ABSTRACT

Title of Dissertation:	Improving External Validity of Epidemiologic Analyses by Incorporating Data from Population-Based Surveys
	Lingxiao Wang, Doctor of Philosophy, 2020
Dissertation directed by:	Professor, Yan Li The Joint Program in Survey Methodology

Many epidemiologic studies forgo probability sampling and turn to volunteer-based samples because of cost, confidentiality, response burden, and invasiveness of biological samples. However, the volunteers may not represent the underlying target population mainly due to self-selection bias. Therefore, standard epidemiologic analyses may not be generalizable to the target population, which is called lack of "external validity." In survey research, propensity score (PS)-based approaches have been developed to improve representativeness of the nonprobability samples by using population-based surveys as references. These approaches create a set of "pseudo-weights" to weight the nonprobability sample up to the target population. There are two main types of PS-based approaches: (1) PS-based weighting methods using PSs to estimate participation rates of the nonprobability sample; for example, the inverse of PS weighting (IPSW); (2) PS-based matching methods using PSs to measure similarity between the units in the nonprobability sample and the reference survey sample, such as PS adjustment by subclassification (PSAS). Although the PS-based weighting methods reduce the bias, they are sensitive to propensity model misspecification and can be inefficient. The PS-based matching methods are more robust to the propensity model misspecification and can avoid extreme weights. However, matching methods such as PSAS are less effective at bias reduction. This dissertation proposes a novel PS-based matching method, named the kernel weighting (KW) approach, to improve the external validity of epidemiologic analyses that gain a better bias–variance tradeoff. A unifying framework is established for PS-based methods to provide three advances. First, the KW method is proved to provide consistent estimates, yet generally has smaller mean-square error than the IPSW. Second, the framework reveals a fundamental strong exchangeability assumption (SEA) underlying the existing PS-based matching methods that has previously been unknown. The SEA is relaxed to a weak exchangeability assumption that is more realistic for data analysis. Third, survey weights are scaled in propensity estimation to reduce the variance of the estimated PS and improve efficiency of all PS-based methods under the framework. The performance of the proposed PS-based methods is evaluated for estimating prevalence of diseases and associations between risk factors and disease in the finite population.

IMPROVING EXTERNAL VALIDITY OF EPIDEMIOLOGIC ANALYSES BY INCORPORATING DATA FROM POPULATION-BASED SURVEYS

by

Lingxiao Wang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:

Professor Yan Li, Chair
Dr. Hormuzd A. Katki
Dr. Barry I. Graubard
Professor Sunghee Lee
Professor Xin He, Dean’s Representative

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Significance

Large-scale cohort studies are the gold standard of design in epidemiology. However, establishing large-scale cohort studies has become more difficult in the United States in recent years because of increasing costs and declining response rates (Morton et al., 2006; Nohr et al., 2006), often due to concerns about confidentiality, volunteer burden, and invasiveness of biological samples.

To optimize resources, current epidemiologic cohorts are beginning to assemble samples within integrated health care systems that have electronic health records and a large pre-existing base of volunteers to recruit, such as the UK Biobank in the National Health Service (Collins, 2012). Unfortunately, volunteer-based cohorts generally have difficulty representing the target population. Many cohorts are well known to have "healthy volunteer effects" (Pinsky et al., 2007), usually resulting in lower disease incidence and mortality in the volunteers than in the general population. For example, the all-cause mortality rate in the UK Biobank was only half that of the UK population (Fry et al., 2017), and it is not representative of the UK population with regard to many sociodemographic, physical, lifestyle, and health-related characteristics. In another example, Katki et al. (2016) found that the lung cancer death risk calculated from the National Lung Screening Trial (NLST) cohort is seriously underestimated when compared to estimates based on nationally representative survey data.

Representative samples are important for generalizing statistical results from cohorts to the underlying target population ("external validity") in many circumstances. Studies of disease surveillance strongly require population representativeness to ensure that trends in disease incidence and mortality observed in the study are actually occurring in the target population. Studies of "translational epidemiology," which attempt to project the impact of epidemiologic findings on population health, require population representativeness.

There are two difficulties in improving representativeness of epidemiologic cohorts for external validity. First, the volunteer participants cannot represent the finite target population due to self-selection bias, low response rate, and coverage issues. As a result, the point estimates obtained directly from the cohorts will be biased from the true values in the population. Since participants volunteer to enroll the study, the selection probabilities are unknown, which makes them hard to measure for reducing bias. Second, many national cohort studies usually assemble samples in multiple study centers over the whole country. Examples include The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial and the National Institutes of Health (NIH)–American Association of Retired Persons (NIH-AARP) Diet and Health Study. Correlation may be induced by the homogeneity of the participants from the same study center. The finite population variance can be underestimated without considering correlation of outcomes within study centers. Therefore, the Type I error of significance tests can be inflated and coverage probability of confidence intervals may be too low even if the point estimates are approximately unbiased.

1.2 Research Gaps, Goals, and Innovations

We develop a methodology to improve external validity of non-representative epidemiologic studies using a population representative survey sample as a reference. Our goal is not to eliminate 100% of bias due to non-representativeness (which is probably unrealistic in practice) but rather to gain a better bias–variance tradeoff that improves external validity of the epidemiologic cohorts.

Research Gaps

Population-Based Health Surveys (PBHSs) have been conducted for many years, collecting broad types of data, comprising demographic, socioeconomic, dietary characteristics, health-related data, and examination measurements. Examples of PBHSs include the U.S. National Health and Nutrition Examination Survey (NHANES) and the U.S. National Health Interview Survey (NHIS) both conducted by the U.S. National Center for Health Statistics (NCHS). Some of PBHSs also collect DNA samples and provide genetic information (NHANES 1998-2002). These PBHSs employ stratified multistage cluster sample designs to select samples that represent the finite target population. The resulting samples are less susceptible to selection bias and coverage issues that can occur in volunteer-based cohorts assembled in epidemiologic studies.

It is worth noting the main disadvantages of PBHSs. First, the cost required to ensure population representativeness is high. For example, it cost NCHS around $100 million to conduct NHANES III yet sampled only 16,397 individuals from 1988 to 1994 (NCHS, Korn & Graubard, 1999). Second, the sample sizes of PBHSs are much too small to study even common chronic diseases like cancers. Longitudinal PBHSs are more expensive due to the expense required to follow-up participants and have smaller sample sizes because of panel attrition. Third, most PBHSs do not collect biospecimens, which is critical to modern epidemiologic research. Although PBHSs may not be the best sources for epidemiologic studies of novel exposure-disease associations, their population representativeness can be useful to improve external validity of epidemiologic cohorts.

Attaining population representativeness requires attempting to recruit a probability-based sample. However, formal probability sampling as done in surveys substantially increases the cost of assembling epidemiologic cohorts (LaVange et al., 2001; Duncan, G. J., 2008; Michael & O'Muircheartaigh, 2008). There are debates between epidemiologists and statisticians about the value of probability sampling (Little, 2010; Keiding & Louis, 2016; Ebrahim & Smith, 2013).

Because probability sampling is considered unrealistic for epidemiologic studies, we advocate that statistical research should focus on improving population representativeness of cohorts for external validity. However, to date, there has been little attention paid by biostatisticians to this issue. Powers et al. (2017) proposed a method to evaluate disease prediction models using a non-representative cohort. Keiding & Louis (2016) discussed problems with self-selected entry to epidemiological studies and surveys when making population level inference. Stuart et al. (2011) generalized results from volunteer randomized trials to populations using a PS-based weighting method (discussed in Chapter 2). There is still a lack of general-purpose methods for improving external validity of epidemiologic analyses.

The issues faced by epidemiologists are analogous to those faced in contemporary survey research. As nonprobability samples, such as web panels and big data from social media, become increasingly popular due to their cost- and time-efficiency (Baker et al., 2013; Kennedy et al., 2016), there are growing concerns about lack of population representativeness in survey research. Two major types of PS-based methods have been studied to improve the representativeness of nonprobability samples using probability-based samples as external references. The first type is PS-based weighting methods that use (functions of) PSs to estimate the participation rates of the nonprobability samples (Valliant, & Dever, 2011; Elliott, 2013; Elliott & Valliant, 2016; and Chen et al, 2019). The second type is the PS-based matching methods that use PSs to measure the similarity between the nonprobability and probability survey sample units (Lee & Valliant, 2009; Rivers, 2007). The goal of these methods is to create "pseudo-weights" for the nonprobability samples to improve the representativeness.

Both PS-based weighting and PS-based matching methods have limitations. The PS-based weighting methods can potentially fully correct bias under appropriate propensity models, but they are sensitive to the propensity model specification (Lee et al., 2010). Moreover, PS-based weighting methods are likely to produce extreme weights due to the estimates of PSs close to 0. The variances of the pseudo-weighted Horvitz–Thompson (HT) estimators (Horvitz and Thompson, 1952) can be inappropriately large if extreme weights are related more to the PS estimation procedure than to the true underlying sample selection (Stuart 2010). Weight trimming, which sets weights above some maximum to that maximum, has been proposed as one solution to reducing variance (Potter, 1993). However, the effect of weight trimming on increasing bias or reducing variance is unclear (Lee et al., 2010; Potter and Zheng, 2015), and there is relatively little guidance regarding the trimming level.

Compared to PS-based weighting methods, PS-based matching methods do not require the propensity model to accurately estimate nonprobability sample participation rates. Therefore, matching methods are less sensitive to model misspecification. The PSAS method, as a commonly used matching method (Lee & Valliant, 2009; Valliant & Dever, 2011; Brick, 2015), classifies the combined nonprobability and probability survey sample by quintiles of estimated PSs, and evenly distributes the survey sample weights to the nonprobability units within each subclass. It avoids extreme weights (Rubin 2001), and therefore yields less variable estimates. However, PSAS is less effective at bias reduction (Valliant & Dever, 2011) because of the key assumption that cohort units represent equal number of population units within subclasses. In addition, the measure of similarity of PSs is ad hoc with limited guidance and justification for forming the subclasses.

To gain efficiency, the existing PS-based matching methods fit the propensity model to the combined (cohort vs. unweighted survey) sample when estimating PS (Lee, 2006; Lee & Valliant, 2009; Valliant & Dever, 2011; Brick, 2015; Rivers, 2017). However, as found in this dissertation, they require a critical Strong Exchangeability Assumption (SEA) stating that the expectation of the outcome variable given the PS is the same in the cohort, survey, and the finite population. Without SEA, the resulting estimates of finite population means can be biased even under the correct propensity model fitted to the unweighted sample. Furthermore, there is no general approach for validating the propensity model to estimate the PSs for the matching methods.

Finally, more attention needs to be paid to the variance estimation of the PS-based pseudo-weighted estimates from the cohorts. There are three sources of variability that should be considered in variance estimation: (1) randomness due to estimating PS; (2)

potential clustering effects within the cohort; and (3) differential pseudo-weights of the cohort. The first two sources are not well considered in existing literature. In contrast with classical survey sampling, in which the sample weights are fixed by the sample design, the pseudo-weights for the cohort are functions of the PS estimated from the propensity model. Ignoring the randomness of the estimated PS can lead to severe underestimation of the variance, especially for the PS-based weighting methods, which directly estimate the participation rates from the PS. Many epidemiologic cohort studies recruit volunteers at multiple study centers clustered in the geographical areas where the target population resides. The resulting samples may have geographical effects (clustering and correlation of observations) due to differences in the distribution of variables among study centers, which influences the variance estimation.

Specific Goals

The specific goals of this dissertation that address the concerns and research gaps mentioned above are stated below.

Specific Goal 1: Develop a new PS-based matching method that provides a set of pseudo-weights for the volunteer-based cohort to improve the bias–variance tradeoff in target population disease prevalence estimation.

Specific Goal 2: Establish a unifying framework for both PS-based weighting and matching methods to relax the SEA to a more realistic Weak Exchangeability Assumption.

Specific Goal 3: Provide appropriate variance estimation for the PS-based pseudo-weighted estimates of population means/prevalences that considers all sources of variability.

Specific Goal 4: Apply the new methods to finite population relative risks estimation from the epidemiologic cohorts.

Innovations

A novel KW approach is proposed to improve external validity of epidemiologic cohort analyses. The KW approach first estimates PS from the propensity model fitted to the combined (cohort vs. unweighted survey) sample. Then, the survey sample weights are fractionally distributed to the cohort according to their similarities measured by difference in kernel-smoothed PSs. The KW approach, as a matching method, is less likely to produce extreme weights than the IPSW method. Meanwhile, the KW method is be more efficient at bias reduction than the PSAS approach because it relaxes the assumption of equal representativeness of cohort units within subclasses by PSAS. We prove that the KW estimators of finite population means/prevalences are consistent under SEA and other standard conditions.

Next, we establish an innovative unifying framework for both PS-based weighting and matching methods. For the matching methods, we relax the SEA to a weak exchangeability assumption (WEA) by estimating PSs from the propensity model fitted to the combined (cohort vs. weighted survey) sample. We prove that the enhanced KW method, under the WEA, provides consistent estimators of finite population means, which is usually unachievable for other matching methods such as the PSAS method. Nevertheless, fitting the propensity model to the weighted sample, when compared to the unweighted sample, can greatly increase the variance of the PS-based pseudo-weighted estimators due to the high variability of the estimated PSs. To improve efficiency, we propose scaling the survey weights (i.e., dividing the survey weights by their mean) in the propensity model. This simple scaling recovers much of the statistical efficiency. In our data example, the enhanced KW method, under WEA and using scaling, generally provides the smallest mean-square error, while protecting against large bias that can be incurred by methods that rely on the SEA.

Taylor linearization (TL) and Jackknife replication (JK) methods are developed for estimating finite population variance of the PS-based pseudo-weighted estimates of the finite population means under the framework. Both the TL and JK methods take the variability of estimating PS, unequal pseudo-weights, and the cluster effects of the survey and/or cohort into account.

1.3 Overview of the Chapters

The remainder of this dissertation comprises the following five chapters. Chapter 2 presents the background and rationale for the study through a comprehensive literature review. It starts by discussing epidemiologic cohorts and the issue of representativeness. PBHSs are then introduced, including sampling designs and sample weights. After that, the PS-based approaches are described as bias reduction methods in epidemiology and in survey research. The mathematical notations, advantages, and disadvantages for each method, as well as research gaps, are included. A kernel smoothing technique is then introduced to improve the existing weighting methods. At the end, motivation and recent research of using scaled weights to improve efficiency of logistic regression analyses are reviewed. Chapter 3 proposes the KW method under the SEA that improves the cohort estimates of population means/prevalences, with the proof of consistency and JK variance estimation provided. Chapter 4 first shows the necessity and difficulties of SEA for PS-based matching methods with illustrative examples. The unifying framework is then established for both PS-based weighting and matching methods under which SEA is relaxed by WEA for the matching methods. The KW method is enhanced, as an example of the PS-matching method, under WEA. Chapter 5 investigates the bias in naïve cohort estimates of relative risks for developing diseases in the finite population under different scenarios of implicit cohort participation mechanisms. The performance of the proposed PS-based methods are compared in reducing bias in estimated relative risks. Monte Carlo simulations and real data examples are provided in Chapters 3-5 to evaluate the performance of the proposed PS-based methods. Discussion and future work are given in Chapter 6.

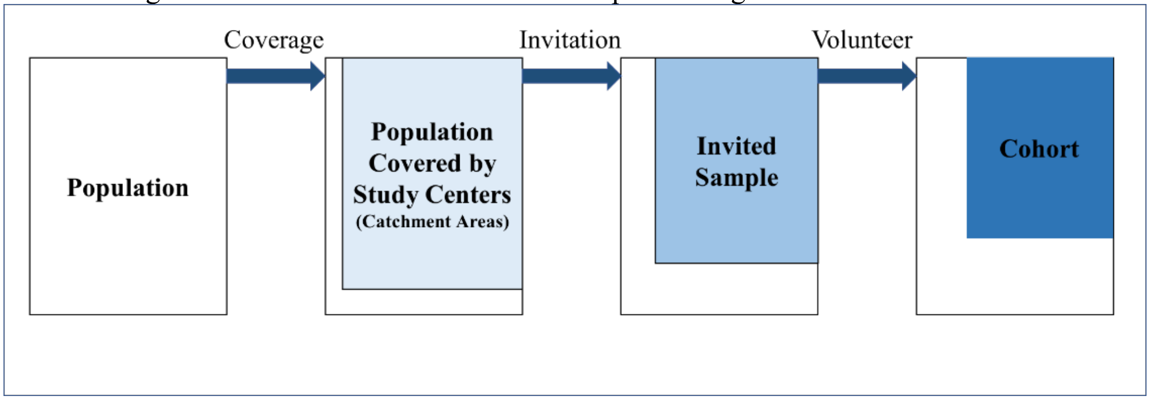
Chapter 2 Literature Review

2.1 Epidemiologic Cohort Studies

Large-scale long-term epidemiological cohorts are an ideal epidemiologic study design. However, they are becoming more difficult to assemble because of increasing costs and declining response rates (Morton et al., 2006; Nohr et al., 2006; Galea & Tracy, 2007), often due to concerns about confidentiality, volunteer burden, and invasiveness of biological samples. To optimize resources, new epidemiological cohorts are being assembled within integrated health care systems that have electronic health records and a large pre-existing base of volunteers to recruit, such as the UK Biobank in the National Health Service (Collins, 2012). Unfortunately, these volunteer-based cohorts generally inadequately represent the target population where the cohorts are obtained due to self-selection bias, low recruitment/response rates, and coverage issues.

For example, the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial, conducted by the National Institutes of Health (NIH), is a large randomized, controlled trial of screening tests for multiple cancers. Upon completion of the trial, participants continued to be followed up as a cohort for the purpose of studying potential risk factors for many types of cancer. The intended target population of this study is all adults aged from 55 to 74 years old in the United States, but as it was a randomized clinical trial, population representativeness and external validity could not be a priority. The age-eligible individuals were first sampled and invited by mail to be in the study. Interested participants were then contacted by one of 10 study centers at their convenience and were assessed for eligibility. The low population coverage of the mailing list used to send invitations can cause severe coverage bias. Also, as volunteers self-selected to participate in the study, selection bias may arise if the volunteered participants were systematically different from the non-participants. The detailed study design of the PLCO trial is described in Prorok et al., (2000). Figure 2.1 below depicts the overall protocol of the cohort recruitment. Due to the coverage issue and self-selection bias, it is difficult to assess if the cohort truly represents the target population.

Figure 2.1 Protocol of Volunteer-Based Epidemiological Cohort Recruitment



Under the context of epidemiology, representative samples are required in many circumstances so that statistical results from the samples can be generalized to the underlying population (external validation). Studies of disease surveillance strongly require population representativeness to ensure that trends in disease incidence and mortality observed in the study are actually occurring in the target population. Studies of "translational epidemiology," which attempt to project the impact of epidemiologic findings on population health, require population representativeness. Lack of population representativeness can have serious consequences for epidemiological analyses that require representativeness. For instance, "healthy volunteer effects" (Pinsky et al., 2007) are not unusual in analyses of these types of cohorts. Often, the disease incidence and mortality rates are much lower in the volunteers than in the general population. For example, the all-cause mortality rate in the UK Biobank was only half that of the UK population (Fry et al., 2017). The distribution of the Biobank sample is quite different from the UK population with regard to many sociodemographic, physical, lifestyle, and health-related characteristics. Katki et al. (2016) found that lung cancer death risk calculated from the National Lung Screening Trial (NLST) is seriously underestimated compared to nationally representative survey data.

However, there is limited statistical research investigating the representativeness of cohorts in epidemiology. Cochran & Chambers (1965) addressed the problem of inferences from sample to population and found that the simple random sample assumption may not be realistic. First, the sampling frame from which the sample is drawn can be quite different from the target population that one wanted to make inference from. Second, the issue of nonresponse can distort the distribution of the sample. As a result, the sample cannot represent the target population and the statistical results cannot be generalized to the target population. Later, Keiding & Louis (2016) discussed issues of volunteer-based samples that are typically used in epidemiologic research, including coverage, response, and selection bias. Improving the representativeness of cohorts is still an open area in epidemiology.

2.2 Population-Based Health Surveys

2.2.1 Overview

Population-Based Health Surveys (PBHS) have been conducted for many years, collecting broad types of data, ranging from demographic to socioeconomic to health-related information such as the occurrence of healthy and unhealthy behaviors, exposures to potential risk factors, dietary intake, physiologic measures of the population, and costs and utilization of health care services. Examples of PBHSs are the National Health and Nutrition Examination Survey (NHANES) and National Health Interview Survey (NHIS) both conducted by the National Center for Health Statistics (NCHS). The large sample sizes enable the study of relatively small but important associations between variables, relatively rare events, and subpopulations of interest. These surveys employ stratified multistage cluster sample designs to select samples that represent the target population. With appropriate statistical methods, the results of analyses can be made for the target population. This section briefly introduces the commonly used sample designs for population-based surveys, sample weights, and variance estimation of weighted estimates accounting for complex survey sampling designs.

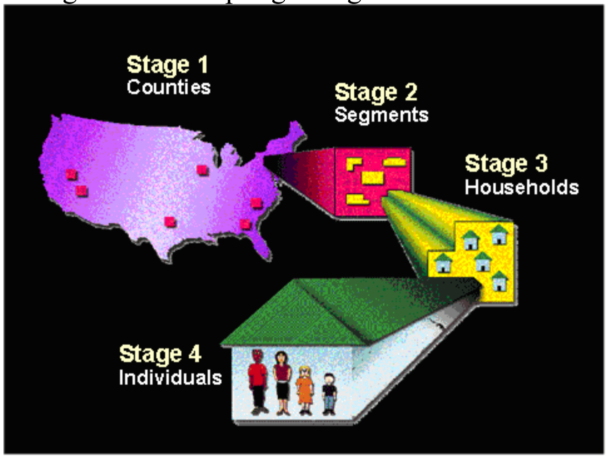
2.2.2 Sample Designs

2.2.2.1 *Multistage Sampling*

Multistage sampling designs are commonly used for PBHSs. The target population is firstly divided into strata defined by geographical areas (e.g., states). Within each stratum, a random sample of clusters (primary sampling units, or PSUs) of individuals is selected at the first stage. Then smaller clusters are successively subsampled within selected PSUs, and finally individuals are successively subsampled. Take the NHANES III as an example. The PSUs of counties (a group of contiguous counties or standard metropolitan area) exhausting the land area of the continental United States are formed and then grouped into mutually exclusive strata. The first stage selected PSUs within each stratum, followed by the other two stages in which segments and households are sampled within selected PSUs and segments sequentially. At the final stage, individuals are randomly selected within each household in the sample

Figure 2.2). At each stage of sampling, units may be sampled with different probabilities (Section 2.2.2.2) based on their characteristics.

Figure 2.2. Sampling Design of NHANES III



2.2.2.2 *Differential Sampling Rates*

Differential sampling rates are usually applied in the complex sampling designs, mainly aiming to decrease the variability of the estimates. They can be used as a single stage sampling, or for any stage of the multistage sampling. Take probability proportional to size (PPS) sampling as an example. For PPS sampling, some known information is selected to be a continuous measure of "size" variable (Z), which the inclusion probabilities are taken to be proportional to. For example, a survey of hospitals may sample hospitals with probability proportional to the bed size. With the assumption that the sampling is without replacement, the inclusion probability of the i-th unit in the population is:

$$\pi_i = \frac{nZ_i}{\sum_{j=1}^{N} Z_j},$$

where n and N are the sample size and population size respectively. PPS sampling may decrease the variance of mean and total estimates if the size variable, Z, is correlated with the variable of interest. Under PPS sampling, the joint inclusion probabilities are needed for variance estimation purpose. However, they are usually hard to obtain. The sample can be treated as if it had been drawn with replacement when the sampling fraction n/N is small enough to be ignored (Cochran, 1997).

With a multistage sampling design, differential sampling rates can be incorporated for any stage. The final sampling rate (inclusion probability) is the production of the sampling rates at each stage.

2.2.3 Sample Weights

The sample weight of a sampled person is the (estimated) number of individuals that the person represents in the population. Korn & Graubard (1999) defined three components of the sample weight as differential sampling rates (base weights), adjustments for nonresponse, and adjustments for inadequate frame coverage. Ignoring the sample weights could lead to invalid inference for the target population if the weights are informative (Fuller, 1999).

- Base Weights

Base weight is the component of the sample weight that accounts for the differential sampling rates. It is defined as the inverse of the inclusion probability of the individual in the sample; that is, $d_i = 1/\pi_i$ for the sample unit i, where π_i is the inclusion probability defined in Section 2.2.2.2.

- Nonresponse and Noncoverage Adjustment

If the unobserved distribution of answers to a question from individuals who do not respond is different from that of those who do respond, then estimates based on the respondents' data alone will be biased (Korn and Graubard, 1999). The nonresponse adjustment can be formed by modeling the probability of responding as a function (e.g., logistic regression) of variables available on all sampled units. The nonresponse adjustment factor to the sample weight for respondent i, denoted by f_i^{NR}, is taken as the inverse of the estimated probability of response (Iannacchione et al., 1991).

The sample frame, in which the individuals are randomly drawn, may not cover the whole target population. There is no information for the uncovered people in the target population who should have been in the sample. If the characteristics of these uncovered people are different from people covered in the sample frame, then the estimates based on covered people only may lead to biased estimates. The poststratification adjustment is usually applied to reduce the coverage error (Holt & Smith, 1979; Kish, 1965; Kalton & Flores-Cervantes, 2003). The adjustment cells for the sampled individuals are formed by demographic variables such as age, sex, and race/ethnicity. Within each cell, the adjustment factor for individual i, denoted by f_i^{NC}, is given by the known census total divided by the sum of the sample weights of the sampled individuals, assuming that the census figures are more accurate than the survey coverage.

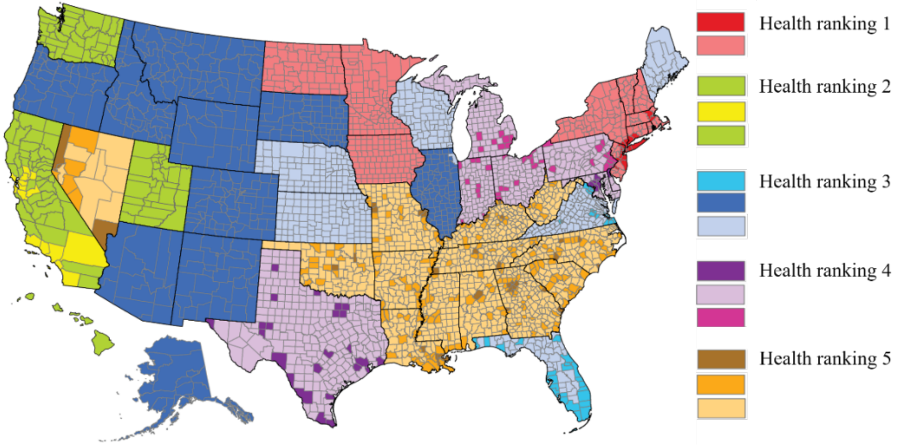
The final sample weight for sample unit i is calculated by the production of the three components, $d_i \cdot f_i^{NR} \cdot f_i^{NC}$.

2.2.4 Consistent Estimation of Finite Population Quantities from Survey Samples

Sample weights are usually incorporated in the analyses to ensure consistency (or approximate unbiasedness) of the population parameter estimation. For sample estimation of population means and proportions, sample weights need to be considered if the variable of interest is correlated with the sample weights. The unweighted sample mean is biased if the variable of interest distributes differently in sample and the population. The weighted sample mean converges to the population mean under suitable conditions, as the sample size increases (Korn & Graubard, 1999).

In regression analyses, sample weights also need to be incorporated. To improve the efficiency of parameter estimates, the sampling design often depends on some key study variables of interest even after conditioning on other covariates. Such a design is referred to as an informative sampling design (Fuller, 1999). For example, in 2011–2014 NHANES, the major strata were formed based on urban–rural measures and health ranking calculated from state-level, health-related variables (Figure 2.3), including death rate, adult high blood pressure, overweight or obese, smoking status, etc. (Johnson et al., 2014).

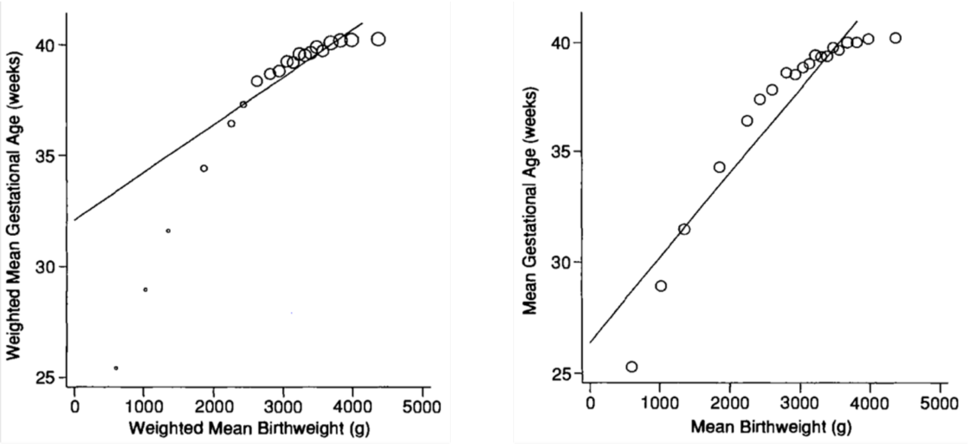
Figure 2.3 State Groupings in NHANES (2011-2014)



The sampling design can be informative when we analyze the association between these health-related variables and some risk factors. The relationship between the covariates and the study variable is distorted in the sample if the design is informative. Hence, ignoring sample weights will lead to biased estimates of association.

If the study variable is independent from design variables conditional on the covariates in the regression analysis, then the sampling design is non-informative. When the relationships between the study variable and the covariates are correctly modelled, the unweighted sample estimates of association are unbiased. However, the weighted and unweighted estimates of regression coefficient may differ when the regression model is misspecified. Korn and Graubard (1999) addressed this problem using an example of regressing gestational age on birthweight in the 1998 National Maternal and Infant Health Survey (NMIHS). The NMIHS is a stratified random sample of women who had a live birth, fetal death, or infant death in the United States. In 1998, the strata were constructed using states, the mother's race, and baby's birth weight. The sampling fractions varied so that Black babies and low-birthweight babies were oversampled (Sanderson et al. 1998).

Figure 2.4 Plots of Mean Gestational Age versus Mean Birthweights with and without Weights in 1998 NMIHS



Source: Korn & Graubard (1999)

This design is non-informative for the analysis of gestational age regressed on baby's birth weight. Figure 2.4 below shows the scatter plot and the regression line with (left) and without (right) sample weights considered. The size of bubbles in the plot on the left indicates the magnitude of sample weights. The regression model is misspecified as a linear regression. Since the sample and population distribution of the independent variable (birthweight) differ, the regressions are attempting to fit a straight line to different parts of a curvilinear relationship. The weighted regression has the advantage that it is at least estimating a population quantity. Although the unweighted slope fits the unweighted sampled data better, the parameter being estimated by the unweighted slope will change depending upon the sample design (Korn & Graubard, 1999).

Although PBHSs are designed to produce valid finite population inferences, they cannot provide ideal data sources for epidemiologic analyses. First, the cost taken to ensure population representativeness is high. For example, For example, it cost NCHS around $100 million to conduct NHANES III yet sampled only 16,397 individuals from 1988 to 1994 (NCHS, Korn & Graubard, 1999). Second, the sample sizes of PBHSs are much too small to study even common chronic diseases such as cancers. Longitudinal PBHSs will be more expensive due to the expense on follow-up and have smaller sample sizes because of panel attrition. Third, PBHSs collect little information that is desired in modern epidemiology (e.g., biospecimens), limiting its usefulness for epidemiologic research. Although PBHSs may not be the best sources for epidemiologic studies of novel exposure–disease associations, they can be considered as good references to improve external validity of epidemiologic cohorts. The next section introduces propensity score (PS) methods that help improve representativeness of the nonprobability sample by using a probability-based survey sample as an external reference.

2.3 PS Adjustment in Epidemiology Studies and Survey Research

PS-based methods were initially developed to match the covariates distributions in the controls with that in the cases for observational studies in epidemiology so that the unbiased estimates of treatment effects can be obtained by comparing the cases and the adjusted controls (Rosenbaum & Rubin, 1983). Analogous to matching the covariate distributions between cases and controls in observation studies, improving representativeness of the nonprobability samples needs the covariate distributions in the nonprobability samples to be close to that in the target finite population. In survey research, similar PS-based methods have been developed by treating the nonprobability sample as the cases, and treating the reference probability survey sample as the controls (Lee, 2004). A set of "pseudo-weights" are calculated for the nonprobability sample so that the distribution of covariates in the pseudo-weighted nonprobability sample are similar to that in the underlying target finite population. This section critically reviews the PS-based methods developed for estimating treatment effects in observation studies, and for improving representativeness of the nonprobability samples in survey research.

2.3.1 PS-Based Adjustments for Treatment Effect Estimation in Observational Studies

2.3.1.1 *Introduction*

Group comparison is a common method for presenting scientific research results in epidemiology. A fundamental problem of the group comparison, though, especially for observational studies, is that the two groups are not randomly selected. The resulting comparison may have confounding bias caused by the different characteristics of the two groups other than the treatment assignment. For example, in an observational study of heart transplant and mortality, patients receiving the transplant were more likely to be older and have worse health conditions, which led to a greater death risk than those who did not receive the transplant. The different mortality rates in the two groups may not be related only to the heart transplant, but also confounders such as age and health conditions. The comparison of mortality will not be convincing unless the confounders are controlled.

2.3.1.2 *Estimating Treatment Effects*

The estimation of causal effects of treatments, firstly formalized in Rubin (1974), is inherently a comparison of potential outcomes. Suppose in a population U, we consider the case of one treatment and one control group, with variables of outcome y_1 and y_0, respectively (1 for presence and 0 for absence of the disease). In principle, each unit $i \in U$ has both responses $y_{1,i}$ and $y_{0,i}$ that would have resulted if it had treatment or not, respectively, under the framework of casual inference. The causal effect of treatment is:

$$\tau = \tau_1 - \tau_0 = E(y_1) - E(y_0), \qquad (2.3.1)$$

where $E(\cdot)$ is the expectation with respect to the distribution of y in U. Suppose in a random sample of the population, s, with size n. Denote $\hat{\tau}_1 = \frac{1}{n}\sum_{i \in s} y_{1,i}$ and $\hat{\tau}_0 = \frac{1}{n}\sum_{i \in s} y_{0,i}$ as the sample estimates of τ_1 and τ_0, respectively. By the law of large numbers, it can be shown that $E(\hat{\tau}_1) = \tau_1$ and $E(\hat{\tau}_0) = \tau_0$.

However, the “fundamental problem of causal inference” is that one can only observe one of the two outcomes, either $y_{1,i}$ or $y_{0,i}$ for individual i (Holland et al., 1985). Let z_i be the binary treatment assignment for unit $i \in s$, (1 for being selected in the treatment group s_1 of size n_1; 0 for being selected in the control group s_0 of size n_0). We can only observe $y_{1,i}$ for $i \in s_1$ and $y_{0,i}$ for $i \in s_0$. The treatment effect τ can be estimated by the average treatment effect from observed data:

$$\hat{\tau} = (\hat{\tau}_1 | z = 1) - (\hat{\tau}_0 | z = 0) = \frac{1}{n_1}\sum_{i \in s_1} y_{1,i} - \frac{1}{n_0}\sum_{i \in s_0} y_{0,i}$$

Under the randomization assumption, the outcome is independent from the group assignment; that is,

$$(y_1, y_0) \perp\!\!\!\perp z. \tag{2.3.2}$$

Lee (2004) proved that $E(\hat{\tau}) = \tau$ as

$$E[E\{(\hat{\tau}_1|z = 1) - (\hat{\tau}_0|z = 0)\}] = E(\hat{\tau}_1 - \hat{\tau}_0) = \tau. \tag{2.3.3}$$

<u>2.3.1.3</u> Balancing Scores and Propensity Score

As it was discussed earlier, in observational studies, Equation (2.3.3), may not hold because the treated individuals can be systematically different from the individuals in the control group. This may confound the outcome and violate assumption (2.3.2). A Balancing score, $b(\boldsymbol{x})$ defined in (2.3.4), should be used to balance the covariates $\boldsymbol{x}$ in the treatment $(z = 1)$ and control group $(z = 0)$ such that the conditional distribution of $\boldsymbol{x}$ given $b(\boldsymbol{x})$ is the same for the two groups (Dawid, 1979; Rosenbaum & Rubin, 1983).

$$\boldsymbol{x} \perp\!\!\!\perp z|b(\boldsymbol{x}). \tag{2.3.4}$$

Rosenbaum & Rubin (1983) proposed to use the PS, which takes the coarsest form of the balancing score. A PS is defined as the probability of a unit being assigned to the treatment group $(z = 1)$ given a set of covariates $\boldsymbol{x}$, denoted as:

$$e(\boldsymbol{x}) = \Pr(z = 1|\boldsymbol{x}). \tag{2.3.5}$$

By assuming that the treatment assignment is strongly ignorable—that is, $(y_1, y_0) \perp\!\!\!\perp z|e(\boldsymbol{x})$—it gives:

$$E\{y_1|e(\boldsymbol{x}), z = 1\} - E\{y_0|e(\boldsymbol{x}), z = 0\} = E\{y_1 - y_0|e(\boldsymbol{x})\}. \tag{2.3.6}$$

Therefore,

$$\begin{aligned} & E_e E_{y|e}\{y_1|e(\boldsymbol{x}), z = 1\} - E_e E_{y|e}\{y_0|e(\boldsymbol{x}), z = 0\} \\ & = E_e E_{y|e}\{y_1 - y_0|e(\boldsymbol{x})\} \\ & = \tau_1 - \tau_0 = \tau. \end{aligned} \tag{2.3.7}$$

where the expectation E_e and $E_{y|e}$ are with respect to the distribution of $e(\boldsymbol{x})$ and the distribution of y given $e(\boldsymbol{x})$ respectively. That is, if $\boldsymbol{x}$ contains all confounders and the distribution of $e(\boldsymbol{x})$ can be correctly modeled, unbiased estimate of treatment effect can be achieved by certain PS-based adjustment.

There are five assumptions under which the PS-based adjustments can reduce the bias of treatment effect estimation in observational studies (Rosenbaum & Rubin, 1983; Rosenbaum, 1984a, 1984b).

(1) Strong ignorability: $(y_1, y_0) \perp\!\!\!\perp z|e(\boldsymbol{x})$: This indicates that the study outcomes (y_1, y_0) and the treatment group assignment are conditionally independent given the PS. Under this assumption, the observed average treatment effect is unbiased given the PS.

(2) No contamination among study units: A treatment assignment of one unit does not affect the assignment for any other units.

(3) Non-zero assignment probability of treatment or control group: All units have a positive probability to be assigned to the treatment or control group for any configuration of covariates $\boldsymbol{x}$.

(4) No missing confounders: The observed covariates $\boldsymbol{x}$ included in the propensity models can handle all confounding in the treatment assignment.

(5) Treatment assignment does not change the covariate values.

In principle, the true PSs in the population should be used as the balancing score so that an unbiased estimate of the treatment effect can be obtained. However, the true PSs are usually not available, and need to be estimated from propensity models fitted to the observed data. Rubin & Thomas (1992, 1996) showed that the using sample estimates of PSs are more effective at bias reduction than using the truth. Many parametric models can be used to estimate the PSs among which the logistic regression is the most commonly used (Stuart, 2010; Lee, 2004). The propensity model is fitted to the combined sample of the treatment and control group ($s = s_1 \cup s_0$) as:

$$\log\left\{\frac{e(\boldsymbol{x})}{1 - e(\boldsymbol{x})}\right\} = \boldsymbol{\beta}^T \boldsymbol{x}, \tag{2.3.8}$$

where $\boldsymbol{\beta}$ is a vector of coefficients to be estimated, and $\boldsymbol{x}$ is a vector of covariates. To obtain statistically reliable estimates, the distribution of the estimated PSs in the treatment and control groups have to be well overlapped (Rosenbaum & Rubin, 1983).

In terms of the covariate selection in Model (2.3.8), literature gives different suggestions. Many of these suggestions recommend using variables related to both the study outcome and the treatment assignment to satisfy the assumption of ignorability (Rosenbaum & Rubin, 1984; Rubin and Thomas, 1996; Heckman et al., 1998b). Shadish et al. (2008) showed that no substantial bias reduction was found if a small set of “predictors of convenience” such as demographics were used. Stuart (2010) summarized these articles and made suggestions on selecting covariates in the propensity model for balancing the bias-variance tradeoff in treatment effect estimation. Excluding important confounders may limit the bias reduction. Including variables unassociated with the outcome can increase the variance, but there is little cost to include variables unassociated with treatment assignment. Hence, Stuart (2010) suggested including as many variables related to treatment assignment and/or outcome when sample size permits. With a small sample size, priority should be given to variables related to the outcome, as there is a higher cost in efficiency loss when including variables unrelated to the outcome but highly related to treatment assignment (Brookhart et al., 2006). However, a potential disadvantage of giving priority to variables related to the outcome is that the propensity model can be selected subjectively for the desirable results of the analysis.

2.3.1.4 *PS-Based Methods for Treatment Effects Estimation*

There are three types of PS-based methods that balance the covariates distributions between treatment and control groups: pair matching, subclassification, and weighting. Both the pair matching and the subclassification methods use estimated PS $\hat{e}(\boldsymbol{x})$, to measure the similarity between individuals in the control group and the treatment group. The pair-matching method matches two individuals based on $\hat{e}(\boldsymbol{x})$—one from the treatment group, and the other from the control group. The treatment effect is estimated from the matched samples only. The subclassification method divides the combined sample of treatment and control group by $\hat{e}(\boldsymbol{x})$ into subclasses. Within each class, an identical weight is assigned to all individuals in the control group so that the weighted control group represents the same number of treated individuals within the class. Different from these two methods, the weighting method uses $\hat{e}(\boldsymbol{x})$ to estimate the inclusion probability of the treatment group versus the control group and uses a function of $\hat{e}(\boldsymbol{x})$ as the weights for controls and/or treated individuals. This section describes the three approaches in details with their advantages and disadvantages.

- PS-based Pair-Matching Method

The purpose of the pair-matching methods in observational studies is to select a sample of untreated individuals that have a close $\boldsymbol{x}$-distribution to the treatment group. Rosenbaum

& Rubin (1985) proposed to use differences in estimated PS as the distance measure. The distance between individuals i and j for matching is defined as:

$$D_{ij} = \left|\hat{e}_i - \hat{e}_j\right|$$

where $\hat{e}_i$ and $\hat{e}_j$ are the predicted PS for units i and j, respectively.

Pair matching by PS has shown to successfully reduce the bias for treatment effect estimation (Rosenbaum & Rubin, 1985). It helps to select a control group that has a similar $\boldsymbol{x}$-distribution with the treatment group **before** the study variables of interest are collected and to save the cost. It can be also applied when the outcome data are already available by selecting a subsample from the control group to match with the treatment group. However, one apparent drawback of the pair-matching method is that some control group members who are not selected to match with any treated individuals are discarded in the analysis and therefore lower the statistical power of the analyses due to the reduced sample size.

Note that $e(\boldsymbol{x})$ has a bounded support $(0, 1)$ and is skewed when the proportion of the controls is small or large in the trial. A tiny difference in $e(\boldsymbol{x})$ may result from large differences in covariates $\boldsymbol{x}$ when $e(\boldsymbol{x})$ is close to the boundary, which can bias the estimates of treatment effects. These boundary problems can be avoided by using the linear propensity score $q(\boldsymbol{x}) = \operatorname{logit} e(\boldsymbol{x})$ (Rubin and Thomas, 1992; Rubin, 2001).

- PS-based Subclassification

Cochran & Chambers (1965) was one of the first uses of subclassification with a single confounder x being used to form the subclasses. Suppose the treatment and the control groups are sampled from the treatment and the control populations. The union of two groups ($s = s_1 \cup s_0$) is divided into G sub-classes according to the distribution of x. The estimate of the mean difference is given by:

$$\Delta = \sum_{g=1}^{G} w_g \left(\bar{y}_{0,g} - \bar{y}_{1,g} \right),$$

where $\bar{y}_{0,g}$ and $\bar{y}_{1,g}$ are the means of outcome y in subclass g for the control and the treatment group, respectively, and w_g is the adjustment weight for subclass $g, g = 1, \cdots, G$. If the variance of y appears constant within sub-classes, then w_g could be taken as $n_{0,g} n_{1,g} / \left(n_{0,g} + n_{1,g} \right)$ by minimizing Δ using the usual least-squares principles.

Cochran (1968) further examined the subclassification method on a single continuous covariate of age in investigating the death rates among smoking groups (non-smokers, cigarettes only, and cigars and/or pipe) in three countries (Canada, U.K., and U.S.). The death rates were found to be much higher for the cigars and/or pipe group, due to the higher age. However, the adjusted death rates exhibited no elevation over those for non-smokers.

Rosenbaum & Rubin (1983) expanded Cochran's work by using estimated PS, $\hat{e}(\boldsymbol{x})$, instead of a single confounder to form the subclasses. This method is widely used in clinical trials (e.g., Lavori & Keller, 1988; Cook & Goldman, 1989; Stone et al., 1995; Rubin, 1997) because of the following advantages: (1) it is easier to operate than the pair matching; (2) the size of the control group is not required to be larger than that of the treatment group; (3) all individuals in the study can be used in the analysis, unlike the pair matching, which discards unmatched individuals; and (4) it is less likely to produce extreme weights than the inverse probability of treatment weighting (IPTW; see next bullet). However, it is unclear how many subclasses should be used. Five-class adjustment was recommended by Cochran (1968) and Rosenbaum and Rubin (1985b), which could remove at least 90% of the bias in the estimated treatment effect in their study. On the contrary, other literature (e.g., Lunceford & Davidian, 2004) suggested considering more subclasses (e.g., 10–20) when sample size permits. More work is needed to determine the optimal number of subclasses that ensures adequate bias reduction without causing variance inflation.

- PS-based Weighting Methods

The weighting method estimates the probability of group membership from (function of) the PS. The inverse of the estimated probabilities are used as the weights so that the $\boldsymbol{x}$-distriution in the weighted control group matches with that in the (weighted) treatment group. This weighting method is known as inverse probability of treatment weighting (IPTW; Czajka et al., 1992; Robins et al., 2000; Lunceford & Davidian, 2004).

There are two types of treatment effects: (1) "average treatment effect" (ATE), or $E(y_1 - y_0)$, which is the treatment effect on individuals in combination of the treatment and control groups (Imbens, 2004); and (2) "average effect of the treatment on the treated" (ATT), or $E(y_1 - y_0| \, z = 1)$, which is the effect for individuals in the treatment group. The IPTW method assigns different weights for estimating the two treatment effects. To estimate ATE, both treatment and control group members get weights:

$$w_i = \frac{z_i}{\hat{e}_i} + \frac{1 - z_i}{1 - \hat{e}_i}, i \in s_1 \cup s_0,$$

where $z_i = 1$ if $i \in s_1$, and $z_i = 0$ if $i \in s_0$ so that either of the group is weighted up to the combined sample. To estimate ATT, the control group is weighted up to the treatment group, with the weight:

$$w_i = z_i + (1 - z_i)\frac{\hat{e}_i}{1 - \hat{e}_i}, i \in s_1 \cup s_0.$$

A potential drawback of IPTW is that the variance can be inflated due to extreme large weights if some estimated PSs are close to 0. Especially when the propensity model is misspecified, the extreme weights can result from the propensity estimation, but not the true probabilities of treatment assignment. Weight trimming, setting weights above some maximum to that maximum, has been proposed as one solution to reducing variance (Potter, 1993). However, the effect of weight trimming on increasing bias or reducing variance is unclear (Lee et al., 2010; Potter and Zheng, 2015), and there is relatively little guidance regarding the trimming level.

2.3.2 PS-Based Methods for Improving Representativeness in Epidemiology and Related Areas

In addition to balancing the different distributions of cases and controls in observational studies, PS-based methods can also be used to improve the representativeness of the study samples. As discussed in Section 2.1, epidemiologic cohorts frequently lack population representativeness, which can have serious consequences for epidemiologic analyses. However, there is limited statistical research investigating the methods of improving external validity and representativeness of the cohorts, among which the PS-based approaches were considered.

Stuart et al. (2011) used a PS-based weighting approach to estimate the participation rates of a randomized trial (denoted by s) so that the estimates of treatment effects in target finite population (denoted by FP) can be obtained from the trial. In their paper, a logistic regression model was fitted to FP (size = N) to estimate the participation rate of the randomized trial for individual i in the finite population, given a set of covariates $\boldsymbol{x}$: $\log\frac{\pi_i}{1-\pi_i} = \boldsymbol{\beta}^T \boldsymbol{x}_i$, for $(i \in FP)$ where $\pi_i = P\{\delta_i = 1|\boldsymbol{x}_i\}$ is the probability of being included in the randomized trial for $i \in FP$, and δ_i is the indicator for being included in the trial (=1 if $i \in s$, 0 for $i \in FP - s$). The estimated participation rate for $i \in FP$ is denoted by $\hat{\pi}_i$.

Stuart et al. (2011) proposed to measure the representativeness of the trial participants to the target population using the difference in averaged predicted PSs between the trial participants and the nonparticipants in the finite population:

$$\Delta_p = \frac{1}{n}\sum_{i \in s} \hat{\pi}_i - \frac{1}{N-n}\sum_{i \in (FP-s)} \hat{\pi}_i$$

If Δ_p is large, which means s is not representative of FP, then the weight of $\frac{1}{\hat{\pi}_i}$ will be assigned to trial participant $i \in s$. Within the trial, the PS-based adjustment can be applied to match the $\boldsymbol{x}$-distribution in the control group to that in the treatment group. For example, individual i in the control group obtains a weight of $\frac{1}{1-\hat{e}(\boldsymbol{x}_i)}$, where $e(\boldsymbol{x}_i) = P(z_i = 1 \mid \boldsymbol{x}_i)$ is the propensity of being in the control group given being selected in the trial for individual i, and z_i is the binary indicator for control/treatment group membership (=0 for controls, =1 for treatments). The final weight for control group member i is given by $\frac{1}{\hat{\pi}_i} \cdot \frac{1}{1-\hat{e}(\boldsymbol{x}_i)}$.

This paper provided a way of obtaining external validity using a volunteer-based randomized trial. However, it is not common that the individual level covariates $\boldsymbol{x}$ are available in the entire population, required for estimating π_i, and therefore limits the application of the proposed method. Typically only a representative probability survey provides information on π_i. Furthermore, this paper only focused on a simple estimate of treatment effect. Performance of the proposed approach can be evaluated for more general analyses. In addition, the variance estimation was not studied.

