

RESEARCH PAPER

Advancing Finite Population Inference: Integrating Sample Surveys and Big Data

Moumita Baishya^{1*}, Ravi Ranjan Kumar², G. Avinash¹, Veershetty¹ and Harish Nayak G.H.¹

¹The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, India

²Guest Faculty, Dept. of Agricultural Statistics, College of Horticulture and Research Station, Saja, Bemetara, India

*Corresponding author: moumitabaishya1194@gmail.com (ORCID ID: 0000-0002-2077-5984)

Received: 14-03-2024

Revised: 23-05-2024

Accepted: 01-06-2024

ABSTRACT

Traditional sample surveys, emphasizing large sample sizes and robust finite population estimates through probability sampling designs, have been a staple in official statistics. However, modernization efforts are underway as statistical agencies explore the integration of big data and web panel information for near real-time estimations. This paper reviews the challenges associated with these endeavours, particularly addressing statistical biases linked to under-coverage in big data and errors in data variables. Kim and Tam (2020) introduced data integration methods, treating big data as an incomplete sampling frame and utilizing calibration weighting. The paper systematically reviews various integration methods, including mass imputation, propensity score, and calibration weighting. Additionally this paper concludes with a simulation study evaluating the performance of Kim and Tam's (2020) proposed estimators, assessing Bias, Standard Error (SE), and Root Mean Square Error (RMSE). The findings contribute to the on-going discourse on modernizing survey methodologies and leveraging diverse data sources for more efficient and timely official statistics.

HIGHLIGHTS

- The findings contribute to the on-going discourse on modernizing survey methodologies and leveraging diverse data sources for more efficient and timely official statistics.

Keywords: Non-Probability Sampling, Big Data, Calibration Weighting, Mass Imputation, Propensity Score, Bias, SE, RMSE

Over the course of time, sample surveys have been systematically conducted to procure dependable estimations of finite population descriptors, encompassing aggregates, averages, and percentiles. Probability-based sampling frameworks often referred to as the design-based methodology, have held a preeminent role, particularly in the domain of official statistics. The design-based paradigm found near-universal favour among practicing official statisticians. Additional techniques for combining two or more probability-based samples have also been devised to enhance estimator efficiency within defined budget constraints. In recent years, significant attention has been directed towards

resolving data collection challenges, primarily to curtail expenditures and uphold response rates through the introduction of innovative data gathering methods. Despite the persistent efforts deployed within the probability sampling framework, there has been a discernible decline in response rates and an escalation in costs (Williams and Brick, 2018). Concurrently, due to technological advancements, voluminous troves of affordable

How to cite this article: Baishya, M., Kumar, R.R., Avinash, G., Veershetty and Harish Nayak, G.H. (2024). Advancing Finite Population Inference: Integrating Sample Surveys and Big Data. *Econ. Aff.*, 69(02): 857-864.

Source of Support: None; **Conflict of Interest:** None



data, commonly known as big data (Administrative data, Transaction data, Social media data, Scrape data from websites, Sensor data and satellite images), and information from non-probability samples, particularly those emanating from online panels, have become readily accessible. Presently, official statistical bodies are actively spearheading modernization efforts, exploring innovative avenues to amalgamate data from diverse origins and generate trustworthy near real-time official statistics. Nonetheless, uninformed utilization of these datasets can result in profound sample selection biases. Without appropriate adjustments to mitigate these biases, the risk of falling into the big data paradox intensifies, as eloquently articulated by Meng (2018).

Probability sampling

Probability sampling stands as the gold standard in survey statistics, offering a benchmark approach for drawing conclusions about finite populations. Often referred to as designed data, these comprehensive surveys face challenges. Precise Estimations for Small Domains, Urgency for Real-time Data, Budget Constraints, Non-Response Bias Concerns, High costs and difficulties in securing participation and maintaining representativeness amid declining response rates are among them. Despite being design-based and closely aligned with the data collection process, large-scale survey programs encounter increasing demands for estimates in smaller population subgroups and the need for more timely data. Budget constraints exacerbate these issues, leading to reductions in sample sizes, while the declining willingness of respondents to participate raises concerns about non-response bias.

