The Survey Statistician, 2022, Vol. 85, 19–29.

New and Emerging Methods

A gentle introduction to data integration in survey sampling

Jae Kwang Kim

Department of Statistics, Iowa State University, jkim@iastate.edu

Abstract

This article provides a systematic review of data integration techniques for combining a probability sample with a non-probability sample when the study variable is observed in the non-probability sample only. We discuss a wide range of integration methods such as mass imputation, propensity score method, calibration weighting, and doubly robust estimation methods. Finally, we highlight important questions for future research.

Keywords: big data, calibration weighting, doubly robust estimation, mass imputation, propensity score.

1 Introduction

Probability sampling is regarded as the gold-standard in survey statistics for finite population inference. Because probability samples are selected under known sampling designs, they 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); Fuller (2009) for textbook discussions. Kalton (2019) provided a comprehensive overview of the survey sampling research in the last 60 years.

On the other hand, 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 typically 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. (2020), among others. Combining the up-to-date information from a probability sample can be viewed as data integration. Rao (2021) and Yang and Kim (2020) provide comprehensive reviews for data integration for finite population inference.

One can view data integration as a missing data problem, and apply the statistical techniques for handling missing data. Specifically, we consider the following setup for data integration. Let A be a probability sample with observations on auxiliary variable X; let B be the non-probability sample with information on both the study variable Y and the auxiliary variables X. Table 1 presents the general setup of the two sample structure for data integration. As indicated in Table 1, sample B is not representative of the target population.

Table 1: Data Structure for Two Samples

Sample	Type	X	Y	Representative?
A	Probability Sample	✓		Yes
B	Non-probability Sample	✓	✓	No

Under the data structure in Table 1, we wish to develop methods for combining information from two samples. To develop statistical methods for data integration, we may require some assumptions on the outcome model or on the sampling mechanism for sample B.

2 Setup and assumptions

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 datasets. The sampling mechanism for sample B is often unknown, and we cannot compute the first-order inclusion probability for Horvitz-Thompson estimation. The naive estimators constructed without adjusting for the sampling process are subject to selection biases. On the other hand, although the probability sample with design weights represents the finite population, it does not contain the study variable. We wish to develop data integration methods that leverage the advantages of both sources.

Let $f(Y \mid X)$ be the conditional distribution of Y given X in the superpopulation model ζ that generates the finite population. Let $\delta_i = 1$ if $i \in B$ and $\delta_i = 0$ otherwise. We make the following assumption.

Assumption 1 *(i) The sampling indicator δ of sample B and the study variable Y are conditionally independent given X; i.e. $P(\delta = 1 \mid X, Y) = P(\delta = 1 \mid X)$; and (ii) $\pi_B(X) \equiv P(\delta = 1 \mid X) > 0$ for all X.*

Assumption 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. Assumption 1 (i) states the ignorability of the selection mechanism to sample B conditional upon the covariates. Under Assumption 1 (i), $E(Y \mid X) = E(Y \mid X, \delta = 1)$ can be estimated based on sample B. Assumption 1 (ii) implies that the support of X 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 Assumption 1 (ii) can be 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: i) the weighting approach of constructing weights for sample B to improve the representativeness of sample B; ii) the imputation approach of 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, 2015; Keiding and Louis; 2016), external validity (Rothwell; 2005) or transportability (Pearl and Bareinboim; 2011; Rudolph and van der Laan; 2017) in the statistics literature.

3 Mass imputation

In mass imputation, we view the probability sample as having 100% missing values for the study variable. We can then use the non-probability sample as training data to develop an imputation model and construct a synthetic dataset for the probability sample. Mass imputation was originally developed in the context of two-phase sampling (Breidt et al.; 1996; Kim and Rao; 2012) to create synthetic data for the probability sample. Rivers (2007), Kim et al. (2021), and Chen et al. (2021) develop mass imputation for a probability sample using observations from a non-probability sample. Even though the observations in the non-probability sample are not necessarily representative of the target population, the relationships among variables in the non-probability sample can be used to develop a predictive model for mass imputation. Thus, the non-probability sample can be used as training data for developing a model for mass imputation.

