Propensity Score Weighting With Survey Weighted Data When Outcomes Are Binary

Chen Yang

Institute for Healthcare Delivery Science, Icahn School of Medicine at Mount Sinai

Meaghan S. Cuerden

London Health Sciences Centre

Jiayi Ji

Department of Biostatistics & Epidemiology, Rutgers School of Public Health

Melissa Aldridge

Brookdale Department of Geriatrics and Palliative Medicine, Icahn School of Medicine at Mount Sinai

Lihua Li (✉ lihua.li@mountsinai.org)

Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai

Research Article

Keywords: Survey data, propensity score weighting, provider-patient discussion, smoking cessation

Posted Date: February 24th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1374473/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Propensity score weighting with survey weighted data when outcomes are binary

Chen Yang^{1,2}, Meaghan S. Cuerden³, Jiayi Ji⁴, Melissa Aldridge⁵, Lihua Li ^{1,2,5}

1. Institute for Healthcare Delivery Science, Icahn School of Medicine at Mount Sinai, New York, New York, USA

2. Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, New York, USA

3. London Health Sciences Centre, London, Ontario, Canada

4. Department of Biostatistics & Epidemiology, Rutgers School of Public Health, Piscataway, NJ, USA

5. Brookdale Department of Geriatrics and Palliative Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Corresponding Author

Lihua Li, PhD

Department of Population Health Science and Policy

Icahn School of Medicine at Mount Sinai

Mount Sinai Hospital

1425 Madison Ave

New York, NY 10029

Phone: 212-659-9663

Email: lihua.li@mountsinai.org

Word count (abstract): 347

Word count (main text): 4,490

Abstract:

Background

Despite that propensity score methods have been widely adopted in observational studies, few studies have focused on how to properly estimate and incorporate propensity scores in complex survey data settings. Particularly, it is lacking that the research on propensity score-based weighting methods when estimating the treatment effect using complex survey data with binary outcomes. The objective of this study is therefore to compare three propensity score weighting methods for complex survey data when outcomes are binary.

Methods

A simulation study was used to compare three propensity score weighting approaches for estimating treatment effects using survey data: 1) no survey weights in the propensity score model or the outcome model; 2) survey weights in the outcome model only; 3) survey weights in both models. Each of the three methods is applied in the context of four measures of the treatment effect: the sample average treatment effect (SATE), the population average treatment effect (PATE), the sample average treatment effect on the treated (SATT), and the population average treatment effect on the treated (PATT). The methods are compared in terms of mean relative bias and coverage probability under different scenarios by varying combinations of sample size and treatment effect, degrees of model misspecification and levels of overlap in propensity score. In addition, using the 2015 National Health Interview Survey (NHIS) data as a real data example, each method is employed to estimate the effect of provider-patient discussion about smoking on smoking cessation.

Results

The methods which account for the survey weight outperform the unweighted method as the degree of misspecification increases, regardless of the sample size, treatment effect and level of overlap in propensity score. The performance of the two weighted methods, where the survey weight is incorporated in the outcome model or both the propensity score model and the outcome model, are similar.

Conclusions

The propensity score weighting methods accounting for the survey weight are necessary for estimating the population-level treatment effects. This paper provides guidance for selecting an appropriate propensity score weighting method based on the estimand of interest when using complex survey data with binary outcomes.

Key Words

Survey data, propensity score weighting, provider-patient discussion, smoking cessation

Background

Observational studies are increasingly used to assess the effect of a particular treatment or exposure, with the advantages of relatively lower cost and fewer ethical issues compared with randomized controlled trials.^{1, 2} However, the fundamental challenge is the presence of confounding, which may lead to biased estimates of the treatment effect. Confounding occurs when there are systematic differences in baseline characteristics between treated and control subjects. Propensity score methods are widely accepted in estimation of treatment or exposure effects to address confounding in observational studies.³⁻⁶ The propensity score can be used in one of four ways: matching on the propensity score, propensity score-based weighting, stratifying by the propensity score, and including the propensity score as a covariate.⁷⁻¹⁰ Propensity score methods ensure balance between two treatment groups on observed factors. The estimands of interest often include the average treatment effect (ATE) and the average treatment effect on the treated (ATT).

Although propensity score methods have been widely researched and adopted in observational studies, few studies have focused on how to properly estimate and incorporate propensity scores in complex survey data settings. Zanutto (2006) and Zanutto, Lu, and Hornik (2005) emphasized the importance of incorporating the survey weight and proposed modifications to accommodate complex survey data.^{11, 12} More recently, Dugoff et al. (2014) and Ridgeway et al. (2015) provided practical guidance in employing different propensity score methods to survey data through the comparison of models with various specification of weights in the context of a continuous outcome with limited settings^{13, 14} Austin et al. (2018) and Lenis et al. (2019)

explored the use of survey weights in the estimation of the propensity score using complex survey data, where the focus was on the use of propensity score matching in the estimation of the ATT.^{15, 16}

What is lacking in the propensity score research is guidance on estimating the treatment effect in the context of propensity score-based weighting using complex survey data with a binary outcome. Binary or dichotomous outcomes often occur in survey studies, for example, self-reported comorbidity, use of a medication, or certain type of health behavior.¹⁵ When the outcome is binary, a logistic regression is often implemented. In the framework of a logistic regression, the estimated ATE is no longer a direct translation of the regression coefficient in a linear regression, such as in the Dugoff et al. (2014) and Ridgeway et al.(2015) studies, where the ATE is equal to the coefficient of treatment β_T in the outcome model $E[Y|T = 1, \vec{X}] - E[Y|T = 0, \vec{X}] = \beta_T$.^{13, 14, 17, 18} Rather, it is a weighted function in terms of the regression coefficients and covariates in the probability scale, where the weight is chosen depending on the estimand of interest. Compared to a linear regression for continuous outcomes, a special case where the ATE can always be explained by β_T , the estimation of treatment effect with binary outcomes is more representative for the propensity score-based weighting method.

The purpose of this study is to examine the performance of three propensity score-based weighting methods in the context of estimating ATEs and ATTs using survey data with a binary outcome. First, we describe the concepts and derivation of the ATE and the ATT for both the population and sample. Second, we assess the performance of three propensity score-based

weighting methods on survey data using a simulation study under different scenarios by varying the combination of sample size and true treatment effect, degree of misspecification, and level of overlap in propensity score. We then use data from the 2015 National Health Interview Survey (NHIS) as an illustrative example to estimate the effect of provider-patient discussion about smoking on the patient's intent and attempt to quit smoking. Lastly, we discuss how to properly conduct propensity score weighting analysis using survey data with a binary outcome and offer guidelines for researchers to appropriately account for survey weights when performing propensity score weighted analyses.