Non-probability sampling

In recent years, probability sampling theory has undergone significant advancements, yet struggles persist in addressing declining response rates and escalating survey costs. Developed countries are grappling with pronounced concerns over traditional survey methods, leading statistical agencies to explore innovative approaches. The allure of large, cost-effective non-probability samples from sources like web surveys and social media has prompted a shift towards modernization.

To counter challenges, agencies are now integrating data from diverse sources, including administrative records not originally intended for statistical use. While these records offer advantages such as timeliness and improved survey quality, concerns linger regarding potential under coverage and privacy issues (Citro, 2014; Holt, 2007; Kalton, 2019).

Big Data

Big data represents the extensive pool of structured and unstructured information emanating from diverse sources such as businesses, social media, and digital platforms. Extracting valuable insights from this vast data landscape is crucial for informed decision-making and strategic planning. While traditional statistical methods rely on well-established sources like censuses and surveys, contemporary statisticians are delving into non-probability samples, particularly big data, to compile official statistics. However, working with big data poses challenges such as under and over coverage, as well as self-selection. Addressing these issues involves employing models to reduce sample selection bias, and in cases where modelling falls short, combining non-probability samples with probability samples may be explored. The challenges intensify with the paradox of big data (Meng, 2018), as larger non-probability samples can lead to more misleading inferences. Overcoming these hurdles requires meticulous statistical adjustments and the integration of other data sources to ensure the validity of statistical inference in the era of big data. Tam & Clarke (2015) and Pfefffermann (2015) addressed methodological uses and challenges of big data in the production of official statistics.

MATERIALS AND METHODS

Data integration

In addressing non-response bias, weighting methods play a crucial role, relying on the missing-at-random (MAR) assumption within Rubin's framework, as elucidated by Valliant and Dever (2011) and Elliot and Valliant (2017). Despite its significance, this assumption lacks empirical verification. A promising alternative emerges through survey data integration, a rapidly evolving field explored by Yang and Kim (2020). This approach involves

merging information from distinct surveys targeting the same population, offering advantages such as cost reduction, reduced respondent burden, and enhanced statistical efficiency. Rao (2020) further expands on the diverse data sources facilitating integration, encompassing surveys based on probability samples, censuses, administrative records, and commercial transactional data.

Combining probability and non--probability samples

Case 1: Study variable observed in both probability sample A and non-probability sample B

Case 2: Study variable not observed in probability sample A

Data integration method for handling big data (Kim and Tam, 2020)

In their 2020 study, Kim and Tam proposed an innovative approach to seamlessly integrate big data with survey sample data. Recognizing potential deviations from missing at random (MAR) in the sampling mechanism for big data, they addressed systematic disparities between the two datasets, even after adjusting for auxiliary variables. Their methodology assumes the presence of survey variables in both datasets but allows for potential inaccuracies in measurements. Treating the big data sample as a finite population with incomplete or imprecise observations, they applied conventional techniques like calibration weighting, incorporating auxiliary information directly. The execution of calibration estimation within the survey data necessitates identifying the subset of the probability sample shared with the big data sample, a process reminiscent of dual frame estimation. In practical applications, the big data sample may exhibit coverage errors, while the survey sample remains error-free. To address imperfect matching scenarios, the researchers introduced an inventive classification method, utilizing observed matching variables from both sources to identify overlapping units. This novel classification procedure enables the derivation of fully nonparametric propensity scores, offering a means to rectify bias in the application of data integration estimators when dealing with inaccurate matching.

