Web-Based Supporting Materials for Improving External Validity of Epidemiologic Cohort Analyses: A Kernel Weighting Approach

Lingxiao Wang,[1, 2] Barry I. Graubard,[2] Hormuzd A. Katki,[2,*] and Yan Li[1,*]

[1]The Joint Program in Survey Methodology, University of Maryland, College Park, U.S.A.

[2]National Cancer Institute, Division of Cancer Epidemiology & Genetics, Biostatistics Branch, U.S.A.

*These co-corresponding authors contributed equally:

HAK: 9609 Medical Center Dr., Room 7e592, Rockville MD 20850. katkih@mail.nih.gov

YL: 1218 Lefrak Hall, 7521 Preinkert Dr, College Park, MD 20742 yli6@umd.edu

Web Appendix A. Consistency of the Kernel Weighted Mean

Following the notation in Section 2 in the main text, we compute k_{ij} given by

$$k_{ij} = \frac{K\left(\frac{p_i - p_j}{h}\right)}{\sum_{j \in s_c} K\left(\frac{p_i - p_j}{h}\right)}, \qquad i \in s_s, \qquad j \in s_c$$

where$K(\cdot)$ is a kernel function, h is the bandwidth, p_i and p_j are arguments of function k_{ij} for the propensity scores that are estimated by $\hat{p}\left(\boldsymbol{x}_i^{(s)}, \widehat{\boldsymbol{\beta}}\right)$ and $\hat{p}\left(\boldsymbol{x}_j^{(c)}, \widehat{\boldsymbol{\beta}}\right)$, with the superscripts (s) and (c) denoting the survey sample and the cohort, respectively. In addition, it follows that $\sum_{j \in s_c} k_{ij} = 1$. The KW pseudo weight for cohort unit j is

$$w_j^{KW} = \sum\nolimits_{i \in s_s} k_{ij} \cdot d_i,$$

where d_i is the sample weight of the survey sample unit i, and $\widehat{N} = \sum_{i \in s_s} d_i$ is an unbiased estimator of the finite population size N. The cohort KW estimator of the population mean ($\bar{Y} = N^{-1} \sum_{k=1}^{N} y_k$) is given by

$$\hat{\bar{Y}}^{KW} = \frac{1}{\widehat{N}^{KW}} \sum\nolimits_{j \in s_c} w_j^{KW} \cdot y_j,$$

where $\widehat{N}^{KW} = \sum_{j \in s_c} w_j^{KW}$. Notice that

$$\widehat{N}^{KW} = \widehat{N} \tag{A.1}$$

because $\sum_{j \in s_c} w_j^{KW} = \sum_{j \in s_c} \sum_{i \in s_s} \left(k_{ij} \cdot d_i\right) = \sum_{i \in s_s} \left(d_i \cdot \sum_{j \in s_c} k_{ij}\right) = \sum_{i \in s_s} d_i$.

Theorem 1.

Suppose in the superpopulation the variable of interest y has an expectation $E(y) = \mu$, where E denotes the expectation with respect to the joint distribution of y and covariates $\boldsymbol{x}$. Assume that the cohort and 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, $E\{y|p(\boldsymbol{x}), \text{cohort}\} = E\{y|p(\boldsymbol{x}), \text{survey}\} = E\{y|p(\boldsymbol{x})\}$;

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

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

then the KW estimator of the population mean $\hat{\bar{Y}}^{KW} = \frac{\sum_{j\in s_c} w_j^{KW} \cdot y_j}{\sum_{j\in s_c} w_j^{KW}} \to \mu$ 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 $\frac{n_c}{N} = O(1)$.

Proof. 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 the random vector (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 $(y_1, p_1), \cdots (y_{n_c}, p_{n_c})$ and survey sample $(y_1, p_1), \cdots (y_{n_s}, p_{n_s})$ 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

$$E\{|y^* - E(y|p)|\} \to 0, \tag{A.2}$$

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 (A.2),

$$E(|y_k^* - \mu_k|) \to 0, \tag{A.3}$$

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 (A.3). By Law of Large Numbers, $\bar{\mu}_k \xrightarrow{P} \mu$. Therefore,

