As shown by this simple example, when cohort units within subgroups have different participation rates, but the same estimated PSs, the PS-based matching methods cannot match the distribution of $y$ in the pseudo-weighted $s_c$ to that in the $FP$. This is because the second equality in SEA is invalid even if the correct propensity model is fitted to the unweighted sample.

4 Propensity Estimation and Weak Exchangeability for Matching Methods

In this section, we propose a general method of propensity estimation that can be used to relax SEA for the matching methods. Suppose we observe the covariates $x_i$ for all $i \in FP$, but we do not observe the cohort participation indicator $\delta_i^{(c)}$ for all $i \in FP$. We construct a pseudo-population of $s_c^{*} \cup FP$, where $s_c^{*}$ is a copy of $s_c$ that has the same joint distribution of covariates $x$ and outcome $y$ as the original $s_c$. In the union of $s_c^{*} \cup FP$, $s_c^{*}$ and $s_c$ are treated as two different sets. Therefore, the number of units in $s_c^{*} \cup FP$ is $n_c + N$. 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 $p_i = P(R_i = 1) = P(i \in s_c^{*} \mid s_c^{*} \cup FP)$. Assume we have a logistic regression model:

$$\log \left\{ \frac{p_i}{1 - p_i} \right\} = \beta^T x_i, \text{ for } i \in s_c^{*} \cup FP. \tag{4.1}$$

The parameter vector $\beta$ differs from $\tilde{\beta}$ in Model (3.1) because the two models define the propensity differently, that is, the probability that individual $i$ is included in $s_c$ versus $FP$ under Model (4.1) as compared with the probability that individual $i$ is included in $s_c$ versus $s_s$ under Model (3.1). The corresponding ‘likelihood’ function is given by

$$L(\beta) = \prod_{i \in s_c^{*} \cup FP} p_i^{R_i} (1 - p_i)^{1 - R_i}. \tag{4.2}$$

In practice, covariates $x_i$ can only be observed in $s_c$ and $s_s$, but not $FP$. To obtain a consistent estimator of $\beta$, we fit Model (4.1) to the combined ($s_c$ vs. weighted $s_s$) where we use the sample weights $d_i$ for $i \in s_s$, in the estimation. The estimator $\hat{\beta}$ is obtained by solving the following weighted estimating equations for $\beta$:
The estimated PS is denoted by $\hat{\beta}_i = \expit(\hat{\beta}^T x_i)$.

4.1 A WEA for PS-Based Matching Methods

We relax the SEA for the PS-based matching methods under the propensity estimation framework described above by using $p = P(R = 1|x)$ (Rosenbaum & Rubin, 1983) or $q = \logit(p)$ (Rubin & Thomas, 1992) as the matching score to measure the similarity among the cohort and survey units.

In general, the cohort $s_c$ is not representative of the finite population $FP$ because $s_c$ is not a random sample from $FP$. Therefore, there are no sample weights for the cohort to equalise the distributions of covariates $x$ in $s_c$ and $FP$ if they differ. Thus, the pseudo-weights, created using the $x$, serve as the sample weights for $s_c$ to weight the $x$ distribution in $s_c$ up to that in $FP$.

The matching methods classify the cohort and survey individuals into ‘matching groups’ with similar $x$-distributions (as measured by certain matching scores) and then distribute the survey weights (evenly by the PSAS method or fractionally by the KW method) to the matched cohort units. As a result, the marginal $x$-distribution in the pseudo-weighted $s_c$ should be closer to the $x$-distribution in the $FP$ (estimated by the sample-weighted $s_s$). The balancing score, defined below, should be used to group (or match) cohort and survey units so that the individuals sharing the same balancing score have the same $x$-distribution in $s_c$ and in $FP$. The balancing score $b(x)$ is a function of covariates $x$ such that the conditional distribution of $x$ given $b(x)$ is the same in the $s_c$ as that in the $FP$. We use the notation in Rosenbaum and Rubin (1983).

\[ x \perp R | b(x), \]  

where $R$ is defined in Equation 4.3. The coarsest balancing score is $p = \Pr(R = 1|x)$, or any one-to-one function of $p$, for example, the participation rate $\pi_c(x) = \frac{p}{1 - p}$ (Wang et al., 2021) or $q = \logit(p)$ (Rubin & Thomas, 1992), which can be estimated from the propensity model (4.1) fitted to the combined ($s_c$ vs. weighted $s_s$) sample.

For estimation of the finite population mean $\mu = \frac{1}{N} \sum_{i \in FP} y_i$, the requirement that the matching score should be a balancing score satisfying (4.4) can be relaxed to the WEA:

\[ E\{y|b(x), s_c\} = E\{y|b(x), FP\}, \]  

where $E$ is the expectation with respect to the distribution of $y$ in $FP$.

The matching scores $p$ and $q$ satisfy WEA (4.5) because they are balancing scores as defined in (4.4). Because $p \in (0,1)$ and the distribution of $p$ is typically very right skewed as the participation rate of the cohort is small, then very small differences in $p$ may result from large differences in covariates $x$, which can bias estimators when PS-based matching methods are applied. These boundary problems can be avoided by using $q = \logit(p)$ (Rubin & Thomas, 1992) as the matching score.

4.2 Enhanced Kernel Weighting (KW.W) Approach under WEA

The KW approach, as a special case of PS-based matching methods, has been proved to provide consistent estimators of finite population means under the SEA along with standard conditions (Wang et al., 2020), whereas other PS-matching methods can require further unrealistic
assumptions to obtain consistency of the estimators of the mean. For example, the PSAS method evenly distributes the survey sample weights to cohort units within subclasses defined by quintiles of the estimated PS that assumes that all cohort units within a subclass represent the same number of population units. This key assumption that ensures consistency of the pseudo-weighted estimator holds only if the units within a subclass have the same value of the PS.

We propose an enhanced KW (referred to as KW.W) method by applying the WEA under the framework of propensity estimation described in the beginning of Section 4 and provide statistical properties of KW.W estimators of finite population means. Similar to the KW method, the KW.W method provides pseudo-weights, denoted by \( w_{K^W}^i, i \in s_c \), but the KW.W method uses \( q = \text{logit}(p) = \beta^T x \) as a matching score, with \( \beta \) estimated under Model (4.1) fitted to the combined \((s_c \text{ vs. weighted } s_s)\) sample by solving pseudo-estimating Equation 4.3. Denote the estimated logit of PSs to be \( q^{(c)}_i \) and \( q^{(s)}_j \) for \( i \in s_c \) and \( j \in s_s \), respectively. The KW.W pseudo-weight, \( w_{K^W}^i \) for \( i \in s_c \), is calculated as

\[
w_{K^W}^i = \sum_{j \in s_s} K \left( \frac{(q^{(c)}_i - q^{(s)}_j)}{h} \right) \cdot d_j,
\]

where \( K(\cdot) \) and \( h \) are the kernel function and the bandwidth (see Equation 3.2) corresponding to the distribution of \( \{q^{(c)}_i, i \in s_c\} \) and the \( K(\cdot) \). The KW.W estimator of \( \mu \) is \( \hat{\mu}_{K^W} = \left( \sum_{i \in s_c} w_{K^W}^i \right)^{-1} \left( \sum_{i \in s_c} w_{K^W}^i y_i \right) \). Based on the theorem in Wang et al. (2020) and changing the SEA (3.3) to the WEA (4.6) and under assumptions A1 and A2 and conditions C1–C5 in Appendix A.1, the \( \hat{\mu}_{K^W} \) is an consistent estimator of the population mean, that is, \( \hat{\mu}_{K^W} = \mu + O_p(n_c^{-1/2}) \). The following theorem provides the expression for the finite population variance of \( \hat{\mu}_{K^W} \):