2.3.3 PS-Based Methods for Improving Representativeness of Nonprobability Samples in Survey Research

2.3.3.1 Introduction

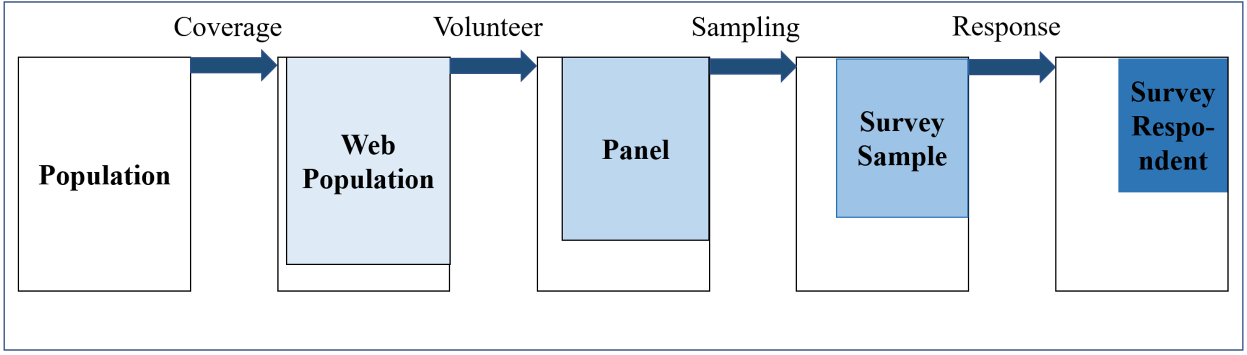
In survey research, the principal goal is to make reliable and accurate inferences to a broader target population. Hence, probability-based sampling designs have been chosen to generate population representative samples for most large-scale surveys (Frankel & Frankel, 1987). However, nonprobability sampling is still widely used in many areas such as polling. Like volunteer-based epidemiologic studies, the nonprobability survey samples are not randomly selected, and they cannot closely represent the target population. Survey researchers have developed several PS-based approaches to improve the representativeness of the nonprobability samples. This section introduces nonprobability samples in survey, existing weighting approaches, and how to apply them to volunteer-based epidemiologic studies.

2.3.3.2 Nonprobability Samples in Surveys

In recent decades, the combination of rapidly increasing costs, declining response rates, and rising concerns about coverage has raised expectations about the potential benefits of web surveys, especially as internet penetration has increased (Couper 2000). However, it is difficult for web surveys to recruit and sample respondents as traditional face-to-face or random digit dialing (RDD) telephone surveys due to unavailability of a sampling frame. Alternative approaches are developed relying on nonprobability methods, most notably opt-in panels composed of volunteers. Figure 2.5 below shows the protocol of volunteer-based panel web surveys.

The protocol of volunteer-based panel web surveys described in Figure 2.5 is similar with that of the volunteer-based cohort studies in epidemiology (Figure 2.1).

Figure 2.5 Protocol of Volunteer-Based Panel Web Surveys



Source: Lee (2004)

Two main bias sources are coverage and self-selection. Considering the similarity between these two types of samples, we can apply the existing weighting methods in survey statistics to epidemiologic studies for representativeness improvement. In the next section, two major types of PS-based methods in survey statistics are critically described.

2.3.3.3 PS-Based Methods for Nonprobability Survey Samples

The existing PS-based adjustments assume that the nonprobability samples have some probability sampling mechanism under which each selected unit has an inclusion probability (or participation rates) and a corresponding sample weight. The goal is to estimate the unknown sample weights (pseudo-weights) relying on a true probability sample that well represents the finite target population. The reference sample is assumed to be independently selected from the same target population with the nonprobability sample and has common variables that explain the unknown sampling mechanism. There are two major types of existing PS-based adjustments in survey statistics: PS-based weighting methods, which use (functions of) PSs to directly estimate participation rates of the nonprobability sample, and PS-based matching methods, which use PS to measure the similarity between the nonprobability sample and the reference sample units.

- PS-Based Weighting Methods

The PS-based weighting approaches attempt to estimate the unknown participation rates for the nonprobability sample units using PS and use the inverse of estimated participation rates as the pseudo-weights. Suppose there are two samples independently (self-) selected from a finite target population (FP): a volunteer-based nonprobability sample (s_c) with n_c units, and a reference probability-based survey sample (s_s) with n_s units, each with a sample weight of d_i, for $i \in s_s$. The ultimate goal is to estimate the participation rate $\pi_i^{(c)} = P\left(\delta_i^{(c)} = 1|\boldsymbol{x}_i\right)$; that is, the probability of being included in s_c for unit $i \in FP$ given some observed covariates $\boldsymbol{x}_i$, where $\delta_i^{(c)}$ (=1 if $i \in s_c$; =0 if $i \in FP - s_c$) is a binary variable indicating whether individual i in the finite population FP is included in the nonprobability sample s_c. This can be done by fitting a logistic regression in FP with all variables that relate to the unknown sampling mechanism of the nonprobability sample being included in the propensity model as covariates (Stuart, 2011). The PS represents the likelihood of being in the nonprobability sample for the population units. However, it is unrealistic to obtain population information. Several alternative PS-based weighting methods are proposed instead.

The first PS-based weighting method is the inverse of PS weighting (IPSW) method proposed by Valliant & Dever (2011). The IPSW method estimates the propensity of being observed in the nonprobability sample by fitting a logistic regression model (2.3.9) to the combined nonprobability sample and *weighted* survey sample:

$$\log\left\{\frac{p_i}{1-p_i}\right\} = \boldsymbol{\beta}^T \boldsymbol{x}_i, \qquad i \in \{s_c \cup^* s_s\} \tag{2.3.9}$$

where p_i is the likelihood of $i \in s_c$ conditional on the cohort and *weighted* survey sample, $\boldsymbol{\beta}$ is a vector of coefficients, and $\boldsymbol{x}_i$ is a vector of observed covariates for $i \in \{s_c \cup^* s_s\}$. The notation $\cup^*$ represents the combination of the two samples that allows people to be selected in both a cohort and the survey. The participation rate $\pi_i^{(c)}$ is then estimated by $\hat{p}_i = \text{expit}\left(\widehat{\boldsymbol{\beta}}^T \boldsymbol{x}_i\right)$, where $\widehat{\boldsymbol{\beta}}$ is the estimate of coefficients $\boldsymbol{\beta}$. The corresponding pseudo-weight $i \in s_c$ is the inverse of predicted pseudo-inclusion probability; that is, $\widetilde{w}_i^{IPSW} = \frac{1}{\hat{p}_i}$.

The IPSW method has been shown to reduce bias of the naïve nonprobability sample estimates of finite population means (Valliant and Dever, 2011). However, the IPSW method implicitly requires that s_s is selected from the complement of s_c, $FP - s_c$, which can be assumed only if the sample fraction of s_c is low. Otherwise, the IPSW pseudo-weighted estimates of finite population quantities can be biased.

Chen et al. (2019) proposed a similar PS-based weighting method, refer to as CLW method, estimating the participation rate $\pi_i^{(c)}$ under a well-defined likelihood function:

$$L(\boldsymbol{\gamma}) = \prod\nolimits_{i \in FP} \left\{\pi_i^{(c)}\right\}^{\delta_i^{(c)}} \left\{1 - \pi_i^{(c)}\right\}^{1-\delta_i^{(c)}} \tag{2.3.10}$$

By assuming a logistic regression for $\pi_i^{(c)}$, that is, $\pi_i^{(c)} = \text{expit}(\boldsymbol{\gamma}^T \boldsymbol{x}_i)$, the consistent estimator of coefficients $\boldsymbol{\gamma}$, denoted by $\widehat{\boldsymbol{\gamma}}$ can be estimated by maximizing the pseudo-log likelihood in the combined nonprobability sample and the *weighted* survey sample:

$$l_p(\boldsymbol{\gamma}) = \sum\nolimits_{i \in s_c} \log \frac{\pi_i^{(c)}}{1 - \pi_i^{(c)}} + \sum\nolimits_{i \in s_s} d_i \log\left\{1 - \pi_i^{(c)}\right\},$$

which is equivalent to solving the pseudo-estimating equations:

$$S_p(\boldsymbol{\gamma}) = \frac{\partial l_p(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \sum\nolimits_{i \in s_c} \boldsymbol{x}_i - \sum\nolimits_{i \in s_s} d_i \pi_i^{(c)} \boldsymbol{x}_i = \mathbf{0}. \qquad (2.3.11)$$

The participation rate is estimated by $\hat{\pi}_i^{(c)} = \text{expit}(\hat{\boldsymbol{\gamma}}^T \boldsymbol{x}_i)$, and the corresponding pseudo-weight is $\widetilde{w}_i^{CLW} = \frac{1}{\hat{\pi}_i^{(c)}}$.

The CLW method provides consistent pseudo-weighted estimates of the finite population means under the correct model of participation rate, regardless of the sample fraction of the nonprobability sample. However, the IPSW method is easier to implement than the CLW method by using existing functions in the software (e.g. “svyglm” in survey package of R, and proc “surveylogistc” in SARS).

Furthermore, both the CLW and the IPSW methods assume a logistic regression for the propensity model fitted to the combined (s_c vs. *weighted* s_s) sample. Due to the highly variable weights in the combined sample (common implicit weight of 1 for s_c, and relatively large differential sample weights for s_s), variances of the estimated coefficients $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ can be large, leading to inefficient pseudo-weighted estimators of the finite population quantities.

Different from the IPSW and the CLW methods that fit propensity models to the combined (s_c vs. <u>*weighted*</u> s_s) sample, Elliott (2009) modeled $\pi_i^{(c)}$ by:

$$\pi_i^{(c)} \dot{\propto} \pi_i^{(s)} \frac{\tilde{p}_i}{1 - \tilde{p}_i}, \text{for } i \in s_c \qquad (2.3.12)$$

where $\pi_i^{(s)}$ is the probability of being selected to s_s from the FP for $i \in s_c$, $\tilde{p}_i = P(i \in s_c | \boldsymbol{x}_i, s_c \cup^* s_s)$ is the propensity of being observed in the nonprobability sample given the combined s_c and *unweighted* s_s, and $\boldsymbol{x}_i$ is the set of common covariates available to both samples that are assumed to fully explain the sampling mechanism for both. Two models are needed to estimate the probabilities $\pi_i^{(s)}$ and $\tilde{p}_i$ separately. Elliott (2009) suggested to estimate $\pi_i^{(s)}$ for $i \in s_c$ by fitting a beta regression model (Ferrari & Cribari-Neto, 2004) of $\pi_j^{(s)}$ on $\boldsymbol{x}_j$ for $j \in s_s$, which ensures the estimated response variable, $\hat{\pi}_i^{(s)}$ for $i \in s_c$ is between 0 and 1, and estimate $\tilde{p}_i$ by fitting the logistic regression model:

$$\log\left\{\frac{\tilde{p}_i}{1-\tilde{p}_i}\right\} = \widetilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i, \qquad i \in \{s_c \cup^* s_s\} \tag{2.3.13}$$

to the combined (s_c vs. *unweighted* s_s) sample. Notice that $\widetilde{\boldsymbol{\beta}}$ in Model (2.3.13) can be different from $\boldsymbol{\beta}$ in Model (2.3.9) because the two models are fitted to different samples. The pseudo-weight for $i \in s_c$ is given by inverse of the estimated participation rate $\hat{\pi}_i^{(c)}$:

$$\widetilde{w}_i^E \propto \frac{1}{\hat{\pi}_i^{(s)}} \cdot \frac{1-\hat{\tilde{p}}_i}{\hat{\tilde{p}}_i}, \tag{2.3.14}$$

One advantage of Elliott's method, compared to the IPSW and CLW methods, is that the two models (one for estimating $\pi_i^{(s)}$, and the other for estimating $\tilde{p}_i$) are fitted to the unweighted samples, which avoid the low efficiency caused by the highly variable weights in the combined (s_c vs. *weighted* s_s) sample to which the IPSW and CLW methods fit the propensity model. Another advantage is that Elliott's method can be adapted to situations in which the nonprobability sample represents only a portion of the population. However, compared to the IPSW and CLW methods, Elliott's method requires one more model to estimate $\pi_i^{(s)}$ for $i \in s_c$. The misspecification of either model may bias the pseudo-weighted estimates.

All three of the PS-based weighting methods described above (IPSW, the CLW method, and Elliott's method) yield approximately unbiased estimates of finite population means if the propensity models are correctly specified and if the sample fraction of the nonprobability sample is small. Nevertheless, there are some common drawbacks to the PS-based weighting methods: (1) the biasness of the pseudo-weighted estimates heavily depend on the propensity model specification since the PSs are directly used to predict participation rates; and (2) extreme weights may occur when the estimated PSs are close to 0, leading to inappropriately inflated variance of the pseudo-weighted estimates.

- PS-Based Matching Methods

Different from the PS-based weighting methods that directly use PSs to estimate the participation rates, the PS-based matching methods use the PS to measure the similarity between units in the nonprobability sample and the survey sample. The survey sample weights are distributed to the nonprobability sample units according to their similarity measure by the PS.

- PS Adjustment by Subclassification (PSAS)

The PSAS approach has been widely used to improve representativeness of nonprobability samples in survey research (Lee, 2004, 2006; Lee & Valliant, 2007; Schonlau et al., 2004; Terhanian & Bremer, 2000). Although the ultimate goal is also to estimate the sample weights, unlike the PS-based weighting methods, PSAS does not use the PS to estimate the inclusion probabilities. Instead, it uses the PS to measure the similarity between the two samples units in terms of the common covariates $\boldsymbol{x}$. To avoid potential inefficiency of the pseudo-weighted estimates due to the highly variable weights in the combined (s_c vs. *weighted* s_s) sample, the propensity model (2.3.13) is fitted to the combined (s_c vs. *unweighted* s_s) sample (Lee & Valliant, 2009). In order to create a single set of weights that can be applied to a wide range of inferences for the population, the covariates should include variables related to sample mechanism of the nonprobability sample and/or the survey.

Then, the combined sample $s = s_c \cup^* s_s$ is sorted by the estimated PS, $\hat{\hat{p}}_i, i \in s$ and partitioned into G subclasses. Cochran (1968) recommended to use quintiles. The assumption here is that all units in each subclass have approximately the same PS. In the g-th subclass of the combined sample, denoted as $s^{(g)}, g = 1, \cdots, G$, suppose there are $n^{(g)} = n_c^{(g)} + n_s^{(g)}$ units, where $n_c^{(g)}$ and $n_s^{(g)}$ are the number of units from the nonprobability sample and survey sample, respectively, in subclass g. The PSAS adjustment weight for nonprobability sample unit i in $s^{(g)}$ is calculated by:

$$\widetilde{w}_i^{PSAS} = \frac{\sum_{j \in \left(s_s^{(g)}\right)} d_j}{n_c^{(g)}}, \tag{2.3.15}$$

where $s_s^{(g)}$ is the set of probability sample unit in the g-th subclass, and d_j is the sample weight for probability sample unit j.

Compared to the IPSW method, the PSAS method is less likely to produce extreme weights if the number of class G is not large. The variance of the PSAS estimate is therefore smaller than that of the IPSW estimate. However, the PSAS method is less effective at bias reduction than the IPSW method (Valliant & Dever, 2011).

- Sample Matching

Rivers (2007) proposed an alternative sample matching approach to create a sample from the nonprobability sample that has a close joint distribution of the covariates, $\boldsymbol{x}$, to the finite target population. This approach also requires a representative sample as a reference that also have covariates $\boldsymbol{x}$ in the nonprobability sample.

Suppose there is a sufficiently large and diverse nonprobability sample s_c and a probability sample s_s with each unit having a sample weight $d_i, i \in s_s$. The variable of interest, y, is available in s_c, but not in s_s. For each unit $j \in s_s$, a closest match i, called matched unit, is found from s_c based on a certain measure of distance (e.g., PS, Mahalanobis distance metric). The set of matched units selected from s_c is called matched sample, denoted by s_m, and it will resemble s_s in terms of the covariate distribution. The pseudo-weight for $i \in s_c$, denoted by $\widetilde{w}_i^M$, is given by:

$$\widetilde{w}_i^M = \begin{cases} d_j \text{ if } i \in s_m \text{ and } i \text{ is the matched unit for } j \in s_s, \\ 0 \quad \text{if } i \in s_c - s_m. \end{cases}$$

Under the following regularity conditions, Rivers showed that the matched sample can be used as if it were a probability sample:

(1) Continuous covariates with overlap: The distribution of $\boldsymbol{x}$ in both selection frames of nonprobability and probability sample is absolutely continuous and the supports of $\boldsymbol{x}$ in the two samples are bounded and well overlapped.

(2) Bounded densities: The density function of $\boldsymbol{x}$ is bounded in both samples. This assumption ensures availability of a close match with a sufficiently large panel.

(3) Smoothness: The density function of $\boldsymbol{x}$ and the conditional expectation $\mu(\boldsymbol{x}) = E(y|\boldsymbol{x})$ is continuous on the support of $\boldsymbol{x}$.

(4) Bounded variance: There exists $c < \infty$ such that $Var(y|\boldsymbol{x}) \leq c$ almost surely.

An optional step is to use the PSAS approach to adjust the matched sample. Rivers showed it through simulations that the further step of the PSAS adjustment helps to reduce more bias of the estimate of population mean obtained from the matched sample.

The Rivers method provides a way to select a subsample from the nonprobability sample to resemble to the reference probability sample. However, the unmatched individuals from the nonprobability sample are discarded, which may reduce the efficiency of the analyses due to limited sample size.

2.3.3.4 *Variance Estimation for Pseudo-Weighted Estimates*

Taylor linearization (TL) and Jackknife replication (JK) are two popular variance estimation methods in design-based inference under complex sample designs that take into account the randomness due to both multistage sample design and unequal sample weights. Both TL and JK variance estimators are design consistent. However, different from the design-based inference for the probability-based sample in which the sample weights are known, the pseudo-weights for the nonprobability samples are estimated. The variance estimation for pseudo-weighted estimates should consider all sources of randomness due to estimating the pseudo-weights, differential pseudo-weights, and the potential complex participation mechanism (e.g., homogeneity among the participants). Ignoring any sources of the randomness may lead to biased variance estimation of the pseudo-weighted estimates. This section introduces the existing TL and JK estimation for variance of the pseudo-weighted estimators of finite population mean, $\mu^{FP} = \frac{1}{N}\sum_{i\in FP} y_i$.

- Naïve Taylor Linearization

The naïve TL approach treats the estimated pseudo-weights for the nonprobability sample as fixed. Only the randomness due to differential pseudo-weights and the potential complex participation mechanism will be considered. Denote the pseudo-weighted estimate of the finite population count (N), and the pseudo-weighted estimate of the finite population total, $Y = \sum_{i\in FP} y_i$, obtained from the nonprobability sample as $\widehat{N}^{(c)} = \sum_{i\in s_c} \widetilde{w}_i$ and $\widehat{Y}^{(c)} = \sum_{i\in s_c} \widetilde{w}_i \cdot y_i$, respectively, where $\widetilde{w}_i$ is the pseudo-weight for $i \in s_c$ provided by a PS-based method. The pseudo-weighted estimate of μ is $\hat{\mu}^{(c)} = \frac{\hat{Y}^{(c)}}{\widehat{N}^{(c)}}$. The native TSL variance estimation is given by:

$$var_{NTL}\left(\hat{\mu}^{(c)}\right) = \left(\hat{\mu}^{(c)}\right)^2 \left\{\frac{var\left(\widehat{N}^{(c)}\right)}{\left(\hat{Y}^{(c)}\right)^2} + \frac{var\left(\hat{Y}^{(c)}\right)}{\left(\widehat{N}^{(c)}\right)^2} - \frac{2cov\left(\hat{Y}^{(c)}, \widehat{N}^{(c)}\right)}{\hat{Y}^{(c)} \cdot \widehat{N}^{(c)}}\right\}.$$

When there is an underlying stratified cluster sample design for the nonprobability sample assembling with H strata and a_h clusters in the sample from stratum h, the variance is estimated by:

$$var_{NTL}\left(\hat{\mu}^{(c)}\right) = \sum_{h=1}^{H} \frac{a_h}{a_h - 1} \sum_{\alpha=1}^{a_h} \left(u_{h\alpha}^{*} - \frac{1}{a_h} \sum_{\alpha}^{a_h} u_{h\alpha}^{*} \right)^2, \qquad (2.3.16)$$

where $u_{h\alpha}^{*} = \frac{\partial\mu}{\partial Y}\Big|_{Y=\hat{Y}^{(c)}} \hat{Y}_{h\alpha}^{(c)} + \frac{\partial\mu}{\partial N}\Big|_{N=\widehat{N}^{(c)}} \widehat{N}_{h\alpha}^{(c)} = \frac{1}{\widehat{N}^{(c)}} \hat{Y}_{h\alpha}^{(c)} - \frac{1}{\left(\hat{Y}^{(c)}\right)^2} \widehat{N}_{h\alpha}^{(c)}$, is the known as the linear substitute (Wolter, 2007, Chapter 6), $\hat{Y}_{h\alpha}^{(c)}$ and $\widehat{N}_{h\alpha}^{(c)}$ are the pseudo-weighted estimates of population total and population size of cluster α in stratum h, respectively.

The naïve TL method tends to underestimate the variance of the pseudo-weighted estimates due to, especially for the variances of the IPSW and the CLW estimates. The two methods fit the propensity model to the sample with highly variable weights between the nonprobability and survey samples, which can lead to large variance of the estimated PSs. Using the inefficient estimates of PSs to estimate the participation rates (i.e., inverse of the pseudo-weights) can substantially increase the variance of the pseudo-weighted estimates of the finite population quantities. Hence, ignoring the randomness of the estimated pseudo-weights can result to severe underestimation of the variance for the IPSW and CLW estimates.

- Taylor Linearization Considering the Randomness of Estimating Pseudo-Weights

Chen et al. (2019) proposed the TL variance estimation for CLW estimates of the finite population means that considers the variability due to estimating PS by using the pseudo-estimating system:

$$\Phi(\boldsymbol{\eta}) = \begin{cases} \frac{1}{N}\sum\nolimits_{i\in FP} \delta_i^{(c)}(y_i - \mu^{FP}) \\ \frac{1}{N}\sum\nolimits_{i\in FP} \delta_i^{(c)}\boldsymbol{x}_i - \frac{1}{N}\sum\nolimits_{i\in FP} \delta_i^{(s)} d_i \pi_i^{(s)} \boldsymbol{x}_i \end{cases} = \mathbf{0}, \tag{2.3.17}$$

where $\boldsymbol{\eta} = (\mu^{FP}, \boldsymbol{\kappa})^T$ is a vector of parameters to be estimated, and $\boldsymbol{\kappa}$ is nuisance parameters for participation rate estimation defined in (2.3.10). Using a first order Taylor expansion, the finite population variance of $\widehat{\boldsymbol{\eta}}$ can be approximated as follows:

$$Var(\widehat{\boldsymbol{\eta}}) \doteq [E\{\phi(\boldsymbol{\eta})\}]^{-1} Var\{\Phi(\boldsymbol{\eta})\}[E\{\phi(\boldsymbol{\eta})\}^T]^{-1} \tag{2.3.18}$$

where $\phi(\boldsymbol{\eta}) = \partial\Phi/\partial\boldsymbol{\eta}$, the expectation E and variance Var is with respect to random selection of the nonprobability and the probability samples. The sample estimate of $Var(\widehat{\boldsymbol{\eta}})$ can be obtained by substituting the finite population quantities in (2.3.18) with the sample estimates.

Chen et al. (2019) provided a framework for estimating the participation rates of the nonprobability sample and for deriving TL variance estimation for pseudo-weighted estimates of finite population means. A similar approach can be applied for more complex estimates such as regression coefficients, which, may require extra tremendous derivation. On the contrary, the JK method can take into account the randomness due to estimating pseudo-weighting by re-estimating all the parameters in replication, which avoids extra computation for different estimates.

- Jackknife Replication

Lee & Valliant (2009) proposed a JK variance estimation for the PSAS estimates of population means. It can be similarly applied to other PS-based pseudo-weighted estimates of finite population quantities. The nonprobability sample is randomly divided into G equal-sized replication groups. The JK estimator for an estimate of population parameter $\hat{\theta}^{PW}$ is:

$$var_{JK}(\hat{\boldsymbol{\omega}}^{PW}) = \frac{G-1}{G}\sum_{g=1}^{G}\left(\hat{\boldsymbol{\omega}}_{(g)}^{PW} - \hat{\boldsymbol{\omega}}^{PW}\right)^2, \qquad (2.3.19)$$

in which $\hat{\boldsymbol{\omega}}_{(g)}^{PW}$ is the pseudo-weighted estimate obtained from the nonprobability sample omitting units in the g-th group, $g = 1, \cdots, G$. In reach replicate, the pseudo-weights are re-calculated. This indirectly reflects the facts that the estimated weights are subject to sample variation. This approach is shown to perform better than the naïve TL method (Lee & Valliant, 2009) which ignores the variability of estimating the pseudo-weights.

However, estimator (2.3.19) does not consider the variability due to randomness of survey sampling or the potential homogeneity in the nonprobability sample. In order to include these two components in the JK variance estimator, the nonprobability sample can be treated as an extra stratum in the combined sample. Each replicate omits one cluster in the probability survey sample or in the nonprobability sample.

2.3.4 Summary

This section summarized two types of PS-based methods (PS-based weighting and matching methods), which improve representativeness of nonprobability samples by using a probability survey sample as the reference. The PS-base weighting methods can reduce bias when the propensity model is correctly specified. Nevertheless, the pseudo-weighted estimates can be inefficient especially if there are extreme pseudo-weights. For the IPSW and the CLW methods, the highly variable weights among the combined nonprobability and probability sample can also lead to inefficient estimates of PSs, which inflate the variance of the pseudo-weighted estimates. Furthermore, they can be sensitive to propensity model specification because the PSs are used to estimate the participation rates. Compared to the PS-based weighting methods, the PS-based matching methods can be less sensitive to the propensity model specification because the PSs are used to measure the similarity between the nonprobability and probability sample units. The sample matching method can provide unbiased estimates under some standard conditions but sacrifices the sample size. The PSAS method avoids extreme weights (Rubin 2001), and therefore yields less variable estimates. However, the PSAS method is less effective at bias reduction (Valliant & Dever, 2011) because the key assumption that nonprobability sample units represent equal numbers of population units within subclasses is hard to be satisfied in reality. Moreover, the measure of similarity of PSs is ad hoc with limited guidance and justification for forming the subclasses.

The TL methods provide close forms of variance estimators, and they are more computing-efficient than replicate methods. The naïve TL variance estimator, however, is well known for possibly underestimating the variance as it ignores the randomness due to estimating PSs (Lee & Valliant, 2009; Landsman & Graubard, 2013; Abadie & Imbens, 2016). The complete TL variance estimator (Chen et al., 2019) should be derived from the estimating system, including both PS estimation and the estimation of the finite population quantities. The complete TL estimator may require tremendous calculation for different estimators.

On the contrary, the JK variance estimators can automatically take into account all sources of variability by re-estimating the pseudo-weights at each replicate. The JK estimator proposed by Lee & Valliant (2009), however, does not consider the randomness due to survey sample selection because the replicates only drop groups of nonprobability sample units. More research is required to test how the randomness of reference survey sample would influence the variance estimation. Moreover, Lee & Valliant (2009) did not consider potential cluster effects of the data. In epidemiologic cohort studies, volunteers are usually recruited from a set of study centers. There may be intra-class correlation due to geographical homogeneity. New variance estimation is required to fill these gaps.

2.4 PS-Based Kernel Smoothing

2.4.1 Introduction

Kernel smoothing is a widely used nonparametric regression technique to estimate the conditional expectation of outcome y given a vector of covariates $\boldsymbol{x}$. Let $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)$ be a sample of n independent and identically distributed observations where y_i's are scalar response variables and the $\boldsymbol{x}_i$'s are covariates. Let the conditional mean of y given $\boldsymbol{x}_0$ be denoted by $m(\boldsymbol{x}_0) = E(y|\boldsymbol{x}_0)$, where $\boldsymbol{x}_0$ is any possible value of the vector of covariates $\boldsymbol{x}$. The kernel estimator of $m(\boldsymbol{x}_0)$ is given by a weighted mean of the observed outcome y:

$$\widehat{m}(\boldsymbol{x}_0) = \frac{\sum_{i=1}^{n} W_i(\boldsymbol{x}_0) y_i}{\sum_{i=1}^{n} W_i(\boldsymbol{x}_0)},$$

where $W_i(\boldsymbol{x})$ is the kernel weight (Nadaraya, 1964; Waston 1964) for sample unit i, estimating the conditional distribution of $(y \mid \boldsymbol{x})$, defined as follows:

$$W_i(\boldsymbol{x}_0) = \frac{K\left(\frac{\|\boldsymbol{x}_0, \boldsymbol{x}_i\|}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\|\boldsymbol{x}_0, \boldsymbol{x}_i\|}{h}\right)}$$

where $K(\cdot)$ is a kernel function (described later), $\|\cdot\|$ represents the distance between the two units (e.g., Euclidean distance) and $h > 0$ is the bandwidth depending on the sample size n and the choice of $K(\cdot)$. The kernel weight $W_i(\boldsymbol{x}_0)$ is a relative distance between the covariate $\boldsymbol{x}_0$ and covariates $\boldsymbol{x}_i$ compared to all other covariates with $\sum_{i=1}^{n} W_i(\boldsymbol{x}_0) = 1$. The more similar the covariates $\boldsymbol{x}_0$ and $\boldsymbol{x}_i$ are, the larger the $W_i(\boldsymbol{x}_0)$ will be.

The kernel function $K(\cdot)$ satisfies the conditions: $\int K(u)du = 1$ and $\int |K(u)|du < \infty$. Some commonly used kernel functions are described in Benedetti (1997):

Table 2.1 Examples of Kernel Functions.

Uniform	Quadratic	Triangular	Gaussian
$K(u) = \frac{1}{2} \cdot I\{\lvert u \rvert < 1\}$	$K(u) = \frac{3}{4}(1 - \lvert u \rvert^2)^+$	$K(u) = (1 - \lvert u \rvert)^+$	$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$

Different kernels may yield different kernel weights $W_i(\boldsymbol{x})$. For example, in Table 2.1, a uniform density kernel gives the same weights to the units whose distances from $\boldsymbol{x}$ are within the support $(-1, 1)$, whereas the other three kernels give more weights to units that are close to $\boldsymbol{x}_0$. With a Gaussian density kernel, all sample units receive positive kernel weights. With a quadratic, or triangular density kernel, units receive a weight of 0 if their distances with $\boldsymbol{x}_0$ are larger than h.

The consistency of the kernel regression estimates requires the bandwidth $h \to 0$ and $n \cdot h \to \infty$ when $n \to \infty$ (Owen, 1987). There are various methods for bandwidth selection. Here we introduce five methods that are most commonly used for kernel density estimation: (1) Silverman's rule of thumb Silverman (1986), (2) Scott's method (Scott, 1992), (3) unbiased cross-validation (UCV; Scott & Terrell, 1987), (4) biased cross-validation (BCV; Scott & Terrell, 1987), and (5) Sheather & Jones's method (S&J;

Sheather & Jones, 1991). The kernel density estimation estimates the unknown probability density function of the random variable x, $f(x)$ by

$$\hat{f}_h(x) = \frac{1}{nh}\sum\nolimits_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{2.4.1}$$

The five methods select a bandwidth by minimizing mean integrated squared error (MISE) or asymptotic mean integrated squared error (AMISE). Minimizing AMISE with respect to h gives the optimal bandwidth:

$$h_{\text{opt}} = \left(\frac{R(K)}{n\sigma_K^4 R(f'')}\right)^{1/5}, \tag{2.4.2}$$

where K is the kernel density function, σ_K is the corresponding standard deviation, $R(K) = \int K^2(z)dz$, n is the sample size, and f is the unknown density distribution to be estimated, with f'' being the second derivative of f. Since f is unknown, $R(f'')$ has to be estimated.

Silverman (1986) uses a density function of $N(\hat{\mu}, \hat{\sigma})$ to replace f in Expression (2.4.2) where $\hat{\mu}$, and $\hat{\sigma}$ are the sample mean and standard deviation. If $K(\cdot)$ is a Gaussian density kernel, then the bandwidth is:

$$h_{opt} = \left(\frac{4}{3n}\right)^{\frac{1}{5}} \hat{\sigma}, \tag{2.4.3}$$

where n is the sample size. Scott (1992) modified Silverman's method by using a more robust estimate of standard deviation, $\min(\hat{\sigma}, IQR/1.35)$ to replace $\hat{\sigma}$ in Expression (2.4.3), where IQR is the interquartile range (difference between third and first quartile).

The S&J method improved Silverman's and Scott's methods by using an empirical estimate of $R(f'')$ instead of using normal density to approximate f in (2.4.2). They applied a two-stage approach to obtain the optimal bandwidth in (2.4.2). At the first stage, f is estimated by (2.4.1) with h selected by Silverman's rule of thumb and obtain an empirical estimate of $R(f'')$, denoted by $\widehat{R(f'')}$. At the second stage, the final optimal bandwidth is obtained by using $\widehat{R(f'')}$ to substitute $R(f'')$ in (2.4.2).

UCV and BCV use cross-validation techniques to estimate the expectation of the estimated density function by minimizing MISE and AMISE, respectively, that is, using $\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i}(x_i)$ to estimate $E\{\hat{f}(x)\}$, with $\hat{f}_{-i}(x_i) = \frac{1}{h(n-1)}\sum_{j\neq i}K\left(\frac{X_i - X_j}{h}\right)$.

2.4.2 KW Methods for Treatment Effect Estimation

In economics, kernel weighting methods have been proposed to match the controls to the treatment units in estimating average treatment effects (Heckman et al., 1997; 1998a; 1998b; Imbens, 2004). The same as notations introduced in Section 2.4, $(y_{0,i}, y_{1,i})$ are the potential outcome for sample unit i; s_0 and s_1 denote the control and treatment group, with sizes n_0 and n_1 respectively; Z_i is a binary variable indicating the treatment group (0 for $i \in s_0$, and 1 for $i \in s_1$). The conditional expectation of outcome y given a vector of covariates $\boldsymbol{x}$, $\mu_z(\boldsymbol{x}) = E[y|z, \boldsymbol{x}]$, is estimated by $\hat{\mu}_z(\boldsymbol{x}) = \sum_{i\in s_z} y_{z,i}k_i / \sum_{i\in s_z} k_i$,where $k_i = K\left(\frac{\|\boldsymbol{x}_i - \boldsymbol{x}\|}{h}\right)$. The potential outcome $y_{0,j}$ for unit $j \in s_1$ is missing, and is imputed with a consistent estimator:

$$\hat{\mu}_0(\boldsymbol{x}_j) = \sum\nolimits_{i\in s_0} k_{ij}y_{0,i} \Big/ \sum\nolimits_{i\in s_0} k_{ij}, \tag{2.4.4}$$

where $k_{ij} = K\left(\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{h}\right)$. The treatment effect, Δ, is calculated by:

$$\widehat{\Delta} = \frac{1}{n_1}\sum\nolimits_{j\in s_1}\left(y_{1,j} - \hat{\mu}_0(\boldsymbol{x}_j)\right). \tag{2.4.5}$$

The unbiasedness of $\widehat{\Delta}$ in (2.4.5) requires the condition of mean independence; that is,

$$E[y_0|\boldsymbol{x}, z = 0] = E[y_0|\boldsymbol{x}, z = 1]. \tag{2.4.6}$$

Heckman et al. (1998a) proposed to use the difference in the propensity of being in the control group versus the treatment group to measure the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ (i.e., using $\{e(\boldsymbol{x}_i) - e(\boldsymbol{x}_j)\}$ to replace $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ in k_{ij}) to avoid the dimensionality problem.

We found that this approach can be viewed as assigning weights to the control group. The estimated treatment effect $\widehat{\Delta}$ in (2.4.5) can be written as:

$$\widehat{\Delta} = \frac{1}{n_1}\sum_{j\in s_1}\left\{y_{1,j} - \sum_{i\in s_0}\frac{k_{ij}y_{0,i}}{\sum_{i\in s_0}k_{ij}}\right\}$$

$$= \frac{1}{n_1}\sum_{j\in s_1} y_{1,j} - \frac{1}{n_0}\frac{n_0}{n_1}\sum_{i\in s_0}\sum_{j\in s_1}\frac{k_{ij}y_{0,i}}{\sum_{i\in s_0}k_{ij}}$$

$$= \frac{1}{n_1}\sum_{j\in s_1} y_{1,j} - \frac{1}{n_0}\sum_{i\in s_0} W_i y_{0,i}$$

where $W_i = \frac{n_0}{n_1}\sum_{j\in s_1}\frac{k_{ij}}{\sum_{i\in s_0}k_{ij}}$ is a PS-based weight for $i \in s_0$, and

$$\sum_{i\in s_0} W_i = \frac{n_0}{n_1}\sum_{i\in s_0}\sum_{j\in s_1}\frac{k_{ij}}{\sum_{i\in s_0}k_{ij}}$$

$$= \frac{n_0}{n_1}\sum_{j\in s_1}\sum_{i\in s_0}\frac{k_{ij}}{\sum_{i\in s_0}k_{ij}} = \frac{n_0}{n_1}\cdot n_1 = n_0.$$

After the transformation, the condition of mean independence (2.4.6)required by the unbiasedness of $\widehat{\Delta}$ becomes:

$$E[y_0|W, z = 0] = E[y_0|W, z = 1]. \qquad (2.4.7)$$

There are two factors affecting the performance of this method in terms of the MSE: kernel function $K(\cdot)$ and bandwidth h. A kernel function with a flatter curve can result in larger bias but smaller variance of $\widehat{\Delta}$ due to smaller variance of the weights $\{W_i, i \in s_0\}$. For example, with a uniform density kernel $U(-1, 1)$, control group member i obtains a PS-based weight $W_i = \frac{n_0}{n_1}\sum_{j\in s_1}\frac{I\{|e(\boldsymbol{x}_i)-e(\boldsymbol{x}_j)|<h\}}{\sum_{i\in s_0} I\{|e(\boldsymbol{x}_i)-e(\boldsymbol{x}_j)|<h\}}$. Control group members can obtain an identical PS-based weight W if the distance between their PSs is smaller than h, which lead to less variable $\{W_i, i \in s_0\}$. However, they may have different values of $\boldsymbol{x}$ and correspondingly different values of y_0. As a result, the condition of mean independence (2.4.7) is violated, leading to a biased estimate of Δ. A large bandwidth can increase bias but reduce variance due to the same reason.

2.4.3 Summary

For treatment effect estimation, the kernel smoothing method, as a PS-based matching method, provides a way to balance the tradeoff between bias reduction and variance inflation. By choosing an appropriate kernel function and bandwidth, it can be less likely to produce extreme weights that inflate the variance than the PS-based weighting methods, such as IPSW. Meanwhile, it can be more effective at bias reduction than the PSAS method. However, there is no existing literature comparing the performance of the kernel smoothing method, IPSW, and PSAS in the context of treatment effect estimation. In finite population inference, there is no kernel smoothing method proposed for improving the population representativeness of nonprobability samples. The kernel weighting (KW) can be addressed to create a set of pseudo-weight for the volunteer-based nonprobability samples by distributing the survey sample weights to the nonprobability sample units based on their similarity measured by kernel smoothed distance in PS. Both theoretical and practical work have to be done to investigate the properties of the KW method and to compare its performance with the existing IPSW and the PSAS methods.

2.5 Improving Efficiency of the PS Estimation by Scaling the Weights

As discussed in Section 2.3.3, fitting propensity model to the combined nonprobability and weighted probability survey sample, as the IPSW method does, may cause high variance of the estimated PSs, which may lead to inefficient pseudo-weighted estimates of the finite population quantities. A similar problem occurs in the population-based case-control studies where the highly variable sample weights among the sample cases and controls lead to large variance of the estimated regression coefficients. Scott and Wild (1986, 2002) showed that although the sample-weighted estimates of the regression coefficients are design consistent and more robust to the outcome model misspecification, they can be inefficient. Scott and Wild (1986, 2002) suggested an alternative approach that rescales the sample weights so that the control or case weights sum up to the sample size of controls or cases. The scaled sample weights between the cases and controls have much lower overall variation than the original sample weights, and thus can improve the efficiency of the regression analyses. Although the scaled sample weighted estimate of the intercept is biased, the bias can be removed by adding an offset. Li et al. (2011) and Landsman & Graubard (2012) extended Scott and Wild's work to more general complex sample designs.

Suppose a sample of n_1 cases (s_1) and a sample of n_0 controls (s_0) are randomly selected from the population of cases and controls respectively. The binary response variable y (1 for cases, and 0 for controls) and a vector of explanatory covariates $\boldsymbol{x}$ follow a logistic regression model in the population:

$$\text{logit}\{\Pr(y = 1 \mid \boldsymbol{x}, \boldsymbol{\theta})\} = \theta_0 + \boldsymbol{\theta}_1^T \boldsymbol{x}$$

where $\boldsymbol{\theta} = (\theta_0, \boldsymbol{\theta}_1^T)^T$ is a vector of the coefficients, with θ_0 being the intercept, and can be consistently estimated by solving the weighted estimation equations:

$$\sum\nolimits_{i \in s_0 \cup s_1} d_i \boldsymbol{x}_i \{y_i - \text{expit}(\theta_0 + \boldsymbol{\theta}_1^T \boldsymbol{x})\} = \mathbf{0} \tag{2.5.1}$$

The scaled sample weight is defined as:

$$d_i^* = y_i d_i / M_1 + (1 - y_i) d_i / M_0$$

where d_i is the sample weight for individual $i \in s_0 \cup s_1$, $M_1 = \sum_{i \in s_1} d_i / n_1$, and $M_0 = \sum_{i \in s_0} d_i / n_0$ are the scaling factors; that is, the mean of the sample weights within cases and controls, respectively. The coefficients $\boldsymbol{\theta}$ estimated by solving the scaled-weighted estimating equations with d_i^* replacing d_i in estimating equations (2.5.1), denoted by $\widehat{\boldsymbol{\theta}}_{sw}$, is consistent to $\boldsymbol{\theta} + \log\left(\frac{M1}{M0}\right) \cdot (1, 0, \cdots, 0)^T$ (Scott & Wild, 1986). Scott & Wild (2001) showed, by simulations, that the variance of $\widehat{\boldsymbol{\theta}}_{sw}$ is much smaller than that of $\widehat{\boldsymbol{\theta}}_w$ estimated from (2.5.1).

Scaling the sample weights for logistic regression analyses in case-control studies has been shown to gain great efficiency of the regression coefficients estimation only at the sacrifice of the biased estimate of the intercept, which can be adjusted by a known offset (Scott & Wild, 2001; Li et al., 2011; Landsman and Graubard, 2012). This approach can be applied to any logistic regression analyses that have estimating equations with the same form of (2.5.1). For example, the IPSW method fits the logistic regression model to the combined sample of nonprobability and sampled weighted probability sample. The resulting estimates are known to have large variance due to highly variable weights in the combined sample. However, no weighting adjustment has been developed to solve this problem. Scaling the survey sample weights may help reduce variance of the estimated propensity model coefficients and, therefore, improve efficiency of the IPSW estimates.

Chapter 3 A Kernel Weighting Approach for Estimating Finite Population Means from Volunteer-Based Epidemiological Cohorts

3.1 Introduction

As discussed in Section 2.3.3, IPSW and PSAS are two of the commonly used approaches that improve the representativeness of the nonprobability sample in survey research. They also can be applied to improve external validity of epidemiologic studies. However, both approaches are known to have disadvantages.

The IPSW approach, as a PS-based weighting method, uses (functions of) PS to estimate participation rates of the nonprobability sample. Hence, the estimates can be sensitive to propensity model misspecification. In addition, by fitting the propensity model to the combined (nonprobability vs. *weighted* survey) sample, the highly variable weights between the nonprobability and the survey sample may lead to inefficient pseudo-weighted estimates of the finite population quantities. Furthermore, the IPSW method tends to produce extreme weights when the estimated PSs are close to 0. These extreme weights may not be caused by the true underlying small selection probabilities, but by the propensity model misspecification. As a result, variance of weighted estimates would be inappropriately inflated. Though weight trimming may help to reduce the variance, its effect on reducing bias and variance is unclear (Lee et al., 2010; Potter & Zheng, 2015), and there is relatively little guidance regarding the trimming level.

Different from the IPSW method, the PSAS method, as a PS-based matching method, is less sensitive to the propensity model misspecification because it uses the PS to measure the similarity between the units in the nonprobability and survey sample. In addition, the PSAS method avoids extreme weights (Rubin 2001), and therefore yields less variable estimates. Nevertheless, the PSAS method is less effective at bias reduction (Valliant & Dever, 2011) because of the key assumption that nonprobability sample units represent equal numbers of population units within subclasses. Moreover, the measure of similarity of propensity scores is ad-hoc with limited guidance and justification for forming the subclasses.

Moreover, little attention has been paid to the effect that geographic clustering within the nonprobability sample has on variance estimation. Different from the web surveys that can recruit people almost everywhere, the epidemiologic cohort studies usually recruit volunteers at multiple study centers that are in various locations in the geographical areas where the target population resides. The resulting samples may have geographical effects (clustering and correlation of observations). Ignoring the geographical correlation may lead to invalid variance estimation of association between smoking and lung cancer. However, existing PS-based methods do not take into account geographic clustering effect for the variance estimation.

In this chapter, we propose a new PS-based matching method, the kernel weighting (KW) approach, to improve the representativeness of the volunteer-based epidemiologic cohort (cohort for simplicity) by using population-based survey sample as a reference. The KW pseudo-weighted cohort will be used to improve the external validity of the disease prevalence rate estimation. The new method is not expected to remove bias due to non-representativeness, but to gain a better bias-variance tradeoff in estimating disease prevalence in the population. Under certain regular conditions, the KW pseudo-weighted estimator is consistent to the true finite target population mean. The naïve Taylor linearization (TL) and jackknife replication (JK) methods are applied to provide variance estimation for the KW estimates of population prevalence.

Monte Carlo simulation studies are conducted to evaluate performance of the KW estimates of disease prevalence comparing to the existing IPSW and PSAS estimates. The three weighting methods (IPSW, PSAS, and KW) are also applied to estimating nine-year disease incidence and mortality from the National-Institutes-of-Health-American-Association-of-Retired-Persons (NIH-AARP) cohort, using the 1997 National Health Interview Survey (NHIS) as the reference.

3.2 Method

3.2.1 Basic setting

Let the target finite population (FP) consist of N individuals indexed by $i \in \{1, \cdots, N\}$, where each individual i has values for the outcome variable of interest y_i and for the vector of covariates $\boldsymbol{x}_i$. We focus on estimating the FP mean of y, i.e., $\mu_{FP} = N^{-1} \sum_{i \in FP} y_i$. Let $s_c \subset FP$ denote a cohort with n_c individuals. We define a random indicator variable $\delta_i^{(c)}$ ($= 1$ 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

$$\pi_i^{(c)} \equiv P(i \in s_c \mid FP) = E_c\left(\delta_i^{(c)} \mid FP \right),$$

where the expectation E_c is with respect to the unknown random cohort sample selection process from FP. The corresponding cohort implicit sample weight is $w_i = 1/\pi_i^{(c)}$ for $i \in$

s_c. All the finite population units are assumed to have a positive participation rate, i.e., $\pi_i^{(c)} > 0$ 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 $\delta_i^{(s)}(= 1$ if $i \in s_s$; 0 otherwise), $\pi_i^{(s)} = E_s\Big(\ \delta_i^{(s)} \mid FP\ \Big)$, and $d_i = 1/\pi_i^{(s)}$, 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, the inclusion probability and the sample weights are assumed to be adjusted by the nonresponse and calibrated to the known population quantities.

3.2.2 Kernel Weighting Method to Create Pseudo weights for a Cohort

In this section, we propose a new PS-based matching method, the kernel weighting (KW) approach, to create pseudo-weights for the cohort by using a probability survey sample as a reference. Analogous to the PSAS method, KW uses PS to measure the similarity of the covariate distributions between the cohort and the survey samples. Accordingly, the propensity model (3.2.1) is fitted to the combined (s_c vs. *unweighted* s_s) sample.

$$\log\left\{\frac{\tilde{p}_i}{1-\tilde{p}_i}\right\} = \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i, \qquad i \in \{s_c \cup^* s_s\} \tag{3.2.1}$$

The PS for $i \in s_c$ and $j \in s_s$ and are denoted by $\tilde{p}_i^{(c)} = \text{expit}\Big\{\tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i^{(c)}\Big\}$ and $\tilde{p}_j^{(s)} = \exp\Big\{\tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i^{(c)}\Big\}$, with the superscripts (s) and (c) denoting that unit i and unit j are in the survey and in the cohort, respectively.

For $i \in s_s$, we compute the (signed) distance of its estimated PS from each $j \in s_c$, $\tilde{p}_j^{(s)} - \tilde{p}_i^{(c)}$, which ranges from -1 to 1. We apply a kernel function centered at zero to smooth the distances. The closer to zero the distance is, the more similar the pair of units is with respect to the covariates, and accordingly the KW method assigns a larger portion of the survey sample weight d_i to the cohort unit j based on the kernel weight:

$$k_{ij} = \frac{K\left\{\left(\tilde{p}_j^{(s)} - \tilde{p}_i^{(c)}\right)/h\right\}}{\sum_{i\in s_c} K\left\{\left(\tilde{p}_j^{(s)} - \tilde{p}_i^{(c)}\right)/h\right\}} \qquad \text{for } i \in s_c, \tag{3.2.2}$$

where $K(\cdot)$ is a zero-centered kernel function (Epanechnikov, 1969) (e.g. uniform, standard normal, or triangular density), and h is the bandwidth corresponding to the selected kernel function (see Section 2.4.1 for discussion of various bandwidth selection methods). Note that $\sum_{i\in s_c} k_{ij} = 1$ and $k_{ij} \in [0, 1]$. The larger the k_{ij} is, the more similar the propensity scores are between cohort unit j and survey unit i.

Finally, the KW pseudo-weight w_i^{KW} for $i \in s_c$, is a sum of the survey sample weights, $\{d_j, j \in s_s\}$, that are weighted by the cohort unit i's kernel weights, $\{k_{ij}, j \in s_s\}$, given by

$$\mathrm{w}_i^{KW} = \sum\nolimits_{j\in s_s} k_{ij} \cdot d_j \tag{3.2.3}$$

Note the pseudo weight w_i^{KW} takes larger proportion of the survey weights associated with survey members whose PSs are closer to cohort unit i. The KW estimator of the finite population prevalence, μ_{FP}, is

$$\hat{\mu}^{KW} = \frac{1}{\hat{N}^{KW}} \sum\nolimits_{i\in s_c} w_i^{KW} \cdot y_i,$$

where $\hat{N}^{KW} = \sum_{i\in s_c} w_i^{KW}$ is the sum of the cohort pseudo-weights. Notice that $\hat{N}^{KW}$ is an unbiased estimator of the finite population count N. It can be shown

$$\hat{N}^{KW} = \hat{N}^{SVY}, \tag{3.2.4}$$

where $\widehat{N}^{SVY} = \sum_{j \in s_s} d_j$ is a unbiased estimator of N from the survey sample, because $\sum_{i \in s_c} w_i^{KW} = \sum_{i \in s_c} \sum_{j \in s_s} (k_{ij} \cdot d_j) = \sum_{j \in s_s} (d_j \cdot \sum_{i \in s_c} k_{ij}) = \sum_{j \in s_s} d_j$. Hence,

$$E(\widehat{N}^{KW}) = E(\widehat{N}^{SVY}) = N$$

Furthermore, the KW estimators of population means or prevalences are design consistent, under regularity conditions (**Theorem 3.1**)

Theorem 3.1 Consistency of the KW estimate of the finite population mean

Suppose, in the superpopulation, the variable of interest y has an expectation $E(y) = \mu < \infty$, where E denotes the expectation with respect to the joint distribution of y and covariates $\boldsymbol{x}$. Assume that the cohort and the survey sample are selected from a finite population (a simple random sample from a superpopulation) and the distributions of the estimated propensity scores are well overlapping between the two samples. If the following conditions are satisfied:

(a) for the kernel function $K(u)$, $\int K(u)du = 1$, $\sup_u |K(u)| < \infty$, and $\lim_{|u| \to \infty} |u| \cdot |K(u)| = 0$;

(b) for the bandwidth $h = h(n_c)$, $h \to 0$, but $n_c \cdot h \to \infty$ as $n_c \to \infty$;

(c) exchangeability, $E\{y|\tilde{p}, cohort\} = E\{y|\tilde{p}, survey\} = E\{y|\tilde{p}\}$;

(d) bounded second moment, $E(y^2) < \infty$; and

(e) bounded survey sample weights, $w_i < M$ for some $M \in \mathbb{R}_{>0}$, $i \in s_s$;

then the KW estimator of the population mean $\hat{\mu}^{KW} = \frac{\sum_{i \in s_c} w_i^{KW} \cdot y_i}{\sum_{i \in s_c} w_i^{KW}} \to \mu$ in probability as the finite population size $N \to \infty$, the survey sample size $n_s \to \infty$, the cohort sample size $n_c \to \infty$, with $\frac{n_c}{N} = O(1)$, and $\frac{n_s}{N} = O(1)$ (proof in Section 3.6.1).