We use $\mathbf{x}$ and y to denote the realized value of X and Y in the sample, respectively. In a parametric approach, let $m(\mathbf{x};\boldsymbol{\beta})$ be the posited model for $m(\mathbf{x}) = E(Y \mid \mathbf{x})$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown parameter. Under Assumption 1, a consistent estimator of $\boldsymbol{\beta}$ can be obtained by fitting the model to sample B. Thus, we can estimate $\boldsymbol{\beta}$ by finding the minimizer of

$$Q(\boldsymbol{\beta}) = \sum_{i \in B} \left\{ y_i - m(\mathbf{x}_i;\boldsymbol{\beta}) \right\}^2 / v(\mathbf{x}_i;\boldsymbol{\beta}) = 0$$

for some $v(\mathbf{x};\boldsymbol{\beta}) = V(Y \mid \mathbf{x};\boldsymbol{\beta})$. Thus, we use the observations in sample B to obtain $\widehat{\boldsymbol{\beta}}$ and construct $\widehat{y}_i = m(\mathbf{x}_i;\widehat{\boldsymbol{\beta}})$ for all $i \in A$.

Using $\widehat{y}_i = m(\mathbf{x}_i;\widehat{\boldsymbol{\beta}})$ for all $i \in A$, we can construct

$$\widehat{\mu}_{\mathrm{I}} = N^{-1} \sum_{i \in A} d_{A,i} \widehat{y}_i$$

as the mass imputation estimator of $\mu = N^{-1} \sum_{i=1}^{N} y_i$, where $d_{A,i}$ is the design weight of unit i for sample A. The justification for $\widehat{\mu}_{\mathrm{I}}$ relies on correct specification of $m(\mathbf{x};\boldsymbol{\beta})$ and the consistency of $\widehat{\boldsymbol{\beta}}$. For variance estimation, either linearization method or bootstrap method can be used. See Kim et al. (2021) for more details.

Instead of using parametric mass imputation with a parametric model, we can develop non-

parametric mass imputation using nonparametric models. Rivers (2007) first proposed using nearest neighbor imputation for mass imputation and its asymptotic theory is rigorously discussed by Yang et al. (2021).

4 Propensity Score Method

Under Assumption 1, we can further build a model for $P(\delta = 1 \mid \mathbf{x})$ and use it to construct the propensity score weights for sample B. Suppose that $\pi(\mathbf{x}) = P(\delta = 1 \mid \mathbf{x})$ has a parametric form such that $\pi(\mathbf{x}) = \pi(\mathbf{x}; \phi)$ for some ϕ. The population log-likelihood function for ϕ can be written as

$$l(\phi) = \sum_{i=1}^{N} \left[\delta_i \log \pi(\mathbf{x}_i; \phi) + (1 - \delta_i) \log\{1 - \pi(\mathbf{x}_i; \phi)\}\right].$$

Thus, the (population-based) maximum likelihood estimator of ϕ can be obtained by solving

$$S_p(\phi) \equiv \sum_{i=1}^{N} \left\{\frac{\delta_i}{\pi(\mathbf{x}_i; \phi)} - \frac{1 - \delta_i}{1 - \pi(\mathbf{x}_i; \phi)}\right\} \dot{\pi}(\mathbf{x}_i; \phi) = 0,$$

which is equivalent to solving

$$\sum_{i=1}^{N} \delta_i h(\mathbf{x}_i; \phi) = \sum_{i=1}^{N} \pi(\mathbf{x}_i; \phi) h(\mathbf{x}_i; \phi) \tag{1}$$

for ϕ, where

$$h(\mathbf{x}_i; \phi) = \frac{\dot{\pi}(\mathbf{x}_i; \phi)}{\pi(\mathbf{x}_i; \phi)\{1 - \pi(\mathbf{x}_i; \phi)\}}$$

and $\dot{\pi}(\mathbf{x}; \phi) = \partial \pi(\mathbf{x}; \phi)/\partial \phi$. The left side of (1) can be constructed from sample B. Thus, we have only to estimate the right side of (1). Using the sampling weights, we can use

$$\sum_{i=1}^{N} \delta_i h(\mathbf{x}_i; \phi) = \sum_{i \in A} d_{A,i} \pi(\mathbf{x}_i; \phi) h(\mathbf{x}_i; \phi), \tag{2}$$

which does not require identification of the elements in both samples. Chen et al. (2020) first proposed estimation using (2) for propensity score method for voluntary samples. The final propensity score (PS) estimator for μ is

$$\hat{\mu}_{PS} = \frac{\sum_{i \in B} \hat{\pi}_i^{-1} y_i}{\sum_{i \in B} \hat{\pi}_i^{-1}}, \tag{3}$$

where $\hat{\pi}_i = \pi(\mathbf{x}_i; \hat{\phi})$. If $n_B = |B|$ is small compared with N, then the estimated probability $\hat{\pi}(\mathbf{x}_i)$ can take small values, and the resulting PS estimator in (3) can be unstable.

