Health Services and Outcomes Research Methodology (2023) 23:21–44
https://doi.org/10.1007/s10742-022-00283-x

Adjustment for biased sampling using NHANES derived propensity weights

Olivia M. Bernstein[1] · Brian G. Vegetabile[2] · Christian R. Salazar[3] · Joshua D. Grill[3,4,5] · Daniel L. Gillen[1,3]

Received: 3 November 2021 / Revised: 15 April 2022 / Accepted: 14 June 2022 /
Published online: 8 July 2022

Abstract

The Consent-to-Contact (C2C) registry at the University of California, Irvine collects data from community participants to aid in the recruitment to clinical research studies. Self-selection into the C2C likely leads to bias due in part to enrollees having more years of education relative to the US general population. Salazar et al. (Alzheimer's Dementia Transl Res Clin Interv 6(1):e120023, 2020, https://doi.org/10.1002/trc2.12023) recently used the C2C to examine associations of race/ethnicity with participant willingness to be contacted about research studies. To obtain representative estimates from C2C we use weighted estimation of associations of interest where the weights are related to the probability of self-selection into the convenience sample. The selection probabilities are estimated using data from the National Health and Nutrition Examination Survey (NHANES). We create a combined dataset of C2C and NHANES subjects and evaluate the trade-offs of different approaches (logistic regression, covariate balancing propensity score, entropy balancing, and random forest) for estimating the probability of membership in C2C relative to NHANES. We further propose methods to estimate the variance of parameter estimates that account for uncertainty that arises from estimating propensity weights. Simulation studies explore the impact of propensity weight estimation on uncertainty. We demonstrate the approach by repeating the analysis by Salazar et al. (Alzheimer's Dementia Transl Res Clin Interv 6(1):e120023, 2020, https://doi.org/10.1002/trc2.12023) with the deduced propensity weights for the C2C subjects and contrast the results of the two analyses. This method can be implemented using our `estweight` package in `R` available on GitHub.

Keywords Biased sample · Convenience sample · Propensity weight · NHANES

✉ Olivia M. Bernstein
obernste@uci.edu

Extended author information available on the last page of the article

1 Introduction

The Consent-to-Contact (C2C) registry at the University of California, Irvine (https://c2c.uci.edu) enrolls potential participants to aid in clinical research recruitment strategies (Grill et al. 2018). Participants are recruited into the registry through a variety of outreach strategies including emails, community talks, postcards, and other methods. Due to this the C2C is not expected to be representative of the United States (US) population. For example, C2C participants tend to report more years of education relative to the general population, are more likely to be non-Hispanic White, have lower rates of comorbidities and higher rates of exercise (see Fig. 3). C2C participants self-report demographic and clinical characteristics during the enrollment process. They also indicate their willingness to be contacted for potential participation in studies that involve various procedures or requirements such as lifestyle/behavioral modification, medication use, blood collection, brain imaging, autopsy, or lumbar punctures. Depending on the requirements and enrollment criteria of a study, participants who specifically report willingness to be contacted about required procedures are likely to be eligible and can be invited to participate, increasing the efficiency of recruitment.

Due to the under-representation of racial/ethnic minority populations in clinical research (Oh et al. 2015), members from our research team used data from the C2C to examine differences in willingness to participate by race/ethnicity (Salazar et al. 2020). Since the C2C is unrepresentative of the target population, using it directly could lead to potentially biased estimates and limited generalizability of estimated associations. For example, one natural way this bias could arise is through a differential relationship between race/ethnicity and willingness to participate by education level.

Convenience samples, like the C2C, are widely used to answer scientific questions because samples representative of the population may be impractical or unethical to collect. Most statistical methods assume representative sampling, but may be naively applied to biased samples. Participant self-selection into convenience samples may, however, reduce the degree to which such samples are representative of target populations of interest leading to a biased sample.

Some common approaches to address biased samples from self-selection, or selection bias, incorporate outside information about a population to obtain more generalizable inference. MRP (multilevel regression and poststratification) derives subpopulation estimates from national surveys. Iterative proportional fitting (or raking) adjusts subpopulation counts to match known marginal counts. For examples of MRP see Gelman and Little (1997) and Park et al. (2004), for an application of MRP see Schell et al. 2020 and for raking see Bishop et al. (1975) and Little and Rubin (2014). Additionally, there are several methods developed to combine estimates from multiple surveys, such as small-area estimation, joint-modeling, and imputation based methods. For a summary see Elliott et al. (2018).

Inverse probability of sampling weights is a popular solution for obtaining more generalizable inference and this approach is the focus of this paper. Inverse probability weights are commonly used to adjust for design-based sampling of subpopulations and provide generalizable inference (Lumley 2010). For design-based survey sampling, inverse probability weights are generally prespecified and fixed by design as the inverse of the sampling probability for each unit. Weights can also be used to account for selection bias in convenience samples, but in this context they are not fixed or known. Convenience samples often do not reflect the prespecified target population of interest because subjects self-select to participate. One solution is to use a representative sample to estimate sampling probabilities and the corresponding sampling weights for subjects in convenience samples (see for example Chen et al. 2020; Ackerman et al. 2021; Robbins et al. 2020; Elliott and Valliant 2017; Zadrozny 2004; O'Muircheartaigh and Hedges 2014).

Estimating sampling weights for convenience samples can only be done if a relevant representative dataset is available. Most of the previous work on calibrating sampling weights assume that a representative sample or a sample frame was readily available (as in Zadrozny 2004; O'Muircheartaigh and Hedges 2014). The National Health and Nutrition Examination Survey (NHANES) is a practical dataset for estimating propensity weights in biomedical convenience samples such as the C2C because it is representative of the US population, it contains medical measurements, and the data are open access (https://wwwn.cdc.gov/nchs/nhanes/Default.aspx). NHANES recruits approximately 5,000 individuals from across the US each year and over-samples people over 65 and minority groups, i.e. Hispanic, non-Hispanic (NH) Black and NH Asian subjects. NHANES data comes with sampling weights for each subject that are a function of the sampling probabilities and can be used to obtain a dataset that is representative of the US. Medical and dietary information are collected through structured questionnaires and in-person measurements ([9]). It is common to compare estimates of population parameters, such as the prevalence of diabetes in the US, to estimates from NHANES as a diagnostic check for selection bias (Funk et al. 2017; Greenblatt et al. 2019; Bailey et al. 2013), but it is not often used for estimating sampling weights. Concurrent work from Ackerman et al. (2021) proposed the use of national survey samples, including NHANES, as a practical way to obtain a representative sample to generalize randomized trial results. They used sampling weights for the complex survey sample in the propensity weight estimation model and the final outcome model, but we utilize them as frequency weights and generate a representative pseudopopulation.

The goal of this paper is to obtain generalizable estimates of a scientific association from a convenience sample while correcting for sampling bias using a representative sample. In our setting, we are unable to obtain a pooled estimate across the convenience sample and representative sample because the outcome of scientific interest is not observed in the representative dataset. In this paper, we focus on estimating inverse probability weights or, equivalently, propensity weights for inclusion into a convenience sample. We do this by combining a convenience and representative sample and use the commonly collected covariates between the two to estimate the probability of membership in the convenience sample versus the representative sample. Our goal is to make this method easy to implement by using NHANES as a representative dataset and providing a package in *R*. We further derive an analytic variance estimator that extends the sandwich estimator for survey weighted generalized linear models (Lumley and Scott 2017) to account for uncertainty from estimating the propensity weights. We follow a similar approach to Schildcrout and Rathouz (2010) and treat the sampling weight estimation model and final outcome model as being simultaneously estimated. Recent work by Wang et al. (2021) and work by Chen et al. (2020) also use a simultaneous estimation procedure for estimating the variance in the final parameter estimates when using estimated sampling weights. Our approach is similar, but we derive the variance of coefficient estimates in the final outcome model, instead of for population mean estimates. Alternatively, Ackerman et al. (2021) used a double-bootstrap to account for uncertainty from estimating propensity weights and from the impact of non-response on the complex survey sample weights. We apply these methods to obtain valid population-level inference for the C2C registry.

The remainder of the manuscript is organized as follows: In Sect. 2, we develop the proposed methodology for calibrating propensity weights and quantifying uncertainty in the final scientific model of interest. In Sect. 3, we present a simulation study investigating the impact of estimated propensity weights on bias and variance. In Sect. 4, we apply our method to an analysis of racial and ethnic differences in research willingness. Finally, we conclude with a discussion of the advantages and limitations of the proposed method.

2 Methodology

Consider two collections of variables, $\mathcal{X}_R$ and $\mathcal{X}_C$ and let $\mathcal{X}$ be the set of variables in both $\mathcal{X}_R$ and $\mathcal{X}_C$, or $\mathcal{X} \equiv \mathcal{X}_R \cap \mathcal{X}_C$. Further, let $\mathcal{Y}$ be a subset of the variables in $\mathcal{X}_C$ that are not in $\mathcal{X}_R$, i.e., $\mathcal{Y} \in \mathcal{X}_C \setminus \mathcal{X}$. We will assume that it is possible to collect data sets on variables from both $\mathcal{X}_R$ and $\mathcal{X}_C$, but that it is not possible to collect random samples for all of the variables from $\mathcal{X}_C$ and thus difficult to obtain population inference. A data set obtained on $\mathcal{X}_C$ will be "convenience" and thus subject to potential bias (such as self-selection bias among other potential issues). Our ultimate goal is population-based inference on the set of variables in $\mathcal{Y}$ that are only collected in the convenience sample. For example, we may be interested in the estimation of the association between some subset of variables from $\mathcal{X}$, defined as $\mathcal{Z}$, i.e., $\mathcal{Z} \subset \mathcal{X}$, and $\mathcal{Y}$. In our context $\mathcal{Z}$ is race/ethnicity (available in both NHANES and C2C) and $\mathcal{Y}$ is willingness to participate in research (available only in C2C). This is not possible using $\mathcal{X}_C$ alone and so our goal is to leverage data sets collected through random sampling ($\mathcal{X}_R$) and convenience ($\mathcal{X}_C$) to obtain valid population-based inference for $\mathcal{Y}$.