In practice, if a cohort or a survey sample includes only specific subgroups of people in the population (e.g. a women's health cohort), then both samples should be constrained to the same subgroup. Otherwise, the estimated propensity scores of the two samples may not overlap well for important covariates, which can lead to unreliable pseudo-weighted estimates (Stuart 2011; Stürmer et al., 2010). We recommend checking on the extent of overlap of the PSs used to compute the pseudo-weights from the IPSW, PSAS, and KW methods. Another issue is the covariate selection for Model (3.2.2). Following Stuart (2010), we suggest including as many variables that could be related to the unknown (self-)selection scheme of the cohort, as possible. All cohort selection-related variables that are common to both samples and their two-way interactions might be initially included in the model. Model selection criteria such as a stepwise procedure (D'Agostino, 1998) with Akaike information criteria (AIC) can be applied to obtain a final model.

Also note that PSAS is a special case of the KW method, with a uniform kernel function in each subclass of estimated PSs, assuming that cohort units within subclasses represent equal numbers of population units (proof in Section 3.6.2). In contrast, the KW method relaxes the key PSAS assumption by assigning various portions of the survey weights to the cohort units according to the similarity of covariates considered in the propensity model, measured by the PS.

3.2.3 Variance Estimation

In this section, the naïve Taylor Linearization (TL) and the Jackknife replication (JK) methods are discussed for variance estimation. JK variance estimators that considers all sources of variability are proposed for the pseudo-weighted estimators using the IPSW, PSAS, and KW methods.

As it was discussed in Section 2.3.3.4, the naïve TL method treats the pseudo-weights as the fixed sample weights. Suppose the cohort is collected from C study centers which are treated as clusters for variance estimation. The naïve TL variance estimator for the estimated population prevalence is given by Expression (2.3.16), with the linear substitute for study center α being $u_\alpha^* = \frac{1}{\hat{N}^{KW}} \hat{Y}_\alpha^{KW} - \frac{\hat{Y}^{KW}}{(\hat{N}^{KW})^2} \hat{N}_\alpha^{KW}$, where $\hat{Y}_\alpha^{KW} = \sum_{i \in \alpha} w_i^{KW} y_i$, $\hat{N}_\alpha^{KW} = \sum_{i \in \alpha} w_i^{KW}$, and $\hat{Y}^{KW} = \sum_{i \in s_c} w_i^{KW} \cdot y_i$.

Though the naïve TL method provides a close form of the variance estimation which is computing efficient, it can underestimate the true variance due to ignoring the variability for estimating PSs. Lee & Valliant (2009) addressed this underestimation for PSAS. To improve variance estimation, we propose a JK method to account for all sources of variability (Ch. 2.5, Korn & Graubard, 1999).

Suppose that the survey sample s_s be randomly selected from a target population by a stratified multistage sample design with L strata in the population as described in Section 2.1 of the main text. At the first stage of sampling, m_l clusters (i.e., PSUs) are randomly selected (approximated by sampling with replacement) from stratum l, for $l = 1, \cdots, L$. The cohort is recruited from C study centers, which are treated as a random sample of clusters (i.e., PSUs) from the finite population.

We combine the cohort with the survey sample and treat the cohort as the $(L+1)$-th stratum in the combined sample. The leave-one-out jackknife (JK) variance estimation procedure involves leaving one PSU out of the combined sample at a time, adjusting the weights in the survey or cohort for the smaller number of sampled PSUs, recomputing new pseudo-weights for the cohort with these adjusted weights, re-estimating the quantity of interest, e.g., prevalence, and then estimating the variance as the variability across the reestimated quantities of interest. The modified sample and weights after removal of each PSU are called jackknife replicates. The total number of replicates is $R = \sum_{l=1}^{L+1} m_l$, where $m_{L+1} = C$, i.e., the total number of PSUs and study centers in the survey and cohort. Formally the jackknife variance estimation procedure follows as:

Step 1. Leave out α-th PSU (a survey sample cluster or a cohort study center) in stratum l, with $\alpha = 1, \cdots, m_l$, and $l = 1, \cdots, L+1$. Then weight up the units in remaining PSU's in stratum l by the ratio of the number of PSUs in l to the number of remaining PSUs, i.e., $\frac{m_l}{m_l - 1}$. This weight adjustment factor for unit $r \in s_c \cup^* s_s$ in replicate-$l\alpha, l = 1, \cdots, L + 1$ and $\alpha = 1, \cdots, m_l$ can be written as

$$f_{r(l\alpha)} = \begin{cases} 0, & \text{for unit } r \text{ in stratum } l \text{ cluster } \alpha; \\ \dfrac{m_l}{m_l - 1}, & \text{for unit } r \text{ in stratum } l \text{ cluster } \alpha' \neq \alpha; \\ 1, & \text{otherwise.} \end{cases}$$

Step 2. Refit Model (2.1.1) in the main text with weights of $f_{r(l\alpha)}$, and then re-estimate the PS for each unit in the replicate-$l\alpha$ sample, denoted by $\hat{\hat{p}}_i^{(c)}$ and $\hat{\hat{p}}_j^{(s)}$ for cohort unit i and survey sample unit j.

Step 3. Compute pseudo-weights. The smoothed kernel weight for cohort unit i borrowed from survey unit j is

$$k_{ij(l\alpha)} = \frac{K\left\{\left(\hat{\hat{p}}_j^{(s)} - \hat{\hat{p}}_i^{(c)}\right)/h\right\}}{\sum_{i \in s_{c(l\alpha)}} K\left\{\left(\hat{\hat{p}}_j^{(s)} - \hat{\hat{p}}_i^{(c)}\right)/h\right\}}, \text{for } i \in s_{c(l\alpha)}; j \in s_{s(l\alpha)}$$

where the bandwidth h is the same as obtained from the original combined sample (Korn & Graubard, 1999 page 89); $s_{s(l\alpha)}$ and $s_{c(l\alpha)}$ denote the cohort and survey sample in replicate-$l\alpha$, respectively. The KW pseudo-weight for cohort unit i in replicate-$l\alpha$ is

$$w_{i(l\alpha)}^{KW} = \sum\nolimits_{j \in s_{s(l\alpha)}} k_{ij(l\alpha)} \cdot d_j \cdot f_{j(l\alpha)}, \qquad \text{for } i \in s_{c(l\alpha)}.$$

Step 4. Re-estimate the population mean/prevalence as

$$\hat{\mu}_{(l\alpha)}^{KW} = \left(\sum\nolimits_{i \in s_{c(l\alpha)}} w_{i(l\alpha)}^{KW}\right)^{-1} \cdot \sum\nolimits_{i \in s_{c(l\alpha)}} w_{i(l\alpha)}^{KW} \cdot y_i.$$

The JK variance estimator for $\hat{\bar{Y}}^{KW}$ is

$$var(\hat{\mu}^{KW}) = \sum_{l=1}^{L+1} \frac{m_l - 1}{m_l} \sum_{\alpha=1}^{m_l} \left(\hat{\mu}_{(l\alpha)}^{KW} - \hat{\mu}^{KW}\right)^2.$$

The PSAS and IPSW JK variance estimators are calculated similarly as described above, but differ at Steps 2 and 3. At Step 2, the IPSW method estimates propensity scores with weights of $f_{j(l\alpha)} d_j$ for each survey unit j. At Step 3, the PSAS method creates pseudo-weights by partitioning the replicate-$l\alpha$ sample into quintiles of predicted propensity scores, and then dividing the sum of survey replicate weights (i.e., $\sum_{j \in sub_g} f_{j(l\alpha)} d_j$, where sub_g is the g-th subclass, $g = 1, \cdots, G$) by the number of cohort units in each quintile. At Step 3, the IPSW method uses the inverse of predicted odds as the pseudo-weights.

3.3 Simulations

3.3.1 Finite Population Generation

A finite population of $M = 3{,}000$ clusters with each cluster composed of $3{,}000$ units was generated (population total $N = 9{,}000{,}000$). The 2015 one-year estimates at county level from the American Community Survey (ACS) were used to generate the finite population of clusters of individuals. For example, the four-category race/ethnicity (Non-Hispanic White, Non-Hispanic Black, Other Non-Hispanic, and Hispanic) from the ACS had the weighted proportions of $o_k^{(\alpha)}, k = 1, \cdots, 4$ for the α-th county, $\alpha = 1, \cdots, M$. Accordingly, the race/ethnicity for individuals in α-th cluster in the simulated finite population is generated by a multinomial distribution with parameter $o^{(\alpha)} = \left(o_1^{(\alpha)}, o_2^{(\alpha)}, o_3^{(\alpha)}, o_4^{(\alpha)}\right)$. The other variables generated from the ACS estimates included age, using a normal distribution with cluster specified mean and variance sex (*sex*), household income level (hh_inc), and urban/rural area (*urb*). We further generated a continuous environmental factor $Env \sim \min\{4.5, \mathrm{LogNormal}(\lambda_\alpha,\ 0.5)\}$, where $\lambda_\alpha \sim \mathrm{Uniform}(0, 0.5)$, for $\alpha = 1, \cdots, M$, resulting in an intra-class correlation (ICC) within the clusters of 0.054 for the finite population.

The disease status y (1 for presence and 0 for absence) was generated to have an ICC within the clusters of 0.07 for the finite population, with the probability of disease generated by $\mu_0 = \mathrm{expit}(\boldsymbol{\theta v})$ (Hunsberger et al., 2008; Oman & Zucker, 2001). The parameter $\boldsymbol{\theta} = (-5, 0.5, -1, 1, 0.3, 0.10)^T$ where the intercept was -5, and the variables in vector $\boldsymbol{v}$ included age level (=1 if 10-19yrs; =2 if 20-29yrs; =3 if 30-39yrs; =4 if 40-49yrs; =5 if 50-59yrs; =6 if >=60 yrs), sex (1 = male and 0 = female), Hispanic (1=Hispanic and 0= otherwise), Env, and an interaction between age and Env. The disease prevalence in the population was 9.59%. A substitute of μ_0 was generated by $z = \mu_0 + e$, with $e \sim \mathrm{Normal}(0, 0.085^2)$ in the finite population to reflect situations occurs in real data when μ is not available but related variables are available. The correlation between z and y was $\rho = 0.30$.

3.3.2 Sampling from the Finite Population to Assemble Survey Sample and Cohort

We conducted two-stage cluster sampling to select the cohort and the survey sample independently to ensure that the true propensity models for all three methods (IPSW, PSAS, and KW) had the same functional form. This sample design enabled us to form a fair comparison among the three methods because each of them would achieve the greatest bias reduction under the same true propensity model.

A survey sample of $n_s = 1{,}500$ individuals (150 clusters of each 10 individuals) was selected by two-stage cluster sampling. At the first stage, 150 clusters were sampled by probability proportional to size (PPS) sampling, with the measure of size (MOS) for $i \in FP$ defined by

$$\sum\nolimits_{i \in u_\alpha} r_i^b,$$

where u_α is the set of individuals from the α-th cluster for $\alpha = 1, \cdots, M$; $b \in \mathbb{R}_{>0}$; and

$$r_i = \exp(\gamma_0 + \boldsymbol{\gamma}^T \boldsymbol{x}_i), \qquad (3.3.1)$$

where $\gamma_0 = 0$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)^T = (0.3, -0.4, 0.7, 0.7)^T$, and the vector of covariates $\boldsymbol{x}_i$ included $x_{i,1} = age$, $x_{i,2} = hh\_inc$, $x_{i,3} = Env$, and $x_{i,4} = z$. At the second stage, 10 individuals were selected by PPS sampling within each sampled cluster with MOS of r_i^b. The final sampling weight (i.e., the reciprocal of the selection probability, $\frac{1}{\pi_i^{(s)}}$) for $i \in FP$ was $\frac{\sum_{i \in FP} r_i^b}{n_s \cdot r_i^b}$. Using this MOS implies that clusters and individuals with larger values of r_i (older people who have lower household income, higher environmental exposure, or larger probability of having disease) were sampled at higher rates to form the survey sample.

A cohort sample of size $n_c = 11{,}250$ people (75 clusters of each 150 individuals) was sampled independently using a similar two-stage PPS design but with different MOSs in the PPS sampling at stages one and two, given as $\sum_{i \in u_\alpha} r_i^a$ and r_i^a, respectively, $a \in \mathbb{R}_{<0}$. As such, clusters and individuals with smaller r_i were sampled at higher rates in the cohort. Table 3.1 below describes the two-stage PPS cluster sampling for both cohort and survey sample selection.

Table 3.1 Two-stage PPS cluster sampling applied in the simulations.

Sample	**Design**	**Measure of Size† (MOS)**	**Inclusion Probability**
Cohort	Stage 1- clusters selected by PPS	$\sum_{i\in u_\alpha} r_i^a$	$\pi_i^{(c)} = \frac{n_c \cdot r_i^a}{\sum_{i\in FP} r_i^a}$
	Stage 2- subjects selected by PPS	$r_i^a, i \in u_\alpha$	
Survey	Stage 1- clusters selected by PPS	$\sum_{i\in u_\alpha} r_i^b$	$\pi_i^{(s)} = \frac{n_s \cdot r_i^b}{\sum_{i\in FP} r_i^b}$
	Stage 2- subjects selected by PPS	$r_i^b, i \in u_\alpha$	

†u_α is the set of individuals from α-th cluster ($\alpha = 1, \cdots, M$).

Under the two-stage PPS sampling described above, the true propensity models for $p_i = P(\, i \in s_c \mid i \in s_c \cup^* FP\,)$ used by the IPSW method, and $\tilde{p}_i = P(\, i \in s_c \mid i \in s_c \cup^* s_s\,)$ used by the PSAS and KW methods were

$$\begin{aligned} \operatorname{logit}\{p_i\} &= \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{x}_i, \text{and} \\ \operatorname{logit}\{\tilde{p}_i\} &= \tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}_1^T \boldsymbol{x}_i \end{aligned} \qquad (3.3.2)$$

respectively, where β_0 and $\tilde{\beta}_0$ were the intercepts, $\boldsymbol{\beta}_1 = a \cdot \boldsymbol{\gamma}$, and $\tilde{\boldsymbol{\beta}}_1 = (a - b) \cdot \boldsymbol{\gamma}$ (proof in Section 3.6.4). The consistent estimates of the regression coefficients can be obtained by fitting the two models to the combined (s_c vs. *weighted* s_s) sample (used by the IPSW method), and to the combined (s_c vs. *unweighted* s_s) sample, respectively (proof in Section 3.6.5). Note that both models had the same functional form of the covariates $\boldsymbol{x}$. This ensures that, the IPSW and KW methods would result in unbiased estimation under the true PS models with the same functional form. Otherwise, a simulation could result in unbiased estimation from one method, but not the other.

The constants a and b are real numbers that control the difference of the covariate distributions between the cohort and the survey. We let $a \cdot b \leq 0$ so that the cohort and survey oversample people with different characteristics, which generally is what occurs in real data. For example, population-based surveys tend to oversample minority subpopulations such as Hispanics, but minorities tend to be grossly under-represented in cohort studies. According to the cohort and survey sample selection probabilities in Table 3.1, when $a = -1$ and $b = 1$, population units with larger values of r_i^{-1} (or r_i) tend to be oversampled in the cohort (or survey). The larger $|a - b|$ is, the more different the covariate distributions are between the cohort and the survey. Results with $|a - b| = 1.5$ ($a = -1$ and $b = 0.5$) are presented in Figure 3.1, Table 3.2, and Table 3.4. The results under an extreme case $|a - b| = 3.7$ are presented in Table 3.3.

3.3.3 Evaluation Criteria

We compared the KW estimates ($\hat{\mu}^{KW}$) of the population disease prevalence (μ^{FP}), with 1) the survey estimates ($\hat{\mu}^{SVY}$), which were approximately unbiased, 2) the unweighted naïve cohort estimates ($\hat{\mu}^{Naive}$) ignoring the sample designs, 3) the IPSW estimates ($\hat{\mu}^{IPSW}$), and 4) the PSAS estimates ($\hat{\mu}^{PSAS}$). The IPSW method used the inverse of estimated odds as the pseudo-weights. The PSAS method used quintiles of estimated propensity scores to form subclasses. For the KW method, the kernel was the symmetric triangular density on (-3, 3) with the bandwidth selected by Silverman's Rule (see Section 3.6.3); other kernel functions and bandwidths performed similarly (see Section 3.3.6).

We used relative bias (%RB), empirical variance (V), mean squared error (MSE) of the estimators, defined by %RB $= \frac{1}{B}\sum_{b=1}^{B}\frac{\widehat{\mu}^{(b)}-\mu^{FP}}{\mu^{FP}}100\%$, V $= \frac{1}{B-1}\sum_{b=1}^{B}\Big\{\hat{\mu}^{(b)} -$

$\frac{1}{B}\sum_{b=1}^{B}\hat{\mu}^{(b)}\Big\}^2$, and MSE $= \frac{1}{B}\sum_{b=1}^{B}\{\hat{\mu}^{(b)} - \mu^{FP}\}^2$, respectively, to evaluate the performance of the prevalence estimators, where $B = 1{,}000$ is the number of simulations, $\hat{\mu}^{(b)}$ is the estimate of the prevalence obtained from the b-th simulated samples.

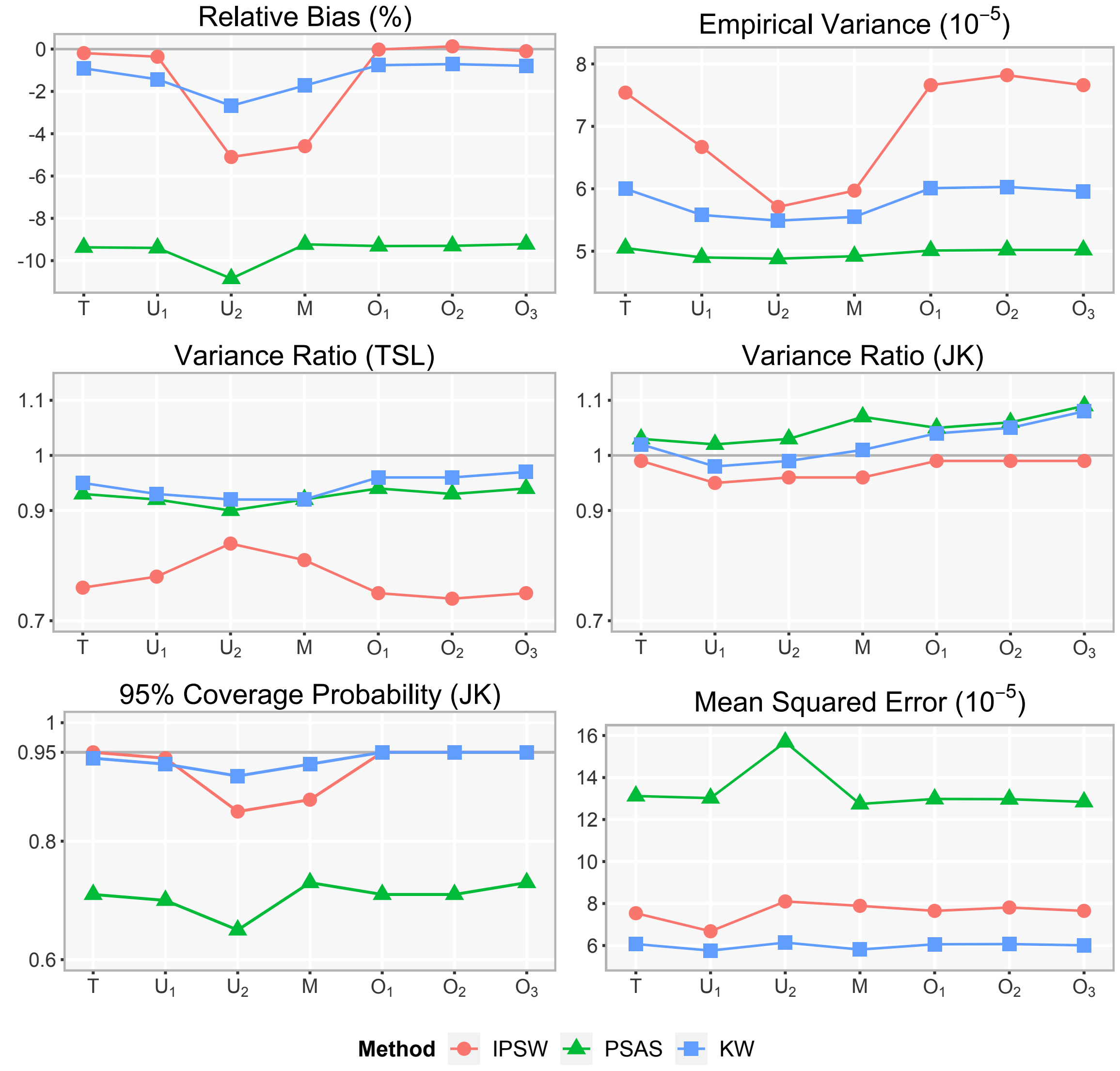
For each mean estimator, two variance estimators were evaluated, i.e., the naive TL estimator and the JK estimator using the variance ratio (VR), and coverage probabilities (CP) of the corresponding 95% confidence intervals, defined by $VR = \frac{\frac{1}{B}\sum_{b=1}^{B}\hat{v}^{(b)}}{V}$, and $CP = \frac{1}{B}\sum_{b=1}^{B} I\left(\mu^{FP} \in CI^{(b)}\right)$ respectively, where $\hat{v}^{(b)}$ is the variance estimate of $\hat{\mu}^{(b)}$, and $CI^{(b)} = \left(\hat{\mu}^{(b)} - 1.96\sqrt{\hat{v}^{(b)}},\ \hat{\mu}^{(b)} + 1.96\sqrt{\hat{v}^{(b)}}\right)$ is the 95% confidence interval from the b-th simulated samples.

3.3.4 Results under Correctly Specified and Six Misspecified Propensity Models

The naïve cohort prevalence, $\hat{\mu}^{Naive}$, was biased by -42.48% (Table 3.2). Figure 3.1 shows the results under the correctly specified propensity model (Model T) and six misspecified models.

The KW estimates, $\hat{\mu}^{KW}$, tended to have the smallest mean squared error and maintained the nominal coverage probability the best. Although IPSW removed slightly more bias than KW when all variables correlated with both sample selection and the outcome y were included in the model, the estimates were much more variable. The bias reduction and variance of $\hat{\mu}^{IPSW}$ were sensitive to propensity model specification. The PSAS estimates, $\hat{\mu}^{PSAS}$, had the smallest variance, but also the smallest bias reduction. The JK variance estimates were approximately unbiased for all three methods. The naïve TL method underestimated the variances of the IPSW estimates by 16%-26%, and the variances of $\hat{\mu}^{PSAS}$ or $\hat{\mu}^{KW}$ by <10% (Table 3.2).

Figure 3.1 Simulation results from 1,000 simulated cohorts and survey samples with each cohort and survey sample fitted to the correct propensity model and six misspecified propensity models†.



†The labels of the x-axises represent the propensity models as follows
Model T: true model. $\text{logit}(p) \sim age, hh\_inc, Env, z.$
Model U_1: underfitted model $\text{logit}(p) \sim age, Env, z;$
Model U_2: underfitted model. $\text{logit}(p) \sim ag, Env;$
Model M: misspecified model $\text{logit}(p) \sim age, Env, Hisp, sex;$
Model O_1: overfitted model $\text{logit}(p) \sim age, hh\_inc, Env, z, Hisp;$
Model O_2: overfitted model $\text{logit}(p) \sim age, hh\_inc, Env, z, Hisp, sex;$
Model O_3: overfitted model $\text{logit}(p) \sim age, hh\_inc, Env, z, urb.$

Models U_1 and U_2 were incorrectly under-fitted: Model U_1 did not include hh_inc that was uncorrelated with disease status y, while Model U_2 also excluded z that was highly predictive of y. The bias of three pseudo-weighted estimates under Model U_1 were all close to the bias under Model T (the true model). However, the empirical variance of $\hat{\mu}^{IPSW}$ was dramatically reduced because the missing variable hh_inc was uncorrelated with the outcome y (similar to the findings in Stuart, 2010). Also, the empirical variances of $\hat{\mu}^{PSAS}$, and $\hat{\mu}^{KW}$ were slightly smaller under Model U_1 than the variances under Model T. In contrast, under Model U_2 with missing z, all three estimates had higher biases but smaller variances, especially $\hat{\mu}^{IPSW}$.

Model M did not include the highly predictive variable z along with hh_inc that were in Model T, but added two extra variables, being Hispanic and sex, which were predictive of y. Comparing results under Models U_2 and M, we observed that adding additional predictors of the outcome y in the under-fitted model reduced, but did not eliminate, the bias. Adding these extra variables increased the variance of $\hat{\mu}^{IPSW}$ but did not affect the variances of $\hat{\mu}^{PSAS}$ and $\hat{\mu}^{KW}$.

Models O_1 , O_2 , and O_3 were incorrectly over-fitted, including unnecessary variables. Model O_1 and O_2 had one (being Hispanic) and two (sex and being Hispanic) additional predictors of y, respectively. Model O_3 included on extra variable (urb) unrelated to y. Under these three models, the bias reduction was similar for all three estimates compared to the bias reduction under the true model respectively. However, adding extra variables resulted in higher variance of $\hat{\mu}^{IPSW}$. Though the variances of $\hat{\mu}^{PSAS}$ and $\hat{\mu}^{KW}$. did not increase, the JK variance estimates were slightly inflated when the propensity model included covariate(s) unrelated to the propensity modeling or the outcome variable.

Table 3.2 Simulation results from 1,000 simulated cohorts and survey samples with each cohort and survey sample fitted to the correct propensity score model and six misspecified propensity score models.

Model	**Estimate** ($h \times 10^3$)	**%RB**	**V** ($\times 10^5$)	**VR** (TL)	**VR** (JK)	**CP** (JK)	**MSE** ($\times 10^5$)
	$\hat{\mu}^{Naive}$	-42.48	2.39	0.19	NA	NA	168.45
	$\hat{\mu}^{SVY}$	-0.11	6.42	1.06	1.06	0.96	6.42
Model T (True model): $\text{logit}(p) \sim age, hh\_inc, Env, z$							
	$\hat{\mu}^{IPSW}$	-0.19	7.54	0.76	0.99	0.95	7.54
	$\hat{\mu}^{PSAS}$	-9.37	5.05	0.93	1.03	0.71	13.12
	$\hat{\mu}^{KW}$(9.43)	-0.91	6.00	0.95	1.02	0.94	6.07
Model U_1 (Underfitted model 1): $\text{logit}(p) \sim age, Env, z$							
	$\hat{\mu}^{IPSW}$	-0.36	6.67	0.78	0.95	0.94	6.68
	$\hat{\mu}^{PSAS}$	-9.40	4.90	0.92	1.02	0.70	13.02
	$\hat{\mu}^{KW}$(9.90)	-1.43	5.58	0.93	0.98	0.93	5.76
Model U_2 (Underfitted model 2): $\text{logit}(p) \sim age, Env$							
	$\hat{\mu}^{IPSW}$	-5.10	5.71	0.84	0.96	0.85	8.10
	$\hat{\mu}^{PSAS}$	-10.85	4.88	0.90	1.03	0.65	15.69
	$\hat{\mu}^{KW}$(10.85)	-2.68	5.49	0.92	0.99	0.91	6.14
Model M (Underfitted + Overfitted model): $\text{logit}(p) \sim age, Env, Hisp, sex$							
	$\hat{\mu}^{IPSW}$	-4.59	5.96	0.81	0.96	0.87	7.89
	$\hat{\mu}^{PSAS}$	-9.23	4.91	0.92	1.07	0.73	12.74
	$\hat{\mu}^{KW}$(10.14)	-1.72	5.54	0.92	1.01	0.93	5.81
Model O_1 (Overfitted model 1): $\text{logit}(p) \sim age, hh\_inc, Env, z, Hisp$							
	$\hat{\mu}^{IPSW}$	-0.02	7.66	0.75	0.99	0.95	7.65
	$\hat{\mu}^{PSAS}$	-9.31	5.01	0.94	1.05	0.71	12.98
	$\hat{\mu}^{KW}$(9.42)	-0.76	6.01	0.96	1.04	0.95	6.06
Model O_2 (Overfitted model 2): $\text{logit}\{\Pr(\boldsymbol{x})\} \sim age, hh\_inc, Env, z, Hisp, sex$							
	$\hat{\mu}^{IPSW}$	0.13	7.82	0.74	0.99	0.95	7.81
	$\hat{\mu}^{PSAS}$	-9.30	5.02	0.93	1.06	0.71	12.97
	$\hat{\mu}^{KW}$(9.41)	-0.71	6.03	0.96	1.05	0.95	6.07
Model O_3 (Overfitted model 3): $\text{logit}\{\Pr(\boldsymbol{x})\} \sim age, hh\_inc, Env, z, urb$							
	$\hat{\mu}^{IPSW}$	-0.10	7.66	0.75	0.99	0.95	7.65
	$\hat{\mu}^{PSAS}$	-9.22	5.02	0.94	1.09	0.73	12.84
	$\hat{\mu}^{KW}$(9.38)	-0.79	5.96	0.97	1.08	0.95	6.01

We also considered the situation where the propensity model includes the variable of interest, y as the single predictor, though it usually does not occur in practice. The KW and the IPSW methods can exactly the same estimator as the survey sample estimator regardless the correct propensity model (proof in Section 3.6.6).

3.3.5 Results under Extreme Selection Probabilities

As noted in Section3.3.2, we changed values of a and b to -2.5 and 1.2, respectively so that the cohort was an extremely non-representative sample of the finite population. Some of the selection probabilities for the cohort sample were close to zero due to extremely small MOS of r_i^a. For example, the minimum selection probability was as small as 7.44×10^{-12}, corresponding to an extremely large weight. Such large weights increased the variability of the pseudo-weighted estimates (Table 3.3). As a result, $\hat{\mu}^{IPSW}$ had an inflated variance, and the largest MSE among the three pseudo-weighted estimates. In contrast, $\hat{\mu}^{KW}$ had much smaller MSE than the others. The variances for all three estimates were overestimated by the JK method due to small sample bias that was likely induced mainly by highly variable weights.

Table 3.3 Simulation results from 1,000 simulated cohorts and survey samples with the true propensity model fitted to each cohort and survey sample under extreme selection probabilities.

Method	**%RB**	$\boldsymbol{V}$ $(\times 10^{-5})$	**MSE** $(\times 10^{-5})$	**VR** (TL)	**VR** (JK)	**CP** (JK)
$\hat{\mu}^{Naive}$	-71.02	1.17	465.26	0.21	NA	NA
$\hat{\mu}^{SVY}$	-0.69	8.36	8.40	1.06	1.06	0.95
$\hat{\mu}^{IPSW}$	7.60	392.29	397.24	0.29	1.56	0.88
$\hat{\mu}^{PSAS}$	-35.16	5.88	119.60	0.90	1.78	0.09
$\hat{\mu}^{KW}$	-6.85	33.31	37.59	0.97	2.40	0.96

3.3.6 Choice of Kernel and Bandwidth

We compared $\hat{\mu}^{KW}$ using two kernel functions: (1) a standard normal density kernel, and (2) a truncated triangular density kernel with support on (-3, 3). For either kernel function, the bandwidth was selected assuming a standard normal density kernel function using five methods: Silverman method (Silverman, 1986), Scott method (Scott, 1992), unbiased cross-validation (Scott & Terrell, 1987), biased cross-validation (Scott & Terrell, 1987), and Sheather & Jones' method (Sheather & Jones, 1991).

Table 3.4 Simulation results from 1,000 simulated cohorts and survey samples, comparing effects of two kernel functions and five bandwidth selection methods†

Kernel	**Bandwidth**	**%RB**	**V**	**VR**	**VR**	**CP**	**MSE**
Function	(Method)		$(\times 10^{-5})$	(TL)	(JK)	JK	$(\times 10^{-5})$
	$\hat{\mu}^{Naive}$	-42.48	2.39	0.19	--	--	168.45
	$\hat{\mu}^{SVY}$	-0.11	6.42	1.06	1.06	0.96	6.42
$N(0,1)$							
	0.00987 (Silv)	-0.26	6.18	0.96	1.08	0.95	6.18
	0.01162 (Scott)	-0.42	6.13	0.96	1.05	0.95	6.14
	0.00188 (UCV)	0.05	6.61	0.96	4.67	1.00	6.61
	0.00775 (BCV)	-0.14	6.28	0.95	1.23	0.96	6.28
	0.00149 (S&J)	0.03	6.70	0.96	7.60	1.00	6.79
$T(-3,3,0)$							
	0.00987 (Silv)^{j}	-0.75	6.02	0.96	1.04	0.95	6.07
	0.01162 (Scott)	-0.94	5.99	0.96	1.02	0.95	6.06
	0.00188 (UCV)	-2.00	6.22	0.93	3.55	1.00	6.58
	0.00775 (BCV)	-0.70	6.08	0.95	1.14	0.96	6.12
	0.00149 (S&J)	-2.52	6.23	0.93	5.74	1.00	6.81

†The fitted propensity model is $\text{logit}\{\Pr(\boldsymbol{x})\} \sim age, hh\_inc, Env, z, Hisp, sex$, which includes two extra covariates $Hisp$ and sex compared to the true model. The bandwidth selection methods include Silverman's rule of thumb (Silv), Scott's method (Scott), unbiased cross validation (UCV), biased cross validation (BCV), and S&J's method (S&J). Notice: these results are slightly different from those in Table 3.2 under Model O2 because the bandwidths were different.

The results of bandwidth selection in Table 3.4 were consistent with the existing literature (Terrell & Scott, 1985; Jones et al. 1996): the Silverman's and Scott's methods tend to give larger bandwidths than the other three. Based on the simulation results, either of these two methods is recommended because the other methods tend to result in smaller bandwidths that increase the empirical variance and inflate the JK variance estimation due to more variable pseudo-weights across replicates. With the same bandwidth, it is observed that the standard normal density kernel, compared to the triangular density kernel, resulted in smaller bias but larger variances of $\hat{\mu}^{KW}$. This is because the standard normal density kernel uses more extreme values for the distances than the truncated triangular density kernel.

Hence, the combination of triangular density kernel and Silverman's bandwidth appears to behave the best with regard to its overall MSE reduction.

3.4 Data Analysis: The NIH-AARP Cohort Study

We estimated (1) prevalence of eight self-reported diseases, (2) prospective nine-year rates of all-cause mortality and (3) all-cancer mortality for people aged 50 to 71 using the US National Institutes of Health and the American Association of Retired Persons (NIH-AARP) Diet and Health Study. These prevalences and mortalities were also available in the US National Health Interview Survey (NHIS), serving as the gold standard that allowed us to examine how much bias in the NIH-AARP estimates can be corrected by the pseudo-weighting methods in practice.

The NIH-AARP cohort recruited 567,169 AARP members from 1995-1996, aged 50 to 71 years, who resided in California, Florida, Pennsylvania, New Jersey, North Carolina, or Louisiana, or in metropolitan Atlanta, Georgia, and Detroit (NIH-AARP, 2006) in the US. The NIH-AARP cohort is linked with Social Security Administration Death Match File and National Death Index (NDI) (NCHS 2013) by standard record linkage methods up to 2011 (NIH-AARP 2006), providing mortality and cause of death ascertainment. AARP members were mailed questionnaires, but only 17.6% returned questionnaires, raising further questions about the representativeness of the NIH-AARP cohort for the US population.

For the reference survey, we used the NHIS, a cross-sectional household interview survey of the civilian noninstitutionalized US population. To make the two samples comparable, we chose the contemporaneous 1997 NHIS respondents aged 50 to 71 years (9,306 participants). The 1997 NHIS has a multistage stratified cluster sample design (see

Section 2.2.2) with 339 strata of each consisting of two sampled PSUs (NCHS, 2000). NHIS was also linked to NDI through 2006 for mortality information (NCHS 2009). All the links were treated as true and no linkage error were considered in this analysis.

After harmonizing variables between NIH-AARP and NHIS, the distribution of common variables and variables of interests are described in Table 3.5 and Table 3.6 respectively.

Table 3.5 Distribution of selected common variables in NIH-AARP and NHIS

	NIH-AARP (1995-96)		**NHIS (1997)**			
	n	%	n	%	weighted n	weighted %
Total	529708	100	9306	100	49761895	100
DEMOGRAPHIC						
Age in years						
50-54	69207	13.07	2637	28.34	15064732	30.27
55-59	117417	22.17	2091	22.47	11480359	23.07
60-64	148726	28.08	1861	20.00	9995586	20.09
65-69	174567	32.96	1944	20.89	9474745	19.04
70-71	19791	3.74	773	8.31	3746473	7.53
Gender						
Male	314269	59.33	4059	43.62	23528092	47.28
Female	215439	40.67	5247	56.38	26233803	52.72
Race						
NH-White	485486	91.65	6693	71.92	39565812	79.51
NH-Black	19576	3.70	1249	13.42	4758442	9.56
Hispanic	9628	1.82	1055	11.34	3468003	6.97
NH-Other	15018	2.84	309	3.32	1969638	3.96
Marital Status						
Married or living as married	366327	69.16	5381	57.82	35937686	72.61
Widowed	58296	11.01	1365	14.67	4765959	9.58
Divorced or Separated	79545	15.02	1919	20.62	5613727	13.26
Never married	25540	4.82	641	6.89	2267497	4.56
SOCIOECONOMIC STATUS						
Education						
High school or less	200498	37.85	5382	57.83	27564686	55.39
Post-high school/some college	123325	23.28	2052	22.05	11440010	22.99
College graduate/postgraduate	205885	38.87	1872	20.12	10757199	21.62
HEALTH BEHAVIOR						
BMI						
<18.5	4233	0.80	130	1.40	654914	1.32
18.5-24.9	182946	34.54	3208	34.47	17143743	34.45
>=25	342529	64.66	5968	64.13	31963238	64.23
Smoking (quit years or dose)						
Never	184416	34.81	4026	43.26	21264038	42.73
Former, quit>=10 years	213657	40.33	2235	24.02	12747525	25.62

Former, quit<10 years	69108	13.05	935	10.05	4926262	9.90
Current, <=1 pack/day	40396	7.63	1644	17.67	8215497	16.51
Current, >1 pack/day	22131	4.18	466	5.01	2608573	5.24
Physical Activity						
<3 times/week	286822	54.15	7775	83.55	40930891	82.25
>=3 times/week	242886	45.85	1531	16.45	8831004	17.75
Health Status (Self-reported)						
Excellent	87439	16.51	1837	19.74	10954418	22.04
Very good	191114	36.08	2578	27.70	14943138	30.06
Good	182621	34.48	2664	28.63	14738240	29.65
Fair	58741	11.09	1273	13.68	6597770	13.27
Poor	9793	1.85	540	5.80	2471456	4.97

Of note is the importance of self-reported health status, a variable often excluded in epidemiologic analyses as being a proxy for disease, but which turns out to be strongly predictive of the propensity to be selected in NIH-AARP versus NHIS. This is expected because cohorts often recruit healthier people (Pinsky et al., 2007; Fry et al., 2017).

We used a stepwise procedure based on the AIC to choose the propensity model fitted to the combination of the NIH-AARP cohort and *unweighted* NHIS sample, which initially included all main effects of five common demographic characteristics (age, sex

Table 3.6 Distribution of self-reported diseases at baseline and nine-year mortality in NIH-AARP and NHIS

	NIH-AARP (1995-96)		**NHIS (1997)**			
	n	%	n	%	weighted n	weighted %
Total	529708		9306		49761895	
Self-Reported Diseases						
Diabetes	48471	9.15	1064	11.43	5215661	10.48
Emphysema	14530	2.74	325	3.49	1794778	3.61
Stroke	11272	2.13	377	4.05	1879697	3.78
Heart Disease	74532	14.07	660	7.09	3608156	7.25
Stroke or Heart Disease	81468	15.38	930	9.99	4920432	9.89
Colon Cancer	4797	0.91	67	0.72	344287	0.69
Breast Cancer (Female)	10285	4.77	187	3.56	903296	3.44
Prostate Cancer (Male)	10154	3.23	83	2.04	493470	2.10
Nine-Year Mortality						
All-Cause Mortality						
Overall	65732	12.41	1324	14.89	6794116	13.67
Age 50-54	2863	4.85	167	6.65	945836	6.27
Age 55-59	8226	7.19	215	10.73	1116271	9.71
Age 60-64	16489	11.39	296	16.55	1563759	15.66
Age 65-72	38154	18.04	646	24.97	3168250	24.09

All-Cancer Mortality						
Overall	42458	8.02	499	5.61	2688875	5.41
Age 50-54	2366	4.01	72	2.87	61728	2.83
Age 55-59	6607	5.77	79	3.94	56952	3.92
Age 60-64	12181	8.41	119	6.65	67972	6.80
Age 65-72	24641	11.65	229	8.85	81622	8.61
Male	29775	9.47	267	6.82	1540510	6.56
Female	16020	7.44	232	4.66	1148365	4.38
Age 50-54, male	1409	4.27	39	3.38	254351	3.47
Age 55-59, male	4072	6.12	48	5.36	286824	5.36
Age 60-64, male	7758	9.10	62	7.52	362227	7.41
Age 65-72, male	16536	12.77	118	11.32	637110	10.78
Age 50-54, female	957	3.68	33	2.43	172198	2.23
Age 55-59, female	2535	5.29	31	2.80	164006	2.67
Age 60-64, female	4423	7.43	57	5.91	316953	6.22
Age 65-72, female	8105	9.89	111	7.18	495208	6.84

race/ethnicity, etc.), three lifestyle factors (smoking status, physical activities, and body mass index [body weight (kg)/height (m) squared]), self-reported health status, and 31 two-way interactions. Table 3.7 below shows the final model estimated by fitting the propensity model with (for IPSW) and without (for PSAS and KW) NHIS sample weights. Note that all the following analyses used the model described in Table 3.7.

Table 3.7 Results for Main effects of the fitted propensity model with ($\hat{\boldsymbol{\beta}}$) or without ($\hat{\hat{\boldsymbol{\beta}}}$) NHIS sample weights†

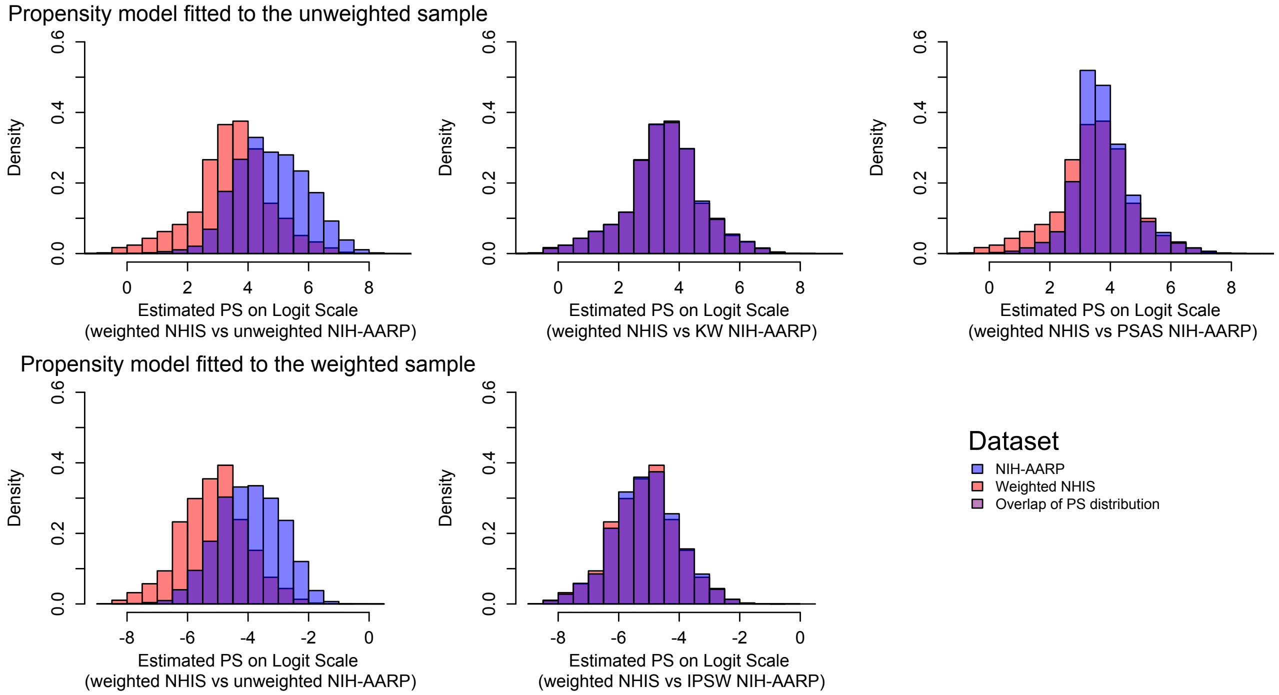
	$\hat{\hat{\boldsymbol{\beta}}}$ (unweighted sample)		$\hat{\boldsymbol{\beta}}$ (weighted sample)	
Coefficients:	Estimate	Std. Error	Estimate	Std. Error
Age	-0.045	0.016**	-0.004	0.022
Sex (ref: male)				
Female	-0.69	0.18***	-1.28	0.28***
Race/Ethnicity (ref: NH-White)				
NH-Black	-2.22	0.45***	-1.61	0.82*
Hispanic	-6.20	0.51***	-3.06	1.06**
NH-Other	-5.29	0.801	-2.91	1.19*
Marital Status (ref: married or living as married)				
Widowed	2.35	0.46***	0.47	0.70
Divorced or Separated	-1.08	0.37**	-1.05	0.59.
Never married	-1.02	0.57.	-0.36	0.96
Education level	-0.53	0.20**	-0.31	0.29
BMI	-0.15	0.027***	-0.11	0.043*
Smoking (ref: Never)				
Former, quit>=10 years	1.32	0.3993***	1.48	0.60*
Former, quit<10 years	0.31	0.5496	0.41	0.84
Current, <=1 pack/day	-3.04	0.4655***	-1.46	0.79.

Current, >1 pack/day	-3.75	0.7680**	-2.64	1.06*
Physical Activity (ref: <3 times/week)				
>=3 times/week	-1.62	0.4337***	0.15	0.54
Self-reported health status	0.93	0.1557***	0.51	0.23*

†The 31 pairwise interactions included in the model are age:race/ethnicity, age:marital status, age:education, age:bmi, age:smoking, age:physical activities, age:health status, sex:race/ethnicity, sex:matrital status, sex:education, sex:bmi, sex:smoking, sex:physical activities, sex:health status, race/ethnicity:marital status, race/ethnicity:education, race/ethnicity:smoking, race/ethnicity:physical activities, race/ethnicity:health status, marital status:education, marital status:physical activities, marital status:health status, education:bmi, education:smoking education:physical activities,, education:health status, bmi:smoking, bmi:physical activities, smoking:physical activities, smoking:health status, physical activities:health status. The magnitude of the p-values are represented by '***' p-value< 0.001; '**' p-value< 0.01; '*' p-value< 0.05; '.' p-value< 0.1.

Figure 3.2 plots the distributions of the estimated propensity score on the logit scale in the unweighted NIH-AARP cohort, and the three pseudo-weighted NIH-AARP cohorts by the IPSW, PSAS and KW methods, compared to the sample-weighted NHIS sample. The percentage of overlapped propensity scores in the data from NHIS and NIH-AARP exceeded 99.9%. All three pseudo-weighted distributions of propensity scores were close to the weighted NHIS sample, among which KW was the closest, followed by IPSW with

Figure 3.2 Comparison of Distributions of Estimated Propensity Scores on Logit Scale



some right-skewness, and PSAS with excess kurtosis. Because the KW and PSAS methods fitted a propensity model to the *unweighted* sample, their estimated propensity scores were close to 1 due to the predominance of cohort units in the combined cohort-survey sample. In contrast, the IPSW method used the propensity model to estimate NIH-AARP cohort membership in the combined cohort and the *weighted* NHIS sample (representing the underlying US population), resulting in small propensities and thus large pseudo-weights.

We used the relative difference from the NHIS estimates (μ^{NHIS}): $\%RD = (\hat{\mu} - \mu^{NHIS})/\mu^{NHIS} \cdot 100\%$, and the percent of bias reduction from the naïve NIH-AARP estimates ($\hat{\mu}^{Naive}$): $\%BR = (\hat{\mu}^{Naive} - \hat{\mu})/(\hat{\mu}^{Naive} - \mu^{NHIS}) \cdot 100\%$ to evaluate the performance of the estimators, where $\hat{\mu}$ is one of the IPSW ($\hat{\mu}^{IPSW}$), PSAS ($\hat{\mu}^{PSAS}$), and KW ($\hat{\mu}^{KW}$) estimates.

Table 3.8 shows the results for estimating prevalences eight self-reported diseases. The naïve NIH-AARP cohort disease prevalence estimates, $\hat{\mu}^{AARP}$, were biased on average

Table 3.8 Estimated population prevalences of eight self-reported diseases at baseline using four methods†

		%**RD**				%**BR**		
Self-reported Disease	μ^{NHIS}(%)	$\hat{\mu}^{Naive}$	$\hat{\mu}^{IPSW}$	$\hat{\mu}^{PSAS}$	$\hat{\mu}^{KW}$	$\hat{\mu}^{IPSW}$	$\hat{\mu}^{PSAS}$	$\hat{\mu}^{KW}$
Diabetes	10.48	-12.70	-12.48	-17.64	-8.94	1.74	-38.86	29.61
Emphysema	3.61	-24.03	-29.79	-25.03	-28.25	-23.99	-4.16	-17.55
Stroke	3.78	-43.61	-45.87	-47.05	-42.85	-5.18	-7.89	1.75
Heart Disease	7.25	94.05	45.13	43.41	46.09	52.01	53.84	50.99
Stroke or Heart Disease	9.89	55.54	19.72	18.25	20.84	64.49	67.13	62.48
Breast Cancer (Female)	3.44	38.53	16.19	21.51	17.75	57.98	44.19	53.92
Colon Cancer	0.69	31.52	3.91	5.91	3.88	87.61	81.24	87.69
Prostate Cancer (Male)	2.10	54.00	18.05	11.80	11.50	66.58	78.15	78.70
Average		44.25	23.89	23.83	22.51	46.00	46.16	49.12

†The propensity model included nine main effects of age, sex, race/ethnicity, marital status, education, BMI, smoking, physical activities, and self-reported health status, as well as 31 interactions. The estimates closest to the corresponding NHIS estimates are in bold.

by ~44%, assuming the NHIS estimates as the truth. All three weighting methods removed roughly half the bias across the eight diseases. The KW method removed slightly more bias than the IPSW and PSAS methods, including a ~88% bias reduction for colon cancer, and ~79% bias reduction for prostate cancer. However, for all three methods, there was little bias reduction for stroke, and the bias increased for emphysema, possibly due to lack of covariates predictive of cohort membership, or accuracy of self-reported disease status (e.g., measurement errors).

