Journal of Data Science 0 (0), 1–20
??? 2026
DOI: 10.6339/26-JDS1234
Computing in Data Science

An Estimation Framework for Combining Probability and Non-probability Samples

Mahmoud Elkasabi[1,*], Taylor Lewis[1], and Matthew Williams[1]
[1]*Center for Official Statistics, RTI International, Durham, NC 27713, USA*

Abstract

Survey researchers are increasingly adopting hybrid sampling designs to address the limitations of traditional probability sampling, especially when studying rare or hard-to-reach populations. Challenges such as high screening costs, low statistical efficiency, and operational constraints make purely probability-based approaches impractical in many contexts. This article uses public data from the National Health and Nutrition Examination Survey to demonstrate how one can make population estimates from a hybrid sampling strategy that combines data from a stratified, multistage probability sample with data from a non-probability sample within the same primary sampling units as the probability sample. We outline a framework and discuss methods for analyzing data from a hybrid sample such as this, where covariates and survey outcomes are observed in both the probability and non-probability samples. We present a case study to illustrate the framework. We provide the case study R code in the supplementary material.

Keywords *hard-to-reach populations; non-probability sample; rare populations*

1 Background

Survey researchers often face significant challenges when studying rare populations due to the high screening costs, statistical inefficiency, and practical limitations of traditional probability sampling methods (Tourangeau et al., 2014). When a rare subpopulation of interest is underrepresented in a survey dataset, its small sample size reduces estimate precision and statistical power. With non-probability sampling continuing to gain legitimacy (e.g., Baker et al., 2013), particularly as a complementary data collection approach, we introduce a framework to integrate data from probability and non-probability samples (e.g., convenience sample, snowball sample, or quota sample). This hybrid approach leverages the statistical rigor of probability sampling while exploiting the efficiency of non-probability methods to effectively over-sample rare populations. A range of statistical techniques have been proposed recently in the literature integrating data from probability and non-probability samples, including propensity score adjustments, calibration weighting, kernel weighting, mass imputation, Bayesian approaches, and combinations thereof. In our setting, we assume that the probability sample is a stratified, multi-stage probability sample, and that the non-probability sample is drawn from the same primary sampling units (PSUs) selected as part of the probability sample.

This paper aims to present a general framework for estimation from data from a hybrid design that combines probability and non-probability samples, with particular emphasis on cases

*Corresponding author. Email: melkasabi@rti.org.

Received September 8, 2025; Accepted May 29, 2026

where non-probability units are selected from PSUs drawn for the probability sample. As we aim to present a structured approach to reasoning about similar data structures, the focus is on estimation from non-probability samples, methods for combining probability and non-probability data/estimates, appropriate analytical tools (software), and performance evaluation. While related issues, such as the ignorability condition, model fit, and variance estimation, are important and worthy of further investigation, they fall outside the scope of this paper and are left for future research. Additionally, our focus is not on comparing different methods, since such a comparison would entail a broader experimental framework and more detailed evaluation than is within the current scope. Instead, our primary objective is to provide practitioners with a clear operational workflow. We delineate the critical stage of the integration process detailing the specific methods and tools suitable for each stage.

Our framework consists of two stages: The first is the estimation stage, which covers methods for deriving estimates from non-probability samples, including techniques for generating survey weights for estimation from the non-probability samples, as well as methods for integrating estimates from both probability and non-probability samples. We illustrate several of these methods using R code, with a particular focus on methods that generate survey weights that can be used for estimation and inference. (The R code is provided in the supplementary materials.) The second is the evaluation stage, which covers measures that can be used to evaluate and compare methods when multiple approaches are used.

This paper is organized as follows. Section 2 describes the data structure and the assumptions underlying the analysis. Section 3 introduces methods for combining non-probability data with probability-based data, beginning with Stage 1, which focuses on estimation from the non-probability sample and distinguishes between compatible and incompatible methods, followed by Stage 2, which details approaches for combining data from non-probability and probability samples. Section 4 introduces measures to evaluate the performance of the proposed methods. Section 5 presents a case study and Section 6 concludes with the recommended next steps and directions for future research.

2 Data Structure and Assumptions

We assume a sampling design that integrates a stratified multistage probability sample, as might be the case with a nationally representative in-person health survey, with a non-probability sample drawn from the same PSUs. Within each stratum, PSUs are selected with known probabilities, followed by subsequent stages of sampling (e.g., segments, households, individuals), also with known probabilities. We assume data from the non-probability sample is collected with unknown probabilities from the probability sample PSUs, using strategies such as social media advertising, community engagement through local organizations, intercept sampling in public spaces, venue-based sampling, or respondent-driven sampling. Ultimately, the choice of respondent recruitment method(s) depends on the characteristics of the targeted rare population, study objectives, and available resources.

With respect to statistical notation, we assume two samples: (1) Sample A, a probability sample of the general population, which includes n_A cases from the targeted rare population; and (2) Sample B, a non-probability sample of n_B cases from the targeted rare population. Both samples include a shared set of auxiliary variables X and outcome variables Y but only Sample A includes survey weights w that can be used to produce weighted estimates of Y. As shown in Figure 1, the objective is to integrate data from Samples A and B into a single dataset comprised

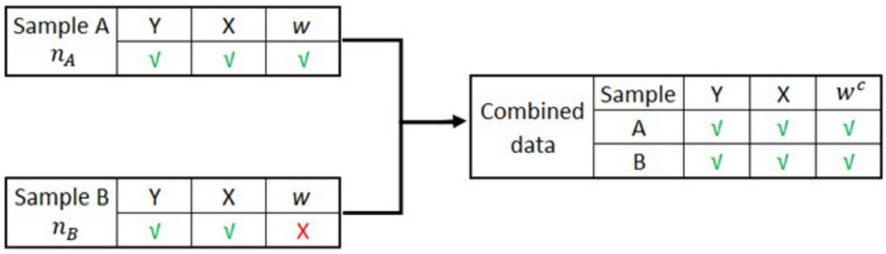

Figure 1: Data structure visualization – samples A and B and the combined sample.

of cases from the targeted rare population such that survey estimates can be generated. The combined dataset will have an analysis weight w^c that can be used to produce weighted estimates of Y based on records from both samples. This involves first generating survey weights for Sample B.

To date, most data integration methods in the literature focus on a problem that combines information from a non-probability sample with a probability sample using shared auxiliary variables X appearing on both samples, assuming the outcome variables only appear in the non-probability sample (e.g. Rivers, 2007; Lee and Valliant, 2009; Valliant and Dever, 2011; Elliott and Valliant, 2017; Chen et al., 2019). Our data integration problem is somewhat different because we assume X and Y appear on both probability and non-probability samples. This structure has been discussed in a limited number of other studies, including Elliott and Haviland (2007), DiSogra et al. (2011), Dever (2018), Wiśniowski et al. (2020), Robbins et al. (2021), Rueda et al. (2023), and Rueda et al. (2023). However, works in the literature considering this data structure often focus on estimation from the non-probability sample, rather than combined estimation from both probability and non-probability samples. In the case of a targeted rare population, we consider the combined estimation approach to be more appropriate because it leverages the strengths of both probability and non-probability samples, allowing the probability sample to provide population representativeness while the non-probability sample contributes additional coverage to the rare subpopulation.