To accomplish this goal, we employ weighted estimators for samples on variables from $\mathcal{X}_C$ that are constructed to obtain population-based inference. To do this, we estimate weights, w_C, that leverage information about the differences in the sample distributions between data sets on the variables from $\mathcal{X}_C$ and $\mathcal{X}_R$. Let X, Y, and Z be samples of observations on variables from the sets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ respectively. The general approach for estimating weights is to first collect data sets, X_C and X_R on variables from $\mathcal{X}_C$ and $\mathcal{X}_R$, respectively, and then estimate a model to discriminate between the two datasets. To accomplish this we stack X_C onto X_R to a create a single combined dataset. We then construct an auxiliary variable, C, which is an indicator that an observation was in dataset X_C. Thus, C evaluates to 1 for units from X_C and 0 otherwise. We can use this stacked data set to estimate the probability of a given unit arising from each sample and use this information to weight X_C to be similar to X_R and obtain inference on variables from $\mathcal{Y}$.

Let the subscript i denote the observation for subject i. Our goal is to obtain a weighted estimator for population-based inference such that,

$$E_{\mathcal{P}|C}[w_{Ci}Y_i|C_i = 1] = E_{\mathcal{P}}[Y_i] \tag{1}$$

where $E_{\mathcal{P}}$ is the expectation taken over the distribution of the target population, $\mathcal{P}$. Now, we define,

$$P_{Ci} = \Pr(C_i = 1|X_i = x_i), \tag{2}$$

which allows us to construct weights

$$w_{Ci} \propto \frac{1 - P_{Ci}}{P_{Ci}} \tag{3}$$

so that Eq. 1 holds under appropriate assumptions on the set $\mathcal{X}$ (see Sect. 2.1). Weights will be normalized so they sum to one. Equation 2 is analogous to the propensity score from the causal inference literature (Rosenbaum and Rubin 1983) and Eq. 3 would correspond to the weights for the average treatment effect on the control (ATC) (Li et al. 2018). Thus, we refer to the estimated weights w_C as inverse propensity weights for self-selection into the convenience sample, or propensity weights for short.

2.1 Assumptions

While our focus throughout will be obtaining population inference from a convenience sample, there are many analogs between the methods here (and their assumptions) and those from the field of causal inference. Therefore, we will describe the assumptions needed for estimating population parameters through the lens of estimating causal effects and discuss how these assumptions do, and do not, apply in our context. The methods here relate most closely to those involving the "propensity score" (Rosenbaum and Rubin 1983).

A propensity score as defined by Rosenbaum and Rubin (1983) is the probability of receiving a treatment when conditioned on observed covariates. In the field of causal inference, typically three assumptions are needed for making causal conclusions using propensity scores: (1) unconfoundedness, (2) positivity, and (3) the stable unit treatment value assumption (SUTVA) (See Chapter 1 and 12 of Imbens and Rubin 2015 or Appendix 1 of Greenland et al. 1999 for an overview). First, the (1) unconfoundedness assumption requires that potential outcomes be conditionally independent of the treatment assignment given the observed covariates. Second, (2) positivity requires that each unit has a positive probability of receiving both treatment and control treatments. More formally, if T is an indicator for receiving a treatment and X are covariates, then $0 < \Pr(T = 1|X = x) < 1$ for all subjects. Finally, (3) SUTVA requires that each subjects treatment assignment does not affect any other subject's potential outcomes and there is no hidden variability in the treatment.

Versions of two of these assumptions are relevant in our context with propensity weights. First, we assume (1) unconfoundedness, that the response is independent of the selection probability conditional on the collected covariates. In other words, any covariate (or a proxy of the covariate) related to both the response or selection probability must be measured in both the representative and biased samples. We are unable to balance on unmeasured covariates. For (2) we assume that each subject must have a non-zero probability of being selected into the convenience sample, or $0 < \Pr(C = 1|X = x) < 1$. Although C2C registration is open to any adult, participants are primarily enrolled from Southern California. Although theoretically possible, the probability of people from outside of the Southern California region being sampled in the C2C is close to zero. Thus, we must assume that the relationship between race/ethnicity and research willingness does not vary by state within the US.

The SUTVA assumption (3) has less applicability in our setting. In particular, SUTVA is typically used to make a consistency argument to map potential outcomes to observed outcomes. In our setting, we assume that the response of an individual would be the same whether or not they are in the convenience sample or the random sample, and therefore the need for this assumption in our setting is diminished. More explicitly, we assume no effect of survey participation on the potential outcome of a participant and our propensity weights are being used to reweight participants back to their population prevalence based on the rates from the random sample.

In practice, it is important to carefully design the data collection to include any covariates hypothesized to be related to the sampling probability. If there are any missing covariates, accounting for the sampling bias due to measured covariates should be better than ignoring the selection mechanism completely (Masten and Poirier 2018), but the unconfoundedness assumption is not testable (Imbens and Rubin 2015). The non-zero sampling probability assumption should motivate researchers to collect participants from each subpopulation based on variables related to selection. We can upweight underrepresented subpopulations, but we are never able to learn about subpopulations that were never studied.

2.2 Dataset construction for estimating propensity weights

We want to estimate propensity weights for the convenience sample X_C with n_C observations. Let m and p be the number of variables in $\mathcal{X}$ and $\mathcal{Z}$, respectively. We include a column of 1s in the sets $\mathcal{X}$ and $\mathcal{Z}$ to be able to estimate an intercept. When using a complex survey sample, such as NHANES, as the representative sample we need to first incorporate design weights to ensure it is representative of the population of interest because certain subpopulations may be oversampled by design. Let X_S be a survey sample with n_S observations and P_{Si} denote the sampling probability for subject i in the survey sample. In NHANES, the sampling probabilities for each subject account for both the survey design and both item and subject level non-response. To obtain a representative dataset, we utilize frequency weights, $w_{si} = P_{Si}^{-1}$, which represent the number of subjects each sampled subject represents in the population and replicate each subject according to their frequency weight. To obtain the smallest representative dataset with whole numbers of subjects, we divide each frequency weight by the smallest observed weight and take the ceiling of it to obtain the number of replications: $w^*_{Si} = \text{Ceiling}(w_{Si}/\min[w_S])$. We implement the frequency weights to obtain a representative sample X_R with dimension $n_R \times m$, where each subject from X_S is replicated w^*_{Si} times for a total of $n_R = \sum_{i=1}^{n_S} w^*_{Si}$ observations. While this approach does not fully account for clustering in the NHANES dataset because information on clustering is not publically available, it is a pragmatic solution because the constructed sample will be more representative than most convenience samples thereby leading to improved inference on the target population. Additionally, clustering may slightly impact the variance of parameters, but in this project, we do not account for variability in the NHANES sampling weights.

Recall that X is the sample of variables collected in both X_R and X_C and containing observations from both representative and convenience sample subjects. Specifically, X is an $n \times m$ dimension matrix where $n = n_R + n_C$. For notational convenience, let $\mathcal{C}$ be the set of subjects from the convenience sample with $|\mathcal{C}| = n_C$ and and $\mathcal{R}$ be the set of subjects from the representative sample with $|\mathcal{R}| = n_R$. To obtain X in practice, we concatenate the convenience and representative samples for the variables in $\mathcal{X}$. We can derive the indicator for membership in the convenience sample, C, and append it to X. To estimate the propensity weights, w_C, defined in Eq. 3 we can directly estimate the probability of convenience sample membership, P_{Ci}, defined in Eq. 2. Many types of propensity weight estimation methods can be considered and we discuss their advantages and disadvantages in the following section.

2.3 Classes of propensity weight estimation methods

In this section we compare different sampling weight estimation strategies and provide examples of each that we will use as test cases. Our goal is to assess the relative performance of different estimation strategies and their strengths and weaknesses. We discuss likelihood based methods and use logistic regression as an example, covariate balancing methods with the covariate balancing propensity score and entropy balancing as examples, and algorithmic methods with random forest as an example. We will explore these four examples of methods for estimating propensity weights and their impact on covariate balance of convenience samples and on bias and variance of estimated associations.

Likelihood-based methods such as linear regression, logistic regression, probit regression, and penalized regression minimize the negative log-likelihood. We focus on logistic regression which takes the form logit $(P_{Ci}) = X_i\gamma$, where logit $(\cdot)$ is the logit function, γ is a $m \times 1$ vector of regression coefficients and X_i is the $m \times 1$ vector of observed covariates for subject i. When implementing logistic regression, we include second order terms and use forward-selection with Akaike's Information Criterion (AIC) for selecting predictors using the `step` function in the `stats` package.

By definition, propensity scores are balancing scores (Rosenbaum and Rubin 1983) and another strategy for propensity weight estimation is to directly balance covariate distributions across the two classes of a binary outcome. Covariate balancing methods, such as the covariate balancing propensity score (CBPS) and entropy balancing (EB), optimize weights by directly targeting covariate balance between the the convenience sample ($C_i = 1$) and the representative sample ($C_i = 0$).

CBPS extends the logistic regression model by incorporating additional moment balancing constraints (Imai and Ratkovic 2014). CBPS is a common method used for estimating propensity scores (see for example Yu et al. 2021; Gearan et al. 2021; Mazzinari et al. 2021). Researchers may be familiar with CBPS and interested in using it for estimating propensity weights, so we will assess how it performs relative to logistic regression. In our context, estimated propensity weights should balance the covariate distribution of the convenience sample so it matches the representative sample (Li et al. 2018). Thus, we estimate weights for the average treatment effect on the control (ATC) so that the representative sample is the reference population. The CBPS method solves the estimating equations as well as covariate balancing conditions. For the ATC, the balancing conditions are

$$E_{\mathcal{P}}\left\{ \frac{(1 - P_{Ci})C_i f(X_{ij})}{P_{Ci}} - (1 - C_i)f(X_{ij}) \right\} = 0, \tag{4}$$

where $f(X_{ij})$ is a function of X_{ij}. For example, $f(X_{ij}) = (X_{ij}; X_{ij}^2)$ would ensure the first and second moments of each covariate will be balanced. We fit the CBPS model with the `CBPS` package and included balancing constraints on second order orthogonal polynomial terms.

