Japanese Journal of Statistics and Data Science (2020) 3:625–650
https://doi.org/10.1007/s42081-020-00093-w

SURVEY PAPER

Theory and Practice of Surveys

Statistical data integration in survey sampling: a review

Shu Yang[1] · **Jae Kwang Kim**[2]

Received: 6 January 2020 / Accepted: 13 September 2020 / Published online: 15 October 2020

Abstract
Finite population inference is a central goal in survey sampling. Probability sampling is the main statistical approach to finite population inference. Challenges arise due to high cost and increasing non-response rates. Data integration provides a timely solution by leveraging multiple data sources to provide more robust and efficient inference than using any single data source alone. The technique for data integration varies depending on types of samples and available information to be combined. This article provides a systematic review of data integration techniques for combining probability samples, probability and non-probability samples, and probability and big data samples. We discuss a wide range of integration methods such as generalized least squares, calibration weighting, inverse probability weighting, mass imputation, and doubly robust methods. Finally, we highlight important questions for future research.

Keywords Generalizability · Meta-analysis · Missing at random · Transportability

1 Introduction

Probability sampling is regarded as the gold standard in survey statistics for finite population inference. Fundamentally, probability samples are selected under known sampling designs and, therefore, are representative of the target population. Because the selection probability is known, the subsequent inference from a probability sample is often design-based and respects the way in which the data were collected; see Särndal et al. (2003), Cochran (1977) and Fuller (2009) for textbook discussions. Kalton (2019) provided a comprehensive overview of the survey sampling research in the last 60 years.

However, many practical challenges arise in collecting and analyzing probability sample data (Baker et al. 2013; Keiding and Louis 2016). Large-scale survey

✉ Jae Kwang Kim
jkim@iastate.edu

[1] Department of Statistics, North Carolina State University, Raleigh, USA

[2] Department of Statistics, Iowa State University, Ames, USA

programs continually face heightened demands coupled with reduced resources. Demands include requests for estimates for domains with small sample sizes and desires for more timely estimates. Simultaneously, program budget cuts force reductions in sample sizes, and decreasing response rates make non-response bias an important concern.

Data integration is a new area of research to provide a timely solution to the above challenges. The goal is multi-fold: (1) minimize the cost associated with surveys, (2) minimize the respondent burden, and (3) maximize the statistical information or equivalently the efficiency of survey estimation. Narrowly speaking, survey integration means combining separate probability samples into one survey instrument (Bycroft 2010). Broadly speaking, one can consider combining probability samples with non-probability samples. Recently, in survey statistics, non-probability data become increasingly available for research purposes and provide unprecedented opportunities for new scientific discovery; however, they also present additional challenges such as heterogeneity, selection bias, high dimensionality, etc. The past years have seen immense progress in theories, methods, and algorithms for surmounting important challenges arising from non-probability data analysis. This article provides a systematic review of data integration for combining probability samples, probability and non-probability samples, and probability and big data samples.

Section 2 establishes notation and reviews these methods in the context of combining multiple probability samples. Existing methods for probability data integration can be categorized into two types depending on the level of information to be combined: a macro approach combining the summary statistics from multiple surveys and a micro approach creating synthetic imputations.

Section 3 describes the motivation, challenges, and methods for integrating probability and emergent non-probability samples. We also draw connections of survey data integration to combine randomized clinical trials and real-world data in Biostatistics. We then discuss a wide range of integration methods including calibration weighting, inverse probability weighting, mass imputation, and doubly robust methods.

We then consider data integration methods for combining probability and big non-probability samples. Depending on the roles in statistical inference, there are two types of *big data*: one with large sample sizes (large n) and the other with rich covariates (large p). In the first type, the non-probability sample can be large in sample size. How to leverage the rich information in the big data to improve the finite population inference is an important research. In the second type, there are a large number of variables. There is a large literature on variable selection methods for prediction, but little work on variable selection for data integration that can successfully recognize the strengths and the limitations of each data source and utilize all information captured for finite population inference. Section 4 presents robust data integration and variable selection methods in this context.

To summarize, Sect. 5 describes the direction of future research along the line of data integration including sensitivity analysis to assess the robustness of study conclusions to unverifiable assumptions, hierarchical modeling, and some cautionary remarks.

2 Combining probability samples

2.1 Multiple probability samples and missingness patterns

Combining two or more independent survey probability samples is a problem frequently encountered in the practice of survey sampling. For simplicity of exposition, let $\mathcal{U} = \{1, \ldots, N\}$ be the index set of N units for the finite population, with N being the known population size. Let $(x_i^{\mathrm{T}}, y_i)^{\mathrm{T}}$ be the realized value of a vector of random variables $(X^{\mathrm{T}}, Y)^{\mathrm{T}}$ for unit i, where X consists of auxiliary variables and Y is the study variable of interest. The parameter of interest is the finite population mean of Y, i.e., $\mu_y = N^{-1} \sum_{i=1}^{N} Y_i$ throughout the article. Let I_i be the sample indicator, such that $I_i = 1$ indicates the selection of unit i into the sample and $I_i = 0$ otherwise. The probability $\pi_i = P(I_i = 1 \mid i \in \mathcal{U})$ is called the first-order inclusion probability and is known by the sampling design. The design weight is $d_i = \pi_i^{-1}$. The joint probability $\pi_{ij} = P(I_i I_j = 1 \mid i, j \in \mathcal{U})$ is called the second-order inclusion probability and is often used for variance estimation of the design-weighted estimator. In particular, $\pi_{ii} = \pi_i$ for all i. The sample size is $n = \sum_{i=1}^{N} I_i$.

The main advantage of probability sampling is to ensure design-based inference. For example, the Horvitz–Thompson (HT) estimator of the population mean of y, denoted by μ_y, is $\widehat{\mu}_{\mathrm{HT}} = N^{-1} \sum_{i:I_i=1} \pi_i^{-1} y_i$, and the design-variance estimator is:

$$\widehat{V}_{\mathrm{HT}} = nN^{-2} \sum_{i:I_i=1} \sum_{j:I_j=1} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}.$$

We consider multiple sources of probability data. For multiple datasets, we use the subscript letter to indicate the respective sample; for example, we use $d_{A,i}$ as the design weight of unit i in sample A.

Depending on the available information from multiple data sources, each sample has planned missingness by design. As illustrated in Table 1, the combined sample exhibits different missingness patterns: monotone and non-monotone. For monotone missingness, our framework covers two common types of studies. First, we have a large main dataset, and then collect more information on important variables for a subset of units, e.g., using a two-phase sampling design (Neyman 1938; Cochran 1977; Wang et al. 2009). Consider the U.S. Census of housing and population as an example. The short form consists of 100% sample, for which basic demographic information was obtained. The long form consists of about 16% sample, for which other social and economic information as well as demographic information were obtained. Deming and Stephan (1940) considered this setup as a classical two-phase sampling problem and use calibration weighting for demographic variable to match the known population counts from the short form.

Second, we have a smaller and carefully designed validation dataset with rich covariates, and then link it to a larger main dataset with fewer covariates. The setup of two independent samples with common items is often called non-nested two-phase sampling. Consider the US consumer expenditure survey as an example. Two independent samples were selected from the same finite population, including a diary survey sample, referred to as sample A, and a face-to-face

Table 1 Missingness patterns in the combined samples: "✓" means "is measured"

Monotone missingness				
	d	X	Y	
Sample A	✓	✓	✓	
Sample B	✓	✓		
Non-monotone missingness I				
	d	X	Y_1	Y_2
Sample A	✓	✓	✓	✓
Sample B	✓	✓	✓	
Sample C	✓	✓		✓
Non-monotone missingness II				
	d	X	Y_1	Y_2
Sample A	✓	✓	✓	
Sample B	✓	✓		✓

d is the design weight, where the subscript indicates the sample, X is the vector of auxiliary variables, and Y, Y_1, and Y_2 are scalar outcome variables

survey sample, referred to as sample B. In sample A, observe auxiliary information X and outcome Y, whereas in sample B, observe common auxiliary information X. Zieschang (1990) considered using sample weighting to estimate detailed expenditure and income items combining sample A and sample B. Another example is the Canadian Survey of Employment, Payrolls, and Hours considered by Hidiroglou (2001). Sample A is a small sample from Statistics Canada Business Register, in which the study variables Y, number of hours worked by employees, and summarized earnings were observed. Sample B is a large sample drawn from a Canadian Customs and Revenue Agency administrative data, in which auxiliary variables X were observed.

Finally, we will consider combining two independent surveys with non-monotone missing patterns. Statistical matching technique will be introduced in Sect. 2.2.1 as a general statistical tool under this setup.

2.2 Two approaches for probability data integration

We classify probability data integration methods based on the level of information to be combined: a macro approach and a micro approach. In the macro approach, we obtain summary information such as the point and variance estimates from multiple data sources and combine those to obtain a more efficient estimator of the parameter of interest, such as population means or totals. In the micro approach, we create single synthetic data that contain all available information from all data sources. The synthetic data can be used to estimate various types of the parameters.

2.2.1 Macro approach: generalized least-squares (GLS) estimation

Renssen and Nieuwenbroek (1997), Hidiroglou (2001), Merkouris (2004), Wu (2004), Ybarra and Lohr (2008), and Merkouris (2010) considered the problem of combining data from two independent probability samples to estimate totals at the population and domain levels. Merkouris (2004) and Merkouris (2010) provided a rigorous treatment of the survey integration through the generalized method of moments.