**Theorem** Under the WEA (4.6), conditions A1 and A2 and C1–C5, C6, C8 and C9 in Appendix A.1 and assuming the logistic regression model (4.1) for the propensity scores \( p = \Pr(R = 1 | x) \), the finite population variance of \( \hat{\mu}_{K^W} \) is \( \text{Var}(\hat{\mu}_{K^W}) = V_{K^W} + o(n_c^{-1}) \), where

\[
V_{K^W} = N^{-2} \sum_{i \in Fp} \pi_i^{(c)} \left( 1 - \pi_i^{(c)} \right) \left\{ w_{K^W}^i (y_i - \mu) - (1 - p) b^T x_i \right\}^2 + b^T Db
\]

with \( b = \left\{ \sum_{i \in Fp} \pi_i^{(c)} (y_i - \mu) \omega_{K^W}^i / \partial \beta \right\} \left( \sum_{i \in Fp} p_i x_i^T \right)^{-1} \), \( D = N^{-2} V_p \left( \sum_{i \in Fp} \delta^{(s)} d_i p_i x_i \right) \), and \( V_p \) denoting the design-based finite population variance under the probability sampling design for \( s_c \). Notice that \( \omega_{K^W}^i / \partial \beta \) depends on the choice of kernel function \( K(\cdot) \) (proof in Appendix A.2).

A consistent sample estimator of \( V_{K^W} \) can be obtained by substituting the finite population quantities by consistent sample estimators (Appendices B.1 and B.2).

### 4.3 Improving Efficiency by Scaling the Survey Weights in Propensity Estimation

The propensity estimation \( \hat{\mu} \) can be inefficient because of the generally large variability of the weights among the combined \( s_c \) (with common weight of one) and sample-weighted \( s_s \) (with survey weights \( d_i, i \in s_s \)) (Wang et al., 2020). An analogous situation occurs in

\[
\text{International Statistical Review (2022), 90, 1, 146–164}
\]

© 2021 International Statistical Institute. This article has been contributed to by US Government employees and their work is in the public domain in the USA.
population-based case–control studies where the weights are often highly variable among cases and controls, where cases have small sample weights (often near the value 1) and controls have large sample weights. In these situations, scaling the weights has been suggested to improve efficiency of estimators (Landsman & Graubard, 2013; Li et al., 2010; Scott & Wild, 1986). Following this rationale, we propose scaling the survey weights \( \{d_i, i \in s_c\} \) by the scaling factor \( a = n_s/(\sum_{i \in s} d_i) \) and denote the scaled weight for the survey unit \( i \in s_c \) by \( d_i^* = a \cdot d_i \) so that \( \sum_{i \in s} d_i^* = n_s \). The propensity model (4.1) is fitted to the combined \((s_c, \text{vs. scaled-weighted } s_s)\) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) population, \( \text{minus} \) the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the pseudo-estimating Equation 4.3 with the population-based case \( s_s \) sample and the proof of Lemma in Appendix A.3).

The Lemma shows that rescaling survey weights in the combined sample for the propensity modelling only affects the intercept of the coefficients, which can be corrected by the offset of loga. The resulting KW.S weights, denoted by \( w_i^{KW.S} \), is not affected by the biased intercept, \( \beta_0^* \), because it cancels out in the distance measure of the matching score:

\[
q_i^{*(c)} - q_j^{*(s)} = \beta_1^T \left\{ x_i^{(c)} - x_j^{(s)} \right\} = q_i^{(c)} - q_j^{(s)},
\]

where \( \beta_1^* \) and \( \beta_1 \) are slopes in \( \beta^* \) and \( \beta \), respectively. Therefore, \( \hat{\mu}_{KW.S} = \mu + Op\left(n_c^{-1/2}\mu\right) \), based on Wang et al. (2020), under the WEA (4.6), conditions A1 and A2 and C1–C5 in Appendix A.1. The following corollary provides the expression for the finite population variance of \( \hat{\mu}_{KW.S} \).

Corollary Under WEA, conditions A1 and A2 and C1–C5 and C7–C9 in Appendix A.1, and assuming the logistic regression model (4.1), the finite population variance of \( \hat{\mu}_{KW.S} \) is \( \text{Var}(\hat{\mu}_{KW.S}) = V_{KW.S} + o\left(n_c^{-1}\right) \), with

\[
V_{KW.S} = N^{-2} \sum_{i \in FP} \left(1 - \pi_i^{(c)}\right) \left\{ w_i^{KW.S} (y_i^c - \mu) - (1 - p_s^c) b^* x_i \right\}^2 + b^* D^* b^*,
\]

where \( b^* \) and \( D^* \) are obtained by replacing \( w_i^{KW.W} \), \( p_i \) and \( d_i \) with \( w_i^{KW.S} \), \( p_i^s \) and \( d_i^s \) in \( b \) and \( D \) defined in the Theorem, respectively (proof and details in Appendix A.4).

Notice the trade-off between bias and variance for matching methods when we relax the SEA to WEA. Using \( q \) or \( q^* \) as the matching score relaxes the SEA to WEA, the resulting KW estimators \( \hat{\mu}_{KW.W} \) and \( \hat{\mu}_{KW.S} \) can have larger variance than the original KW estimator, \( \hat{\mu}_{KW} \) (Wang et al., 2020), due to variable weights among cohort and survey sample for the propensity estimation. However, using \( \tilde{q} = \logit \tilde{p} \) as the matching score, the estimator \( \hat{\mu}_{KW} \) is more efficient than the estimators \( \hat{\mu}_{KW.W} \) and \( \hat{\mu}_{KW.S} \) but can be biased if SEA is invalid. Therefore, we identify situations where SEA is satisfied under WEA so that \( \hat{\mu}_{KW} \) is unbiased and has the smallest MSE. Under the propensity model (3.1), we have \( x \perp \perp T \tilde{q} \) with \( T \) indicating the group membership of \( s_c \) versus \( s_s \) (Rosenbaum & Rubin, 1983) and \( E\{y|q, s_c\} = E\{y|\tilde{q}, s_s\} \). However, \( \tilde{q} \) may not satisfy the second equality in the SEA, that is, \( E\{y|\tilde{q}, s_c\} \neq E\{y|\tilde{q}, FP\} \). Assuming \( q \) satisfies WEA (4.5), if \( \tilde{q} \) is a one-to-one function (e.g. Cases 1 and 2 in Figure 1) or many-to-one function (e.g. Case 3 in Figure 1) of \( q \), the second equality of the SEA \( E\{y|\tilde{q}, s_c\} = E\{y|\tilde{q}, FP\} \) holds
as $E[y|q, s_c] = E[y|q, FP]$. Otherwise, $\tilde{q}$ will not satisfy the second equality the SEA (e.g. Cases 4 and 5 in Figure 1). As a result, $\tilde{\mu}^{KW}$ can be biased.