Entropy balancing (EB) is a non-parametric approach to weight estimation that incorporates moment balancing conditions into model selection (Hainmueller 2012). Entropy balancing allows for the inclusion of initial base weights that contain information about population prevalence. Entropy balancing estimates weights that minimize the divergence between estimated weights and base weights. If there are no base weights available, they can be treated as uniform for all units ($b_i = 1/n$). Similar to the CBPS, we want to estimate weights for the ATC where the representative sample is considered the reference group. Assuming Kullback-Leibler divergence, EB minimizes $H(w_C) = \sum_{i \in C} w_{Ci} \log(w_{Ci}/b_i)$, subject to balancing constraints,

$$\frac{1}{n_R} \sum_{i \in \mathcal{R}} x_{ij}^d = \sum_{i \in C} w_{Ci} x_{ij}^d, \text{ for } d = 1, \ldots, D,$$

where x_{ij} is the j-th covariate for subject i, $\sum w_{Ci} = 1$, and $w_{Ci} > 0$. We implemented entropy balancing with $D = 3$ using the `entbal` package for `R` (https://github.com/bvegetabile/entbal) available on GitHub (Vegetabile et al. 2021).

Unconstrained algorithmic methods, such as support-vector machines or random forest, do not require specifying a model. They predict class probabilities that minimize some out-of-sample measure of goodness of fit (e.g. prediction error) and do not assume a distribution or any moments of the response. As an example, we focus on random forest (Breiman 2001) which is a flexible model that does not require the user to specify a functional form of predictors. Random forest (RF) is an extension of classification and regression trees or CART (Breiman et al. 1984) and limits susceptibility to overfitting by introducing stochasticity. RF builds separate decision trees on bootstrapped samples and averages the prediction across trees for each subject, a technique known as bagging, and for each node of a decision tree, only a random sample of predictors are considered for splitting. To prevent extreme weight values, we trim RF estimates of the probability of convenience sample membership that are 0 or 1 and replace them estimate with 0.01 and 0.99, respectively (Lee et al. 2011). We fit RF using the `randomForest` package in `R` (Liaw and Wiener 2002).

To demonstrate the advantages and disadvantages of the different approaches, we will evaluate the performance of the different propensity weight estimation methods described above: logistic regression, covariate balancing propensity score (CBPS), entropy balancing (EB), and random forest (RF).

2.4 Quantifying uncertainty

There are several common ways to estimate the variance of coefficients from models with weights. Standard analytic variance estimates for coefficient estimates from weighted GLMs are an extension of the sandwich, or Huber-White, estimator (Freedman 2006). These design based variance estimates are included in most survey sampling software packages such as the `svyglm` function in the `survey` package (Lumley and Scott 2017). This approach assumes the propensity weights are fixed, i.e. not estimated, but uncertainty from propensity weight estimation will likely impact the uncertainty of the parameter estimates in the scientific model. If there is little uncertainty in $\widehat{w}_C$, then the design based errors used in survey sampling methodology will likely perform well. Alternatively, resampling methods, such as the bootstrap, are more computationally intensive, but can be used to account for the impact of the uncertainty in the propensity weight estimation procedure on the variance of the parameter estimates by reestimating weights within each bootstrap sample.

In this section, we derive an analytic variance estimate that accounts for uncertainty from the propensity weight estimation method. We use a simultaneous estimating equation approach for variance estimation and extend the approach of Schildcrout and Rathouz (2010). We treat both the propensity weight estimation and final scientific model as if they are being estimated simultaneously and derive the sandwich estimator for the parameters of the scientific model. Of the four propensity weight estimation methods we have considered, only the logistic model allows for a readily tractable analytic variance estimate that accounts for uncertainty in propensity weights. The design based variance estimate can be used for other weight estimation methods.

Suppose the scientific outcome model with response, Y_i, and the $p \times 1$ vector of covariates, Z_i, for subject i is $\eta_i = g(\mu_i) = z_i\beta$ where η_i is the linear predictor, $g(\cdot)$ is a link function, $\mu_i = E(Y_i|Z_i = z_i)$, and β is a $p \times 1$ vector of parameters. Recall, when applying this method to the analysis of the C2C, Z is the subset of covariates in the C2C sample needed for the final analysis, which includes race/ethnicity and adjustment variables, and Y is the willingness to participate. Assuming a representative sampling scheme, the i-th observation's contributions to the k-th element of the score equation is given by

$$U_{ki}(\beta) = \left(\frac{\partial \mu_i}{\partial \beta_k}\right)\left(\frac{Y_i - \mu_i}{V(\mu_i)}\right)$$

where $V(\mu_i) = Var(Y_i)$ and for $k = 1...p$ (Nelder and Wedderburn 1972). When using an unrepresentative sample, each subjects contribution to the score is weighted by their propensity weight, w_{Ci} as follows,

$$\overline{U}_k(\beta) = \sum_{i\in\mathcal{C}} w_{Ci} U_{ki}(\beta) = 0.$$

Define the logistic regression propensity weight model, ψ_i, as $\psi_i = \text{logit}\,(P_{Ci}) = x_i\gamma$ where γ is a $m \times 1$ vector of coefficients. Let x_i be the i-th row of X, the combined $n \times m$ matrix including covariates from the convenience sample and representative sample. The estimated propensity weights w_{Ci} are a function of the probability of convenience sample membership (Eq. 3) and sum to one. The $m \times 1$ score equation has element j,

$$T_j(\gamma) = \sum_{i\in\mathcal{C}\cup\mathcal{R}} T_{ji}(\gamma) = \sum_{i\in\mathcal{C}\cup\mathcal{R}} (C_i - P_{Ci})x_{ij} = 0, \tag{5}$$

where $j = 1...m$. We consider both the score equation for the propensity weight estimation, $T_i(\gamma)$, and the score equation for the scientific outcome model, $\overline{U}_i(\beta,\gamma) = \overline{U}_i(\beta)$. We include γ in the notation to emphasize that the score for the final scientific outcome is a function of the γ through the propensity weights as in Schildcrout and Rathouz (2010). To simplify notation, we sometimes refer to $T_i(\gamma)$ and $\overline{U}_i(\beta,\gamma)$ as T_i and $\overline{U}_i$, respectively. We combine the two estimation equations into a stacked estimating equation

$$\begin{pmatrix} \sum_{i\in\mathcal{C}\cup\mathcal{R}} T_i(\gamma) \\ \\ \sum_{i\in\mathcal{C}} \overline{U}_i(\beta,\gamma) \end{pmatrix} = 0. \tag{6}$$

Using a first order Taylor series expansion of the stacked estimating equation (Eq. 6) we obtain the variance estimate,

$$\widehat{V}_{Prop}[(\widehat{\gamma},\widehat{\beta})] = \widehat{I}^{-1}\widehat{Q}\widehat{I}^{-1}, \tag{7}$$

where the parameters have been replaced with maximum likelihood estimates. In Eq. 7, I is Fisher's information matrix under the assumed distributions of Y_i and C_i such that,

$$I = \begin{pmatrix} I_{TT} & 0 \\ I_{UT} & I_{UU} \end{pmatrix}$$

and Q is the true variance of the score, where

$$Q = \text{Var}\left(\begin{matrix} \sum_{i\in\mathcal{C}\cup\mathcal{R}} T_i(\gamma) \\ \\ \sum_{i\in\mathcal{C}} \overline{U}_i(\beta,\gamma) \end{matrix}\middle| X = x, Z = z\right) = \begin{pmatrix} E_{\mathcal{P}}[TT^T|X=x] & R^T \\ \\ R & E_{\mathcal{P}}[\overline{U}\,\overline{U}^T|Z=z] \end{pmatrix}.$$

Derivations of the components of I and Q are provided in "Appendix 1". The reader may notice similarities to the design based variance used in the survey sampling literature without the finite population correction factor (Lumley and Scott 2017). Define $\widehat{A} = \widehat{I}_{UU}$ and $\widehat{B} = \overline{UU^T}$ so the proposed variance estimator is,

$$\widehat{V}_{Prop}(\widehat{\beta}) = \widehat{A}^{-1}\widehat{B}\widehat{A}^{-1} - \widehat{A}^{-1}\widehat{I}_{UT}\widehat{I}_{TT}^{-1}\widehat{R}^T\widehat{A}^{-1}.$$

Thus the proposed variance estimate can be expressed as the standard design based variance estimator plus a correction factor. We have provided a `estweight` package available for `R` on GitHub (https://github.com/oliviabern/estweight). The `estweight` function takes a representative sample, convenience sample, and the final outcome model and provides weighted parameter estimates. If the user selects a logistic propensity weight estimation method, the function returns the proposed variance estimate, otherwise it provides standard design-based variance estimates.

3 Simulation studies

We considered the impact of our proposed weight estimation on bias of estimated associations and the accuracy of uncertainty quantification procedures through empirical simulation studies. We designed a simulation study to be similar to the analysis of Salazar et al. (2020). NHANES collects cross-sectional data on a 2-year cycle so we combined data from the 2013-2014 and 2015-2016 surveys. All simulations utilized the NHANES data, and like Salazar et al., we excluded all subjects with a reported race or ethnicity of "other" for a total of $n_S = 4,471$ subjects. All subjects had complete data on age, sex, education, race, ethnicity, medical history (high blood pressure, diabetes, kidney disease, liver disease, coronary heart disease, cancer, major depression, prescription drug use), exercise, and amount of sleep. To obtain a representative sample we replicated each observation according to their frequency weight for a final sample size of $n_R = 38,811$ observations. We refer to this representative dataset as NHANES-REP. Code for creating NHANES-REP and for reproducing the simulation study is available on GitHub (https://github.com/oliviabern/estweight_simulationstudy).