$$E|\bar{Y}^* - \mu| \to 0, \tag{A.4}$$

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

$$\bar{Y}^* - \mu \xrightarrow{P} 0. \tag{A.5}$$

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

$$\hat{\bar{Y}}^* = \bar{Y}^* + O_p\left(n_s^{-\frac{1}{2}}\right), \tag{A.6}$$

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

By (A.5) and (A.6),

$$\hat{\bar{Y}}^* \xrightarrow{P} \mu. \tag{A.7}$$

By the Law of Large Numbers,

$$\bar{Y} = \mu + O_p\left(N^{-\frac{1}{2}}\right). \tag{A.8}$$

(A.7) and (A.8) together implies

$$\hat{\bar{Y}}^* - \bar{Y} \xrightarrow{P} 0. \quad \text{(A.9)}$$

(A.4) and (A.6) together implies

$$E\left(\hat{\bar{Y}}^*\right) \to \mu, \quad \text{(A.10)}$$

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

Notice that by applying (A.1), the KW estimator of finite population mean $\hat{\bar{Y}}^{KW} = \hat{\bar{Y}}^*$ because

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

Web Appendix B. Jackknife Variance Estimation for Pseudo-Weighted Estimators

Suppose that the survey sample 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., PSU's) 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 re-estimated 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 PSU's in l to the number of remaining PSU's, 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 propensity score for each unit in the replicate-$l\alpha$ sample.

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

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

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 j in replicate-$l\alpha$ is

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

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

$$\hat{\bar{Y}}_{(l\alpha)}^{KW} = \left(\sum\nolimits_{j \in s_{c(l\alpha)}} w_{j(l\alpha)}^{KW}\right)^{-1} \cdot \sum\nolimits_{j \in s_{c(l\alpha)}} w_{j(l\alpha)}^{KW} \cdot y_j.$$

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

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

The PSAS and IPSW jackknife variance estimates 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 $f_{i(l\alpha)}d_i$ for each survey unit i. 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., $\sum_{i\in sub_g} f_{i(l\alpha)}d_i$, 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.

Web Appendix C. Simulations: 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 $o_r^{(\alpha)}, r = 1, \cdots, 4$ 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 $Env \sim \min\{4.5, \text{LogNormal}(\mu_a,\ 0.5)\}$, where $\mu_\alpha \sim \text{Uniform}(0, 0.5)$, 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 $\mu = \text{expit}(\boldsymbol{\gamma v})$ (Hunsberger et al., 2008; Oman & Zucker, 2001). The parameter $\boldsymbol{\gamma} =$

$(-5, 0.5, -1, 1, 0.3, 0.10)^T$ 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 was generated by $z = \mu + e$, with $e \sim \text{Normal}(0, 0.085^2)$ 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 $\rho = 0.30$.

Web Appendix D. Simulations: True Propensity-Score Models Fitted to Weighted and Unweighted Sample under Two-Stage Cluster Sampling

D.1 Description of Sampling Designs in the Simulation

The cohort and survey sample were randomly selected from the finite population by two-stage cluster sampling with each stage using a probability proportional to size (PPS) sampling. The measures of size (MOS) of the PPS sampling for the cohort and survey sample selection were functions of $q_k^a = \exp\{a \cdot (\beta_0 + \boldsymbol{\beta}\boldsymbol{x}_k)\}$ and $q_k^b = \exp\{b \cdot (\beta_0 + \boldsymbol{\beta}\boldsymbol{x}_k)\}$, respectively, for population unit k, where $\boldsymbol{x}_k$ is a vector of covariates, $(\beta_0, \boldsymbol{\beta})$ is the vector of parameters, a and b are two real numbers (described later). Web Table 1 below describes the two-stage PPS cluster sampling.

Web Table 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	$\sum_{k \in u_\alpha} q_k^a$	$p_k^{(c)} = \frac{n_c \cdot q_k^a}{\sum_{k=1}^N q_k^a}$
	Stage 2- subjects selected by PPS	$q_k^a, k \in u_\alpha$	
Survey	Stage 1- clusters selected by PPS	$\sum_{k \in u_\alpha} q_k^b$	$p_k^{(s)} = \frac{n_s \cdot q_k^b}{\sum_{k=1}^N q_k^b}$
	Stage 2- subjects selected by PPS	$q_k^b, k \in u_\alpha$	

In the table, n_c and n_s are the size of cohort and survey sample, respectively; u_α is the set of individuals from α-th cluster ($\alpha = 1, \cdots, M$), N is the population size; 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 Web Table 1, when $a = -1$ and $b = 1$, population units with larger values of q_k^{-1} and q_k tend to be oversampled in the cohort and survey, respectively. The larger $|a - b|$ is, the more different the covariate distributions are between the cohort and the survey.