We focus on the monotone missingness pattern. The same discussion applies to the other patterns. From each probability sample, we obtain different estimators for the means of common items. The GLS approach combines those estimates as an optimal estimator. Let $\widehat{\mu}_{x,A}$ and $\widehat{\mu}_{x,B}$ be unbiased estimators of μ_x from sample A and sample B, respectively. Let $\widehat{\mu}_B$ be an unbiased estimator of μ_y from sample B.

To combine the multiple estimates, we can build a linear model of three estimates with two parameters as follows:

$$\begin{pmatrix} \widehat{\mu}_{x,A} \\ \widehat{\mu}_{x,B} \\ \widehat{\mu}_B \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}, \tag{1}$$

where $(e_1, e_2, e_3)^{\mathrm{T}}$ has mean $(0, 0, 0)^{\mathrm{T}}$, variance–covariance:

$$V = \begin{pmatrix} \mathrm{var}(\widehat{\mu}_{x,A}) & \mathrm{cov}(\widehat{\mu}_{x,A}, \widehat{\mu}_{x,B}) & \mathrm{cov}(\widehat{\mu}_{x,A}, \widehat{\mu}_B) \\ \mathrm{cov}(\widehat{\mu}_{x,A}, \widehat{\mu}_{x,B}) & \mathrm{var}(\widehat{\mu}_{x,B}) & \mathrm{cov}(\widehat{\mu}_{x,B}, \widehat{\mu}_B) \\ \mathrm{cov}(\widehat{\mu}_{x,A}, \widehat{\mu}_B) & \mathrm{cov}(\widehat{\mu}_{x,B}, \widehat{\mu}_B) & \mathrm{var}(\widehat{\mu}_B) \end{pmatrix},$$

and $\mathrm{var}(\cdot)$ and $\mathrm{cov}(\cdot)$ are the variance and covariance induced by the sampling probability. If the two samples are independently obtained from the sample population, we have $\mathrm{cov}(\widehat{\mu}_{x,A}, \widehat{\mu}_{x,B}) = 0$ and $\mathrm{cov}(\widehat{\mu}_{x,A}, \widehat{\mu}_B) = 0$.

Based on model (1), treat $(\widehat{\mu}_{x,A}, \widehat{\mu}_{x,B}, \widehat{\mu}_B)$ as observations and define a sum of squared error term:

$$Q(\mu_x, \mu_y) = \begin{pmatrix} \widehat{\mu}_{x,A} - \mu_x \\ \widehat{\mu}_{x,B} - \mu_x \\ \widehat{\mu}_B - \mu_y \end{pmatrix}^{\mathrm{T}} V^{-1} \begin{pmatrix} \widehat{\mu}_{x,A} - \mu_x \\ \widehat{\mu}_{x,B} - \mu_x \\ \widehat{\mu}_B - \mu_y \end{pmatrix}.$$

The optimal estimator of (μ_x, μ_y) that minimizes $Q(\mu_x, \mu_y)$ is:

$$\widehat{\mu}_x^* = \alpha^* \widehat{\mu}_{x,A} + (1 - \alpha^*) \widehat{\mu}_{x,B} \tag{2}$$

and

$$\widehat{\mu}_{\mathrm{GLS}} = \widehat{\mu}_B + \begin{pmatrix} \widehat{\mathrm{cov}}(\widehat{\mu}_{x,A}, \widehat{\mu}_B) \\ \widehat{\mathrm{cov}}(\widehat{\mu}_{x,B}, \widehat{\mu}_B) \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \widehat{\mathrm{var}}(\widehat{\mu}_{x,A}) & \widehat{\mathrm{cov}}(\widehat{\mu}_{x,A}, \widehat{\mu}_{x,B}) \\ \widehat{\mathrm{cov}}(\widehat{\mu}_{x,A}, \widehat{\mu}_{x,B}) & \widehat{\mathrm{var}}(\widehat{\mu}_{x,B}) \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\mu}_x^* - \widehat{\mu}_{x,A} \\ \widehat{\mu}_x^* - \widehat{\mu}_{x,B} \end{pmatrix}, \tag{3}$$

where

$$\alpha^* = \frac{\widehat{\text{var}}(\widehat{\mu}_{x,B}) - \widehat{\text{cov}}(\widehat{\mu}_{x,A}, \widehat{\mu}_{x,B})}{\widehat{\text{var}}(\widehat{\mu}_{x,A}) + \widehat{\text{var}}(\widehat{\mu}_{x,B}) - 2\widehat{\text{cov}}(\widehat{\mu}_{x,A}, \widehat{\mu}_{x,B})}.$$

To see the efficiency gain of $\widehat{\mu}_{\text{GLS}}$ over $\widehat{\mu}_B$, using (2), we express:

$$\widehat{\mu}_{\text{GLS}} = \widehat{\mu}_B - \widehat{\text{cov}}(\widehat{\mu}_B, \widehat{\mu}_{x,B} - \widehat{\mu}_{x,A})\{\widehat{\text{var}}(\widehat{\mu}_{x,B} - \widehat{\mu}_{x,A})\}^{-1}(\widehat{\mu}_x^* - \widehat{\mu}_{x,B}).$$

The variance of $\widehat{\mu}_{\text{GLS}}$ is:

$$\text{var}(\widehat{\mu}_B) - \text{cov}(\widehat{\mu}_B, \widehat{\mu}_{x,B} - \widehat{\mu}_{x,A})\{\text{var}(\widehat{\mu}_{x,B} - \widehat{\mu}_{x,A})\}^{-1}\text{cov}(\widehat{\mu}_B, \widehat{\mu}_{x,B} - \widehat{\mu}_{x,A}),$$

which is not larger than $\text{var}(\widehat{\mu}_B)$. The GLS estimator for non-monotone missingness can be constructed similarly. See Fuller and Breidt (1999) for an application in the National Resource Inventory.

2.2.2 Micro approach: mass imputation

Mass imputation (also called synthetic data imputation) is a technique of creating imputed values for items not observed in the current survey by incorporating information from other surveys. Breidt et al. (1996) discussed mass imputation for two-phase sampling. Rivers (2007) proposed a mass imputation approach using nearest-neighbor imputation, but the theory is not fully developed. Schenker and Raghunathan (2007) reported several applications of synthetic data imputation, using a model-based method to estimate totals and other parameters associated with variables not observed in a larger survey but observed in a much smaller survey. Legg and Fuller (2009) and Kim and Rao (2012) developed synthetic imputation approaches to combining two surveys. Chipperfield et al. (2012) discussed composite estimation when one of the surveys is mass imputed. Bethlehem (2016) discussed practical issues in sample matching for mass imputation.

The primary goal is to create a single synthetic dataset of proxy values $\widehat{y}_i$ for the unobserved y_i in sample B and then use the proxy data together with the associated design weights of sample A to produce projection estimators of the population mean μ_y. This is particularly useful when sample B is a large-scale survey and item Y is very expensive to measure. The proxy values $\widehat{y}_i$ are generated by first fitting a working model relating Y to X, $E(Y \mid X) = m(X;\beta_0)$ based on the data $\{(x_i, y_i) : i \in A\}$ from sample A. Then, the synthetic values of Y can be created by $\widehat{y}_i = m(x_i;\widehat{\beta})$ for $i \in B$. Thus, sample A is used as a training sample for predicting Y in sample B. The mass imputation estimator of μ_y is $\widehat{\mu}_{\text{I}} = N^{-1}\sum_{i\in B} d_{B,i}\widehat{y}_i$. Kim and Rao (2012) showed that $\widehat{\mu}_{\text{I}}$ is asymptotically design-unbiased if $\widehat{\beta}$ satisfies:

$$\sum_{i\in A} d_{A,i}\{y_i - m(x_i;\widehat{\beta})\} = 0. \tag{4}$$

With (4):

$$\begin{aligned}\widehat{\mu}_{\mathrm{I}} &= N^{-1}\sum_{i\in B} d_{B,i}\widehat{y}_i + N^{-1}\sum_{i\in A} d_{A,i}(y_i - \widehat{y}_i)\\ &= N^{-1}\sum_{i\in B} d_{B,i}m(x_i;\beta_0) + N^{-1}\sum_{i\in A} d_{A,i}\{y_i - m(x_i;\beta_0)\} = \widehat{P}_B + \widehat{Q}_A,\end{aligned}$$

and

$$\mathrm{var}(\widehat{\mu}_{\mathrm{I}}) = \mathrm{var}(\widehat{P}_B) + \mathrm{var}(\widehat{Q}_A).$$

The asymptotic unbiasedness holds regardless of whether the regression model is true or not. However, a good regression model will reduce the variance of $\widehat{\mu}_{\mathrm{I}}$. For variance estimation, either linearization or replication-based sampling (Kim and Rao 2012) can be used.

2.3 Mass imputation with non-monotone missingness

For non-monotone missingness, the mass imputation method of Kim and Rao (2012) is not directly applicable as the sample with partial observations may contain additional information for parameter estimation. Often, one can consider a joint model of all variables and use the EM algorithm to estimate the model parameters. The joint model deduces the conditional distribution of the missing variables given the observed values for imputation.