3.1 Simulation set up

To investigate the potential impact of underrepresentation in samples and estimated propensity weights on bias and variance of estimated associations we used NHANES-REP as a finite population and drew both representative and deliberately biased samples. Subjects who are Hispanic, NH Black, NH Asian, or who have lower education levels and do not exercise tend to be underrepresented in the C2C and so we generated smaller sampling probabilities for these subpopulations. We use $\mathbb{1}$ to denote an indicator variable. Let P_{Ci} be the biased sampling probability for subject i where logit $(P_{Ci}) = \psi_i$ with

$$\begin{aligned}\psi_i = {} & .15\mathbb{1}_{Female,i} + .25\mathbb{1}_{HighSchool,i} + .1\mathbb{1}_{<HighSchool,i} + .4\mathbb{1}_{SomeCollege,i} \\ & + .85\mathbb{1}_{Hispanic,i} + .45\mathbb{1}_{NHAsian,i}\mathbb{1}_{NHAsian,i}\mathbb{1}_{SomeCollege,i} \\ & + .05\mathbb{1}_{NHBlack,i} + .75\mathbb{1}_{NHBlack,i}\mathbb{1}_{Exercise,i} - .001Age_i^2 + 4.\end{aligned}$$

Within each simulation, we drew a representative simple random sample of size 500 and a biased sample of size 500 with sampling probabilities P_{Ci}. We simulated $Y_i \sim$ Bernoulli(μ_i) with logit (μ_i) = η_i and

$$\eta_i = 1 + log(2)\mathbb{1}_{Hispanic,i} - log(3)\mathbb{1}_{NHAsian,i} + log(1.5)\mathbb{1}_{NHBlack,i} - log(2)P_{Ci} \\ + log(2)\mathbb{1}_{Hispanic,i}P_{Ci} + log(4)\mathbb{1}_{NHAsian,i}P_{Ci} - log(3)\mathbb{1}_{NHBlack,i}P_{Ci}.$$

We estimated propensity weights for subjects in the biased sample with each of the four propensity weight estimation methods described in Sect. 2.3. Similar to our applied example, we were interested in a model of the the marginal relationship between race/ethnicity where,

$$\text{logit }(\Pr[Y_i = 1|X_i = x_i]) = \beta_0 + \beta_1\mathbb{1}_{Hispanic,i} + \beta_2\mathbb{1}_{NHAsian,i} + \beta_3\mathbb{1}_{NHBlack,i}.$$

For each simulation, we (1) fit the above model in the representative sample with the objective of obtaining a similar estimate using a biased sample. (2) We then fit the model in the biased sample without any weighting, (3) with the true propensity weights ($w_{Ci} \propto P_{Ci}^{-1}(1 - P_{Ci})$), and (4) with the estimated propensity weights from each of the four estimation methods and compared the estimates to those obtained in the representative sample. For CBPS and EB, we balanced continuous variables on the first and second moments such that $f(X_{ij}) = (X_{ij}, X_{ij}^2)$ for CBPS and $d = 2$ for EB. For the logistic regression model, we included second-order terms in the model scope and used stepwise AIC for variable selection. Note that we did not include interactions in the model scope so even though the data generating model is logistic, we were unable to correctly specify it. We computed and compared analytic and bootstrap estimates of the standard error to the empirical Monte Carlo standard error. To prevent under-representation of small subpopulations in bootstrap samples, we used race/ethnicity as a stratification variable for sampling. When stratifying the bootstrap sample failed to provide adequate representation of subpopulations leading to extreme weights and inestimable coefficients, we removed the parameter estimates from bootstrap variance estimates. We conducted 1,000 simulations and used 200 bootstrap samples within each simulation.

3.2 Simulation results

3.2.1 Coefficient estimates

The average log odds ratios for representative and biased samples for each of the propensity weight types are shown in Fig. 1. The goal of incorporating estimated weights is to match estimates fit using a biased sample (columns 2-7 of the tabulated results) to those estimated using a representative sample (column 1). Note that estimates derived from a biased sample that fail to account for the sampling scheme (i.e. no weighting, column 2) did not match those from the representative sample. Incorporating both true (column 3) and estimated (columns 4-7) propensity weights allowed us to match the representative sample estimates. The type of propensity weight model did not have an appreciable impact as all of the weighted estimates did not vary much in comparison to the representative sample and to each other. Logistic regression performed well and allowed us to obtain weighted estimates fit in a biased samples that matched the association in the target population. It is a practical choice because it is parsimonious and easy to implement. It is important to note that the true sampling probabilities were generated from a logistic model, but the model

Mean Estimated Log Odds Ratios

	SRS	Biased Sample					
		None	True	Log.	CBPS	EB	RF
Intercept	0.64 (0.11)	0.56 (0.11)	0.65 (0.14)	0.64 (0.14)	0.63 (0.13)	0.64 (0.15)	0.67 (0.22)
Hispanic	1.09 (0.42)	1.19 (0.36)	1.10 (0.39)	1.12 (0.43)	1.12 (0.42)	1.12 (0.45)	1.09 (0.51)
NH Asian	−0.30 (0.50)	−0.17 (0.46)	−0.30 (0.52)	−0.29 (0.55)	−0.28 (0.53)	−0.29 (0.57)	−0.30 (0.63)
NH Black	−0.41 (0.31)	−0.48 (0.28)	−0.41 (0.33)	−0.39 (0.34)	−0.40 (0.33)	−0.39 (0.36)	−0.45 (0.43)

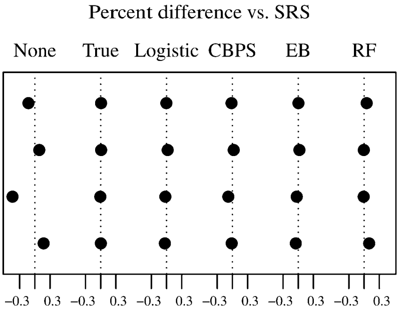


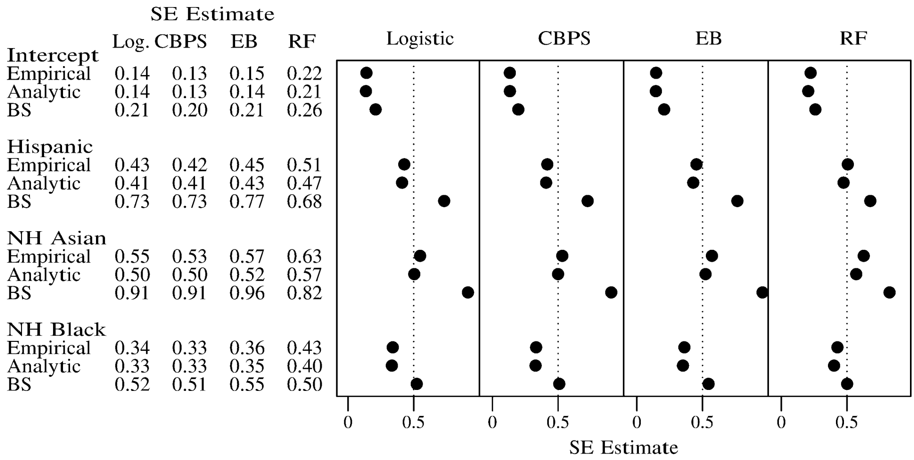
Fig. 1 Results for the simulation study described in Sect. 3.1. The average estimated log odds ratios (empirical standard errors) are presented in the table on the left. Estimates for the marginalized model fit in the simple representative sample (SRS) are in column 1 and estimates fit in a biased sample along with different types of propensity weights are in the other columns. Results compare models fit in a biased sample that do not include weights (None), incorporate the true propensity weights (True), or incorporate propensity weights estimated with a logistic (Log.), covariate balancing propensity score (CBPS), entropy balancing (EB), or random forest (RF) approach. Percent bias comparing average estimates fit in a biased sample to estimates fit in a simple random sample (SRS) are presented in the figure on the right

included interactions that were not in the scope of the logistic model we used for estimation. In this example, incorporating estimated propensity weights was an effective method for obtaining inference on the target population.

3.2.2 Uncertainty estimates

In this simulation, empirical standard errors were larger for weighted estimates fit in biased samples compared to unweighted estimated fit in a representative sample (Fig. 1). We also investigated the impact of the propensity weight estimation method on uncertainty and the performance of analytic and bootstrap variance estimates. Standard error estimates for the four different propensity weight estimation methods are reported in Fig. 2. For the analytic variance estimate, we used the proposed analytic standard error estimate ($\widehat{V}_{Prop}$) when using weights estimated via logistic regression. When using weights estimated with CBPS, EB, and RF methods we used the design based errors from the `survey package` that ignore the propensity weight estimation (Lumley and Scott 2017). The average analytic standard error estimate was generally comparable to the empirical standard error across simulations, but the bootstrap generally overestimated the true uncertainty. The design based and proposed variance estimates resulted in similar standard error estimates. Compared to the bootstrap estimate, analytic estimates more closely approximated the empirical standard error even though they did not account for uncertainty from variable selection when estimating propensity weights. Additionally, they were computationally easier than the bootstrap.

In 21 out of 1,000 simulations, coefficients were not estimable in some bootstrap samples due to extreme values of estimated weights. There was one bootstrap sample in two simulations where the association was inestimable when weights were estimated using EB. This was slightly more common when weights were estimated using RF–out of 1,000 simulations, 19 simulations had at most 5 bootstrap sample that were unable to estimate the association. We hypothesize that this occurs due to uniquely sparse bootstrap samples



	SE Estimate			
	Log.	CBPS	EB	RF
Intercept				
Empirical	0.14	0.13	0.15	0.22
Analytic	0.14	0.13	0.14	0.21
BS	0.21	0.20	0.21	0.26
Hispanic				
Empirical	0.43	0.42	0.45	0.51
Analytic	0.41	0.41	0.43	0.47
BS	0.73	0.73	0.77	0.68
NH Asian				
Empirical	0.55	0.53	0.57	0.63
Analytic	0.50	0.50	0.52	0.57
BS	0.91	0.91	0.96	0.82
NH Black				
Empirical	0.34	0.33	0.36	0.43
Analytic	0.33	0.33	0.35	0.40
BS	0.52	0.51	0.55	0.50

Fig. 2 Standard error estimates for the simulation study described in Sect. 3.1. The different standard error estimates (empirical, analytic and bootstrap) are reported for the coefficient estimates in the marginalized model fit in a biased sample. Standard error estimates are reported for models implementing propensity weights estimated using a logistic (Log.), covariate balancing propensity score (CBPS), entropy balancing (EB) or random forest (RF) method. Details of the standard error calculations are reported in Sect. 2.4

with little to no representation of some subpopulations, but this does not occur for logistic regression or CBPS. The RF method draws bootstrap samples to fit each tree and this bootstrap within a bootstrap may lead to subpopulations without any variation in the response, and thus extreme weights. Entropy balancing targets covariate balance which can be difficult if certain subpopulations are only observed in either the representative or convenience sample. Although CBPS also targets covariate balance, we did not observe any extreme weights and we hypothesize the focus on maximizing the likelihood may prevent this. Although this occurs infrequently, it may arise in practice. We excluded any bootstrap samples with insufficient information for estimating the association for a given sampling weight estimation method from the bootstrapped estimate of the variance.