Methods

- Description of propensity score weighting method

The propensity score is the probability of being treated conditional on observed factors. In the propensity score weighting analysis, the first stage is to estimate the propensity score by regressing the binary treatment variable on the observed baseline covariates, often using a logistic regression.¹⁹ The estimated propensity score is incorporated into the outcome model in the second stage using inverse probability weighting to estimate the ATE or ATT.

Let T denote the binary treatment variable and $Y(T)$ denote the potential binary outcome under treatment T . The ATE is defined as $E[Y(1)] - E[Y(0)]$, which is the difference in the expected value of the potential outcome under treatment $T = 1$, and the expected value of the potential outcome without treatment $T = 0$.²⁰ The ATT is defined as $E[Y(1) | T = 1] - E[Y(0) | T = 1]$, which is the difference in the expected value of the outcome for subjects

in the treated group receiving treatment $T = 1$, and the expected value of the potential outcome for those in the treated group if receiving no treatment $T = 0$.^{21, 22}

In our simulation study, to estimate the ATE, the treated individuals were assigned a weight equal to $\frac{1}{e}$, while control individuals were assigned a weight of $\frac{1}{1-e}$, where e is the estimated propensity score. This weighting is often called “inverse probability of treatment weighting” (IPTW), where the weight is the inverse of the conditional probability of having the subject’s own exposure status (treated/untreated).^{23,24} To estimate the ATT, the treated individuals were assigned a weight equal to 1, while control individuals were assigned a weight of $\frac{e}{1-e}$. This weighting is sometimes referred as “weighting by odds” or standardized mortality ratio weighting,^{24, 25} where the weight can be interpreted as the product of the inverse of the conditional probability of having the subject’s own exposure status and the conditional probability of receiving exposure regardless of one’s actual exposure status. The difference between these two weighting approaches lies in the target group: IPTW weights each group based on the distribution of covariates in the combined sample of both treated and untreated group (the target group), while standardized mortality ratio weighting weights each group based on the distribution of covariates in the treated group (the target group).

- Propensity score weighting and survey data

Survey data often contain complex survey design elements (i.e., strata, cluster, and survey weights). When survey weights are appropriately accounted for, effect estimates can be generalized to the target population.²⁶ When applying propensity score methods to a survey

sample, the estimands of interest include: the sample average treatment effect (SATE), the sample average treatment effect on the treated (SATT), the population average treatment effect (PATE), and the population average treatment effect on the treated (PATT). The former two estimands are used to estimate the treatment effect in the analytic sample, while the latter two refer to the treatment effects in the population from which the sample was drawn.²⁷

²⁸These four estimands are used commonly in health service.¹⁵ Suppose a sample $\mathcal{S}$ is drawn from the target population $\mathcal{P}$ following the survey design such that $\mathcal{S} \subseteq \mathcal{P}$. The SATE is defined as $E[Y(1) - Y(0)|\delta(\mathcal{S}) = 1]$, where $\delta(\mathcal{S}) \in \{0,1\}$ indicates whether the current subject is included in sample $\mathcal{S}$. When $\mathcal{S} = \mathcal{P}$, then $\delta(\mathcal{P}) = 1$. Hence $E[Y(1) - Y(0)|\delta(\mathcal{P}) = 1]$ reduces to $E[Y(1) - Y(0)]$, which is the definition of the PATE. Similarly, the SATT is defined as $E[Y(1) - Y(0)|\delta(\mathcal{S}) = T = 1]$ while the PATT is defined as $E[Y(1) - Y(0)|T = 1]$.

In our simulation study, we explore three approaches of incorporating survey weights in the context of estimating the PATE, PATT, SATE, and SATT: 1) “Unweighted” where neither the propensity score model nor the outcome model is weighted by the survey weight and accounts for the survey design; 2) “Stage II weighted” where the propensity score estimation does not involve survey weights or survey design, but the outcome model uses a weighted regression while accounting for the survey design, where the weight is the product of the survey weight and the propensity score-based inverse probability of treatment weight (often referred as “the composite weight”);²⁹ 3) “Both weighted” where the propensity score estimation uses a survey weighted regression and the outcome model uses a weighted regression, where the weight is

the product of the survey weight and the propensity score-based inverse probability of treatment weight. Both models account for the survey design.

- Simulation study

Following the simulation study design of Dugoff et al. (2014), we simulated population data containing 90,000 observations, which consisted of 3 strata with a sample size of 30,000 each.

We simulated three continuous covariates $X_i \sim N(\mu_{i,s}, 1)$ where $\mu_{1,s} = -0.25, 0, 0.25$; $\mu_{2,s} = -0.5, -0.25, 0$; and $\mu_{3,s} = 0.5, 0.25, 0$ and two binary covariates $X_i \sim \text{Bernoulli}(\mu_{i,s})$ where $\mu_{4,s} = 0.2, 0.4, 0.6$ and $\mu_{5,s} = 0.4, 0.6, 0.8$ for $s = 1, 2, 3$ where s denotes the strata. The probability of being surveyed was a function of $\vec{X} = (X_1, X_2, X_3, X_4, X_5)$, which can be expressed by

$$\pi(\vec{x}) = P(\delta = 1 | \vec{X} = \vec{x}) = \frac{\exp(\gamma_0 + \gamma_1 \cdot x_1 + \gamma_2 \cdot x_2 + \gamma_3 \cdot x_3 + \gamma_4 \cdot x_4 + \gamma_5 \cdot x_5)}{1 + \exp(\gamma_0 + \gamma_1 \cdot x_1 + \gamma_2 \cdot x_2 + \gamma_3 \cdot x_3 + \gamma_4 \cdot x_4 + \gamma_5 \cdot x_5)}$$

where δ is an indicator for being surveyed. We set $\gamma_0 = -1.7$, $\gamma_1 = -\log(1.1)$, $\gamma_2 = -\log(1.25)$, $\gamma_3 = -\log(1.5)$, $\gamma_4 = -\log(1.75)$, and $\gamma_5 = -\log(2)$ to make the overall probability of being surveyed close to 0.1. The probability of being treated was also a function of $\vec{X}$, which can be expressed as

$$P(T = 1 | \vec{X} = \vec{x}) = \frac{\exp(\alpha_0 + \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2 + \alpha_3 \cdot x_3 + \alpha_4 \cdot x_4 + \alpha_5 \cdot x_5)}{1 + \exp(\alpha_0 + \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2 + \alpha_3 \cdot x_3 + \alpha_4 \cdot x_4 + \alpha_5 \cdot x_5)}$$