For illustration, consider the non-monotone missingness I structure in Table 1. The goal is to develop mass imputation for both Y_2 in sample B and Y_1 in sample C. It is attempting to specify the conditional distribution of Y_2 given (X, Y_1) to impute Y_2 in sample B and the conditional distribution of Y_1 given (X, Y_2) to impute Y_1 in sample C. However, this approach may result in model incompatiability. That is, there does not exist a joint model of (Y_1, Y_2) given X that leads to the corresponding conditional distributions. To avoid model incompatibility, we use a joint model for (Y_1, Y_2) given X for prediction though specifying the sequential conditional distribution:

$$f(Y_1, Y_2 \mid X;\theta) = f_1(Y_1 \mid X;\theta_1)f_2(Y_2 \mid X, Y_1;\theta_2), \tag{5}$$

where $\theta = (\theta_1^{\mathrm{T}}, \theta_2^{\mathrm{T}})^{\mathrm{T}}$, θ_1, and θ_2 are unknown parameters.

For parameter estimation, it suffices to use observations in sample A; however, this approach ignores the partial information in sample B and sample C and, therefore, is not efficient. Let the joint set of sampling indexes be $S = A \cup B \cup C$. Assuming no overlap between the samples, we define:

$$\pi_{S,i} = P(i \in S \mid i \in \mathcal{U}) = \begin{cases} \pi_{A,i} & \text{if } i \in A \\ \pi_{B,i} & \text{if } i \in B \\ \pi_{C,i} & \text{if } i \in C, \end{cases}$$

and let d_i be the design weight for unit $i \in S$ without specifying which sample it belongs to. That is, $d_i = d_{A,i}$ if $i \in A$. To incorporate all available information, the EM algorithm can be used as follows.

E-step Let $\theta^{(t)}$ be the parameter estimate at iteration t. Compute the conditional expectation of the pseudo-log-likelihood functions:

$$Q_1(\theta_1 \mid \theta^{(t)}) = \sum_{i \in S} d_i E\left\{\log f_1(y_{1i} \mid x_i;\theta_1) \mid x_i, y_{i,\text{obs}};\theta^{(t)}\right\}$$
$$Q_2(\theta_2 \mid \theta^{(t)}) = \sum_{i \in S} d_i E\left\{\log f_2(y_{2i} \mid x_i, y_{1i};\theta_2) \mid x_i, y_{i,\text{obs}};\theta^{(t)}\right\},$$

where $y_{i,\text{obs}}$ is the observed part of (y_{1i}, y_{2i}).

M-step Update the parameter θ by maximizing $Q_1(\theta_1 \mid \theta^{(t)})$ and $Q_2(\theta_2 \mid \theta^{(t)})$ with respect to θ_1 and θ_2.

The E-step and M-step can be iteratively computed until convergence, leading to the pseudo maximum likelihood estimator $\widehat{\theta}$.

Given $\widehat{\theta}$, mass imputation can be done for both Y_2 in sample B and Y_1 in sample C. The imputation model for Y_2 in sample B is $f_2(Y_2 \mid X, Y_1;\widehat{\theta}_2)$. Also, the imputation model for Y_1 in sample C is:

$$f(Y_1 \mid X, Y_2;\widehat{\theta}) = \frac{f_1(Y_1 \mid X;\widehat{\theta}_1) f_2(Y_2 \mid X, Y_1;\widehat{\theta}_2)}{\int f_1(Y_1 \mid X;\widehat{\theta}_1) f_2(Y_2 \mid X, Y_1;\widehat{\theta}_2) \mathrm{d}Y_1}. \tag{6}$$

To generate imputed values from (6), one may use Markov Chain Monte Carlo methods or the parametric fractional imputation of Kim (2011).

We now consider the non-monotone missingness II structure in Table 1. Sample A and sample B are probability samples which were selected from the same finite population. In sample A, observe (X, Y_1) and in sample B, observe (X, Y_2). The question of interest is the associational relationship of Y_1 and (X, Y_2). If (X, Y_1, Y_2) were jointly observed, one can fit a simple regression model of Y_2 on (X, Y_2). However, based on the available data, Y_1 and Y_2 were not available simultaneously.

This problem fits into the statistical matching framework (D'Orazio et al. 2006). In statistical matching, the goal is to create Y_1 for each unit in sample B by finding a "statistical twin" from the sample A. Typically, one assumes the conditional independence assumption that Y_1 and Y_2 are conditionally independent given X, or equivalently:

$$f(Y_1 \mid X, Y_2) = f(Y_1 \mid X). \tag{7}$$

Then, the "statistical twin" is solely determined by "how close" they are in terms of X's. However, in a regression model of Y_1 on (X, Y_2), (7) sets the regression coefficient associated with Y_2 to be zero a priori, which is contrary to the study question of interest.

For a joint modeling of (X, Y_1, Y_2) without assuming (7), identification is an important issue. Consider the following joint model of (Y_1, Y_2) given X :

$$Y_1 = \alpha_0 + \alpha_1 X + e_1, \tag{8}$$

$$Y_2 = \beta_0 + \beta_1 X + \beta_2 Y_1 + e_2, \tag{9}$$

where $\text{cov}(e_1, e_2) = 0$. Because (X, Y_1) is observed in sample A, (α_0, α_1) is identifiable. Because (X, Y_2) is observed in sample B, $f(Y_2 \mid X)$ is identifiable.

Coupling (8) and (9) leads to:

$$Y_2 = (\beta_0 + \alpha_0\beta_2) + (\beta_1 + \alpha_1\beta_2)X + \beta_2 e_1 + e_2.$$

Thus, only $\beta_0 + \alpha_0\beta_2$ and $\beta_1 + \alpha_1\beta_2$ are identifiable, and $(\beta_0, \beta_1, \beta_2)$ is not.

In general, non-linear relationships can help achieve identification. For example, if the linear relationship of X–Y_1in (8) is:

$$Y_1 = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + e_1. \tag{10}$$

Again, $(\alpha_0, \alpha_1, \alpha_2)$ is identifiable from sample A. Coupling (9) and (10) leads to:

$$Y_2 = (\beta_0 + \alpha_0\beta_2) + (\beta_1 + \alpha_1\beta_2)X + (\alpha_2\beta_2)X^2 + \beta_2 e_1 + e_2.$$

Thus, $\beta_0 + \alpha_0\beta_2$, $\beta_1 + \alpha_1\beta_2$ and $\alpha_2\beta_2$ are identifiable from sample B. As long as $\alpha_2 \neq 0$, $(\beta_0, \beta_1, \beta_2)$ is then identifiable. For an identifiable model, parameter estimation can be implemented either using the EM algorithm or GLS.

Other assumptions can be invoked to achieve model identification. Kim et al. (2016) used an instrumental variable assumption for model identification and develop fractional imputation methods for statistical matching. Park et al. (2016) presented an application of the statistical matching technique using fractional imputation in the context of handling mixed-mode surveys. Park et al. (2017) applied the method to combine two surveys with measurement errors.

3 Combining probability and non-probability samples

3.1 Combining a probability sample with a non-probability sample

Statistical analysis of non-probability survey samples faces many challenges as documented by Baker et al. (2013). Non-probability samples have unknown selection/inclusion mechanisms and are typically biased, and they do not represent the target population. A popular framework in dealing with the biased non-probability samples is to assume that auxiliary variable information on the same population is available from an existing probability survey sample. This framework was first used by Rivers (2007) and followed by a number of other authors including Vavreck and Rivers (2008), Lee and Valliant (2009), Valliant and Dever (2011), Elliott and Valliant (2017), and Chen et al. (2018), among others. Combining the up-to-date information from a non-probability sample and auxiliary information from a probability sample can be viewed as data integration, which is an emerging area of research in survey sampling (Lohr and Raghunathan 2017).

Data integration for finite population inference is similar to the problem of combining randomized experiments and non-randomized real-world evidence studies

for causal inference of treatment effects (Keiding and Louis 2016). In randomized clinical trial, the treatment assignment mechanism is known and, therefore, treatment effect evaluation based on randomized clinical trial is unconfounded. However, due to restrictive inclusion and exclusion criteria, the trial sample may be narrowly defined and can not represent the real-world patient population. On the other hand, by the real-world data collection mechanism, the real-world evidence study is often representative of the target population. Combining trial and real-world evidence studies can achieve more robust and efficient inference of treatment effect for a target patient population. Table 2 draws a parallel comparison of data sources between data integration in survey sampling and that in treatment effect evaluation.