3 Methods to Combine Non-Probability Data with Probability-Based Data

3.1 Stage 1: Estimation from the Non-Probability Sample

Several methods for survey estimation based on non-probability samples have been proposed in the literature. Common methods include propensity score adjustments (Chen et al., 2020; Schonlau and Couper, 2017; Valliant, 2020; Wang et al., 2021), calibration weighting (Ferri-García and MdM, 2018; Kott, 2016; Liu and Valliant, 2023), kernel weighting (Kern et al., 2021; Wang et al., 2020), mass imputation (Kim et al., 2021), super-population modelling and prediction, and Bayesian approaches (Rafei, 2021; Sakshaug et al., 2019; Salvatore et al., 2024). Some methods involve multiple methods, such as so-called doubly robust estimation (Chen et al., 2020; Yang et al., 2020). Comprehensive discussions of these methods appear in review articles by Elliott and Valliant (2017), Yang and Kim (2020), Rao (2021), Wu (2022), and Kim (2022). Additionally, Cobo et al. (2024) provides a helpful review of software packages in R and Python that support inference from non-probability samples, including *NonProbEst* (Castro-

Martin et al., 2020a; Rueda et al., 2020), *nonprobsy* (Chrostowski and Beręsewicz, 2024), *nppR* (Beaumont and Dhushenthen, 2024), and KWML (Wang and Kern, 2023) in R, as well as *inps* (Castro-Martin, 2024) in Python. In this paper, however, we restrict our attention to R packages.

We classify methods that produce survey weights for non-probability samples into two categories: (1) *compatible* with our data structure and (2) *likely incompatible* with our data structure. Compatible methods are those that generate survey weights suitable for downstream statistical analyses after integrating probability and non-probability samples, and for which an accessible statistical package exists to support implementation. Methods are considered likely incompatible if they do not produce survey weights or if there is currently no statistical software available to implement them. We recognize that some of these methods may become compatible in the future as software tools are developed.

Let X and Y denote vectors of auxiliary variables and outcome variables, respectively, with data that are available from both Sample A and Sample B. With I_B denoting the sampling indicator of Sample B, for most methods summarized in Yang and Kim (2020), Kim (2022), and Wu (2022), the following two assumptions regarding the selection probability $\pi_B(X)$ need to hold:

1. $\pi_B(X) = P(I_B = 1|X, Y) = P(I_B = 1|X)$. This means that the selection mechanism into Sample B is conditional upon the covariate vector X, and therefore the expected value of Y can be estimated from Sample B after accounting for $\pi_B(X)$, $E(Y|X) = E(Y|X, I_B = 1)$. This condition is tantamount to the missing at random (MAR) assumption (Little and Rubin, 2019).
2. $\pi_B(X) > 0$. This implies that all units have a non-zero chance to be selected into Sample B.

Yang and Kim (2020) note that these assumptions "...constitute the strong ignorability condition (Rosenbaum and Rubin, 1983). This assumption holds if the set of covariates contains all predictors for the outcome that affect the possibility of being selected in sample B." This might be a limitation if the available auxiliary variables X do not include all predictors that explains the distribution of Y (Meng, 1994; Kang and Schafer, 2007).

3.1.1 Compatible Methods

Propensity Score Adjustments

Propensity scores adjustments (PSAs) were originally proposed to adjust for self-selection biases in observational studies (Rosenbaum and Rubin, 1983), and have since been adopted extensively to compensate for survey nonresponse (Little and Rubin, 2019). Studies by Lee (2006), Lee and Valliant (2009), Valliant and Dever (2011), Elliott and Valliant (2017), and Dever (2018) have extended their use to facilitate estimation from non-probability samples. In essence, the probabilities, or propensities, of self-selection into the non-probability sample are first estimated and then converted into analysis weights, resulting in a family of methods such as inverse probability weighting (IPW), matching, calibration, and model-based adjustment.

In the PSA approach, Sample A is considered representative of the target population and treated as a *reference* sample. A prediction model is then exploited to estimate the propensity scores (probability of inclusion) in Sample B, using covariates X that are common between both samples. These propensity scores are utilized via different techniques to construct pseudo-weights for Sample B that enable estimation and inference from the non-probability sample. Data analysis can also be performed on a combined dataset containing both Samples A and B, assimilating the survey weights and pseudo-weights, respectively. Equivalently, one can generate two sets of point estimates and variances for Samples A and B and combine using rules discussed

in Section 3.2.

One can use either parametric or non-parametric methods to estimate $\pi_B(X)$. Parametric methods require a link function for estimating $\pi_B(X) = \pi_B(x_i, \alpha)$ such as (1) the inverse logit function $\pi_B(X) = 1 - \{1 + \exp(x_i^{'}\alpha)\}^{(-1)}$, (2) the inverse probit function $\pi_B(X) = \Phi(x_i^{'}\alpha)$ where $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$, or (3) the inverse complementary log-log function $\psi_B(X) = 1 - \exp\{-\exp(x_i^{'}\alpha)\}$. As the parametric techniques require a correctly specified model, estimated propensities might be biased and highly variable if the model is misspecified. On the other hand, nonparametric techniques, such as kernel weighting do not require an explicit distributional form for estimating $\phi_B(X)$.

Several strategies were proposed for estimating $\pi_B(x_i, \alpha)$:

1. Chen et al. (2020) proposed estimating $\pi_B(x_i, \alpha)$ via maximum likelihood by solving the pseudo log-likelihood function:

$$l(\alpha) = \sum_{i \in B} \log\left(\frac{\pi_{B,i}}{1 - \pi_{B,i}}\right) + \sum_{i \in A} d_i^A \log(1 - \pi_{B,i}), \tag{1}$$

   then propensities can then be estimated by solving:

$$\sum_{i \in B} x_i - \sum_{i \in A} d_i^A \pi(x_i, \alpha) x_i = 0. \tag{2}$$

   where d_i^A is the design weight of Sample A.
2. Valliant and Dever (2011) proposed to fit a survey-weighted logistic regression model to the pooled dataset of Sample A and Sample B with the dependent variable is R_i that is $R_i = 1$, $i \in B$ and $R_i = 0$, $i \in A$, and weights are defined as $d_i = 1$, $i \in B$ and $d_i = d_i^A(1 + n_B/\hat{N}_A)$, $i \in A$, where n_B is the number of participants in Sample B, and $\hat{N}_A = \sum_{i \in A} d_i^A$.
3. Wang et al. (2021) proposed a two-stage adjusted logistic propensity (ALP) weighting method. In the first stage, initial propensities $\hat{p}_i$ are estimated from a weighted logistic model similar to the one proposed by Valliant and Dever (2011) but with survey weights $d_i = d_i^A$, $i \in A$ and $d_i = 1$, $i \in B$. In the second stage, final propensities are estimated as $\hat{\pi}_{B,i} = \hat{p}_i/(1 - \hat{p}_i)$.
4. Wu (2022) proposed using the generalized estimation equations, where propensities can be estimated by solving:

$$\sum_{i \in B} \frac{x_i}{\pi(x_i, \alpha)} - \sum_{i \in A} d_i^A x_i = 0. \tag{3}$$

   Although there is a great deal of similarity between the maximum likelihood approach and the generalized estimation equations approach, the latter only requires the estimated population totals for the auxiliary variables X, and not necessarily the microdata from Sample A.

