Received: 8 January 2020 | Accepted: 20 March 2021

DOI: 10.1111/rssa.12696

ORIGINAL ARTICLE

Combining non-probability and probability survey samples through mass imputation

Jae Kwang Kim[1] | **Seho Park**[2] | **Yilin Chen**[3] | **Changbao Wu**[3]

[1]Department of Statistics, Iowa State University, Ames, IA 50011, USA

[2]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, USA

[3]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

Correspondence
Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA 50011, USA. Email: jkim@iastate.edu

Funding information
US National Science Foundation, Grant/Award Number: 1733572

Abstract

Analysis of non-probability survey samples requires auxiliary information at the population level. Such information may also be obtained from an existing probability survey sample from the same finite population. Mass imputation has been used in practice for combining non-probability and probability survey samples and making inferences on the parameters of interest using the information collected only in the non-probability sample for the study variables. Under the assumption that the conditional mean function from the non-probability sample can be transported to the probability sample, we establish the consistency of the mass imputation estimator and derive its asymptotic variance formula. Variance estimators are developed using either linearization or bootstrap. Finite sample performances of the mass imputation estimator are investigated through simulation studies. We also address important practical issues of the method through the analysis of a real-world non-probability survey sample collected by the Pew Research Centre.

KEYWORDS

auxiliary variables, bootstrap variance estimator, data integration, ignorable sample selection, model transportability, selection bias

1 | INTRODUCTION

Probability sampling is a classical tool for obtaining a representative sample from a target population. Because the first-order inclusion probabilities are known under the given survey design, probability sampling can provide design unbiased estimators and valid statistical inferences for finite population parameters. Although probability samples have been a tremendously successful story over the past 60 years, the approach has encountered many challenges in recent years, including high cost, low response rate and inability to provide up-to-date information on variables of specific studies. On the other hand, non-probability samples, such as those collected from web panels, have become increasingly popular due to the efficient recruitment process, quick responses and low maintenance expenses. See, for instance, Tourangeau et al. (2013), for examples of web-based non-probability survey samples.

Statistical analysis of non-probability survey samples, however, faces many challenges as documented by Baker et al. (2013). The first major challenge is that the selection/inclusion mechanism for non-probability samples is typically unknown. Treating non-probability samples as if they are a simple random sample often leads to biased results. The second major challenge is that the estimation of the propensity scores (i.e. the probabilities of participating in the non-probability sample) under an assumed model requires auxiliary information at the population level. A popular framework used in recent years is to assume that the required auxiliary information on the target 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), Brick (2015), Elliott and Valliant (2017) and Chen et al. (2020), among others.

Generally speaking, there are three possible approaches to analysing non-probability survey samples: (i) inverse probability weighting (IPW) using estimated propensity scores; (ii) model-based prediction using an outcome regression model; and (iii) doubly robust procedures involving both the propensity scores and the outcome regression model. A central point to all approaches is the assumption that the selection/inclusion mechanism for non-probability survey samples is ignorable. It is essentially the missing at random (MAR) assumption of Rubin (1976). It is well known in the literature on missing data analysis that the MAR assumption cannot be formally tested using data from the sample itself. Practical applications of the propensity score-based methods require a careful examination of the auxiliary variables included in the sample to see whether participation in the non-probability sample could be characterized by those variables. Calibration weighting methods for non-probability samples are also discussed in the literature. See, for instance, Dever and Valliant (2016) and Elliott and Valliant (2017). The proposed methodologies, however, are closely related to the propensity score-based methods and require the same ignorability assumption on the selection mechanism (Chen et al., 2020).

Mass imputation, on the other hand, is a model-based prediction method that includes both the non-probability and the probability samples. The existing probability sample is assumed to have no measurement on the study variable of interest and can be viewed as having 100% missing values for the study variable. The non-probability sample contains values observed on both the study variable and the auxiliary variables. The non-probability sample is then used as training data to develop a prediction model that will be used to provide imputations for the probability sample. The key assumption for the validity of the mass imputation method is that the prediction model built based on the non-probability sample can be transported to the probability sample for imputation. See Section 2 for detailed discussions. Mass imputation has also been developed in the context of two-phase sampling (Breidt et al., 1996; Kim & Rao, 2012; Park & Kim, 2019), but it is not fully investigated in the context of survey integration for combining the non-probability sample with a probability sample. One notable exception is the sample matching method of Rivers (2007), but he did not provide any theoretical justifications for the proposed method. Furthermore, the sample matching method of Rivers (2007) is based on the nearest neighbour imputation method, which suffers from the curse of dimensionality when the number of auxiliary variables is large. 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.

In this paper, we aim to fill the important research gap in survey sampling on mass imputation for analysing non-probability survey samples. From a theoretical point of view, if the training data for building the imputation model were a probability sample, then the theory of Kim and Rao (2012) could be directly applicable. Under the assumptions that the support of auxiliary variables from the non-probability sample is the same as the sample space and the conditional mean function from the non-probability sample can be transported to the probability sample, we show that the method of Kim and Rao (2012) can be applied even when non-probability samples are used as the training data. We develop rigorous asymptotic theory for the mass imputation estimator along with a linearization variance estimator and a proposed bootstrap method for variance estimation.

Our investigation on practical aspects of the method was motivated by the analysis of a real non-probability survey sample collected in 2015 by the Pew Research Centre (PRC) from the United States of America. It turned out that there existed at least two probability survey samples taken in the same year from the same finite population with some common auxiliary variables included in both samples. In addition to demonstrations of the mass imputation method using the PRC non-probability sample, we also address major practical questions through the example, which includes (i) the assessment of auxiliary variables with regard to the ignorability assumption and the imputation model; (ii) the impact of relative sample sizes between the non-probability and the probability samples; and (iii) factors to be considered in choosing a probability sample when multiple samples are available.

The paper is organized as follows. The basic setting is described in Section 2 and the mass imputation is presented in Section 3. Main theoretical results on consistency and the asymptotic variance formula are presented in Section 4. A practically useful bootstrap variance estimator is proposed in Section 5. Results from a limited simulation study on the finite sample performances of the mass imputation estimator are reported in Section 6. Important practical aspects of the method are discussed in Section 7 through an application to the PRC non-probability survey sample. Some concluding remarks are given in Section 8. Proofs are presented in the online Supplementary Material.

2 | PROBLEM SETUP

Let $(\mathbf{x}_i, y_i)$ be the measurements of the auxiliary variables $\mathbf{X}$ and the study variable Y associated with unit i, i=1, ⋯, N, where N is the population size. Let $\theta_N = N^{-1} \sum_{i=1}^{N} y_i$ be the finite population mean of the study variable, which is the parameter of interest. Let A be a probability sample with observations on the auxiliary variables $\mathbf{X}$; let B be the non-probability sample with information on both the study variable Y and the auxiliary variables $\mathbf{X}$. Let $n_A = |A|$ and $n_B = |B|$ be the respective sample sizes. Table 1 presents the general setup of the two sample structure for data integration. The non-probability sample B is not representative of the target population due to the unknown sample selection/inclusion mechanism.

TABLE 1 Data structure for the two survey samples

Sample	Type	X	*Y*	Representativeness
A	Probability Sample	✓	×	Yes
B	Non-probability Sample	✓	✓	No

The two sample datasets can also be represented by $\{(\mathbf{x}_i, w_i), i \in A\}$ and $\{(\mathbf{x}_i, y_i), i \in B\}$, where $w_i = \pi_i^{-1}$ are the survey weights for the probability sample A and $\pi_i = P(i \in A)$ are the first-order inclusion probabilities. If the study variable Y were observed for the probability sample A, the Horvitz–Thompson estimator of θ_N would be given by $\widehat{\theta}_{HT} = N^{-1}\sum_{i\in A} w_i y_i$. The mass imputation estimator to be introduced in the next section replaces the unobserved y_i by an imputed $\widehat{y}_i$ using data from both samples A and B.

Our discussions on important practical issues related to the mass imputation estimator are motivated by the analysis of the survey sample collected by the Pew Research Centre in 2015. The dataset, denoted by PRC, contains 56 variables which aim to reveal the relations between people and their communities. There is a total of 9,301 cases in the PRC dataset which are provided by eight different vendors with unknown sampling and data collection strategies. We treat the PRC dataset as a non-probability survey sample with the sample size $n_B = 9301$. Four variables of the PRC dataset are of particular interest to our analysis, which are pertinent to the research questions on ‘people and their communities’, and are treated as study variables. They are listed at the bottom of Table 2, among them three are binary variables: Talk with neighbours frequently (Y_1), Participated in school groups (Y_2), Participated in service organizations (Y_3), and one is treated as a continuous variable: Days had at least one drink last month (Y_4).

The general framework for mass imputation requires the availability of an existing probability sample taken from the same target population and containing a set of auxiliary variables which are common to the non-probability sample. It turns out that there are (at least) two such probability survey samples available, both were taken in 2015 from the general US population. This raises an important practical question on how mass imputation can be implemented under such scenarios, which will be further discussed in Sections 7 and 8.