where T is the indicator of treatment. We set $\alpha_0 = -1.5$, $\alpha_1 = \log(1.1)$, $\alpha_2 = \log(1.25)$, $\alpha_3 = \log(1.5)$, $\alpha_4 = \log(1.75)$, and $\alpha_5 = \log(2)$ so that the overall prevalence of the treated is close to 0.3. The binary outcome Y was generated from another logistic model

$$P(Y = 1|T = t, \vec{X} = \vec{x}) = \frac{\exp(\beta_0 + \beta_t t + \beta_* t x_1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5)}{1 + \exp(\beta_0 + \beta_t t + \beta_* t x_1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5)}$$

where we assume that the effect of treatment T on potential outcomes $Y(0)$ and $Y(1)$ is modified by the covariate x_1 . We set $\beta_0 = -1.05$, $\beta_1 = 1$, $\beta_2 = -\log(1.25)$, $\beta_3 = -\log(1.5)$, $\beta_4 = -\log(1.75)$, and $\beta_5 = -\log(2)$ so that the average outcome of the untreated group is approximately 0.2. The potential outcomes $Y(0)$ and $Y(1)$ were simulated for each subject corresponding to $T = 0$ and $T = 1$, respectively. To relax the assumption of independence between $Y(0)$ and $Y(1)$ for the same subject, we considered the following simulation steps: First, we simulated $Y(0)$ from the Bernoulli distribution with probability

$$p_0(\vec{x}) = P(Y = 1|T = 0, \vec{X} = \vec{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5)}.$$

Next, to simulate potential outcome $Y(1)$, we leveraged the conditional probability of $Y(1)$ given $Y(0)$, namely, a Bernoulli distribution with the parameter

$$P(Y(1) = 1|T = 1, \vec{X} = \vec{x}, Y(0) = y_0) = p_1(\vec{x})[1 + (1 - p_1(\vec{x}))(y_0 - p_0(\vec{x}))]$$

where

$$p_1(\vec{x}) = P(Y = 1|T = 1, \vec{X} = \vec{x}) = \frac{\exp((\beta_0 + \beta_t) + (\beta_1 + \beta_*)x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5)}{1 + \exp((\beta_0 + \beta_t) + (\beta_1 + \beta_*)x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5)}.$$

This conditional probability method of simulation guarantees that:

1. $p_1(\vec{x})[1 + (1 - p_1(\vec{x}))(y_0 - p_0(\vec{x}))] \in (0,1)$ and
 2. the unconditional expectation of $Y(1)$ is $E[P(Y(1) = 1|T = 1, \vec{X} = \vec{x}, Y(0))] = p_1(\vec{x})$
- (Appendix A.1).

Using stratified sampling, subjects were sampled from Stratum 1, 2, and 3 respectively. The sampling was repeated 1,000 times, generating 1,000 datasets. Given the value of β_* is set to $\log(4)$, we considered 9 different scenarios by varying the sample size and average treatment effect. In these 9 scenarios, the sizes of survey subjects sampled were 3,000 (1,300 from Stratum 1, 1,000 from Stratum 2 and 700 from Stratum 3), 6,000 (2,600 from Stratum 1, 2,000 from Stratum 2 and 1,400 from Stratum 3), and 9,000 (3,900 from Stratum 1, 3000 from Stratum 2 and 2,100 from Stratum 3). The true average treatment effect took the values approximately 0.1, 0.2 and 0.3 by adjusting the value of β_t to 0.16, 0.98 and 1.73 respectively. We then considered three additional scenarios by varying a metric of the degree of model misspecification (DoM) in Lenis et al.(2018)'s paper.³⁰ This metric is defined as the population mean of the standardized absolute distance between the correctly specified and potentially misspecified models, namely:

$$\eta = \frac{1}{N} \sum_{i=1}^N \frac{|\hat{g}(\vec{X}_i) - \hat{g}^c(\vec{X}_i)|}{\sigma_{\hat{g}^c}}$$

where $\hat{g}$ and $\hat{g}^c$ represent the predicted conditional means under the proposed model g and the correct model g^c , respectively. By varying the values of β_* , we consider three scenarios for the degree of model misspecification such that η takes values approximately 0.10, 0.15, 0.20, while keeping $\beta_t = 1.08$. The sample size was kept the same as that of Dugoff et al. (2014), with the value of 9,000 (4,000 from Stratum 1, 3,000 from Stratum 2 and 2,000 from Stratum 3). Furthermore, we considered three scenarios for different levels of overlap in propensity score. Following Hu et al. (2021)'s work, $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$ was set as $(-0.7\psi, \log(1.1)\psi, \log(1.25)\psi, \log(1.5)\psi, \log(1.75)\psi, \log(2)\psi)$, where ψ reflects the

strength of confounding, which further results in different level of overlap in propensity score.³¹

We maintained the sample size of 9,000 and set the values $(\beta_t, \beta_*) = (-0.2, \log(2.46))$ for all three levels of overlap in propensity score. We set the value of ψ to 1, 2.5 and 5 to demonstrate when there is strong, moderate, and weak overlap, depicted by the three panels respectively in Figure 1.

Under each scenario, we used two-stage regressions to estimate the treatment effect: 1) a logistic regression model was used to estimate the propensity score, with two alternatives depending on whether or not it was weighted by the survey weight and accounted for strata; 2) a logistic regression model with inverse probability weights based on the estimated propensity score from stage 1 was used to estimate the treatment effect, with also two alternatives depending on whether or not the propensity score based weight was multiplied by the survey weight and accounted for strata. We considered three approaches defined in the previous section: 1) “Unweighted”, 2) “Stage II weighted”; and 3) “Both weighted”.

The PATE, PATT, SATE and SATT were estimated for each scenario based on the predicted means $\hat{y}_i(0)$ and $\hat{y}_i(1)$ for all $i = 1, \dots, N_{sample}$, from the outcome model. The estimated PATE is the average treatment effect weighted by survey weights of the entire sample, calculated as

$$\widehat{PATE} = \frac{1}{\sum_{i=1}^{N_{sample}} w(\vec{x})} \sum_{i=1}^{N_{sample}} w(\vec{x}) [\hat{y}_i(1) - \hat{y}_i(0)]$$

and the estimated PATT is the average treatment effect weighted by survey weights of those who were treated, calculated as

$$\widehat{PATT} = \frac{1}{\sum_{i=1}^{N_{sample}} t_i w(\vec{x})} \sum_{i=1}^{N_{sample}} t_i w(\vec{x}) [\hat{y}_i(1) - \hat{y}_i(0)]$$

Correspondingly, the other two estimands are the sample versions of the above two, calculated as

$$\widehat{SATE} = \frac{1}{N_{sample}} \sum_{i=1}^{N_{sample}} [\hat{y}_i(1) - \hat{y}_i(0)]$$

and

$$\widehat{SATT} = \frac{1}{\sum_{i=1}^{N_{sample}} t_i} \sum_{i=1}^{N_{sample}} t_i [\hat{y}_i(1) - \hat{y}_i(0)]$$

where $t_i \in \{0,1\}$ is the observed treatment for subject i in the sample and $w(x_i) = \pi(x_i)^{-1}$ is the sampling weight (Appendix A.2). To obtain the standard error of each estimated treatment effect $\widehat{TE}$, denoted $se(\widehat{TE})$, bootstrapping with replacement was used, while maintaining the sample size of each stratum, with a total of 400 bootstrap samples. The 95% confidence interval is computed as $\widehat{TE} \pm 1.96 * se(\widehat{TE})$.