4 Application to the C2C willingness analysis

The work of Salazar et al. (2020) investigated differences in research willingness by race and ethnicity using logistic regression. The methods are described in detail in that paper and we summarize them here. They performed logistic regression models to assess racial/ethnic differences in willingness to participate in research in participants 50 years or older. They separately evaluated 9 different responses: willingness to be contacted about studies involving (1) physical activity/diet modification, (2) cognitive testing, (3) magnetic resonance imaging (MRI), (4) positron emission tomography (PET) scans, (5) blood draws, (6) approved medications, (7) investigational medications, (8) lumbar punctures and (9) autopsy. They adjusted for age, sex, educational attainment, number of comorbidities, number of medications, cognitive function instrument score (Amariglio et al. 2015; Walsh et al. 2006) and research attitudes questionnaire score (Rubright et al. 2011) and used multiple imputation to handle missing data. C2C data is updated as more participants enroll and can be requested at https://c2c.uci.edu/request-c2c-data/. When replicating this analysis, we started with the same dataset of C2C participants, but performed our own multiple imputation.

4.1 Identifying matching covariates

We first identified covariates likely to modify the relationship between race/ethnicity and willingness to participate in research and were collected in both the C2C and NHANES. Some covariates were recorded with differing levels of granularity in the two datasets so we collapsed them into comparable subgroups. For example, the question regarding exercise was phrased differently in the two datasets. NHANES participants were asked if they participated in vigorous or moderate recreational activities in a typical week for at least 10 minutes.The C2C participants were asked if they participated in the following activities for at least 15 minutes/day at least once/week for the last year: walking, hiking, biking, aerobics, calisthenics, swimming, water aerobics, weight training, stretching, or another form of exercise. We decided to exclude the question about walking for the C2C subjects because there were high agreement rates for this question and we were concerned participants may have reported walking for purposes other than recreational exercise. To compare across groups, we created an indicator for exercise. We used NHANES participants who were 50 years or older to match inclusion criteria from Salazar et al. (2020). In total, we included 14 variables: age, sex, education level [Educ.] (less than 12 years [< 12], high school/GED [12], some college [12-16], college graduate [16]), race/ethnicity (NH White, Hispanic, NH Asian, NH Black), high blood pressure (BP), kidney disease, liver disease, congestive heart failure (CHD), past cancer diagnosis, major depression, average hours of sleep per night, prescription medicine use (Presc. meds), and exercise.

Excluding subjects with a reported race or ethnicity of “other" resulted in $n_C = 2,749$ observations in the C2C and excluding 917 NHANES participants with missing data (out of $n_S = 5,605$) resulted in $n_R = 38,811$ observations in NHANES-REP. Most of the missingness in the C2C was confined to a few covariates: 576 subjects were missing sleep, 207 were missing prescription drugs, 167 were missing exercise, and 29 were missing history of cancer. Matching covariates for each dataset are summarized in Fig. 3. Continuous covariates were summarized by mean (standard deviation) and the proportion was reported for categorical variables. Weighted sample statistics using estimated propensity weights for the C2C dataset were also presented. The propensity weights were estimated using logistic regression, CBPS, EB, and RF with one imputed C2C dataset. Logistic regression, CBPS, and EB balance the covariate distributions well, but RF weighted estimates are similar to unweighted ones.

4.2 Estimating propensity weights

We repeated the analysis performed by Salazar et al. and imputed 5 C2C datasets and used Rubin’s rules to aggregate across datasets (Rubin 1987). Within each dataset we estimated propensity weights for each subject using logistic regression, CBPS, EB, and RF. We fit the outcome models that quantify the relationship between race/ethnicity and each of the 9 outcomes and incorporated the estimated propensity weights. We report the estimated odds ratios (OR) and 95% confidence intervals using the analytic variance estimates for each racial/ethnic group for each of the 9 responses. We report the full results with no weighting and each of the four types of propensity weights in “Appendix 2” (Table 1). A selection of these results are depicted using forest plots to discuss the impact of weighting.

	NHANES –REP	C2C				
		None	Log.	CBPS	EB	RF
Age	63.6	65.9	62.4	62.5	63.6	64.2
	(9.3)	(8.7)	(9.2)	(9.2)	(9.3)	(9.7)
Sleep (Hr)	7.4	6.7	7.2	7.2	7.5	7.2
	(1.4)	(1.5)	(1.2)	(1.2)	(1.4)	(1.0)
Educ: < 12	0.16	0.01	0.12	0.12	0.16	0.02
Educ: 12	0.22	0.07	0.23	0.23	0.22	0.10
Educ: 12–16	0.31	0.19	0.31	0.31	0.31	0.26
Educ: 16	0.31	0.73	0.34	0.34	0.31	0.62
NH White	0.75	0.87	0.74	0.75	0.75	0.88
Hispanic	0.11	0.07	0.13	0.13	0.11	0.06
NH Asian	0.04	0.05	0.04	0.04	0.04	0.04
NH Black	0.11	0.01	0.08	0.08	0.11	0.02
Female	0.53	0.62	0.49	0.49	0.53	0.59
High BP	0.52	0.34	0.50	0.50	0.52	0.37
Diabetes	0.19	0.10	0.17	0.17	0.19	0.06
Kidney	0.05	0.03	0.04	0.04	0.05	0.01
Liver	0.05	0.02	0.06	0.06	0.05	0.01
CHD	0.09	0.06	0.09	0.09	0.09	0.03
Cancer	0.20	0.32	0.24	0.24	0.21	0.25
Exercise	0.46	0.83	0.51	0.51	0.48	0.75
Depression	0.10	0.12	0.07	0.07	0.10	0.04
Presc. meds	0.79	0.78	0.75	0.76	0.78	0.78

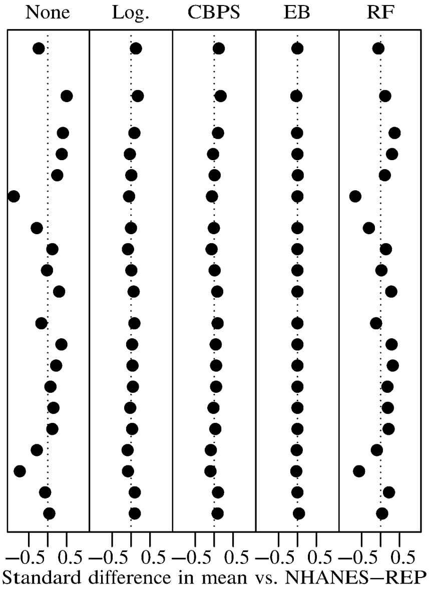


Fig. 3 Covariates were summarized as mean (standard deviation) for continuous variables and as proportions for categorical variables in the table on the left for NHANES-REP (38,811 observations), the unweighted C2C (2,749 subjects) dataset (None), and weighted C2C datasets. Weights were estimated using a logistic (Log.), covariate balancing propensity score (CBPS), entropy balancing (EB), or random forest (RF) method. Continuous covariates were summarized by mean (standard deviation) and categorical covariates by proportion. Standardized difference in means relative to NHANES-REP are presented in the figure on the right (Stuart 2010). Propensity weights were estimated with missing values in the C2C imputed once. Covariates are described in Sect. 4.1

Across all forest plots (Fig. 4), we observed that the standard errors increased with weighting, but this added variance better reflects the true uncertainty in the estimates and their ability to generalize to an external population. For example, the odds ratio comparing Hispanics to NH Whites for willingness to be contacted about studies with lumbar puncture (LP) had a noticeably wider confidence interval for the weighted estimates. The C2C underrepresents Hispanic subjects relative to the US population and the wider confidence intervals reflect this lack of information on the subpopulation. Several statistically significant odds ratios were no longer significant after incorporating estimated propensity weights. NH Asians had significantly higher odds of being willing to be contacted about studies involving LP compared to NH Whites in the original analysis, but this relationship was no longer statistically significant after weighting. Salazar et al. (2020) found it surprising that NH Asians would be more willing to undergo an LP because previous studies had found them less willing relative to NH Whites (Moulder et al. 2017). They speculated that NH Asians in the C2C had been exposed to more education about the LP procedure through outreach events for older Chinese adults. Accounting for sampling bias with estimated weights has attenuated this relationship to the null which aligns with previous findings.

The models using logistic and CBPS estimated weights tended to have similar estimates and confidence intervals. The CBPS model uses a logistic regression model but incorporates

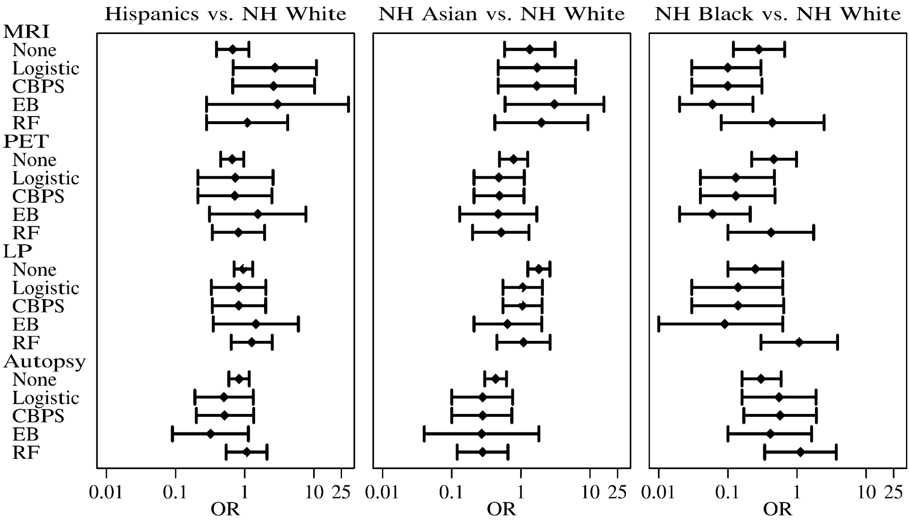


Fig. 4 Forest plots of the estimated odds ratios (OR) and 95% confidence intervals for the racial ethnic differences analysis with MRI, PET, lumbar puncture (LP) and autopsy as the response. Results are presented for unweighted analysis (None) along with the weighted analysis using propensity weights estimated with logistic, covariate balancing propensity score (CBPS), entropy balancing (EB), and random forest (RF) methods

moment balancing conditions into the model fitting. These additional constraints did not impact the final result substantially when compared to the standard logistic regression derived estimates. The estimates using RF and EB weights had high variability and differed from the results using logistic and CBPS weights. Additionally, the point estimates from the models using RF weights tended to differ the most from the other 3 weighted models. Random forest is unique because it is both non-parametric and does not target covariate balancing. An advantage of decision trees is they naturally include interactions in modeling, but in sparse data with little representation of subpopulations this can lead to increased variability. Weight trimming, where estimated probabilities of 0 or 1 were replaced with 0.01 and 0.99, may have also impacted bias and variance estimates as well as the population of inference (Lee et al. 2011).