The first probability sample is the volunteer supplement survey data from the Current Population Survey (CPS), which is one of the most recognized surveys in the United States. The CPS dataset contains $n_A = 80,075$ cases with measurements on volunteerism, which is highly relevant to the study variables considered in the PRC dataset. The second probability sample is the Behavioral Risk Factor Surveillance System (BRFSS) survey data. It is designed to measure behavioural risk factors for US residents and has a large sample size $n_A = 441,456$. Neither of the two probability samples contains measurements of the study variables, but both share a rich set of common survey items with the PRC dataset as shown in Table 2.

As a preliminary analysis, we examine marginal distributions of the variables contained in the three datasets. Table 2 contains the estimated population mean using each of the three datasets. For the PRC dataset, the sampling strategy is unknown and no survey weights are available, so the estimates presented are the unadjusted simple sample means. For the BRFSS and the CPS datasets where survey weights are available, survey weighed estimates are computed. The ‘NA’ in the table indicates that the variable is not available from the dataset. It can be seen that there are noticeable differences between the naive sample means from the PRC sample and the survey weighted estimates from the two probability samples for variables such as Origin (Hispanic/Latino), Education (High school or less), Household (with children), Health (Smoking) and Volunteer works. It is clear evidence that the PRC dataset is not a representative sample for the US population.

3 | THE MASS IMPUTATION ESTIMATOR

We now formally define the mass imputation estimator of the population mean θ_N under the two-sample setup presented in Table 2. Let δ be the indicator variable for the unit being included

TABLE 2 Estimated Population Means of Survey Items from the Three Samples

		PRC	CPS	BRFSS
Age category	<30	0.183	0.212	0.209
	>=30, <50	0.326	0.336	0.333
	>=50, <70	0.387	0.326	0.327
	>=70	0.104	0.126	0.131
Gender	Female	0.544	0.518	0.513
Race	White only	0.823	0.786	0.750
	Black only	0.088	0.125	0.126
Origin	Hispanic/Latino	0.093	0.156	0.165
Region	Northeast	0.200	0.180	0.177
	South	0.275	0.373	0.383
	West	0.299	0.235	0.232
Marital status	Married	0.503	0.528	0.508
Employment	Working	0.521	0.589	0.566
	Retired	0.243	0.143	0.179
Education	High school or less	0.216	0.407	0.427
	Bachelor's degree and above	0.416	0.309	0.263
Household	Presence of child in household	0.289	NA	0.368
	Home ownership	0.654	NA	0.672
Health	Smoke everyday	0.157	NA	0.115
	Smoke never	0.798	NA	0.833
Financial status	No money to see doctors	0.207	NA	0.133
	Having medical insurance	0.891	NA	0.878
	Household income < 20K	0.161	0.153	NA
	Household income > 100K	0.199	0.233	NA
Volunteer works	Volunteered	0.510	0.248	NA
Study variables	Talk with neighbours frequently	0.461	NA	NA
	Participated in school groups	0.210	NA	NA
	Participated in service organizations	0.141	NA	NA
	Days had at least one drink last month	5.301	NA	NA

Note: 'NA' indicates that the variable is not measured in the survey.

in the non-probability sample B. Note that $\delta = 1$ for all $i \in B$ and $\delta = 0$ for all $i \notin B$. This is fundamentally different from the standard setting for missing data analysis where the response indicator variable equals to 1 for respondents and 0 for nonrespondents, all part of the sample. It is also similar to missing data problems if we treat the entire finite population as a sample, which intuitively justifies the need for auxiliary information at the population level in analysing non-probability survey samples.

We first assume that each unit in the population has a non-zero probability to be included in the sample B, that is,

$$P(\delta = 1 \,|\, \mathbf{X} = \mathbf{x}) > 0 \tag{1}$$

for all $\mathbf{x}$ in the support of $\mathbf{X}$. Condition (1) is often referred to as the *Positivity Assumption*, which means that for any possible value $\mathbf{x}$, there is a positive probability for this type of units to be selected for sample B. It implies that the support of $\mathbf{X}$ in sample B coincides with the sample space of $\mathbf{X}$ in the population. The condition can easily be verified if all components of $\mathbf{X}$ are discrete. If certain components of $\mathbf{X}$ are continuous, it is still plausible to check the condition based on the distributions of $\mathbf{X}$ between the weighted distribution in the probability sample A and the empirical distribution in the non-probability sample B. If the two sample supports overlap to each other, then (1) is satisfied. Crump et al. (2009) proposed a systematic way of subsampling to handle the cases with limited overlaps in the context of causal inference.

The most crucial part of the mass imputation approach is the prediction model $f(y|\mathbf{x})$, which can be estimated by using the observed $(Y, \mathbf{X})$ from the non-probability sample B if

$$f(y\,|\,\mathbf{x},\ \delta = 1) = f(y\,|\,\mathbf{x}). \tag{2}$$

The prediction model $f(y|\mathbf{x})$ can then be used to create mass-imputed values of y using observed $\mathbf{x}$ for all the units in the probability sample A. Equation (2) can be termed as the *transportability* condition in the sense that the imputation model built from sample B is transportable to sample A. A sufficient condition for (2) to hold is the ignorability condition for the non-probability sample B, which is similar to the missing at random (MAR) assumption (Rubin, 1976) widely used in the literature on missing data analysis:

$$P(\delta = 1\,|\,\mathbf{X}, Y) = P(\delta = 1\,|\,\mathbf{X}). \tag{3}$$

In other words, transportability is a weaker condition than ignorability. Unfortunately, the ignorability condition cannot be formally tested using the sample itself.

Under assumptions (1)–(2), it is possible to consider a mass imputation estimator based on nearest neighbour imputation as suggested by Rivers (2007). Nearest neighbour imputation is a nonparametric method that does not require any parametric model assumptions. While nonparametric imputation methods can provide robust estimation, it suffers from curse of dimensionality, and the asymptotic bias of the nearest neighbour imputation is not negligible if the dimension of $\mathbf{x}$ is greater than one (Yang & Kim, 2020). In this paper, we consider a semi-parametric model for sample B with the first conditional moment specified as

$$E(Y\,|\,\mathbf{X} = \mathbf{x}) = m(\mathbf{x}; \boldsymbol{\beta}_0) \tag{4}$$

for some unknown $p\times 1$ vector $\boldsymbol{\beta}_0$ and a known function $m(\cdot;\ \cdot)$. The specification of $m(\mathbf{x};\ \boldsymbol{\beta})$ typically follows the mean function for generalized linear models. If Y is continuous, we can use a linear regression model and let $m(\mathbf{x}; \boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$; if Y is binary, we can use a logistic regression model and let $m(\mathbf{x}; \boldsymbol{\beta}) = \{1 + \exp(-\mathbf{x}'\boldsymbol{\beta})\}^{-1}$; if Y represents count data, we can use a log-linear model and let $m(\mathbf{x}; \boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$. Under the assumed transportability condition, the model parameters $\boldsymbol{\beta}$ can be estimated by the non-probability sample, B. We assume that $\widehat{\boldsymbol{\beta}}$ is the unique solution to the estimating equations

$$\widehat{U}(\boldsymbol{\beta}) = \frac{1}{n_B}\sum_{i\in B}\left\{y_i - m(\mathbf{x}_i; \boldsymbol{\beta})\right\}\mathbf{h}(\mathbf{x}_i; \boldsymbol{\beta}) = \mathbf{0} \tag{5}$$

for some p-dimensional vector of functions $\mathbf{h}(\mathbf{x};\ \boldsymbol{\beta})$. The estimating equations given in (5) may be constructed as the quasi-score equations from generalized linear models. Note that $E\{\widehat{U}(\boldsymbol{\beta}_0)\} = 0$ under the assumption (2) and the model (4). Under suitable moment conditions on $(Y,\mathbf{X})$ and smoothness conditions on $m(\mathbf{x};\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ similar to those discussed in Section 3.2 of Tsiatis (2006), it can be shown that the solution to (5), denoted as $\widehat{\boldsymbol{\beta}}$, is consistent for $\boldsymbol{\beta}_0$ under the ignorability assumption.

Once $\widehat{\boldsymbol{\beta}}$ is obtained from the non-probability sample B where both Y and $\mathbf{X}$ are observed, we can obtain predicted values $\widehat{y}_i = m(\mathbf{x}_i;\widehat{\boldsymbol{\beta}})$ for all $i \in A$, which is the so-called 'mass imputation'. The mass imputation estimator for the finite population mean $\theta_N = N^{-1}\sum_{i=1}^{N} y_i$ is computed as

$$\widehat{\theta}_I = \frac{1}{N}\sum_{i\in A} w_i \widehat{y}_i \tag{6}$$

using the survey weights w_i for the probability sample A. If the population size N is unknown, it can be estimated by $\widehat{N} = \sum_{i\in A} w_i$. Some asymptotic properties of the mass imputation estimator given in Equation (6) under the assumed semiparametric model (4) are presented in the next section.