Elliott and Valliant (2017) proposed a different approach of propensity score method for data integration. Note that

$$P(\delta = 1 \mid \mathbf{x}) \propto P(I_A = 1 \mid \mathbf{x}) \cdot \frac{f(\mathbf{x} \mid \delta = 1)}{f(\mathbf{x} \mid I_A = 1)},$$

where I_A is the sample inclusion indicator function for sample A. Thus,

$$\frac{1}{P(\delta = 1 \mid \mathbf{x})} \propto \{P(I_A = 1 \mid \mathbf{x})\}^{-1} \cdot \frac{f(\mathbf{x} \mid I_A = 1)}{f(\mathbf{x} \mid \delta = 1)} := \tilde{w}(\mathbf{x}) \cdot R(\mathbf{x}).$$

Elliott and Valliant (2017) proposed estimating two terms separately. To estimate the first term $\tilde{w}(\mathbf{x}_i)$, using

$$E(w_i \mid \mathbf{x}_i, I_{A,i} = 1) = \frac{1}{P(I_{A,i} = 1 \mid \mathbf{x}_i)},$$

one can apply regression of w_i on $\mathbf{x}_i$ from sample A. To estimate the second term, Elliott and Valliant (2017) proposed using

$$R(\mathbf{x}) \equiv \frac{f(\mathbf{x} \mid I_A = 1)}{f(\mathbf{x} \mid \delta = 1)} \propto \frac{P(I_A = 1 \mid \mathbf{x}, I_A + \delta \geq 1)}{P(\delta = 1 \mid \mathbf{x}, I_A + \delta \geq 1)}.$$

One can apply a suitable classification method from the combined sample to estimate $R(x)$. The final pseudo weight for sample B is then

$$\hat{w}_i = \tilde{w}_i \hat{R}(\mathbf{x}_i).$$

Rafei et al. (2020) uses Bayesian Additive Regression Trees (BART) to estimate the two components in the pseudo weights for voluntary big data sample.

5 Calibration weighting

The second weighting strategy is calibration weighting, or benchmarking weighting (Deville and Särndal; 1992; Kott; 2006; Breidt and Opsomer; 2017). 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 (Lee and Valliant; 2009).

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

$$\sum_{i \in B} \omega_{B,i} \mathbf{x}_i = \sum_{i \in A} d_{A,i} \mathbf{x}_i, \tag{4}$$

where $\sum_{i \in A} d_{A,i} \mathbf{x}_i$ is a design-weighted estimate of the population total of X from the probability sample. Constraint (4) is referred to as the covariate balancing constraint (Imai and Ratkovic; 2014), and weights $\mathcal{Q}_B = \{\omega_{B,i} : i \in B\}$ satisfying (4) 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 use model calibration (Wu and Sitter; 2001). In this approach, one can posit a parametric model for $E(Y \mid \mathbf{x}) = m(\mathbf{x}; \boldsymbol{\beta})$ and estimate the unknown parameter $\boldsymbol{\beta}$ from sample B. The model-based calibration specifies the constraints for $\mathcal{Q}_B$ as

$$\sum_{i \in B} \omega_{B,i} m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}) = \sum_{i \in A} d_{A,i} m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}). \tag{5}$$

Suppose that the finite population follows the following superpopulation model:

$$y_i = m(\mathbf{x}_i) + e_i \tag{6}$$

with $E(e_i \mid \mathbf{x}_i) = 0$ and $V(e_i \mid \mathbf{x}_i) = \sigma^2$. If we can express $m(\mathbf{x}) = \sum_{k=1}^{L} \beta_k b_k(\mathbf{x})$ for some $\beta_k, k = 1, 2, \cdots, L$, that is $m(\mathbf{x}) \in \text{span}\{b_1(\mathbf{x}), \cdots, b_L(\mathbf{x})\}$, then we may use

$$\sum_{i \in B} \omega_{B,i} \left[b_1(\mathbf{x}_i), \cdots, b_L(\mathbf{x}_i)\right] = \sum_{i \in A} d_{A,i} \left[b_1(\mathbf{x}_i), \cdots, b_L(\mathbf{x}_i)\right] \tag{7}$$

in the calibration estimation. As long as $m(\mathbf{x}) \in \text{span}\{b_1(\mathbf{x}), \cdots, b_L(\mathbf{x})\}$ holds, the calibration weights in (7) satisfy (5) without estimating $\boldsymbol{\beta}$. The dimension L may increase with the sample size. In this case, some regularization method can be used to choose L. For example, Montanari and Ranalli (2005) used neural network models and Breidt et al. (2005) used penalized Spline models for nonparametric calibration estimation.