Once the propensities are estimated, there are several competing methods to use them in estimation:

1. Construct a pseudo-weight for Sample B based on the inverse of the estimated response propensities. This is the inverse propensity weighting (IPW) of Valliant (2020):

$$\hat{Y}_{PSA1} = \frac{N}{n_B} \sum_{i \in B} \frac{y_i}{\hat{\pi}_i}. \tag{4}$$

2. Schonlau and Couper (2017) propose obtaining weights for a Horvitz-Thompson type estimator as follows:

$$\hat{Y}_{PSA2} = \sum_{i \in B} y_i \left(\frac{1 - \hat{\pi}_i}{\hat{\pi}_i}\right). \tag{5}$$

3. Valliant and Dever (2011) propose estimating the propensities and then post-stratifying the sample (Deville and Särndal, 1992) in a multi-step process. Specifically, one first sorts the data according to the estimated propensities $\hat{\pi}_i$ and splits the combined sample into $g = 5$ classes, each of which has about the same number of cases in the combined sample. Next, an average propensity is calculated within each subclass $\bar{\pi}_g$, and the inverse is used as a weighting adjustment factor as follows:

$$\hat{Y}_{PSA3} = \frac{N}{n_B} \sum_g \sum_{i \in B_g} \frac{y_i}{\bar{\pi}_g}. \tag{6}$$

Researchers have found that PSA can effectively reduce bias in certain situations, though it often increases variance as a trade-off (Lee and Valliant, 2009). Valliant and Dever (2011) demonstrated that, in order to minimize bias, PSA-based estimates should be supplemented with additional weighting adjustments. Studies by Lee and Valliant (2009) and Ferri-García and Rueda (2020) examined PSA in combination with calibration, concluding that additional calibration adjustments can be beneficial when appropriate covariates are used. When the distribution of auxiliary variables X is similar between both samples A and B, the PSA methods above seem to perform well under this "high-overlap" scenario. However, when the distributions are very different between samples ("low-overlap") these methods can be highly variable (Savitsky et al., 2023, 2025). When available through modeling or direct observation, additional information about the inclusion probabilities of individuals from sample B being selected into sample A can be used to mitigate this additional variability via the Core Relationship for Independent Sampling Properties (CRISP) formulation: $\pi_{B,i}/(\pi_{B,i} + \pi_{A,i})$, $i \in A \cup B$ (Elliot, 2009; Beresovsky et al., 2025). Further refinement may be possible when direct linkage of overlapping records between samples is available (Gershunskaya et al., 2025). However, all PSA methods depend on a model for estimating $\pi_{B,i}$.

With respect to software, *NonProbEst*::`propensities` of the R package *NonProbEst* can be used to estimate propensities using parametric and non-parametric methods. These propensities can then be used to create pseudo-weight using the three methods defined above: 1) *NonProbEst*::`valliant_weights`; 2) *NonProbEst*::`sc_weights`; or 3) *NonProbEst*::`vd_weights`, or as a variant *NonProbEst*::`lee_weights` attributable to Lee and Valliant (2009). Variance estimates for PSA methods can be obtained using the delete-a-group jackknife estimator (*NonProbEst*::`jackknife_variance`). See Rueda et al. (2020) for more details about *NonProbEst*.

Unlike *NonProbEst* that requires using several functions to implement PSA, *nonprobsvy*::`nonprob` can be used directly to produce estimates from the non-probability samples with their estimated standard errors using an analytic version of the variance estimator. The *nonprobsvy*::`nonprob` can be used to implement PSA (IPW) with possible calibration constraints using logistic regression with logit, probit, and complementary log-log link functions.

Calibration Weighting

In the calibration weighting technique, auxiliary variables X from the two samples are calibrated to known population totals such that the weighted totals match those of the target population (Deville and Särndal, 1992; Kott, 2016). These totals may be either taken from a benchmark data source or derived from the probability sample. Under the calibration weighting approach, survey weights for the non-probability sample are directly generated as a product of

a calibration process with the key property

$$\sum_{i \in B} \omega_i^B x_i = \sum_{i \in A} d_i^A x_i, \tag{7}$$

where d_i^A is the design weight of Sample A, and ω_i^B is the calibration weight of Sample B.

In practice, the calibration is performed by exploiting a model of auxiliary variables X, $m(X;\beta)$, and finding ω_i^B by solving the following optimization problem:

$$\min \quad \sum_{i \in A,B} G(\omega_i^B, d_i^A) = \sum_{i \in B} \omega_i^B \log \omega_i^B - \lambda^T \left\{ \sum_{i \in B} \omega_i^B x_i - \sum_{i \in A} d_i^A x_i \right\}. \tag{8}$$

This provides a set of weights of the form

$$\omega_i^B = \omega_B(x_i; \hat{\lambda}) = \frac{N \exp(\hat{\lambda}^T x_i)}{\sum_{i \in B} \exp(\hat{\lambda}^T x_i)}, \tag{9}$$

where $\hat{\lambda}$ satisfies the following equation:

$$\sum_{i \in B} \exp(\lambda^T x_i) \left\{ x_i - \frac{1}{N} \sum_{i \in A} d_i^A x_i \right\} = 0. \tag{10}$$

Calibration weighting can be implemented using in either the *NonProbEst* or *nonprobsvy* R package. In *NonProbEst*, *NonProbEst*::`calib_weights` computes the g-weights for the calibration estimator using *sampling*::`calib`, whereas *nonprobsvy*::`nonprob` can be used to directly produce calibrated weights or calibrated IPW weights.

Kernel Weighting

The kernel weighting (KW) method, developed by Wang et al. (2020), is similar to the IPW variant of the PSA method. Both techniques produce pseudo-weights for the non-probability sample using auxiliary variables from a reference probability sample. However, the KW method differs in how these weights are generated. Like PSA methods, KW relies on estimated propensities for inclusion in Sample B, which can be obtained using various approaches. Logistic regression is commonly used, but an alternative is a machine learning (ML) algorithm. ML algorithms can require fewer assumptions than parametric models and thus be more robust to model misspecification. Recent examples of ML algorithms applied to non-probability sampling estimation and inference can be found in Ferri-García and Rueda (2020), Buelens et al. (2018), Kern et al. (2021), and Castro-Martin et al. (2022). Collectively, these studies suggest that ML methods may be more effective than logistic regression in minimizing self-selection biases in certain scenarios. However, when the underlying sizes of either Sample A or Sample B are small, the performance of ML methods, particularly in terms of predictive accuracy and statistical efficiency, may deteriorate due to increased variance and model instability.