4 | MAIN THEORY

We now discuss asymptotic properties of the mass imputation estimator given in Equation (6). Kim and Rao (2012) investigated the asymptotic properties of the mass imputation estimator in the context of combining two independent probability samples. We now consider an extension to mass imputation using a non-probability sample as a training sample. For the asymptotic framework, we assume that there is a sequence of finite populations and a sequence of samples as discussed in Fuller (2009), which provides a framework to allow the sample sizes go to infinity. Let $\widehat{\boldsymbol{\beta}}$ be the solution to (5) and define $\boldsymbol{\beta}^* = plim\,\widehat{\boldsymbol{\beta}}$, where the reference distribution is the sampling mechanism for sample B. Roughly speaking, if (3) holds, then $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, where $\boldsymbol{\beta}_0$ are the true parameters in the conditional model (4). Under certain regularity conditions including $V\{\widehat{U}(\boldsymbol{\beta}^*)\} = O_p(n_B^{-1})$, we can establish that

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = O_p(n_B^{-1/2}). \tag{7}$$

The following theorem presents an asymptotic expression of the mass imputation estimator.

Theorem 1. *Suppose that* ($\mathbf{x}$, y) *has bounded fourth moments over the sequence of finite populations and that (7) holds. Under the regularity conditions stated in the Supplementary Material, the mass imputation estimator (6) satisfies* $\widehat{\theta}_I = \tilde{\theta}_I + o_p(n_B^{-1/2})$, *where*

$$\tilde{\theta}_I = N^{-1}\sum_{i\in A} w_i m(\mathbf{x}_i;\boldsymbol{\beta}^*) + n_B^{-1}\sum_{i\in B}\left\{y_i - m(\mathbf{x}_i;\boldsymbol{\beta}^*)\right\} g(\mathbf{x}_i;\boldsymbol{\beta}^*) \tag{8}$$

with $g(\mathbf{x}_i;\boldsymbol{\beta}^*) = \mathbf{h}(\mathbf{x}_i;\boldsymbol{\beta}^*)'\mathbf{c}$,

$$\mathbf{c} = \left\{ n_B^{-1}\sum_{i\in B} \dot{\mathbf{m}}(\mathbf{x}_i;\boldsymbol{\beta}^*)\mathbf{h}(\mathbf{x}_i;\boldsymbol{\beta}^*)' \right\}^{-1} N^{-1}\sum_{i=1}^{N} \dot{\mathbf{m}}(\mathbf{x}_i;\boldsymbol{\beta}^*), \tag{9}$$

and $\dot{\mathbf{m}}(\mathbf{x};\boldsymbol{\beta}) = \partial m(\mathbf{x};\boldsymbol{\beta})/\partial\boldsymbol{\beta}$.

Proof of Theorem 1 is presented in the online Supplementary Material. The expression (8) is essentially the first-order Taylor linearization of $\widehat{\theta}_I$ in Equation (6) and the Taylor expansion is made around

$\boldsymbol{\beta} = \boldsymbol{\beta}^*$, not around the true parameter value $\boldsymbol{\beta}_0$. Note that we do not use the ignorability assumption in (3) to establish (8). The theorem does not imply that the estimator is unbiased.

Following (8), we can express

$$\tilde{\theta}_I - \theta_N = N^{-1}\left(\sum_{i\in A} w_i m_i^* - \sum_{i=1}^{N} m_i^*\right) + \left(n_B^{-1}\sum_{i=1}^{N}\delta_i g_i^* e_i^* - N^{-1}\sum_{i=1}^{N} e_i^*\right), \qquad (10)$$

where $m_i^* = m(\mathbf{x}_i; \boldsymbol{\beta}^*)$, $e_i^* = y_i - m(\mathbf{x}_i; \boldsymbol{\beta}^*)$ and $g_i^* = g(\mathbf{x}_i; \boldsymbol{\beta}^*)$. Thus, ignoring the higher order terms, the asymptotic bias is given by

$$\begin{aligned} E(\tilde{\theta}_I - \theta_N) &= E\left(n_B^{-1}\sum_{i=1}^{N}\delta_i g_i^* e_i^* - N^{-1}\sum_{i=1}^{N} e_i^*\right) \\ &= -E\left(N^{-1}\sum_{i=1}^{N} e_i^*\right) \\ &= E\left\{Cov_N(\delta\pi_B^{-1}, e^*)\right\}, \end{aligned} \qquad (11)$$

where $\pi_B = n_B/N$ and the second equality follows from $E\left[\sum_{i=1}^{N}\delta_i\{y_i - m(\mathbf{x}_i; \boldsymbol{\beta}^*)\}g_i^*\right] = 0$ by the definition of $\boldsymbol{\beta}^*$. Here, we used the notation $Cov_N(x, y) = N^{-1}\sum_{i=1}^{N}(x_i - \bar{x}_N)(y_i - \bar{y}_N)$, where $\bar{x}_N$ and $\bar{y}_N$ are the finite population means. If $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$, then $e_i^* = y_i - m(\mathbf{x}_i; \boldsymbol{\beta}_0) = e_i$ and the mass imputation estimator in (6) is unbiased. Otherwise, the bias is non-zero.

On the other hand, the asymptotic bias of the naive estimator, the simple sample mean $\bar{y}_B = n_B^{-1}\sum_{i\in B} y_i$, is given by

$$\begin{aligned} E(\bar{y}_B - \theta_N) &= E\left\{n_B^{-1}\sum_{i=1}^{N}\delta_i(y_i - \bar{y}_N)\right\} \\ &= E\{Cov_N(\delta\pi_B^{-1}, y)\}. \end{aligned} \qquad (12)$$

Thus, comparing (11) with (12), we find that the absolute value of the bias of $\tilde{\theta}_I$ is smaller than that of $\bar{y}_B$ if the variance of e_i^* will be smaller than the variance of y_i, which is quite possible when the prediction model of y estimated from sample B is reasonable, even if it is not entirely correct. Thus, the mass imputation estimator reduces the bias even when the sampling mechanism for sample B is non-ignorable.

We now derive the asymptotic variance formula for the mass imputation estimator. By Equation (10), we have $V(\tilde{\theta}_I - \theta_N) = V_A + V_B$, where

$$V_A = V\left(N^{-1}\sum_{i\in A} w_i m_i^* - N^{-1}\sum_{i=1}^{N} m_i^*\right) \qquad (13)$$

is the variance component for sample A and

$$V_B = V[E\{N^{-1}\sum_{i=1}^{N}(\delta_i g_i \pi_B^{-1} - 1)e_i^* \mid \mathbf{X}, \boldsymbol{\delta}\}] + E[V\{N^{-1}\sum_{i=1}^{N}(\delta_i g_i \pi_B^{-1} - 1)e_i^* \mid \mathbf{X}, \boldsymbol{\delta}\}] \qquad (14)$$

is the variance component for sample B. In Equation (13), only the design-based variance under the probability sampling design for sample A remains. The reference distribution in Equation (14) is the joint distribution of the sampling mechanism for sample B and the superpopulation model in (4). The conditional expectation $E\{\cdot \mid \mathbf{X}, \boldsymbol{\delta}\}$ is with respect to the model for$e_i^* = y_i - m(\mathbf{x}_i; \boldsymbol{\beta}^*)$ conditional on $\boldsymbol{\delta}$. The first term of (14) is unknown in general because the sampling mechanism for sample B is unknown.

If the sampling mechanism for sample B is ignorable as defined in Equation (3), we have $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $e_i^* = e_i$. In this case, the first term of (14) disappears because $E(e_i \mid \mathbf{x}_i, \delta_i) = E\{y_i - m(\mathbf{x}_i; \boldsymbol{\beta}_0) \mid \mathbf{x}_i, \delta_i\} = E\{y_i - m(\mathbf{x}_i; \boldsymbol{\beta}_0) \mid \mathbf{x}_i\} = 0$. Thus, under the ignorability assumption, we can simplify V_B as

$$\begin{aligned} V_B &= E\left[V\left\{N^{-1}\sum_{i=1}^{N}(\delta_i g_i \pi_B^{-1} - 1)e_i \middle| \mathbf{X}, \boldsymbol{\delta}\right\}\right] \\ &= E\left\{N^{-2}\sum_{i=1}^{N}(\delta_i g_i \pi_B^{-1} - 1)^2 e_i^2\right\}. \end{aligned} \tag{15}$$

The second equality follows by the independence of e_i's in the superpopulation model. Thus, the first term of (14) can be understood as the additional variance increase under model misspecification. Furthermore, if $n_B/N = o(1)$ then we can simply use

$$V_B = E\left[n_B^{-2}\sum_{i\in B}\left(e_i g_i\right)^2\right]. \tag{16}$$

Note that we can easily estimate V_B in Equation (16) even though the sampling mechanism for sample B is unknown. The asymptotic variance $V(\tilde{\theta}_I - \theta_N)$ consists of two parts. The first term V_A is of order $O(n_A^{-1})$ and the second term V_B is of order $O(n_B^{-1})$. If $n_A/n_B = o(1)$, that is, the sample size n_B is much larger than n_A, the term V_B is of smaller order and the leading term of the total variance is V_A. Otherwise the two variance components both contribute to the total variance.