Because self-reported diseases had potential measurement errors, we also examined nine-year all-cause mortality as it was obtained from linkage of NIH-AARP (and NHIS) to the NDI (Table 3.9). Surprisingly, the naive NIH-AARP estimate of nonage-specific all-cause mortality had only ~9% bias. However, stratifying mortality by age revealed that the NIH-AARP estimates had a ~25% bias in each age group, which was reduced to 18% by KW (26% bias reduction), the most reduction among the three methods. Thus, the all-cause mortality was confounded by the age distribution: NIH-AARP oversampled older people (Table 3.5), which artificially inflated its overall mortality rate and offset the lower age-specific mortality in the cohort.

Table 3.9 Estimated all-cause nine-year mortality (overall, and by age) using three PS-based methods†

		%**RD**				%**BR**		
Age Group	μ^{NHIS}(%)	$\hat{\mu}^{Naive}$	$\hat{\mu}^{IPSW}$	$\hat{\mu}^{PSAS}$	$\hat{\mu}^{KW}$	$\hat{\mu}^{IPSW}$	$\hat{\mu}^{PSAS}$	$\hat{\mu}^{KW}$
Overall	13.67	-9.21	-16.9	-15.37	-15.51	-83.81	-66.91	-68.39
50-54	6.27	-22.64	-19.6	-23.87	-18.00	13.28	-5.44	20.50
55-59	9.71	-26.03	-20.6	-22.22	-18.90	20.77	14.63	27.37
60-64	15.66	-27.28	-22.5	-21.47	-19.95	17.68	21.29	26.87
64+	24.09	-25.09	-18.7	-17.95	-17.07	25.57	28.44	31.95
Average		25.26	20.3	21.38	18.48	19.45	15.36	26.83

†The estimates closest to the corresponding NHIS estimates are in bold.

The results for all-cancer nine-year mortality differed from all-cause mortality (Table 3.10). The KW estimate had lowest bias for the overall all-cancer mortality (30% bias reduction). When stratifying cancer mortality by age, the PSAS method had slightly more bias reduction, and when stratifying by sex, the KW method reduced more bias. When we categorized mortality by age and sex, different weighting methods removed the most bias in different categories without clear patterns, including the naïve NIH-AARP estimates having the least bias in three of the categories. Part of the reason was the small sample sizes of all-cancer deaths by age and sex in the NHIS sample. In addition, cancer mortality was not as well predicted as all-cause mortality from the covariates in the

Table 3.10 Estimated all-cancer nine-year mortality (overall, and by age and/or sex) using three PB-based methods†

		%RD				**%BR**		
Group	μ^{NHIS}(%)	$\hat{\mu}^{Naive}$	$\hat{\mu}^{IPSW}$	$\hat{\mu}^{PSAS}$	$\hat{\mu}^{KW}$	$\hat{\mu}^{IPSW}$	$\hat{\mu}^{PSAS}$	$\hat{\mu}^{KW}$
Overall	5.41	48.25	35.86	35.34	33.98	25.67	26.76	29.57
50-54	2.83	41.76	41.63	32.13	40.87	0.31	23.05	2.13
55-59	3.92	47.11	42.12	40.92	42.56	10.59	13.14	9.65
60-64	6.80	23.69	21.10	20.40	20.63	10.90	13.88	12.88
64+	8.61	35.37	28.48	26.30	26.31	19.48	25.64	25.62
Average		36.98	33.33	29.94	32.59	9.86	19.05	11.86
Male	6.56	44.47	32.13	29.38	29.63	27.76	33.94	33.37
Female	4.38	69.74	42.38	46.14	42.16	39.24	33.85	39.55
Average		57.11	37.25	37.76	35.90	34.77	33.88	37.14
50-54, male	3.47	23.28	25.96	18.96	26.79	-11.48	18.58	-15.05
55-59, male	5.36	14.23	22.24	16.24	22.82	-56.21	-14.09	-60.30
60-64, male	7.41	22.85	30.98	31.37	30.88	-35.57	-37.30	-35.17
64+, male	10.78	18.40	21.70	22.83	21.39	-17.93	-24.08	-16.26
50-54, female	2.23	65.31	67.15	53.00	63.91	-2.82	18.85	2.15
55-59, female	2.67	97.87	77.42	84.58	77.83	20.89	13.58	20.47
60-64, female	6.22	19.41	13.37	13.08	12.91	31.13	32.60	33.50
64+, female	6.84	44.75	34.37	34.08	33.67	23.20	23.84	24.75
Average		38.26	36.65	34.27	36.27	4.22	10.44	5.20

†The estimates closest to the corresponding NHIS estimates are in bold.

propensity model, thereby reducing the effectiveness of bias correction for all three weighting methods.

3.5 Summary

The KW approach, as a PS-based matching method, is proposed to improve external validity of cohort analyses, using a representative survey sample as a reference of the target population. In brief, the KW approach produces a pseudo-weight for each cohort member in three steps: (1) estimate the PS for each unit in the combined cohort and survey sample, (2) fractionally distribute the sample weight of each survey sample unit to all cohort units based on their similarity measure by the kernel smoothed distance in estimated PS, and (3) create the pseudo-weights for the cohort units as the sum of the sample weights they obtained from all the survey sample units. The sum of the cohort pseudo-weights equals the sum of survey weights. The KW method provides a consistent estimator of population mean/prevalence under the true propensity model and some standard assumptions.

Unlike the naïve TL method, our JK variances take account for all sources of variability in creating pseudo-weights. The three PS-based methods (IPSW, PSAS, and KW) were applied to reduce bias in prevalence estimates from the NIH-AARP cohort using the weighted 1997 NHIS sample as the reference. The KW method generally removed more bias than the IPSW or PSAS method, illustrating the potential benefits of the method. In a few cases, small samples or possibly lack of factors predictive of cohort membership and outcome diseases could increase bias, illustrating practical limitations.

In simulations, the KW estimates had smaller mean squared errors and better confidence interval coverages than the IPSW and PSAS estimates under both properly- and mis-specified propensity models that we considered. The IPSW estimates had the lowest bias among the three pseudo-weighted estimates when the propensity model was properly specified. However, the IPSW method tended to produce extreme weights that inflate variances, as noted previously (Stuart, 2010). Furthermore, the bias reduction and variance of the IPSW estimator can be sensitive to propensity model specification. PSAS is a special case of the KW method, with a uniform density kernel function in each subclass of estimated propensity scores that generally oversmoothed the pseudo-weights. Thus, PSAS tends to produce the least variable weights, resulting in the smallest variances, but also the least bias reduction (also noted by Valliant & Dever, 2011).

The naïve TL variances worked well for the KW and PSAS estimates but failed for the IPSW estimates. The naïve TL method substantially underestimated the variance of the IPSW estimates by ignoring variability due to estimating propensity scores. Since the IPSW method fits the propensity model to the combined sample of cohort and *weighted* survey sample, the estimated model coefficients and propensity scores can have large variance due to variable survey sample weights as well as the naturally high variability among cohort weights of 1 and the survey weights (Li et al., 2011). In contrast, the PSAS and KW methods fit a propensity model to the *unweighted* sample, which yields less variable estimates of PSs and more effective pseudo-weighted estimates of finite population means/prevalences. The JK variance estimation is recommended for the IPSW estimates.

For the NIH-AARP cohort, the KW method reduced bias by 49% on average for estimating the prevalences of eight self-reported diseases (3% more than IPSW and PSAS methods). For nine-year nonage-specific all-cause mortality, the naïve cohort estimate had the smallest bias. However, mortality is strongly confounded by age. For age-specific mortality rates, the KW estimates had a greater averaged bias reduction (27%) than the IPSW (19.45%) and PSAS (15.36%) estimates. Thus, the better performance of the naïve cohort estimator for nonage-specific mortality was caused by disproportionately older volunteer recruitment in the NIH-AARP cohort.

For overall nine-year all-cancer mortality, KW reduced bias the most (~30% reduction). But when stratifying on key confounders (age and sex), no one method worked best for all categories, and PSAS had slightly higher averaged bias reduction than the other two methods across the eight age by sex categories. This result could be due to small sample bias (few cancer deaths in each age by sex category of NHIS sample) or the lack of factors predictive of all-cancer mortality in the propensity model.

The KW method was developed to reduce bias when estimating population prevalence of outcome variables available in cohorts but not in surveys, such as novel molecular or genetic risk factors. In our data example, we purposely selected outcome variables available in both cohort and survey, allowing for quantifying the relative bias by assuming the survey estimates as the gold standard. However, survey estimates can vary from the truth due to sampling errors, and non-sampling errors such as undercoverage and nonresponse bias. Unfortunately, there are few examples where a census of reported diseases is available in the United States.

The simulations provide guidance for choosing propensity model predictors, the kernel function, and bandwidth for using the KW method. For the propensity model, Stuart (2010) suggested including all variables that may be associated with treatment assignment and the outcomes to reduce bias, but for small samples, it is useful to prioritize variables related to the outcome to control the variance (Brookhart et al., 2006). The simulations in the dissertation agree that adding extra predictors of the outcome in the propensity model reduces bias, but at a cost of potential increase in variance, especially for the IPSW method. We suggest that the propensity models aim for maximal bias reduction by including all variables distributed differently in the cohort and the survey sample, all significant interaction terms, and all variables predictive of the outcome. Then, to control variance, we found that the triangular kernel effectively removed the influence of extreme imprecisely estimated weights for the KW method. Finally, we found that the Silverman and Scott bandwidth selection methods provided bias reduction yet controlled variance in our simulations.

3.6 Proofs

3.6.1 Proof of Theorem 3.1 (Consistency of the KW Estimate of FP Mean)

Suppose in the superpopulation, variable $(y, p(\boldsymbol{x}))$ has the joint distribution function F. The finite population consists of $(y_1, p_1), \cdots (y_N, p_N)$ with (y_k, p_k) being a realization of a pair of random variable (y, p), and with $(y_1, p_1), \cdots (y_N, p_N)$ being independent and identically distributed (i.i.d) from F. The cohort $(y_1, p_1), \cdots (y_m, p_m)$ and survey sample $(y_1, p_1), \cdots (y_n, p_n)$ are two random samples of the finite population.

Under the conditions (a), (b) and (c), it can be proved by applying **Theorem 2.1** and **3.1** in Noda (1976) that

$$y^* = \frac{\sum_{j \in s_c} K\left(\frac{p - p_j}{h}\right) \cdot y_j}{\sum_{j \in s_c} K\left(\frac{p - p_j}{h}\right)} \xrightarrow{P} E(y|p),$$

and

$$E\{|y^* - E(y|p)|\} \to 0, \qquad (3.6.1)$$

Denote $y_k^* = \frac{\sum_{j\in s_c} K\left(\frac{p_k - p_j}{h}\right)\cdot y_j}{\sum_{j\in s_c} K\left(\frac{p_k - p_j}{h}\right)}$, $\mu_k = E(y|p_k)$ for $k = 1, \cdots, N$ in the finite population. By applying (3.6.1),

$$E(|y_k^* - \mu_k|) \to 0, \qquad (3.6.2)$$

Let $\bar{Y}^* = N^{-1}\sum_{k=1}^{N} y_k^*$ and $\bar{\mu}_k = N^{-1}\sum_{k=1}^{N} \mu_k$ in the finite population, and then it follows $E(|\bar{Y}^* - \bar{\mu}_k|) \xrightarrow{P} 0$ as $N \to \infty$ based on (3.6.2). By Law of Large Numbers, $\bar{\mu}_k \xrightarrow{P} \mu$. Therefore,

$$E|\bar{Y}^* - \mu| \to 0, \qquad (3.6.3)$$

as $N \to \infty$ and $n_c \to \infty$.

Under condition (d), it follows that $Var(\bar{Y}^* - \mu) \to 0$ as $n_c \to \infty$ and $N \to \infty$. Hence, by Chebyshev's inequality we have

$$\bar{Y}^* - \mu \xrightarrow{P} 0. \qquad (3.6.4)$$

Denote the Hajek estimator (Hajek 1971) for $\bar{Y}^*$ as $\hat{\bar{Y}}^* = \frac{1}{\hat{N}^{SVY}}\sum_{i\in s_s} d_i y_i^*$. According to Isaki and Fuller (1982), with condition (e) we have

$$\hat{\bar{Y}}^* = \bar{Y}^* + O_p\left(n_s^{-\frac{1}{2}}\right), \qquad (3.6.5)$$

as $N \to \infty, n_s \to \infty$.

By (3.6.4) and (3.6.5),

$$\hat{\bar{Y}}^* \xrightarrow{P} \mu. \qquad (3.6.6)$$

By the Law of Large Numbers,

$$\mu^{FP} = \mu + O_p\left(N^{-\frac{1}{2}}\right), \qquad (3.6.7)$$

where $\mu^{FP} = \frac{1}{N}\sum_{i\in FP} y_i$ is the finite population mean. (3.6.6) and (3.6.7) together implies

$$\hat{\hat{Y}}^{*} - \mu^{FP} \overset{P}{\rightarrow} 0. \tag{3.6.8}$$

(3.6.3) and (3.6.5) together implies

$$E\left(\hat{\hat{Y}}^{*}\right) \rightarrow \mu, \tag{3.6.9}$$

as $N \rightarrow \infty, n_s \rightarrow \infty$, and $n_c \rightarrow \infty$.

Notice that by applying $\hat{N}^{KW} = \hat{N}^{SVY}$ in Equality (3.2.4), the KW estimator of finite population mean $\hat{\mu}^{KW} = \hat{\hat{Y}}^{*}$ because

$\hat{\hat{Y}}^{*} = \frac{1}{\hat{N}}\sum_{i\in s_s}\{d_i \cdot (\sum_{j\in s_c} k_{ij}y_j)\} = \frac{1}{\hat{N}^{KW}}\sum_{j\in s_c}\{y_j \cdot (\sum_{i\in s_s} k_{ij}d_i)\} = \frac{1}{\hat{N}^{KW}}\sum_{j\in s_c} w_j^{KW} \cdot y_j.$

This completes the proof of **Theorem 3.1**.

3.6.2 PSAS method as a special case of KW method

As described in Section 2.3.3.3, suppose the PSAS method divides, $s = s_c \cup^{*} s_s$, the combined sample of nonprobability cohort and the probability-based survey sample into G subclasses according to the predicted PSs, and $s^{(g)}$ is the combined sample in subclass g, $g = 1, \cdots, G$. The PSAS pseudo-weights for cohort unit $i \in s^{(g)}$ is

$$w_i^{PSAS} = \frac{\sum_{j\in s^{(g)}} d_j}{n_c^{(g)}}, \tag{3.6.10}$$

where the subscript j indicates a unit in the survey sample, and $n_c^{(g)}$ is the number of cohort units in subclass g. Expression (3.6.10) can be written as

$$\sum\nolimits_{j\in s_s} I_{j\in s^{(g)}} \frac{1}{n_c^{(g)}} d_j,$$

where $I_{j\in s^{(g)}}$ is the indicator for subclass membership (1 if survey unit $j \in s^{(g)}$, 0 otherwise). the term $I_{j\in s^{(g)}} \frac{1}{n_c^{(g)}}$ can be treated as the kernel weight k_{ij} in KW pseudo-weight, and it is equivalent to

$$\frac{K\left\{\left(\tilde{p}_j^{(s)} - \tilde{p}_i^{(c)}\right)/h\right\}}{\sum_{i\in s_c} K\left\{\left(\tilde{p}_j^{(s)} - \tilde{p}_i^{(c)}\right)/h\right\}}.$$

where the kernel function $K(\cdot) = \kappa \cdot I_{i\in s^{(g)}} \cdot I_{j\in s^{(g)}}$ with κ being a constant such that $\int K(u)du = 1$. The bandwidth h is decided depending on how the subclasses are formed. The more subclasses are formed, the smaller h will be. However, with a fixed number of subclasses (e.g. $G = 5, 20, 30$), the bandwidth does not satisfy the condition (b), $h \to 0$, but $n_c \cdot h \to \infty$ as $n_c \to \infty$ in **Theorem 3.1**. Hence, the PSAS estimators of finite population means/prevalences are not consistent.

3.6.3 Optimal bandwidth minimizing asymptotic mean integrated squared error

One of the most commonly used optimality criteria for bandwidth selection is the Asymptotic Mean Integrated Squared Error (AMISE) (Silverman, 1986; Scott, 1992; Sheather, 2004). Minimizing AMISE with respect to h gives the optimal bandwidth

$$h_{\text{opt}} = \left(\frac{R(K)}{n\sigma_K^4 R(f'')}\right)^{1/5}, \qquad (3.6.11)$$

where $K(\cdot)$ is the kernel density function, σ_K is the corresponding standard deviation, $R(K) = \int K^2(z)dz$, n is the sample size, and f is the unknown density function to be estimated, with f'' being the second derivative of f. Since f is unknown, $R(f'')$ needs to be estimated. Silverman (1986), and Scott (1992) approximate f by a normal density with the sample estimates $\hat{\mu}$, and $\hat{\sigma}$ used for the mean and standard deviation. After some calculation, it can be shown that $R(f'') = \frac{3}{8}\hat{\sigma}^{-5}/\sqrt{\pi}$.

As shown by the formula (3.6.11)s, the optimal bandwidth h_{opt} will change based on the given kernel function $K(\cdot)$. Here we give two examples of kernel functions: a normal density with mean 0 and standard deviation σ_K, $N(0, \sigma_K)$, and a symmetric triangular density on the support of $(-t, t)$, $T(-t, t, 0)$.

<u>3.6.3.1</u> $N(0, \sigma_K)$

It can be shown that $R(K) = \frac{1}{2\sqrt{\pi}\sigma_K}$. Then the optimal bandwidth is

$$h_{\text{opt}} = \left(\frac{\frac{1}{2\sqrt{\pi}\sigma_K}}{n\sigma_K^4 \cdot \frac{3}{8}\hat{\sigma}^{-5}/\sqrt{\pi}} \right)^{1/5} \approx 1.06 \frac{\hat{\sigma}}{\sigma_K} \cdot n^{-\frac{1}{5}}, \tag{3.6.12}$$

Silverman's rule of thumb (Silverman, 1986) and Scott's method (Scott, 1992) used the smaller value of $\hat{\sigma}$ and $\frac{IQR}{1.34}$ where IQR is the interquartile range of the sample. Silverman (1986) further recommended reducing the constant 1.06 in Equality (3.6.12) to 0.9 to avoid missing bimodality.

When $\sigma_K = 1$, i.e., $K(\cdot)$ is the density function of a standard normal distribution (i.e., $\sigma_K = 1$), Silverman's rule of thumb and Scott's method give the bandwidths $h_{\text{svlm}} = 0.9 \cdot \min\left(\hat{\sigma}, \frac{IQR}{1.34}\right) \cdot n^{-1/5}$, and $h_{\text{scott}} = 1.06 \cdot \min\left(\hat{\sigma}, \frac{IQR}{1.34}\right) \cdot n^{-1/5}$ respectively.

_3.6.3.2_ $T(-t, t, 0)$

With $K(\cdot)$ being the density function of a symmetric triangular distribution, $T(-t, t, 0)$, we have the standard deviation $\sigma_K = \frac{t}{\sqrt{6}}$, and $R(K) = \frac{2}{3t}$. As before, the normal density is assumed for f. The optimal bandwidth is

$$h_{\text{opt}} = \left(\frac{\frac{2}{3\sqrt{6}\sigma_K}}{n\sigma_K^4 \cdot \frac{3}{8}\hat{\sigma}^{-5}/\sqrt{\pi}} \right)^{1/5} \approx 1.05 \frac{\hat{\sigma}}{\sigma_K} \cdot n^{-\frac{1}{5}},$$

or, $2.57\frac{\hat{\sigma}}{t} \cdot n^{-\frac{1}{5}}$. Following the same logic of Silverman (1986) and Scott (1992), we use the smaller one of $\hat{\sigma}$ and $\frac{IQR}{1.34}$ to replace $\hat{\sigma}$. The resulting optimal bandwidth is $h_{T(t)} = 2.57\frac{\hat{\sigma}}{t} \cdot \min\left(\hat{\sigma}, \frac{IQR}{1.34}\right) \cdot n^{-\frac{1}{5}}$.

As can be seen in the two examples, 3.6.3.1 and 0, the optimal bandwidth h_{opt} changes with the value of the scale parameter of the kernel density function. However, the corresponding kernel density $K\left\{\left(\tilde{p}_i^{(s)} - \tilde{p}_j^{(c)}\right)/h_{opt}\right\}$ remains invariant to changes in the value of the scale parameter, which results in the KW pseudo-weights being also unaffected by scale parameter.

3.6.4 True propensity models under two-stage cluster sampling designs

3.6.4.1 True Propensity-Score Model assumed by PSAS and KW methods

Using the same notation in the main text, we denote s_c and s_s as the cohort and survey sample respectively, and denote FP as the finite population from which the s_c and s_s are selected. Define $\tilde{p}_i$ as the probability of being self-selected in the cohort for $i \in FP$ given it has been selected into the combined sample of cohort and survey sample, given by

$$\tilde{p}_i = P\{\, i \in s_c \mid i \in s_c \cup^* s_s\,\} = \frac{P\{\, i \in s_c \mid FP\,\}}{P\{\, i \in s_c \cup^* s_s \mid FP\,\}}$$

$$= \frac{\pi_i^{(c)}}{\pi_i^{(c)} + \pi_i^{(s)}}, \qquad (3.6.13)$$

where $\pi_i^{(c)} = \frac{n_c \cdot r_i^a}{\sum_{i \in FP} r_i^a}$ and $\pi_i^{(s)} = \frac{n_s \cdot r_i^b}{\sum_{i \in FP} r_i^b}$ (Table 3.1) are the inclusion probabilities of cohort and survey sample respectively for $i \in FP$ under the two-stage PPS sampling in the simulations with $r_i = \exp(\gamma_0 + \boldsymbol{\gamma}^T \boldsymbol{x}_i)$ defined in Formula (3.3.1). The notation $\cup^*$ represents the combination of the two samples that allows population units to be selected in both cohort and survey. The duplicates of s_c and s_s will be counted twice in the combined sample $s_c \cup^* s_s$.

Accordingly, $1 - \tilde{p}_i = \pi_i^{(s)} / \left(\pi_i^{(c)} + \pi_i^{(s)}\right)$ is the probability of $i \in FP$ being selected in the survey conditional on being selected into the combined cohort and survey sample. Hence, the log-odds of the $\tilde{p}_i$ can be written as

$$\log\left(\frac{\tilde{p}_i}{1 - \tilde{p}_i}\right) = \log\left\{\frac{\pi_i^{(c)} / \left(\pi_i^{(c)} + \pi_i^{(s)}\right)}{\pi_i^{(s)} / \left(\pi_i^{(c)} + \pi_i^{(s)}\right)}\right\} = \log\left\{\frac{\pi_i^{(c)}}{\pi_i^{(s)}}\right\}$$

$$= \log\left(\frac{n_c \cdot r_i^a / \sum_{i \in FP} r_i^a}{n_s \cdot r_i^b / \sum_{i \in FP} r_i^b}\right) = \log\left(\frac{n_c \cdot \sum_{i \in FP} r_i^b}{n_s \cdot \sum_{i \in FP} r_i^a}\right) + \log\left(\frac{r_i^a}{r_i^b}\right),$$

Therefore, the true model for $\tilde{p}_i$ is

$$\log\left(\frac{\tilde{p}_i}{1-\tilde{p}_i}\right) = \tilde{\beta}_0 + (a-b)\cdot\boldsymbol{\gamma}^T\boldsymbol{x}_i, \tag{3.6.14}$$

where $\tilde{\beta}_0 = \log\left(\frac{n_c\cdot\sum_{i\in FP} r_i^b}{n_s\cdot\sum_{i\in FP} r_i^a}\right) + (a-b)\cdot\gamma_0$ is the intercept. Note that the vector of model coefficients $\tilde{\boldsymbol{\beta}} = (a-b)\cdot\boldsymbol{\gamma}$ can be estimated by fitting the propensity model (3.6.14) to the combined (s_c vs. *unweighted* s_s) sample. The estimated PS, i.e., $\hat{\tilde{p}}_i = \text{expit}\left(\hat{\tilde{\beta}}_0 + \hat{\tilde{\boldsymbol{\beta}}}^T\boldsymbol{x}\right)$ is used by the PSAS and KW methods to measure the similarity between cohort units and survey units.

<u>3.6.4.2</u> *True Propensity-Score Model assumed by the IPSW method*

As defined in Model (2.3.9), p_i, for $i \in FP$, the probability of being in the cohort conditional on the sample of cohort combined with the finite population (*FP*), written as

$$p_i = P(i \in s_c | i \in s_c \cup^* FP) = \frac{\pi_i^{(c)}}{\pi_i^{(c)}+1}, \text{for } i \in FP.$$

Again, the notation $\cup^*$ means the combination of s_c and FP, allowing duplicated s_c in the combined set $s_c \cup^* FP$. Accordingly,

$$\log\left(\frac{p_i}{1-p_i}\right) = \log\left\{\frac{\pi_i^{(c)}/\left(\pi_i^{(c)}+1\right)}{1/\left(\pi_i^{(c)}+1\right)}\right\} = \log\left(\pi_i^{(c)}\right)$$

$$= \log\left(\frac{n_c}{\sum_{i\in FP} r_i^a}\right) + \log(r_i^a).$$

Therefore, the true model for p_i is

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + a\cdot\boldsymbol{\gamma}^T\boldsymbol{x}_i, \tag{3.6.15}$$

where $\beta_0 = \log\left(\frac{n_c}{\sum_{i \in FP} r_i^a}\right) + a \cdot \gamma_0$ is the intercept. Note that model coefficients $\boldsymbol{\beta} = a \cdot \boldsymbol{\gamma}$ can be estimated by fitting the propensity model (3.6.15) to the combined (s_c vs. *weighted* s_s) sample. The predicted odds, i.e. $\frac{\hat{p}_i}{1-\hat{p}_i} = \exp\left(\hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}\right)$ is used by the IPSW method to estimate the self-selection probability for cohort units.

3.6.5 Approximate unbiasedness of $\hat{\hat{\boldsymbol{\beta}}}$ and $\hat{\boldsymbol{\beta}}$ under two-stage cluster sampling designs in the simulation with the defined MOS

<u>3.6.5.1</u> *Proof of the approximate unbiasedness of $\hat{\hat{\beta}}$ and $\hat{\beta}$ under the situation of single binary covariate in MOS*

The theoretical justification suppose there is a sequence of finite population FP_k of size N_k, for $k = 1, 2, \cdots$. Cohort $s_{c,k}$ of size $n_{c,k}$ and survey sample $s_{s,k}$ of size $n_{s,k}$ are sampled from each FP_k. The sequences of the finite populations, cohort samples, and survey samples have sizes satisfy $\lim_{k \to \infty} \frac{n_{d,k}}{N_k} \to \gamma_d$, where $d = c, s$ and $0 < \gamma_d \leq 1$. In the following, the index k is suppressed for simplicity (Krewski & Rao 1981; Chen et al., 2019). As such, $O_p(N^{-1}) = O_p(n_c^{-1}) = O_p(n_s^{-1})$

Suppose the MOS of the PPS sampling is r_i^a and r_i^b for survey and cohort sample selected respectively for $i \in FP$, with $r_i = \exp(\theta_0 + \theta_1 x_i)$ as defined in Equality (3.3.1) where x is a binary covariate. Accordingly, the true propensity model fitted to the combined (s_c and *unweighted* s_s) is $\log\left(\frac{\tilde{p}_i}{1-\tilde{p}_i}\right) = \tilde{\beta}_0 + \tilde{\beta}_1 x_i$, with $\tilde{\beta}_1 = (a - b)\theta_1$. Suppose the binary covariate x has the distribution in FP shown below.

x	0	1	Total
Count	N_0	N_1	N

Then r_i has two possible values depending on the binary covariate x, defined by r_1 and r_0 as below

$$\begin{cases} r_i|(x_i = 0) = e^{\theta_0} \quad \triangleq r_0, \\ r_i|(x_i = 1) = e^{\theta_0+\theta_1} \triangleq r_1 \end{cases}$$

The (self-)selection probabilities of the cohort and the survey samples for $i \in FP$, $\pi_i^{(c)}$ and $\pi_i^{(s)}$, are defined as follows:

$$\text{cohort:} \begin{cases} \left(\pi_i^{(c)}\middle|r_i = r_0\right) \triangleq \pi_0^{(c)} \\ \left(\pi_i^{(c)}\middle|r_i = r_1\right) \triangleq \pi_1^{(c)} \end{cases}; \quad \text{survey:} \begin{cases} \left(\pi_i^{(s)}\middle|r_i = r_0\right) \triangleq \pi_0^{(s)} \\ \left(\pi_i^{(s)}\middle|r_i = r_1\right) \triangleq \pi_1^{(s)} \end{cases}$$

Suppose in the combined (s_c and *unweighted* s_s) sample, the table of sample membership T ($T_i = 0$ if $i \in s_s$; 1 if $i \in s_c$) and x is as follows

	T (sample membership)	
x	0 (s_s)	1 (s_c)
0	n_{00}	n_{01}
1	n_{10}	n_{11}
Total	n_s	n_c

A propensity model logit$(\tilde{p}) = \tilde{\beta}_0 + \tilde{\beta}_1 x$ is fitted to the *unweighted* combined sample. As $e^{\hat{\tilde{\beta}}_1} = \frac{n_{00} \cdot n_{11}}{n_{01} \cdot n_{10}}$, the expectation of $\hat{\tilde{\beta}}_1$ is

$$E\left(\hat{\tilde{\beta}}_1\right) = E(\log n_{00}) + E(\log n_{11}) - E(\log n_{01}) - E(\log n_{10}). \qquad (3.6.16)$$

By Taylor linearization, we have

$$E(\log n_{11}) = \log E(n_{11}) - \frac{Var(n_{11})}{2E^2(n_{11})} + \xi, \qquad (3.6.17)$$

where ξ is the remainder that has lower order than $\frac{Var(n_{11})}{2E^2(n_{11})}$. The first term of (3.6.17)can be written as $\log E(n_{11}) = \log\left\{N_1 \pi_1^{(c)}\right\}$, because

$$E(n_{11}) = \sum\nolimits_{i\in FP} E\left\{x_i \cdot \delta_i^{(c)}\right\} = \sum\nolimits_{i=1}^{N} P\left\{\delta_i^{(c)} = 1 | x_i = 1\right\} \cdot P\{x_i = 1\}$$

$$= \sum\nolimits_{i=1}^{N} \pi_1^{(c)} \cdot \frac{N_1}{N} = N_1 \pi_1^{(c)}$$

In the second term of (3.6.17), $Var[n_{11}]$ is

$$Var(n_{11}) = \sum\nolimits_{i\in FP} E\left\{x_i \cdot \delta_i^{(c)}\right\}\left[1 - E\left\{x_i \cdot \delta_i^{(c)}\right\}\right] = E(n_{11})\left\{1 - \frac{E(n_{11})}{N}\right\}$$

Therefore, the second term of (3.6.17) is $\frac{Var(n_{11})}{2E^2(n_{11})} = \frac{1-E(n_{11})/N}{2E(n_{11})} = O_p(n_c^{-1})$. Hence,

$$E[\log n_{11}] = \log N_1 \pi_1^{(c)} + O_p(n_c^{-1}) \qquad (3.6.18)$$

It can be similarly shown that

$$E(\log n_{01}) = \log N_0 \pi_0^{(c)} + O_p(n_c^{-1}), E(\log n_{00}) = \log N_0 \pi_0^{(s)} + O_p(n_c^{-1}),$$
$$\text{and } E(\log n_{10}) = \log N_1 \pi_1^{(s)} + O_p(n_c^{-1}). \qquad (3.6.19)$$

Equalities (3.6.16), (3.6.18) and (3.6.19) together give

$$E\left(\hat{\tilde{\beta}}_1\right) = \log\left\{\frac{\pi_1^{(c)}\pi_0^{(s)}}{\pi_0^{(c)}\pi_1^{(s)}}\right\} + O_p(n_c^{-1})$$

Under the two-stage PPS design described in Section 3.3.2, $\frac{\pi_1^{(c)}}{\pi_0^{(c)}} = \frac{r_1^{(c)}}{r_0^{(c)}} = e^{b\theta_1}$, and $\frac{\pi_0^{(s)}}{\pi_1^{(s)}} = \frac{r_1^{(s)}}{r_0^{(s)}} = e^{-a\theta_1}$. Hence,

$$E\left(\hat{\tilde{\beta}}_1\right) = \log\{e^{(b-a)\theta}\} + O_p(n_s^{-1}) = \tilde{\beta}_1 + O_p(n_s^{-1}).$$

The intercept $\hat{\tilde{\beta}}_0 = \log\frac{n_{01}}{n_{00}}$. Based on Equalities (3.6.19), $E\left(\hat{\tilde{\beta}}_0\right) = \log\frac{\pi_0^{(c)}}{\pi_0^{(s)}} + O_p(n_s^{-1})$, where

$$\log\frac{\pi_0^{(c)}}{\pi_0^{(s)}} = \log\left(\frac{n_c \cdot \sum_{i\in FP} r_i^b}{n_s \cdot \sum_{i\in FP} r_i^a}\right) + (a-b)\theta_0 = \tilde{\beta}_0$$

This completes the proof.

For the propensity model fitted to the combined (s_c vs. *weighted* s_s) sample, it can be similarly proved that $E\left[\hat{\beta}_1\right] = \beta_1 + O_p(n_s^{-1})$, and $E\left[\hat{\beta}_0\right] = \beta_0 + O_p(n_s^{-1})$ by using the following table of sample membership R ($R_i = 0$ if $i \in s_c$; 1 if $i \in FP$) and x in the combined (s_c vs. *weighted* s_s) sample as

	R	
x	0 ($weighted\ s_s$)	1 (s_c)
0	$\widehat{N}_0 = n_{00}/\pi_0^{(s)}$	n_{01}
1	$\widehat{N}_1 = n_{00}/\pi_0^{(s)}$	n_{11}
Total	$\widehat{N} = \widehat{N}_0 = \widehat{N}_1$	n_c

<u>3.6.5.2</u> *Empirical justification of the approximate unbiasedness of $\hat{\hat{\beta}}$ and $\hat{\beta}$ in simulations*

In simulations, we empirically verify the approximate unbiasedness of $\hat{\hat{\beta}}$ and $\hat{\beta}$. The MOS for cohort and survey sample selection were r_i^a and r_i^b respectively, for $i \in FP$ where $r_i = \exp(\gamma_0 + \gamma_1 age_i + \gamma_2 hh\_inc_i + \gamma_3 Env_i + \theta_4 z_i)$ with $\gamma_0 = 0$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (0, 0.3, -0.4, 0.7, 0.7)$, $a = -1$, and $b = 0.5$ (detailed in Section 3.3).

As derived in (3.6.14) and (3.6.15), the regression coefficients of the true propensity models of $p_i = P(\, i \in s_c \mid s_c \cup^* FP\,)$ and $\tilde{p}_i = P(\, i \in s_c \mid s_c \cup^* s_s\,)$ are

$$\boldsymbol{\beta} = -\boldsymbol{\gamma} = -(\gamma_1, \gamma_2, \gamma_3, \gamma_4) \text{ and } \widetilde{\boldsymbol{\beta}} = (a - b) \cdot \boldsymbol{\gamma} = -1.5(\gamma_1, \gamma_2, \gamma_3, \gamma_4),$$

respectively. The percent of relative bias (%RB), empirical variance (V) of the estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\widehat{\boldsymbol{\beta}}}$ are shown in Table 3.11.

As expected, the unweighted and weighted sample produced approximately unbiased estimates of $\widetilde{\boldsymbol{\beta}} = -1.5\boldsymbol{\gamma}$ and $\boldsymbol{\beta} = -\boldsymbol{\gamma}$**,** respectively. Thus, the three methods can achieve greatest bias reduction under the true propensity models that have the same

Table 3.11 Results of coefficients of propensity model fitted to unweighted and weighted combined cohort and survey sample over 1,000 simulation runs

	$\gamma_1 = 0.3$	$\gamma_2 = -0.4$	$\gamma_3 = 0.7$	$\gamma_4 = 0.7$
Estimate				
$\hat{\hat{\boldsymbol{\beta}}}$	-0.44	0.59	-1.04	-1.06
$\hat{\boldsymbol{\beta}}$	-0.29	0.39	-0.68	-0.71
Relative Bias%				
$\hat{\hat{\boldsymbol{\beta}}}$	-2.2	-1.7	-0.6	0.6
$\hat{\boldsymbol{\beta}}$	3.3	-2.5	2.9	1.4
Empirical Variance ($\times 10^3$)				
$\hat{\hat{\boldsymbol{\beta}}}$	0.71	1.98	8.61	93.02
$\hat{\boldsymbol{\beta}}$	1.03	3.53	15.43	189.29

$\hat{\hat{\boldsymbol{\beta}}}$ is estimated from the propensity model fitted to the combined (s_c and *unweighted* s_s), which is used by the PSAS and KW methods.$E\left(\hat{\hat{\boldsymbol{\beta}}}\right) \approx -1.5\boldsymbol{\gamma} = (-0.45, 0.6, -1.05, -1.05)$.

$\hat{\boldsymbol{\beta}}$ is estimated from the propensity model fitted to the combined (s_c and *weighted* s_s), which is used by the IPSW method.$E(\hat{\boldsymbol{\beta}}) \approx -\boldsymbol{\gamma} = (-0.3, 0.4, -0.7, -0.7)$.

functional form of covariates $\boldsymbol{x}$ in the simulation (the IPSW and KW estimates are expected to be approximately unbiased while the PSAS estimates can be biased under the true propensity model due to invalid assumption of the equal representativeness of cohort units within subclasses). This allows for a fair comparison among the three methods in the simulation.

However, the coefficients estimated from the propensity model fitted to the *weighted* sample had much larger empirical variances than the coefficients estimated from the model fitted to the *unweighted* sample due to the highly variable weights (weights of 1 for cohort units, and the sample weights for survey units). Hence, we expect that the naïve Taylor linearization (TL) method, which ignores variability due to estimating propensity scores, may substantially underestimate the variance of the IPSW estimates.

3.6.6 The KW and IPSW Estimators of Population Means When the Propensity Model Includes Variable of Interest Only

Regardless the true propensity model, we fit the model logit$(\tilde{p}) \sim y$ to the combined (s_c vs. unweighted s_s) sample, where y is the variable of interest. Suppose the distribution of y in the combined sample is as follows.

Table 3.12 Distribution of variable of interest in the cohort and survey sample

	T sample membership		
y	0 (s_s)	1 (s_c)	Total
0	n_{00}	n_{01}	$n_{0\cdot}$
1	n_{10}	n_{11}	$n_{1\cdot}$
Total	$n_{\cdot 0}(n_s)$	$n_{\cdot 1}(n_c)$	$n_{\cdot\cdot}$

Suppose the estimated coefficients of the propensity model are $\hat{\hat{\boldsymbol{\beta}}} = \left(\hat{\hat{\beta}}_0, \hat{\hat{\beta}}_1\right)$. Since y is the only covariate in the propensity model, there are only two values of the estimated PSs: $\hat{\hat{p}}_0 = \text{expit}(\hat{\hat{\beta}}_0)$, and $\hat{\hat{p}}_1 = \text{expit}\left(\hat{\hat{\beta}}_0 + \hat{\hat{\beta}}_1\right)$. Assume that the kernel function $K(\cdot)$ and bandwidth h satisfy conditions (a) and (b) in **Theorem 3.1**so that $K\{(\hat{\hat{p}}_0 - \hat{\hat{p}}_0)/h\} = K\{(\hat{\hat{p}}_1 - \hat{\hat{p}}_1)/h\} = K(0) > 0$ and $K\{(\hat{\hat{p}}_0 - \hat{\hat{p}}_1)/h\} \doteq 0$. The KW estimator of the total number of diseased people in the population, $\hat{Y}^{KW}$ is equal to the survey estimator, $\hat{Y}^{SVY} = \sum_{j=1}^{n_{10}} d_j$. The proof is as follows:

$$\hat{Y}^{KW} = \sum_{i \in s_c} w_i^{KW} \cdot y_i$$

$$= \sum_{j \in s_s} d_j \left[\frac{\sum_{i \in s_c} K\left\{\left(\hat{\hat{p}}_j^{(s)} - \hat{\hat{p}}_i^{(c)}\right)/h\right\} y_i}{\sum_{i \in s_c} K\left\{\left(\hat{\hat{p}}_j^{(s)} - \hat{\hat{p}}_i^{(c)}\right)/h\right\}} \right] \text{(switching order of summation)}$$

$$= \sum_{j=1}^{n_{00}} d_j \frac{\sum_{i=1}^{n_{01}} K(0) \cdot 0 + \sum_{i=1}^{n_{11}} 0 \cdot 0}{\sum_{i=1}^{n_{01}} K(0) + \sum_{i=1}^{n_{11}} 0} + \sum_{j=1}^{n_{10}} d_j \frac{\sum_{i=1}^{n_{01}} 0 \cdot 1 + \sum_{i=1}^{n_{11}} K(0) \cdot 1}{\sum_{i=1}^{n_{01}} 0 + \sum_{i=1}^{n_{11}} K(0)}$$

$$= \sum_{j=1}^{n_{10}} d_j \frac{\sum_{i=1}^{n_{11}} K(0)}{\sum_{i=1}^{n_{11}} K(0)} = \sum_{j=1}^{n_{10}} d_j = \hat{Y}^{SVY}.$$

Meanwhile, since $\hat{N}^{KW} = \hat{N}^{SVY}$ (proof in Equality (3.2.4)),

$$\hat{\bar{Y}}^{KW} = \frac{\hat{Y}^{KW}}{\hat{N}^{KW}} = \frac{\hat{Y}^{SVY}}{\hat{N}^{SVY}} = \hat{\bar{Y}}^{SVY}.$$

Under the IPSW method, we expand the survey sample by their population weights, hence n_{00} and n_{10} are replaced by $n_{00} = \sum_{j=1}^{n_{00}} d_j$ and $n_{10} = \sum_{j=1}^{n_{10}} d_j$. The estimated PSs are $\{\hat{p}|(y = 1)\} = n_{11}/\left\{\left(\sum_{j=1}^{n_{10}} d_j\right) + n_{11}\right\}$, and $\{\hat{p}|(y = 0)\} = n_{01}/\left\{\left(\sum_{j=1}^{n_{00}} d_j\right) + n_{01}\right\}$.

By using $\hat{p}^{-1}$ as the pseudo weight proposed by Dever & Valliant (2009), the resulting pseudo-weighted estimator of population mean, $\bar{Y}$ is

$$\hat{\mu}^* = \frac{\sum_{i=1}^{n_{11}}\{\hat{p}|(y = 1)\}}{\sum_{i=1}^{n_{11}}\{\hat{p}|(y = 1)\} + \sum_{i=1}^{n_{01}}\{\hat{p}|(y = 0)\}} = \frac{\left(\sum_{j=1}^{n_{10}} d_j\right) + n_{11}}{\sum_{j \in s_s} d_j + n_c}.$$

Compared to the survey estimator of μ_{FP}, $\hat{\mu}^{SVY} = \frac{\sum_{j=1}^{n_{10}} d_j}{\sum_{j \in s_S} d_j}$, $\hat{\mu}^*$ is biased. The bias goes to 0 when the sample fraction of the cohort is small, i.e., $\frac{n_c}{N} \to 0$. To get unbiased estimators, the IPSW method should use the inverse of odds as the pseudo weights, i.e., $w^{IPSW} = \frac{1-\hat{p}}{\hat{p}}$.

$$\hat{\mu}^{IPSW} = \frac{\sum_{i \in s_c} w_i^{IPSW} y_i}{\sum_{i \in s_c} w_i^{IPSW}} = \frac{\frac{\sum_{j=1}^{n_{10}} d_j}{n_{11}} \cdot n_{11}}{\frac{\sum_{j=1}^{n_{10}} d_j}{n_{11}} \cdot n_{11} + \frac{\sum_{j=1}^{n_{10}} d_i}{n_{01}} \cdot n_{01}} = \frac{\sum_{j=1}^{n_{10}} d_j}{\sum_{j \in s_S} d_j} = \hat{\mu}^{SVY}$$

Hence, when y is the only predictor in the propensity model, the KW and the IPSW methods will give exactly the same estimator as the survey estimator, which is design consistent to the finite population mean.

Chapter 4 Efficient and Robust Propensity-Score-Based Weighting Methods for Finite Population Inference from Nonprobability Epidemiologic Cohort

4.1 Introduction

In this chapter, 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 the finite population means. The SEA states that the expectation of the outcome variable given the estimated PS is the same in all three of the cohort, the survey, and the finite population. We prove that, without the SEA, current PS-based matching estimates are biased, even when a correctly specified propensity model is fitted to the combined (cohort vs. *unweighted* survey) sample. We establish a unifying framework for both PS-based weighting and matching methods. We relax the SEA to a weak exchangeability assumption (WEA) by defining matching scores that are (functions of) PSs estimated from the propensity models fitted to the combined (cohort vs. *weighted* survey) sample.

However, fitting the propensity model to the weighted sample (when compared to the unweighted sample) increases variability in PS estimation and can greatly increase the variance of pseudo-weighted estimates of finite population quantities. To increase efficiency, we propose **scaling** survey weights by their mean in PS estimation. Scaling 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 (Scott & Wild, 1986; Li et al., 2011; Landsman & Graubard, 2013). We demonstrate that this simple scaling greatly reduces variance while retaining the consistency of the estimators of finite population means. We derive TL variances for pseudo-weighted estimates of finite population means that take the variability of estimating the pseudo-weights into account under the framework. Two types of TL sample estimators of the finite population variances are given, depending on the underlying self-selection process of the cohort sample.

Monte Carlo simulation studies are conduced to evaluate the performance of the proposed PS-based methods under the SEA and the WEA with correct and misspecified propensity models. We apply our methods to an example where we use the naïve (unweighted) 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. This example is contrived, but the use of naïve NHANES allows the assessment of the bias reduction obtained by the proposed PS-based methods, without differences in the population coverage and measurement errors between a cohort and the reference survey that usually occur.

4.2 Basic Setting

We use notation consistent with Section 3.2.1 in Chapter 3. In addition, we require the following standard assumptions for the cohort participation:

A1. The cohort participation indicator $\delta^{(c)}$ is independent of the outcome variable y given the covariates $\boldsymbol{x}$, i.e., $\Pr\left(\delta^{(c)} = 1 \mid y, \boldsymbol{x} \right) = \Pr\left(\delta^{(c)} = 1 \mid \boldsymbol{x} \right)$.

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

For the reference survey, it is assumed that the sampling is also independent of the outcome variable y given the covariates $\boldsymbol{x}$, i.e., $\Pr\left(\delta^{(s)} = 1 \mid y, \boldsymbol{x} \right) = \Pr\left(\delta^{(s)} = 1 \mid \boldsymbol{x} \right)$.

4.3 Strong Exchangeability Assumption

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 the matching scores accurately to estimate the participation rates $\pi_i^{(c)}$ for the cohort units. To avoid low efficiency of pseudo-weighted estimates due to high variability of the estimated pseudo-weights, the existing PS-based matching methods, including the PSAS method (Lee & Valliant, 2009), the Rivers' matching method (Rivers et al., 2007), and the KW method proposed in Chapter 3, use the propensity of participating in the cohort (s_c) versus being selected in the survey sample (s_s) as the matching score, defined by,

$$\begin{aligned}\tilde{p}_i &= P\{i \in s_c \mid i \in s_c \cup^* s_s\} = \frac{P\{i \in s_c \mid FP\}}{P\{i \in s_c \cup^* s_s \mid FP\}} \\ &= \frac{\pi_i^{(c)}}{\pi_i^{(c)} + \pi_i^{(s)}}\end{aligned} \tag{4.3.1}$$

The union $s_c \cup^* s_s$ allows for duplication of individuals in both s_c and s_s. In practice, the set of duplicates is usually small and also it is not possible to identify the duplicates. Assume that the relationship between $\tilde{p}_i$ and $\boldsymbol{x}$ follows a logistic regression model

$$\log\frac{\tilde{p}_i}{1-\tilde{p}_i} = \widetilde{\boldsymbol{\beta}}^T \boldsymbol{x}_i, i \in \{s_c \cup^* s_s\}, \tag{4.3.2}$$

where $\widetilde{\boldsymbol{\beta}}$ is a vector of unknown regression coefficients, and can be estimated by fitting Model (4.3.2) to the combined (s_c vs. *unweighted* s_s) sample. As proved in Section 3.6.1, using $\tilde{p}$ in the matching score requires the strong exchangeability assumption (SEA)

$$E(y \mid \tilde{p}, s_s) = E(y \mid \tilde{p}, s_c) = E(y \mid \tilde{p}, FP), \tag{4.3.3}$$

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 three 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 (equal expectations for the cohort and survey samples) automatically holds under Model (4.3.2) (Rosenbaum & Rubin, 1983), but the second equality (equal expectations for the cohort and finite population) may not necessarily hold, resulting in biased pseudo-weighted estimation. We use simple examples to illustrate the cases of unbiased mean estimation when the SEA is valid, and also the cases of biased mean estimation when the SEA is violated even if the propensity model is correctly specified.

Simple Examples

Suppose the covariates $\boldsymbol{x}$ 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 $\mu_{jk} = E(y \mid \text{age} = j, \text{sex} = k)$ for $j, k = 0, 1$. We assume μ_{jk} differs by the four categories of age by sex. The overall finite population mean of y is $\mu = \frac{1}{N}\sum_{j,k} N_{jk} \cdot \mu_{jk}$, where N_{jk} is the number of individuals in the FP for category jk (age $= j, \text{sex} = k$, for $j, k = 0, 1$). We assume s_c and s_s are two independent samples randomly selected by stratified simple random sample designs. The participation/sampling rates of s_c and s_s are, respectively, $a_j^{(c)}$ and $a_j^{(s)}$ for age group j, and $b_k^{(c)}$ and $b_k^{(s)}$ for sex group k. The final participation/sampling rates of s_c and s_s for a population unit in category jk are $\pi_{jk}^{(c)} = a_j^{(c)} b_k^{(c)}$ and $\pi_{jk}^{(s)} = a_j^{(s)} b_k^{(s)}$, respectively. Accordingly, the propensity of being included in s_c versus s_s is $\tilde{p}_{jk} = \frac{a_j^{(c)} b_k^{(c)}}{a_j^{(c)} b_k^{(c)} + a_j^{(s)} b_k^{(s)}}$ by Equality (4.3.1). The sample sizes in category jk are $n_{jk}^{(c)} = N_{jk}\pi_{jk}^{(c)}$ for s_c and $n_{jk}^{(s)} = N_{jk}\pi_{jk}^{(s)}$ for s_s.

1. SEA Valid Case

In the SEA valid case, the values of $\tilde{p}$ are different in the four categories defined by age and sex. The SEA is satisfied because $E(\, y \mid \tilde{p}\,) = E(\, y \mid \text{age} = j, \text{sex} = k\,) = \mu_{jk}$ for s_c, s_s, and FP. The PS-based matching methods distribute survey weights to the cohort units within each of the four matching groups defined by $\tilde{p}$, i.e., pseudo-weights for all cohort units in category jk are commonly $\widetilde{w}_{jk} = \frac{N_{jk}}{n_{jk}^{(c)}}$. Denote the pseudo-weighted estimate of μ by $\hat{\mu} = \left(\sum_{j,k} \widetilde{w}_{jk} \cdot n_{jk}^{(c)}\right)^{-1} \sum_{i \in (j,k)} \widetilde{w}_{jk,i} \cdot y_i$. As a result, $\hat{\mu}$ an is unbiased estimator of μ with $E(\hat{\mu}) = \frac{1}{N} \sum_{j,k} N_{jk} \cdot \mu_{jk} = \mu$.