4.4 PS-Based Matching Methods Compared with Weighting Methods

Based on the propensity estimation method described in Section 4, the cohort participation rate can be estimated by $\tilde{\pi}_i^{(c)} = \tilde{\pi}_i/(1 - \tilde{\pi}_i)$ for $i \in s_c$ (Wang et al., 2021). A commonly used PS-based weighting method is the IPSW, which uses $w_i^{IPSW} = 1/\tilde{\pi}_i^{(c)}$ as the pseudo-weight for $i \in s_c$. Consistency and variance estimation of $\tilde{\mu}^{IPSW} = (\sum_{i \in s_c} w_i^{IPSW} y_i)/(\sum_{i \in s_c} w_i^{IPSW})$ are provided in Wang et al. (2021). Similar to the KW.S method, rescaled IPSW, called IPSW.S method, where the pseudo-weights are $w_i^{IPSW.S} = \{a \cdot \tilde{\pi}_i/(1 - \tilde{\pi}_i)\}^{-1}$, can improve efficiency of weighted estimation over using IPSW (see Remark in Appendix A.4 about the consistencty and finite population variance of the $\tilde{\mu}^{IPSW.S} = (\sum_{i \in s_c} w_i^{IPSW.S} y_i)/(\sum_{i \in s_c} w_i^{IPSW.S})$).

The consistency of PS-based weighting estimators $\tilde{\mu}^{IPSW}$ and $\tilde{\mu}^{IPSW.S}$ do not require SEA or WEA as the matching method do but require that the propensity model (4.1) is correctly specified for propensity estimation. Moreover, they can have larger variance compared with the PS-based matching estimators due to potential extreme pseudo-weights. Estimators $\tilde{\mu}^{IPSW}$ and $\tilde{\mu}^{IPSW.S}$ are included for comparison purposes in the simulation and data example.

5 Simulations

We conducted simulation studies to evaluate the finite sample performance of the proposed PS-based pseudo-weighted estimators of finite population mean, $\mu$, that is, $\tilde{\mu}^{KW.W}$ and $\tilde{\mu}^{KW.S}$, compared with the naïve cohort estimator $\tilde{\mu}^{Naive}$, $\tilde{\mu}^{KW}$, $\tilde{\mu}^{IPSW}$ and $\tilde{\mu}^{IPSW.S}$.

5.1 Generating the Finite Population

We generated a finite population (FP) of size $N = 200,000$, with four covariates $x_1 \sim N(1,1)$, $x_2 \sim N(1,1)$, $x_3 \sim \text{LogNormal}(0,0.7)$ and $x_4 (=1$ if $x_1+x_2 > 2; 0$ otherwise). Note $x_1$ and $x_2$ are correlated with $x_4$, but independent of $x_3$. The outcome $y$ for $i \in FP$ was generated by $y_i = 2 + x_{1,i} + x_{2,i} + x_{4,i} + \epsilon_i$, $i \in FP$, where the error terms $\epsilon_i$ were independent and identically distributed (i.i.d.) as $N(0,1)$. The finite population mean of $y$ is $\mu = 4.50$.

For each $i \in FP$, we created two variables, $x_{1,i}^*$ and $x_{1,i}^{**}$, as functions of $x_1$: $x_{1,i}^* = x_{1,i} + 0.15x_{1,i}$ and $x_{1,i}^{**}$ was defined as a categorical variable (=1 if $x_{1,i} \leq 10^{th}$ percentile; 2 if $10^{th} < x_{1,i} \leq 40^{th}$ percentiles; 3 if $40^{th} < x_{1,i} \leq 70^{th}$ percentiles; 4 if $70^{th} < x_{1,i} \leq 90^{th}$ percentiles; and 5 if $x_{1,i} > 90^{th}$ percentile of $x_1$ in the FP). The variables of $x_1^*$ and $x_1^{**}$ were used in the simulations as a substitute of the covariate $x_1$ to reflect cases when $x_1$ is not available but related variables are available.
5.2 Sampling from the Finite Population to Assemble the Survey Sample and Cohort

Probability proportional to size (PPS) sampling is commonly applied in large-scale national surveys (such as NHIS and NHANES) to select the first-stage sampling units. Guided by real data, we conducted a PPS sampling to select the cohort and the survey sample in simulation studies.

This PPS sample design enabled us to form a fair comparison among the five pseudo-weighting methods (KW, IPSW, KW.W, IPSW.S and KW.S) because they have the same functional form as explained below and in Appendix A.5.

A cohort of size $n_c = 2400$ individuals was randomly selected from the $FP$ by PPS sampling with the measure of size (MOS) for individual $i \in FP$ defined by $m_i^{(c)} = \exp (a_1 x_{1,i} + a_2 x_{2,i} + a_3 x_{3,i})$, where $a = (a_1, a_2, a_3) = (0.6, 0.15, 0.24)$. The sample weight (i.e. the reciprocal of the selection probability) for individual $i$ in the cohort was $w_i^{(c)} = \sum_{c=1}^{N_c} m_i^{(c)}/n_c \cdot m_i^{(c)}$. A survey sample of size $n_s = 2000$ individuals was sampled independently of the sampling of the cohort where a similar PPS sampling design was used, but with a different MOS $m_i^{(s)} = \exp (\gamma_1 x_{1,i} + \gamma_2 x_{2,i} + \gamma_3 x_{3,i})$.

Under the PPS sampling described above, the true propensity model of a population unit included in $s_c$ versus $FP$ (assumed by the IPSW, KW.W, IPSW.S and KW.S methods) and in $s_c$ versus $s_s$ (assumed by the original KW) were, respectively,

$$\text{logit}\{P(i \in s_c | i \in s_c \cup s_s)\} = \beta_0 + \beta_1^T x_i,$$  \hspace{1cm} (5.1)

and

$$\text{logit}\{P(i \in s_s | i \in s_c \cup s_s)\} = \tilde{\beta}_0 + \tilde{\beta}_1^T x_i,$$  \hspace{1cm} (5.2)

respectively, where $\beta_1 = a, \tilde{\beta}_1 = a - \gamma, \beta_0$ and $\tilde{\beta}_0$ are the intercepts (Appendix A.5) and $x_i = (x_{1,i}, x_{2,i}, x_{3,i})^T$. The two propensity models have the same functional form so that the proposed PS-based weighting and matching methods can be fairly compared. Notice that values of $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ in MOS of survey sample selection can be varied to control for the validity of the SEA assumed by the original KW method (Section 5.3).