Basic Setup

Case 1: Study variable observed in both samples B and A

In the ideal case, the study variable y is observed within the non-probability sample B and independently within a probability sample A of size n , which is independently chosen from the target population. It is postulated that the units in sample A not included in sample B can be distinguished. Considering, a finite population U with a cardinality of N . Within this finite population, two distinct samples, denoted as A (probability sample) and B (big data sample obtained through an undisclosed selection mechanism) are taken. In both samples, the variable Y is measured. Initially, the assumption is that Y is measured in sample A without any measurement error. However, in sample B , Y may not be precisely measured. Thus, instead of observing y_i , observe y_i^* (contaminated version of y_i) from sample B . For more simplicity, we assume that

$$y_i^* = \beta_0 + \beta_1 y_i + e_i \quad \dots(1)$$

Where, (β_0, β_1) are unknown parameters and $e_i \sim (0, \sigma^2)$. Model (1) implies that y_i^* can be systematically different from y_i . In the special case of $(\beta_0, \beta_1) = (0, 1)$ there is no measurement bias in y_i^* . In addition, because the selection mechanism for the big data sample is unknown, it is subject to selection bias.

To address the selection bias and rectify measurement errors inherent in big data, Kim and Tam (2020) posited the availability of a gold standard survey sample. While acquiring survey sample data is typically a costly endeavour, leveraging the gold standard can significantly enhance the quality of the big data sample. To make sample A gold standard sample, a probability sampling design for selecting sample A is employed, and y_i are accurately observed from the sample.

From sample A , we can compute $\hat{T}_a = \sum_{i \in A} d_i y_i$, a design-unbiased estimator of $T = \sum_{i=1}^N y_i$, where $d_i = \pi_i^{-1}$ is the design weight of unit i and π_i is the first-order inclusion probability of unit i in sample A . Table 1 presents the data structure of our setup. Assuming that it is possible to identify elements in

sample A also belonging to sample B . That is, we can create δ_i for $i \in A$, where,

$$\delta = \begin{cases} 1 & \text{if } i \in B \\ 0 & \text{otherwise} \end{cases} \dots(2)$$

The objective is to combine the form both datasets in order to derive an enhanced estimator for T . Through judicious application of weighting techniques, we can achieve a superior estimator for T , denoted as \hat{T}_a , which completely ignores the information in the big data sample. Tam and Kim (2020) provided methods for adjusting such bias by using data integration.

Data Integration for Handling Selection Bias

1. Post-stratified data integration estimator

They first considered the simple case of no measurement errors in Y , that is, $y_i^* = y_i$.

Then, conceptually defined δ_i in (2) throughout the finite population. Thus, the set of elements with $\delta_i = 1$ is the big data sample.

According to Hartley's screening estimator of the total;

$$\hat{Y}_H = \sum_{i \in A} d_i(1 - \delta_i)y_i + \sum_{i \in B} y_i$$

$$T = \sum_{i=1}^N y_i = T_b + T_c$$

Where, $T_b = \sum_{i=1}^N \delta_i y_i$ and $T_c = \sum_{i=1}^N (1 - \delta_i)y_i$

Because T_b can be obtained from sample B , we only have to estimate T_c from sample A . Thus, $\hat{T}_{DI} = T_b + \sum_{i \in A} (1 - \delta_i)y_i$, a design-based estimator of T obtained from two samples.

If the population size N is known, a better estimator is

$$\hat{T}_{PDI} = T_b + (N - N_b) \frac{\sum_{i \in A} d_i(1 - \delta_i)y_i}{\sum_{i \in A} d_i(1 - \delta_i)}, \dots(3)$$

Where, $N_b = \sum_{i=1}^N \delta_i$ is the size of sample B . \hat{T}_{PDI} in (3) is a post-stratified estimator with the two post-strata defined by $\delta_i = 1$, and $\delta_i = 0$.