D.2 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 p_k as the probability of being self-selected in the cohort for $k \in FP$ given it has been selected into the combined sample of cohort and survey sample, given by

$$p_k = \Pr\{k \in s_c | k \in s_c \cup^* s_s\} = \frac{p_k^{(c)}}{p_k^{(c)} + p_k^{(s)}}, \qquad \text{for } k \in FP.$$

where $p_k^{(c)}$ and $p_k^{(s)}$ (defined in Web Table 1) are the inclusion probabilities of cohort and survey sample respectively for $k \in FP$. The notation $\cup^*$ represents the combination of the two samples that allows population units to be selected in both cohort and survey. The overlap of s_c and s_s will be counted twice in the combined sample $s_c \cup^* s_s$.

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

$$\log\left(\frac{p_k}{1-p_k}\right) = \log\left\{\frac{p_k^{(c)}/\left(p_k^{(c)}+p_k^{(s)}\right)}{p_k^{(s)}/\left(p_k^{(c)}+p_k^{(s)}\right)}\right\} = \log\left\{\frac{p_k^{(c)}}{p_k^{(s)}}\right\}.$$

Since $p_k^{(c)} = \frac{n_c \cdot q_k^a}{\sum_{k=1}^N q_k^a}$, and $p_k^{(s)} = \frac{n_s \cdot q_k^b}{\sum_{k=1}^N q_k^b}$ as defined in Web Table 1, we have

$$\log\left(\frac{p_k}{1-p_k}\right) = \log\left(\frac{n_c \cdot q_k^a / \sum_{k=1}^N q_k^a}{n_s \cdot q_k^b / \sum_{k=1}^N q_k^b}\right)$$

$$= \log\left(\frac{n_c \cdot \sum_{k=1}^N q_k^b}{n_s \cdot \sum_{k=1}^N q_k^a}\right) + \log\left(\frac{q_k^a}{q_k^b}\right),$$

Because $q_k^a = \exp\{a \cdot (\beta_0 + \boldsymbol{\beta x}_k)\}$ and $q_k^b = \exp\{b \cdot (\beta_0 + \boldsymbol{\beta x}_k)\}$, we have

$$\log\left(\frac{p_k}{1-p_k}\right) = \omega + (a-b) \cdot \boldsymbol{\beta x}_k, \qquad \text{(D.1)}$$

where $\omega = \log\left(\frac{n_c \cdot \sum_{k=1}^N q_k^b}{n_s \cdot \sum_{k=1}^N q_k^a}\right) + (a-b) \cdot \beta_0$ is the intercept. Note that the vector of model coefficients $(a-b) \cdot \boldsymbol{\beta}$ can be estimated by fitting the propensity model (D.1) to the sample of cohort and *unweighted* survey. The predicted p_k, i.e., $\hat{p}_k = \text{expit}\left(\hat{\omega} + (a-b) \cdot \hat{\boldsymbol{\beta}}\boldsymbol{x}\right)$ is used by the PSAS and KW methods to measure the similarity between cohort units and survey units.

D.3 True Propensity-Score Model assumed by the IPSW method

Define p_k^*, for $k \in FP$, the probability of being in the cohort conditional on the sample of cohort combined with the finite population (*FP*), written as

$$p_k^* = \Pr\{k \in s_c | k \in s_c \cup^* FP\} = \frac{p_k^{(c)}}{p_k^{(c)}+1}, \text{for } k \in FP.$$

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,

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

$$= \log\left(\frac{n_c}{\sum_{k=1}^{N} q_k^a}\right) + \log(q_k^a).$$

Since $q_k^a = \exp\{a \cdot (\beta_0 + \boldsymbol{\beta}\boldsymbol{x}_k)\}$, we have

$$\log\left(\frac{p_k^*}{1-p_k^*}\right) = \omega^* + a \cdot \boldsymbol{\beta}\boldsymbol{x}_k, \tag{D.2}$$

where $\omega^* = \log\left(\frac{n_c}{\sum_{k=1}^{N} q_k^a}\right) + a \cdot \beta_0$ is the intercept. Note that model coefficient $a \cdot \boldsymbol{\beta}$ can be estimated by fitting the propensity model (D.2) to the combined cohort and *weighted* survey sample. The predicted odds, i.e. $\frac{\hat{p}_k^*}{1-\hat{p}_k^*} = \exp\left(\hat{\omega}^* + a \cdot \hat{\boldsymbol{\beta}}_w \boldsymbol{x}\right)$ is used by the IPSW method to estimate the self-selection probability for cohort units.