5.3 Methods and Evaluating Criteria

We considered two scenarios. In Scenario 1, we set $\gamma = (-0.4, -0.1, -0.16)$ for MOS of survey sample selection in Scenario 1 so that SEA is satisfied for the KW method under Model (5.2). This is because $q_i = \tilde{\beta}_0 + x_{1,i} + 0.25 x_{2,i} + 0.4 x_{3,i}$ under the KW true propensity model (5.2) is a one-to-one function of the balancing score $\gamma_1 = \beta_0 + 0.6 x_{1,i} + 0.15 x_{2,i} + 0.24 x_{3,i}$ under the true model (5.1) (i.e. $q_i = 0.6 (q_i - \tilde{\beta}_0) + \beta_0$), Figure 2). In Scenario 2, $\gamma = (-0.65, 0.2, 0.0)$ so that SEA was not satisfied even under the true propensity model as the matching score $\tilde{q}$ is not a balancing score (Figure 2).

Under Scenario 2, we considered three mis-specified propensity models in addition to the true propensity model (Model T): (1) underfitted Model U, excluding $x_3$ in the true propensity model; (2) mis-specified Model $M_1$, substituting the true covariate $x_1$ in Model U by variable...
\(x^*_1\), a one-to-one function of \(x_1\); and (3) mis-specified Model \(M_2\), substituting the true covariate \(x_1\) in Model \(U\) by a categorical version of \(x_1\), that is, \(x^{**}_1\).

We used criteria of relative bias (\(\%RB\)), empirical variance (\(V\)) and MSE of the estimators to evaluate performance of the five PS-based methods, defined by

\[
MSE = \frac{1}{B} \sum_{b=1}^{B} \left\{ \hat{\mu}^{(b)} - \mu \right\}^2,
\]

where \(B = 10,000\) is the number of simulations and \(\hat{\mu}^{(b)}\) is the estimate of FP mean \(\mu\), obtained from the \(b\)-th simulated samples.

For each mean estimator, we evaluated two variance estimators, that is, the proposed TL estimator (described in Appendix B.2) and the jackknife replication (JK) estimator (Appendix B.3), using the variance ratio (VR), and coverage probabilities (CP) of the corresponding 95% confidence intervals, defined by

\[
VR = \frac{1}{B} \sum_{b=1}^{B} \hat{\gamma}^{(b)},
\]

\[
CP = \frac{1}{B} \sum_{b=1}^{B} I \left( \mu \in CI^{(b)} \right),
\]

where \(\hat{\gamma}^{(b)}\) is the variance estimate of \(\hat{\mu}^{(b)}\) and \(CI^{(b)} = \left( \hat{\mu}^{(b)} - 1.96 \sqrt{\hat{\gamma}^{(b)}}, \hat{\mu}^{(b)} + 1.96 \sqrt{\hat{\gamma}^{(b)}} \right)\) is the 95% confidence interval from the \(b\)-th simulated sample.
5.4 Results under Scenario 1: The valid SEA

Table 1 shows the results under the SEA. The unweighted naïve cohort mean, $\hat{\mu}^{Naive}$, was biased by 20.97%, whereas the survey estimate $\hat{\mu}^{SVY}$ was approximately unbiased. All KW and IPSW methods yielded approximately unbiased estimates of $\mu$. The KW estimator, $\hat{\mu}^{KW}$, had the smallest variance because no sample weights were considered in estimating PSs. The $\hat{\mu}^{KW}$ required the SEA and naturally had the smallest MSE (in bold) and maintained the nominal CP in Scenario 1.

5.5 Results under Scenario 2: The invalid SEA

Table 2 shows results under four fitted propensity models that includes different sets of covariates. The method providing smallest MSE is in bold. Under the correctly specified propensity model (Model T), though the KW estimator, $\hat{\mu}^{KW}$, had the smallest variance among the five pseudo-weighted estimators, it had the largest bias, leading to low CP and the largest MSE, whereas $\hat{\mu}^{KW,W}$ and $\hat{\mu}^{KW,S}$ had smaller biases. Similar to the results in Table 1, scaling survey weights in the propensity model yielded more efficient estimates, especially for the IPSW method. Though $\hat{\mu}^{IPSW,S}$ had ~60% smaller variance than $\hat{\mu}^{IPSW}$, it was not as efficient as $\hat{\mu}^{KW,S}$. As a result, $\hat{\mu}^{KW,S}$ performed the best in terms of MSE.

Excluding $x_4$ in the propensity model (Model U) did not affect the extent of the bias of the estimates because $x_4$ was uncorrelated with the outcome variable $y$. However, the empirical variances of $\hat{\mu}^{IPSW}$ and $\hat{\mu}^{KW,W}$ were substantially reduced compared with the variances under Model T (similar findings as in Wang et al., 2020). In contrast, the variances of $\hat{\mu}^{IPSW,S}$ and $\hat{\mu}^{KW,S}$ were nearly unchanged.

In Model M1, the true covariate $x_1$ in Model U was substituted by $x_1^*$, which is an one-to-one function of $x_1$. The IPSW estimators $\hat{\mu}^{IPSW}$ and $\hat{\mu}^{IPSW,S}$ were biased because the cohort participation rates cannot be accurately estimated from Model M1. However, the matching methods with matching scores $\tilde{q} = \tilde{\beta}_1x_1^* + \tilde{\beta}_2x_2$ still worked well because it was close to a one-to-one function of the true participation rate. As a result, the KW estimates were less biased than the

Table 1. Results from 10,000 simulated cohorts and survey samples under SEA.

| Estimator          | % RB $^a$ | $V^b \times 10^3$ | VR $^c$ (TL) | VR (JK) | CP $^d$ (TL) | CP (JK) | MSE $^e \times 10^3$ |
|--------------------|-----------|-------------------|--------------|---------|--------------|---------|----------------------|
| $\hat{\mu}^{Naive}$ | 20.97     | 1.71              | 1.02         | 1.02    | 0.00         | 0.96    | 889.76               |
| $\hat{\mu}^{SVY}$  | 0.07      | 2.82              | 1.01         | 1.02    | 0.96         | 0.96    | 2.83                 |
| Model T (true) logit[Pr(y)] $\sim x_1, x_2, x_4$ | -0.12     | 9.96              | 0.92         | 1.01    | 0.94         | 0.95    | 9.99                 |
| $\hat{\mu}^{IPSW}$ | 0.07      | 4.59              | 0.99         | 0.99    | 0.95         | 0.95    | 4.60                 |
| $\hat{\mu}^{IPSW,S}$ |          |                   |              |         |              |         |                      |
| $\hat{\mu}^{KW}$   | 0.18      | 2.54              | 1.07         | 1.06    | 0.96         | 0.95    | 2.61                 |
| $\hat{\mu}^{KW,W}$ | 0.66      | 4.02              | 1.01         | 1.07    | 0.93         | 0.94    | 4.90                 |
| $\hat{\mu}^{KW,S}$ | 0.63      | 3.12              | 1.03         | 1.07    | 0.93         | 0.93    | 3.92                 |

$^a$%RB is the relative bias of the sample mean of 10,000 simulation estimates from the true population mean.