Survey statisticians and biostatisticians have provided different methods for combining information from multiple data sources. Lohr and Raghunathan (2017) and Rao (2020) provided comprehensive reviews of statistical methods for finite population inference. In biostatistics, meta-analysis has been a long-standing method to synthesize evidences from multiple trial and observational data. Meta-analysis combines *aggregate information* to accommodate heterogeneity in treatment effects estimated from trial and observational data; see Verde and Ohmann (2015) for an overview of different modeling techniques in meta-analysis. Existing methods for data integration of a probability sample and a non-probability sample can be categorized into three types as follows. The first type is the so-called propensity score adjustment (Rosenbaum and Rubin 1983). In this approach, the probability of a unit being selected into the non-probability sample, which is referred to as the propensity or sampling score, is modeled and estimated for all units in the non-probability sample. The subsequent adjustments, such as propensity score weighting or stratification, can then be used to adjust for selection biases; see, e.g., Lee and Valliant (2009), Elliott and Valliant (2017) and Chen et al. (2018). Stuart et al. (2011, (2015) and Buchanan et al. (2018) used propensity score weighting to generalize results from randomized trials to a target population. O'Muircheartaigh and Hedges (2014) proposed propensity score stratification for analyzing a non-randomized social experiment. One notable disadvantage of the propensity score methods is that they rely on an explicit propensity score model and are biased and highly variable if the model is mis-specified (Kang and Schafer 2007). The second type uses calibration weighting (Deville and Särndal 1992; Kott 2006). This technique calibrates auxiliary information in the non-probability sample with that in the probability sample, so that

Table 2 Data integration in survey sampling and biostatistics

Survey sampling	Treatment effect evaluation	Representative of the finite population	Unbiased estimation[a]
Probability sample	Real-world evidence study	✓	
Non-probability sample	Randomized experiment		✓

[a]In survey sampling, some probability samples may not observe the study variable of interest; for treatment effect evaluation, randomized experiments provide unbiased estimation of treatment effect due to treatment randomization

after calibration, the weighted distribution of the non-probability sample is similar to that of the target population. The third type is mass imputation, which imputes the missing values for all units in the probability sample. In the usual imputation for missing data analysis, the respondents in the sample constitute a training dataset for developing an imputation model. In the mass imputation, an independent non-probability sample is used as a training dataset, and imputation is applied to all units in the probability sample; see, e.g., Breidt et al. (1996), Rivers (2007), Kim and Rao (2012), Chipperfield et al. (2012), Bethlehem (2016) and Yang and Kim (2018).

3.2 Setup and assumptions

Non-probability samples become increasingly popular in survey statistics, but may suffer from selection bias that limits the generalizability of results to the target population. We consider integrating a non-probability sample with a carefully designed probability sample which provides the representative covariate information of the target population.

Let $X \in \mathbb{R}^p$ be a vector of auxiliary variables (including an intercept) that are available from two data sources, and let $Y \in \mathbb{R}$ be the study variable of interest. We consider combining a probability sample with X, referred to as sample A, and a non-probability sample with (X, Y), referred to as sample B, to estimate μ_y the population mean of Y. We focus on the case where the study variable Y is observed in sample B only, but the other auxiliary variables are commonly observed in both data. Although the big data source has a large sample size, the sampling mechanism is often unknown, and we cannot compute the first-order inclusion probability for Horvitz–Thompson estimation. The naive estimators without adjusting for the sampling process are subject to selection biases, as illustrated in Table 3. On the other hand, although the probability sample with design weights represents the finite population, it does not observe the study variable. The complementary features of probability samples and non-probability samples raise the question of whether it is possible to develop data integration methods that leverage the advantages of both sources.

Because the sampling mechanism of a non-probability sample is unknown, the target population quantity is not identifiable in general. Unlike the previous case in Sect. 2, the sampling mechanism of sample B is unknown and, therefore, μ_y is not identifiable in general.

Table 3 Illustration of the total error from the simple mean estimator of $\bar{Y}_N$ based on probability simple random sample and big non-probability sample

Total error (MSE)	=	Variance	+	Bias^2
Probability sample		$\{(1-f_A)/n_A\}S_Y^2$		0
Non-probability sample (Big data)		≈ 0		$r_B^2\{(1-f_B)/f_B\}S_Y^2$

$f_A = n_A/N$ and $f_B = n_B/N$ are the sampling fractions of sample A and sample B, respectively; r_B is the correlation between the outcome Y and the inclusion indicator I_B; S_Y is the population variance of Y

Two datasets were considered from the 2005 Pew Research Centre (PRC) and the 2005 Behavioral Risk Factor Surveillance System (BRFSS). The goal of the PRC study was to evaluate the relationship between individuals and community (Chen et al. 2018; Kim et al. 2018). The 2005 PRC data are non-probability sample data provided by eight different vendors, which consist of $n_B = 9301$ subjects. Yang et al. (2019) focus on two study variables, a continuous Y_1 (days had at least one drink last month) and a binary Y_2 (an indicator of voted local elections). The 2005 BRFSS sample is a probability sample, which consists of $n_A = 441{,}456$ subjects with survey weights. This dataset does not have measurements on the study variables of interest; however, it contains a rich set of common covariates with the PRC dataset. The covariate distributions from the PRC sample and the BRFSS sample are considerably different, e.g., age, education (high school or less), financial status (no money to see doctors, own house), retirement rate, and health (smoking). Therefore, the PRC dataset is not representative of the target population, and the naive analyses of the study variables are subject to selection biases.

Let $f(Y \mid X)$ be the conditional distribution of Y given X in the superpopulation model ζ that generates the finite population. We make the following assumption.

Assumption 1 (i) The sampling indicator I_B of sample B and the study variable Y is independent given X; i.e., $P(I_B = 1 \mid X, Y) = P(I_B = 1 \mid X)$, referred to as the sampling score $\pi_B(X)$; and (ii) $\pi_B(X) > 0$ for all X.

Assumptions 1 (i) and (ii) 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 setup has previously been used by several authors; see, e.g., Rivers (2007) and Vavreck and Rivers (2008). Assumption 1 (i) states the ignorability of the selection mechanism to sample B conditional upon the covariates. Under Assumption 1 (i), $E(Y \mid X) = E(Y \mid X, I_B = 1)$, denoted by $m(X)$, can be estimated based on sample B. Assumption 1 (ii) implies that the support of in sample B is the same as that in the finite population. Assumption 1 (ii) does not hold if certain units would never be included in the non-probability sample. The plausibility of this assumption can be easily checked by comparing the marginal distributions of the auxiliary variables in sample B with those in sample A.

Under the sampling ignorability assumption, there are two main approaches: (1) the weighting approach by constructing weights for sample B to improve the representativeness of sample B; (2) the imputation approach by creating mass imputation for sample A using the observations in sample B. There is considerable interest in bridging the findings from a randomized clinical trial to the target population. This problem has been termed as generalizability (Cole and Stuart 2010; Stuart et al. 2011; Hernan and VanderWeele 2011; Tipton 2013; O'Muircheartaigh and Hedges 2014; Stuart et al. 2015; Keiding and Louis 2016; Buchanan et al. 2018), external validity (Rothwell 2005), or transportability (Pearl and Bareinboim 2011; Rudolph and van der Laan 2017) in the statistics literature, and has connections to the covariate shift problem in machine learning (Sugiyama and Kawanabe 2012).

3.3 Propensity score weighting

Under Assumption 1 (i) and (ii), we can build a model for $\pi_B(X) = P(I_B = 1 \mid X)$ and use it to adjust for the selection bias in sample B. In practice, the propensity score function $\pi_B(X)$ is unknown and needs to be estimated from the data. Let $\pi_B(X;\alpha)$ be the posited models for $\pi_B(X)$, where α is the unknown parameter. Several authors have proposed different estimation strategies. For example, $\widehat{\alpha}$ can be obtained by a weighted regression of $I_{B,i}$ on x_i combining sample A and sample B ($I_{B,i} = 0$ for $i \in A$ and $I_{B,i} = 1$ for $i \in B$), weighted by the design weights from sample A, which is valid if the size of sample B is relatively small (Valliant and Dever 2011). Chen et al. (2018) proposed estimating α by solving:

$$\widehat{S}_1(\alpha) = \sum_{i \in B} x_i - \sum_{i \in A} d_{A,i} \pi_B(x_i;\alpha) x_i = 0, \tag{11}$$

which is a sample version of the population estimating equation $S(\alpha) = \sum_{i \in U} \left\{ I_{B,i} - \pi(x_i;\alpha) \right\} x_i = 0$. Instead of using (11), one can also use:

$$\widehat{S}_2(\alpha) = \sum_{i \in B} \frac{1}{\pi_B(x_i;\alpha)} x_i - \sum_{i \in A} d_{A,i} x_i = 0,$$

which is closely related to the calibration weighting approach for nonresponse adjustment.

Given $\widehat{\alpha}$, the inverse probability of sampling weighting estimator of μ_y is:

$$\widehat{\mu}_{\mathrm{IPW}} = \widehat{\mu}_{\mathrm{IPW}}(\widehat{\alpha}) = N^{-1} \sum_{i=1}^{N} \frac{I_{B,i}}{\pi_B(x_i;\widehat{\alpha})} y_i. \tag{12}$$

Variance estimation of $\widehat{\mu}_{\mathrm{IPW}}$ can be obtained by the standard M-estimation theory.

One of the notable disadvantages of the propensity score methods is that they rely on an explicit propensity score model and are biased if the model is misspecified (Kang and Schafer 2007). Moreover, if the estimated propensity score is close to zero, $\widehat{\mu}_{\mathrm{IPW}}$ will be highly unstable.

3.4 Calibration weighting

The second weighting strategy is calibration weighting or bench marking weighting (Deville and Särndal 1992; Kott 2006). This technique can be used to calibrate auxiliary information in the non-probability sample with that in the probability sample, so that, after calibration, the non-probability sample is similar to the target population.