2. SEA Invalid Case

SEA can be invalid if the value of $\tilde{p}$ cannot differentiate the four age-by-sex categories. This can happen even when the true propensity model (including main effects of 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, i.e., $\frac{b_0^{(s)}}{b_0^{(c)}} = \frac{b_1^{(s)}}{b_1^{(c)}} = b$, the PS in Equality (3.3.1) becomes $\tilde{p}_{jk} = \frac{a_j^{(c)}}{a_j^{(c)} + b \cdot a_j^{(s)}}$, which is identical within age group j (denoted by $\tilde{p}_{j\cdot}$) regardless of the sex group. The first equality in the SEA holds, i.e.,

$$E\{y|\tilde{p}_{j\cdot}, s_c\} = \frac{\mu_{j0} n_{j0}^{(c)} + \mu_{j1} n_{j1}^{(c)}}{n_{j0}^{(c)} + n_{j1}^{(c)}} = \frac{\mu_{j0} n_{j0}^{(s)} + \mu_{j1} n_{j1}^{(s)}}{n_{j0}^{(s)} + n_{j1}^{(s)}} = E\{y|\tilde{p}_{j\cdot}, s_s\}$$

as $\frac{n_{j0}^{(s)}}{n_{j1}^{(s)}} = \frac{n_{j0}^{(c)}}{n_{j1}^{(c)}}$. However, the second equality in the SEA does not hold because

$$E\{y|\tilde{p}_{j\cdot}, FP\} = \frac{\mu_{j0} N_{j0} + \mu_{j1} N_{j1}}{N_{j0} + N_{j1}} \neq E\{y|\tilde{p}_{j\cdot}, s_c\}$$

as $\mu_{j0} \neq \mu_{j1}$. Applying a PS-based matching method, the survey sample weights would be distributed to the cohort units according to age only, resulting in common pseudo-weights within age categories, i.e. $\widetilde{w}_{j0} = \widetilde{w}_{j1} = \frac{N_{j0}+N_{j1}}{n_{j0}^{(c)}+n_{j1}^{(c)}}$ for $j = 0, 1$. As a result, the pseudo-weighted estimate $\hat{\mu}$ is biased as shown below

$$E(\hat{\mu}) = \frac{1}{N}\left\{\frac{N_{00}+N_{01}}{n_{00}^{(c)}+n_{01}^{(c)}} \cdot \left(\mu_{00}n_{00}^{(c)} + \mu_{01}n_{01}^{(c)}\right) + \frac{N_{10}+N_{11}}{n_{10}^{(c)}+n_{11}^{(c)}}\left(\mu_{10}n_{10}^{(c)} + \mu_{11}n_{11}^{(c)}\right)\right\} \neq \mu.$$

As shown by the simple examples, when cohort units within a subgroup have different participation rates, but the same estimated PS, 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 propensity model is correctly fitted to the unweighted sample.

4.4 Unifying Framework for Using Propensity Scores in PS-Based Methods

We observe the covariates $\boldsymbol{x}_i$ for all $i \in FP$, but we do not observe the cohort participation indicator $\delta_i^{(c)}$ for all $i \in FP$. Instead of directly modeling the cohort participation rate, $\pi_i^{(c)}$, we define $p_i = P(\, i \in s_c \mid i \in s_c \cup^* FP\,)$, where the notation $\cup^*$ represents the union of s_c and FP which includes the duplication of the individuals in s_c that, of course, are in FP. Therefore, $s_c \cup^* FP$ contains $N + n_c$ individuals. As to be shown, duplicating the units in s_c is a computational device that allows us to recover an estimate of the underlying inclusion probability $\pi_i^{(c)}$ from a propensity model. According to the definition of p_i, it gives

$$\frac{p_i}{1-p_i} = \frac{P(i \in s_c \mid i \in s_c \cup^* FP)}{P(i \in FP \mid i \in s_c \cup^* FP)} \tag{4.4.1}$$
$$= \frac{P(i \in s_c)/P(i \in s_c \cup^* FP)}{P(i \in FP)/P(i \in s_c \cup^* FP)} = P(i \in s_c \mid i \in FP) = \pi_i^{(c)}.$$

Assume a logistic regression model for p_i

$$\log\left\{\frac{p_i}{1-p_i}\right\} = \boldsymbol{\beta}^T \boldsymbol{x}_i, \text{for } i \in s_c \cup^* FP. \tag{4.4.2}$$

Combining (4.4.1) and (4.4.2) gives $\pi_i^{(c)} = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)$, allowing us to obtain the cohort participation rate via a logistic propensity model. The log-likelihood function is given by

$$l(\boldsymbol{\beta}) = \sum\nolimits_{i \in s_c \cup^* FP} R_i \cdot \log p_i + (1-R_i)\log(1-p_i)$$
$$= \sum\nolimits_{i \in s_c} \log p_i + \sum\nolimits_{i \in FP} \log(1-p_i), \tag{4.4.3}$$

where R_i indicates the membership of s_c in $s_c \cup^* FP$ (i.e., $R_i = 1$ if $i \in s_c$, and $= 0$ if $i \in FP$), and the propensity score p_i can be rewritten as $p_i = P(R_i = 1 \mid \boldsymbol{x}_i)$ for simplicity. Furthermore, $p_i = \text{expit}(\boldsymbol{\beta}^T \boldsymbol{x}_i)$ based on Model (4.4.2). Note that $\boldsymbol{\beta}$ differs from $\widetilde{\boldsymbol{\beta}}$ in Model (4.3.2) since the two models define the propensity differently, i.e., the probability that individual i is included in s_c vs. FP under Model (4.4.2) as compared to the probability that individual i is included in s_c vs. s_s under Model (4.3.2). To obtain a consistent estimator of $\boldsymbol{\beta}$, we fit Model (4.4.2) 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 through the pseudo log-likelihood function

$$\tilde{l}(\boldsymbol{\beta}) = \sum\nolimits_{i \in s_c} \log p_i + \sum\nolimits_{i \in s_s} d_i \log(1-p_i). \tag{4.4.4}$$

Heuristically, we are substituting the sample weighted s_s for FP. The estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by solving the following weighted estimating equations for $\boldsymbol{\beta}$

$$\tilde{S}(\boldsymbol{\beta}) = \frac{\partial \tilde{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum\nolimits_{i \in s_c} (1 - p_i)\, \boldsymbol{x}_i - \sum\nolimits_{i \in s_s} d_i p_i \boldsymbol{x}_i = \mathbf{0}.$$

According to Equation (4.4.1), the participation rate $\pi_i^{(c)}$ for unit $i \in s_c$ is estimated by $\hat{\pi}_i^{(c)} = \frac{\hat{p}_i}{1-\hat{p}_i}$, with $\hat{p}_i = \text{expit}\left(\widehat{\boldsymbol{\beta}}^T \boldsymbol{x}_i\right)$ being the estimated PS under the propensity model (4.4.2).

4.4.1 Inverse PS Weighting (IPSW) Method as a PS-Based Weighting Method

The IPSW method (Valliant and Dever, 2011) uses the inverse of estimated participation rate as the pseudo-weight, i.e., $w_i^{IPSW} = 1/\hat{\pi}_i^{(c)}$. The corresponding IPSW estimator of finite population mean, $\mu^{FP} = \frac{1}{N} \sum_{i \in FP} y_i$, is $\hat{\mu}^{IPSW} = \frac{\sum_{i \in s_c} w_i^{IPSW} y_i}{\sum_{i \in s_c} w_i^{IPSW}}$.

4.4.2 A Weak Exchangeability Assumption for PS-Based Matching Methods

We relax the SEA for the PS-based matching methods under the framework described above by using $p = P(\,R = 1 \mid \boldsymbol{x}\,)$ (Rosenbaum & Rubin, 1983) or $q = \text{logit}(p)$ (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. There are no cohort sample weights that can be used to equalize the distributions of covariates $\boldsymbol{x}$ in s_c and FP if they are different. The pseudo-weights created using the $\boldsymbol{x}$ serve as the sample weights for s_c to weight the $\boldsymbol{x}$ distribution in s_c up to that in FP.

The matching methods classify the cohort and survey individuals into "matching groups" with similar $\boldsymbol{x}$-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 $\boldsymbol{x}$-distribution in the pseudo-weighted s_c becomes closer to the $\boldsymbol{x}$-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 $\boldsymbol{x}$-distribution in s_c and in FP. The balancing score $b(\boldsymbol{x})$ is a function of covariates $\boldsymbol{x}$ such that the conditional distribution of $\boldsymbol{x}$ given $b(\boldsymbol{x})$ is the same in the s_c as that in the FP. We use the notation in Rosenbaum & Rubin (1983)

$$\boldsymbol{x} \perp\!\!\!\perp R|b(\boldsymbol{x}), \tag{4.4.5}$$

where R is defined in log-likelihood (4.4.3). The coarsest balancing score is $p = \Pr(R = 1 \mid \boldsymbol{x})$, or any one-to-one functions of p, e.g., the participation rate $\pi^{(c)} = \frac{p}{1-p}$ or $q = \operatorname{logit} p$ (Rubin & Thomas, 1992), which can be estimated from the propensity model (4.4.2) fitted to the combined (s_c vs *weighted* s_s) sample.

For estimating μ^{FP}, the requirement that the matching score should be a balancing score satisfying Definition (4.4.5) can be relaxed to the WEA

$$E\{y \mid b(\boldsymbol{x}), s_c\} = E\{y \mid b(\boldsymbol{x}), FP\}, \tag{4.4.6}$$

where E is the expectation with respect to the distribution of y in FP. The matching scores p and q satisfy WEA (4.4.6) because they are balancing scores defined in (4.4.5). Note that p has a bounded support $(0, 1)$ and is right skewed when the participation rate of the cohort is small. A tiny difference in p may be resulted from large differences in covariates $\boldsymbol{x}$, which can bias the estimates by the PS-based matching methods. These boundary problems can be avoided by using $q = \operatorname{logit}(p)$ (Rubin & Thomas, 1992) as the matching score.

4.4.3 Applying WEA to Kernel Weighting (KW) Approach

The kernel weighting (KW) approach, as a special case of PS-based matching methods, has been proved to provide consistent estimators of finite population means under SEA along with standard conditions (Section 3.6.1), whereas other PS-matching methods such as PSAS may not result in consistent estimators. In this section, we propose an enhanced KW (referred to as KW.W) method by applying the WEA under the framework in Section 4.4.2, 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 $w_i^{KW.W}$ for each individual $i \in s_c$, but the KW.W method uses $q = \text{logit}(p) = \boldsymbol{\beta}^T \boldsymbol{x}$ as a matching score, with $\boldsymbol{\beta}$ estimated under Model (4.4.2) fitted to the combined (s_c vs. *weighted* s_s) by maximizing the pseudo-loglikelihood (4.4.4). Denote the estimated logit of propensity scores to be $q_i^{(c)}$ and $q_j^{(s)}$ for $i \in s_c$ and $j \in s_s$, respectively. The KW.W pseudo-weight, $w_i^{KW.W}$ for $i \in s_c$, is calculated as

$$w_i^{KW.W} = \sum\nolimits_{j \in s_s} \left[\frac{K\left\{\left(q_i^{(c)} - q_j^{(s)}\right)/h\right\}}{\sum_{i \in s_c} K\left\{\left(q_i^{(c)} - q_j^{(s)}\right)/h\right\}} \cdot d_j \right], \qquad (4.4.7)$$

where $K(\cdot)$ is a zero-centered kernel function (Epanechnikov, 1969) (e.g. uniform, standard normal, or triangular density), h is the bandwidth corresponding to the selected kernel function. The KW.W estimate of μ is $\hat{\mu}^{KW.W} = \frac{\sum_{i \in s_c} w_i^{KW.W} y_i}{\sum_{i \in s_c} w_i^{KW.W}}$. Under the WEA and the conditions in **Theorem 4.1**. below, $\hat{\mu}^{KW.W}$ is design consistent with the finite population variance approximation $V^{KW.W}$(see definition below).

We consider the following limiting process (Krewski & Rao 1981; Chen et al., 2019) for the theoretical justification of **Theorem 4.1**. Suppose there is a sequence of finite population FP_k of size N_k, for $k = 1, 2, \cdots$. Cohort $s_{c,k}$ of size $n_{c,k}$ and survey sample $s_{s,k}$

of size $n_{s,k}$ are sampled from each FP_k. The sequences of the finite populations, cohort samples, and survey samples have sizes satisfy $\lim_{k\to\infty}\frac{n_{d,k}}{N_k}\to\gamma_d$, where $d = c, s$ and $0 < \gamma_d \leq 1$. In the following, the index k is suppressed for simplicity.

Theorem 4.1 Consistency of the KW.W estimate of the finite population mean.

Under the WEA (4.4.6)*, conditions* **A1**, **A2** *and* **C1-C5** *in Section 4.10.2, the KW.W estimate of the finite population mean* $\hat{\mu}^{KW.W} = \mu^{FP} + O_p\left(n_c^{-1/2}\right)$*. Assuming the logistic regression model* (4.4.2) *for the propensity scores* $p = Pr(R = 1 \mid \boldsymbol{x})$*, and under conditions* **C6**, **C8**, **C9** *in Section 4.10.2, we have the finite population variance* $Var(\hat{\mu}^{KW.W}) = V^{KW.W} + o(n_c^{-1})$*, where*

$$V^{KW.W} = N^{-2}\sum_{i\in FP}\pi_i^{(c)}\left(1-\pi_i^{(c)}\right)\{w_i^{KW.W}(y_i-\mu^{FP}) - (1-p_i)\boldsymbol{b}^T\boldsymbol{x}_i\}^2 + \boldsymbol{b}^T D\boldsymbol{b},$$

$$\boldsymbol{b}^T = \left\{\sum\nolimits_{i\in FP}\pi_i^{(c)}(y_i-\mu^{FP})\frac{\partial w_i^{KW.W}}{\partial\boldsymbol{\beta}}\right\}\{\sum\nolimits_{i\in FP}p_i\boldsymbol{x}_i\boldsymbol{x}_i^T\}^{-1},\ D = N^{-2}V_p\left(\sum\nolimits_{i\in FP}\delta_i^{(s)}d_ip_i\boldsymbol{x}_i\right),$$

and V_p *denoting the design-based finite population variance under the probability sampling design for* s_s*. Notice that* $\frac{\partial w_i^{KW.W}}{\partial\boldsymbol{\beta}}$ *depends on the choice of kernel function* $K(\cdot)$ *(proof in Section 4.10.3).*

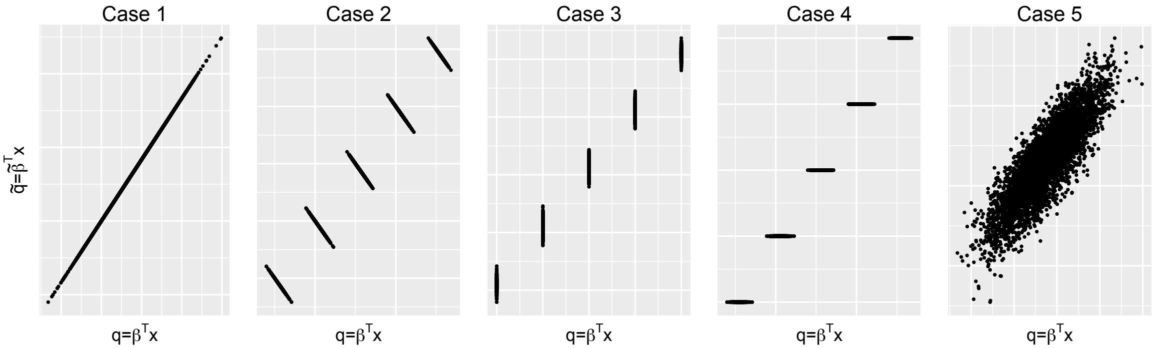
4.4.4 Checking if SEA is satisfied, assuming WEA holds

Although the matching methods such as the KW.W method can produce unbiased estimates by using the matching scores q, the variance can be inflated due to the highly variable weights in the combined sample. Using matching score $\tilde{p}$ defined in (4.3.1) may yield more efficient estimators, as the survey weights are not considered in the PS estimation. However, using the $\tilde{p}$ as a matching score can bias estimates of population means due to the potentially invalid SEA. Under the propensity model (4.3.2), we have $\boldsymbol{x} \perp\!\!\!\perp T|\tilde{p}$ with T

indicating the group membership of s_c versus s_s (Rosenbaum & Rubin, 1983), and $E\{\, y \mid \tilde{p}, s_c \,\} = E\{\, y \mid \tilde{p}, s_s \,\}$. However, $\tilde{p}$ may not satisfy the second equality in the SEA (4.3.3), i.e., $E\{\, y \mid \tilde{p}, s_c \,\} \neq E\{\, y \mid \tilde{p}, FP \,\}$. Therefore, the resulting estimates of population means can be biased. Hence, satisfying the second equality in the SEA is crucial for using $\tilde{p}$ as the matching score.

We recommend using the following scatter plots to empirically determine whether the SEA is satisfied, if at least WEA holds. Plot $\tilde{q} = \text{logit}(\tilde{p}) = \widetilde{\boldsymbol{\beta}}^T \boldsymbol{x}$, estimated by fitting Model (4.3.2) to the unweighted sample, versus $q = \boldsymbol{\beta}^T \boldsymbol{x}$, estimated by fitting Model (4.4.2) to the weighted sample. If $\tilde{q}$ is a one-to-one function (e.g. cases 1 and 2 in Figure 4.1) or many-to-one function (e.g. case 3 in Figure 1) of q, the second equality of the SEA $E\{\, y \mid \tilde{q}, s_c \,\} = E\{\, y \mid \tilde{q}, FP \,\}$ holds as $E\{\, y \mid q, s_c \,\} = E\{\, y \mid q, FP \,\}$. Otherwise $\tilde{q}$ will not satisfy the second equality the SEA (e.g. cases 4 and 5 in Figure 4.1). As a result, using matching scores of $\tilde{p}$ (or $\tilde{q}$) in the matching methods can produce biased estimates of population means for cases 4 and 5.

Figure 4.1 Hypothetical scatter plots of linear propensity scores for SEA diagnoses.



4.5 Improving Efficiency of the IPSW and KW Estimators by Scaling Survey Weights in PS Estimation

The IPSW and KW.W estimators of population means 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 weight of d_i, $i \in s_s$). Scaling the weights has been suggested to improve efficiency of estimators in population-based case-control studies when the weights are highly variable among cases and controls (Scott & Wild, 1986; Li et al., 2010). Following the rationale of Scott & Wild (1986), we propose scaling the survey weights $\{d_i, i \in s_s\}$ by the scaling factor $a = \frac{n_s}{\sum_{i \in s_s} d_i}$ and denote the scaled weight for the survey unit $i \in s_s$ by $d_i^* = a \cdot d_i$, so that $\sum_{i \in s_s} d_i^* = n_s$. The propensity model (4.4.2) is fitted to the combined (s_c vs. *scaled-weighted* s_s) sample and the pseudo log-likelihood (4.4.4)with the d_i replaced by d_i^* is maximized to solve for $\boldsymbol{\beta}$. The resulting estimator is denoted by $\widehat{\boldsymbol{\beta}}^*$.

Lemma 4.1. $\widehat{\boldsymbol{\beta}}^*$ is a consistent estimator of $\boldsymbol{\beta}^* = \boldsymbol{\beta} + \log a \cdot \boldsymbol{e}_1$, where $\boldsymbol{e}_1 = (1, 0, \cdots, 0)^T$, a is the scaling factor for survey weights, and $\boldsymbol{\beta}$ is the vector of regression coefficients defined in Model (4.4.2) (see proof of **Lemma 4.1** in Section 0).

Lemma 4.1 shows that rescaling survey weights in the combined sample for the propensity modeling only affects the intercept of the coefficients, which can be corrected by the offset of $\log a$. Therefore, as shown in **Theorem 4.2**, the resulting IPSW.S and KW.S estimators of the population mean μ when using the weights d_i^* for propensity estimation, denoted by $\hat{\mu}^{IPSW.S}$ and $\hat{\mu}^{KW.S}$, are also consistent estimators of μ as were $\hat{\mu}^{IPSW}$ and $\hat{\mu}^{KW.W}$.

Theorem 4.2. Consistency of the IPSW.S and KW.S estimates of the finite population means.

Under the WEA (4.4.6)*, conditions* **A1**, **A2** *and* **C1-C5** *in Section 4.10.2*, *and assuming the logistic regression model* (4.4.2)*, we have $\hat{\mu}^* = \mu^{FP} + O_p\left(n_c^{-1/2}\right)$, with $\hat{\mu}^*$ being either $\hat{\mu}^{IPSW.S}$ or $\hat{\mu}^{KW.S}$. Under conditions* **C7-C9** *in Section 4.10.2, we have the finite population variance $Var(\hat{\mu}^*) = V^* + o(n_c^{-1})$, with*

$$V^* = N^{-2} \sum_{i \in FP} \pi_i^{(c)}\left(1 - \pi_i^{(c)}\right)\left\{\widetilde{w}_i^*(y_i - \mu^{FP}) - (1 - p_i^*)\boldsymbol{b}^{*T}\boldsymbol{x}_i\right\}^2 + \boldsymbol{b}^{*T} D^* \boldsymbol{b}^*,$$

where V^*can be $V^{IPSW.S}$ or $V^{KW.S}$ depending on the choice of $\{\widetilde{w}_i^*, i \in FP\}$ being a set of IPSW.S or KW.S pseudo weights, $\boldsymbol{b}^*$and $D^*$are obtained by replacing $w_i^{KW.W}$, p_i and d_i with $\widetilde{w}_i^*$, p_i^* and d_i^* in $\boldsymbol{b}$ and D defined in **Theorem 4.1**, *respectively (proof and details in Section 4.10.5).*

4.6 TL Variance Estimation

4.6.1 Plug-in variance estimator for independent selection of the cohort units

In this section we summarize the steps for obtaining the consistent estimators of finite population variances for the four IPSW and KW estimators ($\hat{\mu}^{IPSW}$, $\hat{\mu}^{KW.W}$, $\hat{\mu}^{IPSW.S}$, $\hat{\mu}^{KW.S}$), and provide their expressions. Following the proof of **Theorem 4.1** (Section 4.10.3), both finite population variances of $\hat{\mu}^{IPSW}$ and $\hat{\mu}^{KW.W}$ can be approximated by

$$V = N^{-2} \sum_{i \in FP} \pi_i^{(c)}\left(1 - \pi_i^{(c)}\right)\{\widetilde{w}_i(y_i - \mu^{FP}) - (1 - p_i)\boldsymbol{b}^T\boldsymbol{x}_i\}^2 + \boldsymbol{b}^T D \boldsymbol{b}, \qquad (4.6.1)$$

where $\{\widetilde{w}_i, i \in FP\}$ is the set of IPSW (w_i^{IPSW}) or enhanced KW ($w_i^{KW.W}$) pseudo-weights. The first summand of (4.6.1) can be consistently estimated from the pseudo-weighted s_c assuming Poisson sampling of the cohort sample:

$$\{\widehat{N}^{(c)}\}^{-2} \sum_{i \in s_c} \left(1 - \frac{1}{\widetilde{w}_i}\right) \{\widetilde{w}_i (y_i - \hat{\mu}) - (1 - \hat{p}_i) \widehat{\boldsymbol{b}}^T \boldsymbol{x}_i\}^2, \tag{4.6.2}$$

where $\frac{1}{\widetilde{w}_i}$ is the estimate of $\pi_i^{(c)}$, $\hat{p}_i$ is the PS estimated from Model (4.4.2) for $i \in s_c$, $\widehat{N}^{(c)} = \sum_{i \in s_c} \widetilde{w}_i$, $\hat{\mu}$ is the pseudo-weighted estimate of μ^{FP} (either $\hat{\mu}^{IPSW}$ or $\hat{\mu}^{KW.W}$ depending on the choice of $\widetilde{w}_i$), and $\widehat{\boldsymbol{b}}^T = \left\{ \sum_{i \in s_c} (y_i - \hat{\mu}) \frac{\partial \widetilde{w}_i}{\partial \boldsymbol{\beta}} \Big|_{\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}} \right\} \{\sum_{i \in s_c} \hat{p}_i \boldsymbol{x}_i \boldsymbol{x}_i^T\}^{-1}$. The second summand, $\boldsymbol{b}^T D \boldsymbol{b}$, in (4.6.1) is estimated by $\widehat{\boldsymbol{b}}^T \widehat{D} \widehat{\boldsymbol{b}}$, where $\widehat{D}$ is the survey design consistent variance estimator that takes the complex sample design of the reference survey into account. For example, under a stratified multistage clustering sampling with H strata and u_h primary sampling units (PSUs) in stratum h, $\widehat{D}$ is estimated by

$$\{\widehat{N}^{(s)}\}^{-2} \cdot \sum_{h=1}^{H} \frac{u_h}{u_h - 1} \sum_{l=1}^{u_h} (\boldsymbol{z}_l - \bar{\boldsymbol{z}})(\boldsymbol{z}_l - \bar{\boldsymbol{z}})^T \tag{4.6.3}$$

where $\widehat{N}^{(s)} = \sum_{i \in s_s} d_i$, $\boldsymbol{z}_l = \sum_{i \in (lh)} d_i \hat{p}_i \boldsymbol{x}_i$ is the (lh) PSU total and $\bar{\boldsymbol{z}} = \frac{1}{a_h} \sum_l^{a_h} \boldsymbol{z}_l$ is the mean of PSU total in stratum h.

4.6.2 Other underlying complex sampling schemes for a cohort sample

We extend the sample estimate of V to more complex situations where the cohort can be assembled under a cluster sampling design (NIH-AARP, 2006) with a fixed sample size (usually assumed in the software such as “survey” package in R (Lumley, 2020), and SAS). The first summand of V in (4.6.1) can be estimated by

$$\left\{\widehat{N}^{(c)}\right\}^{-2} \cdot \frac{l}{l-1} \sum\nolimits_{\alpha=1}^{l} \left\{ (u_l - \bar{u}) - \widehat{\boldsymbol{b}}^T (\boldsymbol{v}_l - \bar{\boldsymbol{v}}) \right\}^2 \qquad (4.6.4)$$

where l is the number of clusters in the cohort, $u_l = \sum_{i \in \alpha} \widetilde{w}_i (y_i - \hat{\mu})$ and $\boldsymbol{v}_l = \sum_{i \in l} (1 - \hat{p}_i) \boldsymbol{x}_i$ are the cluster totals, $\bar{u} = \frac{1}{l} \sum_{\alpha=1}^{l} u_l$, and $\bar{\boldsymbol{v}} = \frac{1}{l} \sum_{\alpha=1}^{l} \boldsymbol{v}_l$ are the sample means of cluster totals. Terms $\widehat{\boldsymbol{b}}^T$ and $\widehat{D}$ can be estimated in the same way as Section 6.1.

In both Section 4.6.1 and 4.6.2, the sample estimate variance estimator of $\hat{\mu}^{IPSW.S}$ (or $\hat{\mu}^{KW.S}$) can be simply obtained by substituting $\widetilde{w}_i$, $\hat{\mu}$, $\widehat{\boldsymbol{\beta}}$, $\hat{p}_i$ and d_i by $\widetilde{w}_i^*$, $\hat{\mu}^*$, $\widehat{\boldsymbol{\beta}}^*$, $\hat{p}_i^*$ and d_i^*, respectively, in formulas (4.6.2)-(4.6.4).

4.7 Simulations

4.7.1 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 (=1 if $x_1 + x_2 > 2$; 0 otherwise), and $x_4 \sim \text{LogNormal}(0, 0.7)$. Note x_1 and x_2 are correlated with x_3, but independent of x_4. The outcome y for $i \in FP$ were generated by $y_i = 2 + x_{1,i} + x_{2,i} + x_{3,i} + \epsilon_i, i \in FP,$ where the error terms ϵ_i were independent and identically distributed (iid) as $N(0,1)$. The finite population mean of y is $\mu^{FP} = 4.50$.

For each $i \in FP$, we created two variables, x_1^* and x_1^{**} as functions of x_1: $x_{1,i}^* = x_{1,i} + 0.15 x_{1,i}^3$, and $x_{1,i}^{**}$ was defined as a categorical variable (=1 if $x_{1,i} \leq 10^{\text{th}}$ percentile; 2 if $10^{\text{th}} < x_{1,i} \leq 40^{\text{th}}$ percentiles; 3 if $40^{\text{th}} < x_{1,i} \leq 70^{\text{th}}$ percentiles; 4 if $70^{\text{th}} < x_{1,i} \leq 90^{\text{th}}$ percentiles; and 5 if $x_{1,i} > 90^{\text{th}}$ 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.

4.7.2 Sampling from the Finite Population to Assemble the Survey Sample and Cohort

A cohort of size $n_c = 2{,}400$ individuals was randomly selected from the FP by Probability Proportional to Size (PPS) sampling with measure of size (MOS) for individual $i \in FP$ defined by $r_i^{(c)} = \exp\left(\alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{4,i}\right)$, where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3) = (0.6, 0.15, 0.24)$. The sample weight (i.e., the reciprocal of the selection probability) for individual i in the cohort was $w_i^{(c)} = \frac{\sum_{i=1}^{N} r_i^{(c)}}{n_c \cdot r_i^{(c)}}$. A survey sample of size $n_s = 2{,}000$ individuals was sampled independently of the sampling of the cohort where a similar PPS sampling design was used, but with a different MOS $r_i^{(s)} = \exp\left(\gamma_1 x_{1,i} + \gamma_2 x_{2,i} + \gamma_3 x_{4,i}\right)$.

Under PPS sampling described above, the true propensity model of a population unit included in s_c vs. FP (assumed by the IPSW and KW.W methods), and that in s_c vs. s_s (assumed by the original KW) are

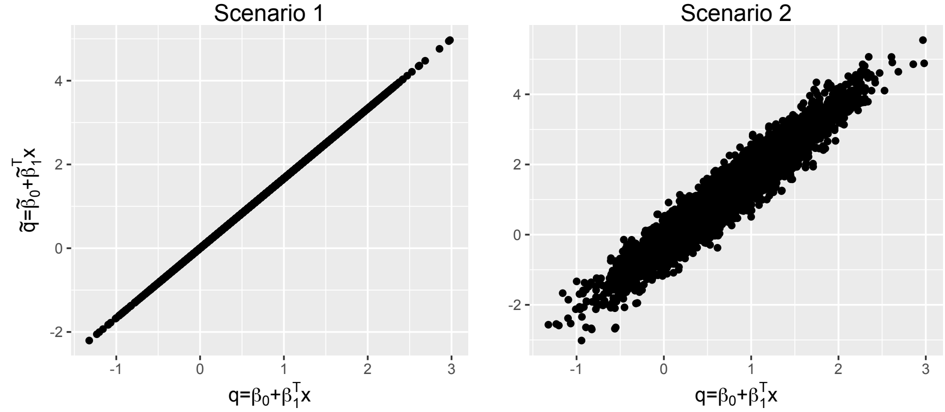
$$\begin{aligned} \operatorname{logit}\{p(i \in s_c | i \in s_c \cup^* FP)\} &= \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{x}_i, \qquad \text{and} \\ \operatorname{logit}\{p(i \in s_c | i \in s_c \cup^* s_s)\} &= \tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}_1^T \boldsymbol{x}_i, \end{aligned} \tag{4.7.1}$$

respectively, where $\boldsymbol{\beta}_1 = \boldsymbol{\alpha}$, $\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{\alpha} - \boldsymbol{\gamma}$; β_0 and $\tilde{\beta}_0$ are the intercepts (Section 4.10.6). 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 $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$ in the MOS of survey sample selection can be varied to control for the validity of the SEA assumed by the original KW method introduced in Chapter 3. We considered two scenarios with $\boldsymbol{\gamma} = (-0.4, -0.1, 0.16)$ in Scenario 1, and $\boldsymbol{\gamma} = (-0.65, 0.2, 0)$ in Scenario 2. Following Section 4.4.4, we made a scatter plot of $\tilde{q} = \tilde{\beta}_0 + \tilde{\boldsymbol{\beta}}_1^T \boldsymbol{x}$ versus $q = \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{x}$ under both scenarios. As shown in

Figure 4.2, $\tilde{q}$ was a one-to-one-function of the balancing score q in Scenario 1, and therefore satisfied the second equality $E(\, y \mid \tilde{q}, s_c\,) = E(\, y \mid \tilde{q}, FP\,)$ in the SEA. However, in Scenario 2, $\tilde{q}$ was not a balancing score (similar to case 5 in Figure 4.1), which violated SEA.

Figure 4.2 Scatter plots of linear propensity scores for SEA diagnoses in the simulation



4.7.3 Evaluating Criteria

We examined the performance of the five PS-based estimators of μ_{FP}: two IPSW estimates ($\hat{\mu}^{IPSW}$, $\hat{\mu}^{IPSW.S}$) and three KW methods ($\hat{\mu}^{KW}$, $\hat{\mu}^{KW.W}$, $\hat{\mu}^{KW.S}$), which are compared to the naïve unweighted cohort estimator ($\hat{\mu}^{Naive}$) and the weighted survey estimator ($\hat{\mu}^{SVY}$). We used criteria of relative bias (%RB), empirical variance (V), mean squared error (MSE) of the estimators, defined by

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

$$\text{MSE}=\frac{1}{B}\sum_{b=1}^{B}\left\{\hat{\mu}^{(b)}-\mu_{FP}\right\}^2,$$

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

For each mean estimator, we evaluated two variance estimators, i.e., the Taylor linearization (TL, described Section 4.6.2) estimator and the Jackknife replication (JK) estimator (Section 4.10.7), using the variance ratio (VR), and coverage probabilities (CP) of the corresponding 95% confidence intervals, defined by

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

where $\hat{v}^{(b)}$ is the variance estimate of $\hat{\mu}^{(b)}$, and $CI^{(b)}=\left(\hat{\mu}^{(b)}-1.96\sqrt{\hat{v}^{(b)}},\ \hat{\mu}^{(b)}+1.96\sqrt{\hat{v}^{(b)}}\right)$ is the 95% confidence interval from the b-th simulated sample.

4.7.4 Results under Scenario 1: the valid SEA

Table 4.1 shows the results under the SEA. The unweighted naïve cohort mean, $\hat{\mu}^{Naive}$, was biased by 20.97% while the survey estimate $\hat{\mu}^{SVY}$ is approximately unbiased. All KW and IPSW methods yielded approximately unbiased estimates of μ^{FP}. Although the original IPSW estimate $\hat{\mu}^{IPSW}$ had small bias, it was inefficient. The $\hat{\mu}^{IPSW.S}$, by fitting the propensity model to the scaled weighted sample, halved the variance of $\hat{\mu}^{IPSW}$, without increasing the bias. The extended KW estimator, $\hat{\mu}^{KW.W}$, also had smaller variance than $\hat{\mu}^{IPSW}$ because the estimated PSs were used to measure the similarity instead of estimating

Table 4.1 Results of from 10,000 simulated cohorts and survey samples under SEA.

Estimator	%RB	V ($\times \mathbf{10^3}$)	VR(TL)	VR(JK)	CP(TL)	CP(JK)	MSE ($\times \mathbf{10^3}$)
$\hat{\mu}^{Naive}$	20.97	1.71	1.02		0.00		889.76
$\hat{\mu}^{SVY}$	0.07	2.82	1.01	1.02	0.96	0.96	2.83
Model T (True) logit{Pr($\boldsymbol{x}$)} ~ x_1, x_2, x_4							
$\hat{\mu}^{IPSW}$	-0.12	9.96	0.92	1.01	0.94	0.95	9.99
$\hat{\mu}^{IPSW.S}$	0.07	4.59	0.99	0.99	0.95	0.95	4.60
$\hat{\mu}^{KW}$	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

participation rates. The KW.S estimator $\hat{\mu}^{KW.S}$, which used the scaled survey weights to estimate the PSs, further reduced the variance and improved MSE compared to $\hat{\mu}^{KW.W}$. The KW estimate, $\hat{\mu}^{KW}$ had the smallest variance because no sample weights were considered in estimating PSs. The $\hat{\mu}^{KW}$ required the SEA, which held in this scenario, and thus naturally had the smallest MSE and maintained the nominal CP in Scenario 1. The TL and JK methods gave similar variance estimates.

4.7.5 Results under Scenario 2: the invalid SEA

We changed the MOS for the sample selection of s_s while the sample selection of s_c remained the same as given in Section 4.7.4 so that the SEA was invalid (Scenario 2 in Figure 2). We discuss the results under four propensity models that includes different sets of covariates.

Correct propensity Model T with $\boldsymbol{x} = (x_1, x_2, x_4)$

As shown in Table 4.2 under the correctly specified propensity model (*Model T*), though the KW estimate, $\hat{\mu}^{KW}$, had the smallest variance among the five pseudo-weighted estimates, 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 4.1, using scaled 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.

Underfitted propensity Model U with $\boldsymbol{x} = (x_1, x_2)$

Model U was an underfitted propensity model where the covariate x_4 was missing. Excluding x_4 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 to the variances under *Model* T (same findings as in Chapter 3, and Stuart, 2010). In contrast, the variances of $\hat{\mu}^{IPSW.S}$ and $\hat{\mu}^{KW.S}$ were nearly unchanged.

Misspecified propensity Model M_1 *with* $\boldsymbol{x} = (x_1^*, x_2)$

In *Model* M_1, the true covariate x_1 in *Model* U was substituted by x_1^* which was a nonlinear function of x_1. The IPSW estimates $\hat{\mu}^{IPSW}$ and $\hat{\mu}^{IPSW.S}$ were biased because the cohort participation rates cannot be accurately estimated from *Model* M_1. However, the matching methods with matching scores $\hat{q} = \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2$ still worked well because $x_1^* = x_1 + 0.15x_1^3$ was a one-to-one function of x_1, and therefore $\hat{q}$ was close to a one-to-one function of the true participation rate. As a result, the KW estimates were less biased than the IPSW estimates. Furthermore, using scaled survey weights, $\hat{\mu}^{KW.S}$ outperformed $\hat{\mu}^{IPSW.S}$ with smaller bias, variance, and nearly nominal CP.

Misspecified propensity Model M_2 *with* $\boldsymbol{x} = (x_1^{**}, x_2)$

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 misspecified 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 estimates by matching methods were biased. For this scenario, $\hat{\mu}^{IPSW.S}$ had smaller MSE and more accurate CP than $\hat{\mu}^{KW.S}$ due to the smaller bias.

TL and JK variance estimation

The TL variance estimates were close to the truth (with the VR close to 1) for all estimates of μ^{FP} except for $\hat{\mu}^{IPSW}$ with its corresponding VR<<1. This is due to the finite sample bias caused by large variability of the sample weights in the combined (s_c vs. weighted s_s) sample (with the common value of one for the cohort weights vs. values ranging from 23 to 618 for the survey weights.) The coefficient of variance (CV) of the weights was ~160% in the combined sample, indicating highly variable sample weights. The finite sample bias of the TL variance estimator for $\hat{\mu}^{IPSW}$ became smaller as N, n_c, and n_s increased (results not shown). This is consistent with the previous findings from population-based case-control studies where the TL method underestimates the variance of logistic regression coefficients because of the large variability in the sample weights for combined sample of cases and controls (Li et al. 2010; Landsman & Graubard, 2012). In contrast, the TL variance estimate for $\hat{\mu}^{IPSW.S}$ worked well since the variability of the weights in the combined (s_c vs. scaled-weighted s_s) sample was reduced. The scaled survey sample weights range from 0.2 to 5.1 and the CV of weights decreases to 50%.

The JK method consistently had larger estimates of variances compared to the TL variance estimates (similar results were shown by Efron & Gong, 1983), and the JK estimates were more accurate for the variance of $\hat{\mu}^{IPSW}$. However, in some of the simulations the JK overestimated the variance of $\hat{\mu}^{KW}$ and $\hat{\mu}^{KW.S}$. Under *Model* M_1, values of the matching score, i.e., $\hat{q}$ can be quite different across the replicates due to the highly variable covariate x_1^* in the propensity model, which could slow down the convergence of the JK variance estimates.

In summary, scaling the survey weights not only substantially 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 misspecification, 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.

Table 4.2 Results from 10,000 simulated cohorts and survey samples with each cohort and survey sample fitted to the correct propensity model and three misspecified propensity models with violated SEA.

Estimator	**%RB**	**V ($\times$ $\mathbf{10^3}$)**	**VR (TL)**	**VR (JK)**	**CP (TL)**	**CP (JK)**	**MSE ($\times$ $\mathbf{10^3}$)**
$\hat{\mu}^{Naive}$	20.97	1.72	1.02		0.00		889.89
$\hat{\mu}^{SVY}$	0.04	3.61	1.02	1.02	0.95	0.95	3.62
Model $\boldsymbol{T}$ (True) logit$\{\Pr(\boldsymbol{x})\} \sim x_1, x_2, x_4$							
$\hat{\mu}^{IPSW}$	-0.36	14.73	0.86	1.03	0.93	0.94	14.99
$\hat{\mu}^{IPSW.S}$	0.03	5.92	0.98	1.00	0.95	0.95	5.92
$\hat{\mu}^{KW}$	4.84	2.66	0.92	1.04	0.01	0.02	50.03
$\hat{\mu}^{KW.W}$	0.84	4.83	1.04	1.09	0.93	0.94	6.24
$\hat{\mu}^{KW.S}$	0.65	3.55	1.02	1.08	0.93	0.93	**4.39**
Model $\boldsymbol{U}$ (Underfitted) logit$\{\Pr(\boldsymbol{x})\} \sim x_1, x_2$							
$\hat{\mu}^{IPSW}$	-0.24	13.62	0.89	1.02	0.93	0.94	13.73
$\hat{\mu}^{IPSW.S}$	0.04	5.76	0.98	1.00	0.95	0.95	5.77
$\hat{\mu}^{KW.W}$	0.80	3.92	1.12	1.13	0.94	0.94	5.22
$\hat{\mu}^{KW.S}$	0.57	3.39	1.02	1.08	0.93	0.94	**4.05**
Model $\boldsymbol{M_1}$ (Misspecified variable) logit$\{\Pr(\boldsymbol{x})\} \sim x_1^*, x_2$							
$\hat{\mu}^{IPSW}$	4.91	16.94	0.62	1.08	0.44	0.55	65.61
$\hat{\mu}^{IPSW.S}$	3.22	7.10	0.95	1.00	0.58	0.58	28.04
$\hat{\mu}^{KW.W}$	0.58	4.54	1.08	1.45	0.87	0.95	5.22
$\hat{\mu}^{KW.S}$	0.52	3.18	0.94	1.29	0.92	0.96	**3.72**
Model $\boldsymbol{M_2}$ (Misspecified variable) logit$\{\Pr(\boldsymbol{x})\} \sim x_1^{**}, x_2$							
$\hat{\mu}^{IPSW}$	1.54	7.46	0.98	1.00	0.86	0.86	12.28
$\hat{\mu}^{IPSW.S}$	1.58	4.58	0.99	1.00	0.82	0.82	**9.62**
$\hat{\mu}^{KW.W}$	2.08	4.10	0.99	1.12	0.70	0.75	12.87
$\hat{\mu}^{KW.S}$	2.04	3.62	0.92	1.10	0.65	0.72	12.01

4.8 Data Analysis: The U.S. National Health and Nutrition Examination Survey

Note that, even if common variables are available in the cohort and the survey, and the propensity model is appropriately specified, there can be many other factors influencing performance of the proposed KW.S and IPSW.S estimates. First, the cohort may produce quite different estimates from the survey estimates even if the implicit self-selection weights are known. This can be caused by sampling errors in the estimates, under-coverage of cohort study centers for the finite population, and measurement errors due to different data collection modes or questionnaires conducted by the cohort and the survey. Second, different questionnaires between the cohort and the survey can make data harmonization difficult and imprecise, resulting in biased estimation. For example, a question about the same topic may provide respondents with different categories to select in their responses between the cohort and the survey.

In this analysis, in order to reduce the influence of these factors, we used the Third U.S. National Health and Nutrition Examination Survey (NHANES III) as the volunteer-based "cohort" (ignoring sample weights) and the contemporaneous U.S. National Health Interview Survey (NHIS) as the reference survey. Although this example is contrived, it has a key advantage for illuminating the performance of our methodology, namely that 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", we could potentially truly recover US-representative estimates, and thus enables us to characterize the performance of our methodology in real data. Although problems with misaligned target populations and data harmonization 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 US using the adult sample of the 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-institutionalized population of the United States. NHANES III oversampled poverty areas, children under age 5, adults age 60 and over, non-Hispanic blacks, and Mexican Americans (Ezzati et al., 1992). The CV of sample weights is 125%, indicating highly variable selection probabilities, and potential low representativeness of the unweighted sample. We ignored all complex design features of NHANES III to treat 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 = 19738). NHIS is also a cross-sectional household interview survey with 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 (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

(NCHS 2013), allowing us to quantify the relative bias of unweighted NHANES estimates, assuming that the NHIS estimates are the gold standard.

We first compared the distributions of selected common covariates in the two samples (Table 4.3). 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; (2) health status, which was self-reported in NHANES but reported by the proxy of the household

Table 4.3 Distribution of selected common variables in NIH-AARP and NHIS

		NHIS 1994		**NHANES III**	
	Total Count	n =19738	$\widehat{N}$ =189608549	n =20050	$\widehat{N}$ =187647206
		%	Weighted %	%	Weighted %
Age Group	18-24 years	10.5	13.3	15.8	15.8
	25-44 years	42.9	43.7	35.4	43.7
	45-64 years	26.1	26.6	22.6	24.6
	65 years and older	20.5	16.4	26.2	16.0
Race	NH-White	76.1	75.9	42.3	76.0
	NH-Black	12.6	11.2	27.4	11.2
	Hispanic	8.0	9.0	28.9	9.3
	NH-Other	3.3	4.0	1.5	3.5
Region	Northeast	20.7	20.5	14.6	20.8
	Midwest	26.1	25.1	19.2	24.1
	South	31.5	32.5	42.7	34.3
	West	21.6	21.9	23.5	20.9
Poverty	No	79.1	82.3	67.9	80.3
	Yes	13.1	10.6	21.4	12.1
	Unknown	7.8	7.0	10.7	7.6
Education	Lower than High school	60.5	62.0	39.0	51.6
	High School/Some College	25.7	25.7	35.9	32.7
	College or higher	13.8	12.3	25.1	15.7
Health Status (Self-Rprtd)	Excellent/Very good	60.5	62.0	39.0	51.6
	Good	25.7	25.7	35.9	32.7
	Fair/Poor	13.8	12.3	25.1	15.7

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 used 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), socioeconomic 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 4.4 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).

Table 4.4 Main effects of the fitted Propensity score model with or without NHIS sampling weights (interactions not shown)

	Weighted Sample[1]		Scale-weighted Sample[2]		Unweighted Sample[3]	
Coefficients	Estimate[4] (× 100)	Std. Err.[5] (× 100)	Estimate (× 100)	Std. Err. (× 100)	Estimate (× 100)	Std. Err. (× 100)
(Intercept)	-718.0	35.95***	137.8	26.89***	250.7	24.89***
Age (in years)	-11.8	1.97***	-9.7	1.48***	-16.7	1.34***
Age^2	0.2	0.04***	0.2	0.03***	0.3	0.03***
Sex (ref: male)						
Female	-1.5	3.63	2.1	3.27	-10.2	3.14**
Race/Ethnicity (ref: NH-White)						
NH-Black	224.3	16.50***	216.5	15.31***	267.2	13.77***
Hispanic	23.4	21.33	5.7	18.41	28.5	16.70•
NH-Other	21.2	36.27	16.6	34.05	28.6	32.54
Region (ref: Northeast)						
Midwest	-31.3	13.89*	-46.2	10.78***	-47.7	10.16***
South	22.8	11.26*	11.3	9.98	16.9	9.32•
West	-56.1	14.46***	-74.2	11.82***	-77.3	11.10***
Marital Status (ref: married or living as married)						
Previously married	-35.3	10.11***	-13.8	7.79•	-68.3	7.31***
Never married	-5.3	12.00	-7.2	8.92	-75.2	8.15***
Education level	-33.6	3.57***	-24.9	2.73***	-23.8	2.57***

Poverty (ref: No)						
Yes	-105.6	18.09***	-103.4	12.83***	-93.5	11.68***
Unknown	-438.3	39.76***	-384.5	22.10***	-360.0	19.67***
Household Income	-26.4	2.80***	-23.7	2.22***	-14.7	2.08***
BMI (ref: normal)						
Under-weight	3.6	8.64	-9.7	7.25	-4.2	6.81
Over-weight	2.7	4.00	0.4	2.89	0.7	2.67
Obese	-0.4	4.61	-4.5	3.62	-4.6	3.38
Health Status	40.4	4.67***	47.0	3.77***	48.6	3.50***
Smoking (ref: Non-smoker)						
Current smoker	17.3	4.36***	11.0	3.24***	9.8	2.99**
Former smoker	13.9	4.43**	13.3	3.16***	8.3	2.99**
Chewing tobacco (ref: No)						
Yes	-19.1	5.24***	-25.5	4.28***	-23.0	4.04***

[1]Weighted Sample: The combined NHANES and weighted NHIS sample, with the weights being the NHIS population weights. The fitted propensity model is used by the IPSW and KW.W approaches.

[2]Scale-weighted Sample: The combined NHANES and scale-weighted NHIS sample, with the weights being the scaled NHIS population weights. The fitted propensity model is used by the IPSW.S and KW.S approaches.

[3]Unweighted Sample: The combined NHANES and weighted NHIS sample. The fitted propensity model is used by the KW approach.

[4]Estimates: Estimated model coefficients on log-odds scale.

[5]Std. Err.: Square root of estimated variance of estimated model coefficients.

'***' p-value< 0.001; '**' p-value< 0.01; '*' p-value< 0.05; '.' p-value< 0.1.