The design variance of \hat{T}_{PDI} in (3) is:

$$\text{Var}(\hat{T}_{PDI}) = (N - N_b)^2 \text{Var} \left\{ \frac{\sum_{i \in A} d_i(1 - \delta_i)y_i}{\sum_{i \in A} d_i(1 - \delta_i)} \right\} \approx \text{Var} \left\{ \sum_{i \in A} d_i(1 - \delta_i)(y_i - \bar{Y}_c) \right\}$$

Where, $\bar{Y}_c = \sum_{i=1}^N (1 - \delta_i)y_i / (N - N_b)$

Here, the approximate equality follows from Taylor linearization applied to the ratio component in (3). If the sampling design for sample A is simple random sampling of size n with $n/N \approx 0$, we have;

$$\text{Var}(\hat{T}_{PDI}) \approx (1 - W_b) \frac{N^2}{n} S_c^2 \dots(4)$$

Where, $W_b = \frac{N_b}{N}$ and

$$S_c^2 = (N - N_b)^{-1} \{ \sum_{i \in A} (1 - \delta_i)(y_i - \bar{Y}_c)^2 \}$$

Thus, the variance reduction of $\text{Var}(\hat{T}_{PDI})$ compared with $\hat{T}_a = \sum_{i \in A} d_i y_i$ is;

$$\frac{\text{Var}(\hat{T}_{PDI})}{\text{Var}(\hat{T}_a)} = (1 - W_b) \frac{S_c^2}{S^2}$$

If, $S_c^2 \approx S^2$ then data integration estimator is always more efficient than the design-based estimator using sample A only.

2. Regression data integration estimator

Kim and Tam (2020) additionally deliberated strategies for enhancing the efficacy of the data integration estimator. One such approach involves employing the concept of ratio estimation for T , by treating $x_i = \delta_i y_i$ as the auxiliary variable, which is observed throughout the finite population.

$$\hat{R} = \frac{\sum_{i=1}^N x_i}{\sum_{i \in A} d_i x_i}$$

Thus, it can be multiplied to direct estimator to reduce the variance, that is, to improve efficiency. The resulting ratio estimator is:

$$\hat{T}_{RatDI} = \hat{T}_a \hat{R} = T_b \frac{\hat{T}_a}{\hat{T}_b} \quad \dots(6)$$

Where, $\hat{T}_b = \sum_{i \in A} d_i \delta_i y_i$ and $\hat{T}_a = \sum_{i \in A} d_i y_i$

Thus, in (6) is called the ratio data integration estimator. \hat{T}_{RatDI} can be express as:

$$\hat{T}_{RatDI} = \sum_{i \in A} d_i \left(\frac{T_b}{\hat{T}_b} \right) y_i = \sum_{i \in A} w_i y_i$$

Where, w_i satisfies,

$$\sum_{i \in A} w_i x_i = \sum_{i \in A} d_i \left(\frac{T_b}{\hat{T}_b} \right) \delta_i y_i = \sum_{i=1}^N \delta_i y_i = \sum_{i=1}^N x_i \quad \dots(7)$$

Thus, equality (7) implies that the ratio data integration estimator satisfies the calibration property of the auxiliary variable in the sense that the estimator applied to x_i matches the known population total of x_i .

More generally, we can apply the calibration estimation method to $x_i = (1, \delta_i y_i)^T$, because $\sum_{i=1}^N (1, \delta_i y_i) = (N, N_b, T_b)$ is known.

Specifically, we can find $\{w_i : i \in A\}$ that minimises an objective function $Q(d, w)$ subject to the calibration equation $\sum_{i \in A} w_i x_i = \sum_{i=1}^N x_i$.

The regression estimator is based on objective function:

$$Q(d, w) = \sum_{i \in A} d_i \left(\frac{w_i}{d_i} - 1 \right)^2$$

The solution to the optimisation problem is:

$$w_i = d_i X_N^T \left(\sum_{i \in A} d_i x_i x_i^T \right)^{-1} x_i \quad \dots(8)$$

