Statistical Journal of the IAOS 38 (2022) 647–659
DOI 10.3233/SJI-210924
IOS Press

Error analysis for hybrid estimates of proportions using big data[1]

Siu-Ming Tam[a,*], Dennis Trewin[b] and Lyndon Ang[c]
[a]*National Institute of Applied Statistical Research, University of Wollongong, Northfields Avenue, Wollongong, NSW, Australia*
[b]*Blairgowrie, VIC, Australia*
[c]*Australian Bureau of Statistics, Belconnen, ACT, Australia*

Abstract. Big data, including administrative data, is seen as a new data source for official statistics especially given the increasing difficulty of getting acceptable response rates in sample surveys. It might be used directly or perhaps with the use of models to adjust for shortcomings in the big data. Hybrid estimates using complementary survey data are another technique for overcoming these shortcomings. To make decisions on how big data might be used, we need to understand the nature of the errors in the big data source. The paper describes an Error Framework for the analysis of errors in big data and hybrid estimates. The paper also describes the circumstances under which hybrid estimates will provide more accurate estimates than big data in isolation or survey data. A case study is provided to illustrate the application of hybrid estimates in practice. A potential application of hybrid estimation is also described to address the upward biases that often exist in epidemiological modelling.

Keywords: Big data, hybrid estimates, error framework, epidemiological modelling

1. Introduction

High quality traditional probability surveys are becoming more difficult and expensive to conduct. Non-response is a growing problem especially among some segments of the population and therefore there is an increasing risk of bias in estimates derived from sample surveys. At the same time, there is a plethora of data from other sources as a result of the digital revolution. These data can have their own quality problems e.g. important parts of the population may not be covered, the data generation process is unknown and self-selection biases often exist, or the concepts measured do not align with the target measures. However, they do have advantages over probability samples such as a greater number of observations, lower data collection costs and no additional reporting burden. These types of data are often referred to as big data. For the purposes of this paper, we are including administrative data as one example of big data.

How can big data be used in official statistics? There are four possibilities as described below and Option 4 is the subject of this paper.

1. Directly unmodified to produce statistics (more common for the use of administrative data).
2. Directly unmodified to update a benchmark for a base period (e.g. derived from a sample survey) where big data is available on a more regular basis and can be used to provide updated estimates of the benchmark.
3. Directly but transformed through the use of models but still likely to need other data (e.g. derived from a probability sample) or strong assumptions to support the models.

[1]The views expressed in this paper are those of the authors and do not necessarily represent the views of the Australian Bureau of Statistics.

*Corresponding author: Siu-Ming Tam, National Institute of Applied Statistical Research, University of Wollongong, Northfields Avenue, Wollongong, NSW 2522, Australia. E-mail: stattam@gmail.com.

4. Combined data sources (integrated data sources such as hybrid estimates).

Probability surveys and big data have different strengths and weaknesses. As they are different, it does raise the question of whether they can be used in combination to build on their respective strengths and to compensate for their different weaknesses. Estimates combining both data sources are referred to as hybrid estimates.

The layout of this paper is as follows: Section 2 provides an Error Frameworks for Big Data extended to cover Hybrid Estimates. Sections 3 to 6 provide a mathematical exposition of hybrid estimates for finite population proportions, discuss when they are more accurate than either the big data themselves or survey data, and consider the case where there are measurement errors in the big data. Section 7 provides a Case Study. Section 8 discusses a possible epidemiological application with concluding remarks outlined in Section 9.

2. Total error framework for big data and hybrid estimates

For this paper, we define hybrid data as survey and big data sets that complement each other and are combined in some way to provide more robust estimates. Estimates derived in the way outlined later in this paper from hybrid data are called hybrid estimates. For example, a sample survey could be used to complement a big data source to address its weaknesses e.g. under-coverage, self-selection errors of certain populations or providing the data to enable the adjustments for any validity or measurement errors in the big data. Of course, one can have more complex hybrid estimates that combine more than two data sets.

The starting point for the consideration of hybrid estimates is understanding the sources of error in both the probability survey and the big data source. There is a lot of literature on Total Survey Error (TSE) that covers the former, e.g. [1]. A common representation of TSE is to categorise Errors of Measurement as specification error, measurement error, processing error or model/estimation error and Errors of Representation as frame error, sampling error or non-response error. Total Error is a similar concept to Total Survey Error but recognises that the term 'survey' is misleading in the case of big data. Furthermore, the sources of error will be different, requiring a modified framework for specifying the errors.

There have been some recent approaches to describe Total Error for Big Data [2]. We have chosen to adapt the work of [3]. Their work was targeted at administrative data but can be easily extended to big data. An important initiative in their work is to describe errors separately for each of three phases of the production process – (1) single input data, (2) integrated data sources, and (3) output. We will use the same structure except that the integration aspects of phase 2 are considered as part of input phase and the output phase comprises the hybrid estimates discussed in this paper. Only high-level detail is shown in Table 1. Also, the explanations are most relevant to applications of hybrid estimates for official statistics.