Under the linear regression model $Y_i = \mathbf{x}_i'\boldsymbol{\beta} + e_i$ with $e_i \sim (0, \sigma_e^2)$, independent among all i, we have $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}} = \left(\sum_{i\in B}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i\in B}\mathbf{x}_i y_i$. The mass imputation estimator of (6) under the regression model is given by $\hat{\theta}_{I,reg} = N^{-1}\sum_{i\in A} w_i \mathbf{x}_i'\hat{\boldsymbol{\beta}}$. If the probability sample A is selected by simple random sampling, the asymptotic variance of $\hat{\theta}_{I,reg}$ is given by $V(\hat{\theta}_{I,reg} - \theta_N) \approx V_A + V_B$, where

$$\begin{aligned} V_A &= V\left(n_A^{-1}\sum_{i\in A}\mathbf{x}_i'\boldsymbol{\beta}\right) = n_A^{-1}\boldsymbol{\beta}'\Sigma_{xx}\boldsymbol{\beta}, \\ V_B &= V\left\{N^{-1}\sum_{i=1}^{N}\left(\delta_i \mathbf{x}_i'\mathbf{c}\pi_B^{-1} - 1\right)e_i\right\} \cong E\left\{n_B^{-2}\sum_{i\in B}\left(\mathbf{x}_i'\mathbf{c}\right)^2 e_i^2\right\}, \end{aligned}$$

$\Sigma_{xx} = N^{-1}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'$, $\mathbf{c} = \left(n_B^{-1}\sum_{i\in B}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\bar{\mathbf{x}}_N$ and $\bar{\mathbf{x}}_N = N^{-1}\sum_{i=1}^{N}\mathbf{x}_i$. If $\mathbf{x}_i = (1, x_i)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ and assuming $\pi_B = n_B/N$ is negligible, the asymptotic variance reduces to

$$V\left(\hat{\theta}_{I,reg} - \theta_N\right) \approx \frac{1}{n_A}(\beta_1)^2\sigma_x^2 + \frac{1}{n_B}\sigma_e^2 + E\left[\frac{(\bar{x}_N - \bar{x}_B)^2}{\sum_{i\in B}(x_i - \bar{x}_B)^2}\right]\sigma_e^2,$$

where $\sigma_x^2 = N^{-1}\sum_{i=1}^{N}(x_i - \bar{x}_N)^2$ and $\bar{x}_B = n_B^{-1}\sum_{i\in B} x_i$. If sample B is a random sample from the population, then the third term is of order $O(n_B^{-2})$ and becomes negligible. However, since sample B is a non-probability sample, the third term might not be negligible.

Variance estimation for the mass imputation estimator (6) requires estimation of the two components V_A and V_B. The first component can be estimated by

$$\hat{V}_A = \frac{1}{N^2}\sum_{i\in A}\sum_{j\in A}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} w_i m(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) w_j m(\mathbf{x}_j;\hat{\boldsymbol{\beta}}),$$

where $\pi_{ij} = P(i, j \in A)$ are the joint inclusion probabilities.

The second component can be estimated by

$$\hat{V}_B = \frac{1}{n_B^2}\sum_{i\in B}\hat{e}_i^2\{\mathbf{h}(\mathbf{x}_i;\hat{\boldsymbol{\beta}})'\hat{\mathbf{c}}\}^2, \tag{17}$$

where $\hat{e}_i = y_i - m(\mathbf{x}_i;\hat{\boldsymbol{\beta}})$ and $\hat{\mathbf{c}} = \left\{n_B^{-1}\sum_{i\in B}\dot{\mathbf{m}}(\mathbf{x}_i;\boldsymbol{\beta}^*)\mathbf{h}(\mathbf{x}_i;\boldsymbol{\beta}^*)'\right\}^{-1} N^{-1}\sum_{i\in A} w_i\dot{\mathbf{m}}(\mathbf{x}_i;\boldsymbol{\beta}^*)$. Under the linear regression model, $\mathbf{h}(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) = \mathbf{x}_i$ and $\hat{\mathbf{c}} = \left(n_B^{-1}\sum_{i\in B}\mathbf{x}_i\mathbf{x}_i'\right)^{-1} N^{-1}\sum_{i\in A} w_i\mathbf{x}_i$.

The total variance of $\hat{\theta}_I$ can be estimated by $\hat{V}(\hat{\theta}_I - \theta_N) = \hat{V}_A + \hat{V}_B$.

5 | BOOTSTRAP VARIANCE ESTIMATION

The variance estimator presented in Section 4 is based on the linearization method. The closed-form formula for the asymptotic variance is simple to implement. However, to compute $\hat{V}_B$ in Equation (17), we need to use individual observations of $(\mathbf{x}_i, y_i)$ in sample B, which is not necessarily available when only the sample A with mass imputed responses is released to the public data users. Note that one of the goals of mass imputation is to produce a representative sample A with synthetic observations on the response variable using sample B as a training dataset. Once the mass imputation is performed, the training data are no longer necessary in computing point estimators. It is therefore desirable to develop a variance estimation method that does not require access to observations in sample B.

To achieve this goal, we propose a bootstrap method (Rao & Wu, 1988) for variance estimation that creates a replicated set of synthetic data $\{\hat{y}_i^{(k)},\ i \in A\}$ corresponding to each set of bootstrap weights $\{w_i^{(k)},\ i \in A\}$, k=1, ⋯, L associated with sample A only. The method enables users to correctly estimate the variance of the mass imputation estimator $\hat{\theta}_I$ without access to the training data $\{(y_i, \mathbf{x}_i)\colon i \in B\}$ from sample B. The data file will contain additional columns of $\{\hat{y}_i^{(k)}\colon i \in A\}$ associated with the columns of bootstrap weights $\{w_i^{(k)}; i \in A\}$, k=1, ⋯, L, where L is the number of replicates created from sample A. Kim and Rao (2012) also considered a similar method in the context of survey integration from non-nested two-phase sampling.

In order to develop a valid bootstrap method for the mass imputation estimator $\hat{\theta}_I$ in Equation (6), it is critical to develop a valid bootstrap method for estimating $V(\hat{\boldsymbol{\beta}})$ when $\hat{\boldsymbol{\beta}}$ is computed from (5). Under assumptions (2) and (4), we have

$$V(\hat{\boldsymbol{\beta}}) \doteq J^{-1}\Omega J^{-1\prime}, \tag{18}$$

where $J = E\left\{n_B^{-1}\sum_{i\in B}\dot{\mathbf{m}}_i\mathbf{h}_i'\right\}$, $\Omega = E\left\{n_B^{-2}\sum_{i\in B}E(e_i^2|\mathbf{x})\mathbf{h}_i\mathbf{h}_i'\right\}$ with $\dot{\mathbf{m}}_i = \dot{\mathbf{m}}(\mathbf{x}_i;\boldsymbol{\beta}_0)$ and $\mathbf{h}_i = \mathbf{h}(\mathbf{x}_i;\boldsymbol{\beta}_0)$. The reference distribution for (18) is the joint distribution of the superpopulation model (4) and the unknown sampling mechanism for the non-probability sample B. Interestingly, the variance formula in (18) equals exactly to the variance of $\widehat{\boldsymbol{\beta}}$ when sample B is selected by simple random sampling (SRS). That is, even though the sampling design for sample B is not SRS, its effect on the variance of $\widehat{\boldsymbol{\beta}}$ is essentially the same with SRS. This is due to the MAR assumption in (2) which makes the effect of the sampling design for estimating $\boldsymbol{\beta}$ ignorable even though it is still not ignorable for $\theta_N = N^{-1} \sum_{i=1}^{N} y_i$. Therefore, we can safely ignore the selection mechanism for sample B when estimating $\boldsymbol{\beta}$ and develop a valid bootstrap method for estimating the variance of $\widehat{\boldsymbol{\beta}}$ using the bootstrap method for SRS.

Our proposed bootstrap method can be described as the following four steps:

1. Create the kth set of replication weights $\{w_i^{(k)}, i \in A\}$ based on the sampling design for the probability sample A.
2. Generate the kth bootstrap sample of size n_B from sample B using simple random sampling with replacement and compute $\widehat{\boldsymbol{\beta}}^{(k)}$ using the same estimation Equation (5) applied to the bootstrap sample.
3. Use $\widehat{\boldsymbol{\beta}}^{(k)}$ obtained from Step 2 to compute $\widehat{y}_i^{(k)} = m(\mathbf{x}_i; \widehat{\boldsymbol{\beta}}^{(k)})$ for each $i \in A$ and create a new column $\{\widehat{y}_i^{(k)}, i \in A\}$ alongside $\{w_i^{(k)}, i \in A\}$ for the sample A dataset.
4. Repeat Steps 1–3, independently, for k=1, ⋯, L with a pre-chosen L.