Instead of estimating the propensity score model and inverting the propensity score to correct for the selection bias of the non-probability sample, the calibration strategy estimates the weights directly. Toward this end, we assign a weight $\omega_{B,i}$ to each unit i in the sample B, so that:

$$\sum_{i\in B} \omega_{B,i} x_i = \sum_{i\in A} d_{A,i} x_i. \tag{13}$$

where $\sum_{i\in A} d_{A,i} x_i$ is a design-weighted estimate of the population total of X from the probability sample. Constraint (13) is referred to as the covariate balancing constraint (Imai and Ratkovic 2014), and weights $\mathcal{Q}_B = \{\omega_{B,i} : i \in B\}$ are the calibration weights. The balancing constraint calibrates the covariate distribution of the non-probability sample to the target population in terms of X. Instead of calibrating each X, one can calibrate model-based calibration (McConville et al. 2017; Chen et al. 2018, 2019). In this approach, one can posit a parametric model for $E(Y \mid X) = m(X;\beta)$ and estimate the unknown parameter β based on sample B. The model-based calibration specifies the constraints for $\mathcal{Q}_B$ as:

$$\sum_{i\in B} \omega_{B,i} m(x_i;\widehat{\beta}) = \sum_{i\in A} d_{A,i} m(x_i;\widehat{\beta}). \tag{14}$$

We estimate $\mathcal{Q}_B$ by solving the following optimization problem:

$$\min_{\mathcal{Q}_B} \left\{ L(\mathcal{Q}_B) = \sum_{i\in B} \omega_{B,i} \log \omega_{B,i} \right\}, \tag{15}$$

subject to $\omega_{B,i} \geq 0$, for all $i \in B$; $\sum_{i\in B} \omega_{B,i} = N$, and the balancing constraint (13) or (14).

The objective function in (15) is the negative entropy of the calibration weights; thus, minimizing this criteria ensures that the empirical distribution of calibration weights is not too far away from the uniform, such that it minimizes the variability due to heterogeneous weights. This optimization problem can be solved using convex optimization with Lagrange multiplier. Other objective functions, such as $L(\mathcal{Q}_B) = \sum_{i\in B} \omega_{B,i}^2$, can also be considered. This optimization problem can be solved using convex optimization with Lagrange multiplier. By introducing Lagrange multiplier λ, the objective function becomes:

$$L(\lambda, \mathcal{Q}_B) = \sum_{i\in B} \omega_{B,i} \log \omega_{B,i} - \lambda^{\mathrm{T}} \left\{ \sum_{i\in B} \omega_{B,i} x_i - \sum_{i\in A} d_{A,i} x_i \right\}. \tag{16}$$

Thus, by minimizing (16), the estimated weights are:

$$\omega_{B,i} = \omega_B(x_i;\widehat{\lambda}) = \frac{N \exp\left(\widehat{\lambda}^{\mathrm{T}} x_i\right)}{\sum_{i\in B} \exp\left(\widehat{\lambda}^{\mathrm{T}} x_i\right)},$$

and $\widehat{\lambda}$ solves the equation:

$$U(\lambda) = \sum_{i\in B} \exp\left(\lambda^{\mathrm{T}} x_i\right) \left\{ x_i - \frac{1}{N} \sum_{i\in A} d_{A,i} x_i \right\} = 0, \tag{17}$$

which is the dual problem to the optimization problem (15).

The calibration weighting estimator is:

$$\widehat{\mu}_{\text{cal}} = \frac{1}{N}\sum_{i=1}^{N} \omega_{B,i} I_{B,i} y_i. \tag{18}$$

Variance estimation of $\widehat{\mu}_{\text{cal}}$ can be obtained by the standard M-estimation theory by treating λ as the nuisance parameter and (17) as the corresponding estimating equation.

The justification for $\widehat{\mu}_{\text{cal}}$ subject to constraint (13) relies on the linearity of the outcome model, i.e., $m(X) = X^{\mathrm{T}}\beta^*$ for some β^*, or the linearity of the inverse probability of sampling weight, i.e., $\{\pi_B(X)\}^{-1} = X^{\mathrm{T}}\alpha^*$ for some α^* (Fuller 2009; Theorem 5.1). The linearity conditions are unlikely to hold for non-continuous variables. In these cases, $\widehat{\mu}_{\text{cal}}$ may be biased. The justification for $\widehat{\mu}_{\text{cal}}$ subject to constraint (14) relies on a correct specification of $m(X;\beta)$ in the data integration problem.

Chan et al. (2016) generalize this idea further to develop a general calibration weighting method that satisfies the covariate balancing property with increasing dimensions of the control variables $m(x)$. Zhao (2019) developed a unified approach of covariate balancing PS method using tailored loss functions. The regularization techniques using penalty terms into the loss function can be naturally incorporated into the framework and machine learning methods, such as boosting, can be used. The covariate balancing condition, or calibration condition, in (13) can be relaxed. Zubizarreta (2015) relaxed the exact balancing constraints to some tolerance level. Wong et al. (2019) used the theory of reproducing Kernel Hilbert space to develop an uniform approximate balance for covariate functions.

3.5 Mass imputation approach

The third type is mass imputation, where the imputed values are created for the whole elements in the probability sample. In the usual imputation for missing data analysis, the respondents in the sample provide a training dataset for developing an imputation model. In the mass imputation, an independent big data sample is used as a training dataset, and imputation is applied to all units in the probability sample. While the mass imputation idea for incorporating information from big data is very natural, the literature on mass imputation itself is very sparse.

In a parametric approach, let $m(X;\beta)$ be the posited model for $m(X)$, where $\beta \in \mathcal{R}^p$ is the unknown parameter. Under Assumption 1, $\widehat{\beta}$ can be obtained by fitting the model to sample B. We assume that $\widehat{\beta}$ is the unique solution to:

$$\widehat{U}(\beta) = \sum_{i\in B} \left\{y_i - m(x_i;\beta)\right\} h(x_i;\beta) = 0$$

for some p-dimensional vector $h(x_i;\beta)$. Thus, we use the observations in sample B to obtain $\widehat{\beta}$ and use it to construct $\widehat{y}_i = m(x_i;\widehat{\beta})$ for all $i \in A$.

Under some regularity conditions, the mass imputation estimator

$$\widehat{\mu}_{\mathrm{I}} = \widehat{\mu}_{\mathrm{I}}(\widehat{\beta}) = N^{-1} \sum_{i \in A} d_{A,i} m(x_i; \widehat{\beta})$$

satisfies: $\widehat{\mu}_{\mathrm{I}} = \widehat{\mu}_{\mathrm{I}}(\beta_0) + o_P(n_B^{-1/2})$, where

$$\widehat{\mu}_{\mathrm{I}}(\beta) = N^{-1} \sum_{i \in A} d_{A,i} m(x_i;\beta) + n_B^{-1} \sum_{i \in B} \{y_i - m(x_i;\beta)\} h(x_i;\beta)^{\mathrm{T}} c^*,$$

$$c^* = \left\{ n_B^{-1} \sum_{i \in B} \dot{m}(x_i;\beta_0) h^{\mathrm{T}}(x_i;\beta_0) \right\}^{-1} \left\{ N^{-1} \sum_{i=1}^{N} \dot{m}(x_i;\beta_0) \right\},$$

where β_0 is the true value of β and $\dot{m}(x;\beta) = \partial m(x;\beta)/\partial\beta$.

Also:

$$E\{\widehat{\mu}_{\mathrm{I}}(\beta_0) - \mu_y\} = 0,$$

and

$$\mathrm{var}\{\widehat{\mu}_{\mathrm{I}}(\beta_0) - \mu_y\} = \mathrm{var}\left\{ N^{-1} \sum_{i \in A} d_{A,i} m(x_i;\beta_0) - N^{-1} \sum_{i \in U} m(x_i;\beta_0) \right\} + E\left[n_B^{-2} \sum_{i \in B} E(e_i^2 \mid x_i) \{h(x_i;\beta_0)^{\mathrm{T}} c^*\}^2 \right],$$

where $e_i = y_i - m(x_i;\beta_0)$. The justification for $\widehat{\mu}_{\mathrm{I}}$ relies on a correct specification of $m(X;\beta)$ and the consistency of $\widehat{\beta}$. If $m(X;\beta)$ is mis-specified or $\widehat{\beta}$ is inconsistent, $\widehat{\mu}_{\mathrm{I}}$ can be biased. For variance estimation, either linearization method or bootstrap method can be used. See Kim et al. (2018) for more details.