$^b$V is the empirical (sample) variance of the 10,000 simulation estimates.

$^c$VR is the ratio of analytical variance (sample mean of 10,000 simulation variance estimates, with TL and JK representing the Taylor linearisation and the jackknife methods, respectively) and the empirical variance ($V$).

$^d$CP is the coverage probability calculated by the proportion of times that the 95% confidence interval covers the true population mean, with standard error estimated by the TL and the JK methods.

$^e$MSE is the mean squared error calculated by the sample mean of 10,000 simulation squared bias. Method providing smallest MSE is in bold.

International Statistical Review (2022), 90, 1, 146–164
© 2021 International Statistical Institute. This article has been contributed to by US Government employees and their work is in the public domain in the USA.
Table 2. Results from 10,000 simulated cohorts and survey samples with each cohort and survey sample fitted to the correct propensity model and three mis-specified propensity models with violated SEA.

| Estimator | %RB | V^b (×10^4) | VR^c (TL) | VR (JK) | CP^d (TL) | CP (JK) | MSE^e (×10^4) |
|-----------|-----|-------------|-----------|---------|-----------|---------|---------------|
| \(\hat{\mu}_{\text{Naive}}\) | 20.97 | 1.72 | 1.02 | 1.02 | 0.00 | 0.95 | 889.89 |
| \(\hat{\mu}_{\text{SVY}}\) | 0.04 | 3.61 | 1.02 | 1.02 | 0.95 | 0.95 | 3.62 |
| Model T (true) \(\logit\{Pr(x)\} \sim x_1, x_2, x_4\) | -0.36 | 14.73 | 0.86 | 1.03 | 0.93 | 0.94 | 14.99 |
| \(\hat{\mu}_{\text{PSW}}\) | 0.03 | 5.92 | 0.98 | 1.00 | 0.95 | 0.95 | 5.92 |
| \(\hat{\mu}_{\text{KW}}\) | 4.84 | 2.66 | 0.92 | 1.04 | 0.01 | 0.02 | 50.03 |
| \(\hat{\mu}_{\text{KW W}}\) | 0.84 | 4.83 | 1.04 | 1.09 | 0.93 | 0.94 | 6.24 |
| \(\hat{\mu}_{\text{KW S}}\) | 0.65 | 3.55 | 1.02 | 1.08 | 0.93 | 0.93 | 4.39 |
| Model U (underfitted) \(\logit\{Pr(x)\} \sim x_1, x_2\) | -0.24 | 13.62 | 0.89 | 1.02 | 0.93 | 0.94 | 13.73 |
| \(\hat{\mu}_{\text{PSW S}}\) | 0.04 | 5.76 | 0.98 | 1.00 | 0.95 | 0.95 | 5.77 |
| \(\hat{\mu}_{\text{KW W}}\) | 0.80 | 3.92 | 1.12 | 1.13 | 0.94 | 0.94 | 5.22 |
| \(\hat{\mu}_{\text{KW S}}\) | 0.57 | 3.39 | 1.02 | 1.08 | 0.93 | 0.94 | 4.05 |
| Model M_1 (mis-specified variable) \(\logit\{Pr(x)\} \sim x_1^*, x_2\) | -0.24 | 13.62 | 0.89 | 1.02 | 0.93 | 0.94 | 13.73 |
| \(\hat{\mu}_{\text{PSW S}}\) | 0.04 | 5.76 | 0.98 | 1.00 | 0.95 | 0.95 | 5.77 |
| \(\hat{\mu}_{\text{KW W}}\) | 0.80 | 3.92 | 1.12 | 1.13 | 0.94 | 0.94 | 5.22 |
| \(\hat{\mu}_{\text{KW S}}\) | 0.57 | 3.39 | 1.02 | 1.08 | 0.93 | 0.94 | 4.05 |
| Model M_2 (mis-specified variable) \(\logit\{Pr(x)\} \sim x_1^*, x_2\) | -0.24 | 13.62 | 0.89 | 1.02 | 0.93 | 0.94 | 13.73 |
| \(\hat{\mu}_{\text{PSW S}}\) | 0.04 | 5.76 | 0.98 | 1.00 | 0.95 | 0.95 | 5.77 |
| \(\hat{\mu}_{\text{KW W}}\) | 0.80 | 3.92 | 1.12 | 1.13 | 0.94 | 0.94 | 5.22 |
| \(\hat{\mu}_{\text{KW S}}\) | 0.57 | 3.39 | 1.02 | 1.08 | 0.93 | 0.94 | 4.05 |

%RB is the relative bias of the sample mean of 10,000 simulation estimates from the true population mean.

V^b is the empirical (sample) variance of the 10,000 simulation estimates.

VR is the ratio of analytical variance (sample mean of 10,000 simulation variance estimates, with TL and JK representing the Taylor linearisation and the jackknife methods, respectively) and the empirical variance (V).

CP is the coverage probability calculated by the proportion of times that the 95% confidence interval covers the true population mean, with standard error estimated by the TL and the JK methods.

MSE is the mean squared error calculated by the sample mean of 10,000 simulation squared bias. Method providing smallest MSE is in bold.

IPSW estimates. Furthermore, by scaling survey weights, \(\hat{\mu}_{KW S}\) performed best with smallest bias, variance and nearly nominal CP.

In contrast, Model M_2 substituted \(x_1\) by \(x_1^*\), which was a categorical variable that was coarser than \(x_1\) in Model U. This model did not accurately estimate the cohort participation rates or provide an adequate balancing score used for matching, because individuals in the same category of \(x_1^*\) took on the same values of the matching scores \(\hat{q}\) and were incorrectly assigned the same pseudo-weights. Hence, all of the pseudo-weighted estimates were biased.

The TL variance estimates were close to the truth (with VR \(\approx 1\)) for all estimates except for \(\hat{\mu}_{PSW}\) with its VR \(<1\). This result is due to the finite sample bias caused by the large variability of the sample weights in the combined \((s_c, s_{sc})\) with the common value of one for the cohort weights versus values ranging from 23 to 618 for the survey weights (same findings in Landsman & Graubard, 2013; Li et al., 2010). In contrast, the TL variance estimate for \(\hat{\mu}_{PSW S}\) worked well because the variability of the weights in the combined \((s_c, s_{sc})\) sample was reduced. The scaled survey sample weights range from 0.2 to 5.1.