Writing $\hat{\mu}_w = N^{-1} \sum_{i \in B} \omega_{B,i} y_i$, we can express

$$\begin{aligned} \hat{\mu}_w - \mu &= N^{-1} \left\{ \sum_{i \in B} \omega_{B,i} m(\mathbf{x}_i) - \sum_{i=1}^{N} m(\mathbf{x}_i) \right\} + N^{-1} \left\{ \sum_{i \in B} \omega_{B,i} e_i - \sum_{i=1}^{N} e_i \right\} \\ &:= C + D. \end{aligned}$$

Since $E(D) = 0$ by model (6), we may require $E(C) = 0$ to get unbiased estimation. A sufficient condition for $E(C) = 0$ under model (6) is the model calibration condition in (5) or (7). To find the optimal calibration estimator that minimizes variance of $\hat{\mu}_w$ in the class of unbiased estimators under model (6), we have only to minimize $E(D^2)$ subject to the calibration constraints. Note that

$$\begin{aligned} E(D^2) &= \text{var} \left\{ N^{-1} \sum_{i=1}^{N} (\delta_i \omega_{B,i} - 1) e_i \right\} \\ &= N^{-2} \sum_{i=1}^{N} (\delta_i \omega_{B,i} - 1)^2 \sigma^2 = \sigma^2 N^{-2} \sum_{i \in B} (\omega_{B,i} - 1)^2 + \text{constant}. \end{aligned}$$

Thus, we can formulate the calibration weighting problem as finding the minimizer of $Q_0(\boldsymbol{\omega}_B) = \sum_{i \in B} (\omega_{B,i} - 1)^2$ subject to (4) or (7) with $\boldsymbol{\omega}_B = \{\omega_{B,i}; i \in B\}$. However, using $Q_0(\boldsymbol{\omega}_B)$ as the objective function for the calibration problem can lead to negative calibration weights.

To avoid negative calibration weights, following Hainmueller (2012), we may consider the entropy divergence

$$Q(\boldsymbol{\omega}_B) = \sum_{i \in B} \omega_{B,i} \log(\omega_{B,i}) \tag{8}$$

as the objective function for optimization. Thus, we find the minimizer of $Q(\boldsymbol{\omega}_B)$ subject to $\omega_{B,i} \geq 0,$ for all $i \in B$; $\sum_{i \in B} \omega_{B,i} = N$, and the balancing constraint (4) or (7). This optimization problem can be solved using convex optimization with a Lagrange multiplier. Other objective functions can also be considered. By introducing Lagrange multiplier $\boldsymbol{\lambda}$, the objective function becomes

$$L(\boldsymbol{\omega}_B, \boldsymbol{\lambda}) = \sum_{i \in B} \omega_{B,i} \log \omega_{B,i} - \boldsymbol{\lambda}' \left\{ \sum_{i \in B} \omega_{B,i} \mathbf{x}_i - \sum_{i \in A} d_{A,i} \mathbf{x}_i \right\}. \tag{9}$$

Thus, by minimizing (9), the estimated weights are

$$\omega_{B,i} = \omega_B(\mathbf{x}_i; \widehat{\boldsymbol{\lambda}}) = N \frac{\exp\left(\widehat{\boldsymbol{\lambda}}' \mathbf{x}_i\right)}{\sum_{i \in B} \exp\left(\widehat{\boldsymbol{\lambda}}' \mathbf{x}_i\right)},$$

where $\widehat{\boldsymbol{\lambda}}$ solves

$$U(\boldsymbol{\lambda}) \equiv \sum_{i \in B} \exp\left(\boldsymbol{\lambda}' \mathbf{x}_i\right) \left\{ \mathbf{x}_i - N^{-1} \sum_{i \in A} d_{A,i} \mathbf{x}_i \right\} = 0. \tag{10}$$

Finally, the calibration weighting estimator is

$$\widehat{\mu}_{\text{cal}} = \frac{1}{N} \sum_{i \in B} \omega_{B,i} y_i. \tag{11}$$

Variance estimation of $\widehat{\mu}_{\text{cal}}$ can be obtained by the standard M-estimation theory by treating $\boldsymbol{\lambda}$ as the nuisance parameter and (10) as the corresponding estimating equation.