For model evaluation, we summarized the mean relative bias in absolute value (MRB) of each estimate and plotted the boxplots for the distribution of relative bias. The mean relative bias in absolute value was defined as $\frac{1}{1000} \sum_{i=1}^{1000} (|1 - \widehat{TE}_i / TE|)$ for each treatment effect. We also compared the coverage probability across three methods. The coverage probability is defined as $\frac{1}{1000} \sum_{i=1}^{1000} I(TE \text{ is in } \widehat{TE} \pm 1.96 * se(\widehat{TE}))$.

Simulation results

- Sample size and treatment effect

The MRB and coverage probability obtained by the three methods for different combinations of sample size and true ATE are reported in Table 1. The method with a lower MRB and coverage

probability closer to 95% is desired. It is observed that the MRB obtained by two weighted methods were very close and consistently lower than that obtained by the “Unweighted” method for all measures, regardless of the sample size and true ATE. Increasing the sample size helped reduce the MRB for the two weighted methods; however, this was not necessarily the case for the “Unweighted” method. In addition, as the value of true ATE increased from 0.1 to 0.3, the MRB obtained by each of the three methods consistently decreased for all four measures (Table 1, Supplement Figure 1).

Table 1: Mean relative bias (in absolute value) and coverage probability obtained by three propensity score weighting approaches under different scenarios by varying the sample size and magnitude of true average treatment effect (MRB is in scale of 10^{-2} , CP is in the scale of percentage)

		True ATE=0.1								True ATE=0.2								True ATE=0.3							
		PATE		SATE		PATT		SATT		PATE		SATE		PATT		SATT		PATE		SATE		PATT		SATT	
S	M	MRB	CP	MRB	CP	MRB	CP	MRB	CP	MRB	CP	MRB	CP	MRB	CP	MRB	CP	MRB	CP	MRB	CP	MRB	CP	MRB	CP
3,000	I	2.58	59.4	2.41	65.4	3.70	62.2	3.52	69.6	1.32	66.6	1.14	73.4	2.04	68.2	1.83	76.0	1.09	47.6	0.96	62.2	1.53	68.6	1.39	76.2
	II	1.46	91.8	1.63	91.0	1.53	93.4	1.65	90.6	0.79	92.8	0.89	90.6	0.91	92.8	0.99	89.8	0.50	94.2	0.54	91.8	0.65	94.2	0.70	90.8
	III	1.46	92.4	1.63	91.8	1.54	93.8	1.66	90.8	0.79	93.8	0.89	91.2	0.92	93.2	1.00	90.4	0.50	94.6	0.54	92.2	0.66	95.0	0.70	91.0
6,000	I	2.46	34.0	2.26	47.2	3.67	43.4	3.49	54.8	1.29	33.6	1.07	55.6	2.06	45.6	1.84	56.4	1.08	21.8	0.93	36.2	1.54	42.6	1.38	58.2
	II	1.17	90.6	1.37	86.8	1.17	92.6	1.30	88.8	0.57	91.0	0.70	87.4	0.66	92.0	0.74	89.0	0.35	92.4	0.41	88.6	0.48	93.2	0.53	90.0
	III	1.17	91.4	1.36	87.0	1.18	92.8	1.31	89.6	0.57	92.2	0.70	88.2	0.66	92.6	0.75	89.6	0.35	92.6	0.41	88.8	0.48	93.4	0.54	90.2
9,000	I	2.43	24.8	2.21	36.4	3.64	28.6	3.46	40.8	1.26	26.4	1.03	46.8	2.02	37.6	1.80	54.4	1.04	8.80	0.89	24.4	1.49	32.6	1.34	47.2
	II	0.98	89.4	1.21	84.4	0.93	91.0	1.09	87.8	0.49	90.0	0.65	85.4	0.56	91.0	0.66	87.6	0.28	91.0	0.36	86.2	0.39	92.0	0.47	88.6
	III	0.98	89.8	1.20	84.8	0.94	91.2	1.10	88.4	0.49	90.8	0.64	85.8	0.56	91.2	0.67	88.2	0.28	91.6	0.35	87.0	0.39	92.2	0.48	89.2

Abbreviations: S, Sample size; M, Method; I, Unweighted; II, Stage II weighted; III, Both weighted; MRB, Mean relative bias (in absolute value); CP, Coverage probability ;PATE population average treatment effect; PATT population average treatment effect on the treated; SATE sample average treatment effect; SATT sample average treatment effect on the treated.

The coverage rates of the bootstrapped confidence intervals from the two weighted methods were very close for all measures, with the “both weighted” method resulting in slightly higher coverage probabilities. Compared to the coverage rates from the “Unweighted” method, the coverage rates from the two weighted methods were consistently higher, no matter which specification of sample size and ATE was used. As the sample size increased, the coverage rates obtained by three methods consistently decreased for all four measures, maintaining approximately 90% coverage probability for PATE and PATT across three specifications of true ATEs when the sample size was 9,000. In contrast, as the ATE increased, the coverage rates from the two weighted methods consistently increased for all four measures, regardless of the sample size (Table 1).

- Degree of model misspecification

The MRB and coverage rates of the bootstrapped confidence intervals under three scenarios of model misspecification are reported in Table 2. As the degree of model misspecification increased, the MRB from the “Unweighted” method increased for all measures drastically, while the MRB from the two weighted methods changed by a much smaller magnitude. Similarly, the coverage rates from the “Unweighted” method decreased consistently for all measures, while the coverage rates obtained by the two weighted methods remained relatively stable. Under Scenario 1 where the degree of model misspecification was low, the “Unweighted” method yielded both lower MRB and higher coverage rates for SATE and SATT, compared to the two weighted methods. As the value of η increased from 0.10 to 0.15 in Scenario 2, the two weighted methods outperformed the “Unweighted” one, with consistently

315 lower MRB and higher coverage rates for all measures. Under Scenario 3, the two weighted
316 methods were favored again, with advantages more evident in lower MRB and higher coverage
317 rates compared to the “Unweighted” method. Overall, the results confirm that the weighted
318 methods outperform the unweighted one when the misspecification is not negligible. In
319 addition, the two weighted methods consistently yielded coverage rates close to 95% for
320 estimating PATE and PATT, regardless of the degree of model misspecification (Table 2,
321 Supplement Figure 2).