The JK method consistently had larger estimates of variances compared with the TL variance estimates (similar results were shown by Efron & Gong, 1983), and the JK estimates were more accurate for the variance of \(\hat{\mu}_{PSW}\). However, in some of the simulations, the JK overestimated
the variance of $\hat{\mu}^{KW}$ and $\hat{\mu}^{KW,S}$ due to highly variable covariates in the propensity model (Model $M_1$).

In summary, scaling the survey weights substantially not only decreased the variance of the mean estimates but also reduced the finite sample bias of the TL variance estimates. The resulting estimates $\hat{\mu}^{IPSW,S}$ and $\hat{\mu}^{KW,S}$ outperformed $\hat{\mu}^{IPSW}$ and $\hat{\mu}^{KW}$, respectively. The proposed $\hat{\mu}^{KW,S}$ generally had the smallest variance among the four methods, and its variance changed least among all the four estimates as the fitted propensity model varied. The proposed $\hat{\mu}^{KW,S}$ had the smallest MSE when the propensity model was appropriately specified (Models $T$ and $U$). Under Model $M_1$, when the variable(s) in the fitted propensity model was no coarser than the correct variable(s), $\hat{\mu}^{KW,S}$ was robust to model mis-specification and therefore unbiased and more efficient than $\hat{\mu}^{IPSW,S}$. Under Model $M_2$, the performance of $\hat{\mu}^{KW,S}$ and $\hat{\mu}^{IPSW,S}$ was comparable, and $\hat{\mu}^{IPSW,S}$ had slightly smaller MSE than $\hat{\mu}^{KW,S}$ due to the smaller bias.

6 Data Analysis: The US NHANES

For our example, we used the Third US National Health and Nutrition Examination Survey (NHANES III) as the volunteer-based ‘cohort’ (ignoring sample weights) and the contemporaneous US NHIS as the reference survey. This example has several advantages for illuminating the performance of our methodology. The ‘cohort’ and reference survey have approximately the same target population, data collection mode and questionnaires. This ensures that when applying our methodology to the ‘cohort’ in an effective manner, we should be able approximately recover US representative estimates, enabling us to characterise the performance of our methodology with real data. Although problems with misaligned target populations and data harmonisation are serious practical issues, they are beyond the scope of our methodology.

We estimated prospective 15-year all-cause mortality rates for adults in the United States using the adult sample of household interview part of NHANES III conducted in 1988–1994, with sample size = 20,050. NHANES III is partly a cross-sectional household interview survey and partly a medical examination survey of the civilian, non-institutionalised population of the United States. NHANES III oversampled poverty areas, children under age 5, adults aged 60 and over, non-Hispanic blacks and Mexican Americans (Ezzati et al., 1992). The coefficient of variation (CV) of sample weights is 125%, indicating highly variable selection probabilities and thus potential low representativeness of the unweighted sample. We ignored all complex design features of NHANES III and treated it as a cohort. For estimating mortality rates, we approximate 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 the 1994 NHIS respondents to the supplement for monitoring achievement of the Healthy People Year 2000 objectives, aged 18 and older (sample size = 19,738). NHIS is also a cross-sectional household interview survey of the same target finite population as the NHANES III. The 1994 NHIS had a multistage stratified cluster sample design, with over sampling of the aged, low income and Black and Hispanic populations (Massey et al., 1989). There were 125 strata and 248 pseudo-PSUs in the sample. We collapsed strata with only one PSU with the next nearest strata for variance estimation purpose (Hartley et al., 1969). The CV of sample weights in 1994 NHIS sample is 58%. NHANES III and NHIS were linked to National Death Index (NDI) for mortality (National Center for Health Statistics, 2013), allowing us to quantify the relative bias of unweighted NHANES estimates, assuming the NHIS estimates as the gold standard.

International Statistical Review (2022), 90, 1, 146–164
© 2021 International Statistical Institute. This article has been contributed to by US Government employees and their work is in the public domain in the USA.
We first compare the distributions of selected common covariates in the two samples (Table C.1 in Appendix C). As expected, the covariates in the weighted samples of NHANES and 1994 NHIS have very close distributions because both weighted samples represent approximately the same finite population. There are two exceptions: (1) education level, probably due to differences in how the question was asked in the two surveys, and (2) health status, which was self-reported in NHANES but reported by the proxy of household representative in NHIS. As expected, the covariates distribute quite differently in the unweighted NHANES from the weighted samples, especially for design variables such as age, race/ethnicity, poverty and region.

We use an AIC-based stepwise procedure (Lumley, 2020) to choose the propensity model fitted to combined sample of unweighted NHANES and weighted NHIS. This initially included main effects of common demographic characteristics (age, sex/race/ethnicity, region and marital status), socio-economic status (education level, poverty and household income), tobacco usage (smoking status and chewing tobacco), health variables (body mass index [BMI] and self-reported health status), a quadratic term for age and all two-way interactions. Table C.2 in Appendix C shows the final propensity models fitted to the weighted sample (for IPSW and KW.W), scaled weighted sample (for IPSW.S and KW.S) and unweighted sample (for KW).

To evaluate the performance of the five PS-based methods, we used relative difference from the NHIS estimate \( \%RD = \left( \frac{\hat{\mu} - \hat{\mu}_{\text{NHIS}}}{\hat{\mu}_{\text{NHIS}}} \right) \times 100 \), bias reduction from the naïve (unweighted) NHANES estimates \( \%BR = \left( \frac{\hat{\mu}_{\text{Naive}} - \hat{\mu}}{\hat{\mu}_{\text{Naive}}} \right) \times 100 \), TL variance estimate \( (V) \) and estimated MSE \( = \left( \frac{\hat{\mu} - \hat{\mu}_{\text{NHIS}}}{\hat{\mu}_{\text{NHIS}}} \right)^2 + V \), which treated the NHIS estimates as truth.

Table 4 shows that the weighted 1994 NHIS and the sample-weighted NHANES III estimates (TW) of 15-year all-cause mortality were very close (%RD = 2.6% for overall estimate and %RD = 4.5–7.3% on average for the estimates by subgroups). In contrast, the naïve NHANES III estimate of overall mortality was ~52% biased from the NHIS estimate because older people who have higher mortalities were oversampled and the bias in subgroup-specific mortality reached 96.8% for non-Hispanic Whites. All KW and IPSW methods substantially reduced the bias from the naïve estimates. The four methods that fit propensity models to the (scaled-)weighted sample (IPSW, IPSW.S, KW.W and KW.S) provided the closest estimates. The bias in the naïve estimate of overall mortality was almost eliminated by the KW.W and KW.S methods (~96% bias removed). KW.S on average had the least bias for the subgroup-specific mortality among the four methods. Similar to the simulation results, KW.W and KW.S estimates had smaller variance (estimated by TL method) than the IPSW and IPSW.S estimates. As a result, the KW.S estimates had on average the smallest MSE.