Where, $X_N = \sum_{i=1}^N x_i$

To understand the solution in (8), if we write $x_i = (1 - \delta_i, x_{ii}^T)^T$ with $x_{ii} = \delta_i (1, y_i)^T$, the regression weight in (8) reduces to:

$$w_i = \begin{cases} d_i X_1^T \hat{\Sigma}_{xx}^{-1} x_{ii} & \text{if } \delta_i = 1 \\ d_i (N_c / \hat{N}_c) & \text{if } \delta_i = 0, \end{cases} \quad \dots(9)$$

$$X_1 = \sum_{i=1}^N x_{1i}, \hat{\Sigma}_{xx11} = \sum_{i \in A} d_i x_{1i} x_{1i}^T, N_c = N - N_b,$$

and

$$\hat{N}_c = \sum_{i \in A} d_i (1 - \delta_i)$$

The weights in (9) satisfy $\sum_{i \in A} w_i (\delta_i, \delta_i y_i) =$

$$(N_b, T_b), \sum_{i \in A} w_i (1 - \delta_i) = N_c$$

The regression data integration estimator is then defined as;

$$\hat{T}_{RegDI} = \sum_{i \in A} w_i y_i \quad \dots(10)$$

Where w_i is defined in (9). Inserting (9) into (10), we can write

$$\hat{T}_{RegDI} = \sum_{i=1}^N \delta_i (1, y_i)^T \hat{\beta}_1 + N_c \frac{\hat{T}_c}{\hat{N}_c} \quad \dots(11)$$

where, $\hat{T}_c = \sum_{i \in A} d_i (1 - \delta_i) y_i$ and

$$\hat{\beta}_1 = \left\{ \sum_{i \in A} d_i \delta_i (1, y_i) (1, y_i)^T \right\}^{-1} \sum_{i \in A} d_i \delta_i (1, y_i)^T y_i$$

Therefore, the regression data integration estimator in (11) is algebraically equivalent to the post-stratified data integration estimator in (3).

A linearization variance estimator for (10)

$$\hat{V}(\hat{T}_{RegDI}) = \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\hat{e}_i \hat{e}_j}{\pi_i \pi_j} \quad \dots(12)$$

Where π_{ij} is the joint inclusion probability of unit i and j , $\hat{e}_i = y_i - x_i^T \hat{\beta}$ and $\hat{\beta} = \left(\sum_{i \in A} d_i x_i x_i^T \right)^{-1} \sum_{i \in A} d_i x_i y_i$

RESULTS AND DISCUSSION (Simulation study)

Firstly, continuous Y variable is considered from the following model

$$y_i = 3 + 0.7 (x_i - 2) + e_i,$$

where, $x_i \sim N(2, 1)$, $e_i \sim N(0, 0.51)$ and e_i is independent of x_i . We generate a finite population of size $N = 1000000$ from this model. Also, we generate;

Table 1: Results of the four estimators for simulation study based on a Monte Carlo sample of size 1000

Scenario	Estimator	Bias	SE	RMSE
I	Mean A	0.00	0.031	0.031
	Mean B	-0.11	0.001	0.113
	PDI	0.00	0.022	0.022
	RegDI	0.00	0.022	0.022
II	Mean A	0.00	0.031	0.031
	Mean B	-1.10	0.001	1.101
	PDI	-0.49	0.022	0.495
	RegDI	0.00	0.024	0.024
III	Mean A	-1.00	0.033	1.001
	Mean B	-0.11	0.001	0.113
	PDI	-0.51	0.023	0.507
	RegDI	0.00	0.028	0.028

Table 2: Results of the three estimators for simulation study

Sample Size	Estimators	Est Mean	Bias	RMSE	SE	% CV
250	Mean A_S1	2.9751	-0.0208	0.0570	0.0179	1.7865
	PDI_S1	2.9953	-0.0006	0.0272	0.0091	0.9068
	RegDI_S1	2.9953	-0.0006	0.0272	0.0091	0.9068
	PDI_S2	2.5039	-0.4920	0.4927	0.0108	1.0847
	PDI_S3	2.4852	-0.5107	0.5113	0.0103	1.0294
500	Mean A_S1	2.9758	-0.0201	0.0391	0.0113	1.1289
	PDI_S1	2.9956	-0.0003	0.0172	0.0057	0.5730
	RegDI_S1	2.9956	-0.0003	0.0172	0.0057	0.5730
	PDI_S2	2.5043	-0.4916	0.4919	0.0069	0.6855
	PDI_S3	2.4858	-0.5101	0.5103	0.0066	0.6585
750	Mean A_S1	2.9765	-0.0194	0.0322	0.0087	0.8667
	PDI_S1	2.9960	0.0001	0.0132	0.0044	0.4400
	RegDI_S1	2.9960	0.0001	0.0132	0.0044	0.4400
	PDI_S2	2.5046	-0.4913	0.4914	0.0053	0.5264
	PDI_S3	2.4859	-0.5100	0.5102	0.0050	0.5013