The 11 two-way interactions included in the propensity models are poverty: household income, race/ethnicity: region, age^2: race/ethnicity, race/ethnicity: health status, age: health status, region: household income, marital status: household income, age^3, education level: household income, race/ethnicity: poverty, age^2: poverty, sex: race/ethnicity

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

Table 4.5 shows that the weighted 1994 NHIS and the weighted NHANES III estimates (TW) of 15-year all-cause mortality were very close (%RD = 2.6% for overall

Table 4.5 Estimates of all-cause 15-year mortality (overall, and by subgroups)

	Est	%Relative Difference from the NHIS Estimate						
	NHIS	TW	Naïve	KW	IPSW	IPSW.S	KW.W	KW.S
Overall	17.6	-2.6	52.2	17.7	-3.8	-3.2	**-2.0**	-2.2
(%BR)				(66.0)	(92.7)	(93.8)	(**96.1**)	(95.8)
Age group								
18-24 yrs	2.2	-16.1	**0.5**	-35.9	-33.4	-30.9	-32.4	**-30.2**
25-44 yrs	3.9	-7.9	30.9	**-4.5**	-14.8	-14.3	**-12.5**	-14.0
45-64 yrs	17.7	5.8	30.6	**1.3**	-3.8	**-3.2**	-4.1	-3.7
65-69 yrs	45.5	0.9	9.5	**-1.3**	-6.4	-5.4	-5.8	**-4.7**
70-74 yrs	60.0	3.5	6.4	**-0.4**	-1.4	-1.3	-1.3	**-1.1**
>=75 yrs	86.2	1.1	4.3	3.3	3.3	3.2	3.2	**3.1**
Average		5.9	13.7	**7.8**	10.5	9.7	9.9	**9.4**
(%BR)				**(43.2)**	(23.3)	(29.2)	(27.7)	**(31.0)**
Sex								
Male	18.8	-7.1	58.1	15.9	**0.0**	0.8	1.8	2.0
Female	16.5	1.9	46.5	21.0	-6.5	-6.8	**-4.8**	-5.9
Average		4.5	52.3	18.4	**3.3**	3.8	3.3	3.9
(%BR)				(64.8)	**(93.8)**	(92.7)	(93.7)	(92.5)
Race								
NH-White	18.7	-1.7	96.8	17.8	-1.9	-1.8	**-0.2**	-0.8
NH-Black	18.9	-5.6	19.4	17.3	-4.4	-6.9	**-4.0**	-7.0
Hispanic	10.2	-9.1	62.2	13.2	-15.3	-9.3	-13.0	**-7.9**
NH-Other	9.0	-12.8	63.8	**-14.1**	-31.1	-23.8	-24.2	**-19.0**
Average		7.3	60.5	15.6	13.2	10.5	10.4	**8.7**
(%BR)				(74.2)	(78.2)	(82.7)	(82.9)	**(85.7)**

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 ~52.2% biased from the NHIS estimate because older people who have higher mortalities were oversampled, and the bias insubgroup-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 close estimates. The bias in the naïve estimate of overall mortality was almost eliminated by the KW.W and KW.S methods (~96.1% and 95.8% 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.

Table 4.6 Taylor Linearization variance estimates and mean squared errors of the of all-cause 15-year mortality estimates (overall, and by subgroups)

	TL Variance Estimate ($\times 10^5$)					**MSE ($\times 10^5$)**				
	KW	IPSW	IPSW.S	KW.W	KW.S	KW	IPSW	IPSW.S	KW.W	KW.S
Overall	1.2	1.4	1.0	1.0	**1.0**	98.2	5.8	4.3	**2.3**	2.5
Age group										
18-24 yrs	**0.8**	0.8	0.9	0.8	0.9	7.2	6.4	5.7	6.1	**5.5**
25-44 yrs	0.8	0.7	0.7	0.7	**0.7**	**1.1**	4.0	3.8	**3.1**	3.6
45-64 yrs	5.2	5.4	5.1	**4.9**	4.9	**5.7**	9.8	**8.2**	10.2	9.2
65-69 yrs	35.9	35.6	34.4	33.5	**33.3**	**39.5**	120.8	95.1	103.3	**78.1**
70-74 yrs	32.1	31.6	30.3	30.4	**29.6**	**32.7**	39.2	36.9	36.9	**33.9**
>=75 yrs	6.0	6.1	5.8	5.9	**5.8**	85.5	85.7	80.2	83.4	**76.2**
Average	13.5	13.4	12.9	12.7	**12.5**	**28.6**	44.3	38.3	40.5	**34.4**
Sex										
Male	2.9	2.8	2.5	**2.4**	2.4	91.7	2.8	**2.7**	3.5	3.8
Female	2.1	2.2	1.7	1.7	**1.7**	121.7	13.7	14.3	**7.9**	11.1
Average	2.5	2.5	2.1	2.1	**2.0**	106.7	8.3	8.5	**5.7**	7.4
Race										
NH-White	1.9	2.2	1.7	1.6	**1.6**	112.1	3.5	2.8	**1.6**	1.8
NH-Black	3.8	6.6	5.7	3.0	**2.7**	110.9	13.6	22.9	**8.6**	20.3
Hispanic	2.6	6.1	5.7	**1.8**	1.9	20.7	30.5	14.5	19.4	**8.3**
NH-Other	19.2	14.6	14.4	**13.2**	14.4	**35.2**	92.1	59.9	60.2	**43.5**
Average	6.9	7.4	6.9	**4.9**	5.1	69.7	34.9	25.1	22.5	**18.5**

Interestingly, 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. This paradox is caused by the validity of the SEA for the age-specific mortality estimation, but not for overall mortality estimation. As shown in Table 4.5, the small biases in the KW estimates of age-specific mortality imply that the SAE held, meaning $E(y \mid \text{age}, \tilde{p}, s_c) \doteq E(y \mid \text{age}, \tilde{p}, FP)$. As shown in Table 4.7, the KW pseudo-weighted age distribution in s_c (unweighted NHANES sample) differed from that in FP (represented by the weighted NHIS), indicating $P(\text{age} \mid \tilde{p}, s_c) \neq$

$P(\text{ age } \mid \tilde{p}, FP\,)$. As a result, the SEA was invalid for the overall mortality estimation using the original KW method, that is

$$E(\,y \mid \tilde{p}, s_c\,) = \sum_{\text{age}}\{E(\,y \mid \text{age}, \tilde{p}, s_c\,)P(\text{ age } \mid \tilde{p}, s_c\,)\}$$

$$\neq \sum_{\text{age}} E(\,y \mid \text{age}, \tilde{p}, FP\,)P(\text{ age } \mid \tilde{p}, FP\,) = E(\,y \mid \tilde{p}, FP\,).$$

Table 4.7 Relative difference of age group proportion estimates from the 1994 NIHS estimates

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

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 other four methods (IPSW, IPSW.S, KW.W, and KW.S) had similar estimated mortality rates. The IPSW estimates had the largest variances, followed by the IPSW.S estimates. The KW.S estimates had the smallest variances with the smallest MSE in most cases. The results of the Jackknife replication and the TL variance estimates were similar in this real data example (results not shown).

4.9 Summary

In this Chapter, a unifying framework is established for both PS-based weighting and matching methods to improve estimates of finite population means from non-representative cohort data, by using a reference representative survey sample of the target population. Three contributions are made under this unifying framework. First, the underlying Strong Exchangeability Assumption (SEA) assumed by the existing PS-based matching methods is identified. The simulations and data example demonstrate that the PS-based matching methods that rely on the SEA, such as the original KW estimator introduced in Chapter 3, have smallest MSE when the SEA holds, but have large bias when the SEA fails. As in our data example, SEA failed for estimating overall mortality, but held for age-specific mortality estimation. Second, as a remedy, PS-based matching methods are proposed without requiring the SEA, but a Weak Exchangeability Assumption (WEA). Third, the efficiency of PS-based estimates is further improved by scaling the survey weights to sum to the survey sample size. Scaling reduces the variance of the estimated PSs and thus markedly improves efficiency of the pseudo-weighted estimates, especially for the IPSW method. The recommended method, kernel-weighting with scaling (KW.S), is more robust by only requiring the WEA, yet has smallest MSE.

For the variance estimation, the JK method is recommended 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.

The 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-life. For example, most cohort studies only recruit people in a few study centers in a target population (e.g., the US), while 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.

4.10 Proofs

We consider the following limiting process (Krewski and Rao 1981; Chen et al., 2019)for the theoretical justification of **Theorem 4.1**. Suppose there is a sequence of finite population FP_k of size N_k, for $k = 1, 2, \cdots$. Cohort $s_{c,k}$ of size $n_{c,k}$ and survey sample $s_{s,k}$ of size $n_{s,k}$ are sampled from each FP_k. The sequences of the finite population, the cohort and the survey sample have their sizes satisfy $\lim_{k\to\infty} \frac{n_{d,k}}{N_k} \to \gamma_d$ where $d = c, s$ and $0 < \gamma_d \leq 1$. In the following the index k is suppressed for simplicity.

4.10.1 Fundamental Assumptions for Cohort Participation

A1. The cohort participation indicator $\delta^{(c)}$ is independent of the outcome variable y given the covariates $\boldsymbol{x}$, i.e., $\Pr\big(\,\delta^{(c)} = 1 \mid y, \boldsymbol{x}\,\big) = \Pr\big(\,\delta^{(c)} = 1 \mid \boldsymbol{x}\,\big)$.

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

4.10.2 Regularity Conditions

C1 The FP and the sampling design for selecting s_s satisfy $N^{-1}\sum_{i\in s_s} d_i \boldsymbol{v}_i - N^{-1}\sum_{i\in FP} \boldsymbol{v}_i = O_p\big(n_s^{-1/2}\big)$ with $\boldsymbol{v}_i$ being a function of $\boldsymbol{x}_i$ and outcome y_i.

C2 For the probability of being randomly selected in the survey sample, $\pi_i^{(s)}$, and self-selected in the cohort, $\pi_i^{(c)}$, there exists an $M \in \mathbb{R}_{>0}$ such that $1 < {\pi_i^{(s)}}^{-1} < M$, and $1 < {\pi_i^{(c)}}^{-1} < M$, for $i \in FP$.

C3 The kernel function $K(u)$ satisfies $\int K(u)du = 1$, $\sup_u |K(u)| < \infty$, and $\lim_{|u|\to\infty} |u| \cdot |K(u)| = 0$.

C4 The bandwidth corresponding to $K(u)$, $h = h(n_c)$, satisfies the conditions that $h \to 0$ and $n_c \cdot h \to \infty$ as $n_c \to \infty$.

C5 y has bounded second moment in the FP, i.e, $N^{-1}\sum_{i\in FP} y_i^2 = O(1)$, and $\boldsymbol{x}$ has bounded third moment in FP, i.e., $N^{-1}\sum_{i\in FP}\|\boldsymbol{x}_i\|^3 = O(1)$,

C6 The $\boldsymbol{x}$, and the propensity score p_i in Model (4.10.1) below satisfy $N^{-1}\sum_{i\in FP} p_i^2 \boldsymbol{x}_i \boldsymbol{x}_i^T$ is positive definite.

C7 The $\boldsymbol{x}$, and the propensity score p_i^* in Model (4.10.5) below satisfy $\frac{a}{N}\sum_{i\in FP} p_i^{*2} \boldsymbol{x}_i \boldsymbol{x}_i^T$ is positive definite.

C8 The cohort participation and survey sample selection are uncorrelated given $\boldsymbol{x}$, i.e., $cov\left(\delta_i^{(c)}, \delta_j^{(c)} \mid \boldsymbol{x}_i, \boldsymbol{x}_j \right) = 0$ for $i, j \in FP$.

C9 The cohort participation are uncorrelated given $\boldsymbol{x}$, i.e., $cov\left(\delta_i^{(c)}, \delta_j^{(c)} \mid \boldsymbol{x}_i, \boldsymbol{x}_j \right) = 0$ for $i \neq j$.

C1 and **C2** are regularity conditions for sample selection and finite population inference that are commonly used. **C1** gives the rate of convergence of the estimated means of the y and the $\boldsymbol{x}$. **C2** indicates the (self-) selection rates of cohort and sample inclusion rates of the survey are asymptotically bounded. **C3** and **C4** are standard conditions for kernel function and bandwidth in kernel regression (Noda, 1976). **C5-C7** are standard conditions involving bounded moments that are used to obtain consistent estimators of FP means and the TL FP variance. **C8** and **C9** assume uncorrelated selection between cohort and survey sample and uncorrelated cohort participation in the FP respectively. **C1**-**C9** are used for deriving the closed form of TL FP variance of the FP mean estimators from the cohort.

4.10.3 Proof of Theorem 4.1

Consistency of $\hat{\mu}^{KW.W}$ is derived in a similar way as the proof **Theorem 4.1** under **C1-C5**. Notice that the consistency does not require modeling of the propensity score $p = P(i \in s_c \mid s_c \cup^* FP)$. To obtain the finite population variance, we assume the logistic regression model

$$\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{\beta}^T \boldsymbol{x}_i, i \in s_c \cup^* FP, \tag{4.10.1}$$

and consider the following system of estimating equations

$$\Phi(\boldsymbol{\eta}) = \begin{pmatrix} U(\mu^{FP}) = \frac{1}{N}\sum_{i\in FP} \delta_i^{(c)} w_i^{KW.W}(y_i - \mu^{FP}) \\ S(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i\in FP} \delta_i^{(c)}(1-p_i)\boldsymbol{x}_i - \frac{1}{N}\sum_{i\in FP} \delta_i^{(s)} d_i p_i \boldsymbol{x}_i \end{pmatrix} = \mathbf{0} \tag{4.10.2}$$

where $\boldsymbol{\eta} = (\mu^{FP}, \boldsymbol{\beta})$, the KW.W pseudo-weight $w_i^{KW.W}$ for $i \in FP$ is a function of $\boldsymbol{\beta}$. The solution of estimating equation (4.10.2) is denoted by $\widehat{\boldsymbol{\eta}} = (\hat{\mu}, \widehat{\boldsymbol{\beta}})$. Under **C1**, **C2**, **C5**, using the first order Taylor expansion, we have

$$\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta} = -[E\{\phi(\boldsymbol{\eta})\}]^{-1}\Phi(\boldsymbol{\eta}) + o_p\left(n_c^{-1/2}\right)$$

where $\phi(\boldsymbol{\eta}) = \partial\Phi(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$, and $[E\{\phi(\boldsymbol{\eta})\}]^{-1} = \begin{pmatrix} U_{\mu^{FP}}^{-1} & \boldsymbol{b}^T \\ \mathbf{0} & S_{\boldsymbol{\beta}}^{-1} \end{pmatrix}$, with

$U_{\mu^{FP}} = E(\partial U/\partial \mu^{FP}) = -\frac{1}{N}\sum_{i\in FP}\pi_i^{(c)} w_i^{KW.W} \doteq - 1$ as $w_i^{KW.W} \doteq {\pi_i^{(c)}}^{-1}$,

$U_{\boldsymbol{\beta}} = E(\partial U/\partial \boldsymbol{\beta}^T) = \frac{1}{N}\sum_{i\in FP}\pi_i^{(c)}(y_i - \mu^{FP})\frac{\partial w_i^{KW.W}}{\partial \boldsymbol{\beta}^T}$,

$$S_{\boldsymbol{\beta}} = E(\partial S/\partial \boldsymbol{\beta}) = -\frac{1}{N}\sum_{i\in FP}\pi_i \cdot p_i(1-p_i)\boldsymbol{x}_i\,\boldsymbol{x}_i^T - \frac{1}{N}\sum_{i\in FP} p_i(1-p_i)\boldsymbol{x}_i\boldsymbol{x}_i^T$$

$$= -\frac{1}{N}\sum_{i\in FP} p_i^2\boldsymbol{x}_i\,\boldsymbol{x}_i^T - \frac{1}{N}\sum_{i\in FP} p_i(1-p_i)\boldsymbol{x}_i\boldsymbol{x}_i^T$$

$= -\frac{1}{N}\sum_{i\in FP} p_i\boldsymbol{x}_i\boldsymbol{x}_i^T$, is negative definite under **C6** and thus invertible,

and $\boldsymbol{b}^T = -U_{\mu^{FP}}^{-1} U_{\boldsymbol{\beta}} S_{\boldsymbol{\beta}}^{-1}$.

It follows that

$$Var(\widehat{\boldsymbol{\eta}}) = [E\{\phi(\boldsymbol{\eta})\}]^{-1} Var\{\Phi(\boldsymbol{\eta})\}[E\{\phi(\boldsymbol{\eta})\}^T]^{-1} + o_p(n_c^{-1}), \qquad (4.10.3)$$

We decompose $Var\{\Phi(\boldsymbol{\eta})\}$, denoted by V_Φ, into two parts under **C8**:

$$V_\Phi = Var\begin{pmatrix} 0 \\ \frac{1}{N}\sum_{i\in FP}\delta_i^{(s)} d_i p_i \boldsymbol{x}_i \end{pmatrix} + Var\begin{Bmatrix} \frac{1}{N}\sum_{i\in FP}\delta_i^{(c)} w_i^{KW.W}(y_i - \mu^{FP}) \\ \frac{1}{N}\sum_{i\in FP}\delta_i^{(c)}(1-p_i)\boldsymbol{x}_i \end{Bmatrix},$$

The first summand, defined by V_1, involves s_s selection only: $V_1 = \begin{pmatrix} 0 & \boldsymbol{0}^T \\ \boldsymbol{0} & D \end{pmatrix}$ with $D = N^{-2} V_p\left(\sum_{i\in FP}\delta_i^{(s)} d_i p_i \boldsymbol{x}_i\right)$, where V_p denotes the design-based finite population variance under the probability sampling design for s_s. The second summand V_2 involves s_c selection only. Under **C9**,

$$V_2 = N^{-2}\sum_{i\in FP}\pi_i^{(c)}\left(1-\pi_i^{(c)}\right)\begin{Bmatrix} \left(w_i^{KW.W}\right)^2 (y_i - \mu^{FP})^2 & (1-p_i) w_i^{KW.W}(y_i - \mu^{FP})\boldsymbol{x}_i^T \\ (1-p_i) w_i^{KW.W}\left(y_i - \mu^{FP}\right)\boldsymbol{x}_i & (1-p_i)^2 \boldsymbol{x}_i \boldsymbol{x}_i^T \end{Bmatrix}.$$

Based on Equality (4.10.3), the finite population variance of $\hat{\mu}$ is given by

$$Var(\hat{\mu}^{KW.W}) = V^{KW.W} + o_p(n_c^{-1}), \text{ with}$$

$$V^{KW.W} = \begin{pmatrix} U_{\mu^{FP}}^{-1} & \boldsymbol{b}^T \end{pmatrix} \cdot V_\Phi \cdot \begin{pmatrix} U_{\mu^{FP}}^{-1} \\ \boldsymbol{b} \end{pmatrix} \qquad (4.10.4)$$

$$= \frac{1}{N^2} \sum\nolimits_{i \in FP} \pi_i^{(c)} \left(1 - \pi_i^{(c)}\right) \{ w_i^{KW.W} (y_i - \mu^{FP}) - (1 - p_i) \boldsymbol{b}^T \boldsymbol{x}_i \}^2 + \boldsymbol{b}^T D \boldsymbol{b}.$$

In sample estimate of $V^{KW.W}$, with the standard normal density as the kernel function, $K(u) \propto \exp\left(\frac{u^2}{2}\right)$, we have $w_i^{KW.W} = \sum_{j \in s_s} d_i \frac{e_{ij}}{\sum_{i \in s_c} e_{ij}}$, and $\frac{\partial w_i^{KW.W}}{\partial \boldsymbol{\beta}^T} = \sum_{j \in s_s} d_i \left\{ \frac{e_{ij} \cdot \partial e_{ij} / \partial \boldsymbol{\beta}^T}{\sum_{i \in s_c} e_{ij}} - \frac{e_{ij} \cdot \sum_{i \in s_c} (e_{ij} \cdot \partial e_{ij} / \partial \boldsymbol{\beta}^T)}{\left(\sum_{i \in s_c} e_{ij}\right)^2} \right\}$, where $e_{ij} = \exp\left\{ \frac{1}{2h^2} \left(\boldsymbol{\beta}^T \boldsymbol{x}_i - \boldsymbol{\beta}^T \boldsymbol{x}_j \right)^2 \right\}$, and $\frac{\partial e_{ij}}{\partial \boldsymbol{\beta}^T} = e_{ij} \cdot \frac{1}{h^2} \left(\boldsymbol{\beta}^T \boldsymbol{x}_i - \boldsymbol{\beta}^T \boldsymbol{x}_j \right) \cdot \left(\boldsymbol{x}_i - \boldsymbol{x}_j \right)^T$.

Remark: Notice that the finite population variance of the IPSW estimate of finite population mean, $\hat{\mu}^{IPSW}$, can be obtained by replacing $w_i^{KW.W}$ and $\frac{\partial w_i^{KW.W}}{\partial \boldsymbol{\beta}^T}$ in by w_i^{IPSW} and $\frac{\partial w_i^{IPSW}}{\partial \boldsymbol{\beta}^T} = -w_i^{IPSW} \boldsymbol{x}_i^T$ respectively in (4.10.4).

4.10.4 Proof of Lemma 4.1

Under the logistic regression Model (4.10.5) fitted to the combined sample of s_c vs. s_s weighted by the *scaled* sample weights,

$$\log\left(\frac{p_i^*}{1 - p_i^*} \right) = \boldsymbol{\beta}^{*T} \boldsymbol{x}_i, \qquad (4.10.5)$$

the propensity score p_i^* is different from $p_i = P(\, i \in s_c \mid s_c \cup^* FP\,)$ in Model (4.10.1), and can be defined as $p_i^* = P(\, i \in s_c \mid s_c \cup^* S\,)$, where S is a given simple random sample selected from FP with a sampling rate a. Then the relationship between p_i^* and the cohort participation rate $\pi_i^{(c)}$ can be obtained by

$$\frac{p_i^*}{1 - p_i^*} = \frac{P(\, i \in s_c \mid s_c \cup^* S\,)}{P(\, i \in S \mid s_c \cup^* S\,)} = \frac{P(\, i \in s_c \mid FP\,)}{P(\, i \in S \mid FP\,)} = \frac{\pi_i^{(c)}}{a}$$

Meanwhile, we know the cohort participation rate $\pi_i^{(c)} = \frac{p_i}{1 - p_i}$ under Model (4.10.1). Hence, Model (4.10.5) can be re-parameterized as

$$\begin{aligned}\log\left(\frac{p_i^*}{1-p_i^*}\right) &= \log\left(\frac{\pi_i^{(c)}}{a}\right) = \log\left\{\frac{p_i}{a(1-p_i)}\right\} \\ &= -\log a + \boldsymbol{\beta}^T \boldsymbol{x}_i\end{aligned} \qquad (4.10.6)$$

where $\boldsymbol{\beta}$ is the vector of coefficients in Model (4.10.1). Comparing (4.10.5) and (4.10.6) gives $\boldsymbol{\beta}^* = \boldsymbol{\beta} + \log a \cdot \boldsymbol{e}_1$, with $\boldsymbol{e}_1 = (1, 0, \cdots, 0)^T$that is the result of **Lemma 4.1**.

4.10.5 Proof of Theorem 4.2

The properties of $\hat{\mu}^{IPSW.S}$ and $\hat{\mu}^{KW.S}$ can be proved via the estimating equations

$$\Phi^*(\boldsymbol{\eta}) = \begin{pmatrix} U^*(\mu^{FP}) = \frac{1}{N}\sum_{i\in FP} \delta_i^{(c)} \widetilde{w}_i^* (y_i - \mu^{FP}) \\ S^*(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i\in FP} \delta_i^{(c)} (1-p_i^*)\boldsymbol{x}_i - \frac{a}{N}\sum_{i\in FP} \delta_i^{(s)} d_i p_i^* \boldsymbol{x}_i \end{pmatrix} = \boldsymbol{0}, \quad (4.10.7)$$

where $\widetilde{w}_i^*$ is the pseudo sample weight $w_i^{IPSW.S}$ or $w_i^{KW.S}$. Under Conditions **C1-C5**, we have $\Phi^*(\widehat{\boldsymbol{\eta}}^*) = \boldsymbol{0}$, and $\Phi^*(\boldsymbol{\eta}^*) = O_p\left(n_c^{-1/2}\right)$, where $\boldsymbol{\eta}^* = (\mu^{FP}, \boldsymbol{\beta}^*)$ and $\hat{\mu}^*$ is the IPSW.S or the KW.S estimator of μ. Using a first order Taylor expansion, we have

$$\widehat{\boldsymbol{\eta}}^* - \boldsymbol{\eta}^* = [E\{\phi^*(\boldsymbol{\eta}^*)\}]^{-1}\Phi^*(\boldsymbol{\eta}^*) + o_p\left(n_c^{-1/2}\right),$$

where $[E\{\phi^*(\boldsymbol{\eta}^*)\}]^{-1} = \left\{\frac{\partial\Phi^*(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}}\right\}^{-1}\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}^*} = [E\{\phi(\boldsymbol{\eta})\}]^{-1} = \begin{pmatrix} U_{\mu^{FP}}^{*-1} & \boldsymbol{b}^{*T} \\ \boldsymbol{0} & S_{\boldsymbol{\beta}^*}^{*-1} \end{pmatrix}$, with

$U_{\mu^{FP}}^* = E(\partial U^*/\partial\mu^{FP}) \doteq -1$ as $\widetilde{w}_i^* \doteq {\pi_i^{(c)}}^{-1}$,

$U_{\boldsymbol{\beta}^*}^* = E(\partial U^*/\partial\boldsymbol{\beta}^T)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \frac{1}{N}\sum_{i\in FP} \pi_i^{(c)} (y_i - \mu^{FP}) \frac{\partial \widetilde{w}_i^*}{\partial\boldsymbol{\beta}^T}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$,

$S_{\boldsymbol{\beta}^*}^* = E\{\partial S^*(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\}|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = -\frac{1}{N}\sum_{i\in FP} a p_i^* \boldsymbol{x}_i \boldsymbol{x}_i^T$ (negative definite and invertible under **C7**), and $\boldsymbol{b}^{*T} = -U_{\mu^{FP}}^{*-1} U_{\boldsymbol{\beta}^*}^* S_{\boldsymbol{\beta}^*}^{*-1}$.

Notice that based on **Lemma 4.1**, $U_{\mu^{FP}}^* = U_{\mu^{FP}}$ and $U_{\boldsymbol{\beta}^*}^* = U_{\boldsymbol{\beta}}$ for $U_{\mu^{FP}}^*, U_{\boldsymbol{\beta}^*}^*$ in the scaled estimating equations (4.10.7) and $U_{\mu^{FP}}, U_{\boldsymbol{\beta}}$ in the original estimating equations (4.10.2).

Accordingly,

$$Var(\widehat{\boldsymbol{\eta}}^*) = [E\{\phi^*(\boldsymbol{\eta}^*)\}]^{-1}Var\{\Phi^*(\boldsymbol{\eta}^*)\}[E\{\phi^*(\boldsymbol{\eta}^*)\}^T]^{-1} + o_p(n_c^{-1}).$$

The calculation of $Var\{\Phi^*(\boldsymbol{\eta}^*)\}$ is similar to calculating $Var\{\Phi(\boldsymbol{\eta})\}$ in Equality (4.10.3), with $w_i^{KW.W}$, p_i, and d_i replaced by $\widetilde{w}_i$, p_i^*, and $a \cdot d_i$ respectively. Finally, we have $Var(\hat{\mu}^*) = V^* + o_p(n_c^{-1})$, with

$$V^* = \begin{pmatrix} U_{\mu^{FP}}^{-1} & \boldsymbol{b}^{*T} \end{pmatrix} \cdot Var(\widehat{\boldsymbol{\eta}}^*) \cdot \begin{pmatrix} U_{\mu^{FP}}^{-1} \\ \boldsymbol{b}^* \end{pmatrix}$$

$$= \frac{1}{N^2}\sum\nolimits_{i\in FP} \pi_i^{(c)}\left(1-\pi_i^{(c)}\right)\left\{\widetilde{w}_i(y_i - \mu^{FP}) - (1-p_i^*)\boldsymbol{b}^{*T}\boldsymbol{x}_i\right\}^2 + \boldsymbol{b}^{*T}D^*\boldsymbol{b}^*,$$

where D^* replaces d_i and p_i in D by $a \cdot d_i$ and p_i^* respectively.

4.10.6 True Propensity Models in Simulations

The propensity of unit $i \in FP$ being included in the cohort (s_c) vs. the finite population (FP), based on Equality (4.4.1), is $P(i \in s_c | i \in s_c \cup^* FP) = \frac{\pi_i^{(c)}}{1+\pi_i^{(c)}}$. Hence, the true propensity model is

$$\text{logit}\{P(i \in s_c | i \in s_c \cup^* FP)\} = \log \pi_i^{(c)} = \log \frac{n_c \cdot r_i^{(c)}}{\sum_{i=1}^N r_i^{(c)}},$$

where $r_i^{(c)} = \exp(\boldsymbol{\alpha}^T\boldsymbol{x}_i)$ is the measure of size (MOS) of the Probability Proportional to size (PPS) Sampling for the cohort selection. Therefore,

$$\text{logit}\{p(i \in s_c | i \in s_c \cup^* FP)\} = \log \frac{n_c}{\sum_{i=1}^N r_i^{(c)}} + \log r_i^{(c)} = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}_i,$$

where $\beta_0 = \log \frac{n_c}{\sum_{i=1}^N r_i^{(c)}}$, and $\boldsymbol{\beta} = \boldsymbol{\alpha}$.

The propensity of unit $i \in FP$ being included in (s_c) vs. the *unweighted* survey sample (s_s) is $P(i \in s_c | i \in s_c \cup^* s_s) = \frac{\pi_i^{(c)}}{\pi_i^{(c)}+\pi_i^{(s)}}$. Hence, the true propensity score model is

$$\text{logit}\{p(i \in s_c | i \in s_c \cup^* s_s)\} = \log \frac{\pi_i^{(c)}}{\pi_i^{(s)}} = \log \left\{ \frac{n_c \cdot r_i^{(c)}}{\sum_{i=1}^{N} r_i^{(c)}} \bigg/ \frac{n_s \cdot r_i^{(s)}}{\sum_{i=1}^{N} r_i^{(s)}} \right\},$$

where $r_i^{(s)} = \exp(\boldsymbol{\gamma}^T \boldsymbol{x}_i)$ is the MOS of the PPS sampling for the survey sample selection. Therefore,

$$\text{logit}\{p(i \in s_c | i \in s_c \cup^* s_s)\} = \log \frac{n_c \sum_{i=1}^{N} r_i^{(s)}}{n_s \sum_{i=1}^{N} r_i^{(c)}} + \log \frac{r_i^{(c)}}{r_i^{(s)}} = \beta_0 + \boldsymbol{\beta}_1^T \boldsymbol{x}_i,$$

where $\beta_0 = \log \frac{n_c \sum_{i=1}^{N} r_i^{(s)}}{n_s \sum_{i=1}^{N} r_i^{(c)}}$, and $\boldsymbol{\beta}_1 = \boldsymbol{\alpha} - \boldsymbol{\gamma}$.

4.10.7 Jackknife Variance Estimation for Pseudo-Weighted Estimates in the Simulations

We treat the cohort and survey sample as 2 strata in the combined sample for Jackknife (JK) variance estimation since the two samples are independently selected. To reduce the number of replicates, we randomly grouped the cohort and survey sample into $g_1 = 120$ and $g_2 = 100$ groups in each simulation run which ensures sufficient degrees of freedom for the JK variance estimation (Fay, 1985). Formally, the JK variance estimation procedure for $\hat{\mu}^{KW.W}$follows as:

Step 1. Leave out α-th random group in stratum r, with $\alpha = 1, \cdots, g_r$, and $r = 1, 2$. Then weight up the units in remaining groups in stratum l by the ratio of the number of groups in l to the number of remaining groups, i.e., $\frac{g_r}{g_r - 1}$. This weight adjustment factor for unit i in replicate $r\alpha, r = 1, 2$ and $\alpha = 1, \cdots, g_r$, can be written as

$$f_{i(l\alpha)} = \begin{cases} 0, & \text{for unit } i \text{ in stratum } r \text{ group } \alpha; \\ \dfrac{g_r}{g_r - 1}, & \text{for unit } i \text{ in stratum } r \text{ group } \alpha' \neq \alpha; \\ 1, & \text{otherwise.} \end{cases}$$

Step 2. Refit Model 4.2 to the combined (s_c vs. weighted s_s) with weights of $f_{i(1\alpha)}$ for $i \in s_c$, and weights of $f_{i(2\alpha)} \cdot d_i$ for $i \in s_s$. Then re-estimate the propensity score for each unit in the replicate-$r\alpha$ sample.

Step 3. Create the KW.W pseudo-weight for cohort unit i in replicate-$r\alpha$ is

$$w_{i(r\alpha)}^{KW.W} = \sum\nolimits_{j \in s_{s(r\alpha)}} \left\{ \frac{K\left\{\left(q_i^{(c)} - q_j^{(s)}\right)/h\right\}}{\sum_{i \in s_{c(r\alpha)}} K\left\{\left(q_i^{(c)} - q_j^{(s)}\right)/h\right\}} \cdot d_j \cdot f_{j(r\alpha)} \right\}$$

where the bandwidth h is the same as obtained from the original combined sample (Korn & Graubard, 1999 page 89); $s_{s(r\alpha)}$ and $s_{c(r\alpha)}$ denote the cohort and survey sample in replicate-$r\alpha$, respectively.

Step 4. Re-estimate the population mean/prevalence estimate as

$$\hat{\mu}_{(r\alpha)}^{KW.W} = \left(\sum\nolimits_{i \in s_{c(r\alpha)}} w_{i(r\alpha)}^{KW.W}\right)^{-1} \cdot \sum\nolimits_{i \in s_{c(r\alpha)}} w_{i(r\alpha)}^{KW.W} \cdot y_i.$$

The JK variance estimate for KW.W estimate of population mean/prevalence, $\hat{\mu}^{KW.W}$, is

$$var(\hat{\mu}^{KW.W}) = \sum_{r=1}^{2} \frac{g_r - 1}{g_r} \sum_{\alpha=1}^{g_r} \left\{\hat{\mu}_{(r\alpha)}^{KW.W} - \hat{\mu}^{KW.W}\right\}^2.$$

JK estimators for the variance of the KW, IPSW, IPSW.S, and KW.S estimates are calculated similarly as described above, but differ at Steps 2 and 3. At Step 2, the KW method fits the propensity model with weights of $f_{i(r\alpha)}$ for each cohort and survey unit i, while the IPSW.S and KW.S methods fit the propensity model with the weights of $a \cdot$

$f_{i(2\alpha)} \cdot d_i$ for survey unit i, with a being the scaling factor. At Step 3, the IPSW and IPSW.S methods take the inverse of predicted odds as the pseudo-replicate weights.

Chapter 5 Improving External Validity of Association/Relative Risk Estimation from Nonprobability Cohorts

5.1 Introduction

In epidemiology, associations between risk factors and diseases are important to study for human diseases. There is fractious debate about the value of population-representative samples for external validity of association estimation. Some argues that lack of representative samples may not lead to large bias in association estimation if the confounders are appropriately controlled (Pizzi et al., 2011; Richiardi et al., 2013) while others advocate for the necessity of representative sample (Little, 2010; Keiding & Louis, 2016). Ignoring the representativeness of the sample may lead to biased estimates of associations for the target finite population if the sample selection is informative, that is, the probabilities of sample selection are correlated with the outcome variable conditional on the risk factors included in the analysis model (Fuller, 1999). When the probabilities of sample selection are noninformative, (i.e., the sample selection is uncorrelated with the outcome variable conditional on the risk factors), the naïve estimates of associations, assuming simple random sampling, can also be biased for the target finite population quantities if the analic model is misspecified (Korn & Graubard, 1999). Moreover, significance tests for associations in the naïve sample can be different from the target finite population. For example, Kennedy et al. (2016) found that the significant marginal effects associated with race/ethnicity shown in a benchmark survey sample were rarely captured by nine non-probability samples. However, there is limited literature investigating the performance of PS-based methods in reducing bias of the nonprobability sample estimates of associations. This chapter focuses on how unrepresentativeness of the cohort influences estimates of associations between risk factors and certain diseases for the target finite population, and whether or not the PS-based methods can obtain less biased and efficient estimates from the cohort.

The remainder of this chapter is arranged as follows: Section 5.2 introduces the basic setup for regression analyses, including the justification of biasedness of naïve cohort estimates of regression coefficients under the informative and noninformative cohort participation mechanism. PS-based estimates of regression coefficients and variance estimation methods are also be discussed in this section. Section 5.3 presents simulation studies that evaluate performance of the proposed PS-based methods under two situations where the cohort participation is (1) informative; and (2) noninformative. The robustness of the PS-based methods to propensity model misspecification is examined in both scenarios. Section 5.4. presents estimates of relative risk of 15-year mortality for obese adults compared to nonobese adults in the U.S. from the naïve NHANES III sample by using 1994 NHIS sample as a reference sample.

5.2 Method

5.2.1 Setup for regression analyses

We consider the regression of a variable y on a vector of covariates $\boldsymbol{x}$. Let the target finite population (FP) consist of N individuals indexed by $i \in \{1, \cdots, N\}$, where each individual i has values for the outcome variable of interest y_i and for the vector of covariates $\boldsymbol{x}_i$. Suppose that the outcome variable and the covariates follow the regression model

$$E_m(y_i \mid \boldsymbol{x}_i) = g^{-1}(\theta_0 + \boldsymbol{\theta}_1^T \boldsymbol{x}_i), i \in FP \tag{5.2.1}$$

where the expectation E_m is respective to the distribution of the outcome y given the covariates $\boldsymbol{x}$ in the superpopulation (defined in Section 3.6.1), $g(\cdot)$ is a known link function, and $\boldsymbol{\theta} = (\theta_0, \boldsymbol{\theta}_1^T)^T$ is an unknown vector of coefficients We are interested in estimating $\boldsymbol{\theta}_1$, the parameters of association. Following notation in Chapter 3 and Chapter 4, we let $s_c \subset FP$ to denote a cohort with n_c individuals. The cohort participation indicator, implicit cohort participation rate, and the corresponding implicit cohort sample weight for $i \in FP$ are defined by $\delta_i^{(c)}$(= 1 if $i \in s_s$; 0 otherwise), $\pi_i^{(c)} = E_c\left(\delta_i^{(c)} \mid FP \right)$, and $w_i = 1/\pi_i^{(c)}$, respectively, where the expectation E_c is with respect to the unknown random cohort participation process from FP.

Under the assumption **A1** in Section 4.2, i.e., $\Pr\left(\delta^{(c)} = 1 \mid y, \boldsymbol{x} \right) = \Pr\left(\delta^{(c)} = 1 \mid \boldsymbol{x} \right)$, the cohort participation is non-informative for the regression model (5.2.1) . Under A1, the regression parameters $\boldsymbol{\theta}$ estimated from the naïve cohort are unbiased because

$$E_m\left(y_i \mid \boldsymbol{x}_i, \delta_i^{(c)} = 1\right) = E_m(y_i \mid \boldsymbol{x}_i).$$

However, if the outcome model is misspecified, the regression parameters estimated from the naïve cohort may not be approximately unbiased for the finite population quantities. Suppose we misspecify the outcome model as follows

$$E_m(y_i \mid \boldsymbol{x}_i^*) = g^{-1}\left(\theta_0^* + \boldsymbol{\theta}_1^{*T} \boldsymbol{x}_i^*\right), i \in FP \tag{5.2.2}$$

where $\boldsymbol{x}^*$ is a set of predictors, and $\boldsymbol{\theta}^* = (\theta_0^*, \boldsymbol{\theta}_1^{*T})^T$ is a vector of parameters to be estimated. Note that $\boldsymbol{\theta}^*$ can be different from $\boldsymbol{\theta}$ if $\boldsymbol{x}^*$ and $\boldsymbol{x}$ are not identical. Suppose the predictors $\boldsymbol{x}^*$ do not include all covariates that are correlated with both y and $\delta^{(c)}$ (i.e., confounders) in $\boldsymbol{x}$. Let $\boldsymbol{x}_m$ to be a vector of the missing confounders. Then we have $\Pr\left(\delta^{(c)} = 1 \mid y, \boldsymbol{x}^*, \boldsymbol{x}_m\right) = \Pr\left(\delta^{(c)} = 1 \mid \boldsymbol{x}^*, \boldsymbol{x}_m\right)$, but $\Pr\left(\delta^{(c)} = 1 \mid y, \boldsymbol{x}^*\right) \neq \Pr\left(\delta^{(c)} = 1 \mid \boldsymbol{x}^*\right)$, and therefore

$$E_m\left(y_i \mid \boldsymbol{x}_i^*, \delta_i^{(c)} = 1\right) \neq E_m\left(y_i \mid \boldsymbol{x}_i^*\right).$$

Hence, the naïve cohort estimates of $\boldsymbol{\theta}^*$ are biased.

If the cohort participation is informative, i.e., $\Pr\left(\delta^{(c)} = 1 \mid y, \boldsymbol{x}\right) \neq \Pr\left(\delta^{(c)} = 1 \mid \boldsymbol{x}\right)$, the regression parameters $\boldsymbol{\theta}$ estimated from the naïve cohort are biased because

$$E_m\left(y_i \mid \boldsymbol{x}_i, \delta_i^{(c)} = 1\right) \neq E_m\left(y_i \mid \boldsymbol{x}_i\right).$$

PS-based methods can be applied to improve the representativeness of the cohort so that the regression coefficients $\boldsymbol{\theta}$ estimated from the pseudo-weighted cohort are less biased. PS-based methods are first applied to create a set of pseudo weights for the cohort by using a survey sample $s_s \subset FP$ as the reference sample. Then the outcome model (5.2.1) or (5.2.2) is fitted to the pseudo-weighted cohort. We consider five PS-based pseudo-weighted estimators of $\boldsymbol{\theta}$: the original IPSW ($\widehat{\boldsymbol{\theta}}^{IPSW}$) and KW ($\widehat{\boldsymbol{\theta}}^{KW}$) estimates introduced in Chapter 3, the enhanced KW estimate ($\widehat{\boldsymbol{\theta}}^{KW.W}$), the IPSW and KW estimates with scaled survey weights in PS estimation, proposed in Chapter 4, refer to as $\widehat{\boldsymbol{\theta}}^{IPSW.S}$ and $\widehat{\boldsymbol{\theta}}^{KW.S}$, respectively.

5.2.2 Variance Estimation

Two variance estimators for the pseudo-weighted estimates of $\boldsymbol{\theta}$ are considered: the naïve Taylor linearization (TL) and Jackknife replication (JK) method. Similar to Section 3.2.3, the naïve TL method treats the pseudo weights as fixed, and ignore the randomness due to estimating PS. The JK method, on the contrary, takes all sources of variability into account by recalculating pseudo-weights for each replicate.

5.3 Simulations

We examined the performance of the five PS-based estimators of the log-odds ratios, $\boldsymbol{\theta}$ (i.e., where $g(\cdot)$ is the logit function), of developing a disease compared to the naïve cohort estimates ($\widehat{\boldsymbol{\theta}}^{Naive}$) and the sample weighted survey estimate ($\widehat{\boldsymbol{\theta}}^{SVY}$) in two scenarios where the cohort participation was : (1) informative; and (2) non-informative.

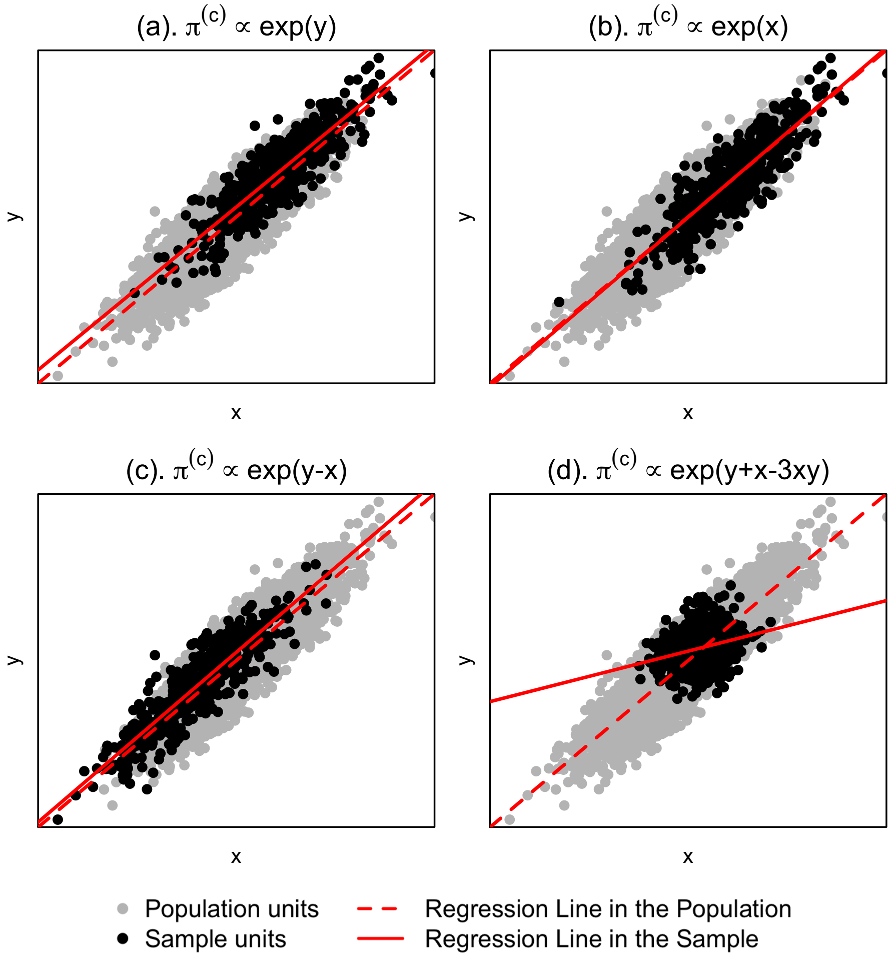
Under Scenario (1), $\widehat{\boldsymbol{\theta}}^{Naive}$ may not be approximately unbiased while the pseudo-weighted estimators would be approximately unbiased if the propensity model is correctly specified. Under Scenario (2), all the estimates would be unbiased under the correct outcome model. Under the misspecified outcome model, however, $\widehat{\boldsymbol{\theta}}^{Naive}$ may not be approximately unbiased but the pseudo-weighted estimators would be approximately unbiased if the propensity model was correct. The robustness of the pseudo-weighted estimators to the propensity model misspecification was also examined under the two scenarios.

In both scenarios, we used a finite population of $M = 3{,}000$ clusters with each cluster composed of $3{,}000$ units (population total $N = 9{,}000{,}000$) generated in Chapter 3 and applied similar two-stage cluster PPS sample designs for both cohort participation and survey sample selection to ensure the true propensity models for all five PS-based methods have the same functional form so that a fair comparison can be made among these methods.

Following the framework established in Section 4.4, we considered the cohort participation rate $\pi^{(c)} \propto \exp(\boldsymbol{\beta}^T \boldsymbol{v})$, where $\boldsymbol{v}$ can include the outcome y, and the covariates

$\boldsymbol{x}$ predictive to y, and their interactions. Figure 5.1 shows a simple example of a linear regression model of continuous y on x as the analytic model. If $\boldsymbol{v}$ only includes the main effect(s) of x, or (and) y, the cohort participation is noninformative and the regression line between x and y is approximately the same in the sample and in the population (situations a, b, and c). The cohort participation is informative only if $\boldsymbol{v}$ includes the interaction of x and y (situation d).

Figure 5.1 Noninformative and informative cohort participation rates.



5.3.1 Scenario 1: informative cohort selection

5.3.1.1 Disease Outcome Model in the Finite Population

A binary variable for disease status, y (1 for presence, and 0 for absence) was generated to have an ICC within the clusters of 0.07 for the finite population, with the probability of having disease generated by $\mu = \text{expit}(-6 + 0.5age + 1.5Env)$ (Hunsberger et al., 2008; Oman & Zucker, 2001). The prevalence in the population was 14.6%. The outcome model of y in the finite population was

$$\log\left\{\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right\} = \theta_0 + \theta_1 age_i + \theta_2 Env_i, i \in FP \qquad (5.3.1)$$

where θ_0 and $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are the unknown parameters to be estimated, with the true values $\theta_0 = -6$, and $\boldsymbol{\theta} = (0.5, 1.5)$. We are interested in estimating the relative risk of y associated with the environmental factor Env after controlling for the age group (i.e., e^{θ_2}). A proxy of y was generated by $y^* = y + \epsilon$, with $\epsilon \sim \text{Normal}(0, 0.05^2)$ in the finite population to reflect situations in real data when y is not available for sample selection but related variables are available.

5.3.1.2 Sampling from the Finite Population to Assemble Survey Sample and Cohort

Two-stage cluster sample designs similar with those in Section 3.3.2 were applied to randomly select the cohort and survey sampled independently as to ensure that true propensity models for all PS-based methods (IPSW, KW, KW.W, IPSW.S, and KW.S) had the same functional form.

A cohort sample of size $n_c = 11{,}250$ people (75 clusters of each 150 individuals) was sampled by a two-stage PPS design with the MOS in the PPS sampling at stages one and two being $\sum_{i \in u_\alpha} r_i^a$ and r_i^a, respectively, where u_α is the set of individuals from the α-th cluster for $\alpha = 1, \cdots, M$; $a = 1$ is a constant, and

$$r_i = \exp(\gamma_0 + \gamma_1 Env_i + \gamma_2 \, y_i^* \cdot Env_i), \qquad (5.3.2)$$

where $\gamma_0 = 0.7, \boldsymbol{\gamma} = (\gamma_1, \gamma_2)^T = (0.5, -1)^T$. The final cohort participation rate for $i \in FP$ was $\pi_i^{(c)} = \frac{n_c \cdot r_i}{\sum_{i \in FP} r_i}$. The cohort participation rate included the interaction between y and the predictor Env (situation (d) of Figure 5.1). This was an informative design for regression model (5.3.1), because the correlation between the cohort participation rate $\pi^{(c)}$ and the covariate Env was different depending on the value of the outcome y, as shown below.

$$\pi_i^{(c)} = E\left(\delta_i^{(c)} \mid y_i, \boldsymbol{x}_i \right) \propto r_i(y_i, \boldsymbol{x}_i) \propto \begin{cases} \exp\{(0.5 - \epsilon) Env_i\}, & \text{if } y_i = 0, \\ \exp\{(-0.5 - \epsilon) Env_i\}, & \text{if } y_i = 1. \end{cases}$$

$$\neq E\left(\delta_i^{(c)} \mid \boldsymbol{x}_i \right)$$

Hence, the naïve cohort estimator of θ_2 would not be approximately unbiased. On the contrary, the naïve cohort estimator of θ_1 would be approximately unbiased because age was predictive to y, but not correlated with the cohort inclusion indicator $\delta_i^{(c)}$ given y.

A survey sample of $n_s = 1{,}500$ individuals (150 clusters of each 10 individuals) was sampled independently from the cohort selection using a similar two-stage PPS design but with different MOSs in the PPS sampling at stages one and two, given as $\sum_{i \in u_\alpha} r_i^b$ and r_i^b, respectively, with $b = -0.5$.

Under the two-stage PPS sampling described above, the true propensity models for $\tilde{p}_i = P(\, i \in s_c \mid i \in s_c \cup^* s_s \,)$ used by the KW method and for $p_i = P(\, i \in s_c \mid i \in s_c \cup^* FP \,)$ used by the IPSW, KW.W, IPSW.S, and KW.S methods were

$$\begin{aligned}\text{logit}\{\tilde{p}_i\} &= \tilde{\beta}_0 + 1.5\gamma_1 Env_i + 1.5\gamma_2 y_i^* \cdot Env_i, \text{and} \\ \text{logit}\{p_i\} &= \beta_0 + \gamma_1 Env_i + \gamma_2 y_i^* \cdot Env_i\end{aligned} \tag{5.3.3}$$

where $\beta_0 = \log\left(\frac{n_c}{\sum_{i \in FP} r_i}\right) + \gamma_0$ and $\tilde{\beta}_0 = \log\left(\frac{n_c \cdot \sum_{i \in FP} r_i^{-0.5}}{n_s \cdot \sum_{i \in FP} r_i}\right) + 1.5\gamma_0$ are the intercepts (proof in Section 3.6.4).

5.3.1.3 *Results under the Correctly Specified and Three Misspecified Propensity Models*

Table 5.1 shows the results under the correct propensity model. As expected, all the estimators, including the naïve cohort estimator, of θ_1 were approximately unbiased, while the naïve cohort estimator of θ_2 was biased by -64.96%. All the examined PS-based methods yielded approximately unbiased estimators of θ_2 when the propensity model was correctly specified. Consistent with the results observed for estimating finite population means/prevalences, the IPSW estimate had a much higher empirical variance than the other PS-based estimates, leading to the largest MSE. The KW estimate had the smallest variance and MSE, taking the advantage of fitting the propensity model to the combined (cohort and unweighted survey) sample. The other three (KW.W, IPSW.S, and KW.S) estimators performed equally well, with slightly higher variances and MSEs than the KW estimate.