Chan et al. (2016) generalize the calibration idea further to develop a general calibration weighting method that satisfies the covariate balancing property with increasing dimensions of the control variables for $m(\mathbf{x})$. Zhao (2019) developed a unified approach of covariate balancing method using Tailored loss functions. The regularization techniques using penalty terms in the loss function can be incorporated into the framework. The covariate balancing condition, or calibration condition, in (4), can be relaxed using soft calibration (Rao and Singh; 1997; Guggemos and Tille; 2010). Wong and Chan (2018) used the theory of reproducing Kernel Hilbert space to develop a uniform approximate balance for covariate functions.

6 Doubly robust estimation

To improve the robustness against model misspecification, one can consider combining the weighting and imputation approaches (Kim and Haziza; 2014). The doubly robust (DR) estimator employs both the propensity score and the outcome models, which is given by

$$\widehat{\mu}_{\text{dr}} = \widehat{\mu}_{\text{dr}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\delta_i}{\pi_B(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}})} \{ y_i - m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}) \} + I_{A,i} d_{A,i} m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}) \right]. \tag{12}$$

The estimator $\widehat{\mu}_{\text{dr}}$ 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 $\widehat{\mu}_{\text{HT}} = N^{-1} \sum_{i \in A} d_{A,i} y_i$ be the Horvitz–Thompson estimator that could be used if y_i were observed in sample A. Note that

$$\widehat{\mu}_{\text{dr}} - \widehat{\mu}_{\text{HT}} = -\frac{1}{N} \sum_{i=1}^{N} \left\{ I_{A,i} d_{A,i} - \delta_i \{ \pi_B(\mathbf{x}_i; \widehat{\boldsymbol{\alpha}}) \}^{-1} \right\} \widehat{e}_i,$$

where $\widehat{e}_i = y_i - m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}})$. To show the double robustness of $\widehat{\mu}_{\text{dr}}$, we consider two scenarios. In

the first scenario, if $\pi_B(\mathbf{x};\boldsymbol{\alpha})$ 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 for zero. In the second scenario, if $m(\mathbf{x};\boldsymbol{\beta})$ is correctly specified, then $E(\widehat{e}_i) \cong 0$. In both cases, $\widehat{\mu}_{\mathrm{dr}} - \widehat{\mu}_{\mathrm{HT}}$ is unbiased for zero and therefore $\widehat{\mu}_{\mathrm{dr}}$ is unbiased for μ_y. Asymptotic expansion of the DR estimator is simplified if the model parameters satisfy the orthogonality condition of Randles (1982). That is, if

$$\frac{\partial}{\partial\boldsymbol{\alpha}}\hat{\mu}_{\mathrm{dr}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathbf{0} \;\text{ and }\; \frac{\partial}{\partial\boldsymbol{\beta}}\hat{\mu}_{\mathrm{dr}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathbf{0} \tag{13}$$

at $(\boldsymbol{\alpha},\boldsymbol{\beta}) = (\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$, then we can safely ignore the uncertainty of estimating $(\boldsymbol{\alpha},\boldsymbol{\beta})$ in the final DR estimation. We can impose (13) in constructing the estimating equation for model parameters.

Yang et al. (2019) extended DR estimation to the high dimensional covariate problem. If both the outcome model and the propensity score model are nonparametric, then the estimator of the form (12) is no longer doubly robust. In this case, estimation using sample splitting can be considered. See Chernozhukov et al. (2017) for details.

7 Discussion

Data integration is a new field of study with a wide range of prospective research subjects. We have considered the situation of merging data from two samples, one from probability sampling and the other from non-probability sampling, where the probability sample lacks the study variable of interest. As a result, information bias affects the probability sample, whereas selection bias affects the non-probability sample. We can adjust for selection bias in the non-probability sample or adjust for information bias in the probability sample using statistical procedures for handling missing data. The majority of data integration methods are based on the unverifiable assumption that the sampling mechanism for the non-probability sample is non-informative. Suppose the non-probability sample is big data. In that case, we can develop the dual frame estimator approach as in Kim and Tam (2021), and the non-informativeness assumption of the sampling mechanism is unnecessary.