322 Table 2: Mean relative bias (in absolute value) and coverage probability obtained by three
323 propensity score weighting approaches under different scenarios by varying the degree of
324 model misspecification (MRB is in scale of 10^{-2} , CP is in the scale of percentage)

Degree of model misspecification	Method	PATE		SATE		PATT		SATT	
		MRB	CP	MRB	CP	MRB	CP	MRB	CP
Low ($\eta = 0.10$)	Unweighted	6.79	85.8	4.46	94.2	9.38	85.4	5.76	88.8
	Stage II weighted	4.79	94.6	6.59	87.2	6.36	93.8	8.32	85.2
	Both weighted	4.78	95.0	6.58	87.8	6.41	94.0	8.37	85.4
Medium ($\eta = 0.15$)	Unweighted	9.18	79.8	6.25	86.8	14.46	58.8	11.28	64.2
	Stage II weighted	4.62	94.4	6.25	87.0	5.70	94.0	7.20	87.2
	Both weighted	4.62	94.6	6.23	87.2	5.75	94.2	7.27	87.4
High ($\eta = 0.20$)	Unweighted	12.98	60.2	11.08	54.8	20.34	20.2	18.57	28.8
	Stage II weighted	4.56	94.4	5.71	90.8	5.14	94.2	6.01	88.8
	Both weighted	4.54	94.8	5.68	91.0	5.22	94.4	6.07	89.0

325 Abbreviations: MRB, Mean relative bias (in absolute value); CP, Coverage probability; PATE population
326 average treatment effect; PATT population average treatment effect on the treated; SATE sample
327 average treatment effect; SATT sample average treatment effect on the treated.

328

- 329
- Level of overlap in propensity score

330 The MRB and coverage rates of the bootstrapped confidence intervals under the three levels of
331 overlap in propensity score are reported in Table 3. As the level of overlap changed from strong
332 to weak, the MRB for all four measures increased and the coverage rates decreased
333 consistently. This was observed for all three methods. Compared to the “Unweighted” method,
334 the two weighted methods yielded consistently lower MRB and higher coverage rates for all
335 measures, regardless of the level of overlap in propensity score (Table 3, Supplement Figure 3).
336 Similar to the results in other scenarios, the performance of the two weighted methods was
337 very close, with the “Both weighted” method yielding slightly higher coverage rates in most
338 cases (Table 3).

339 Table 3: Mean relative bias (in absolute value) and coverage probability obtained by three
340 propensity score weighting approaches under different scenarios by varying the level of overlap
341 in propensity score (MRB is in scale of 10^{-2} , CP is in the scale of percentage)

		PATE		SATE		PATT		SATT	
Level of overlap in propensity score	Method	MRB	CP	MRB	CP	MRB	CP	MRB	CP
Strong ($\psi = 1$)	Unweighted	5.50	63.2	5.27	45.2	6.54	65.2	6.36	41.2
	Stage II weighted	2.58	90.2	2.93	89.4	2.42	93.2	2.72	89.4
	Both weighted	2.58	89.8	2.93	89.8	2.42	93.4	2.71	89.6
Moderate ($\psi = 2.5$)	Unweighted	6.11	52.8	5.96	39.2	7.49	50.2	7.40	25.6
	Stage II weighted	3.78	88.0	4.24	87.4	3.84	91.0	4.20	87.4
	Both weighted	3.78	88.2	4.23	87.8	3.84	91.4	4.20	87.8
Weak ($\psi = 5$)	Unweighted	10.51	43.6	10.60	12.2	9.83	43.2	9.78	13.4
	Stage II weighted	7.28	85.2	7.85	85.4	6.65	88.4	6.95	84.2
	Both weighted	7.27	86.2	7.83	85.2	6.63	88.8	6.92	85.0

342
343 Abbreviations: MRB, Mean relative bias (in absolute value); CP, Coverage probability; PATE population
344 average treatment effect; PATT population average treatment effect on the treated; SATE sample
345 average treatment effect; SATT sample average treatment effect on the treated.

346

Real data application

To encourage smoking cessation, the 2008 U.S. Public Health Service's Clinical Practice Guideline recommends healthcare professionals consistently identify smokers and treat them using a brief intervention known as the '5A' model: 1) ask (identify and document tobacco use status); 2) advise (urge every tobacco user to quit); 3) assess (is the tobacco user willing to make a quit attempt at this time?); 4) assist (for the patient willing to make a quit attempt, use counselling and pharmacotherapy to help them quit); and 5) arrange (schedule follow-up contact within the first week after the quit date).³² To assess the effect of the first two of the 5 A's on smoking behaviors among current smokers, we used data from 4071 adult (age ≥ 18 years) current smokers in the 2015 National Health Interview Survey (NHIS) who had seen a doctor or other health professional in the past 12 months. The NHIS is an on-going cross-sectional survey of the civilian non-institutionalized population in the U.S. using a stratified multistage sample design, where sampling weights are constructed so that the survey sample can be expanded to represent the total population of the U.S.; the datasets are publicly available. We used both the core survey components and the Tobacco Section in the Cancer Control Supplement to obtain comprehensive information about current smokers' smoking and quitting behavior, as well as their interactions with healthcare providers.

The first question of interest was whether recent provider-patient discussions about smoking were associated with an attempt to quit smoking. The second question of interest was whether recent provider-patient discussions about smoking habits were associated with intent to quit smoking. Smokers who answered 'yes' to "DURING THE PAST 12 MONTHS, has a doctor or

369 other health professional talked to you about your smoking?” or ‘yes’ to “In the PAST 12
370 MONTHS, has a medical doctor, dentist, or other health professional ADVISED you to quit
371 smoking, or to quit using other kinds of tobacco?” were considered treated. Those who
372 answered ‘no’ to both questions were untreated. The two main outcomes were: attempt to
373 quit (defined as stopped smoking for more than one day because of trying to quit smoking
374 during the past 12 months), and intent to quit (defined as would like to completely quit
375 smoking cigarettes).

376

377 We used the three methods considered in the simulations to estimate the treatment effect.
378 The variables in the propensity score model included the following commonly adjusted socio-
379 demographic factors: age, sex, race/ethnicity group, marital status, region, education,
380 employment status, health insurance type, and the ratio of family income to the applicable
381 Federal poverty thresholds, as well as three smoking-related comorbidities: lung disease,
382 cardiovascular diseases, and cancer. Similar to previous studies, we included an indicator for
383 serious psychological distress using the Kessler Scale.³³ In addition, we included two measures
384 of smoking exposure: years smoked and numbers of cigarettes smoked daily. In total, we
385 included 16 covariates for the analysis.