*PDI is the post stratified data integration estimator, RegDI is the regression data integration estimator, S1 = situation I, S2 = situation II and S3= situation III.

$$y_i^* = 2 + 0.9 (y_i - 3) + u_i,$$

Where, $u_i \sim N(0, 0.52)$ and u_i is independent of y_i

In this simulation, two samples, denoted by A and B obtained repeatedly, by simple random sampling of size $n = 1000$ and by an unequal probability sampling of size $N_b = 500000$, respectively.

Under this sampling mechanism, the sample mean of B is smaller than the population mean.

We consider the following three scenarios:

- Scenario I: No measurement errors in both samples. Thus, we observe y_i in both samples.
- Scenario II: Measurement errors in sample B. Thus, we observe y_i in sample A and y_i^* in sample B.

- Scenario III: Measurement errors in sample A. Thus, we observe y_i^* in sample A and y_i in sample B.

In addition, it is assumed that we observe the matching indicator δ_i in sample A. If $i i \delta_i = 1$ in sample A, we observe (y_i, y_i^*)

The following four estimators for the population mean of Y considered for comparison:

1. Mean A. Mean of sample A observations.
2. Mean B. Mean of sample B observations.
3. Post-stratified data integration estimator of the form (3).
4. Regression data integration estimator of the form (10).

In Scenario II, the post-stratified data integration estimator is computed using;

$$\hat{\theta}_{PDI} = \frac{1}{N} \left\{ \sum_{i=1}^N \delta_i y_i^* + (N - N_b) \frac{\sum_{i \in A} d_i (1 - \delta_i) y_i}{\sum_{i \in A} d_i (1 - \delta_i)} \right\}$$

In Scenario III, the post-stratified data integration estimator is computed using

$$\hat{\theta}_{PDI} = \frac{1}{N} \left\{ \sum_{i=1}^N \delta_i y_i + (N - N_b) \frac{\sum_{i \in A} d_i (1 - \delta_i) y_i^*}{\sum_{i \in A} d_i (1 - \delta_i)} \right\}$$

Table 2 displays the outcomes of a simulation study conducted with 1000 Monte Carlo samples. The mean *A* estimator is unbiased across most scenarios, except for scenario III, wherein systematic measurement errors affect sample *A*. The mean *B* estimator consistently exhibits bias due to selection bias in sample *B*, with the most substantial bias observed in scenario II, where measurement errors compound the issue. The variance of the mean *B* estimator is notably reduced owing to the considerable sample size in sample *B* ($N_b = 500000$). Furthermore, the post-stratified estimator's variance is approximately half that of the mean *A* estimator due to $N_b/N = 0.5$, and a larger $W_b = N_b/N$ would result in even lower variance than the variance estimator post-stratified estimator, as implied by equation (4). Nevertheless, in scenario II, the post-stratified data integration estimator becomes biased because T_b is estimated without correcting for the measurement errors. In scenario III, it is biased because T_c is estimated from sample *A* without correcting for the measurement errors. The regression data integration estimator is unbiased for all scenarios. It is the same as the post-stratified data integration estimator under scenario I, as discussed in (11).

Additionally, we computed variance estimators for the regression data integration estimator.