3.6 Doubly robust estimation

To improve the robustness against model mis-specification, one can consider combining the weighting and imputation approaches (Kim and Wang 2018). The doubly robust estimator employs both the propensity score and the outcome models, which is given by:

$$\widehat{\mu}_{\mathrm{dr}} = \widehat{\mu}_{\mathrm{dr}}(\widehat{\alpha}, \widehat{\beta}) = N^{-1} \sum_{i=1}^{N} \left[\frac{I_{B,i}}{\pi_B(x_i;\widehat{\alpha})} \{y_i - m(x_i;\widehat{\beta})\} + I_{A,i} d_{A,i} m(x_i;\widehat{\beta}) \right]. \quad (19)$$

The estimator $\widehat{\mu}_{\mathrm{dr}}$ is doubly robust in the sense that it is consistent if either the propensity score model or the outcome model is correctly specified, not necessarily both. Moreover, it is locally efficient if both models are correctly specified (Bang and Robins 2005; Cao et al. 2009). Let $\widehat{\mu}_{\mathrm{HT}} = N^{-1} \sum_{i \in A} d_{A,i} y_i$ be the Horvitz–Thompson estimator that could be used if y_i were observed in sample A. We express $\widehat{\mu}_{\mathrm{dr}} - \widehat{\mu}_{\mathrm{HT}} = -\sum_{i \in A} d_{A,i} \widehat{e}_i + \sum_{i \in B} \{\pi_B(x_i;\widehat{\alpha})\}^{-1} \widehat{e}_i$, where $\widehat{e}_i = y_i - \widehat{y}_i$. To show the double robustness of $\widehat{\mu}_{\mathrm{dr}}$, we consider two scenarios. In the first scenario, if $\pi_B(X;\alpha)$ is correctly specified, then:

$$E\left(\widehat{\mu}_{\mathrm{dr}} - \widehat{\mu}_{\mathrm{HT}} \mid \mathcal{F}_N\right) \cong -\sum_{i\in A} d_{A,i}\widehat{e}_i + \sum_{i\in U}\widehat{e}_i,$$

which is design-unbiased of zero. In the second scenario, if $m(X;\beta)$ is correctly specified, then $E(\widehat{e}_i) \cong 0$. In both cases, $\widehat{\mu}_{\mathrm{dr}} - \widehat{\mu}_{\mathrm{HT}}$ is unbiased of zero and, therefore, $\widehat{\mu}_{\mathrm{dr}}$ is unbiased of μ_y.

If either $\pi_B(X^{\mathrm{T}}\alpha)$ or $m(X^{\mathrm{T}}\beta)$ is correctly specified:

$$n^{1/2}\left\{\widehat{\mu}_{\mathrm{dr}}(\widehat{\alpha},\widehat{\beta}) - \mu\right\} \to \mathcal{N}(0, V),$$

as $n \to \infty$, where $V = \lim_{n\to\infty}(V_1 + V_2)$:

$$\begin{aligned} V_1 =& E\left\{\frac{n}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{A,ij} - \pi_{A,i}\pi_{A,j})\frac{m(x_i^{\mathrm{T}}\beta^*)}{\pi_{A,i}}\frac{m(x_j^{\mathrm{T}}\beta^*)}{\pi_{A,j}}\right\},\\ V_2 =& \frac{n}{N^2}\sum_{i=1}^{N} E\left[\left\{\frac{I_{B,i}}{\pi_{B,i}(x_i^{\mathrm{T}}\alpha^*)} - 1\right\}^2\left\{y_i - m(x_i^{\mathrm{T}}\beta^*)\right\}^2\right]. \end{aligned}$$

To estimate V_1, we can use the design-based variance estimator applied to $m(X_i^{\mathrm{T}}\widehat{\beta})$ as:

$$\widehat{V}_1 = \frac{n}{N^2}\sum_{i\in A}\sum_{j\in A}\frac{(\pi_{A,ij} - \pi_{A,i}\pi_{A,j})}{\pi_{A,ij}}\frac{m(X_i^{\mathrm{T}}\widehat{\beta})}{\pi_{A,i}}\frac{m(X_j^{\mathrm{T}}\widehat{\beta})}{\pi_{A,j}}. \tag{20}$$

To estimate V_2, we further express V_2 as:

$$V_2 = \frac{n}{N^2}\sum_{i=1}^{N} E\left[\left\{\frac{I_{B,i}}{\pi_{B,i}(X_i^{\mathrm{T}}\alpha^*)^2} - \frac{2I_{B,i}}{\pi_{B,i}(X_i^{\mathrm{T}}\alpha^*)}\right\}\left\{Y_i - m(X_i^{\mathrm{T}}\beta^*)\right\}^2 + \left\{Y_i - m(X_i^{\mathrm{T}}\beta^*)\right\}^2\right]. \tag{21}$$

Let $\sigma^2(X_i^{\mathrm{T}}\beta^*) = E\left[\left\{Y_i - m(X_i^{\mathrm{T}}\beta^*)\right\}^2\right]$, and let $\widehat{\sigma}^2(X_i)$ be a consistent estimator of $\sigma^2(X_i^{\mathrm{T}}\beta^*)$. We can then estimate V_2 by:

$$\widehat{V}_2 = \frac{n}{N^2}\sum_{i=1}^{N}\left[\left\{\frac{I_{B,i}}{\pi_B(X_i^{\mathrm{T}}\widehat{\alpha})^2} - \frac{2I_{B,i}}{\pi_B(X_i^{\mathrm{T}}\widehat{\alpha})}\right\}\left\{Y_i - m(X_i^{\mathrm{T}}\widehat{\beta})\right\}^2 + I_{A,i}d_{A,i}\widehat{\sigma}^2(X_i)\right]. \tag{22}$$

By the law of large numbers, $\widehat{V}_2$ is consistent for V_2 regardless of whether one of $\pi_{B,i}(X_i^{\mathrm{T}}\alpha)$ or $\pi_{B,i}(X_i^{\mathrm{T}}\beta)$ is mis-specified, and therefore, it is doubly robust.

4 Combining probability and big data

4.1 Big data sample

To meet the new challenges in the probability sampling, statistical offices face the increasing pressure to utilize convenient but often uncontrolled big data sources, such as satellite information (McRoberts et al. 2010), mobile sensor data (Palmer et al. 2013), and web survey panels (Tourangeau et al. 2013). Couper (2013), Citro (2014), Tam and Clarke (2015), and Pfeffermann et al. (2015) articulated the promise of harnessing big data for official and survey statistics, but also raised many issues regarding big data sources. While such data sources provide timely data for a large number of variables and population elements, they are non-probability samples and often fail to represent the target population of interest because of inherent selection biases. Tam and Kim (2018) also covered some ethical challenges of big data for official statisticians and discuss some preliminary methods of correcting for selection bias in big data.

Combining information from several sources to improve estimates for population parameters is an important practical problem in survey sampling. In the past decade, more and more auxiliary information became available, including large administrative record datasets and remote-sensing data derived from satellite images. How to combine such information with survey data to provide better estimates for population parameters is a new challenge that survey statisticians face today. Tam and Clarke (2015) presented an overview of some initiatives of big data applications in official statistics of the Australian Bureau of Statistics. Such big data are becoming increasingly popular, and they come from a variety of sources such as remote-sensing data, administrative data such as tax data, so on.

Suppose that there are two data sources, one from a probability sample, referred to as sample A, and the other from a big data source, referred to as sample B. Table 4 illustrates the observed data structure.

Table 4 Data structure for data integration with big data

	d	X	Y
Scenario 1			
Sample A	✓	✓	✓
Sample B		✓	
Scenario 2			
Sample A	✓	✓	
Sample B		✓	✓

Sample A is a probability sample, Sample B is a big data sample, which may not be representative of the population, and d is the design weight

4.2 Scenario 1: leverage auxiliary information in big data to improve efficiency

In Scenario 1, the probability sample contains Y observations. Therefore, μ_y is identifiable and can be estimated by the commonly-used estimator solely from sample A, denoted by $\widehat{\mu}_A$. We can leverage the X information in the big data sample to improve the sample A estimator. We consider the case where additionally the membership to the big data can be determined throughout the probability sample. The key insight is that the subsample of units in sample A with the big data membership constitutes a second-phase sample from the big data sample, which acts as a new population. We calibrate the information in the second-phase sample to be the same as the new acting population. The calibration process in turn improves the accuracy of the mass imputation estimator without specifying any model assumptions. Let $h = (I_B, 1 - I_B, I_B X)$.

Following Yang and Ding (2018), we can consider a class of estimators satisfying:

$$n_A^{1/2} \begin{pmatrix} \widehat{\mu}_A - \mu_y \\ \widehat{h}_A - \widehat{h}_B \end{pmatrix} \to \mathcal{N}\left\{ 0, \begin{pmatrix} V_{yy,A} & \Gamma^{\mathrm{T}} \\ \Gamma & V \end{pmatrix} \right\}, \tag{23}$$

in distribution, as $n_A \to \infty$, where $\widehat{h}_A = N^{-1} \sum_{i \in A} d_{A,i} h_i$ and $\widehat{h}_B = N^{-1} \sum_{i \in B} h_i$. Heuristically, if (23) holds exactly rather than asymptotically, by the multivariate normal theory, we have the following conditional distribution:

$$n_A^{1/2}(\widehat{\mu}_A - \mu_y) \Big| n_A^{1/2}(\widehat{h}_A - \widehat{h}_B) \sim \mathcal{N}\left\{ n_A^{1/2} \Gamma^{\mathrm{T}} V^{-1} (\widehat{h}_A - \widehat{h}_B), V_{yy,A} - \Gamma^{\mathrm{T}} V^{-1} \Gamma \right\}.$$

Let $\widehat{V}_{yy,A}$, $\widehat{\Gamma}$ and $\widehat{V}$ be consistent estimators for $V_{yy,A}$, Γ, and V. We set $n_A^{1/2}(\widehat{\mu}_A - \mu_y)$ to equal its estimated conditional mean $n_A^{1/2} \widehat{\Gamma}^{\mathrm{T}} \widehat{V}^{-1} (\widehat{h}_A - \widehat{h}_B)$, leading to an estimating equation for μ_y:

$$n_A^{1/2}(\widehat{\mu}_A - \mu_y) = n_A^{1/2} \widehat{\Gamma}^{\mathrm{T}} \widehat{V}^{-1} (\widehat{h}_A - \widehat{h}_B).$$

Solving this equation for μ_y, we obtain the estimator:

$$\widehat{\mu} = \widehat{\mu}_A - \widehat{\Gamma}^{\mathrm{T}} \widehat{V}^{-1} (\widehat{h}_A - \widehat{h}_B).. \tag{24}$$

Under certain regularity conditions, if (23) holds, then $\widehat{\mu}$ is consistent for μ_y, and:

$$n_A^{1/2}(\widehat{\mu} - \mu_y) \to \mathcal{N}(0, V_{yy,A} - \Gamma^{\mathrm{T}} V^{-1} \Gamma), \tag{25}$$

in distribution, as $n_A \to \infty$. Given a nonzero Γ, the asymptotic variance, $V_{yy,A} - \Gamma^{\mathrm{T}} V^{-1} \Gamma$, is smaller than the asymptotic variance of $\widehat{\mu}_A$, $V_{yy,A}$.