The framework focusses on the residual errors after the hybrid estimates have been created by combining the survey and big data sources. The residual errors will contain both systematic (bias) and random components. It is important to understand both. It is almost certain that some random components will still exist even when there have been adjustments for systematic bias.

Having defined the relevant frameworks, how do we use them to make decisions on the choice of survey, big data or hybrid data source for estimation? The first step is to understand the errors in the data sources irrespective of whether they are survey or big data based. Wherever possible, this should be based on quantitative information, but we appreciate this is not always possible. Informed guesses may be necessary. Independent and expert third party advice can also be especially useful. The aim here is to identify the most important sources of error, not every source of error, and the extent to which they can be mitigated. These are the errors that need to be considered in determining whether the big data can be used or not, the design of any complementary data collections, and the development of the hybrid estimates.

The focus should then be on understanding how these errors might be measured, controlled or mitigated. The framework described in Table 1 is relevant to the error analysis of the hybrid estimates. If it is decided to consider hybrid estimates, the next step is to design the hybrid estimates themselves. This will include consideration of important residual errors and how they might be further mitigated. It may be necessary to commission some special studies to support this analysis. For adjustment of coverage errors, as seen below, representivity ratios of both the big data and the survey data are very important considerations. For the adjustment of stochastic measurement errors, statistical models will also be necessary.

Table 1
Total error framework for big data and hybrid estimates

	Explanation of errors in big data	Explanation of residual errors in hybrid estimates
Errors of measurement – type of error		
Validity error	Where the available measures differ from the target concept. Similar to specification error in TSE.	Where there is any residual error (systematic or random) after adjustments are made to the big data measures.
Measurement error	When the actual measures contain errors e.g. poor form design when collecting the big data.	As above plus any measurement errors in the survey source.
Processing error	Where errors are made in processing e.g. Extract, Transform and Load (ETL) activities, editing, and coding.	ETL, editing and coding errors in the survey and big data sources.
Modelling error	Systematic and random errors in modelling used to produce outputs from the big data inputs.	Residual systematic and random errors after any modelling done to produce hybrid estimates.
Linkage error	Linking errors where multiple data sources are used to create the big data set.	Residual impact of errors on hybrid estimates that are caused in the process of linking common units, and errors in linking survey data with big data
Errors of representation – type of error		
Frame coverage or selection errors	Includes under-coverage (e.g. because of self-selection bias), over-coverage and duplication of units in the big data.	Net under-coverage, over-coverage and duplication of units used in hybrid estimates e.g. units not covered by either source.
Sampling error	Not applicable	Due to the use of a sample to support hybrid estimates
Missing data/non-response error	Includes both unit and item non-responses.	Residual impacts from missing data and survey non-response that are not adjusted through the hybrid estimation process.

3. How to determine when to use hybrid estimates?

If one has a probability sample, A, with partial response, and a big data set B with under-coverage/self-selection errors and, possibly measurement errors as well, which estimate out of the three, from the partially responding sample, big data set or a combination of the two, would be preferred?

[4] showed that, for simple random sampling with negligible finite population correction and full response, a hybrid estimator of the finite population total is preferred as it generally has smaller mean squared error (MSE) than either the big data estimator or the survey estimator. The hybrid estimator will have a smaller variance than that of the estimator from the probability sample, A, if $\left(1-\frac{N_B}{N}\right) S_C^2 < S_U^2$ where N_B, N, S_U and S_C are the size of the big data set, size of the finite population, and the standard deviation of the population, and of the population segment missed by the big data set, C, respectively. The inequality is almost always true when N_B is large in comparison with N. The methodology can also be examined from a dual sampling frame perspective [5].

In this paper, we extend the idea of [4] in the following ways. First, we extend the comparison of the efficiency (in MSE terms) of the hybrid estimator with the Horvitz-Thompson estimator of the finite population proportion from a responding probability sample adjusted for non-response errors, and the big data estimator of the proportion adjusted for under-coverage/self-selection errors. Second, the adjustment is carried out using "representivity ratios". Third, when the big data is subject to measurement errors, instead of using the erroneous big data as benchmarks as in [4] for the hybrid estimate, we use a statistical method to adjust for the measurement errors before we combine the big data with survey data.

To illustrate ideas, let A_R and B denote the responding sample and big data respectively. We consider the simple case of estimating the finite population proportion, $P_U = \frac{\sum_{i\in U} Y_i}{N}$, where $Y_i = 1$ or 0. We also assume initially that there are no measurement errors in the big data, and then extend our results to the measurement error case.