Step 1 needs to follow standard practice in survey sampling on creating replication weights for design-based variance estimation. See, for instance, Chapter 10 of Wu and Thompson (2020). The final sample A dataset contains additional columns for the replicate versions of mass imputed values $\{\widehat{y}_i^{(k)}, i \in A\}$ and the sets of replication weights $\{w_i^{(k)}, i \in A\}$, k=1, ⋯, L. The replicate versions of the mass imputation estimator $\widehat{\theta}_I$ are computed as

$$\widehat{\theta}_I^{(k)} = \frac{1}{N} \sum_{i \in A} w_i^{(k)} \widehat{y}_i^{(k)}, \quad k = 1, \cdots, L, \tag{19}$$

and the resulting bootstrap variance estimator of $\widehat{\theta}_I$ is computed as

$$\widehat{V}_b(\widehat{\theta}_I) = \frac{1}{L} \sum_{k=1}^{L} \left(\widehat{\theta}_I^{(k)} - \widehat{\theta}_I \right)^2 . \tag{20}$$

Note that once $\{(w_i^{(k)}, \widehat{y}_i^{(k)})$: $i \in A\}$ are obtained for k=1, ⋯, L, in addition to the original mass imputation data $\{(w_i, \widehat{y}_i)$: $i \in A\}$, the users do not need to have access to sample B.

The following theorem establishes the consistency of the proposed bootstrap variance estimator. Proof of the theorem is presented in the online Supplementary Material.

Theorem 2. *Suppose that the conditions of Theorem 1 hold and assume that the sampling mechanism for sample B is ignorable as defined in Equation (3). Under the additional assumption stated in the Supplementary Material, the bootstrap variance estimator given by (20) satisfies*

$$\widehat{V}_b(\widehat{\theta}_I) = V(\widehat{\theta}_I - \theta_N) + o_p(n_B^{-1}). \tag{21}$$

6 | SIMULATION STUDY

A limited simulation study is performed to (i) evaluate the departure of the imputation model to the performance of mass imputation estimator, (ii) check the validity of the proposed variance estimators, and (iii) compare the mass imputation estimator with the inverse probability weighted (IPW) estimator of Chen et al. (2020) using a logistic regression model for estimating the propensity scores.

The setup for the simulation study employed a 3×2 factorial structure with two factors. The first factor is the superpopulation model that generates the finite population. The second factor is the sample size for sample B. We used two levels for n_B, where $n_B = 500$ or $n_B = 1000$. We generated the following three models for finite populations of size N=100,000. The variables, x_i and e_i, are independently generated from $N(2, 1)$ and $N(0, 1)$, respectively.

The study variable y_i are constructed differently for each model:

Model I $\quad y_i = 1 + 2x_i + e_i$.
Model II $\quad y_i = 3 + x_i + 2e_i$.
Model III $\quad y_i = 2.5 + 0.5x_i^2 + e_i$.

Model I generates a finite population with a high correlation between x and y ($r^2 = 0.8$), Model II generates a finite population with a low correlation ($r^2 = 0.2$), and Model III generates a finite population where the linear relationship fails. Model III is included to check the effect of model mis-specification for the imputation model.

From each of the three populations, we generated two independent samples. We use simple random sampling of size $n_A = 500$ to obtain sample A. In selecting sample B of size n_B, where $n_B \in \{500, 1000\}$, we create two strata where Stratum 1 consists of elements with $x_i \leq 2$ and Stratum 2 consists of elements with $x_i > 2$. Within each stratum, we select n_h elements by simple random sampling, independent between the two strata, where $n_1 = 0.7n_B$ and $n_2 = 0.3n_B$. We assume that the stratum information is unavailable at the time of data analysis. Using the two samples A and B, we compute four estimators of $\theta_N = N^{-1}\sum_{i=1}^{N} y_i$:

1. The sample mean from sample A: $\widehat{\theta}_A = n_A^{-1}\sum_{i \in A} y_i$.
2. The naive estimator (sample mean) from sample B: $\widehat{\theta}_B = n_B^{-1}\sum_{i \in B} y_i$.
3. The mass imputation estimator from sample A given in Equation (6) using $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ where $(\widehat{\beta}_0, \widehat{\beta}_1)$ are the estimated regression coefficients obtained from sample B.
4. The IPW estimator proposed by Chen et al. (2020): $\widehat{\theta}_{IPW} = N^{-1}\sum_{i \in B}\widehat{\pi}_i^{-1} y_i$, where the propensity scores, $\pi_i = \pi(\mathbf{x}_i;\boldsymbol{\phi}) = \{1 + \exp(-\phi_0 - \phi_1 x_i)\}^{-1}$ with $\mathbf{x}_i = (1, x_i)'$ and $\boldsymbol{\phi} = (\phi_0, \phi_1)'$, are estimated by using $\widehat{\boldsymbol{\phi}}$ which solves the following score equations:

$$U(\boldsymbol{\phi}) = \sum_{i \in B} \mathbf{x}_i - \sum_{i \in A} w_i \pi(\mathbf{x}_i;\boldsymbol{\phi})\mathbf{x}_i = \mathbf{0}. \tag{22}$$

The sample mean of sample A, not available in practice but computed in the simulation, serves as a gold standard estimator. Results are based on 1000 repeated simulation runs. Table 3 presents the Monte Carlo bias and variance, and the relative mean squared error of the four point estimators. The relative mean squared error of estimator $\widehat{\theta}$ is defined as

$$ReMSE = \frac{MSE(\widehat{\theta})}{MSE(\widehat{\theta}_A)}.$$

TABLE 3 Monte Carlo bias, variance and relative mean square error (ReMSE)

		Model I			Model II			Model III		
n_B	Estimator	Bias	Var ($\times 10^3$)	ReMSE ($\times 10^2$)	Bias	Var ($\times 10^3$)	ReMSE ($\times 10^2$)	Bias	Var ($\times 10^3$)	ReMSE ($\times 10^2$)
500	$\widehat{\theta}_A$	0.00	9.54	100	0.00	9.50	100	0.00	10.81	100
	$\widehat{\theta}_B$	−0.64	4.68	4,299	−0.32	7.98	1,151	−0.64	4.43	3,791
	$\widehat{\theta}_I$	0.00	10.03	105	0.00	10.26	108	−0.06	10.32	128
	$\widehat{\theta}_{IPW}$	−0.03	13.81	156	−0.01	11.27	121	−0.04	21.03	205
1000	$\widehat{\theta}_A$	0.00	9.69	100	0.00	9.87	100	0.00	10.62	100
	$\widehat{\theta}_B$	−0.64	2.22	4,266	−0.32	4.03	1,065	−0.64	22.13	3,861
	$\widehat{\theta}_I$	0.00	8.93	93	0.00	6.20	63	−0.06	8.61	118
	$\widehat{\theta}_{IPW}$	−0.04	11.77	137	−0.02	6.82	72	−0.04	16.11	167

Note that the sample mean $\widehat{\theta}_A$ from sample A is not available in practice but is computed here as the gold standard. Table 3 shows that, with models I and II where the linear regression model holds, the mass imputation estimator is unbiased for the population mean. The naive mean estimator of sample B underestimates the population mean for all scenarios considered in the simulation. When the size of sample B for training data is larger than the size of sample A ($n_B = 1000$), it is possible that the mass imputation estimator has a smaller MSE than the gold standard. Under the simple regression model, the asymptotic variance of the mass imputation estimator is

$$V(\widehat{\theta}_I - \theta_N) \approx \frac{1}{n_A}\sigma_x^2\beta_1^2 + \frac{1}{n_B}\sigma_e^2 E\left\{1 + \frac{(\overline{x}_N - \overline{x}_B)^2}{s_{x,B}^2}\right\},$$

where $s_{x,B}^2 = (n_B - 1)^{-1}\sum_{i\in B}(x_i - \overline{x}_B)^2$, while the variance of sample mean of sample A is $V(\widehat{\theta}_A) = \sigma_y^2/n_A = (\beta_1^2\sigma_x^2 + \sigma_e^2)/n_A$. Thus, if n_B is much larger than n_A, the mass imputation estimator can be more efficient than the sample mean of A. The IPW estimator is less efficient than the mass imputation estimator for all cases considered in the current simulation setup.

The imputation model using the simple linear regression is incorrectly specified for model III. The mass imputation estimator is modestly biased. Nonetheless, the performance in terms of MSE is better than the IPW estimator because the mass imputation estimator has much smaller variance than the IPW estimator, even though the absolute bias is larger, when the linear relationship fails.