386

387 Out of 4,071 subjects who visited a doctor during the past 12 months, 2,909 (weighted 71.46%)
388 had a discussion with their provider about smoking and 1,162 (weighted 28.54%) did not. Of
389 those who had the provider-patient discussion, 1,517 (weighted 52.32%) attempted to quit

390 smoking, compared to 543 (weighted 46.55%) of those who did not have the discussion
 391 attempted to quit smoking (Table 4).

392 Table 4. The effect of provider-patient discussion about smoking on patients' attempt to quit
 393 smoking from a propensity score-weighted analysis

Method	Group	No. attempted to quit	Weighted ^a outcome%	PATE [95% CI]	PATT [95% CI]	SATE [95% CI]	SATT [95% CI]
Unweighted	No Discussion (N=1162)	543	46.55%	reference	reference	reference	reference
	Discussion (N=2909)	1,517	52.32%	10.55 [6.99-14.10]	10.71 [6.90-14.51]	10.55 [6.99-14.10]	10.71 [6.90-14.51]
Stage II weighted	No discussion (N=1162)	543	46.55%	reference	reference	reference	reference
	Discussion (N=2909)	1,517	52.32%	11.70 [8.14-15.26]	12.20 [8.39-16.00]	11.70 [8.15-15.26]	12.20 [8.39-16.00]
Both weighted	No discussion (N=1162)	543	46.55%	reference	reference	reference	reference
	Discussion (N=2909)	1,517	52.32%	11.58 [8.02-15.13]	12.03 [8.23-15.84]	11.58 [8.02-15.14]	12.03 [8.23-15.84]

394 Abbreviations: PATE population average treatment effect; PATT population average treatment effect on
 395 the treated; SATE sample average treatment effect; SATT sample average treatment effect on the
 396 treated.

397 ^a The weighted percentage of subjects with the outcome of interest, weighted by the survey weight.
 398

399 Table 4 reports the estimates of the four measures of the effect of having a provider-patient
 400 discussion on a patient's attempt to quit smoking. For example, the estimated PATE in the
 401 "both weighted" method is 11.58 (95% CI [8.02-15.13]) which indicates that provider-patient
 402 discussion was associated with an average increase in the probability to attempt to quit

403 smoking of approximately 12%. The “Unweighted” method consistently yields lower estimates
 404 for all four measures, with PATE, PATT, SATE and SATT 10.55 (95% CI [6.99-14.10]), 10.71 (95%
 405 CI [6.90-14.51]), 10.55 (95% CI [6.99-14.10]) and 10.71 (95% CI [6.90-14.51]) respectively.

406 In an analysis of the effect of provider-patient discussion about smoking on intent to quit
 407 smoking, 2,089 (weighted 71.81%) of those who had the provider-patient discussion intended
 408 to quit smoking; while only 726 (weighted 62.48%) of those who did not have the discussion
 409 intended to quit smoking (Table 5). Table 5 reports the four measures of the effect of having a
 410 provider-patient discussion on patient’s intent to quit smoking. The “Unweighted” method
 411 consistently produces higher estimates for all four measures. Similar to the results in Table 4,
 412 both weighted methods yield similar estimates. For example, the estimated PATEs of having a
 413 patient-provider discussion on patient’s intent to quit smoking are 8.31 (95% CI [4.78-11.83])
 414 and 8.39 (95% CI [4.87-11.91]), respectively from the “Stage II weighted” and “Both weighted”
 415 models.

416 Table 5. The effect of provider-patient discussion about smoking on patients’ intent to quit
 417 smoking from a propensity score-weighted analysis

Method	Group	No. intended to quit	Weighted ^a outcome %	PATE [95% CI]	PATT [95% CI]	SATE [95% CI]	SATT [95% CI]
Unweighted	No Discussion (N=1162)	726	64.62%	reference	reference	reference	reference
	Discussion (N=2909)	2,089	71.10%	12.23 [8.70-15.75]	12.89 [9.06-16.71]	12.42, [8.85-15.99]	13.10, [9.22-16.97]
	No discussion	726	64.62%	reference	reference	reference	reference

Stage II weighted	(N=1162)						
	Discussion (N=2909)	2,089	71.10%	8.31 [4.78-11.83]	8.74 [4.91-12.56]	8.48 [4.91-12.05]	8.93 [5.05-12.80]
Both weighted	No Discussion (N=1162)	726	64.62%	reference	reference	reference	reference
	Discussion (N=2909)	2,089	71.10%	8.39 [4.87-11.91]	8.66 [4.84-12.49]	8.57 [5.00-12.14]	8.85 [4.98-12.73]

418 Abbreviations: PATE population average treatment effect; PATT population average treatment effect on
419 the treated; SATE sample average treatment effect; SATT sample average treatment effect on the
420 treated.

421 ^a The weighted percentage of subjects with the outcome of interest, weighted by the survey weight.

422

423

424 Discussion

425

426 In this study, we conducted a series of simulations to evaluate three propensity score weighting
427 methods when using complex survey data to estimate the effect of a treatment on a binary
428 outcome. Under different scenarios by varying sample size, treatment effect, degree of model
429 misspecification, and level of covariate overlap, we considered three methods where the survey
430 weight was incorporated or not incorporated into the estimation of either the propensity score
431 or the treatment effect.

432

433 We found that a larger treatment effect and stronger overlap in propensity score tended to be
434 associated with smaller relative biases and larger coverage rates from all three propensity score
435 weighting methods for all four measures of treatment effects. This suggests that preliminary
436 analysis of treatment effect and overlap in propensity score may be helpful to assess the

performances of propensity score weighting methods. Researchers may decide whether additional procedures such as trimming extreme scores or overlap weighting are needed to further reduce the potential biases before the propensity score weighting methods are implemented.³⁴

We also found that compared to the two weighted methods, the “Unweighted” method tended to generate drastically reduced coverage rates but not reduced bias as the sample size increased. The performance of the “Unweighted” method was also sensitive to the degree of model misspecification, while the two weighted methods yielded robust estimators with reasonable bias and coverage rates under various degrees of model misspecification. When the misspecification is not negligible, both weighted methods obtained favorable results for all four measures, regardless of the sample size, magnitude of treatment effect, and level of overlap in propensity score. Therefore, we recommend that the survey weight should always be incorporated into the estimation of the treatment effect and that there are marginal gains in terms of coverage rates when the survey weight is incorporated into the estimation of the propensity score.