The KW method is implemented using the following steps:

1. Estimate propensity scores using the machine learning model as:

$$\hat{\pi}_i = E_M[\hat{I}_{B;i} = 1 | x_i], \tag{11}$$

where M is one of the machine learning models to estimate the propensity, and

$$\hat{I}_{B;i} = \begin{cases} 1 & i \in B \\ 0 & i \in A \end{cases}, \quad i \in A \cup B. \tag{12}$$

2. Calculate the signed distance between the two individuals belonging to the different samples as:
$$\Delta_{ij} = \hat{\pi}_i - \hat{\pi}_j, \quad i \in B, j \in A. \tag{13}$$
3. Smooth the distance Δ_{ij} between individuals using a kernel function centered at zero, and calculate the kernel weight accordingly as:
$$k_{ij} = \frac{K\{\Delta_{ij}/h\}}{\sum_{i \in B} K\{\Delta_{ij}/h\}}, \quad i \in B, j \in A, \tag{14}$$
so that
$$\sum_{i \in B} k_{ij} = 1, \quad k_{ij} \in [0, 1], \tag{15}$$
where $K\{\cdot\}$ is the zero-centered kernel function (Epanechnikov, 1969), and h is the band-width corresponding to that kernel function.
4. Calculate pseudo-weights as sum of the survey sample weights, d_j, $j \in A$, that are weighted by the cohort unit i's kernel weights k_{ij}, $i \in B$ as follows
$$w_i^{KW} = \sum_{j \in A} d_j k_{ij}, \quad i \in B, j \in A. \tag{16}$$

Kernel weighting can be done using the R package *KWML* (Wang and Kern, 2023). The package computes pseudo-weights for non-probability samples using the KW techniques, and also offers an option for propensities estimated using logistic regression models.

3.1.2 Likely Incompatible Methods

Mass Imputation

Mass imputation was developed to create synthetic data within the context of two-phase sampling (Kim and Rao, 2012). Studies by Rivers (2007), Kim et al. (2021), and Chen et al. (2022, 2023) extended this technique so that it can be used to create synthetic data of the outcome variables for a probability sample using data from a non-probability sample. Under the mass imputation approach, the Sample A is treated as a sample that has all values of the outcome variable Y missing. Since both X and Y are observed in Sample B, this supplemental sample is exploited as a training dataset to develop an imputation model that generates synthetic values of Y for the probability sample. In this context, Sample B does not need to represent the target population, as the focus is on capturing the relationships between variables to train the imputation model. Once the synthetic values of the outcome are generated, inferences are made based solely on these synthetic values, without any contribution from the original Y values in the non-probability sample.

Although mass imputation has been proposed as a novel application of conventional imputation methods (Little and Rubin, 2019), we argue that it is not well suited to our data structure. Mass imputation typically assumes a setting in which the outcome variable is observed exclusively in the nonprobability sample, an assumption that does not hold in our case. Moreover, data from the nonprobability sample are used solely to train the imputation model and do not directly contribute to inference for the target population. In addition, mass imputation does not produce pseudo-weights for units in the nonprobability sample, making it incompatible with weighting-based inference approaches. Finally, because it does not effectively

integrate nonprobability units into the target population, mass imputation does not lead to any increase in the effective sample size for population-level estimation. Taken together, these limitations make mass imputation unsuitable for the hybrid data structure considered in this study.

Sample Matching

Another (likely) incompatible method is sample matching. For some example applications, see Rivers (2007), Baker et al. (2013), Mulrow et al. (2007), Castro-Martin et al. (2020b), and Liu and Valliant (2023). The idea behind sample matching is to assign pseudo-weights to units in the non-probability sample based on matched units from the probability sample, using a shared set of auxiliary variables $\boldsymbol{X}$. Note that neither the KW method nor sample matching are appropriate for small sample sizes, as we have in the targeted rare population setting. And aside from the general-purpose R package *matching* (Sekhon, 2011), we are not aware of any statistical package specifically developed to implement statistical matching for inference from non-probability samples.

Superpopulation Modeling/Prediction Approaches

Similar to the disqualifying reason for mass imputation, all of these methods assume a data structure where the outcome variables are only available in the non-probability sample. The notion is to model the outcome variable based on a covariance structure with auxiliary variables $\boldsymbol{X}$ appearing on the non-probability sample, and use the model to predict the (missing) outcome values in the probability sample. Some of these prediction methods can be implemented using the *NonProbEst* R package. For more details about these methods, see Elliott and Valliant (2017), Sakshaug et al. (2019), Valliant (2020), Cornesse et al. (2020), Wiśniowski et al. (2020), Rafei (2021), Rafei et al. (2022), Nandram et al. (2021), Nandram and Rao (2023) and Salvatore et al. (2024).

Doubly Robust Estimation

To avoid the dilemma of potential model misspecification under a PSA method, the doubly robust estimation method combines a given PSA method with another method such as mass imputation or a super-population prediction method. Under this approach, a model for propensities and a prediction model for the outcome variable are both incorporated into the adjustment. Even if one of the models is misspecified, the doubly robust estimator remains consistent. The doubly robust methods can be implemented using the R packages *nonprobsvy* and *nppR*. For more details about this method, see Kim and Haziza (2014), Kim and Wang (2019), Yang et al. (2020), Chen et al. (2020, 2023), Chen and Haziza (2022), Castro-Martin et al. (2022), Castro-Martin et al. (2020b, 2022).

3.2 Stage 2: Combining Data from Sample A and Sample B

In the second stage, data from Sample A and Sample B are combined into a single dataset. This requires the analyst to create a combined weight, derived from the survey and pseudo-weights, respectively, and append to the dataset. A combined estimator $\hat{Y}_{com}$ from the two samples can be estimated as

$$\hat{Y}_{com} = \alpha \hat{Y}_A + (1 - \alpha)\hat{Y}_B, \tag{17}$$

where $\hat{Y}_A = \sum_A w_i y_i$ is the weighted, estimated total of Y from the probability sample A, $\hat{Y}_B = \sum_B w_i y_i$ is the like from the non-probability sample B, and α is a composite factor with

$0 \leqslant \alpha \leqslant 1$. It is important to note that the combined estimator $\hat{Y}_{com}$ relies on the assumption that $\hat{Y}_B$ provides an unbiased estimate of Y. This is a particularly strong assumption and may be difficult to satisfy in practice, especially when Sample B is subject to selection biases or other sources of error.

The literature on multi-frame sample designs (e.g. Lohr, 2011; Skinner and Rao, 1996) offers several choices for determining α:

1. Denoting V_A as the variance of the estimated total from A and V_B the variance of the estimated total from B, the composite factor

$$\alpha_{opt} = \frac{V_B}{V_A + V_B} \tag{18}$$

will produce a combined estimate with a minimized estimated variance. However, this method is dependent on a specific outcome variable, so different values of α would be needed for different outcome variables.