Incorporating estimated propensity weights not only impacted the uncertainty, it also changed the direction of point estimates. For example, in the original analysis Hispanics had lower odds of being willing to be contacted about studies that involve an MRI scan, but most of the weighted point estimates suggested Hispanics may actually have higher odds. The results for the original analysis were close to being statistically significant but the weighted models showed little evidence of an effect which leads to a different interpretation of the results. The four weighted point estimates were not covered by the unweighted confidence interval.

5 Discussion

Our results demonstrated the utility of using convenience samples in concert with a representative sample to estimate weights that can be used to estimate population representative parameters of interest. Convenience samples are widely available and often used in research studies, but failing to account for the selection mechanism can lead to biased estimates and underestimation of the true variance of estimates. It is important to carefully select a target population and design studies and analyses that generalize to this population. If researchers are not able to obtain a representative sample because of ethical or practical considerations, they are forced to use a convenience sample. Since estimated propensity weights can only balance a convenience sample on observed covariates, researchers must take care to collect any covariates that they hypothesize are related to both the outcome and sampling probability. Additionally, any subpopulation with a convenience sample membership probability of zero and thus not represented in the convenience sample cannot be included in the target population. Researchers must carefully consider which covariates should be collected and which subpopulations are being sampled into a convenience sample to allow for valid estimates of associations in the desired target population.

In the analysis of racial/ethnic differences of research willingness, weighted confidence intervals were generally at least twice as wide as unweighted confidence intervals. Incorporating propensity weights can increase the variability of parameter estimates because subjects with a low estimated probability of convenience sample membership will have large estimated weights and undue influence on the estimated associations (Little and Rubin 2014). Although propensity weights may increase variance, they can reduce bias of the estimated association in the population of interest. Using an unrepresentative sample provides less information about the target population, and thus the increased variance of our estimator reflects this uncertainty (Lumley 2010). To measure how much the sampling mechanism impacts efficiency, we can compute the design effect, which is the ratio of the variance of the parameter estimate in an unrepresentative sample compared to a simple random sample (Lumley 2010; Kish 1965). When comparing empirical variances from the simulation study for logistic-weighted coefficient estimates fit in a biased sample compared to unweighted estimates in a SRS, the design effect ranges from 1 to 1.6. Thus, we will need a biased sample that is up to 1.6 times bigger than a simple random sample to obtain the same variance.

For estimating the variance of parameter estimates in the outcome model, we compared a resampling approach and our proposed analytic approach for a logistic regression model that accounts for uncertainty arising from the propensity weight estimation. In empirical studies, we found the proposed analytic estimates performed better than the bootstrap estimate even though the analytic estimate does not account for the model selection when estimating propensity weights. Surprisingly, the bootstrap estimate tends to overestimate the uncertainty even though we replicated the estimation method within each bootstrap sample. Perhaps there was more variability in estimated weights within bootstrap samples than within the full simulated data set. Previous work on implementing entropy balancing weights (Vegetabile et al. 2021) and exact matching using the propensity score (Austin and Stuart 2015) also reported conservative bootstrap variance estimates, but the stratified double bootstrap where units are resampled from the survey sample and convenience sample used by Ackerman et al. (2021) provides similar variance estimates to a design-based approach. The proposed analytic variance for model-based propensity weight estimation methods accounts for uncertainty in the estimated weights. The design based standard errors, however, perform well even though they fail to account for the propensity weight estimation process. We suggest using the proposed variance estimator with a model based propensity weight estimation procedure because it may perform better than the design based estimate, but the design based approach should perform well if needed. In our context, the two methods did not diverge substantially, but they could if there is high variability in the propensity weights. It is possible to derive an analytic variance estimator that accounts for the weight estimation for the CBPS model. However, it is made difficult because the CBPS model is overspecified and fit using generalized method of moments (Imai and Ratkovic 2014; Hansen 1982). One might be able to incorporate the final scientific model into the CBPS model as an additional balancing constraint for a simultaneous estimation approach. This may be an interesting area of research to pursue.

All four propensity weight estimation models decreased bias in the simulation study. Algorithmic propensity weight estimation methods are very flexible, but random forest provided the smallest degree of bias reduction and the largest variance in the simulation study. Using RF-derived weights provided the poorest covariate balance (Fig. 3). Models that incorporate covariate balancing into model-selection help ensure covariate balance in the biased sample. EB, however, scales better than forward step-wise model selection for logistic regression, but CBPS tends to be slower due to the additional constraints. Estimates using CBPS did not deviate substantially from those using logistic regression, so the additional balancing constraints did not improve performance. We used the default settings for the `CBPS` package and users can change the settings to focus more on covariate balance. Although EB balanced covariates better than the logistic model in the applied example, they both reduced bias in estimated associations to the same extent in simulation study. Likelihood-based regression models, such as logistic regression, allow for an analytic variance estimate that fully accounts for uncertainty. It can also be easily expanded to include interaction and smoothers to allow for greater flexibility, but the second order terms seemed to perform well enough in our experiments. In practice, we suggest using logistic regression because it effectively reduces bias under our assumptions, is familiar to many scientists, is broadly accessible in different statistical software packages, and allows for an analytic variance estimate that accounts for uncertainty from estimation of propensity weights. The prediction models we implemented are only several examples of many different options one could use. Practitioners can select their preferred prediction model.

NHANES is a practical choice for generalizing biomedical studies to the US population. Different research areas may, however, collect variables that are not recorded in NHANES but are believed to be strongly related to the sampling probability. Other national surveys collect different variables and may be more relevant to different research areas. Other examples of national surveys are the American Community Housing Survey which collects population and housing information, the Behavioral Risk Factor Surveillance System which conducts health-related telephone interviews, the General Social Survey which studies American society, and the Current Population Survey that collects labor force statistics. Additionally, researchers may want to generalize to a population outside of the US. If the target population is a subset of the US population, NHANES can be subset and used as the representative sample. Otherwise, other representative samples need to be obtained. Researchers can consider census data if available, government sponsored national surveys, or international surveys. After specifying the target population, one should consider which samples are most representative and accessible.

Overall, estimated propensity weights reduce bias on parameter estimates from unrepresentative sampling. We, of course, are unable to account for any unmeasured covariates that may contribute to selection bias. Additionally, we are unable to learn about subpopulations that were never sampled. For example, if there are no NH Blacks with less than a high school education in the C2C, we cannot weight this missing subpopulation. We collapsed different variables to match across different datasets and were unable to empirically evaluate if these are equivalent definitions. Implementing estimated propensity weights increases the uncertainty of estimates but this reflects the information available on target population parameters.

Convenience samples are easily collected and are used for research in many disciplines. The NHANES dataset is a rich, open access dataset that will likely have many overlapping covariates with convenience samples. Estimated propensity weights using NHANES is practical and effective at addressing selection bias concerns in convenience samples when trying to generalize to the non-institutionalized US population.

Appendix 1: Derivation of variance estimator

In this section we derive the various components of I and Q introduced in Sect. 2.4. Using iterated expectations we can show the cross-term R is

$$R = E_{\mathcal{P}}\Big[\sum_{i\in\mathcal{C}} \overline{U}_i\Big(\sum_{i\in\mathcal{C}} T_i^T + \sum_{i\in\mathcal{R}} T_i^T)\Big)\Big|Z = z\Big] = E_{\mathcal{P}}\Big[\sum_{i\in\mathcal{C}} \overline{U}_i \sum_{i\in\mathcal{C}} T_i^T\Big|Z = z\Big].$$

So the method of moments estimator for Q is

$$\hat{Q} = \begin{pmatrix} TT^T & \hat{R}^T \\ \hat{R} & \overline{U}\,\overline{U}^T \end{pmatrix}\Bigg|_{\begin{pmatrix}\beta\\ \gamma\end{pmatrix}=\begin{pmatrix}\hat{\beta}\\ \hat{\gamma}\end{pmatrix}}$$

with $\hat{R} = \sum_{i\in\mathcal{C}} \overline{U}_i \sum_{i\in\mathcal{C}} T_i^T$.