In the DuGoff et al. (2014) and Ridgeway et al. (2015) studies with a continuous outcome, the weighted methods yielded consistently better results than the “Unweighted” method. We found that with a binary outcome, results are similar but less consistent as the degree of misspecification varies, which was not examined in the previous studies. When the degree of misspecification was low, the “Unweighted” method produced favorable SATE and SATT with

the lowest MRB and reasonable coverage probability. However, since misspecification is unknown in real data settings, caution must be taken when estimating SATE and SATT using the “Unweighted” method.

Given the results of the simulation studies, we are unable to provide strong recommendations in terms of whether to incorporate survey weights in the first stage or not. DuGoff et al. (2014) suggested not incorporating the survey weights in the first stage, while Ridgeway et al. (2015) disagreed with this assertion. Our results show that both methods yielded very similar results, where the “Both weighted” method resulted in slightly higher coverage probabilities in most settings. Finally, we found that the two weighted methods yielded consistent estimated PATT, with very close MRE values regardless of the degree of misspecification. This finding corroborates with the results in Lenis et al. (2018)’s paper, showing that the doubly robust estimators from propensity weighting analysis yield a consistent estimator of the PATT when either the propensity score model or the outcome model is correctly specified.^{30, 35} In our case, the misspecification was only in the outcome model while the propensity score model was correctly specified.

In summary, we recommend that survey weights should be incorporated into propensity score weighting analyses using survey data: either by incorporating survey weights in the outcome model or incorporating survey weights in both the outcome model and the propensity score model. The unweighted method offers better results for SATE and SATT when the degree of misspecification is negligible, however, the degree of the misspecification is often unknown in

481 reality. The advantage of weighted methods is more evident when the estimand of interest is
482 PATE or PATT.

483
484 Our work focuses on estimating the treatment effect on a binary outcome using propensity
485 score weighing on simulated survey data. Building on this work, future areas of research
486 include designing the simulation of survey data with two levels (strata and cluster), inclusion of
487 more strata and additional predictors, and using other methods in deriving the propensity
488 score, for example bagging or tree-based methods. While our study explored binary outcomes
489 in survey data settings, and other studies have focused on continuous outcomes, this work can
490 be further extended to alternative types of outcomes, such as count outcomes and survival
491 outcomes.

492 493 **Conclusion**

494 This paper compared three propensity score weighting approaches to explore the appropriate
495 application of propensity score-based weights in survey data with binary outcomes. A
496 simulation study with varying combinations of sample size and treatment effect size, degrees of
497 misspecification, and levels of overlap in propensity score is provided along with an illustration
498 through the application to a real data example. The propensity score weighting methods
499 accounting for the survey weight are more appropriate for estimating the population-level
500 effects, particularly the population average treatment effect. We recommend that when using
501 propensity score weighting with survey data, the survey weight is included in the outcome
502 model or in both the propensity score and outcome models.

503 **List of abbreviations**

504 SATE: Sample average treatment effect

505 PATE: Population average treatment effect

506 SATT: Sample average treatment effect on the treated

507 PATT: Population average treatment effect on the treated

508 NHIS: National Health Interview Survey

509 ATE: Average treatment effect

510 ATT: Average treatment effect on the treated

511 IPTW: Inverse probability of treatment weighting

512 DoM: Degree of model misspecification

513 MRB: Mean relative bias

514 RB: Relative bias

515

516 **Declarations:**

517 **Ethics approval and consent to participate:**

518 All methods were performed in accordance with the relevant guidelines and regulations. All

519 methods implementation, simulations and statistical analyses were performed using R

520 Statistical Software (version 4.0.4; R Foundation for Statistical Computing, Vienna, Austria). Our

521 study is not human subjects research as it consists of only simulated datasets and publicly

522 available survey data from National Health Interview Survey (NHIS). Our institution's

523 institutional review board does not require an approval for research involving NHIS data which

524 are completely de-identified and publicly available.

525 **Consent for publication**

526 Not applicable.

527 **Availability of data and materials**

528 Deidentified NHIS data are publicly available at

529 https://www.cdc.gov/nchs/nhis/nhis_2015_data_release.htm

530 **Competing interests**

531 The authors declare that they have no competing interests.

532 **Funding**

533 Research reported in this work was supported in part by the National Cancer Institute Cancer

534 Center Support Grant P30CA196521-01 awarded to the Tisch Cancer Institute (TCI) and the

535 National Institute on Aging Claude D. Pepper Older Americans Independence Center

536 P30AG028741 of the Icahn School of Medicine at Mount Sinai (IS-MMS).

537 **Author's contributions**

538 All authors have read and approved the manuscript. C.C. wrote the final version of the

539 statistical codes, drafted the first version of the manuscript, and reviewed the final version of

540 the paper. M.S.C. helped develop the methods, conceived the simulation study and critically

541 revised the manuscript for important intellectual content. J.J. validated the final version of the

542 codes for simulations and data analyses and reviewed the final version of the paper. M.A.

543 worked on interpreting the results, deciding on its presentation and reviewed the final version

544 of the manuscript. L.L. developed the methods and conceived the simulation study, supervised

545 coding, and finalized the manuscript. The corresponding author ensures that all the listed

546 authors meet authorship criteria.

Acknowledgments

The authors thank Dr. Melanie Besculides for editing the manuscript.

Reference

1. Benson K and Hartz AJ. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine* 2000; 342: 1878-1886. DOI: 10.1056/NEJM200006223422506.
2. Feinstein AR. Epidemiologic analyses of causation: the unlearned scientific lessons of randomized trials. *Journal of Clinical Epidemiology* 1989; 42: 481-489. DOI: 10.1016/0895-4356(89)90142-X.
3. Lee BK, Lessler J and Stuart EA. Improving propensity score weighting using machine learning. *Stat Med* 2010; 29: 337-346. DOI: 10.1002/sim.3782.
4. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70: 41-55. DOI: 10.1093/biomet/70.1.41.
5. Suri RS, Li L and Nesrallah GE. The risk of hospitalization and modality failure with home dialysis. *Kidney International* 2015; 88: 360-368. 2015/03/19. DOI: 10.1038/ki.2015.68.
6. Tremblay D, King A, Li L, et al. Risk factors for infections and secondary malignancies in patients with a myeloproliferative neoplasm treated with ruxolitinib: a dual-center, propensity score-matched analysis. *Leuk Lymphoma* 2020; 61: 660-667. 2019/11/13. DOI: 10.1080/10428194.2019.1688323.
7. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011; 46: 399-424. 2011/06/08. DOI: 10.1080/00273171.2011.568786.

570 8. Austin PC and Mamdani MM. A comparison of propensity score methods: a case-study
571 estimating the effectiveness of post-AMI statin use. *Stat Med* 2006; 25: 2084-2106. 2005/10/13. DOI:
572 10.1002/sim.2328.