2. Assign the composite factor as proportional to the two sample sizes:

$$\alpha_n = \frac{n_a}{n_A + n_B}. \tag{19}$$

3. Assign the composite factor as proportional to the *effective* sample size as:

$$\alpha_{\tilde{n}} = \frac{\tilde{n}_a}{\tilde{n}_A + \tilde{n}_B}, \tag{20}$$

where

$$\tilde{n}_A = \frac{(\sum_A w_i)^2}{\sum_A w_i^2} \quad \text{and} \quad \tilde{n}_B = \frac{(\sum_B w_i)^2}{\sum_B w_i^2}. \tag{21}$$

A weighted form of the combined estimator can be written as:

$$\hat{Y}_{com} = \sum_{i \in A \cup B} w_i^c y_i, \tag{22}$$

where

$$w_i^c = \begin{cases} \alpha_{\tilde{n}} w_i, & i \in A \\ (1 - \alpha_{\tilde{n}}) w_i, & i \in B \end{cases}. \tag{23}$$

Thus the estimated combined variance can be written as:

$$\hat{V}_{com} = \alpha^2 \hat{V}_A + (1 - \alpha)^2 \hat{V}_B. \tag{24}$$

4 Methods Evaluation

After calculating the combined estimate, it is important to assess its accuracy, particularly when multiple estimation methods have been used, and one must be selected as the final estimate. The evaluation can be conducted by comparing the combined estimates against reference benchmarks. These benchmarks may be derived from known population parameters, such as those obtained through a population census, or from estimates produced by reliable population surveys. It is advisable to examine as many benchmarks as possible; however, it is particularly important that these benchmarks are highly correlated with the key variables under study, as this enhances the relevance and validity of the evaluation. For example, when the key study variables pertain

to the health status of participants, it is recommended to include benchmark variables known to be associated with health outcomes, such as smoking behavior, blood pressure, and other health-related indicators. If the combined estimates demonstrate strong performance on these correlated variables, it can reasonably be expected that similar levels of accuracy will be observed for other health-related study variables.

Let Y_R denote the value of the benchmark (reference) for variable Y. Here are examples of measures that can be calculated:

1. **Difference.** This measure retains the direction of the discrepancy and indicates whether the composite estimate systematically overestimates or underestimates the benchmark.

$$\hat{Y}_{com} - Y_R. \tag{25}$$

2. **Absolute difference.** This measure quantifies the magnitude of the discrepancy between the composite estimate $\hat{Y}_{com}$ and the benchmark value Y_R, expressed in the original units of Y.

$$\left|\hat{Y}_{com} - Y_R\right|. \tag{26}$$

When Y_R is based on a population parameter, the t-value of this difference can be calculated as

$$t = \frac{\left|\hat{Y}_{com} - Y_R\right|}{\sqrt{V(\hat{Y}_{com})}}, \tag{27}$$

whereas when Y_R is based on a survey estimate $\hat{Y}_R$, the t-value of this difference can be calculated as

$$t = \frac{\left|\hat{Y}_{com} - \hat{Y}_R\right|}{\sqrt{V(\hat{Y}_{com}) + V(\hat{Y}_R)}}. \tag{28}$$

3. **Relative bias.** This measure expresses the absolute difference between the composite estimate and the benchmark as a percentage of the benchmark value, allowing comparisons across variables with different scales.

$$100 \times \frac{\left|\hat{Y}_{com} - Y_R\right|}{Y_R}. \tag{29}$$

4. **Relative signed bias.** This measure expresses the signed bias as a percentage of the benchmark value, facilitating interpretation of the magnitude and direction of bias on a relative scale.

$$100 \times \frac{\hat{Y}_{com} - Y_R}{Y_R}. \tag{30}$$

5. **Confidence interval overlap indicator.** This indicator assesses whether the benchmark value lies within the confidence interval of the composite estimate.

$$\mathbb{I}\left(Y_R \in \left[\hat{Y}_{com} \pm z_{\alpha/2}\sqrt{V(\hat{Y}_{com})}\right]\right). \tag{31}$$

5 Case Study

In this section, we simulate the following data structure: a stratified multistage probability sample targeting adults aged 18 years and older, with non-citizen adults as a subpopulation of interest. Although the probability sample recruits a sizable number of non-citizens, this number is insufficient to produce reliable survey estimates, particularly because the demographic composition of recruited non-citizens is biased due to low response rates among younger non-citizens. Therefore, the probability sample is supplemented with additional data from non-citizens recruited using nonprobability procedures within the probability sample's PSUs. These procedures are specifically designed to oversample younger non-citizens.

We simulate our data structure using data from the National Health and Nutrition Examination Survey (NHANES) 2017–18 public use data (https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017). We use the nonprob function from the nonprobsvy R package to perform all the estimation methods. Unlike other packages, nonprob handles weighting and produces estimates from the non-probability sample along with relevant standard errors. In other packages, users would need to use multiple functions to achieve the same results. All R code used is provided in the supplemental materials.

5.1 Data

We used data on non-citizen adults aged 18 years and older from the NHANES 2017–18 public-use files (N = 801). The data were split into two equal replicates, which were used to construct Sample A and Sample B. In creating the replicates, we ensured that the complex survey design features, specifically the strata and cluster variables, were preserved within each replicate.

We simulated a probability sample (Sample A) using a selection model in which inclusion probabilities depend on auxiliary variables. The model was specified to under-represent adults aged 18–34 years. Based on this model, we selected 200 adults from the first NHANES replicate. We then simulated a nonprobability sample (Sample B) using a selection model designed to over-represent adults aged 18–34 years. This model was applied to the second replicate of the population data to select 150 adults.

For both samples, we maintained the following variables:

1. Three study variables: ever told you had high blood pressure, smoked at least 100 cigarettes in life, and coverage by health insurance;
2. Six auxiliary variables: sex, age group, race/Hispanic origin, education level, marital status, and presence of children aged 5 years or younger in the household; and
3. Two complex survey design variables: stratum identifier, and PSU identifier.

In addition, we incorporated the survey weights for Sample A and adjusted them to reflect the simulated selection process.

5.2 Methods

To calculate survey weights for Sample B, we considered three methods:

1. IPW with propensities estimated based on a model that considered all the six auxiliary variables, where NHANES data were used as a reference sample for estimating the propensities.
2. Calibration to estimated population totals of the six auxiliary variables, derived from NHANES data.
3. Calibrated IPW: a calibrated version of IPW using aggregate totals of the six auxiliary variables, where NHANES data were used as a reference sample for estimating the propensities

and for aggregating the calibration totals.
It is worth noting that because all weighting methods rely on the same auxiliary data, any differences in the results were expected to be small.

We used the NHANES sample as the reference sample in all three estimation methods. Although Sample A could have been used as an alternative reference, we did not choose it because Sample A does not provide full coverage of young non-citizens. Additionally, we included strata and cluster indicators as auxiliary variables in the adjustment models for both the IPW and calibrated IPW approaches, in order to account for overlapping design variables between Sample B and NHANES.

Furthermore, our demonstrations assumed that no survey weights were available for the non-probability sample. However, this assumption could be modified by deriving and assigning a non-null analysis weight for the non-probability sample, one that is then adjusted to generate the pseudo-weight. In fact, the *nonprobsvy*::`nonprob` function includes an argument for the user to provide variable weights from the non-probability sample.