Table 4 presents Monte Carlo mean and relative bias of the two variance estimators of the mass imputation estimator using linearization and bootstrap. Both variance estimators show negligible relative biases. In particular, the bootstrap variance estimator shows good performance even under model III.

7 | APPLICATION TO THE PRC DATASET

We now apply the proposed mass imputation techniques described in Section 3 and compute point and variance estimates of population means for the four study variables in the PRC dataset, which were introduced in Section 2:

TABLE 4 Monte Carlo mean and relative bias (R.B.) of variance estimators

		Linearization		Bootstrap	
n_B	Model	Mean	R.B.	Mean	R.B.
500	I	0.0102	0.019	0.0103	0.023
	II	0.0109	0.067	0.0110	0.069
	III	0.0099	−0.036	0.0106	0.026
1000	I	0.0091	0.017	0.0091	0.019
	II	0.0064	0.039	0.0065	0.041
	III	0.0082	−0.043	0.0086	−0.009

1. Y_1: Talk with neighbours frequently,
2. Y_2: Participated in school groups,
3. Y_3: Participated in service organizations,
4. Y_4: Days had at least one drink last month.

Note that PRC is the non-probability sample and CPS and BRFSS are the two existing probability samples to be used to provide information on auxiliary variables. The PRC sample was collected through web panels, the BRFSS survey was conducted by telephone interviews, and the CPS survey used a mixed mode with 11% of units surveyed by telephone and the remaining part of the survey by personal interviews.

Prior to any formal analysis of the datasets, we would like to check whether the positivity and the transportability assumptions are reasonable. The support of the auxiliary variables in the PRC dataset can be checked and compared to the support in the CPS sample or the BRFSS sample, which confirms the positivity assumption. The transportability assumption cannot be formally tested. Sensitivity analyses are a useful tool for checking the assumptions.

Our analysis on the PRC dataset focuses on three important practical aspects of the method. We first investigate how the relative sizes of sample A and sample B impact the mass imputation estimator and its associated variance estimator. We then discuss the covariates selection for the prediction model, and apply different strategies based on the availability of auxiliary variables from one or both probability samples. Lastly, we make comparisons between the mass imputation estimators and the IPW estimators. A sensitivity analysis on the model assumptions is also included.

7.1 | Impact of relative sample sizes

Recall from Section 2 that the two probability samples CPS and BRFSS contain respectively $n_A = 80,075$ and 441,456 cases while the non-probability PRC dataset has $n_B = 9301$ units. The sample size n_A is much larger than n_B for both CPS and BRFSS. To investigate other possible scenarios where n_A and n_B has different ratios, we draw three subsamples from the original BRFSS dataset by the simple random sampling without replacement method. The resulting subsamples, denoted by BRFSS$^{(1)}$, BRFSS^{t}(2) and BRFSS$^{(3)}$, have sample size $n_A^* = 80,000$, 8000 and 800, respectively. Survey weights for each of the subsamples are computed as $w_i n_A / n_A^*$, where w_i is the weight of unit i in the original BRFSS sample with $n_A = 441,456$.

The choices of covariates for the prediction model are constrained by the set of covariates observed in both the probability sample and the non-probability sample. The common set of covariates between BRFSS and PRC, however, differs from the set between CPS and PRC. We use the set of

TABLE 5 Mass imputation estimates of population means by different probability samples

Response	PRC	CPS	BRFSS	BRFSS$^{(1)}$	BRFSS$^{(2)}$	BRFSS$^{(3)}$
Y_1	0.461	0.458	0.457	0.456	0.458	0.468
Y_2	0.210	0.206	0.200	0.200	0.203	0.210
Y_3	0.141	0.135	0.133	0.133	0.134	0.137
Y_4	5.301	4.986	4.931	4.921	4.931	4.997

covariates which are available in all three datasets for the current investigation. We use a logistic regression model for binary responses and a linear regression model for the continuous response.

We first compute the point estimates of the population means by the proposed mass imputation method using five different probability samples, and the results are presented in Table 5. The first row specifies which probability sample is used as sample A except for the column under 'PRC' which represents the naive estimates of simple sample means. It can be seen that the three larger probability samples BRFSS, BRFSS$^{(1)}$ and BRFSS$^{(2)}$ with $n_A = 441,456$ and $n_A^* = 80,000$ and 8000 produce almost identical results. The smallest probability sample BRFSS$^{(3)}$ with $n_A^* = 800$ leads to noticeably different results, indicating potential inconsistent estimates when the size of the probability sample is too small. We also notice that, despite the differences between the CPS and the BRFSS samples, their corresponding mass imputation estimators produce similar results with the use of the same set of auxiliary variables.

We further look at variance estimators with different probability samples, using the linearization and bootstrap methods described in Sections 4 and 5. The linearization variance estimator is based on the formula $V(\tilde{\theta}_I - \theta_N) = V_A + V_B$ given by Theorem 1, where V_A is the design-based variance component under the probability sampling design for sample A, and V_B is the variance component for sample B under the prediction model. Unfortunately, detailed design information other than the survey weights is not available for either the BRFSS or the CPS sample. We use an approximate variance formula for V_A by assuming that the survey design is single-stage PPS sampling with replacement, a strategy often used by survey data analyst for the purpose of variance estimation. The bootstrap variance estimator is computed based on L=5000 bootstrap samples.

Results for variance estimators with decomposition of the variance components $(\widehat{V}_A, \widehat{V}_B)$ for the linearization method are reported in Table 6. Variance decomposition cannot be done for the bootstrap method. We have the following major observations from Table 6: (i) the two variance estimation methods, linearization and bootstrap, give similar results in total variance for all cases; (ii) the two original large probability samples CPS and BRFSS produce similar total variances for all cases, with the design variance component $\widehat{V}_A$ making a negligible contribution; (iii) when the size of the probability sample becomes smaller, from BRFSS$^{(1)}$ to BRFSS$^{(2)}$ and then BRFSS$^{(3)}$, the total variance becomes larger, and the variance component $\widehat{V}_A$ from the probability sampling design for sample A becomes dominant.

7.2 | Impact of the prediction model

The second part of the analysis is on covariate selection for the prediction model. We consider following selection strategies: (a) Use the common set of variables available in all three datasets. This is what has been used in Section 7.1, and the resulting model is denoted as ξ(Partial), since the common set of three is a partial set of all available. (b) Use the common set of variables available in the two datasets PRC and CPS or PRC and BRFSS, leading to two larger but different common sets of variables for the two probability samples. The resulting models are denoted as ξ(All). (c) Use the set of covariates which are

TABLE 6 Variance and variance components of the mass imputation estimators

Y	Method	CPS	BRFSS	BRFSS$^{(1)}$	BRFSS$^{(2)}$	BRFSS$^{(3)}$
Y_1	L	4.195	4.323	4.859	11.848	75.193
		(0.087, 4.108)	(0.144, 4.179)	(0.763, 4.096)	(7.732, 4.116)	(70.226, 4.966)
	Boot	4.055	4.187	5.031	12.170	72.559
Y_2	L	2.602	2.599	2.729	4.737	23.802
		(0.037, 2.565)	(0.039, 2.559)	(0.211, 2.518)	(2.121, 2.616)	(20.495, 3.307)
	Boot	2.607	2.615	2.822	4.532	22.840
Y_3	L	1.922	1.910	1.950	2.662	10.626
		(0.016, 1.906)	(0.016, 1.894)	(0.083, 1.867)	(0.791, 1.871)	(8.211, 2.415)
	Boot	1.930	1.886	1.942	2.719	10.818
Y_4	L	978	1,010	1,062	1,785	8,887
		(12, 966)	(16, 994)	(85, 977)	(818, 967)	(7,744, 1,143)
	Boot	952	996	1,091	1,808	8,759

Note: L: Linearization; Boot: Bootstrap. All numbers multiplied by 10^5.

available in both PRC and CPS (or both PRC and BRFSS) but are selected through the backward variable selection algorithm. The resulting models are denoted as ξ(Select). *p*-values for covariates in model ξ(All) and ξ(Select) are listed in Table 7 and Table 8.

Mass imputation-based point estimates $\widehat{\theta}_I$ for the population means of the four response variables using different sets of covariates with the probability Sample *A* as either CPS or BRFSS are presented in Table 9, with the standard errors in parentheses obtained by using the linearization variance estimator. Major observations from Table 9 can be summarized as follows. (i) With a chosen sample *A*, the point estimates under model ξ(All) are very similar to the estimates under model ξ(Select) which drops some covariates using the backward variable selection algorithm; (ii) the point estimates under model ξ(All) behave quite differently to the estimates under model ξ(Partial) which excludes some covariates based on the availability from a second probability sample; (iii) the linearization variance estimates under model ξ(Select) are always smaller than the variance estimates under model ξ(All), showing some efficiency gain by eliminating non-significant factors from the model; (iv) the point estimates using CPS are very similar to the estimates using BRFSS under the same model ξ(Partial) with the same set of covariates; (v) the point estimates using CPS under model ξ(All) are quite different to the estimates using BRFSS under model ξ(All). Note that the two models use two different sets of covariates. Some additional covariates available with one particular probability sample may be highly correlated with some of the response variables. For instance, the covariate 'Volunteer works', which is available in CPS but not in BRFSS, is related to the response variables Y_1, Y_2 and Y_3. Similarly, health related covariates, which are available in BRFSS but not in CPS, likely explain the evident difference of $\widehat{\theta}_I$ for response Y_4 when the BRFSS probability sample is used.