The TL method, which ignores the variability due to estimating the PS substantially underestimated the variance of the pseudo-weighted estimates of θ_2. The underestimation was more severe for the IPSW and the IPSW.S methods, because the estimated PS were directly used to estimate the cohort participation rates. In contrast, the JK method that considered all sources of variability provided approximately unbiased variance estimation for all five PS-based methods. However, the outcome variable y (or the proxy y^*) is usually only available in the cohort, but not in the survey sample. Therefore, PS-based

Table 5.1 Results of regression coefficient estimation from 2,000 simulated cohorts and survey samples selected by informative designs under the correct propensity model

	%RB		$\mathbf{V}(\times 10^3)$		**VR** (TL)		**VR** (JK)		**MSE** $(\times 10^3)$	
Method	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$
CHT	1.46	-64.95	2.30	11.31	0.57	0.53			2.35	960.54
SVY	1.62	0.74	2.69	19.45	1.01	0.99	1.03	1.03	2.76	19.57
IPSW	1.49	1.40	2.37	24.03	0.93	0.49	0.99	1.01	2.43	24.47
KW	1.48	0.18	2.52	16.89	0.93	0.79	1.09	1.06	2.58	16.90
KW.W	1.56	0.00	2.45	17.34	0.92	0.73	1.02	1.01	2.51	17.34
IPSW.S	1.49	0.42	2.36	17.13	0.93	0.69	0.99	1.00	2.42	17.17
KW.S	1.56	-0.03	2.45	17.26	0.93	0.74	1.02	1.02	2.51	17.26

methods cannot provide unbiased estimators of the regression coefficients if the cohort participation is informative. We examined the performance of the PS-based methods under three misspecified propensity models that missed y^*. *Model* U was an underfitted model that only included the main effect of Env. *Model* M_1 included an extra covariate age, which was the predictor of the outcome y, and the interaction of age and Env. *Model* M_2 substituted y^* in the true propensity model using the probability of developing the disease $\mu = \text{expit}(-6 + 0.5age + 1.5Env)$. All the three misspecified propensity models only included (a fixed function of) predictors of the outcome y. Hence, none of the PS-based methods reduced bias from the naïve cohort estimator of θ_2 (RB $= -64.95\%$) under these misspecified propensity models as shown in Table 5.2.

Table 5.2 Relative bias (%) of $\hat{\theta}_2$ from 2,000 simulated cohorts and survey samples selected by informative designs under misspecified propenstiy models

	PS-Based Method				
Propensity Model	IPSW	KW	KW.W	IPSW.S	KW.S
Model U $\text{logit}(p) \sim Env$	-65.07	-64.86	-64.86	-65.10	-64.86
Model M_1: $\text{logit}(p) \sim age, Env, age \cdot Env$	-65.00	-64.92	-65.04	-65.01	-65.02
Model M_2: $\text{logit}(p) \sim Env, \mu \cdot Env$	-64.93	-65.02	-64.87	-64.91	-64.84

5.3.2 Scenario 2: Non-Informative Cohort Selection

_5.3.2.1_ *Disease Outcome Model in the Finite Population*

A binary variable of disease status, y (1 for presence, and 0 for absence) was generated to have an ICC within the clusters of 0.08 for the finite population, with the probability of disease generated by $\mu = \text{expit}(1 + 0.5age + 0.5Env + 0.5age \cdot Env)$ (Hunsberger et al., 2008; Oman & Zucker, 2001). The prevalence in the population was 31.9%. The outcome model of y in the finite population was

$$\log\left\{\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right\} = \theta_0 + \theta_1 age_i + \theta_2 Env_i + \theta_3 age_i \cdot Env_i\,, i \in FP \quad (5.3.4)$$

where θ_0 and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ are the unknown vector of parameters to be estimated, with the true values $\theta_0 = 1$, and $\boldsymbol{\theta} = (0.5, 0.5, 0.5)$. A proxy of y was generated by $z = \mu + \epsilon$, with $\epsilon \sim N(0, 0.05^2)$ in the finite population to reflect situations occurs in real data when μ is not available for sample selection but related variables are available.

_5.3.2.2_ *Sampling from the Finite Population to Assemble Survey Sample and Cohort*

The cohort and survey sample were independently selected using two-stage PPS designs similar to those described in Section 5.3.1.2, but with different r_i for calculation of MOSs.

$$r_i = \exp(\gamma_0 + \gamma_1 age_i + \gamma_2 z_i), \quad (5.3.5)$$

where $\gamma_0 = 1, \boldsymbol{\gamma} = (\gamma_1, \gamma_2)^T = (0.45, -1)^T$. The cohort participation was noninformative for the regression model (5.3.4), because it only depended on the predictor age and a fixed function of the predictors (z) (situation (b) in Figure 5.1). The naïve cohort estimate of the regression coefficients $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ would be unapproximately unbiased.

The true propensity models for $\tilde{p}_i = P(\, i \in s_c \mid i \in s_c \cup^* s_s\,)$ used by the KW method and for $p_i = P(\, i \in s_c \mid i \in s_c \cup^* FP\,)$ used by the IPSW, KW.W, IPSW.S, and KW.S methods were

$$\begin{aligned} \text{logit}\{\tilde{p}_i\} &= \tilde{\beta}_0 + 1.5\gamma_1 age_i + 1.5\gamma_2 z_i, \text{and} \\ \text{logit}\{p_i\} &= \beta_0 + \gamma_1 age_i + \gamma_2 z_i \end{aligned} \tag{5.3.6}$$

where $\beta_0 = \log\left(\frac{n_c}{\sum_{i \in FP} r_i}\right) + \gamma_0$ and $\tilde{\beta}_0 = \log\left(\frac{n_c \cdot \sum_{i \in FP} r_i^{-0.5}}{n_s \cdot \sum_{i \in FP} r_i}\right) + 1.5\gamma_0$ are the intercepts (proof in Section 3.6.4).

<u>5.3.2.3</u> *Results under the Correctly Specified Outcome model*

Table 5.3 shows the results under the correct outcome and propensity model. As expected, all the estimators, including the naïve cohort estimators, of $\boldsymbol{\theta}$ were approximately unbiased. The naïve cohort estimates had much smaller variances and MSEs than any of the PS-based pseudo-weighted estimates due to ignoring the weights. The five PS-based methods had similar performances. The TL method slightly underestimated variance compared to the JK method.

Table 5.3 Results of regression coefficient estimation of the correctly specified outcome model from 2,000 simulated cohorts and survey samples under non-informative designs

	%RB			**V**$(\times 10^2)$			**VR** (TL)			**VR** (JK)			**MSE** $(\times 10^2)$		
Method	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
CHT	0.57	-1.56	0.25	0.55	5.00	0.28	0.70	0.84	0.71	0.00	0.00	0.00	0.55	5.01	0.28
SVY	1.32	-0.40	1.17	3.52	15.82	1.46	0.94	0.91	0.94	0.99	0.97	1.00	3.53	15.82	1.47
IPSW	1.60	-1.19	0.83	0.89	7.17	0.43	0.94	0.92	0.92	1.04	1.03	1.02	0.90	7.17	0.43
KW	1.54	-1.36	0.86	0.88	7.11	0.42	0.94	0.93	0.92	1.06	1.08	1.06	0.88	7.11	0.42
KW.W	1.61	-0.26	0.39	0.89	6.89	0.42	0.93	0.94	0.92	1.02	1.04	1.02	0.89	6.89	0.42
IPSW.S	1.66	-0.16	0.38	0.90	7.02	0.42	0.94	0.94	0.93	1.03	1.05	1.03	0.90	7.02	0.42
KW.S	1.61	-0.26	0.40	0.89	6.91	0.42	0.93	0.94	0.92	1.02	1.03	1.01	0.89	6.91	0.42

5.3.2.4 *Results under a Misspecified Outcome model*

Suppose the outcome model mistakenly left out the covariate age and the interaction between age and Env. The misspecified model in the finite population was

$$\log\left\{\frac{P(y_i=1)}{1-P(y_i=1)}\right\} = \theta_0^* + \theta_2^* Env_i\,, i \in FP \qquad (5.3.7)$$

with θ_0^* and θ_2^* being the unknown census regression coefficients. Although $\theta_0^* = 2.62$ and $\theta_2^* = 1.30$ were different the parameters $\theta_0 = 1$ and $\theta_2 = 0.5$ in the true outcome model (5.3.4), they were finite population quantities. We were interested in estimating θ_2^* from the cohort. Table 5.4 shows the results under the correctly specified propensity model.

Table 5.4 Results of regression coefficient estimation of the misspecified outcome model from 2,000 simulated cohorts and survey samples under non-informative designs.

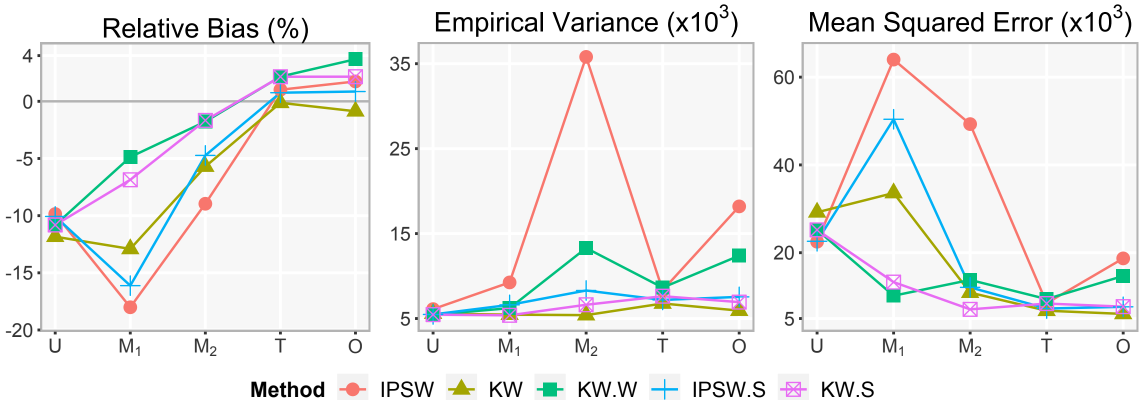
Estimator	%RB	**V**$(\times 10^3)$	**VR** (TL)	**VR** (JK)	**MSE** $(\times 10^3)$
$\hat{\theta}_2^{*\,Naive}$	35.01	4.73	0.48	NA	211.36
$\hat{\theta}_2^{*\,SVY}$	0.60	13.80	1.05	1.06	13.86
$\hat{\theta}_2^{*\,IPSW}$	1.02	8.33	0.70	1.00	8.50
$\hat{\theta}_2^{*\,KW}$	-0.13	6.77	0.83	1.05	6.78
$\hat{\theta}_2^{*\,KW.W}$	2.16	8.64	0.63	1.04	9.43
$\hat{\theta}_2^{*\,IPSW.S}$	0.75	7.19	0.81	1.00	7.28
$\hat{\theta}_2^{*\,KW.S}$	2.15	7.63	0.72	1.03	8.41

The naïve cohort estimator, $\hat{\theta}_2^{*\,Naive}$ was biased by 35.01%. The five PS-based methods had similar performance in terms of relative bias, empirical variance, and MSE. The original KW estimate had slightly smaller variance and MSE than the other methods. The TL variance estimation, consistent with the previous results, underestimated the variance of $\hat{\theta}_2^*$ due to ignoring the variability in estimating PSs, while the JK estimates were approximately unbiased.

When the fitted propensity model was misspecified, the five PS-based method performed differently in estimating θ_2^*. As shown in Figure 5.2, the KW.S estimator,

$\hat{\theta}_2^{*\,KW.S}$, tended to have the smallest MSE among all the five PS-based pseudo-weighted estimators. Although the IPSW ($\hat{\theta}_2^{*\,IPSW}$) and IPSW.S ($\hat{\theta}_2^{*\,IPSW.S}$) estimators had smaller biases than $\hat{\theta}_2^{*\,KW.S}$ under the correctly specified propensity model, they were more sensitive to propensity model misspecification. The original KW estimator, $\hat{\theta}_2^{*\,KW}$, consistently had the smallest variance, but can have large bias under the misspecified propensity models. The bias reduction of the KW.W estimator, $\hat{\theta}_2^{*\,KW.W}$ was also relatively robust to the propensity misspecification. However, it can have inflated variance due to the highly variable weights in the combined (cohort and *weighted* survey) sample.

Figure 5.2 Results of $\hat{\theta}_2^*$ obtained from 2,000 simulated cohorts and survey samples under non-informative designs with each cohort and survey sample fitted to the correct propensity model and four misspecified propensity models†



†The labels of the x-axises represent the propensity models as follows:
Model U: underfitted model logit(p) ~ age
Model M_1: misspecified model logit(p) ~ age, Env
Model M_2: misspecified model logit(p) ~ age, Env, $age \cdot Env$
Model T: true model logit(p) ~ age, z
Model O: overfitted model. logit(p) ~ age, z, Env, $race/ethnicity$

Model U was an underfitted model that did not include z, which was highly correlated with the disease status y. All the PS-based pseudo-weighted estimates had large bias (relative bias ~10%) and similar variances.

Models M_1 also excluded z but added an extra variable of Env, which was a predictor of y. Comparing results under Models M_1 and U, we observe that adding extra covariates that were correlated with the outcome variable in the propensity model did not reduce, but increased biases of $\hat{\theta}_2^{*IPSW}$ and $\hat{\theta}_2^{*IPSW.S}$. The potential reason is that the cohort participation rates were poorly estimated under Model M_1. This result is different from the findings in estimating finite population means where adding predictors of y to the misspecified propensity model usually help reduce the bias (Figure 5.4 for finite population prevalence estimation in Scenario 2, and Figure 3.1 for simulations in Chapter 3). Hence, bias reduction of the IPSW, and the IPSW.S methods were more sensitive to the propensity model misspecification in estimating regression finite population coefficients as compared to estimating the finite population means/prevalences. Moreover, the variance of $\hat{\theta}_2^{*IPSW}$ increased under Model M_1 compared to the results under Model U due to including an extra covariate in the propensity model. On the contrary, $\hat{\theta}_2^{*KW.W}$ and $\hat{\theta}_2^{*KW.S}$ showed greater bias reduction under Model M_1, without inflating the variances.

Model M_2 added the interaction term of age and Env besides the main effect of Env and age in Model M_1. Adding this interaction term reduced, but did not eliminate, the bias for all the pseudo-weighted estimates, among which the $\hat{\theta}_2^{*KW.W}$ and $\hat{\theta}_2^{*KW.S}$ estimators had the smallest bias. However, the variances increased especially for $\hat{\theta}_2^{*IPSW}$ and $\hat{\theta}_2^{*KW.W}$ because of highly variable weights in the combined (cohort and *weighted* survey) sample.

Model O was incorrectly overfitted, including unnecessary variables of Env and race/ethnicity compared to the true propensity model. The bias reduction was similar for all five estimators compared to the bias reduction under the true model. However, adding extra variables resulted in higher variance of $\widehat{\theta}_2^{*IPSW}$ and $\widehat{\theta}_2^{*KW.W}$.

For the variance estimation (Table 5.6), the naïve TL method underestimated the variance, especially for $\widehat{\theta}_2^{*IPSW}$ and $\widehat{\theta}_2^{*KW.W}$, with the variance ratio as low as 0.17 and 0.37 under Model M_2, respectively. The JK estimates were approximately unbiased for variances of $\widehat{\theta}_2^{*IPSW}$ and $\widehat{\theta}_2^{*IPSW.S}$, but can overestimate the variances of $\widehat{\theta}_2^{*KW}$ $\widehat{\theta}_2^{*KW.W}$ and $\widehat{\theta}_2^{*KW.S}$, which is consistent with the results in Chapter 4

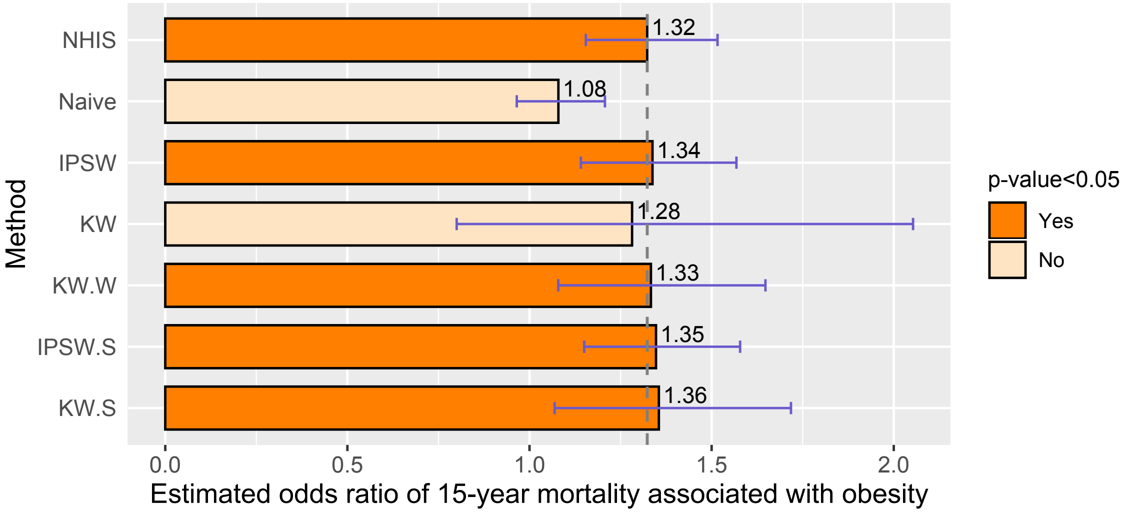
For comparison purpose, the results of finite population prevalence estimation were reported in Figure 5.4 and Table 5.6. The pattern of the results are consistent with those in Chapter 3 and Chapter 4.

5.4 Data Analysis: The U.S. National Health and Nutrition Examination Survey

This data example used the same nonprobability cohort and the reference survey sample as Chapter 4. We estimated the odds ratio of 15-year all-cause mortality associated with obesity (BMI ≥ 30) for adults in the US using the unweighted adult sample of household interview part of NHANES III conducted in 1988-1994 as the cohort. The reference survey sample was from the 1994 NHIS respondents to the supplement for monitoring achievement of the Healthy People Year 2000 objectives, aged 18 and older.

In the outcome model, we controlled the confounders of age (continuous in years), sex (male, and female), race/ethnicity (Non-Hispanic White, Non-Hispanic Black, Hispanic, and Non-Hispanic others), education level (continuous trend), and smoking status (non-smokers, former smokers, and current smokers). Figure 5.3 shows the odds-ratios of 15-year mortality associated with obesity (with 95% confidence interval) estimated by seven methods: NHIS estimates with the complex sample designs considered, the naïve NHANES estimates ignoring all the complex sample designs, and five PS-based pseudo-weighted estimators (IPSW, KW, KW.W, IPSW.S, and KW.S) obtained from the NHANES sample.

Figure 5.3 Odds ratio of 15-year mortality associated with obesity estimated by seven methods



With the complex sample designs of 1994 NHIS considered, the odds of 15-year mortality was 1.32 times (Figure 5.3) higher for obese adults than nonobese adults in the U.S. (p-value 5.6×10^{-5} in Table 5.5). In contrast, there was no significant difference in the odds of 15-year mortality for obese adults than nonobese adults in the naïve NHANES sample (estimated odds ratio = 1.08, with p-value= 0.18 in Table 5.5). The 95% CI of the naïve NHANES estimate was not overlapped with the that of the weighted NHIS estimate. All PS-based pseudo-weighted estimates of odds ratios were close to the NHIS estimate. The NHIS estimate was covered by 95% CIs of the pseudo-weighted estimates. However, the KW method, although it removed most bias of the naïve NHANES estimate, failed to capture the significantly higher risk of 15-year mortality for the obese people compared to the nonobese people (p-value= 0.15) due to a large JK variance (Table 5.5). The large JK variance estimates can be caused by the propensity model misspecification or the overestimation of the JK method as observed in the simulations.

Table 5.5 Estimates of log-odds ratios for all-cause 15-year mortality with Jackknife standard error estimates and p-values

	Age	**Sex**	**Race/Ethnicity**			**Educ**	**Smoking Status**		**Obese**
		Female	NH Black	Hispanic	NH-Other		Former Smoker	Current Smoker	Yes
NHIS Estimate	0.11	-0.55	0.52	-0.13	-0.26	-0.17	0.91	0.21	0.28
%Relative Bias (%Bias reduction)									
Naïve	-0.7	-15.8	-43.7	68.6	176.8	-4.0	-5.2	9.9	-72.8
IPSW	5.3	-30.2	-19.8	48.7	138.5	-8.3	11.2	15.2	3.9
	(911)	(-91)	(55)	(29)	(22)	(-110)	(314)	(-56)	(105)
KW	7.5	-26.4	-26.8	30.7	146.3	-7.6	11.0	10.9	-11.4
	(1254)	(-67)	(39)	(55)	(17)	(-91)	(310)	(-11)	(84)
KW.W	5.4	-29.2	-20.2	42.2	132.4	-8.6	11.3	16.3	2.7
	(923)	(-85)	(54)	(38)	(25)	(-115)	(317)	(-66)	(104)
IPSW.S	5.0	-29.5	-20.1	58.0	130.1	-6.3	11.7	16.4	6.4
	(880)	(-87)	(54)	(15)	(26)	(-57)	(326)	(-67)	(109)
KW.S	5.5	-28.8	-20.5	52.0	128.3	-6.7	12.5	18.6	8.5
	(864)	(-87)	(54)	(15)	(26)	(-56)	(325)	(-66)	(112)
Standard Error × 10^2 (JK)									
NHIS	0.2	5.9	9.4	11.1	22.1	2.2	7.9	6.7	7.0
Naïve	0.2	4.8	5.8	6.2	22.6	1.7	6.1	5.5	5.6
IPSW	0.3	6.4	7.8	8.8	26.0	2.1	10.1	8.4	8.1
KW	1.4	8.8	18.0	30.5	44.5	2.2	11.5	13.7	24.1
KW.W	0.5	8.3	13.8	21.0	28.2	2.5	10.4	8.4	10.8
IPSW.S	0.3	6.4	7.5	8.2	25.3	2.1	9.8	8.3	8.1
KW.S	0.5	7.5	15.9	15.3	26.7	2.3	10.8	8.7	12.1
P-value									
NHIS	0	0	3.1E-08	2.3E-01	2.4E-01	4.7E-14	0	2.1E-03	5.6E-05
Naïve	0	0	4.9E-07	3.2E-04	1.6E-03	0.0E+00	0	2.2E-05	1.8E-01
IPSW	0	1.4E-09	7.0E-08	2.6E-02	1.6E-02	1.6E-13	0	4.4E-03	3.4E-04
KW	0	3.1E-06	3.5E-02	5.7E-01	1.5E-01	5.7E-12	0	9.4E-02	3.0E-01
KW.W	0	2.5E-06	2.6E-03	3.7E-01	3.1E-02	8.5E-10	0	4.1E-03	7.7E-03
IPSW.S	0	8.2E-10	3.2E-08	1.1E-02	1.8E-02	5.3E-14	0	3.8E-03	2.2E-04
KW.S	0	1.6E-07	9.0E-03	1.9E-01	2.5E-02	5.3E-12	0	4.6E-03	1.2E-02

5.5 Summary

In this chapter, we investigated the influence of unrepresentative sample on the naïve cohort estimators of the odds ratios from logistic regression analyses. When the cohort participation is correlated with the outcome variable conditional on the covariates in the regression model (informative participation), the naïve estimates of the regression coefficients would not be approximately unbiased. When the cohort self-selection is independent from the outcome variable given the covariates (noninformative participation), the naïve estimators would be approximately unbiased under the correct outcome model, but would not be approximately unbiased under misspecified outcome models. Since the outcome model is usually unknown in practice, PS-based methods should be applied for relative risk estimation from the nonprobability cohort. Moreover, as shown in the real data example, the results of significance tests obtained from the nonprobability cohorts can be invalid.

As shown in the simulations, the PS-based methods corrected the bias of the native cohort estimators using a correctly specified propensity model under either an informative or a noninformative selection. The original KW estimator had the smallest variances and MSE on average, followed by the KW.S and the IPSW.S estimators. However, the five methods performed differently when the propensity model was misspecified. The KW.S and the KW.W methods were more robust to the model misspecification than the IPSW and IPSW.S methods in terms of the bias reduction. This is because these two methods use the estimated PS to measure the similarity between the cohort and survey sample units, instead of predicting participation rates. Moreover, adding predictors of the outcome variable in an underfitted propensity model can increase bias of the IPSW and IPSW.S estimators, but decrease bias of the KW.W and the KW.S estimators. This finding implies that the IPSW methods are more sensitive to propensity model misspecification for the regression coefficient estimation as compared to the mean/prevalence estimation. The KW.S and IPSW.S estimators had smaller variances than the KW.W and the IPSW estimates, respectively. Therefore, the KW.S estimators on average had smallest MSE among all the PS-based estimates on average.

In the real data example, the risk of mortality in 15-years among the obese people was significantly higher than that among the nonobese people in the US obtained from the 1994 NHIS. However, the risk of 15-year mortality was not significantly different from the obese and nonobese adults in the naïve NHANES III sample. All PS-based methods substantially reduced the bias of the naïve NHANES estimator and captured the significant association between obesity and mortality except for the KW method due to the large JK variance estimate.

As observed in simulations and the real data example, the KW estimators of regression coefficients can be more sensitive to propensity model misspecification than the KW.W and the KW.S estimators. Future research is needed to establish a general framework of PS estimation for regression analyses. The JK method is more likely to overestimate the variance of the KW estimates of regression coefficients when the propensity model is misspecified. In future research, the complete TL variance estimation should be developed to incorporate the randomness due to PS estimation.

This chapter reveals the importance of the cohort representativeness in regression analysis when the cohort participation is informative or when the regression model is misspecified. There are more cases where the cohort representativeness should be considered under the framework of regression analysis. For example, we are interested in estimating the total effect of an exposure on the outcome (having a disease or not). If the cohort participation depends on a mediator of the exposure and the outcome variable, the naïve cohort estimate of the total effect of the exposure would be biased. This is because the cohort participation is correlated with the outcome given the exposure. However, the mediator cannot be controlled in the outcome model, because part of the total effect can be absorbed by the mediator. The PS-based methods can help reduce bias in estimating total effect of the exposure in this situation by including the mediator as a covariate in the propensity model.

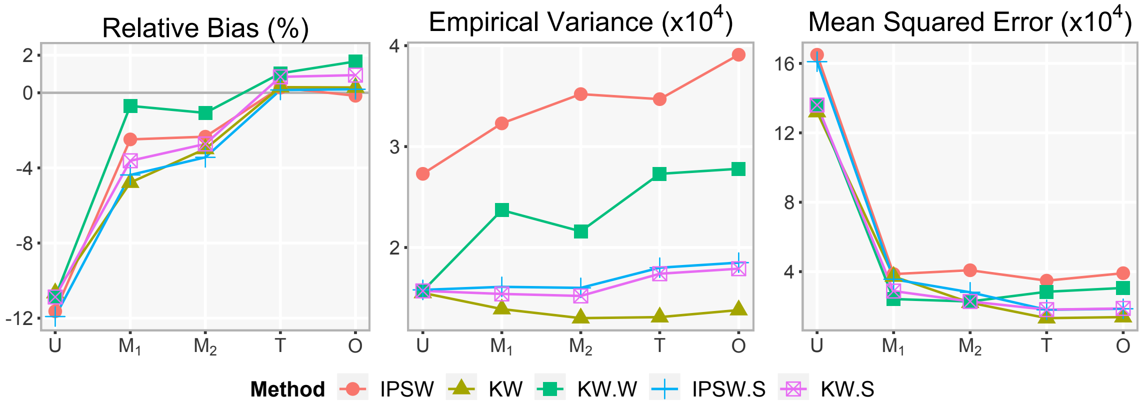
5.6 Supplementary Figures and Tables

Table 5.6 Results of regression coefficient estimation of the misspecified outcome model from 2,000 simulated cohorts and survey samples under non-informative designs with each cohort and survey sample fitted to four misspecified propensity models†

Model	Estimator	%**RB**	**V** $(\times 10^3)$	**VR** (TL)	**VR** (JK)	**MSE** $(\times 10^3)$
	$\hat{\theta}_2^{*\,Naive}$	35.01	4.73	0.48		211.36
	$\hat{\theta}_2^{*\,SVY}$	0.60	13.80	1.05	1.06	13.86
Model U: $\text{logit}(p) \sim age$						
	$\hat{\theta}_2^{*\,IPSW}$	-9.86	6.11	0.84	0.99	22.49
	$\hat{\theta}_2^{*\,KW}$	-11.84	5.58	0.94	0.99	29.21
	$\hat{\theta}_2^{*\,KW.W}$	-10.82	5.46	0.95	0.98	25.18
	$\hat{\theta}_2^{*\,IPSW.S}$	-10.07	5.48	0.93	0.98	22.56
	$\hat{\theta}_2^{*\,KW.S}$	-10.82	5.46	0.95	0.98	25.18
Model M_1: $\text{logit}(p) \sim age, Env$						
	$\hat{\theta}_2^{*\,IPSW}$	-18.01	9.25	0.75	1.01	63.95
	$\hat{\theta}_2^{*\,KW}$	-12.91	5.46	1.02	1.26	33.55
	$\hat{\theta}_2^{*\,KW.W}$	-4.87	6.22	0.80	1.09	10.22
	$\hat{\theta}_2^{*\,IPSW.S}$	-16.12	6.61	0.96	0.98	50.42
	$\hat{\theta}_2^{*\,KW.S}$	-6.86	5.38	0.92	1.09	13.32
Model M_2: $\text{logit}(p) \sim age, Env, age \cdot Env$						
	$\hat{\theta}_2^{*\,IPSW}$	-8.96	35.77	0.17	1.04	49.29
	$\hat{\theta}_2^{*\,KW}$	-5.71	5.40	1.01	1.38	10.89
	$\hat{\theta}_2^{*\,KW.W}$	-1.77	13.27	0.37	1.26	13.80
	$\hat{\theta}_2^{*\,IPSW.S}$	-4.73	8.30	0.67	1.00	12.07
	$\hat{\theta}_2^{*\,KW.S}$	-1.67	6.62	0.75	1.25	7.09
Model O: $\text{logit}(p) \sim age, z, Env, race\_ethnicity$						

$\hat{\theta}_2^{*\,IPSW}$	1.74	18.15	0.33	1.07	18.66
$\hat{\theta}_2^{*\,KW}$	-0.87	5.93	1.00	1.55	6.05
$\hat{\theta}_2^{*\,KW.W}$	3.68	12.41	0.44	1.18	14.70
$\hat{\theta}_2^{*\,IPSW.S}$	0.86	7.55	0.79	1.00	7.67
$\hat{\theta}_2^{*\,KW.S}$	2.16	6.93	0.81	1.13	7.71

Figure 5.4 Results of finite population prevalence estimation in Scenario (2) from 2,000 simulated cohorts and survey samples under non-informative designs with each cohort and survey sample fitted to four misspecified propensity models†



†The labels of the x-axises represent the propensity models as follows

Model U: underfitted model logit$(p) \sim age$

Model M_1: misspecified model logit$(p) \sim age\,, Env$

Model M_2: misspecified model logit$(p) \sim age\,, Env, age \cdot Env$

Model T: true model logit$(p) \sim age\,, z$

Model O: overfitted model. logit$(p) \sim age\,, z, Env, race/ethnicity$

Table 5.7 Results of finite population prevalence estimation in Scenario (2) from 2,000 simulated cohorts and survey samples under non-informative designs with each cohort and survey sample fitted to four misspecified propensity models

Model	Estimator	%RB	**V** $(\times 10^3)$	**VR** (TL)	**VR** (JK)	**MSE** $(\times 10^3)$
	$\hat{\mu}^{Naive}$	42.665	1.69			187
	$\hat{\mu}^{SVY}$	-1.076	1.17	1.05	1.05	1.29
Model U: logit(p) $\sim age, z$						
	$\hat{\mu}^{IPSW}$	0.276	3.47	0.36	1.04	3.48
	$\hat{\mu}^{IPSW.S}$	0.298	1.31	0.95	1.03	1.32
	$\hat{\mu}^{KW}$	1.035	2.73	0.46	1.04	2.84
	$\hat{\mu}^{KW.W}$	0.149	1.80	0.69	1.03	1.80
	$\hat{\mu}^{KW.S}$	0.848	1.74	0.72	1.04	1.81
Model U: logit(p) $\sim age$						
	$\hat{\mu}^{IPSW}$	-11.633	2.73	0.42	1.01	16.5
	$\hat{\mu}^{IPSW.S}$	-10.661	1.55	0.74	0.97	13.2
	$\hat{\mu}^{KW}$	-10.873	1.57	0.73	1.00	13.6
	$\hat{\mu}^{KW.W}$	-11.914	1.58	0.72	0.97	16.1
	$\hat{\mu}^{KW.S}$	-10.873	1.57	0.73	0.99	13.6
Model M_1: logit(p) $\sim age, Env$						
	$\hat{\mu}^{IPSW}$	-2.483	3.23	0.37	1.02	3.86
	$\hat{\mu}^{IPSW.S}$	-4.781	1.39	0.85	1.00	3.72
	$\hat{\mu}^{KW}$	-0.699	2.37	0.51	1.05	2.42
	$\hat{\mu}^{KW.W}$	-4.376	1.61	0.74	1.02	3.56
	$\hat{\mu}^{KW.S}$	-3.621	1.54	0.77	1.04	2.88
Model M_2: logit(p) $\sim age, Env, age \cdot Env$						
	$\hat{\mu}^{IPSW}$	-2.339	3.52	0.34	1.02	4.08
	$\hat{\mu}^{IPSW.S}$	-2.987	1.30	0.92	1.01	2.21
	$\hat{\mu}^{KW}$	-1.071	2.16	0.56	1.06	2.28
	$\hat{\mu}^{KW.W}$	-3.437	1.60	0.75	1.02	2.81
	$\hat{\mu}^{KW.S}$	-2.725	1.52	0.79	1.04	2.28
Model O: logit(p) $\sim age, z, Env, race\_ethnicity$						
	$\hat{\mu}^{IPSW}$	-0.159	3.91	0.32	1.03	3.91
	$\hat{\mu}^{IPSW.S}$	0.282	1.38	0.90	1.00	1.38
	$\hat{\mu}^{KW}$	1.668	2.78	0.46	1.05	3.06
	$\hat{\mu}^{KW.W}$	0.186	1.85	0.67	1.00	1.86
	$\hat{\mu}^{KW.S}$	0.939	1.79	0.70	1.02	1.88

Chapter 6 Discussion and Future Work

6.1 Summary

In this dissertation, A new PS-based KW (matching) approach is proposed to improve external validity of cohort analyses, using a representative survey sample as a reference sample for the target population. The KW method, which is a PS-based matching method, uses PS to measure similarity between the cohort and survey sample units, and therefore is less sensitive to the propensity model misspecification compared to the PS-based weighting methods. In addition, the KW method relaxes the PSAS assumption of identical representativeness of the cohort units within subclasses by fractionally distributing survey sample weights to the cohort units based on their similarity measured by kernel smoothed distance in PS. The KW method provides consistent estimators of population means/prevalences/associations under the true propensity model and some regularity assumptions.

To avoid high variability due to using the survey weighting in PS estimation, the KW approach as well as other existing PS-based matching methods estimates PS by fitting a propensity model to the combined (cohort vs. unweighted survey) sample. However, as we found in our research, it requires the strong exchangeability assumption (SEA) for estimating the finite population means. The SEA states that the expectation of the analysis variable given the PS is the same among all three of the cohort, reference survey, and the finite target population. It is proved in this dissertation that, without the SEA, current PS-based matching estimators can be biased, even under the correct propensity model.

A novel unifying framework is established in the dissertation for both PS-based weighting and matching methods. This unifying framework allows this dissertation to make three contributions. First, the SEA is identified for the original PS-based matching methods. The simulations and data example demonstrate that the PS-based matching methods that rely on the SEA, such as the original KW estimator, have the smallest mean squared error (MSE) when the SEA holds, but have large bias when the SEA fails. Second, as a remedy, the enhanced PS-based matching methods are proposed without requiring WEA under the framework, rather than the SEA. Third, the efficiency of the PS-based estimates is further improved by scaling the survey weights to sum to the survey sample size. Scaling the survey sample weights reduces the variance of the estimated PSs and thus markedly improves the efficiency of the pseudo-weighted estimates, especially under the IPSW method. The kernel weighting with scaling (KWS) method is most recommended because of its robustness to propensity model misspecification, and the smallest MSE in general.

The TL and JK variance estimations are developed under the framework to take into account all sources of variability in the final pseudo-weighted estimates. We recommend JK method for estimating variances of the original IPSW estimator because the empirical results showed that the TL method can have finite sample bias due to highly variable weights in the combined sample. 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.

The PS-based methods were developed to reduce bias when the outcome variables are available in cohorts but not in surveys, such as novel molecular or genetic risk factors.

In the data examples, the outcome variables were purposely selected so that they were available in both the cohort and the survey, allowing us to quantify the relative bias by assuming the survey estimates as the gold standard. However, the survey estimates can vary from the truth due to sampling errors and non-sampling errors such as undercoverage and nonresponse bias. Unfortunately, there is no census of reported diseases in the United States.

The simulations provide guidance for choosing propensity model predictors. For the propensity model, Stuart (2010) suggested including all variables that may be associated with treatment assignment and the outcomes to reduce bias, but for small samples, it is useful to prioritize variables related to the outcome to control the variance (Brookhart et al., 2006). The simulations agree that adding extra predictors of the outcome in the propensity model reduces bias, but at a potential cost of increasing variance, especially for the original IPSW method (Chapter 3). I suggested that the propensity models aim for maximal bias reduction by including all variables distributed differently in the cohort and the survey sample, all significant interaction terms, and all variables predictive of the outcome.

All the PS-based methods assume the final weights of the probability survey sample are the inverse of true inclusion probabilities from the finite population. However, ideal survey weights are likely unachievable due to imperfect undercoverage and nonresponse adjustments. The accuracy of the survey weights may substantially affect the bias reduction of the IPSW method because this method uses survey sample weights for PS estimation and estimates the participation rates directly from (functions of) the PS. On the other hand, the matching methods empirically show less sensitivity to accuracy of the survey sample weights because they use the PS to measure the similarity between the cohort and the survey sample units.

6.2 Future Work

As nonprobability samples become more and more popular in many areas in the era of big data, more attention should be paid to making finite population inferences from the nonprobability samples. Although the PS-based methods can improve the external validity of estimates from the nonprobability samples, there is much room for future research before these methods can be widely used in practice by epidemiologists and medical researchers.

6.2.1 Propensity model diagnosis

The amount of bias reduction of the PS-based methods crucially depends on how well the propensity model predictors predict the outcome. If the propensity model is poorly fitted, then the PS-based estimates can even be more biased than the naïve cohort estimates. Furthermore, including all known variables in the propensity model may not suffice for meaningful bias reductions. Further research is needed for developing propensity model selection and diagnostics to identify situations in which the PS-based method might reduce little bias, or even increase bias.

The PS-based matching methods fitting propensity model to the combined (cohort vs. *unweighted* survey) sample provide efficient and unbiased estimators of finite population means only under SEA. A visual scatterplot is proposed to assess if the SEA holds in a situation where WEA is approximately held. However, there remains much room for developing formal propensity model diagnostics for SEA.

6.2.2 Optimal scaling factor

In Chapter 4, the scaled survey weights are proposed in propensity estimation to improve efficiency of pseudo-weighted estimates of finite population means. The scaling factor is chosen so that the scaled survey weights sum up to the survey sample size. A substantial gain in efficiency is observed in the simulations and the real data example. However, more theoretical justification is needed for quantifying the efficiency improvement. More work is needed to determine optimal scaling when rescaling the survey weights issued in propensity estimation to minimize the variance of the pseudo-weighted cohort estimates. Moreover, the effect of scaling weights on efficiency improvement may differ depending on the propensity model misspecification. For example, Kim & Skinner (2013) found that the weighting adjustments for regression analysis under an informative sample may increase the variance of the estimated regression coefficients if the analysis model is misspecified. Future research is needed for investigating properties of the estimators using scaled survey weights.

6.2.3 Doubly robust estimators

As discussed in Chapter 3 and Chapter 4, the PS-based estimators of finite population means can be biased when the propensity model is misspecified. The PS-based weighting methods, such as IPSW, are more sensitive to model misspecification than the matching methods, especially when there are extreme weights. In order to improve the robustness of their PS-based weighting method, Chen et al. (2019) proposed a doubly robust estimator of finite population means by combining a pseudo-weighted estimator from the nonprobability sample and a survey estimator obtained by the model-based prediction method.

The model-based prediction approach has also been explored for finite population inference from nonprobability samples (Elliott & Valliant, 2016; Chen et al., 2019). Suppose the outcome model in the finite population is:

$$y_i = m(\boldsymbol{x}_i), i \in FP, \tag{6.2.1}$$

where $m(x_i) = E_m(\, y_i \mid \boldsymbol{x}_i\,)$ with the expectation E_m is respective to the distribution of the outcome y given the covariates $\boldsymbol{x}$. Under the assumption **A1** in Section 4.2 that the cohort participation and the outcome variable are independent conditional on the covariates, $E_m(\, y_i \mid \boldsymbol{x}_i, R_i = 1\,) = E_m(\, y_i \mid \boldsymbol{x}_i\,) = m(x_i)$. Hence the outcome model is robust to cohort selection, and $m(\boldsymbol{x})$ can be modeled from the naïve cohort. The estimate of $m(\boldsymbol{x})$ is denoted by $\widehat{m}(\boldsymbol{x})$. The doubly robust estimator (Chen et al., 2019) of the finite population mean ($\mu^{FP} = \frac{1}{N}\sum_{i\in FP} y_i$) combines a pseudo-weighted cohort estimator and a survey estimator as follows:

$$\hat{\mu}^{DR} = \frac{1}{\widetilde{\widehat{N}}}\sum\nolimits_{i\in s_c} \widetilde{w}_i \cdot \{y_i - \widehat{m}(\boldsymbol{x}_i)\} + \frac{1}{\widehat{N}^{SVY}}\sum\nolimits_{i\in s_s} d_i \cdot \widehat{m}(\boldsymbol{x}_i) \tag{6.2.2}$$

where $\widetilde{\widehat{N}} = \sum_{i\in s_c} \widetilde{w}_i$ is the cohort estimator of the finite population count, $\widetilde{w}_i$ is the pseudo-weight for $i \in s_c$ obtained by a PS-based method, and $\widehat{N}^{SVY} = \sum_{i\in s_s} d_i$ is the survey estimate of finite population count.

Chen et al. (2019) proved that their doubly robust estimator is consistent if either the propensity model or the outcome prediction model is correct. Increased efficiency can be achieved if both models are correct. Similar doubly robust estimators can be constructed using the IPSW or the KW pseudo-weights in the first summand of Formula (6.2.2). Future research should investigate further doubly robustness properties, including consistency and finite population variance, when using the IPSW or the KW methods to obtain the pseudo-weights.

6.2.4 Combining weight trimming and the PS-based methods

Weight trimming, which sets weights above some maximum to that maximum, has been proposed as a solution to reducing variance of the sampled weighted estimates (Potter, 1993). It can be applied to the PS-based pseudo weights to improve efficiency of the pseudo-weighted estimates of the finite population quantities, especially for the IPSW methods that are more likely to produce extreme weights. However, the effect of weight trimming on increasing bias or reducing variance is unclear (Lee et al., 2010; Potter and Zheng, 2015), and there is relatively little guidance regarding the trimming level. Future research is needed to investigate how to minimize mean squared errors by trimming the pseudo-weights. The variance estimation also has to be adapted to reflect the trimmed pseudo-weights.

6.2.5 Applying KW methods to adjust nonresponse bias in survey research

Discussed in Section 2.2.3, estimates based on the respondents' data alone can be biased if the unobserved distribution of study variables from nonrespondents (s_{nr}) is different from that of respondents (s_r).

Weighting adjustment cell (WAC) method (described in Little & Vartivarian, 2003) and inverse of response-propensity weighting (IRW) method (Iannacchione et al., 1991) are two commonly used nonresponse adjustment methods (Korn and Graubard, 1999). These two methods have similar advantages and disadvantages of the PSAS and the IPSW method respectively. The WAC method divides the sampled individuals (respondents +

nonrespondents) into disjoint groups ("weighting adjustment cells"). Then it assigns a common nonresponse adjustment (i.e., the number of sampled individuals divided by the number of respondents) by assuming probabilities of responding for all sampled individual are the same within a given cell. Similar to the PSAS method, the WAC method is less likely to produce extreme adjustment factors but can be less efficient in reducing nonresponse bias due to the assumption of equal response propensity within the cell. The IRW method is formed by modeling the probability of responding as a function (e.g., logistic regression) of variables available on all sampled units. The nonresponse adjustment factor is taken as the inverse of the estimated probability of response. This method, like the IPSW method, can reduce more nonresponse bias, but can be sensitive to the response model misspecification and can produce extreme adjustment factor.

The KW method can be applied to reduce the nonresponse bias, aiming for reducing MSE, in two ways. The first way is to fit a responding propensity model (the dependent variable =1 for $i \in s_r$; =0 for$j \in s_{nr}$) to the *unweighted* sample, as the IRW method does. Then, assign the nonresponse adjustment

$$f_i^{KW1} = 1 + \sum\nolimits_{j \in s_{nr}} \frac{K\left\{\left(\tilde{p}_i^{(r)} - \tilde{p}_j^{(nr)}\right)/h\right\}}{\sum_{i \in s_r} K\left\{\left(\tilde{p}_i^{(r)} - \tilde{p}_j^{(nr)}\right)/h\right\}}, \text{ for } i \in s_r, \qquad (6.2.3)$$

where $\tilde{p}_i^{(r)}$ and $\tilde{p}_j^{(nr)}$ are the propensity of responding for $i \in s_r$ and for $j \in s_{nr}$ in the *sample* respectively. The KW adjusted sample weight for $i \in s_r$ is the production of f_i^{KW1} and the original base weight d_i: $w_i^{KW1} = f_i^{KW1} \cdot d_i$.

The second way is to fit the responding propensity model (the dependent variable =1 for $i \in s_r$; =0 for$j \in s_{nr}$) to the *weighted* sample and obtain the KW2 adjusted sample weight

$$w_i^{KW2} = d_i \left(1 + \sum_{j \in s_{nr}} \frac{d_j K\left\{ \left(p_i^{(r)} - p_j^{(nr)} \right) / h \right\}}{\sum_{i \in s_r} d_i K \left\{ \left(p_i^{(r)} - p_j^{(nr)} \right) / h \right\}} \right), \quad \text{for } i \in s_r, \qquad (6.2.4)$$

where $p_i^{(r)}$ and $p_j^{(nr)}$ are the propensity of responding for $i \in s_r$ and for $j \in s_{nr}$ in the *population* respectively.

The two methods require different conditions for consistent estimates of finite population quantities. Moreover, the TL finite population variance and the variance estimation can be harder to obtain because the sets of respondents and nonrespondents are mutually exclusive, not independent. The variance estimation needs to take the correlation between the two samples into account. Future research is needed to examine the performance the KW nonresponse adjustment and to derive the variance estimation.

6.2.6 Extending PS-based methods to different epidemiologic study designs

Using PS-based methods to estimate finite population means and associations from the cohort is a starting point for improving external validity of cohort analyses. The ultimate goal is to extend the PS-based methods to other epidemiologic study designs.

For example, many epidemiologic studies involve sampling within cohorts, such as the nested case-control studies, case-cohort studies, or general two-phase sampling (Li et al., 2016). PS-based methods can be adapted to improve the representativeness of the sub-samples for the full cohort, which is important in situations where the subsampling is not totally under control of study investigators. Accordingly, the propensity model should be fitted to the combined subsample versus the full cohort, using as the full cohort as the reference. We expect that the PS-based methods should perform better for these studies compared to improving representativeness of the cohort to the finite population, because problems related to differences in data processing errors, measurement errors, and coverage errors due to different study designs in the cohort and the external reference survey sample could be mitigated or entirely absent between the subsample and the cohort.

Similarly, the PS-based methods can to be applied to estimating risk models (e.g., probability of disease in a particular time period) not only from cohort studies, but also from case-control studies. Due to potential unrepresentativeness of a cohort, the absolute risk estimates obtained from the naïve cohorts may not be generalizable to the population. The PS-based methods can be applied to improve the external validity of the absolute risk estimation from the cohorts. For example, synthetic population-based case-control studies (Tota et al., 2019) can be constructed to estimate individualized risk of rare diseases, in which the cases are from nonprobability case studies and weighted using a population-based disease registry (e.g., Surveillance, Epidemiology, and End Results [SEER] Program), and the controls are from population-based health surveys (e.g. NHIS, NHANES).

I hope that my work will promote attention to improving external validity of cohort analyses with the goal of developing reliable methodology and software for medical researchers.