5.3 Results

As shown in Figure 2, compared with the NHANES data, individuals in Sample A are, by design, less likely to be young (aged 18–34 years) and more likely to be older (i.e., age groups 35-49, 50-64, and 65 and older.) This age imbalance affects other related characteristics. For example, Sample A individuals are less likely to be Hispanic or Mexican American, to have attained a high school degree or higher, or to have children aged five years or younger in the household. They are more likely to be married, to have a high school degree, and to be non-Hispanic White. Estimates from Sample A are weighted estimates calculated using survey weight accounting for the selection from NHANES but not for differential nonresponse among the young age group. By design, unweighted estimates from Sample B show the opposite pattern. Compared with the NHANES data, individuals in Sample B are more likely to be young (aged 18–34 years) and less likely to be in older age groups, especially 35-49 years, reflecting the design of the selection model. As a result, Sample B includes a higher proportion of those who reported other races, those with less than a high school degree, and, to a lesser degree, those who live in households with children aged five years or younger. Conversely, individuals in Sample B are less likely to be males, married, to have a high school degree or more, and to be non-Hispanic White.

As shown in Figure 2, IPW estimates from Sample B are generally closer to the NHANES estimates than the unweighted estimates. This pattern is observed across nearly all characteristics. As expected, the calibrated IPW estimates from Sample B closely match the NHANES estimates, since the calibration was performed using the auxiliary variables displayed in the figure. Calibrated estimates are not presented in Figure 2 as they match estimates from the calibrated IPW estimates as both are calibrated to the same population totals.

As shown in Figure 3, for all three study variables, point estimates derived from Sample A, Sample B, and the combined sample (Sample A + Sample B) are not statistically different from the corresponding NHANES estimates, as evidenced by the substantial overlap of their confidence intervals. However, clear differences emerge in the precision of the estimates across samples. In particular, the confidence intervals associated with the combined sample are uniformly narrower than those obtained from Sample A or Sample B alone. This pattern is consistent with design-based expectations, as combining the two samples effectively increases the analytic sample size and, consequently, reduces the estimated variances and standard errors of the survey estimators. Importantly, the gain in precision appears to be achieved without in-

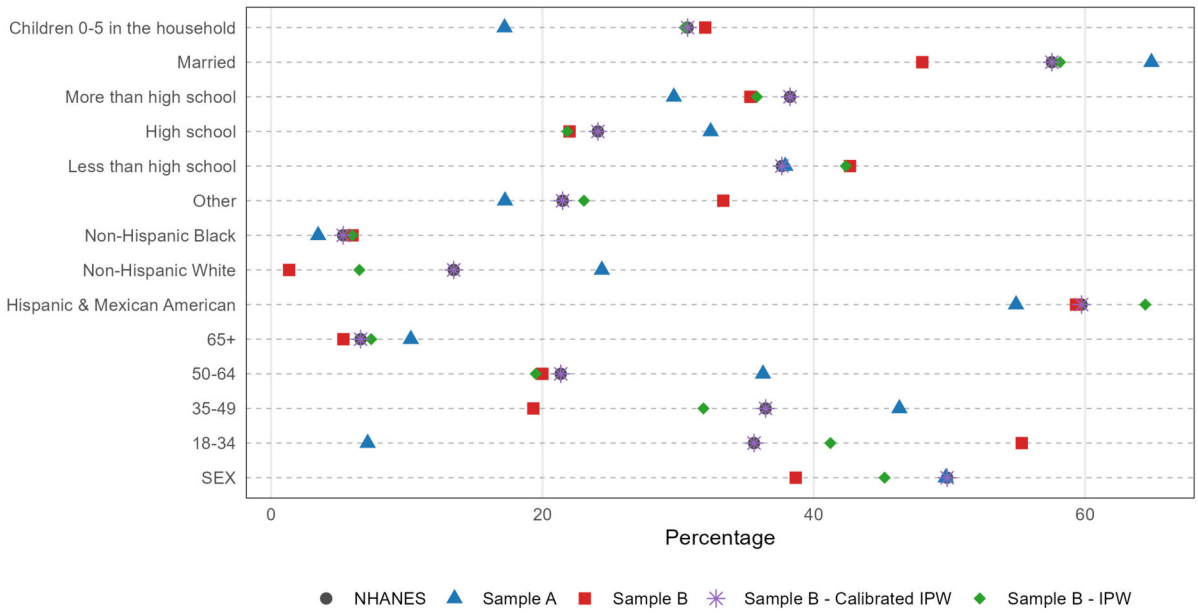

Figure 2: Auxiliary Variable Percentage Distributions – NHANES, and Samples A and B.

troducing detectable bias relative to the NHANES benchmark, suggesting that the combined sample yields more efficient estimates while preserving comparability in terms of point estimation. The supplementry materials provide more tables that include performance measures of the different estimates.

6 Conclusion

This article provides a general framework for estimation under hybrid designs that combine probability and nonprobability samples to oversample rare populations, with particular attention to settings in which nonprobability units are drawn from PSUs selected for a probability sample. By focusing on estimation from nonprobability samples, strategies for integrating probability and nonprobability data or estimates, appropriate analytical tools, and performance evaluation, we offer practical guidance for researchers working with similar data structures. While several important issues, such as ignorability assumptions, model diagnostics, and variance estimation, remain beyond the scope of this study, the framework presented here lays a foundation for more comprehensive methodological development. Future work can build on this foundation to explore these topics in greater depth and to conduct systematic comparisons of alternative approaches within broader experimental settings.

To illustrate the proposed framework and demonstrate its practical implementation, we conducted a simulation-based case study using data from NHANES 2017–18. The case study illustrates how the proposed framework can be applied in practice to a realistic hybrid design involving a probability sample supplemented by a nonprobability sample drawn within the same PSUs. Using NHANES 2017–18 data, we demonstrated how targeted nonprobability recruitment can address deficiencies in coverage and representation (i.e., the under-representation of younger non-citizens in the probability sample), while maintaining coherence with the original complex survey design. The results show that appropriate weighting adjustments substantially mitigate