D.4 Simulation Results

In this simulation, we empirically verify D.2 and D.3. The MOS for cohort and survey sample selection were q_k^a and q_k^b respectively, where $q_k = \exp(\beta_0 + \beta_1 age_k + \beta_2 hh\_inc_k + \beta_3 Env_k + \beta_4 z_k)$ with $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = (0, 0.3, -0.4, 0.7, 0.7)$, $a = -1$, and $b = 0.5$. As derived in D.2 and D.3, $\hat{\boldsymbol{\beta}}_w$, and $\hat{\boldsymbol{\beta}}$, the regression coefficients estimated from the propensity models fitted to the weighted and unweighted combination of the cohort and survey sample are approximate unbiased estimators of $a \cdot \boldsymbol{\beta} = -(\beta_1, \beta_2, \beta_3, \beta_4)$ and $(a-b) \cdot \boldsymbol{\beta} = -1.5(\beta_1, \beta_2, \beta_3, \beta_4)$, respectively. The percent of relative bias relative bias (%RB), empirical variance (V) of $\hat{\boldsymbol{\beta}}_w$, and $\hat{\boldsymbol{\beta}}$ are shown in Web Table 2.

Web Table 2
Results of coefficients of propensity model fitted to unweighted and weighted combined cohort and survey sample over 1,000 simulation runs

	$\beta_1 = 0.3$		$\beta_2 = -0.4$		$\beta_3 = 0.7$		$\beta_4 = 0.7$	
	%RB	**V**(10^{-3})	**%RB**	**V**(10^{-3})	**%RB**	**V**(10^{-3})	**%RB**	**V**(10^{-3})
$\widehat{\boldsymbol{\beta}}$	-2.2%	0.71	-1.7%	1.98	-0.6%	8.61	0.6%	93.02
$\widehat{\boldsymbol{\beta}}_w$	3.3%	1.03	-2.5%	3.53	2.9%	15.43	1.4%	189.29

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

Web Appendix E. 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

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

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 (E.1), 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 $(-t, t)$, $T(-t, t, 0)$.

E.1 $\underline{N(0, \sigma_K)}$

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

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

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 (E.2) 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 bandwidth $h_{\text{silverman}} = 0.9 \cdot \min\left(\hat{\sigma}, \frac{IQR}{1.34}\right) \cdot n^{-1/5}$, and $h_{\text{scott}} = 1.06 \cdot \min\left(\hat{\sigma}, \frac{IQR}{1.34}\right) \cdot n^{-1/5}$ respectively.

E.2 $\underline{T(-t,t,0)}$

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, E.1 and E.2, the optimal bandwidth h_{opt} changes with the value of the scale parameter of the kernel density function. However, the corresponding kernel density $K\left(d\left(\boldsymbol{x}_i^{(s)}, \boldsymbol{x}_j^{(c)}\right)/h_{opt}\right)$ remains invariant to changes in the value of the scale parameter, which results in the kernel weight (defined in (2.2.1) in the main text) being also unaffected by scale parameter.