7.3 | Comparing mass imputation with inverse probability weighting

The third part of the analysis is to make comparisons between the mass imputation estimator $\widehat{\theta}_I$ and the IPW estimator $\widehat{\theta}_{IPW}$ of Chen et al. (2020), which is also examined in the simulation study reported in Section 6. The mass imputation estimators are computed based on model ξ(Select). The IPW estimators are based on the logistic regression model for the propensity scores using the same set of

TABLE 7 p-values for covariates in model ξ(All)

	CPS				BRFSS			
Covariates	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4
Intercept	*	*	*	*	*	*	*	*
Age	*	*	0.188	*	*	0.044	0.305	0.071
Female	*	0.990	*	*	*	0.175	*	*
White only	0.010	0.163	0.665	0.066	0.064	0.074	0.879	0.086
Black only	*	*	*	0.589	0.002	*	*	0.805
Hispanic/Latino	0.122	*	*	0.335	0.164	0.001	*	0.108
Northeast	*	0.514	0.068	0.987	*	0.617	0.125	0.688
South	0.007	*	0.252	0.985	0.011	*	0.461	0.933
West	0.087	0.001	0.668	0.060	0.032	*	0.908	0.013
Married	0.021	*	*	0.769	0.142	*	*	0.030
Working	0.864	0.062	0.010	*	0.406	0.001	*	*
Retired	0.007	0.912	0.002	*	*	0.006	*	*
High school or less	0.364	0.134	0.304	0.016	0.314	*	*	*
Bachelor's degree and above	0.023	0.039	0.189	0.005	0.038	*	*	*
Presence of child in household	NA	NA	NA	NA	*	*	*	*
Home ownership	NA	NA	NA	NA	0.036	*	*	0.001
Smoke everyday	NA	NA	NA	NA	0.847	0.893	0.067	0.311
Smoke never	NA	NA	NA	NA	*	0.031	*	*
No money to see doctors	NA	NA	NA	NA	*	*	*	0.018
Having medical insurance	NA	NA	NA	NA	0.006	*	0.009	0.338
Household income < 20K	0.003	0.025	0.038	0.003	NA	NA	NA	NA
Household income > 100K	0.075	0.057	0.213	*	NA	NA	NA	NA
Volunteered	*	*	*	0.438	NA	NA	NA	NA

Note: '*' indicates that p-value <0.001.

covariates as in the model ξ(Select). The point estimates along with standard errors obtained using the linearization variance estimator for the four response variables are reported in Table 10. The naive estimator has no meaningful variance to report. The two estimators $\widehat{\theta}_I$ and $\widehat{\theta}_{IPW}$ provide comparable results in both bias and variance, regardless whether CPS or BRFSS is used as the probability sample A.

7.4 | Sensitivity analysis

In the final part of the analysis, we present results from a sensitive analysis of the transportability assumption on the imputation model used for analysing the PRC non-probability survey sample and

TABLE 8 p-values for covariates in Model ξ(Select)

	CPS				BRFSS			
Covariates	Y_1	Y_2	Y_3	Y_4	Y_1	Y_2	Y_3	Y_4
Intercept	*	*	*	*	*	*	*	*
Age	*	*	×	*	*	0.044	×	0.072
Female	*	×	*	*	*	×	*	*
White only	0.009	×	×	0.003	0.068	0.074	×	0.011
Black only	*	*	*	×	0.002	*	*	×
Hispanic/Latino	0.118	*	*	×	×	0.001	*	0.094
Northeast	*	×	0.013	×	*	×	0.095	×
South	0.007	*	0.095	×	0.009	*	×	×
West	0.097	*	×	0.009	0.016	*	×	0.003
Married	0.020	*	*	×	0.147	*	*	0.022
Working	×	0.027	0.005	*	×	0.002	*	*
Retired	0.002	×	*	*	*	0.008	*	*
High school or less	×	0.128	×	0.013	×	*	0.001	*
Bachelor's degree and above	0.005	0.031	0.029	0.003	0.004	*	*	*
Presence of child in household	NA	NA	NA	NA	*	*	*	*
Home ownership	NA	NA	NA	NA	0.026	*	*	0.001
Smoke everyday	NA	NA	NA	NA	×	×	0.081	×
Smoke never	NA	NA	NA	NA	*	*	*	*
No money to see doctors	NA	NA	NA	NA	*	*	*	0.010
Having medical insurance	NA	NA	NA	NA	0.004	*	0.008	×
Household income < 20K	0.002	0.027	0.021	0.001	NA	NA	NA	NA
Household income > 100K	0.076	0.048	×	*	NA	NA	NA	NA
Volunteered	*	*	*	×	NA	NA	NA	NA

Note: '*' indicates that p-value <0.001. '×' indicates that the covariate is not selected by the backward variable selection algorithm.

comparisons of the mass imputation estimator to the survey weighted Hájek estimator and the IPW estimator for the population mean. The procedures involve

1. choosing a 'study variable' (Z) which is available in both the PRC non-probability sample and one of the probability samples (so the Hájek estimator can be computed for the corresponding population mean);
2. selecting a set of auxiliary variables ($\mathbf{X}^*$) which are available in both the non-probability sample and the probability samples and can be used as predictors for the choosing study variable;
3. fitting a suitable model $f(z|\mathbf{x}^*)$, depending on the chosen Z, separately using the non-probability sample and the probability sample;

TABLE 9 Mass imputation estimates of population means using different sets of covariates

		$\widehat{\theta}_I$		
Response	**Sample *A***	**ξ(Partial)**	**ξ(All)**	**ξ(Select)**
Y_1	CPS	0.458(0.006)	0.404(0.007)	0.402(0.006)
	BRFSS	0.457(0.007)	0.446(0.007)	0.447(0.006)
Y_2	CPS	0.206(0.005)	0.134(0.004)	0.134(0.004)
	BRFSS	0.200(0.005)	0.198(0.005)	0.198(0.005)
Y_3	CPS	0.135(0.004)	0.088(0.003)	0.087(0.003)
	BRFSS	0.133(0.004)	0.121(0.004)	0.120(0.004)
Y_4	CPS	4.986(0.099)	5.076(0.111)	5.086(0.098)
	BRFSS	4.931(0.100)	4.812(0.101)	4.807(0.099)

TABLE 10 Estimated population means by mass imputation and IPW

Response	**Sample *A***	$\widehat{\theta}_I$	$\widehat{\theta}_{IPW}$
Y_1	CPS	0.402(0.006)	0.396(0.006)
	BRFSS	0.447(0.006)	0.443(0.006)
Y_2	CPS	0.134(0.004)	0.136(0.004)
	BRFSS	0.198(0.005)	0.193(0.006)
Y_3	CPS	0.087(0.003)	0.088(0.003)
	BRFSS	0.120(0.004)	0.120(0.004)
Y_4	CPS	5.086(0.098)	5.059(0.099)
	BRFSS	4.807(0.099)	4.717(0.096)

4. computing the mass imputation estimator, the Hájek estimator and the IPW estimator of the population mean of Z.

Our starting point is Table 2 presented in Section 2. We chose the following three variables from the PRC dataset which are also available from one of the two probability survey samples, and treated them as ‘study variables’: Z_1 – ‘Smoke everyday’ (also available from BRFSS); Z_2 – ‘Smoke never’ (also available from BRFSS); Z_3 – ‘Volunteered’ (also available from CPS). All three variables are binary and the corresponding population means are population proportions. They all measure behaviours which might be related to other characteristics described by the auxiliary variables included in the datasets. The set of auxiliary variables $\mathbf{X}^*$ used for the modelling are those from Table 2 which are measured in both samples (the PRC sample and one of the probability samples). The variable ‘Age category’ is simplified to three groups: [18, 25], (25, 55] and (55, +∞). The conditional model $f(z\,|\,\mathbf{x}^*)$ is specified by a logistic regression. The model building process starts with the full set of available auxiliary variables but the final model only includes the significant auxiliary variables selected by the backward variable selection algorithm.