Even when the non-informativeness assumption (Assumption 1) is true, the proposed data integration methods employ explicit assumptions for the outcome regression model or sample selection model. Modest model misspecification does not necessarily lead to biased point estimation, but may increase the variance. In this case, the proposed variance estimators based on the assumed model may underestimate the true variance of the data integration estimators. Achieving robustness and assessing uncertainty under modest model misspecification is an important future research topic.

If the sampling mechanism is informative, imputation techniques can be developed under the strong model assumptions for the sampling mechanism (Morikawa and Kim; 2020). As in the non-informative sampling case, the informative sampling assumptions are unverifiable. Thus, sensitivity analysis is recommended to evaluate the robustness of the study conclusions to unverifiable assumptions. Or, if budget is allowed, a follow-up subsampling can be used to build a realistic model for the informative sampling mechanism. Developing tools for data integration under informative sampling is another important research topic.

Acknowledgements

The article is a concisely updated version of the review paper of Yang and Kim (2020). The author is grateful to professors Wayne Fuller and Maria Giovanna Ranalli for their very constructive comments. The research was partially supported by 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., Gile, K. J. and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling, *Journal of Survey Statistics and Methodology* **1**: 90–143.

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

Breidt, F. J., Claeskens, G. and Opsomer, J. D. (2005). Model-assisted estimation for complex surveys using penalised splines, *Biometrika* **92**: 831–846.

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

Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques, *Statistical Science* **32**: 190–205.

Cao, W., Tsiatis, A. A. and 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. and 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, S., Yang, S. and Kim, J. K. (2021). Nonparametric mass imputation for data integration, *Journal of Survey Statistics and Methodology* . Accepted for publication.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples, *Journal of the American Statistical Association* **115**: 2011–2021.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects, *The American Economic Review* **107**: 261–265.

Cochran, W. G. (1977). *Sampling Techniques*, 3 edn, New York: John Wiley & Sons, Inc.

Cole, S. R. and Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial, *American Journal of Epidemiology* **172**: 107–115.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association* **87**: 376–382.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples, *Statistical Science* **32**: 249–264.

Fuller, W. A. (2009). *Sampling Statistics*, Wiley, Hoboken, NJ.

Guggemos, F. and Tille, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models, *Journal of statistical planning and inference* **140**: 3199–3212.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies, *Political Analysis* **20**: 25–46.

Imai, K. and 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.

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

Kim, J. K., Park, S., Chen, Y. and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation, *Journal of the Royal Statistical Society, Series A* **184**: 941–963.

Kim, J. K. and Rao, J. N. K. (2012). Combining data from two independent surveys: a model-assisted approach, *Biometrika* **99**: 85–100.

Kim, J. K. and Tam, S. (2021). Data integration by combining big data and survey sample data for finite population inference, *International Statistical Review* **89**: 382–401.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors, *Survey Methodology* **32**: 133–142.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment, *Sociological Methods and Research* **37**: 319–343.

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling, *Journal of the American Statistical Association* **100**: 1429–1442.

Morikawa, K. and Kim, J. K. (2020). Semiparametric optimal estimation with nonignorable nonresponse data, *Annals of Statistics* **49**: 2991–3014.

Pearl, J. and 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.

Rafei, A., Flannagan, C. A. C. and Elliott, M. R. (2020). Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees, *Journal of Survey Statistics and Methodology* **8**: 148–180.

Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters, *Annals of Statistics* **10**: 462–74.

Rao, J. N. K. (2021). On making valid inferences by integrating data from surveys and other sources, *Sankhya B* **83**: 242–272.

Rao, J. N. K. and Singh, A. C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling, *ASA Proceedings of the Section on Survey Research Methods*, pp. 57–85.

Rivers, D. (2007). Sampling for web surveys, *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association (Alexandria, VA).

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**(1): 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. and 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. and Wretman, J. (2003). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Stuart, E. A., Bradshaw, C. P. and 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. and 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.

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

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

Wong, R. K. W. and Chan, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies, *Biometrika* **105**: 199–213.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data, *Journal of the American Statistical Association* **96**: 185–193.

Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: A review, *Japanese Journal of Statistics and Data Science* **3**: 625–650.

Yang, S., Kim, J. K. and Hwang, Y. (2021). Integration of survey data and big observational data for finite population inference using mass imputation, *Survey Methodology* **47**: 29–58.

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

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