Web Appendix F. Supplementary Tables

Web Table 3
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	Method	%RB	V ($\times 10^{-5}$)	VR (TL)	VR (JK)	CP (JK)	MSE_g ($\times 10^{-5}$)
	CHT	-42.48	2.39	0.19	NA	NA	168.45
	SVY	-0.11	6.42	1.06	1.06	0.96	6.42
T							
True model	$\text{logit}\{\Pr(\boldsymbol{x})\} \sim age, hh\_inc, Env, z$						
	IPSW	-0.19	7.54	0.76	0.99	0.95	7.54
	PSAS	-9.37	5.05	0.93	1.03	0.71	13.12
	KW (h=0.00943)	-0.91	6.00	0.95	1.02	0.94	6.07
U1							
Underfitted model 1	$\text{logit}\{\Pr(\boldsymbol{x})\} \sim age, Env, z$						
	IPSW	-0.36	6.67	0.78	0.95	0.94	6.68
	PSAS	-9.40	4.90	0.92	1.02	0.70	13.02
	KW (h =0.00990)	-1.43	5.58	0.93	0.98	0.93	5.76
U2							
Underfitted model 2	$\text{logit}\{\Pr(\boldsymbol{x})\} \sim age, Env$						
	IPSW	-5.10	5.71	0.84	0.96	0.85	8.10
	PSAS	-10.85	4.88	0.90	1.03	0.65	15.69
	KW (h =0.01085)	-2.68	5.49	0.92	0.99	0.91	6.14
M							
Underfitted +	$\text{logit}\{\Pr(\boldsymbol{x})\} \sim age, Env, Hisp, sex$						
Overfitted model	IPSW	-4.59	5.96	0.81	0.96	0.87	7.89
	PSAS	-9.23	4.91	0.92	1.07	0.73	12.74
	KW (h =0.01014)	-1.72	5.54	0.92	1.01	0.93	5.81
O1							
Overfitted model 1	$\text{logit}\{\Pr(\boldsymbol{x})\} \sim age, hh\_inc, Env, z, Hisp$						
	IPSW	-0.02	7.66	0.75	0.99	0.95	7.65
	PSAS	-9.31	5.01	0.94	1.05	0.71	12.98
	KW (h =0.00942)	-0.76	6.01	0.96	1.04	0.95	6.06
O2							
Overfitted model 2	$\text{logit}\{\Pr(\boldsymbol{x})\} \sim age, hh\_inc, Env, z, Hisp, sex$						
	IPSW	0.13	7.82	0.74	0.99	0.95	7.81
	PSAS	-9.30	5.02	0.93	1.06	0.71	12.97
	KW (h =0.00941)	-0.71	6.03	0.96	1.05	0.95	6.07
O3							
Overfitted model 3	$\text{logit}\{\Pr(\boldsymbol{x})\} \sim age, hh\_inc, Env, z, urb$						
	IPSW	-0.10	7.66	0.75	0.99	0.95	7.65
	PSAS	-9.22	5.02	0.94	1.09	0.73	12.84
	KW (h =0.00938)	-0.79	5.96	0.97	1.08	0.95	6.01

Web Table 4

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

Kernel Function	**Bandwidth** (Method)	**%RB**	**V** ($\times 10^{-5}$)	**VR** (TL)	**VR** (JK)	**CP** JK	**MSE**$_i$ ($\times 10^{-5}$)
	CHT	-42.48	2.39	0.19	--	--	168.45
	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 logit{Pr($\boldsymbol{x}$)} ~ $age, hh\_inc, Env, z, Hisp, sex$, 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 Web Table 3 under Model O2 because the bandwidths were different.

Web Table 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

Web Table 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

Web Table 7

Main effects of the fitted propensity model with ($\widehat{\boldsymbol{\beta}}_w$) or without ($\widehat{\boldsymbol{\beta}}$) NHIS sample weights†

	$\widehat{\boldsymbol{\beta}}$		$\widehat{\boldsymbol{\beta}}_w$	
Coefficients:	Estimates	Std. Err.	Estimates	Std. Err.
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.

Web Appendix G. REFERENCE

Hajek, J. (1971) Comment on "An essay on the logical foundations of survey sampling" by Basu, D. in Part 1, *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston, New York.

Hunsberger, S., Graubard, B. I., & Korn, E. L. (2008). Testing logistic regression coefficients with clustered data and few positive outcomes. *Statistics in medicine*, **27**(8), 1305-1324.

Isaki, C. T., & Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, **77**(377), 89-96.

Oman, S. D., & Zucker, D. M. (2001). Modelling and generating correlated binary variables. *Biometrika*, **88**(1), 287-290.

Noda, K. (1976). Estimation of a regression function by the Parzen kernel-type density estimators. *Annals of the Institute of Statistical Mathematics*, **28**(1), 221-234.

Scott, D. W. (1992). The curse of dimensionality and dimension reduction. *Multivariate Density Estimation: Theory, Practice, and Visualization, Second Edition*, 217-240. John Wiley & Sons

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.

Sheather, S. J. (2004). Density estimation. *Statistical Science* **19**(4), 588-597.

Silverman, B. W. (1986). Density estimation for statistics and data analysis (Vol. 26). CRC press.