573 9. Rosenbaum PR. Model-based direct adjustment. *Journal of the American Statistical Association*
574 1987; 82: 387-394. DOI: 10.2307/2289440.

575 10. Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical*
576 *Science* 2010; 25: 1-21. DOI: 10.1214/09-STS313.

577 11. Zanutto E. A comparison of propensity score and linear regression analysis of complex survey
578 data. *Journal of Data Science* 2006; 4: 67-91.

579 12. Zanutto E, Lu B and Hornik R. Using propensity score subclassification for multiple treatment
580 doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*
581 2005; 30: 59-73. DOI: 10.3102/10769986030001059.

582 13. DuGoff EH, Schuler M and Stuart EA. Generalizing observational study results: applying
583 propensity score methods to complex surveys. *Health Services Research* 2014; 49: 284-303. DOI:
584 10.1111/1475-6773.12090.

585 14. Ridgeway G, Kovalchik SA, Griffin BA, et al. Propensity score analysis with survey weighted data.
586 *Journal of Causal Inference* 2015; 3: 237-249. 2015/09/01. DOI: 10.1515/jci-2014-0039.

587 15. Austin PC, Jembere N and Chiu M. Propensity score matching and complex surveys. *Statistical*
588 *Methods in Medical Research* 2018; 27: 1240-1257. DOI: 10.1177/0962280216658920.

589 16. Lenis D, Nguyen TQ, Dong N, et al. It's all about balance: propensity score matching in the
590 context of complex survey data. *Biostatistics* 2019; 20: 147-163. 2018/01/03. DOI:
591 10.1093/biostatistics/kxx063.

592 17. Gomila R. Logistic or linear? Estimating causal effects of experimental treatments on binary
593 outcomes using regression analysis. *Journal of Experimental Psychology: General* 2021; 150: 700-709.
594 DOI: 10.1037/xge0000920.

595 18. Goetghebeur E, le Cessie S, De Stavola B, et al. Formulating causal questions and principled
596 statistical answers. *Stat Med* 2020; 39: 4922-4948. 2020/09/24. DOI: 10.1002/sim.8741.

597 19. Rosenbaum PR and Rubin DB. Reducing bias in observational studies using subclassification on
598 the propensity score. *Journal of the American Statistical Association* 1984; 79: 516-524. DOI:
599 10.2307/2288398.

600 20. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies.
601 *Journal of Educational Psychology* 1974; 66: 688-701. DOI: 10.1037/h0037350.

602 21. Heckman JJ and Robb R. Alternative methods for evaluating the impact of interventions: An
603 overview. *Journal of Econometrics* 1985; 30: 239-267. DOI: 10.1016/0304-4076(85)90139-3.

604 22. Rubin DB. Assignment to treatment group on the basis of a covariate. *Journal of Educational*
605 *Statistics* 1977; 2: 1-26. DOI: 10.2307/1164933.

606 23. Lunceford JK and Davidian M. Stratification and weighting via the propensity score in estimation
607 of causal treatment effects: a comparative study. *Stat Med* 2004; 23: 2937-2960. DOI:
608 10.1002/sim.1903.

609 24. Sato T and Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*
610 2003; 14: 680-686. 2003/10/22. DOI: 10.1097/01.EDE.0000081989.82616.7d.

611 25. Harder VS, Stuart EA and Anthony JC. Propensity score techniques and the assessment of
612 measured covariate balance to test causal associations in psychological research. *Psychological Methods*
613 2010; 15: 234-249. 2010/09/09. DOI: 10.1037/a0019623.

614 26. Pfeiffermann D. The role of sampling weights when modeling survey data. *International*
615 *Statistical Review* 1993; 61: 317-337. DOI: 10.2307/1403631.

27. Imai K, King G and Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2008; 171: 481-502. DOI: 10.1111/j.1467-985X.2007.00527.x.
28. Imbens G. Nonparametric estimation of average treatment effects under exogeneity: a review. *The Review of Economics and Statistics* 2004; 86: 4-29.
29. Cook BL, McGuire TG, Meara E, et al. Adjusting for health status in non-linear models of health care disparities. *Health Services and Outcomes Research Methodology* 2009; 9: 1-21. 2010/03/31. DOI: 10.1007/s10742-008-0039-6.
30. Lenis D, Ackerman B and Stuart EA. Measuring model misspecification: application to propensity score methods with complex survey data. *Comput Stat Data Anal* 2018; 128: 48-57. 2018/05/22. DOI: 10.1016/j.csda.2018.05.003.
31. Hu L, Ji J and Li F. Estimating heterogeneous survival treatment effect in observational data using machine learning. *Stat Med* 2021; 40: 4691-4713. 2021/06/12. DOI: 10.1002/sim.9090.
32. Treating tobacco use and dependence: 2008 update. Agency for Healthcare Research and Quality, <https://www.ahrq.gov/prevention/guidelines/tobacco/index.html>, Accessed on February 4, 2022
33. Henley SJ, Asman K, Momin B, et al. Smoking cessation behaviors among older U.S. adults. *Preventive Medicine Reports* 2019; 16: 100978. 2019/10/30. DOI: 10.1016/j.pmedr.2019.100978.
34. Li F, Thomas LE and Li F. Addressing extreme propensity scores via the overlap weights. *American Journal of Epidemiology* 2019; 188: 250-257. 2018/09/07. DOI: 10.1093/aje/kwy201.
35. Joseph DYK and Joseph LS. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 2007; 22: 523-539. DOI: 10.1214/07-STS227.

Figure 1

Figure 1: Level of overlap in propensity score keeping overall treatment prevalence approximately equal to 0.5. Panel A to C respectively represents strong, moderate, and weak overlap in propensity score by specifying the value of ψ to 1, 2.5 and 5 correspondingly.

Additional Files:

Additional file 1 — Additional Figures

Supplement Figure 1. The boxplots of relative biases (RB, in absolute value) obtained by three propensity score weighting approaches under different scenarios by varying the sample size and magnitude of true average treatment effect (RB is in scale of 10^{-2})

Supplement Figure 2. The boxplots of relative biases (RB, in absolute value) obtained by three propensity score weighting approaches under different scenarios by varying the degree of model misspecification (RB is in scale of 10^{-2})

Supplement Figure 3. The boxplots of relative biases (RB, in absolute value) obtained by three propensity score weighting approaches under different scenarios by varying the level of overlap in propensity score (RB is in scale of 10^{-2})

Additional file 2 — Appendix

A.1 The simulation based on the conditional probability

A.2 Calculation of the estimated PATE and PATT

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [20220216SupplementaryMaterials.docx](#)