Table 11 presents the estimated population mean (with standard error in parentheses) for Z_1, Z_2 and Z_3 using four different methods: (a) Hájek: The Hájek estimator computed from the probability sample (BRFSS for Z_1 and Z_2, CPS for Z_3); (b) Naive: The simple sample mean from the non-probability sample; (c) $\widehat{\theta}_I$: The mass imputation estimator; and (d) $\widehat{\theta}_{IPW}$: The IPW estimator of Chen et al. (2020).

TABLE 11 Estimated population means for Z_1, Z_2 and Z_3 by using different approaches

Response	Sample *A*	Hájek	Naive	$\widehat{\theta}_I$	$\widehat{\theta}_{IPW}$
Z_1	BRFSS	0.115(0.001)	0.157	0.179(0.005)	0.173(0.005)
Z_2	BRFSS	0.833(0.001)	0.798	0.771(0.006)	0.776(0.007)
Z_3	CPS	0.248(0.002)	0.510	0.488(0.006)	0.493(0.007)

The design-based Hájek estimator is based on the observed values of the response variable in the probability sample BRFSS/CPS along with the survey weights and can be viewed as gold standard for the current comparisons.

We have three main observations from Table 11. First, the mass imputation estimator $\widehat{\theta}_I$ is very close to the IPW estimator $\widehat{\theta}_{IPW}$ for all three response variables. Second, the mass imputation estimator of Z_1 and Z_2 using BRFSS is closer to the naive estimator instead of the Hájek estimator, which seems to contradict the theory presented in Section 4. Third, the mass imputation estimator $\widehat{\theta}_I$ for Z_3 using CPS is closer to the Hájek estimator than the naive estimator.

It turns out that there is a major issue of mode effect in this analysis. The PRC non-probability sample was collected through web panels and hence is a web-based survey. The BRFSS was conducted through telephone interviews, and both landline household telephone numbers and cellular telephone numbers were used. There have been reported survey mode effect in the literature that measurements on behaviours such as smoking and drinking differ significantly between web surveys and telephone surveys. See, for instance, Chen et al. (2018) for findings from the International Tobacco Control (ITC) Canada survey. The Hájek estimator in the current analysis provides an estimate for the population means of Z_1 and Z_2 with measurements taken by telephone interviews. The mass imputation estimator $\widehat{\theta}_I$ is computed based on imputed values of the study variable but the imputation model was built using data from the non-probability sample for which the measurements were taken by the self-administered web survey. The value of $\widehat{\theta}_I$ is closer to the IPW estimator $\widehat{\theta}_{IPW}$, and both provide estimates for the population means with measurements taken by self-reported answers through the web. The mixed mode used by CPS, computer-assisted telephone interview (CATI) and computer-assisted personal interview (CAPI), seems also to have an impact on the results of the analysis on Z_3.

Note that results from the sensitivity analysis do not formally prove or disprove the validity of the model transportability assumption. For instance, the transportability of the model $f(x_2 \mid x_1)$ does not imply the transportability of the model $f(y \mid x_1, x_2)$. The interpretations of the results are further complicated by the mode effect as well as various non-sampling characteristics that differ in the two surveys. One might gain certain insides from such analysis but results are often inconclusive.

8 | CONCLUDING REMARKS

The use of non-probability survey samples as an efficient and cost-effective data source has become increasingly popular in recent years. Theoretical developments on analysis of non-probability samples, however, severely lag behind the need of making valid inference from such datasets. Non-probability survey samples are biased and do not represent the target population. Valid inferences require supplementary information on the population. The mass imputation approach proposed here relies on the availability of a probability survey sample from the same target population with information on covariates. The covariates are also measured for the non-probability sample and need to possess two crucial features for the framework discussed in this paper: (a) they characterize the inclusion/exclusion mechanism for units in the non-probability sample so that the ignorability or the transportability assumption could be justified; and (b) they are relevant to the response variable in terms of prediction power, a condition required for the mass imputation estimator.

Our analyses of the PRC non-probability sample shed light on two important practical issues. The first is the choice of an existing probability sample from the same target population if two or more such samples are available. The decision rests on the set of auxiliary variables which is available in both the non-probability and the probability samples. The size of the probability sample is less crucial as long as it is not too small. Two different probability samples tend to produce similar results if the same set of auxiliary variables is used. The second is the interpretation of the mass imputation estimator when the non-probability survey and the probability survey use different modes of data collection. The estimator is computed using the probability sample under mass imputation but the final estimate carries the mode from the non-probability sample since the model is built based on the observed study variable in the non-probability sample.

One of the main advantages of the mass imputation approach over the propensity score approach is that we can use subject matter knowledge on the study variables to build a good prediction model that can lead to better estimation. However, many non-sampling characteristics such as the mode effect that differ in the two surveys can make the required assumptions invalid, resulting in weakened estimates as shown in the sensitivity analysis. The choices between prediction-based approaches such as mass imputation and the propensity score-based approaches such as inverse probability weighting depend heavily on the quality of available auxiliary variables and require balanced considerations on the pros and cons of the approach.

The theoretical results on mass imputation presented in this paper focus on the estimation of finite population means. This is in line with traditional approaches in survey sampling where the basic theory is developed for the Horvitz–Thompson estimator. Inferences on other finite population parameters under mass imputation remain to be a research topic for future development. The main theoretical results are also based on deterministic imputation under the semiparametric model (4), and the model is imposed for a univariate response variable. One research topic is to explore the use of random imputation, including multiple imputation (Rubin, 1987) and fractional imputation (Kim, 2011), for analysing non-probability survey samples. Another research topic is on how to ensure consistency when the analysis involves multivariate response variables.

Statistics Canada has been implementing the modernization initiatives in recent years, which call for a culture switch from the traditional survey-centric approach by the agency to using data from multiple sources. One of the questions arising from the discussions is the role of traditional probability-based surveys, and there are even questions on the necessity of their existence in the future. Our theoretical results presented in this paper call for probability survey samples with rich information on auxiliary variables. A few large scale high-quality probability surveys representing the target population can play significant roles in analysing data from non-probability survey samples. While the use of non-probability samples and data from different sources has become unavoidable for statistical agencies, there needs to be a balanced mindset towards the bias-variance trade-off when combining information from multiple sources. In addition, statistical analysis of combined or fused data is itself an important research problem (Zhang & Chambers, 2019) and requires more comprehensive methodological development.

ACKNOWLEDGEMENTS

The authors are grateful to the two anonymous referees, the AE, and the Editor for their very constructive comments. The research of J.K. Kim was partially supported by a grant from US National Science

Foundation (MMS-1733572). The research of C. Wu was supported by grants from the Natural Science and Engineering Research Council of Canada and the Canadian Statistical Sciences Institute.

ORCID

Jae Kwang Kim https://orcid.org/0000-0002-0246-6029
Yilin Chen https://orcid.org/0000-0002-3510-2555

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 nonprobability sampling. *Journal of Survey Statistics and Methodology*, 1, 90–143.

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.

Brick, J. (2015) Compositional model inference. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 299–307.

Chen, M., Thompson, M.E. & Wu, C. (2018) Empirical likelihood methods for complex surveys with data missing-by-design. *Statistica Sinica*, 28, 2027–2048.

Chen, Y., Li, P. & Wu, C. (2020) Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*. In press.

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

Crump, R.K., Hotz, V.J., Imbens, G.W. & Mitnik, O.A. (2009) Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96, 187–199.

Dever, J.A. & Valliant, R. (2016) General regression estimation adjusted for undercoverage and estimated control totals. *Journal of Survey Statistics and Methodology*, 4, 289–318.

Elliott, M. & Valliant, R. (2017) Inference for nonprobability samples. *Statistical Science*, 32(2), 249–264.

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

Kim, J.K. (2011) Parametric fractional imputation for missing data analysis. *Biometrika*, 98, 119–132.

Kim, J.K. & Rao, J.N. (2012) Combining data from two independent surveys, a model-assisted approach. *Biometrika*, 99(1), 85–100.

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.

Park, S. & Kim, J.K. (2019) Mass imputation for two-phase sampling. *Journal of the Korean Statistical Society*, 48, 578–592.

Rao, J.N.K. & Wu, C.F.J. (1988) Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231–241.

Rivers, D. (2007) Sampling for web surveys. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 1–26.

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, 63(3), 581–592.

Rubin, D.B. (1987) *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Tourangeau, R., Conrad, F. & Couper, M. (2013) *The science of web surveys*, 1st edn. Oxford: Oxford University Press.

Tsiatis, A.A. (2006) *Semiparametric theory and missing data*. New York: Springer.

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.

Wu, C. & Thompson, M.E. (2020) *Sampling theory and practice*. Cham: Springer.

Yang, S. & Kim, J.K. (2020) Asymptotic theory and inference of predictive mean matching imputation using a superpopulation model framework. *Scandinavian Journal of Statistics*, 47, 839–861.

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

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Kim JK, Park S, Chen Y, Wu C. Combining non-probability and probability survey samples through mass imputation. *J R Stat Soc Series A*. 2021;184:941–963. https://doi.org/10.1111/rssa.12696