Now consider the the terms of I. I_{TT} is the derivative of T with respect to γ which is Fisher's information matrix for logistic regression,

$$I_{TT} = \sum_{i\in\mathcal{C}\cup\mathcal{R}} E_{\mathcal{P}}\Big(-\frac{\partial T_i}{\partial\gamma}\Big|X_i = x_i\Big) = \sum_{i\in\mathcal{C}\cup\mathcal{R}} \Big(x_i^T\left(P_{Ci}(1 - P_{Ci})\right)x_i\Big).$$

I_{UU} is similar,

$$I_{UU} = -\sum_{i\in\mathcal{C}} E_{\mathcal{P}}\Big[w_{Ci}\frac{\partial U_j}{\partial\beta_k}\Big|Z = z\Big] = z^T M(\beta) z$$

where

$$M(\beta) = diag\Big(\frac{w_{Ci}(\partial\eta_i/\partial\mu_i)^{-2}}{V(\mu_i|Z_i = z_i)}\Big).$$

Finally, I_{UT} is

$$I_{UTkj} = -\sum_{i\in\mathcal{C}\cup\mathcal{R}} E_{\mathcal{P}}\Big[\frac{\partial U_{ki}}{\partial\gamma_j}\Big|Z_i = z_i\Big] = \sum_{i\in\mathcal{C}\cup\mathcal{R}} E_{\mathcal{P}}\Big[w_{Ci}\Big(\frac{\partial\mu_i}{\partial\beta_j}\Big)\Big(\frac{Y_i - \mu_i}{V(\mu_i)}\Big)X_{ij}\Big|Z_i = z_i\Big].$$

This term does not equal zero, because X_{ij} is not fixed. We can rearrange $I^{-1}QI^{-1}$ to point out the relationship to the design based variance used in the survey sampling literature (Lumley and Scott 2017). Define $\widehat{A} = \widehat{I}_{UU}$ and $\widehat{B} = \overline{UU}^T$ so that,

$$\widehat{Q} = \begin{pmatrix} T \\ \overline{U} \end{pmatrix} \begin{pmatrix} T^T \overline{U}^T \end{pmatrix} = \begin{pmatrix} TT^T & \widehat{R}^T \\ \widehat{R} & \widehat{B} \end{pmatrix}.$$

Using the formula for blockwise inversion (Fact 2.17.1 in Bernstein 2009),

$$I^{-1} = \begin{pmatrix} I_{TT} & 0 \\ I_{UT} & A \end{pmatrix}^{-1} = \begin{pmatrix} I_{TT}^{-1} & 0 \\ -A^{-1}I_{UT}I_{TT}^{-1} & A^{-1} \end{pmatrix}.$$

Combining I^{-1} and Q we obtain the proposed variance estimator,

$$\widehat{V}_{Prop}(\widehat{\beta}) = \widehat{A}^{-1}\widehat{B}\widehat{A}^{-1} - \widehat{A}^{-1}\widehat{I}_{UT}\widehat{I}_{TT}^{-1}\widehat{R}^T\widehat{A}^{-1}.$$

Appendix 2: Bias adjusted C2C results

See Table 1.

Author's contribution OMB, DLG, and BGV contributed to the conception and design of the methodology. Data preparation was performed by OMB, CRS, and JDG. OMB implemented the methodology, wrote the software package, and wrote the first draft of the manuscript. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The C2C Registry was made possible through a philanthropic gift from HCP, Inc and grants P30 AG066519 and UL1 TR000153. OMB was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-183928 and the ARCS foundation. BGV was supported by the National Institute on Drug Abuse of the National Institutes of Health under award number R01DA045049. CRS was supported by the National Institute on Aging of the National Institute of Health under a diversity supplement to award AG059407 and an Alzheimer's Association research fellowship AARFD-20-682432. JDG and DLG were supported by the National Institutes of Health under award P30AG066519. DLG was also supported by the National Institute of Health under award R01AG053555. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data availability NHANES data is publically available online at https://www.cdc.gov/nchs/nhanes/index.htm and C2C data can be requested at https://c2c.uci.edu/request-c2c-data/.

Code availability The proposed method can be implemented in `R` using our package `estweight` available on GitHub (https://github.com/oliviabern/estweight). Code for recreating the simulation study and cleaning NHANES data is available at https://github.com/oliviabern/estweight_simulationstudy.

Declaration

Conflict of Interest No potential conflics of interest are reported by the authors.

Table 1 Bias adjusted C2C results: Odds Ratios and 95% confidence intervals are presented for the models from Salazar et al. (2020) assessing the relationship between race/ethnicity and 9 responses with adjustment variables. The models were fit without any propensity weights and with propensity weights estimated using logistic regression, covariate balancing propensity score (CBPS), entropy balancing (EB), and random forest (RF) methods

Trial type	Model	Hispanic	NH Asian	NH Black
Physical activity/diet modification	Unweighted	1.06 (0.52, 2.16)	0.68 (0.34, 1.35)	1.87 (0.25, 13.95)
	Logistic	1.52 (0.40, 5.83)	0.82 (0.23, 2.97)	4.54 (0.45, 45.38)
	CBPS	1.52 (0.40, 5.76)	0.83 (0.23, 3.02)	4.64 (0.46, 46.28)
	EB	4.80 (0.54, 42.33)	0.35 (0.06, 2.18)	4.61 (0.31, 69.07)
	RF	0.74 (0.16, 3.50)	0.87 (0.19, 4.09)	3.93 (0.36, 43.02)
Cognitive testing	Unweighted	0.50 (0.22, 1.12)	0.52 (0.18, 1.52)	0.71 (0.09, 5.55)
	Logistic	0.73 (0.13, 4.26)	0.47 (0.06, 3.89)	0.95 (0.09, 10.14)
	CBPS	0.72 (0.13, 4.04)	0.45 (0.05, 3.72)	0.99 (0.09, 10.66)
	EB	1.75 (0.09, 33.19)	0.91 (0.03, 23.98)	0.93 (0.08, 11.60)
	RF	0.27 (0.06, 1.27)	0.88 (0.15, 5.36)	1.41 (0.13, 15.37)
MRI scans	Unweighted	0.67 (0.39, 1.15)	1.34 (0.58, 3.12)	0.28 (0.12, 0.66)
	Logistic	2.73 (0.68, 10.90)	1.71 (0.47, 6.22)	0.10 (0.03, 0.30)
	CBPS	2.61 (0.67, 10.25)	1.70 (0.47, 6.17)	0.10 (0.03, 0.31)
	EB	2.99 (0.28, 31.66)	3.06 (0.59, 15.86)	0.06 (0.02, 0.23)
	RF	1.09 (0.28, 4.19)	1.98 (0.42, 9.35)	0.44 (0.08, 2.47)
PET scans	Unweighted	0.66 (0.45, 0.97)	0.78 (0.49, 1.25)	0.46 (0.22, 0.98)
	Logistic	0.73 (0.21, 2.57)	0.48 (0.21, 1.12)	0.13 (0.04, 0.47)
	CBPS	0.72 (0.21, 2.47)	0.49 (0.21, 1.11)	0.13 (0.04, 0.48)
	EB	1.55 (0.31, 7.65)	0.47 (0.13, 1.70)	0.06 (0.02, 0.21)
	RF	0.81 (0.34, 1.94)	0.52 (0.20, 1.31)	0.42 (0.10, 1.74)
Blood draws	Unweighted	0.62 (0.35, 1.10)	0.31 (0.18, 0.53)	0.27 (0.11, 0.67)
	Logistic	1.62 (0.45, 5.75)	0.37 (0.15, 0.89)	0.70 (0.15, 3.20)
	CBPS	1.58 (0.45, 5.54)	0.37 (0.15, 0.88)	0.70 (0.16, 3.16)
	EB	4.17 (0.69, 25.15)	0.24 (0.05, 1.07)	0.61 (0.12, 3.01)
	RF	0.77 (0.17, 3.47)	0.16 (0.04, 0.60)	0.29 (0.05, 1.61)
Approved medications	Unweighted	0.68 (0.42, 1.10)	0.61 (0.36, 1.01)	0.67 (0.25, 1.80)
	Logistic	0.17 (0.06, 0.49)	0.66 (0.29, 1.50)	0.72 (0.21, 2.40)
	CBPS	0.17 (0.06, 0.49)	0.65 (0.29, 1.49)	0.72 (0.21, 2.41)
	EB	0.13 (0.02, 0.79)	0.43 (0.08, 2.44)	0.81 (0.10, 6.69)
	RF	0.64 (0.19, 2.17)	0.44 (0.15, 1.29)	0.45 (0.14, 1.45)
Investigational medications	Unweighted	0.62 (0.42, 0.90)	0.55 (0.36, 0.83)	0.52 (0.24, 1.11)
	Logistic	0.31 (0.10, 0.93)	0.33 (0.14, 0.77)	0.70 (0.25, 1.98)
	CBPS	0.31 (0.10, 0.93)	0.33 (0.14, 0.76)	0.71 (0.25, 1.98)
	EB	0.50 (0.07, 3.73)	0.29 (0.08, 1.04)	0.82 (0.25, 2.75)
	RF	0.67 (0.26, 1.67)	0.40 (0.15, 1.08)	0.36 (0.11, 1.15)
Lumbar puncture	Unweighted	0.95 (0.70, 1.30)	1.82 (1.26, 2.63)	0.25 (0.10, 0.62)
	Logistic	0.82 (0.33, 2.02)	1.06 (0.55, 2.05)	0.14 (0.03, 0.62)
	CBPS	0.82 (0.34, 2.01)	1.05 (0.55, 2.03)	0.14 (0.03, 0.64)
	EB	1.45 (0.35, 5.97)	0.64 (0.21, 2.00)	0.09 (0.01, 0.62)
	RF	1.27 (0.64, 2.50)	1.09 (0.45, 2.64)	1.07 (0.30, 3.85)

Table 1 (continued)

Trial type	Model	Hispanic	NH Asian	NH Black
Autopsy	Unweighted	0.83 (0.59, 1.16)	0.43 (0.30, 0.62)	0.30 (0.16, 0.59)
	Logistic	0.50 (0.19, 1.33)	0.28 (0.10, 0.76)	0.55 (0.16, 1.88)
	CBPS	0.51 (0.20, 1.35)	0.28 (0.10, 0.74)	0.57 (0.17, 1.90)
	EB	0.32 (0.09, 1.13)	0.27 (0.04, 1.82)	0.41 (0.10, 1.62)
	RF	1.07 (0.54, 2.10)	0.28 (0.12, 0.65)	1.12 (0.34, 3.70)

References

Ackerman, B., Lesko, C.R., Siddique, J., et al.: Generalizing randomized trial findings to a target population using complex survey population data. Stat. Med. **40**(5), 1101–1120 (2021)

Amariglio, R.E., Donohue, M.C., Marshall, G.A., et al.: Tracking early decline in cognitive function in older individuals at risk for Alzheimer's disease dementia: the Alzheimer's disease cooperative study cognitive function instrument. JAMA Neurol **72**(4), 446–454 (2015). https://doi.org/10.1001/jamaneurol.2014.3375

Austin, P.C., Stuart, E.A.: Estimating the effect of treatment on binary outcomes using full matching on the propensity score. Stat. Methods Med. Res. **26**(6), 2505–2525 (2015). https://doi.org/10.1177/0962280215601134