CONCLUSION

Data integration stands as a burgeoning area of study, explored through statistical techniques in survey sampling. While probability sampling remains the gold standard, the measurement of the study variable may stem from non-probability samples or big data, necessitating assumptions about sampling

and outcome models. Most methods assume non-informative sampling mechanisms, posing challenges in verification. In cases of informative sampling, imputation techniques (Riddles *et al.* 2016; Morikawa and Kim, 2018) demand stringent model assumptions, prompting sensitivity analysis to assess study robustness to unverifiable assumptions. Navigating data integration complexities mandates careful consideration of assumptions, especially in informative sampling scenarios.

In conclusion, the proposed data integration techniques (Kim and Tam, 2020) offer a promising avenue for addressing the inherent challenges associated with big data sampling. Leveraging an independent probability sample mitigates under-coverage biases. Calibration weighting and treating big data as an incomplete sampling frame help tackle measurement errors. In practical application, these methods prove particularly valuable under certain conditions. Firstly, the presence of a probability sample *A*, capable of measuring *y* or providing a suitable proxy y^* , is essential. While the existence of such a sample is rare, organizations like national statistical offices may deem it worthwhile to strategically design, develop, and implement such a random sample to capture the desired measure. In such cases, the population count of sample units, *N*, is inherently known. Secondly, the efficacy of the calibration method hinges on the substantial coverage of *B*, which is a reasonable assumption when dealing with sizable big data sets. Moreover, for measurement error adjustments, it is safe to assume that $A \cap B$ is not devoid of data, particularly in cases where such adjustments are warranted. The outcomes of our simulation study, based on 1000 Monte Carlo samples conducted by Kim and Tam (2021), underscore the practical value of these methods.

REFERENCES

Brodie, M.A., Pliner, E.M., Li, K., Chen, Z., Gandevia, S.C. and Lord, S.R. 2018. Big data vs accurate data in health research: large-scale physical activity monitoring, smartphones, wearable devices and risk of unconscious bias. *Med. Hypotheses*, **119**: 32-36.

Deville, J.-C. and Särndal, C.E. 1992. Calibration estimators in survey sampling. *J. Am. Stat. Assoc.*, **87**: 376-382.

Elliott, M. and Valliant, R. 2017. Inference for non-probability samples. *Statistical Science*, **32**: 249-264.

- Groves, R. and Peytcheva, E. 2008. The impact of non-response rates on non-response bias: a meta- analysis. *Public Opin. Q.*, **72**: 167-189.
- Hartley, H.O. 1962. Multiple frame surveys. In *Proceedings of the Social statistics Section*. American Statistical Association, pp. 1-24.
- Kaplan, R.M., Chambers, D.A. and Glasgow, R.E. 2014. Big data and large sample size: a cautionary note on the potential of bias. *Am. Soc. for Clin. Pharmac. and Therap.*, **7**: 342-346.
- Kim, J.K. and Tam, S.M. 2021. Data integration by combining big data and survey sample data for finite population inference. *Int. Int. Statistical Rev.*, **89**(2): 382-401.
- Lohr, S. and Raghunathan, T. 2017. Combining survey data with other data sources. *Stat. Sci.*, **32**: 293-312.
- Meng, X.L. 2018. Statistical paradises and paradoxes in big data(i): law of large populations, big data paradox, and 2016 US Presidential Election. *Ann. Appl. Stat.*, **12**: 685-726.
- Pfeffermann, D. 2015. Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture. *J. Surv. Stat. Methodol.*, **3**: 425-483.
- Rao, J.N.K. 2021. On making valid inferences by integrating data from surveys and other sources. *Sankhya B.*, **83**(1): 242-272.
- Tam, S.M. and Clarke, F. 2015. Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *Int. Stat. Rev.*, **83**: 436-448.
- Tam, S.M. and Kim, J.K. 2018. Big data, selection bias and ethics — an official statistician's perspective. *Stat. J. IAOS*, **34**: 577-588.
- Valliant, R. and Dever, J.A. 2011. Estimating propensity adjustments for volunteer web surveys. *Sociol. Methods Res.*, **40**: 105-137.
- Yang, S. and Kim, J.K. 2020. Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, **3**(2): 625-650.