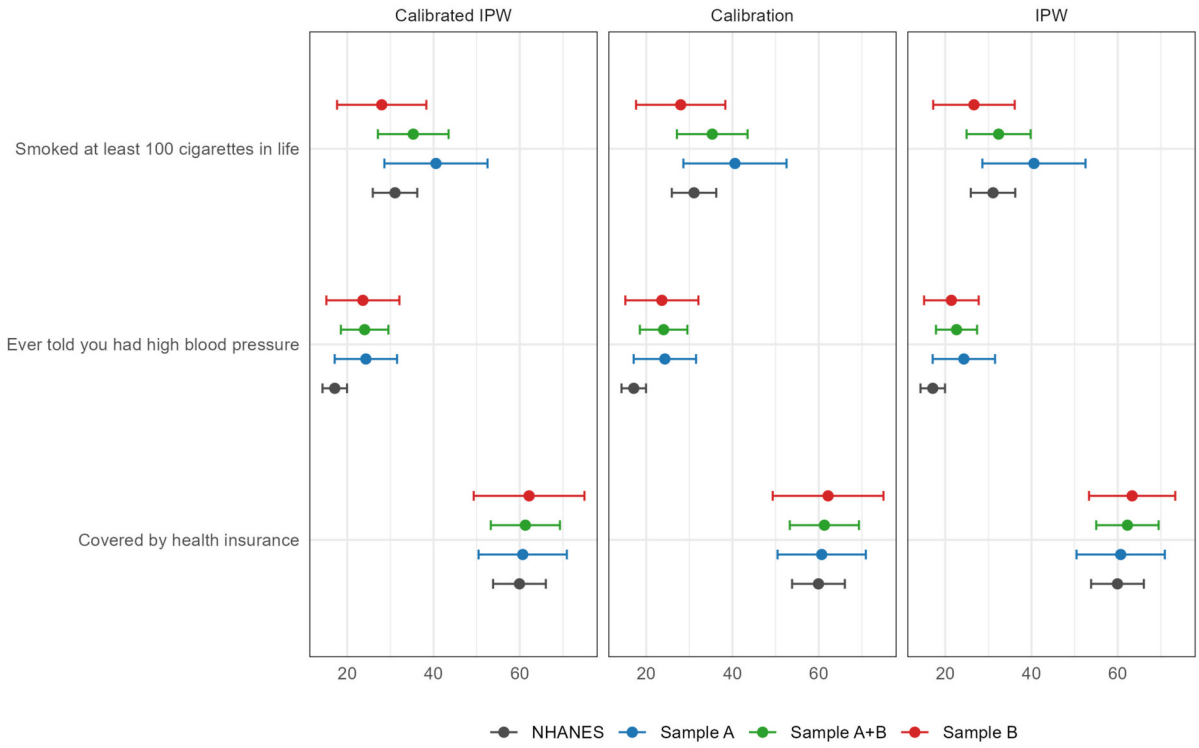

Figure 3: Study variables proportions and confidence intervals – NHANES, samples A and B and the combined sample.

the selection biases inherent in the nonprobability sample, yielding estimates that align closely with the NHANES benchmark. Moreover, combining the probability and nonprobability samples leads to meaningful gains in precision, as evidenced by narrower confidence intervals, without introducing detectable bias.

Finally, it is worth noting that the aim of the case study is not to evaluate the relative performance of the proposed methods, as such an assessment would require more extensive and carefully designed simulation studies, but rather to illustrate the estimation process in a realistic hybrid design setting. In particular, the case study is intended to demonstrate implementation details and practical considerations, supported by reproducible R code provided in the supplementary materials. Several important limitations and directions for future research remain. Additional work is needed to assess how the dependence induced by selecting nonprobability units within probability sample PSUs affects variance estimation, and to determine how such effects can be appropriately incorporated into the estimation methods recently proposed in the literature. Moreover, further research is warranted to investigate the use of replication-based variance estimators—such as balanced repeated replication or jackknife repeated replication—and to evaluate their suitability for accounting for stratification and clustering in both the probability and nonprobability samples within hybrid designs. Finally, future research should examine optimal recruitment strategies for nonprobability samples within this data structure, including how different recruitment mechanisms and targeting criteria influence representativeness, bias, and efficiency when nonprobability sampling is conducted within probability sample PSUs.

Supplementary Material

The online supplementary material contains annotated R syntax and results to illustrate estimation from non-probability samples and combining estimates from probability and non-probability samples.

References

Baker J, Brick J, Bates N, Battaglia M, Couper M , ..., Tourangeau R (2013). Summary report of the aapor task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1(2): 90–143. https://doi.org/10.1093/jssam/smt008

Beaumont J, Dhushenthen J (2024). nppr: Inference on non-probability sample data via integrating probability sample data. https://github.com/StatCan/nppR.

Beresovsky V, Gershunskaya J, Savitsky TD (2025). Review of quasi-randomization approaches for estimation from non-probability samples. *Statistical Science*. Forthcoming.

Buelens B, Burger J, van den Brakel J (2018). Comparing inference methods for non-probability samples. *International Statistical Review*, 86(2): 322–343. https://doi.org/10.1111/insr.12253

Castro-Martin L (2024). Inps: Inference from non-probability samples. python package version 1.0. https://github.com/luiscastro193/inps.

Castro-Martin L, Ferri-Garcia R, Rueda M (2020a). Estimation in nonprobability sampling: Package 'nonprobest'. (version 0.2.4.) https://CRAN.R-project.org/package=NonProbEst.

Castro-Martin L, Rueda M, Ferri-Garcia R (2020b). Inference from non-probability surveys with statistical matching and propensity score adjustment using modern prediction techniques. *Mathematics*, 8(6), 879. https://www.mdpi.com/2227-7390/8/6/879. https://doi.org/10.3390/math8060879

Castro-Martin L, Rueda M, Ferri-Garcia R (2022). Combining statistical matching and propensity score adjustment for inference from non-probability surveys. *Journal of Computational and Applied Mathematics*, 404, 113414. https://www.sciencedirect.com/science/article/pii/S0377042721000339. https://doi.org/10.1016/j.cam.2021.113414

Chen J, Valliant R, Elliott M (2019). Calibrating non-probability surveys to estimated control totals using lasso, with an application to political polling. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 68(3): 657–681. https://doi.org/10.1111/rssc.12327

Chen S, Haziza D (2022). General purpose multiply robust data integration procedures for handling nonprobability samples. *Scandinavian Journal of Statistics*, 50(2): 697–724. https://doi.org/10.1111/sjos.12605

Chen S, Woodruff A, Campbell J, Vesely S, Xu Z, Snider C (2023). Combining probability and nonprobability samples by using multivariate mass imputation approaches with application to biomedical research. *Stats*, 6(2): 617–625. https://doi.org/10.3390/stats6020039

Chen S, Yang S, Kim J (2022). Nonparametric mass imputation for data integration. *Journal of Survey Statistics and Methodology*, 10(1): 1–24. https://doi.org/10.1093/jssam/smaa036

Chen Y, Li P, Wu C (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532): 2011–2021. https://doi.org/10.1080/01621459.2019.1677241

Chen Y, Li P, Wu C (2023). Dealing with undercoverage for non-probability survey samples. *Survey Methodology*, 49(2): 497–515.

Chrostowski L, Beręsewicz M (2024). nonprobsvy: modern inference methods for non-probability samples in r (version 0.1.0). https://cran.r-project.org/package=nonprobsvy.

Cobo B, Ferri-García R, Rueda-Sánchez J, Rueda M (2024). Software review for inference with non-probability surveys. *The Survey Statistician*, 90: 40–47. https://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2024_July_N90_06.pdf.

Cornesse C, Blom A, Dutwin D, Krosnick J, De Leeuw E , ..., Wenz A (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, 8(1): 4–36. https://doi.org/10.1093/jssam/smz041

Dever J (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. In: *Proceedings of the 2018 Federal Committee on Statistical Methodology Research Conference*. https://nces.ed.gov/FCSM/pdf/A4_Dever_2018FCSM.pdf.

Deville JC, Särndal CE (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418): 376–382. https://doi.org/10.1080/01621459.1992.10475217

DiSogra C, Cobb C, Chan E, Dennis J (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. In: *Proceedings of the Survey Research Methods Section of the American Statistical Association*. http://www.asasrms.org/Proceedings/y2011/Files/302704_68925.pdf.

Elliot MR (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2(6).

Elliott M, Haviland A (2007). Use of a web-based convenience sample to supplement a probability sample. *Survey Methodology*, 33(2): 211–215.