Bailey, L.C., Milov, D.E., Kelleher, K., et al.: Multi-institutional sharing of electronic health record data to assess childhood obesity. PLoS ONE **8**(6), e66,192 (2013). https://doi.org/10.1371/journal.pone.0066192

Bernstein D,S.: Basic Matrix Properties. In: Matrix Mathematics: Theory, Facts, and Formulas, 2nd edn. Princeton University Press, Princeton, New Jersey, p. 159 (2009)

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W.: Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge (1975)

Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J., Stone, C.J., et al.: Classification and Regression Trees, 1st edn. Chapman and Hall/CRC, Boca Raton (1984)

Centers for Disease Control and Prevention (CDC).: National Health and Nutrition Examination Survey Data (2013–2016) (2013). https://wwwn.cdc.gov/Nchs/Nhanes/

Chen, Y., Li, P., Wu, C.: Doubly robust inference with nonprobability survey samples. J. Am. Stat. Assoc. **115**(532), 2011–2021 (2020). https://doi.org/10.1080/01621459.2019.1677241

Elliott, M.R., Valliant, R.W.: Inference for nonprobability samples. Stat. Sci. **32**, 249–264 (2017)

Elliott, M.R., Raghunathan, T.E., Schenker, N.: Combining estimates from multiple surveys. In: Wiley StatsRef: Statistics Reference Online. American Cancer Society, pp. 1–10 (2018). https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08079

Freedman, D.A.: On the so-called Huber Sandwich estimator and robust standard errors. Am. Stat. **60**(4), 299–302 (2006). https://doi.org/10.1198/000313006X152207

Funk, L.M., Shan, Y., Voils, C.I., et al.: Electronic health record data versus the national health and nutrition examination survey (NHANES): a comparison of overweight and obesity rates. Med. Care **55**(6), 598–605 (2017). https://doi.org/10.1097/MLR.0000000000000693

Gearan, E.C., Monzella, K., Gola, A.A., et al.: Adolescent participants in the school lunch program consume more nutritious lunches but their 24-hour diets are similar to nonparticipants. J. Adolesc. Health (2021). https://doi.org/10.1016/j.jadohealth.2020.12.003

Gelman, A., Little, T.C.: Poststratification into many categories using hierarchical logistic regression. Surv. Methodol. **23**, 127–135 (1997)

Greenblatt, R.E., Zhao, E.J., Henrickson, S.E., et al.: Factors associated with exacerbations among adults with asthma according to electronic health record data. Asthma Res. Pract. **5**(1), 1 (2019). https://doi.org/10.1186/s40733-019-0048-y

Greenland, S., Robins, J.M., Pearl, J.: Confounding and collapsibility in causal inference. Stat. Sci. **14**(1), 29–46 (1999). https://doi.org/10.1214/ss/1009211805

Grill, J.D., Hoang, D., Gillen, D.L., et al.: Constructing a local potential participant registry to improve Alzheimer's disease clinical research recruitment. J. Alzheimer's Dis. **63**(3), 1055–1063 (2018). https://doi.org/10.3233/JAD-180069

Hainmueller, J.: Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. Polit. Anal. **20**(1), 25–46 (2012). https://doi.org/10.1093/pan/mpr025

Hansen, L.P.: Large sample properties of generalized method of moments estimators. Econometrica **50**(4), 1029–1054 (1982). https://doi.org/10.2307/1912775

Imai, K., Ratkovic, M.: Covariate balancing propensity score. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **76**(1), 243–263 (2014). https://doi.org/10.1111/rssb.12027

Imbens, G.W., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, Cambridge (2015). https://doi.org/10.1017/CBO9781139025751

Kish, L.: Survey Sampling. Wiley, New York (1965)

Lee, B.K., Lessler, J., Stuart, E.A.: Weight trimming and propensity score weighting. PLoS ONE **6**(3), e18,174 (2011). https://doi.org/10.1371/journal.pone.0018174

Li, F., Morgan, K.L., Zaslavsky, A.M.: Balancing covariates via propensity score weighting. J. Am. Stat. Assoc. **113**(521), 390–400 (2018). https://doi.org/10.1080/01621459.2016.1260466

Liaw, A., Wiener, M.: Classification and regression by randomforest. R News **2**(3), 18–22 (2002)

Little, R.J.A., Rubin, DB.: Complete-Case and Available-Case Analysis, Including Weighting Methods. In: Statistical Analysis with Missing Data. John Wiley & Sons, Ltd, pp 41–58 (2014). https://doi.org/10.1002/9781119013563.ch3

Lumley, T.: Complex Surveys: A Guide to Analysis Using R. Wiley, Hoboken (2010)

Lumley, T., Scott, A.: Fitting regression models to survey data. Stat. Sci. **32**(2), 265–278 (2017). https://doi.org/10.1214/16-STS605

Masten, M.A., Poirier, A.: Identification of treatment effects under conditional partial independence. Econometrica **86**(1), 317–351 (2018). https://doi.org/10.3982/ECTA14481

Mazzinari, G., Serpa Neto, A., Hemmes, S.N.T., et al.: The Association of Intraoperative driving pressure with postoperative pulmonary complications in open versus closed abdominal surgery patients-a post-hoc propensity score-weighted cohort analysis of the LAS VEGAS study. BMC Anesthesiol. **21**(1), 84 (2021). https://doi.org/10.1186/s12871-021-01268-y

Moulder, K.L., Besser, L.M., Beekly, D., et al.: factors influencing successful lumbar puncture in Alzheimer research. Alzheimer Dis. Assoc. Disord. **31**(4), 287–294 (2017). https://doi.org/10.1097/WAD.0000000000000209

Nelder, J.A., Wedderburn, R.W.M.: Generalized linear models. J. R. Stat. Soc. Ser. A (Gen.) **135**(3), 370–384 (1972). https://doi.org/10.2307/2344614. (**publisher: [Royal Statistical Society, Wiley]**)

Oh, S.S., Galanter, J., Thakur, N., et al.: Diversity in clinical and biomedical research: a promise yet to be fulfilled. PLOS Med. **12**(12), e1001918 (2015). https://doi.org/10.1371/journal.pmed.1001918

O'Muircheartaigh, C., Hedges, L.V.: Generalizing from unrepresentative experiments: a stratified propensity score approach. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) **63**(2), 195–210 (2014). https://doi.org/10.1111/rssc.12037. (**publisher: John Wiley & Sons, Ltd**)

Park, D.K., Gelman, A., Bafumi, J.: Bayesian multilevel estimation with poststratification: state-level estimates from national polls. Polit. Anal. **12**(4), 375–385 (2004). https://doi.org/10.1093/pan/mph024

Robbins, M.W., Ghosh-Dastidar, B., Ramchand, R.: Blending probability and nonprobability samples with applications to a survey of military caregivers. J. Surv. Stat. Methodol. (2020)

Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983). https://doi.org/10.1093/biomet/70.1.41

Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. Wiley, Hoboken (1987)

Rubright, J.D., Cary, M.S., Karlawish, J.H., et al.: Measuring how people view biomedical research: reliability and validity analysis of the Research Attitudes Questionnaire. J. Empir. Res. Hum. Res. Ethics **6**(1), 63–68 (2011). https://doi.org/10.1525/jer.2011.6.1.63

Salazar, C.R., Hoang, D., Gillen, D.L., et al.: Racial and ethnic differences in older adults' willingness to be contacted about Alzheimer's disease research participation. Alzheimer's Dementia Transl. Res. Clin. Interv. **6**(1):e120023 (2020). https://doi.org/10.1002/trc2.12023

Schell, T.L., Peterson, S., Vegetabile, B.G., et al.: State-level estimates of household firearm ownership (2020). https://www.rand.org/pubs/tools/TL354.html

Schildcrout, J.S., Rathouz, P.J.: Longitudinal studies of binary response data following case-control and stratified case-control sampling: design and analysis. Biometrics **66**(2), 365–373 (2010). https://doi.org/10.1111/j.1541-0420.2009.01306.x

Stuart, E.A.: Matching methods for causal inference: a review and a look forward. Stat. Sci. **25**(1), 1–21 (2010). https://doi.org/10.1214/09-STS313

Vegetabile, B.G., Griffin, B.A., Coffman, D.L., et al.: Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures. Health Serv. Outcomes Res. Method. (2021). https://doi.org/10.1007/s10742-020-00236-2

Walsh, S.P., Raman, R., Jones, K.B., et al.: ADCS prevention instrument project: the mail-in cognitive function screening instrument (MCFSI). Alzheimer Dis. Assoc. Disord. **20**(4 Suppl 3), S170-178 (2006). https://doi.org/10.1097/01.wad.0000213879.55547.57

Wang, L., Valliant, R., Li, Y.: Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts (2020). arXiv:2007.02476 [stat]

Yu, J., Green, M.D., Li, S., et al.: Liver metastasis restrains immunotherapy efficacy via macrophage-mediated T cell elimination. Nat. Med. **27**(1), 152–164 (2021). https://doi.org/10.1038/s41591-020-1131-x

Zadrozny, B .: Learning and evaluating classifiers under sample selection bias. In: Proceedings of the Twenty-First International Conference on Machine learning. Association for Computing Machinery, New York, NY, USA, ICML ’04, p 114. https://doi.org/10.1145/1015330.1015425 (2004)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Olivia M. Bernstein[1] · **Brian G. Vegetabile**[2] · **Christian R. Salazar**[3] · **Joshua D. Grill**[3,4,5] · **Daniel L. Gillen**[1,3]

Brian G. Vegetabile
bvegetab@rand.org

Christian R. Salazar
csalaza4@hs.uci.edu

Joshua D. Grill
jgrill@hs.uci.edu

Daniel L. Gillen
dgillen@uci.edu

[1] Department of Statistics, University of California, Irvine, Irvine, CA, USA

[2] RAND Corporation, Santa Monica, CA, USA

[3] Institute for Memory Impairments and Neurological Disorders, University of California, Irvine, Irvine, CA, USA

[4] Department of Psychiatry and Human Behavior, University of California, Irvine, Irvine, CA, USA

[5] Department of Neurobiology and Behavior, University of California, Irvine, Irvine, CA, USA