The asymptotic variance of $\widehat{\mu}$ can be estimated by:

$$\widehat{V} = \left(\widehat{V}_{yy,A} - \widehat{\Gamma}^{\mathrm{T}} \widehat{V}^{-1} \widehat{\Gamma} \right) / n_A. \tag{26}$$

Kim and Tam (2018) also explored similar ideas. They develop a calibration weighting method to incorporate the big data auxiliary information and apply the method to the official statistics in Australian Bureau of Statistics. In this application, the big

data is the Australian Agricultural Census with 85% response rate and the probability sample is the Rural Environment and Agricultural Commodities Survey used for calibration. In this application, the measurement from Census data is the auxiliary variable used for calibration.

4.3 Scenario 2: leverage probability sampling designs to correct for selection bias

In Scenario 2, we have a similar setup as in Sect. 3. Depending on the roles in statistical inference, there are two types of *big data*: one with large sample sizes (large n) and the other with rich covariates (large p). We review methods for the two types of big data.

4.3.1 Robust mass imputation estimation

In the first type, the non-probability sample can be large in sample size. How to leverage the rich information in the big data to improve the finite population inference is an important research. We review robust mass imputation methods.

When the sample size of the big data is large, mass imputation is more desirable. In mass imputation, we can train a predictive model from the big data and impute the missing y_i in sample A. Instead of a parametric approach, we can also consider non-parametric approaches. To find suitable imputed values, we consider nearest-neighbor imputation; that is, find the closest matching unit from sample B based on the X values and use the corresponding Y value from this unit as the imputed value.

Using sample B (big data) as a training data, find the nearest neighbor of each unit $i \in A$ using a distance measure $d(x_i, x_j)$. Let $i(1)$ be the index of its nearest neighbor, which satisfies:

$$d(x_{i(1)}, x_i) \leq d(x_j, x_i), \forall j \in B.$$

The nearest-neighbor imputation estimator of μ is:

$$\widehat{\mu}_{\text{nni}} = N^{-1} \sum_{i \in A} d_{A,i} y_{i(1)}.$$

Yang and Kim (2018) showed that under some regularity conditions, $\widehat{\mu}_{\text{nni}}$ has the same asymptotic distribution as $\widehat{\mu}_{\text{HT}} = N^{-1} \sum_{i \in A} d_{A,i} y_i$. Therefore, the variance of $\widehat{\mu}_{\text{nni}}$ is the same as the variance of $\widehat{\mu}_{\text{HT}}$. This implies that the standard point estimator can be applied to the imputed data$\{(x_i, y_{i(1)}) : i \in A\}$ as if the $y_{i(1)}$s were observed values. Let $\pi_{A,ij}$ be the joint inclusion probability for units i and j. They showed that the direct variable estimator based on the imputed data:

$$\widehat{V}_{\text{nni}} = \frac{n_A}{N^2} \sum_{i \in A} \sum_{j \in A} \frac{(\pi_{A,ij} - \pi_{A,i}\pi_{A,j})}{\pi_{A,ij}} \frac{y_{i(1)}}{\pi_{A,i}} \frac{y_{j(1)}}{\pi_{A,j}}$$

is consistent for V_{nni}.

Yang and Kim (2018) also considered two strategies for improving the nearest-neighbor imputation estimator, one using K-nearest-neighbor imputation (Mack and Rosenblatt 1979) and the other using generalized additive models (Wood 2006). In K-nearest-neighbor imputation, instead of using one nearest neighbor, they identify multiple nearest neighbors in the big data sample and use the average response as the imputed value. This method is popular in the international forest inventory community for combining ground-based observations with imagines from remote sensors (McRoberts et al. 2010). In the second strategy, they investigated modern techniques of prediction for mass imputation with flexible models. They used generalized additive models (Wood 2006) to learn the relationship of the outcome and covariates from the big data and create predictions for the probability samples. We note that this strategy can apply to a wider class of semi- and non-parametric estimators such as single index models, Lasso estimators (Belloni et al. 2015), and machine learning methods such as random forests (Breiman 2001).

4.3.2 Variable selection in the presence of a large number of covariates

In the second type, when there are a large number of variables, there is a large literature on variable selection methods for prediction, but little work on variable selection for data integration that can successfully recognize the strengths and the limitations of each data source and utilize all information captured for finite population inference.

In practice, subject matter experts recommend a rich set of potentially useful variables but typically will not identify the set of variables to adjust for. In the presence of a large number of auxiliary variables, variable selection is important, because existing methods may become unstable or even infeasible, and irrelevant auxiliary variables can introduce a large variability in estimation. Gao and Carroll (2017) proposed a pseudo-likelihood approach for combining multiple non-survey data with high dimensionality; this approach requires all likelihoods be correctly specified and therefore is sensitive to model mis-specification. Chen et al. (2018) proposed a model-based calibration approach using LASSO; this approach relies on a correctly specified outcome model.

Yang et al. (2019) proposed a doubly robust variable selection and estimation strategy. In the first step, it selects a set of variables that are important predictors of either the sampling score or the outcome model using penalized estimating equations. In the second step, it re-estimates the nuisance parameter (α, β) based on the joint set of covariates selected from the first step and considers a doubly robust estimator of μ, $\widehat{\mu}_{\mathrm{dr}}(\widehat{\alpha}, \widehat{\beta})$ in (19), where the estimating functions are:

$$J(\alpha,\beta)=\begin{pmatrix} J_1(\alpha,\beta) \\ J_2(\alpha,\beta) \end{pmatrix} = \begin{pmatrix} N^{-1}\sum_{i=1}^{N} I_{B,i}\left\{\frac{1}{\pi_B(x_i^{\mathrm{T}}\alpha)}-1\right\}\{y_i - m(x_i^{\mathrm{T}}\beta)\}x_i \\ N^{-1}\sum_{i=1}^{N}\left\{\frac{I_{B,i}}{\pi_B(x_i^{\mathrm{T}}\alpha)}-d_{A,i}I_{A,i}\right\}\partial m(x_i^{\mathrm{T}}\beta)/\partial\beta \end{pmatrix}. \quad (27)$$

Importantly, the two-step estimator allows model mis-specification of either the sampling score or the outcome model. In the existing high-dimensional causal inference literature, the doubly robust estimators have been shown to be robust to

selection errors using penalization (Farrell 2015) or approximation errors using machine learning (Chernozhukov et al. 2018). However, this double robustness feature requires both nuisance models to be correctly specified. Using (27) relaxes this requirement by allowing one of the nuisance models to be mis-specified. This also enables one to construct a simple and consistent variance estimator (20)+(22) allowing for doubly robust inferences.

5 Concluding remarks

Data integration is an emerging area of research with many potential research topics. We have reviewed statistical techniques and applications for data integration in survey sampling context. Probability sampling remains as the gold standard to obtain a representative sample, but the measurement of the study variable can be obtained from an independent non-probability sample or big data. In this case, assumptions about the sampling model or the outcome model are required. Most data integration methods are based on the unverifiable assumption that the sampling mechanism for the non-probability sample (or big data) is non-informative (corresponding to the missingness at random in the missing data literature).

If the sampling mechanism is informative, imputation techniques can be developed under the strong model assumptions for the sampling mechanism (e.g., Riddles et al. 2016; Morikawa and Kim 2018). Like the non-informative sampling case, the informative sampling assumption is unverifiable. In such settings, sensitivity analysis is recommended to assess the robustness of the study conclusions to unverifiable assumptions. This recommendation echoes Recommendation 15 of the National Research Council (NRC) report entitled "The Prevention and Treatment of Missing Data in Clinical Trials" (National Research Council 2010). Chapter 5 of the NRC Report describes "global" sensitivity analysis procedures that rigorously evaluate the robustness of study findings to untestable assumptions about how missingness might be related to the unobserved outcome.

When the training dataset has a hierarchical structure, multi-level or hierarchical models can be used to develop mass imputation. This is closely related to unit-level small area estimation in survey sampling (Rao and Molina 2015). The small area estimation is particularly promising when we apply data integration using big data. That is, when we use big data as a training sample for prediction, the multi-level model can be used to reflect the possible correlation structure among observations. The parameter estimates for the multi-level model computed from the big data can be used for predicting unobserved study variables in the survey sample if the same multi-level model can be made. Further research in this direction, including the mean-squared error estimation for this small area estimation, will be a topic of future research.

Finally, the uncertainty due to errors in record linkage and statistical matching is also an important problem. The matched sample using record linkage techniques (Fellegi and Sunter 1969) is subject to linkage errors. Zhang and Chambers (2019) cover several research topics in the statistical analysis of combined or fused data.