Importantly, the original KW method had the largest bias in overall mortality (BR\% = 66% vs. ≥92.7\%) but had least bias for age-specific mortality (BR\% = 43.2% vs. ≤31.0%) and achieved smallest MSE for most age groups (in italic in Table 4). This paradox can be explained by the invalidity of the SEA for overall mortality, but not for age-specific mortality. As shown in Table 4, the small biases in the KW estimates of age-specific mortality imply that the SEA held with \( E(y|\text{age}, \hat{\rho}c) = E(y|\text{age}, \hat{\rho}, \text{FP}) \). As shown in Table 3, the KW (in bold if yielding largest relative difference) pseudo-weighted age distribution in \( sc \) (unweighted NHANES sample) differed from that in \( FP \) (represented by the weighted NHIS), indicating \( P(\text{age}|\hat{\rho}, s_c) \neq P(\text{age}|\hat{\rho}, \text{FP}) \). As a result, the SEA was invalid for the overall mortality estimation using the original KW method, that is,
This result is consistent with the findings in the simulations: The original KW estimates can have the smallest (or largest) MSE when the SEA is valid (or invalid). The best methods for overall mortality were KW.W and KW.S, both of which require only the WEA to hold.

The other four methods (IPSW, IPSW.S, KW.W and KW.S) had similar estimated mortality rates. The method yielding smallest relative bias, TL variance and MSE among the four was highlighted in bold in Table 4. The IPSW estimates had the largest variances, followed by the IPSW.S estimates. The KW.S estimates had the smallest variances and MSE in most cases. The results of the JK and the TL variance estimates were similar in this example (results not shown).

Table 4. Relative difference of age group proportion estimates from the 1994 NIHS estimates.

| Age group      | IPSW  | IPSW.S | KW   | KW.W | KW.S |
|----------------|-------|--------|------|------|------|
| 18–24 years    | −8.2  | −8.3   | −23.6| −7.3 | −8.3 |
| 25–44 years    | 4.5   | 3.1    | −4.3 | 2.9  | 2.5  |
| 45–64 years    | −1.6  | 0.5    | 4.6  | −0.7 | 0.6  |
| 65–69 years    | −11.7 | −9.0   | 5.4  | −10.6| −8.8 |
| 70–74 years    | −4.0  | −2.1   | 18.8 | −1.8 | 0.1  |
| > = 75 years   | 4.6   | 2.2    | 39.8 | 7.4  | 4.2  |
| Avg.           | 5.8   | 4.2    | 16.1 | 5.1  | 4.1  |

a Estimate with the largest relative difference among the five methods is in bold.

Avg. is the average of the absolute value of relative difference across the subgroups.

\[ E(y|\bar{p}, s_c) = \sum_{\text{age}} \{E(y|\text{age}, \bar{p}, s_c)P(\text{age}|\bar{p}, s_c)\} \]

\[ = \sum_{\text{age}} E(y|\text{age}, \bar{p}, FP)P(\text{age}|\bar{p}, FP) = E(y|\bar{p}, FP). \]

7 Discussion

Epidemiologic cohorts are often a type of non-probability sample that use volunteer-based recruitment for data collection. Improving population representativeness of cohort studies has not received enough attention by biostatisticians or epidemiologists. With an epidemiologic cohort as the illustrative example, this paper aims to bring this area of research to the attention of biostatisticians and epidemiologists while still remaining novel and relevant to survey statisticians.

We established a unified framework for both PS-based weighting and matching methods that improves estimates of finite population means from non-representative cohort data by using a reference representative survey sample of the target population. This unifying framework allows us to make three contributions. First, we identified the underlying SEA, implicitly assumed by existing PS-based matching methods, whose failure invalidates inference even if the PS-model is correctly specified. Our simulations and data example demonstrate that PS-based matching methods that rely on the SEA, such as the original KW estimator (Wang et al., 2020), have smallest MSE when the SEA holds but have large bias when the SEA fails. Second, as a remedy, we proposed PS-based methods that require only a WEA. Third, we further improved the efficiency of PS-based estimates by scaling the survey weights to sum to the survey sample size. Scaling reduces the variance of the estimated PS’s and thus markedly improves efficiency of the pseudo-weighted estimates, especially for the IPSW method. Our recommended method, kernel weighting with scaling the survey weights (KW.S), is more robust by only requiring the WEA, yet the scaling ensures that it has smallest MSE.
**Table 4. Estimates of all-cause 15-year mortality (overall, and by subgroups) with estimated variance and mean squared error.**

|    | Exp$^a$ | %relative difference from the NHIS estimate$^b$ | TL variance estimate$^b$ ($\times 10^5$) | MSE$^bc$ ($\times 10^5$) |
|----|---------|---------------------------------------------|-----------------------------------------|------------------------|
|    | Overall |                               |                                         |                        |
|    |         | %BR$^d$ | TW | Naïve | KW | IPSW | IPSWS | KW.W | KWS | KW  | IPSW | IPSWS | KW.W | KWS |
|    | $^{(66.0)}$ | $^{(92.7)}$ | $^{(93.8)}$ | $^{(96.1)}$ | $^{(95.8)}$ | $^{(5.8)}$ | $^{(4.3)}$ | $^{(2.3)}$ | $^{(2.5)}$ | $^{(7.9)}$ | $^{(8.6)}$ | $^{(10.5)}$ | $^{(11.2)}$ | $^{(12.5)}$ |
| NHIS | Overall | 17.6 | -2.6 | 52.2 | 17.7 | -3.8 | -3.2 | -2.0 | -2.2 | 1.2 | 1.4 | 1.0 | 1.0 | 1.0 | 98.2 | 5.8 | 4.3 | 2.3 |
| Age group | 18–24 | 2.2 | -16.1 | 0.5 | -35.9 | -33.4 | -30.9 | -32.4 | -30.2 | 0.8 | 0.8 | 0.9 | 0.8 | 0.9 | 7.2 | 6.4 | 5.7 | 6.1 |
| | 25–44 | 3.9 | -7.9 | 30.9 | -4.5 | -14.8 | -14.3 | -12.5 | -14.0 | 0.8 | 0.7 | 0.7 | 0.7 | 0.7 | 1.1 | 4.1 | 3.8 | 3.2 |
| | 45–64 | 17.7 | 5.8 | 30.6 | 1.3 | -3.8 | -3.2 | -4.1 | -3.7 | 5.2 | 5.4 | 5.1 | 4.9 | 4.9 | 5.7 | 9.8 | 8.2 | 10.2 |
| | 65–69 | 45.5 | 0.9 | 9.5 | -1.3 | -6.4 | -5.4 | -5.8 | -4.7 | 35.9 | 35.6 | 34.4 | 33.5 | 33.3 | 39.5 | 120.8 | 95.1 | 103.5 |
| | 70–74 | 60.0 | 3.5 | 6.4 | -0.4 | -1.4 | -1.3 | -1.3 | -1.1 | 32.1 | 31.6 | 30.3 | 30.4 | 29.6 | 32.7 | 39.2 | 36.9 | 33.9 |
| | > 74 | 86.2 | 1.1 | 4.3 | 3.3 | 3.3 | 3.2 | 3.1 | 3.1 | 6.0 | 6.1 | 5.8 | 5.9 | 5.8 | 85.5 | 85.7 | 80.2 | 83.4 |
| Avg$\frac{e}{f}$ | 5.9 | 13.7 | 7.8 | 10.5 | 9.7 | 12.9 | 12.7 | 12.5 | 28.6 | 44.3 | 38.3 | 40.5 | 34.4 | 25.7 | 37.0 | 35.8 | 35.9 |

---

$^a$ Est is the sample-weighted NHIS estimate of population mortality.