Elliott M, Valliant R (2017). Inference for nonprobability samples. *Statistical Science*, 32(2): 249–264. https://doi.org/10.1214/16-STS598

Epanechnikov V (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, 14(1): 153–158. https://doi.org/10.1137/1114019

Ferri-García R, Rueda M (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PLoS ONE*, 15(4): e0231500. https://doi.org/10.1371/journal.pone.0231500

Ferri-García R, MdM R (2018). Efficiency of propensity score adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT-Statistics and Operations Research Transactions*, 42(2): 159–162.

Gershunskaya J, Beresovsky V, Savitsky Mason L TD (2025). Estimation from combined probability and non-probability samples under uncertain sampling overlap. In: *The Joint Statistical Meetings*, Nashville, TN, USA. Conference presentation.

Kang J, Schafer J (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4): 523–539. https://doi.org/10.1214/07-STS227

Kern C, Li Y, Wang L (2021). Boosted kernel weighting–using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*, 9(5): 1088–1113. https://doi.org/10.1093/jssam/smaa028

Kim J (2022). A gentle introduction to data integration in survey sampling. *The Survey Statistician*, 85: 19–29. https://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2022_January_N85_03.pdf.

Kim J, Haziza D (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24(1): 375–394.

Kim J, Park S, Chen Y, Wu C (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society. Series A. Statistics in Society*, 184(3): 941–963. https://doi.org/10.1111/rssa.12696

Kim J, Rao J (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika*, 99(1): 85–100. https://doi.org/10.1093/biomet/asr063

Kim J, Wang Z (2019). Sampling techniques for big data analysis. *International Statistical Review*, 87(S1): S177–S191. https://doi.org/10.1111/insr.12290

Kott P (2016). Calibration weighting in survey sampling. *WIREs: Computational Statistics*, 8(1): 39–53. https://doi.org/10.1002/wics.1374

Lee S (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22(2): 329–349.

Lee S, Valliant R (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3): 319–343. https://doi.org/10.1177/0049124108329643

Little R, Rubin D (2019). *Statistical Analysis with Missing Data*. Wiley, Hoboken, NJ, 3 edition.

Liu Z, Valliant R (2023). Investigating an alternative for estimation from a nonprobability sample: Matching plus calibration. *Journal of Official Statistics*, 39(1): 45–78. https://doi.org/10.2478/jos-2023-0003

Lohr S (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology*, 37(2): 197–213.

Meng XL (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9(4): 538–558. https://doi.org/10.1214/ss/1177010269

Mulrow E, Ganesh N, Pineau V, Yang M (2007). Using statistical matching to account for coverage bias when combining probability and nonprobability samples. In: *Proceedings of the Survey Research Methods Section of the American Statistical Association*. http://www.asasrms.org/Proceedings/y2020/files/1505359.pdf.

Nandram B, Choi J, Liu Y (2021). Integration of nonprobability and probability samples via survey weights. *International Journal of Statistics and Probability*, 10(6): 5–21. https://doi.org/10.5539/ijsp.v10n6p5

Nandram B, Rao J (2023). Bayesian predictive inference when integrating a non-probability sample and a probability sample. arXiv preprint: https://arxiv.org/abs/2305.08997.

Rafei A (2021). Robust and efficient bayesian inference for large-scale non-probability samples, Ph.D. thesis, University of Michigan. https://deepblue.lib.umich.edu/handle/2027.42/169715.

Rafei A, Elliott M, Flannagan C (2022). Robust and efficient bayesian inference for non-probability samples. arXiv preprint: https://arxiv.org/abs/2203.14355.

Rao J (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B: The Indian Journal of Statistics*, 83: 242–272. https://doi.org/10.1007/s13571-020-00227-w

Rivers D (2007). Sampling for web surveys. In: *Proceedings of the Survey Research Methods Section of the American Statistical Association*. http://www.websm.org/uploadi/editor/1368187629Rivers_2007_Sampling_for_web_surveys.pdf.

Robbins M, Ghosh-Dastidar B, Ramchand R (2021). Blending probability and nonprobability samples with applications to a survey of military caregivers. *Journal of Survey Statistics and Methodology*, 9(5): 1114–1145. https://doi.org/10.1093/jssam/smaa037

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

Rueda M, Ferri-Garcia R, Castro-Martin L (2020). The R package nonprobest for estimation in non-probability surveys. *The R Journal*, 12(1): 405–417. https://doi.org/10.32614/RJ-2020-015

Rueda M, Pasadas-del Amo S, Rodriguez B, Castro-Martin L, Ferri-Garcia R (2023). Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques: An application to a survey on the impact of the COVID-19 pandemic in Spain. *Biometrical Journal*, 65(2): 2200035. https://doi.org/10.1002/bimj.202200035

Sakshaug J, Wisniowski A, Perez Ruis D, Blom A (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics*, 35(3): 653–681. https://doi.org/10.2478/jos-2019-0027

Salvatore C, Biffignandi S, Sakshaug J, Wisniowski A, Struminskaya B (2024). Bayesian integration of probability and nonprobability samples for logistic regression. *Journal of Survey Statistics and Methodology*, 12(2): 458–492. https://doi.org/10.1093/jssam/smad041

Savitsky T, Williams M, Gershunskaya J, Beresovsky V (2023). Methods for combining probability and nonprobability samples under unknown overlaps. *Statistics in Transition*, 24(5): 1–34. https://doi.org/10.59170/stattrans-2023-061

Savitsky TD, Williams MR, Beresovsky V, Gershunskaya J (2025). Thresholding nonprobability units in combined data for efficient domain estimation. *Statistics in Transition*, 26(2): 1–19. https://doi.org/10.59139/stattrans-2025-013

Schonlau M, Couper M (2017). Options for conducting web surveys. *Statistical Science*, 32(2): 279–292. https://doi.org/10.1214/16-STS597

Sekhon J (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, 42(7): 1–52. https://doi.org/10.18637/jss.v042.i07

Skinner C, Rao J (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91(433): 349–356. https://doi.org/10.1080/01621459.1996.10476695

Tourangeau R, Edwards B, Johnson T, Wolter K, Bates N (2014). *Hard-to-Survey Populations*. Cambridge University Press, Cambridge, UK.

Valliant R (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2): 231–263. https://doi.org/10.1093/jssam/smz003

Valliant R, Dever J (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1): 105–137. https://doi.org/10.1177/0049124110392533

Wang L, Graubard BI, Katki HA, Li Y (2020). Improving external validity of epidemiologic cohort analyses: A kernel weighting approach. *Journal of the Royal Statistical Society. Series A. Statistics in Society*, 183(3): 1293–1311. https://doi.org/10.1111/rssa.12564

Wang L, Kern C (2023). Kwml: Boosted kernel weighting. r package version 1.0.1. https://github.com/chkern/KWML/.

Wang L, Valliant R, Li Y (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40(24): 5237–5250. https://doi.org/10.1002/sim.9122

Wiśniowski A, Sakshaug JW, Perez Ruiz DA, Blom AG (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8(1): 120–147. https://doi.org/10.1093/jssam/smz051

Wu C (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48(2): 283–311.

Yang S, Kim J (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3: 625–650. https://doi.org/10.1007/s42081-020-00093-w

Yang S, Kim J, Song R (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 82(2): 445–465. https://doi.org/10.1111/rssb.12354