Acknowledgements The authors are grateful to the two anonymous referees for the constructive comments. Dr. Yang is partially supported by the National Science Foundation Grant DMS-1811245. Dr. Kim is partially supported by the National Science Foundation Grant MMS-1733572 and the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa.

References

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., et al. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, *1*, 90–143.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*, 962–973.

Belloni, A., Chernozhukov, V., Chetverikov, D., & Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, *186*, 345–366.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching? *Social Science Computer Review*, *34*, 59–77.

Breidt, F. J., McVey, A., & Fuller, W. A. (1996). Two-phase estimation by imputation. *Journal of the Indian Society of Agricultural Statistics*, *49*, 79–90.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., et al. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society, Series A*,. https://doi.org/10.1111/rssa.12357.

Bycroft, C. (2010). *Integrated household surveys: A survey vehicles approach*. Wellington: Statistics New Zealand.

Cao, W., Tsiatis, A. A., & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, *96*, 723–734.

Chan, K. C. G., Yam, S. C. P., & Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society, Series B*, *78*, 673–700.

Chen, Y., Li, P. & Wu, C. (2018). Doubly robust inference with non-probability survey samples. arXiv preprint arXiv:1805.06432 .

Chen, J. K. T., Valliant, R., & Elliott, M. R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Survey Methodology*, *44*, 117–144.

Chen, J. K. T., Valliant, R. L., & Elliott, M. R. (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*, 657–681.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., et al. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*, C1–C68.

Chipperfield, J., Chessman, J., & Lim, R. (2012). Combining household surveys using mass imputation to estimate population totals. *The Australian and New Zealand Journal of Statistics*, *54*, 223–238.

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, *40*, 137–161.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *The American Journal of Epidemiology*, *172*, 107–115.
Couper, M. P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods, Number*, *3*, 145–156.
Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*, 427–444.
Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376–382.
D'Orazio, M., Zio, M. D., & Scanu, M. (2006). *Statistical matching: Theory and practice*. Chichester: Wiley.
Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*, 249–264.
Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, *189*, 1–23.
Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*, 1183–1210.
Fuller, W. A. (2009). *Sampling statistics*. Hoboken: Wiley.
Fuller, W. A., & Breidt, F. J. (1999). Estimation for supplemented panels. *Sankhya: Series B*, *61*, 58–70.
Gao, X., & Carroll, R. J. (2017). Data integration with high dimensionality. *Biometrika*, *104*, 251–272.
Hernan, M. A., & VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology*, *22*, 368.
Hidiroglou, M. (2001). Double sampling. *Survey Methodology*, *27*, 143–54.
Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B*, *76*, 243–263.
Kalton, G. (2019). Developments in survey research over the past 60 years: A personal perspective. *International Statistical Review*, *87*, S10–S30.
Kang, J. D., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 523–539.
Keiding, N., & Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society, Series A*, *179*, 319–376.
Kim, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, *98*, 119–132.
Kim, J., Berg, E., & Park, T. (2016). Statistical matching using fractional imputation. *Survey Methodology*, *40*, 19–40.
Kim, J. K., Park, S., Chen, Y., & Wu, C. (2018). Combining non-probability and probability survey samples through mass imputation, arxiv.org/abs/1812.10694 .
Kim, J. K., & Rao, J. N. K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika*, *99*, 85–100.
Kim, J. K., & Tam, S. (2018). Data integration by combining big data and survey sample data for finite population inference. https://arxiv.org/abs/2003.12156
Kim, J. K., & Wang, Z. (2018). Sampling techniques for big data analysis in finite population inference. *International Statistical Review*, *87*, S177–S191.
Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, *32*, 133–142.
Lee, S., & Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods and Research*, *37*, 319–343.
Legg, J. C., & Fuller, W. A. (2009). *Two-phase sampling, handbook of statistics* (Vol. 29, pp. 55–70). New York: Elsevier.
Lohr, S. L., & Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science*, *32*, 293–312.
Mack, Y., & Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, *9*, 1–15.
McConville, K. S., Breidt, F. J., Lee, T. C., & Moisen, G. G. (2017). Model-assisted survey regression estimation with the LASSO. *Journal of Survey Statistics and Methodology*, *5*, 131–158.
McRoberts, R. E., Tomppo, E. O., & Næsset, E. (2010). Advances and emerging issues in national forest inventories. *Scandinavian Journal of Educational Research*, *25*(4), 368–381.
Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *The Journal of the American Statistical Association*, *99*, 1131–9.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society: Series B*, *72*, 27–48.
Morikawa, K., & Kim, J. K. (2018). A note on the equivalence of two semiparametric estimation methods for nonignorable nonresponse. *Statistics & Probability Letters*, *140*, 1–6.
National Research Council (2010). The Prevention and Treatment of Missing Data in Clinical Trials.
Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, *33*, 101–116.
O'Muircheartaigh, C., & Hedges, L. V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society: Series C*, *63*, 195–210.
Palmer, J. R., Espenshade, T. J., Bartumeus, F., Chung, C. Y., Ozgencil, N. E., & Li, K. (2013). New approaches to human mobility: Using mobile phones for demographic research. *Demography*, *50*, 1105–1128.
Park, S., Kim, J. K., & Park, S. (2016). An imputation approach for handling mixed mode surveys. *Annals of Applied Statistics*, *10*, 1063–1085.
Park, S., Kim, J. K., & Stukel, D. (2017). A measurement error model for survey data integration: Combining information from two surveys. *Metron*, *75*, 345–357.
Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach, Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, IEEE, pp. 540–547.
Pfeffermann, D., Eltinge, J. L., Brown, L. D., & Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics: 24th annual morris hansen lecture. *Journal of Survey Statistics and Methodology*, *3*, 425–483.
Rao, J.N.K. (2020). On Making Valid Inferences by Integrating Data from Surveys and Other Sources. Sankhya B. https://doi.org/10.1007/s13571-020-00227-w
Rao, J. N., & Molina, I. (2015). *Small area estimation*. New York: Wiley.
Renssen, R. H., & Nieuwenbroek, N. (1997). Aligning estimates for common variables in two or more sample surveys. *The Journal of the American Statistical Association*, *92*, 368–75.
Riddles, M. K., Kim, J. K., & Im, J. (2016). A propensity-score-adjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology*, *4*, 215–245.
Rivers, D. (2007). *Sampling for web surveys, ASA proceedings of the section on survey research methods*. Alexandria: American Statistical Association.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.
Rothwell, P. M. (2005). External validity of randomised controlled trials: "to whom do the results of this trial apply?". *The Lancet*, *365*, 82–93.
Rudolph, K. E., & van der Laan, M. J. (2017). Robust estimation of encouragement design intervention effects transported across sites. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*, 1509–1525.
Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. New York: Springer-Verlag.
Schenker, N., & Raghunathan, T. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine*, *26*, 1802.
Stuart, E. A., Bradshaw, C. P., & Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, *16*, 475–485.
Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A*, *174*, 369–386.
Sugiyama, M., & Kawanabe, M. (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. Cambridge: MIT Press.
Tam, S.-M., & Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *International Statistical Review*, *83*, 436–448.
Tam, S.-M., & Kim, J.-K. (2018). Big data ethics and selection-bias: An official statistician's perspective. *Statistical Journal of the IAOS*, *34*(4), 577–588.
Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, *38*, 239–266.
Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. New York: Oxford University Press.

Valliant, R., & Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods and Research*, *40*, 105–137.

Vavreck, L., & Rivers, D. (2008). The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties*, *18*, 355–366.

Verde, P. E., & Ohmann, C. (2015). Combining randomized and non-randomized evidence in clinical research: A review of methods and applications. *Research Synthesis Methods*, *6*, 45–62.

Wang, W., Scharfstein, D., Tan, Z., & MacKenzie, E. J. (2009). Causal inference in outcome-dependent two-phase sampling designs. *Journal of the Royal Statistical Society: Series B*, *71*, 947–969.

Wong, M., Gao, J., Thanarajasingam, G., Sloan, J., Dueck, A., Novotny, P., Jatoi, A., Hurria, A., Wood, W., Feliciano, J., Walter, L., Miaskowski, C., Stinchcombe, T. & Wang, X. (2019). Comparison of chemotherapy toxicity over time according to age and performance status in patients with adadvanced non-small cell lung cancer: A toxicity over time analysis of CALGB 9730, Technical report, Duke University B&B.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Boca Raton: Chapman and Hall/CRC.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *The Canadian Journal of Statistics*, *32*, 15–26.

Yang, S., & Ding, P. (2018). Combining multiple observational data sources to estimate causal effects, arXiv preprint arXiv:1801.00802.

Yang, S., & Kim, J. K. (2018). Integration of survey data and big observational data for finite population inference using mass imputation, arXiv preprint arXiv:1807.02817.

Yang, S., Kim, J. K., & Song, R. (2019). Doubly robust inference when combining probability and nonprobability samples with high-dimensional data. *Journal of the Royal Statistical Society, Series B*, *82*, 445–465.

Ybarra, L., & Lohr, S. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, *95*, 919–31.

Zhang, L.-C., & Chambers, R. L. (2019). *Analysis of integrated data*. Boca Raton: CRC Press.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Annals of Statistics*, *47*, 965–993.

Zieschang, K. D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*, *85*, 986–1001.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, *110*, 910–922.