$^b$ %relative difference from the NHIS estimate is in black and bold for the other four methods (IPSW, KW, W, IPSWS and KWS).

$^c$ %BR is the percentage of bias reduced from the naïve unweighted NHANES estimate. %BR = 100 means removing all bias of the naïve estimate; %BR < 0 means an increased bias; 100 < %BR < 200 means a reduced bias but with the bias in the opposite direction of the naïve estimate; %BR > 200 means an increased bias in the opposite direction.

$^d$ Avg. is the average of the absolute value of bias across the subgroups.

$^e$ Race has four categories: non-Hispanic White (NH-W), non-Hispanic Black (NH-B), Hispanics (Hisp) and non-Hispanic Other (NH-O).
For the variance estimation, we recommend the JK method for the IPSW estimates because our empirical results indicate that the TL method can have greater finite sample bias due to highly variable weights in the combined sample. However, both the JK and the TL methods provided good variance estimation for the IPSW.S estimates. The TL method is recommended for the KW.W and the KW.S estimates because the JK method can overestimate the variance. In our data example, though the original KW method reduced most bias for age-specific mortality rate (~43% bias reduction vs. 23%–31% for the other four methods), it had the largest bias for the overall mortality rate (~18% relative bias vs. −3.8% to −2% for the other four methods). This is because SEA approximately holds for age-specific mortality, but failed for overall mortality (where WEA held). Among all the five PS-based methods, KW.S reduced most bias and obtained smallest MSE on average, due to robustness from matching, WEA and reduction of weight variability from kernel smoothing and scaling survey weights.

Our unifying framework codifies two other key assumptions generally taken for granted. Assumption A1 ensures non-informative sampling of the cohort, allowing for correct estimation of participation rates. Assumption A2 ensures that the cohort and the survey samples cover the same target finite population. Assumption A1 is often reasonable, especially when the outcome is measured after the cohort is assembled, but assumption A2 is generally violated, to some extent, in real data. For example, most cohort studies only recruit people in a few study centres in a target population (e.g. the United States), whereas many surveys are representative of the target population. One solution is to use subgroups of the survey sample that are covered by the cohort as the reference so that the weighted cohort only represents a defined subpopulation. This problem of misaligned coverage between cohort and survey is a critical issue for future research.

There is much room for future research. First, the true propensity model is usually unknown in practice. Criteria such as AIC and BIC can be used for propensity model selection. However, the variable selection procedure can be data driven and depend on covariates available in the two samples. The propensity model selection procedure has not been accounted for in pseudo-weighted analyses. Second, doubly robust estimators (Chen et al., 2019) are a promising approach to enhance robustness to propensity model mis-specification by imputing the completely missing outcome into the survey based on the cohort. Third, diagnostics are needed to assess if the SEA holds in an analysis. Finally, more work is needed to determine optimal scaling factors for rescaling the survey weights used in propensity estimation that minimise the variance of the pseudo-weighted estimators.

ACKNOWLEDGEMENTS

The first author was supported by Predoctoral Fellowship Research Program, the US National Institute of Health/National Cancer Institute and Intramural Research Program of the US National Institutes of Health/National Cancer Institute.

References

Brick, J. M. (2015). Compositional model inference. *JSM Proceedings (Survey Research Methods Section)*, 299–307.
Chen, Y., Li, P., and Wu, C. (2019). Doubly robust inference with nonprobability survey samples. *J. Am. Stat. Assoc.*, 1–11.
Collins, R. (2012). What makes UK Biobank special? *Lancet* 379(9822), 1173–1174.
Duncan, G. J. (2008). When to promote, and when to avoid, a population perspective. *Demography*, 45(4), 763–784.
Efron, B., and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *Amer. Stat.*, 37 (1), 36–48.
Elliott, M. R., and Valliant, R. (2017). Inference for nonprobability samples. *Stat. Sci.*, 32(2), 249–264.
Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory Probab. Appl.*, **14**(1), 153–158.

Ezzati, T. M., Massey, J. T., Waksberg, J., Chu, A., and Maurer, K. R. (1992). Sample design: Third national health and nutrition examination survey. *Vital Health Stat.*, **2** (113), 1–35.

Hartley, H. O., Rao, J. N. K., and Kiefer, G. (1969). Variance estimation with one unit per stratum. *J. Am. Stat. Assoc.*, **64**(327), 841–851.

Landsman, V., and Graubard, B. I. (2013). Efficient analysis of case-control studies with sample weights. *Stat. Med.*, **32**(2), 347–360.

LaVange, L. M., Koch, G. G., and Schwartz, T. A. (2001). Applying sample survey methods to clinical trials data. *Stat. Med.*, **20**(17–18), 2609–2625.

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

Li, Y., Graubard, B., and DiGaetano, R. (2010). Weighting methods for population-based case-control study. *J. Royal Stat. Soc. C.*, **60**, 165–185.

Lumley, T. (2020). Package ‘survey’. http://cran.r-project.org/web/packages/survey/survey.pdf

Massey, J. T., Moore, T. F., Parsons, V. L., and Tadros, W. (1989). Design and estimation for the National Health Interview Survey, 1985-94. National Center for Health Statistics. *Vital Health Stat.*, **2** (110).

National Center for Health Statistics (2013). National Death Index user’s guide. Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/nchs/data/ndi/ndi_users_guide.pdf

Rivers, D. (2007). Sampling for web surveys. In *Joint Statistical Meetings*.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.

Rubin, D. B., and Thomas, N. (1992). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*, **79**(4), 797–809.

Scott, A., and Wild, C. (2001). Case-control studies with complex sampling. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, **50**(3), 389–401.

Scott, A. J., and Wild, C. J. (1986). Fitting logistic models under case-control or choice based sampling. *J. R. Stat. Soc. B. Methodol.*, **48**(2), 170–182.

Valliant, R., and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociol. Method Res.*, **40**(1), 105–137.

Wang, L., Graubard, B. I., Katki, H. A., and Li, Y. (2020). Improving external validity of epidemiologic cohort analyses: A kernel weighting approach. *J. Royal Stat. Soc. Ser. A.*, **183**(3), 1293–1311.

Wang, L., Valliant, R. & Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. arXiv preprint arXiv:2007.02476.

[Received August 2020; accepted August 2021]