APPENDIX

R code Example for generating Table 3.2

```
######################################################################
#                       Simulations in Chapter 3                     #
######################################################################
library(survey)
# Note: please load subfunctions (included at the end) before running
 the main program
source("Sampling.R")
source("getY.R")
source("Kernel functions.R")
source("PS weights.R")
source("JRR.R")
source("simu_fun_c3.R")
# Read dataset of finite population (FP)
fnt.pop = read.table("/Users/wangl29/Box/Research/Lingxiao Projects/KW
 codes/fnt.pop_reg_all.txt",
                     header = T)
# Set up random seeds
seed = read.table("/Users/wangl29/Box/Research/Lingxiao Projects/KW
 codes/seed.txt")
seed1=seed[,1]
seed2=seed[,2]
# Basic setups
NSIMU = 2 # number of simulation runs
N = 9000000 # FP count
Cluster = 3000 # number of clusters in FP
n_c = 100 # cohort size
psu_c = 10 # number of clusters selected by the cohort
n_s = 100 # survey sample size
psu_s = 10 # number of clusters selected by the survey sample

# Generate outcome variable y and the substitute
theta = c(-5,0.5,-1,1,0.3, 0.1)
n.theta = length(theta)
set.seed(97035)
fnt.pop$hisp=(fnt.pop$race_eth==4)
fnt.pop$ageE = fnt.pop$age*fnt.pop$Env1
fnt.pop$ageE[fnt.pop$ageE>15]=15
cov.m = as.matrix(cbind(1, fnt.pop$age, fnt.pop$gender, fnt.pop$hisp,
                        fnt.pop$Env1, fnt.pop$ageE))
out_y = getY(thetas = theta, design.x = cov.m, ICCy=0.07,Cluster=3000,
 clustersize=3000)
```

```
y_r1 = out_y$y; mu=mean(y_r1); mu
py = out_y$py; mean(py)
fnt.pop$y = y_r1; fnt.pop$py=py; fnt.pop$z=py-1.7*fnt.pop$u

# Coefficient for generating MOS of PPS sampling
fnt.pop$Env_h = as.numeric(fnt.pop$Env1>2.5)
gamma.t = c(0,-0.3, 0.4, -0.7, -0.7)
n.gamma = length(gamma.t)
# Calculate individual level MOS
odds_trt = exp(as.matrix(cbind(1, fnt.pop$age, fnt.pop$hh_inc,
                               fnt.pop$Env_h, fnt.pop$z))%*%
               matrix(gamma.t, n.gamma, 1))

#True PS model
Formula_t = as.formula("trt ~ age+hh_inc+Env_h+z")
#Fitted PS model
Formula_fit = as.formula("trt ~ age+hh_inc+Env_h+z")
# number of coefficients in the fitted PS model
n.beta.fit=5
# Calcualte MOS for the first stage (clustering) PPS sampling
size.I_c = aggregate(odds_trt^2.5, list(fnt.pop$psu), sum)[, 2]
size.I_s = aggregate(1/(odds_trt^1.2), list(fnt.pop$psu), sum)[, 2]

# Name the variable storing the output
est    = matrix(0, NSIMU, 6) # Estimates of FP proportion of y=1
var1   = matrix(0, NSIMU, 6) # Naïve TL estimates of standard error
var2   = matrix(0, NSIMU, 6) # JK estimates of standard error
h_out = rep(0, NSIMU)
# Estimates of the propensity model coefficients
beta_est    = matrix(0, NSIMU, n.beta.fit) # PSAS and KW
beta.w_est = matrix(0, NSIMU, n.beta.fit) # IPSW

for(simu in c(1: NSIMU)){
  #Select a cohort
  samp.c = samp.slct(seed     = seed2[simu],  fnt.pop = fnt.pop,
                     n        = n_c,          Cluster = Cluster,
                     Clt.samp = psu_c,        dsgn    = "pps-pps",
                     size     = odds_trt^2.5, size.I  = size.I_c)

  # Naïve cohort estimate
  est [simu, 1]  = mean(samp.c$y); var1[simu, 1]  = var(samp.c$y)/n_c
  # Cohort estimate considering the true sample weights
  ds = svydesign(ids=~psu, data=samp.c, weights=~wt, nest=TRUE)
  ds.rep = as.svrepdesign(ds)
  out_c = svymean(~y, design = ds)
  est [simu, 2] = out_c[1]; var1[simu, 2] = vcov(out_c)
  out_c.jrr = svymean(~y, design = ds.rep)
  var2[simu, 2] = vcov(out_c.jrr)
  # Select a probability sample
  samp.s = samp.slct(seed     = seed2[simu],      fnt.pop  = fnt.pop,
                     n        = n_s,              Cluster  = Cluster,
                     Clt.samp = psu_s,            dsgn     = "pps-pps",
                     size     = 1/(odds_trt^1.2), size.I   = size.I_s)
  # Sample weighted survey sample estimate
  ds = svydesign(ids=~psu, data=samp.s, weights=~wt, nest=TRUE)
  ds.rep = as.svrepdesign(ds)
  # Population estimate of disease prevalence using survey sample
  out_s = svymean(~y, design = ds)
  est [simu, 3] = out_s[1]; var1[simu, 3] = vcov(out_s)
  out_s.jrr = svymean(~y, design = ds.rep)
  var2[simu, 3] = vcov(out_s.jrr)

  # PS-based methods (IPSW, PSAS, and KW)
  samp.c_all = simu_fun_c3(chtsamp = samp.c, svysamp = samp.s,
                           svy_wt  = "wt",   Formula = Formula_fit,
                           krn     = "triang")
  # Coefficients of propensity score models
  beta_est[simu, ]   = samp.c_all$beta
beta.w_est[simu, ] = samp.c_all$beta.w
# IPSW estimate
  ds.ipsw  = svydesign(ids=~psu, data= samp.c_all$chtsamp_adj,
 weights=~ ipsw, nest=TRUE)
  out_ipsw = svymean(~y, design = ds.ipsw)
  est [simu, 4] = out_ipsw[1]; var1[simu, 4] = vcov(out_ipsw)
  # PSAS estimate
  ds.psas  = svydesign(ids=~psu, data= samp.c_all$chtsamp_adj,
 weights=~ psas, nest=TRUE)
  out_psas = svymean(~y, design = ds.psas)
  est [simu, 5] = out_psas[1]; var1[simu, 5] = vcov(out_psas)
  # KW estimate
  ds.kw  = svydesign(ids=~psu, data= samp.c_all$chtsamp_adj, weights=~
 kw, nest=TRUE)
  out_kw = svymean(~y, design = ds.kw)
  est [simu, 6] = out_kw[1]; var1[simu, 6] = vcov(out_kw)
  h_out[simu] = samp.c_all$h

  # JK variances for the IPSW, PSAS, and KW estimates
  theta = JRR_var(chtsamp = samp.c,       svysamp = samp.s,
                  svy_wt  = "wt",         h_in    = h_out[simu],
                  psu     = "psu",        resp    = "y",
                  Formula =  Formula_fit, krn     = "triang")
  sum_sq = (theta$theta - c(out_ipsw[1], out_psas[1], out_kw[1]))^2
  var2[simu, c(4:6)] = apply(t(t(sum_sq)*c(rep((psu_c-1)/psu_c, psu_c),
                                           rep((psu_s-1)/psu_s,
 psu_s))),
                             1, sum)
  print(simu)} #End of simulation

# Generating Table 3.2
mean(h_out) #Bandwidth
# Results
mu_hat = apply(est, 2, mean); relb_mu = (mu_hat- mu)/mu*100 # Relative
 Bais
ev.mu = apply(est, 2, var) # Empirical variance

# Analytical variances (Naïve TL, cand JK), and Variance Ratios
av.mu1 = apply(var1, 2, var); vr.mu1 = av.mu1/ev.mu # Naïve TL
av.mu2 = apply(var2, 2, var); vr.mu2 = av.mu2/ev.mu # JK
mse.mu = ev.mu + (mu_hat-mu)^2 # MSE

ci_lw.ntl = est - 1.96*sqrt(var1)
ci_lw.jk  = est - 1.96*sqrt(var2)
ci_up.ntl = est + 1.96*sqrt(var1)
ci_up.jk  = est + 1.96*sqrt(var2)

cp.ntl = apply(sapply(1:6, function(i) (mu>= ci_lw.ntl[,i])&(mu<=
 ci_up.ntl[,i])), 2, mean)
cp.jk  = apply(sapply(1:6, function(i) (mu>= ci_lw.jk[,i]) &(mu<=
 ci_up.jk[,i])),  2, mean)


# Generate Table 3.2
y_out = t(rbind(RB = relb_mu, V = ev.mu*1e5,
                VR_NTL = vr.mu1, VR_JK = vr.mu2,
                CP_NTL = cp.ntl, CP_JK = cp.jk, MSE = mse.mu*1e5))
rownames(y_out) = c("Naive", "Cht", "Svy", "IPSW", "PSAS", "KW")
round(y_out, 7)


######################################################################
#                            SUBFUNCTIONS                            #
######################################################################

######################################################################
# FUNCTION simu_fun_c3 is a function for 1 simulation run in         #
#         in Chapter 3                                               #
# Input                                                              #
#  chtsamp: cohort sample                                            #
#  svysamp: survey sample                                            #
#  svy_wt:  survey sample weight                                     #
#  rm.s:    whether remove survey sample units that are not matched  #
#           with any cohort units for KW method or not (default is F)#
#  h:       pre-set banddwith. If NULL, h will be calculated         #
# Output                                                              #
#  chtsamp_adj: cohort sample with the IPSW, PSAS, and KW weights    #
#  h:           bandwidth                                             #
#######################################################################
simu_fun_c3 = function(chtsamp,svysamp,svy_wt, Formula, krn = "dnorm",
rm.s = F, h=NULL){
  krn <<- krn; rm.s<<- rm.s; h   <<-h
  # Get names of the response variable and predictors for the
  # propensity score estimation model
  # response variable and covariates in the propensity model
  Fml_names = all.vars(Formula)
  rsp_name = Fml_names[1] # response variable
  mtch_var = Fml_names[-1] # covariates
  # Remove incomplete records in the cohort, if there are any.
  chtsamp_sub = as.data.frame(chtsamp[, mtch_var])
  if(sum(is.na(chtsamp_sub))>0){
    cmplt.indx = complete.cases(chtsamp_sub)
    chtsamp_sub = chtsamp_sub[cmplt.indx, ]
    chtsamp = chtsamp[cmplt.indx,]
    warning("Missing values in covariates are not allowed. Records with
missing values in the cohort are removed.")
  }
  # Remove incomplete survey sample, if there are any.
  svysamp_sub = as.data.frame(svysamp[, mtch_var])
  svy_wt.vec = c(svysamp[, svy_wt])
  if(sum(is.na(svysamp_sub))>0){
    cmplt.indx = complete.cases(svysamp_sub)
    svysamp_sub = svysamp_sub[cmplt.indx, ]
    svy_wt.vec = sum(svysamp[,
svy_wt])/sum(svy_wt.vec[cmplt.indx])*svy_wt.vec[cmplt.indx]
    warning("Missing values in covariates are not allowed. Records with
missing values in the survey sample are removed.
            The complete cases are reweighted. Missing completely at
random is assumed.")
  }
  m = dim(chtsamp_sub)[1] # size of cohort (complete cases)
  n = dim(svysamp_sub)[1] # size of survey sample (complete cases)

  # Combine the two complete samples
  chtsamp_sub[,rsp_name] = 1; svysamp_sub[,rsp_name] = 0
  names(chtsamp_sub) = c(mtch_var, rsp_name)
  names(svysamp_sub) = c(mtch_var, rsp_name)
  psa_dat = rbind(chtsamp_sub, svysamp_sub)

  # Fit logistic regression model to predict propensity scores
  svyds = svydesign(ids =~1, weight = rep(1, m+n), data = psa_dat)
  lgtreg = svyglm(Formula, family = binomial, design = svyds)
  # regression coefficients of the propensity model fitted to the
unweighted sample
  beta = summary(lgtreg)$coeff[, 1]
  p_score = lgtreg$fitted.values
  # Propensity scores for the cohort
  p_score.c = p_score[psa_dat[,rsp_name]==1]
  # Propensity scores for the survey sample
  p_score.s = p_score[psa_dat[,rsp_name]==0]

  # Fit logistic regression model to predict propensity scores (with
weights)
  psa_dat$wt_cmb = c(rep(1, m), svy_wt.vec)
  ds = svydesign(ids=~1, weight = ~ wt_cmb, data = psa_dat)
  lgtreg.w = svyglm(Formula, family = binomial, design = ds)
  p_score.w = lgtreg.w$fitted.values
  p_score.w.c = p_score.w[psa_dat[,rsp_name]==1]
  # regression coefficients of the propensity model fitted to the
weighted sample
  beta.w = summary(lgtreg.w)$coeff[, 1]
  ################# Calculate pseudo weights #######################
  # calculate IPSW weights
  ipsw = ipsw.wt(p_score.c = p_score.w.c, svy.wt = svy_wt.vec)
  # calculate PSAS weights
  psas = psas.wt(p_score.c = p_score.c, p_score.s = p_score.s, svy.wt =
svy_wt.vec, nclass = 5)$pswt
  # calculate KW weights
  kw_out = kw.wt(p_score.c = p_score.c, p_score.s = p_score.s, svy.wt =
svy_wt.vec, Large=F)
  kw = kw_out$pswt; h = kw_out$h
  chtsamp_adj = cbind(chtsamp, ipsw = ipsw, psas = psas, kw = kw)

  return(list(chtsamp_adj = chtsamp_adj, h       = h,
              beta   = beta,beta.w = beta.w,
              p_dat   = data.frame(p   = p_score,
                                   trt = psa_dat[,rsp_name]),
              p.w_dat = data.frame(p   = p_score.w,
                                   trt = psa_dat[,rsp_name])))
}# end simu_fun_c3


######################################################################
# FUNCTION ipsw.wt is a function calculating pseudo weights using    #
#           IPSW method                                              #
# INPUT:                                                             #
#  p_score.c: predicted propensity score for cohort                  #
#  svy.wt:    a vector of survey weights                             #
# OUTPUT: pswt      - IPSW pseudo weights                           #
######################################################################
```

```r
ipsw.wt = function(p_score.c, svy.wt){
  pswt = as.vector((1-p_score.c)/p_score.c)
  pswt/sum(pswt)*sum(svy.wt)}
######################################################################
# FUNCTION psas.wt is a function calculating pseudo weights using    #
#          PSAS methods                                              #
# INPUT:                                                             #
#  p_score.c: predicted propensity score for cohort                  #
#  p_score.s: predicted propensity score for survey                  #
#  svy.wt:    a vector of survey weights                             #
#  nclass:    number of subclasses (by percentiles) for sample       #
#  division                                                          #
# OUTPUT                                                             #
#  pswt:   PSAS pseudo weights                                       #
#  nclass: actual number of subclasses used (empty classes will be   #
#  combined)                                                         #
# WARNINGS:                                                          #
#  If fewer than 2 cohort units in one or more subclasses            #
#  "Extreme weights may occur due to limited number(<=2) of cohort   #
#   units in some cells"                                             #
#  If there are subclasses including no cohort units                 #
#  "Empty subclasses were combined with the neighbor subclass."      #
######################################################################
psas.wt = function(p_score.c, p_score.s, svy.wt, nclass){
  nclass0 = nclass
  m = length(p_score.c); n = length(p_score.s)
  p_score = c(p_score.c, p_score.s); trt = c(rep(1, m), rep(0, n))
  p_score.q = quantile(p_score, prob = seq(0, 1, length = (nclass+1)))
  p_score.q.u = unique(p_score.q)
  nclass = length(p_score.q.u)-1
  subclass = cut(p_score, breaks = p_score.q.u, include.lowest = TRUE)
  levels(subclass) = c(1: nclass)
  nclass.c = length(unique(subclass[trt==1]))
  nclass.s = length(unique(subclass[trt==0]))
  while (nclass.c!= nclass.s){
    nclass = min(nclass.c, nclass.s)
    p_score.q = quantile(p_score,
                         prob = seq(0, 1, length = (nclass+1)))
    p_score.q.u = unique(p_score.q)
    nclass = length(p_score.q.u)-1
    subclass = cut(p_score, breaks = p_score.q.u,
                   include.lowest = TRUE)
    levels(subclass) = c(1: nclass)
    nclass.c = length(unique(subclass[trt==1]))
    nclass.s = length(unique(subclass[trt==0]))}
  p_score_dat = data.frame(id = c(1:m), subclass = subclass[trt==1])
  p_score_dat = p_score_dat[order(p_score_dat$subclass),]
  # Assign pseudo weights to the cohort units
  svy_N = aggregate(svy.wt, by=list(subclass[trt==0]), FUN = sum)[,2]
  cht_n = aggregate(rep(1, m), by=list(subclass[trt==1]),
                    FUN = sum)[,2]
  if(sum(cht_n<=2)>0) warning("Extreme weights may occur due to limited
 number(<=2) of cohort units in some cells")
  if(nclass<nclass0) warning("Empty cells were combined with the
 neighbor cells.")
  wt_f  = svy_N/cht_n
  pswt = rep(wt_f, cht_n)
  p_score_dat$pswt = rep(wt_f, cht_n)
  p_score_dat = p_score_dat[order(p_score_dat$id),]
  return(list(pswt = p_score_dat$pswt, nclass = nclass))
}# end psas.wt

######################################################################
# FUNCTION kw.wt is a function calculating pseudo weights using KW   #
#  methods                                                           #
# INPUT                                                              #
#  p_score.c: propensity score for cohort                            #
#  p_score.s: propensity score for survey                            #
#  svy.wt:    a vector of survey weights                             #
#  h:         bandwidth parameter (will be calculated corresponding  #
#             to kernel function if h is NULL).                      #
#  krnfun:    kernel function                                        #
#             "triang" -triangular density on (-3, 3)                #
#             "dnorm"  -standard normal density                      #
#             "dnorm_t"-truncated standard normal densigy on (-3, 3) #
#  Large:     if the cohort size is so large that it has to be       #
#             divided into pieces                                    #
#  rm.s:      removing unmatched survey units or not.                #
#             Default is FALSE                                       #
# OUTPUT                                                             #
#  psd.wt:   KW pseudo weights                                       #
#  delt.svy: number of unmatched survey sample units                 #
## WARNINGS                                                          #
#  If there are unmatched survey sample units, the program gives     #
#  "The input bandwidth h is too small. Please choose a larger one!" #
#  If rm.s=T, the program deletes unmatched survey sample units, and #
#  gives "records in the prob sample were not used because of a small#
#  bandwidth"   ##
#  If rm.s=F, the program evenly distribute weights of unmatched     #
#  survey sample units to all cohot units.                           #
######################################################################
kw.wt = function(p_score.c, p_score.s, svy.wt, h=NULL, mtch_v = NULL,
 krn="triang", Large = F, rm.s = F){
  # get the name of kernel function
  # calculate bandwidth according to the kernel function
  #triangular density
  if(is.null(h)){
  if(krn=="triang")h = bw.nrd0(p_score.c)/0.9*0.8586768
  if(krn=="dnorm"|krn=="dnorm_t")h = bw.nrd0(p_score.c)
  }
  krnfun = get(krn)
  # create signed distance matrix
  m = length(p_score.c); n = length(p_score.s)
   if (Large == F){
   sgn_dist_mtx = outer(p_score.s, p_score.c, FUN = "-")
   krn_num = krnfun(sgn_dist_mtx/h)
   if(is.null(mtch_v)){
    adj_m = 1
   }else{adj_m=outer(mtch_v[1:n], mtch_v[(n+1):(n+m)], FUN='==')}
   krn_num = krn_num*adj_m; row.krn = rowSums(krn_num)
   sum_0.s = (row.krn==0); delt.svy = sum(sum_0.s)
   if(delt.svy>0){
    warning('The input bandwidth h is too small. Please choose a
larger one!')
    if(rm.s == T){
     warning(paste(sum(sum_0.s), "records in the prob sample were
not used because of a small bandwidth"))
     row.krn[sum_0.s]=1}else{
     krn_num[sum_0.s,]= 1; row.krn[sum_0.s] = m}}
   row.krn = rowSums(krn_num); krn = krn_num/row.krn
   # Final pseudo weights
   pswt_mtx = krn*svy.wt; psd.wt = colSums(pswt_mtx)}else{
   psd.wt = rep(0, m); grp_size =  floor(n/50)
   up = c(seq(0, n, grp_size)[2:50], n); lw = seq(1, n, grp_size)[-51]
   delt.svy = 0
   for(g in 1:50){
    sgn_dist_mtx = outer(p_score.s[lw[g]:up[g]], p_score.c,
                          FUN = "-")
    krn_num = krnfun(sgn_dist_mtx/h)
    if(is.null(mtch_v)){
     adj_m = 1}else{
    adj_m=outer(mtch_v[lw[g]:up[g]], mtch_v[(n+1):(n+m)], FUN='==')}
    krn_num = krn_num*adj_m; row.krn = rowSums(krn_num)
    sum_0.s = (row.krn==0); delt.svy = delt.svy + (sum(sum_0.s)>0)
    if((sum(sum_0.s)>0)){
     warning('The input bandwidth h is too small. Please choose a
larger one!')
     if(rm.s == T){
      warning(paste(sum(sum_0.s), "records in the prob sample were
not used because of a small bandwidth"))
      row.krn[sum_0.s]=1}else{
      krn_num[sum_0.s,]= 1; row.krn[sum_0.s] = m}}
    row.krn = rowSums(krn_num); krn = krn_num/row.krn
    # Final psuedo weights
```

```
      pswt_mtx = krn*svy.wt[lw[g]:up[g]]
      psd.wt = colSums(pswt_mtx) + psd.wt}}
  return(list(pswt = psd.wt, delt.svy = delt.svy, h = h))
} # end of kw.wt

#######################################################################
# FUNCTION samp.slct is a function to select a sample                 #
# INPUT:                                                              #
#  seed:     random seed that enables replication                     #
#  fnt.pop:  the finite population                                    #
#  n:        sample size                                              #
#  Cluster:  total number of clusters in the finite population        #
#  Clt.samp: number of clusters to be selected in the sample          #
#  dsgn:     sampling designs (e.g., pps)                             #
#  size:     MOS for the second stage                                 #
#  size.I:   MOS for the first stage (default NULL)                   #
# OUTPUT: a dataset of the sampled individuals including variables in #
#         fnt.pop and the sample weight (wt)                          #
#######################################################################
samp.slct = function(seed, fnt.pop, n, Cluster=NULL, Clt.samp=NULL,
                     dsgn, size = NULL, size.I = NULL){
  set.seed(seed); N = nrow(fnt.pop); size.Cluster = N/Cluster
  # one-ste sample design
  if(dsgn=="pps"){ fnt.pop$x=size; samp = sam.pps(fnt.pop,size, n)}
  if(dsgn == "pps-pps"){
    # MOS for the second stage
    fnt.pop$x = size
    # First stage: select clusters by pps
    index.psuI = sam.pps(matrix(1:Cluster,,1),size.I, Clt.samp)
    index.psuI = index.psuI[order(index.psuI[,1]),]  #sort selected
 psus
    sample.I = fnt.pop[fnt.pop$psu %in% index.psuI[,1],]
    sample.I$wt.I = rep(index.psuI[,'wt'],each=size.Cluster)
    # Second stage: select individuals within selected psus by pps
    samp=NULL
    for (i in 1: Clt.samp){
      popn.psu.i= sample.I[sample.I$psu==index.psuI[i,1],]
      size.II.i = sample.I[sample.I$psu==index.psuI[i,1],"x"]
      samp.i = sam.pps(popn.psu.i,size.II.i, n/Clt.samp)
      samp.i$wt = samp.i$wt*samp.i$wt.I
      samp = rbind(samp,samp.i)}}
  rownames(samp) = as.character(1:dim(samp)[1])
  return(samp)} # end FUNCTION samp.slct
######################################################################
# FUNCTION sam.pps is a subfunction of samp.slct to select a sample  #
#          under pps sampling                                        #
# INPUT:                                                             #
#  popul: the population including response and covariates           #
#  MSize: MOS                                                        #
#  n:     the sample size                                            #
# OUTPUT: dataset of the selected sample with all variables in fnt.pop
#          and sample weight (wt)                                    #
######################################################################
sam.pps<-function(popul,Msize, n){
  N=nrow(popul); pps.samID=sample(N,n,replace=F,prob=Msize)
  if (dim(popul)[2] == 1){
    sam.pps.data=as.data.frame(popul[pps.samID,])
    names(sam.pps.data) = names(popul)
  }else{sam.pps.data=popul[pps.samID,]}
  sam.pps.data$wt=sum(Msize)/n/Msize[pps.samID]
  return(sam.pps = sam.pps.data)} # End FUNCTION sam.pps

# Function rbern is to generate n random numbers following
# Bernouli(1, prob)
rbern <- function(n,prob){
  x <- runif(n,min = 0,max = 1); x.bern <- ifelse(x <= prob,1,0)
  return(x.bern)}

# Function getY is to generate disease status fnt.y in the finite
#               population
getY <- function(thetas, design.x, ICCy,Cluster, clustersize)
{
  dim(thetas) <- c(length(thetas), 1)   #make it matrix of ncol=1
  imd <- exp( design.x %*% thetas)
  N <- nrow(design.x); p <- imd/(1+imd) #Pr(y=1|D,E)

  if (ICCy==0) y <- ifelse(runif(N)<=p,1,0)
  if (ICCy!=0){
    ei0 <- rep(rnorm(Cluster),each=clustersize)
    eij <- rnorm(N); Uij <- rbern(N,sqrt(ICCy))
    thetaij <- qnorm(p); threshold <- Uij*ei0 + (1-Uij)*eij
    y <- ifelse(threshold <= thetaij,1,0)}
  return(list(y=y, py=p))}

# Kernel functions
# Triangular Density
triang = function(x){ x[abs(x)>3]=3; 1/3-abs(x)/3^2}
# Normal density with mean 0 standard deviation 3
dnorm_3 = function(x) dnorm(x, sd=3)
# Truncated standard Normal density
dnorm_t = function(x){
  c = integrate(dnorm, -3, 3)$value; y=dnorm(x)/c;
  y[y<=dnorm(3)/c]=0; y}

######################################################################
# FUNCTION JRR_var is a function to calcualte JK variances for       #
#          simulations in Chapter 3                                  #
# INPUT                                                              #
#  chtsamp: dataframe of cohort                                      #
#  svysamp: dataframe of survey sample                               #
#  svy_wt:  vector of survey sample weight                           #
#  psu:     name of psu variable in the survey or cohort sample      #
#  Formula: Fitted propensity model                                  #
#  resp:    name of response variable in the propensity model        #
#  h_in:    given bandwidth for KW method (default is NULL)          #
#  krnfun:  kernel function for KW method (default is dnorm)         #
#  rm.s:    whether removing unmatched survey sample units (default  #
#           is F)                                                    #
# OUTPUT theta - JK replicate estimates using IPSW, PSAS, and KW     #
######################################################################
JRR_var = function(chtsamp,svysamp,svy_wt, psu, Formula,
                   resp, h_in = NULL, krn = "dnorm", rm.s = F){
  krn <<- krn; rm.s <<- rm.s; h_in <<- h_in
  # Get names of the response variable and covariates for the
 propensity score estimation model
  Fml_names = all.vars(Formula)
  rsp_name = Fml_names[1]  # response variable
  mtch_var = Fml_names[-1] #covariates
  # Remove incomplete records in the cohort, if there are any.
  chtsamp_sub = as.data.frame(chtsamp[, c(mtch_var, resp)])
  if(sum(is.na(chtsamp_sub))>0){
    cmplt.indx = complete.cases(chtsamp_sub)
    chtsamp_sub = chtsamp_sub[cmplt.indx, ]
    chtsamp = chtsamp[cmplt.indx,]
    warning("Missing values in covariates are not allowed. Records with
 missing values in the cohort are removed.")}
  # Remove incomplete survey sample, if there are any.
  svysamp_sub = as.data.frame(svysamp[, c(mtch_var, resp)])
  svy_wt.vec = c(svysamp[, svy_wt])
  if(sum(is.na(svysamp_sub))>0){
    cmplt.indx = complete.cases(svysamp_sub)
    svysamp_sub = svysamp_sub[cmplt.indx, ]
    svysamp = svysamp[cmplt.indx,]
    svy_wt.vec = sum(svysamp[, svy_wt])/sum(svy_wt.vec[cmplt.indx])*
                                            svy_wt.vec[cmplt.indx]
    warning("Missing values in covariates are not allowed. Records with
 missing values in the survey sample are removed.
            The complete cases are reweighted. Missing completely at
 random is assumed.")}
  m = dim(chtsamp_sub)[1] # size of cohort (complete cases)
  n = dim(svysamp_sub)[1] # size of survey sample (complete cases)

  # number of clusters in the cohort
  uni_psu.c = unique(chtsamp[,psu]); m_psu = length(uni_psu.c)
```

```
# number of clusters in the survey sample
uni_psu.s = unique(svysamp[,psu]); n_psu = length(uni_psu.s)
# total number of clusters in the combined sample
tot_psu = m_psu + n_psu
theta_jk = matrix(0, 3, tot_psu)
rownames(theta_jk) = c("ipsw", "psas", "kw")
h_sil = h_in
for (k in 1:m_psu){
  # remove one psu at each replicate
  chtsamp.k = chtsamp_sub[chtsamp[,psu]!=uni_psu.c[k],]
  m.k = dim(chtsamp.k)[1] # size of the remainder in the cohort
  # Combine the two complete samples
  chtsamp.k[,rsp_name] = 1; svysamp_sub[,rsp_name] = 0
  names(chtsamp.k)   = c(mtch_var, resp, rsp_name)
  names(svysamp_sub) = c(mtch_var, resp, rsp_name)
  psa_dat = rbind(chtsamp.k, svysamp_sub)
  # Fit logistic regression model to predict propensity scores
  svyds = svydesign(ids =~1, weight = c(rep(m_psu/(m_psu-1), m.k),
                                        rep(1, n)),
                    data = psa_dat)
  lgtreg = svyglm(Formula, family = binomial, design = svyds)
  p_score = lgtreg$fitted.values
  # Propensity scores for the cohort
  p_score.c = p_score[psa_dat[,rsp_name]==1]
  # Propensity scores for the survey sample
  p_score.s = p_score[psa_dat[,rsp_name]==0]
  # Fit logistic regression model to predict propensity scores (with
weights)
  psa_dat$wt_cmb = c(rep(m_psu/(m_psu-1), m.k), svy_wt.vec)
  ds = svydesign(ids=~1, weight = ~ wt_cmb, data = psa_dat)
  lgtreg.w = svyglm(Formula, family = binomial, design = ds)
  p_score.w = lgtreg.w$fitted.values
  p_score.w.c = p_score.w[psa_dat[,rsp_name]==1]
  ############### Calculate replicate pseudo weights ###############
  # calculate IPSW weights
  ipsw = ipsw.wt(p_score.c = p_score.w.c, svy.wt = svy_wt.vec)
  # calculate PSAS weights
  psas = psas.wt(p_score.c = p_score.c, p_score.s = p_score.s,
                 svy.wt = svy_wt.vec, nclass = 5)$pswt
  # calculate KW weights
  kw = kw.wt(p_score.c = p_score.c, p_score.s = p_score.s,
             svy.wt = svy_wt.vec, Large=F)$pswt
  chtsamp_adj = cbind(chtsamp.k, ipsw = ipsw, psas = psas, kw = kw)
  ################# Calculate replicate estimates ##################
  theta_jk[1, k] = sum(chtsamp.k[,resp]*ipsw)/sum(ipsw) # IPSW
  theta_jk[2, k] = sum(chtsamp.k[,resp]*psas)/sum(psas) # PSAS
  theta_jk[3, k] = sum(chtsamp.k[,resp]*kw)/sum(kw)     # KW
  #print(k)}
```

```
for (k in 1:n_psu){
  # remove one psu at each replicate
  svysamp.k = svysamp_sub[svysamp[,psu]!=uni_psu.s[k],]
  # updated survey weights
  svy_wt.vec.k = svy_wt.vec[svysamp[,psu]!=uni_psu.s[k]]*
                 n_psu/(n_psu-1)
  n.k = dim(svysamp.k)[1]    # size of the remainder in the survey
  # Combine the two complete samples
  chtsamp_sub[,rsp_name] = 1; svysamp.k  [,rsp_name] = 0
  names(svysamp.k)   = c(mtch_var, resp, rsp_name)
  names(chtsamp_sub) = c(mtch_var, resp, rsp_name)
  psa_dat = rbind(svysamp.k, chtsamp_sub)
  # Fit logistic regression model to predict propensity scores
  svyds = svydesign(ids =~1, weight = c(rep(n_psu/(n_psu-1), n.k),
                                        rep(1, m)),
                    data = psa_dat)
  lgtreg = svyglm(Formula, family = binomial, design = svyds)
  p_score = lgtreg$fitted.values
  # Propensity scores for the cohort
  p_score.c = p_score[psa_dat[,rsp_name]==1]
  # Propensity scores for the survey sample
  p_score.s = p_score[psa_dat[,rsp_name]==0]
  # Fit logistic regression model to predict propensity scores (with
    weights)
  psa_dat$wt_cmb = c(svy_wt.vec.k, rep(1, m))
  ds = svydesign(ids=~1, weight = ~ wt_cmb, data = psa_dat)
  lgtreg.w = svyglm(Formula, family = binomial, design = ds)
  p_score.w = lgtreg.w$fitted.values
  p_score.w.c = p_score.w[psa_dat[,rsp_name]==1]
  #################### Calculate pseudo weights ####################
  # calculate IPSW weights
  ipsw = ipsw.wt(p_score.c = p_score.w.c, svy.wt = svy_wt.vec.k)
  # calculate PSAS weights
  psas = psas.wt(p_score.c = p_score.c, p_score.s = p_score.s,
                 svy.wt = svy_wt.vec.k, nclass = 5)$pswt
  # calculate KW weights
  kw = kw.wt(p_score.c = p_score.c, p_score.s = p_score.s,
                         svy.wt = svy_wt.vec.k, Large=F)$pswt
  ################# Calculate replicate estimates #################
  # IPSW
  theta_jk[1, (k+m_psu)] = sum(ipsw* chtsamp_sub[,resp])/sum(ipsw)
  # PSAS
  theta_jk[2, (k+m_psu)] = sum(psas* chtsamp_sub[,resp])/sum(psas)
  # KW
  theta_jk[3, (k+m_psu)] = sum(kw* chtsamp_sub[,resp])/sum(kw)
  #print(k+m_psu)
}
return(list(theta = theta_jk))}#End FUNCTION JRR_var
```

REFERENCES

Abadie, A., & Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, **84**(2), 781-807.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, **1**(2), 90-143.

Beaumont, J. F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika* **95**, 539–53

Benedetti, J. K. (1977). On the nonparametric estimation of regression functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 248-253.

Benson, K., & Hartz, A. J. (2000). A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, **342**(25), 1878-1886.

Booth, C. M., & Tannock, I. F. (2014). Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. *British Journal of Cancer*, **110**(3), 551.

Bouwmeester, W., Moons, K. G. M., Kappen, T. H., Van Klei, W. A., Twisk, J. W. R., Eijkemans, M. J. C., & Vergouwe, Y. (2013). Internal validation of risk models in clustered data: a comparison of bootstrap schemes. *American journal of epidemiology*, **177**(11), 1209-1217.

Breslow, N. E., & Day, N. E. (1980). *Statistical Methods in Cancer Research*, scientific publication No. 32. Lyon, France, International Agency for Research on Cancer, vol 1

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology* **163**(12), 1149-1156.

Chalmers, T. C., Smith Jr, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., & Ambroz, A. (1981). A method for assessing the quality of a randomized control trial. *Controlled clinical trials*, **2**(1), 31-49.

Chen, Y., Li, P., & Wu, C. (2019). Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association*, 1-11.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 295-313.

Cochran, W. G. (1977). *Sampling Techniques, Third Edition*. New York: Wiley.

Cochran, W. G., & Chambers, S. P. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A* (*General*), **128**(2), 234-266.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 417-446.

Cole, P. (1979). The evolving case-control study. *The Case-Control Study Consensus and Controversy* (pp. 15-27).

Collins, R. (2012). What makes UK Biobank special? *The Lancet* **379**(9822), 1173-1174.

Cook, E. F., & Goldman, L. (1989). Performance of tests of significance based on stratification by a multivariate confounder score or by a propensity score. *Journal of clinical epidemiology*, **42**(4), 317-324.

Couper, M., P. (2000). Web surveys: a review of issues and approaches.” *Public Opinion Quarterly* **64**(4):464–94.

Czajka, J. L., Hirabayashi, S. M., Little, R. J., & Rubin, D. B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business & Economic Statistics*, **10**(2), 117-131.

Dans, A.L., Dans, L.F., Guyatt, G.H., Richardson, S. and Evidence-Based Medicine Working Group, (1998) Users' guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. Jama, 279(**7**), pp.545-549.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, B 41, 1-31.

DiSogra, C., Cobb, C., Chan, E., & Dennis, J. M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. In *Joint Statistical Meetings (JSM), Survey Research Methods* (pp. 4501-4515).

Deville, J.C., and Särndal, C.E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, **87**(418), 376-382.

Deville J.C., Särndal, C.E., and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling, *Journal of the American Statistical Association*, **88**(423), 1013-1020.

DeStefano, F., Ford, E. S., Newman, J., Stevenson, J. M., Wetterhall, S. F., Anda, R. R, and Vinicor, F. (1993). Risk factors for coronary heart disease mortality among persons with diabetes. *Annals of Epidemiology* 3, 27-34.

Digaetano, R., & Graubard, B. I., (2003). Sampling racially matched population controls for case-control studies: using DMV lists and oversampling minorities.

Doll, R., & Hill, A. B. (1950). Smoking and carcinoma of the lung. British medical journal, **2**(4682), 739.

DuMouchel, W. H., & Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, **78**(383), 535-543.

Duncan, G. J. (2008). When to promote, and when to avoid, a population perspective. *Demography*, **45**(4), 763-784.

Dwyer-Lindgren, L., Mokdad, A. H., Srebotnjak, T., Flaxman, A. D., Hansen, G. M., and Murray, C. J. (2014). Cigarette smoking prevalence in US counties: 1996-2012. Population health metrics, 12(1), 1.

Ebrahim, S., and Davey Smith, G. (2013). Commentary: Should we always deliberately be non-representative? *International journal of epidemiology*, **42**(4), 1022-1026.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, **37**(1), 36-48.

Elliott, M. R. (2013). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, **2**(6).

Elliott, M.R., Valliant, R., Chen, J. K-T (2016). Inference for Non-probability Samples. to appear in *Statistical Science*.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. Theory of Probability & Its Applications, **14**(1), 153-158.

Escobedo, L. G., Giles, W. H., and Anda, R. F. (1997). Socioeconomic status, race, and death from coronary heart disease. *American Journal of Preventive Medicine* 13, 123-130.

Ezzati, T. M., Massey, J. T., Waksberg, J., Chu, A., & Maurer, K. R. (1992). Sample design: Third National Health and Nutrition Examination Survey. *Vital and health statistics. Series 2, Data evaluation and methods research*, (**113**), 1-35.

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, **31**(7), 799-815.

Frankel, M. R., & Frankel, L. R. (1987). Fifty years of survey sampling in the United States. *The Public Opinion Quarterly*, 51, S127-S138.

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R. and Allen, N. E. (2017). Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American journal of epidemiology*, **186**(9), 1026-1034.

Fuller, W. A. (1999). *Sampling statistics* (Vol. 560). John Wiley & Sons.

Galea, S., & Tracy, M. (2007). Participation rates in epidemiologic studies. *Annals of epidemiology*, **17**(9), 643-653.

Grimes, D. A., & Schulz, K. F. (2002). Bias and causal associations in observational research. *The Lancet*, **359**(9302), 248-252.

Gu, F., Wacholder, S., Kovalchik, S., Panagiotou, O. A., Reyes-Guzman, C., Freedman, N. D., De Matteis, S, Consonni, D, Bertazzi, P., A., Bergen, A., W., Landi, M., T., and Caporaso, N., E. (2014). Time to smoke first morning cigarette and lung cancer in a case–control study. *Journal of the National Cancer Institute*, **106**(6), dju118.

Haenszel, W., Loveland, D. B., and Sirken, Ì .G. (1962). Lung-cancer mortality as related toresidence and smoking histories. I. white males. *Journal of the National Cancer Institute*, 28,947-1001.

Hartley, H. O., Rao, J. N. K., & Kiefer, G. (1969). Variance estimation with one unit per stratum. Journal of the American Statistical Association, **64**(327), 841-851.

Heckman J. J., Hidehiko H., Todd P. (1997). Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Review of Economic Studies*, 64:605–654.

Heckman J. J., Ichimura H., Smith J., Todd P. (1998a). Characterizing selection bias using experimental data. *Econometrika*, **66**(5):1017–1098.

Heckman J. J., Ichimura H., Todd P. (1998b) Matching as an econometric evaluation estimator. *Review of Economic Studies*, 65:261–294.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). Applied survey data analysis. Chapman and Hall/CRC.

Hernan, M., A., & Robins, J., M. (2015) *Causal Inference*. New York, NY: Chapman & Hall/CRC.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161-1189.

Holland, P. W., Glymour, C., & Granger, C. (1985). Statistics and causal inference. *ETS Research Report Series*, **1985**(2).

Holt, D., & Smith, T. F. (1979). Post stratification. *Journal of the Royal Statistical Society: Series A (General)*, **142**(1), 33-46.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, **47**(260), 663-685.

Iannacchione, V. G., Milne, J. G., and Folsom, R. E. (1991). Response probability weight adjustments using logistic regression. Folsom, R. E. (1991). Response probability weight adjustments using logistic regression. American Statistical Association 1991 *Proceedings of the Section on Survey Research Methods*, 637-642.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and statistics, **86**(1), 4-29.

Johnson, C. L., Dohrmann, S. M., Burt, V. L., & Mohadjer, L. K. (2014). National health and nutrition examination survey: sample design, 2011-2014. *Vital and health statistics*. Series 2, Data evaluation and methods research, (**162**), 1.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. Journal of official

statistics, 19(2), 81.

Katki, H. A., Kovalchik, S. A., Berg, C. D., Cheung, L. C., and Chaturvedi, A. K. (2016). Development and Validation of Risk Models to Select Ever-Smokers for CT Lung Cancer Screening. *JAMA*, **315**(21), 2300-2311.

Keiding, N., & Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **179**(2), 319-376.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., and Gimenez, A. (2016). Evaluating online nonprobability surveys. Washington, DC: Pew Research Center.

Kim, J. K., & Skinner, C. J. (2013). Weighting in survey analysis under informative sampling. *Biometrika*, **100**(2), 385-398.

Kish, L. (1985). Survey sampling. 1965. New Yory: Wiley Pty Ltd.

Korn, E. L., & Graubard, B. I. (1999). Analysis of health surveys (Vol. 323). John Wiley & Sons.

Kott, P. S. (2006), Using calibration weighting to adjust for nonresponse and coverage errors, *Survey Methodology*, **32** (2), 133–142.

Krewski, D., & Rao, J. N. K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 1010-1019.

Kupper, L. L., McMichael, A. J., & Spirtas, R. (1975). A hybrid epidemiologic study design useful in estimating relative risk. *Journal of the American Statistical Association*, **70**(351a), 524-528.

Landsman, V., & Graubard, B. I. (2012). Efficient analysis of case-control studies with sample weights. *Statistics in medicine* **32**(2), 347-360.

Lane-Claypon, J. E. (1926). A Further Report on Cancer of the Breast with Special Reference to its Associated Antecedent Conditions.

Last, J. M., Abramson, J. H., & Freidman, G. D. (Eds.). (2001). A dictionary of epidemiology (Vol. 4). New York: Oxford University Press.

LaVange, L. M., Koch, G. G., and Schwartz, T. A. (2001). Applying sample survey methods to clinical trials data. *Statistics in Medicine*, **20**(17-18), 2609-2623.

Lavori, P. W., & Keller, M. B. (1988). Improving the aggregate performance of psychiatric diagnostic methods when not all subjects receive the standard test. *Statistics in Medicine*, **7**(7), 727-737.

Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PloS one*, **6**(3), e18174.

Lee, S. (2004). Statistical estimation methods in volunteer panel web surveys. Ph.D. dissertation, Joint Program in Survey Methodology, University of Maryland.

Lee, S. (2006). Propensity Score Adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics* 22:329-49.

Lee, S., & Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, **37**(3), 319-343.

Li, Y., Graubard, B. I., & DiGaetano, R. (2011). Weighting methods for population‐based case–control studies with complex sampling. *Journal of the Royal Statistical Society: Series C* (*Applied Statistics*), **60**(2), 165-185.

Little, R., J. (2010). Discussion of Articles on the Design of the National Children's Study. *Statistics in Medicine*, **29**(13), 1388-1390.

Little, R. J., & Vartivarian, S. (2003). On weighting the rates in non‐response weights. *Statistics in medicine*, **22**(9), 1589-1599.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, **23**(19), 2937-2960.

Lunn, A. D., & Davies, S. J. (1998). A note on generating correlated binary variables. *Biometrika*, **85**(2), 487-490.

Lumley, T., 2020. Package 'Survey.' Available at the following link: (http://cran.r-project.org/web/packages/survey/survey.pdf).

Michael, R. T., and O'Muircheartaigh, C. A. (2008). Design priorities and disciplinary perspectives: the case of the US National Children's Study. *Journal of the Royal Statistical Society: Series A* (*Statistics in Society*), **171**(2), 465-480.

Morabia, A. (2013). A History of Epidemiologic Methods and Concepts. Birkhäuser.

Morton, L. M., Cahill, J., & Hartge, P. (2006). Reporting participation in epidemiologic studies: a survey of practice. *American journal of epidemiology*, **163**(3), 197-203.

Muscat, J. E., Ahn, K., Richie, J. P., and Stellman, S. D. (2011). Nicotine dependence phenotype and lung cancer risk. *Cancer*, **117**(23), 5370-5376.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, **9**(1), 141-142.

National Cancer Institute, NCI Dictionary of Cancer Terms, 17.03d April 7, 2017, Bethesda, MD

https://www.cancer.gov/publications/dictionaries/cancer-terms

National Center for Health Statistics. National Death Index user’s guide. Hyattsville, MD. 2013. (Available at the following address:

https://www.cdc.gov/nchs/data/ndi/ndi_users_guide.pdf)

National Center for Health Statistics. Office of Analysis and Epidemiology, The National Health Interview Survey (1986-2004) Linked Mortality Files, mortality follow-up through 2006: Matching Methodology, May 2009. Hyattsville, Maryland. (Available at the following address:

http://www.cdc.gov/nchs/data/datalinkage/matching_methodology_nhis_final.pdf)

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

Nohr, E. A., Frydenberg, M., Henriksen, T. B., and Olsen, J. (2006). Does low participation in cohort studies induce bias?. *Epidemiology*, **17**(4), 413-418.

Nordberg, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics*, **5**(3), 223.

Oman, S. D., & Zucker, D. M. (2001). Modelling and generating correlated binary variables. *Biometrika*, **88**(1), 287-290.

Owen, A. (1987). Nonparametric conditional estimation (No. SLAC-309).

Pappas, G., Queen, S., Hadden, W., & Fisher, G. (1993). The increasing disparity in mortality between socioeconomic groups in the United States, 1960 and 1986. *New England journal of medicine*, **329**(2), 103-109.

Pfeffermann, D. & Sverchkov, M. Y. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya B* 61, 166–86.

Pfeffermann, D. & Sverchkov, M. Y. (2003). Fitting Generalized Linear Models under Informative Sampling. *In Analysis of Survey Data*, Ed. R. L. Chambers & C. J. Skinner. Chichester: Wiley.

Pinsky, P. F. (2006). Racial and ethnic differences in lung cancer incidence: how much is explained by differences in smoking patterns? (United States). *Cancer Causes & Control*, **17**(8), 1017-1024.

Pinsky, P. F., Miller, A., Kramer, B. S., Church, T., Reding, D., Prorok, P., Gelmann, E., Schoen, R. E., Buys, S., Hayes, R. B., and Berg, C. D. (2007). Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. *American journal of epidemiology* **165**(8), 874-881.

Pizzi, C., De Stavola, B., Merletti, F., Bellocco, R., dos Santos Silva, I., Pearce, N., and Richiardi, L. (2011). Sample selection and validity of exposure–disease association estimates in cohort studies. *Journal of Epidemiology & Community Health*, 65(**5**), 407-411.

Potter, F. J. (1993). The effect of weight trimming on nonlinear survey estimates. *In Proceedings of the American Statistical Association, Section on Survey Research Methods* (Vol. 758763).

Potter, F., & Zheng, Y. (2015). Methods and issues in trimming extreme weights in sample surveys. *In Proceedings of the American Statistical Association, Section on Survey Research Method*s (pp. 2707-2719).

Powers, S., McGuire, V., Bernstein, L., Canchola, A. J., and Whittemore, A. S. (2017)

Evaluating disease prediction models using a cohort whose covariate distribution differs from that of the target population. *Statistical methods in medical research,* 1:962280217723945 doi: 10.1177/0962280217723945

Prorok, P. C., Andriole, G. L., Bresalier, R. S., Buys, S. S., Chia, D., Crawford, E. D., ... & Hayes, R. B. (2000). Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. Contemporary Clinical Trials, 21(**6**), 273S-309S.

Richiardi, L., Pizzi, C., and Pearce, N. (2013). Commentary: Representativeness is usually not necessary and often should be avoided. *International journal of epidemiology*, **42**(4), 1018-1022.

Rivers, D. (2007). Sampling for Web Surveys. White paper prepared from presentation given at the 2007 Joint Statistical Meetings, Salt Lake City, Utah, July-August.

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

Rogers, R. G., & Powell-Griner, E. (1991). Life expectancies of cigarette smokers and nonsmokers in the United States. *Social science & medicine*, **32**(10), 1151-1159.

Rogers, R. G. (1992). Living and dying in the USA: sociodemographic determinants of death among blacks and whites. *Demography*, **29**(2), 287-303.

Rosenbaum, P.R. (1984a). From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment. *Journal of the American Statistical Association*, **79**(385), 41-48.

Rosenbaum, P.R. (1984b). The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment, *Journal of the Royal Statistical Society, Series A* (*General*), **147**(5), 656-666.

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

Rosenbaum, P.R., and Rubin, D.B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, **79**(387), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, **39**(1), 33-38.

Rubin, D. B. (1973). Matching to remove bias in observational studies. Biometrics, 159-183.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, **66**(5), 688.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74 (**336**), 318–328

Rubin, D.B. (1997). Estimation from Nonrandomized Treatment Comparisons Using Subclassification on Propensity Scores. *Annals of Internal Medicine*, 127, **8**(2), 757-763.

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

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 249-264.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology* **2**(3), 169-188.

Sanderson, M., & Gonzalez, J. F. (1998). 1988 National Maternal and Infant Health Survey: methods and response characteristics. *Vital and health statistics*. Series 2, Data evaluation and methods research, (125), 1.

Schonlau, M., Zapert, K., Simon L.P., Sanstad, K., Marcus, S., Adams, J., Spranca, M., Kan, H., Turner, R., and Berry, S (2004). A comparison between responses from a propensity-weighted web survey and an identical RDD survey. *Social science computer review*, **22**(1), 128-138.

Scott, A. J., & Wild, C. J. (1986). Fitting logistic models under case-control or choice based sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, **48**(2), 170-182.

Scott, A., & Wild, C. (2001). Case–control studies with complex sampling. *Journal of the Royal Statistical Society: Series C* (*Applied Statistics*)*,* 50(3), 389-401.

Scott, A., & Wild, C. (2009). Population-based case-control studies. In Handbook of Statistics (Vol. 29, pp. 431-453). Elsevier.

Scott, D. W. (1992), Multivariate Density Estimation: Theory, Practice, and Visualization, New York: John Wiley

Scott, D. W., & Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical association* **82**(400), 1131-1146.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* 683-690.

Silverman, B. W. (1986). Density estimation for statistics and data analysis (Vol. 26). CRC press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* **25**(1), 1.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174**(2), 369-386.

Shadish WR, Clark M, Steiner PM. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*;**103**(484):1334–1344.

Stone, R.A., Oborsky, S., Singer, D.E., Kapoor, W.N., and Fine, M.J. (1995). Propensity Score Adjustment for Pretreatment Differences between Hospitalized and Ambulatory Patients with Community-Acquired Pneumonia. *Medical Care,* 33, AS56-66.

Stürmer, T., Rothman, K. J., Avorn, J., & Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. American journal of epidemiology, **172**(7), 843-854.

Taylor, J. M., Ankerst, D. P., & Andridge, R. R. (2008). Validation of biomarker-based risk prediction models. *Clinical Cancer Research*, **14**(19), 5977-5983.

Terhanian, G. and J. Bremer. (2000). Confronting the selection-bias and learning effects problems associated with Internet research. White paper, Harris Interactive, Rochester, NY.

Terrell, G. R., & Scott, D. W. (1985). Oversmoothed nonparametric density estimates. *Journal of the American Statistical Association* **80**(389), 209-214.

The National Death Index. Hyattsville, MD: Division of Vital Statistics, National Center for Health Statistics, 2007. (http:// www.cdc.gov/nchs/ndi.htm).

Tota, J. E., Gillison, M. L., Katki, H. A., Kahle, L., Pickard, R. K., Xiao, W., ... & Chaturvedi, A. K. (2019). Development and validation of an individualized risk prediction model for oropharynx cancer in the US population. *Cancer*, **125**(24), 4407-4416.

Valliant, R., & Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research* **40**(1), 105-137.

Weinkam, J. J., Rosenbaum, W. L., and Sterling, T. D. (1992). Computation of relative risk based on simultaneous surveys: an alternative to cohort and case-control studies.

American Journal of Epidemiology 136, 722-729.

Waston, G. S. (1964). Smooth regression analysis. *Sankhya Ser. A* 26:359–372

Wolter, K. (1985). Introduction to variance estimation. Springer Science & Business Media.