For this set up, we are interested in the choice between estimators from the big data sample, B, $P_B = \frac{\sum_{i\in B} Y_i}{N_B}$, the partially responding probability sample, A_R, $\hat{P}_{A_R} = \frac{\sum_{i\in A_R} Y_i}{n_{A_R}}$, or a hybrid estimate based on B and A_R.

We use a simple example to illustrate the concept of "representivity ratios". Suppose the population comprises 5 million males and 5 million females, and the big data set consists of 5 million males and 4 million females. The population proportion of males is 0.5 but the corresponding proportion in the big data set is 0.556, giving an estimation

error of 0.056. Why does this happen? It happens because whilst the propensity of males included in the big data set is 100%, the inclusion propensity for females is only 80%. If we let r_B denote the representivity ratio of B, i.e. the ratio of the propensity of $Y_i's = 1$ to the propensity of $Y_i's = 0$ to be included in B, it gives us an indication of how balanced the representation of the population of $Y_i's = 1$ and $Y_i's = 0$ is in B. Where this ratio is not equal to 1, there is an imbalance and the proportion computed from B would be biased. In addition, [6] showed that P_U can be recovered from the expression $P_U = (r_B P_B^{-1} - r_B + 1)^{-1}$. In the hypothetical example above, $r_B = 1.25$ and this gives us an indication that males are over-represented in the big data set. Using this information and putting $P_B = 0.556$ into the "recovery" formula gives us back $P_U = 0.5$. Similarly, if we let r_R denote the representivity ratio for A_R, we can also use $P_U = (r_R P_{A_R}^{-1} - r_R + 1)^{-1}$ to recover the finite population proportion based on the partially responding sample.

From $P_U = (r_B P_B^{-1} - r_B + 1)^{-1}$, we have $P_B = r_B P_U\{1+(r_B-1)P_U\}^{-1}$, hence $MSE(P_B) = (P_B - P_U)^2 = \left\{\frac{(r_B-1)P_U(1-P_U)}{1+(r_B-1)P_U}\right\}^2$ which is just the bias squared. Likewise, the bias squared for $\hat{p}_{A_R}$ is $\left\{\frac{(r_R-1)P_U(1-P_U)}{1+(r_R-1)P_U}\right\}^2$ and its sampling variance, ignoring finite population corrections, is $\frac{\hat{p}_{A_R}(1-\hat{p}_{A_R})}{n_{A_R}}$ noting $\hat{p}_{A_R} = r_R P_U\{1+(r_R-1)P_U\}^{-1}$. Hence, $MSE(\hat{p}_{A_R}) \doteq \frac{1}{n_{A_R}} \frac{r_R P_U(1-P_U)}{\{1+(r_R-1)P_U\}^2} + \left\{\frac{(r_R-1)P_U(1-P_U)}{1+(r_R-1)P_U}\right\}^2$ comprising a sampling variance term and a bias squared term.

To construct the hybrid estimator, we can combine the proportion from B, with the estimated proportion for C, through the information from $A_R \cap C$. This assumes that there is sufficient information available for the analyst to identify the units in the responding sample that are not in the big data set, using methods of deterministic or probabilistic matching. Let $\hat{p}_{A_R \cap C}$ be the computed proportion of $Y_i = 1$ in $A_R \cap C$. Then the hybrid estimator of P_U is given by $\hat{p}_H = W_B P_B + W_C \hat{p}_{A_R \cap C}$, where $W_B = \frac{N_B}{N}$ and $W_C = 1 - W_B$. We know that $\hat{p}_{\tilde{H}} = W_B P_B + W_C \hat{p}_{A \cap C}$ is approximately unbiased because $\hat{p}_{A \cap C}$ is an approximately unbiased estimator of the proportion of $Y = 1$ in C. From $\hat{p}_{A \cap C} = R\hat{p}_{A_R \cap C} + (1-R)\hat{p}_{A_{NR} \cap C}$, where $\hat{p}_{A_{NR} \cap C}$ is the unobserved proportion from the non-responding sample in C, we therefore have $|\hat{p}_H - \hat{p}_{\tilde{H}}| = W_C(1-R)|\hat{p}_{A_R \cap C} - \hat{p}_{A_{NR} \cap C}| \leqslant W_C(1-R)$ where R is the response rate, and hence $\hat{p}_H$is almost unbiased to $O\{W_C(1-R)\}$. Therefore, the *MSE* of $\hat{p}_H$ is bounded by: $MSE(\hat{p}_H) \leqslant \frac{W_C^2}{n_{A_R \cap C}} \frac{r_R P_C(1-P_C)}{\{1+(r_R-1)P_C\}^2} + W_C^2(1-R)^2$, where P_C denotes the proportion of $Y_i = 1$ in C. Using arguments similar to deriving $MSE(\hat{p}_{A_R})$ above, we can show that $MSE(\hat{p}_H) \doteq \frac{W_C^2}{n_{A_R \cap C}} \frac{r_R P_C(1-P_C)}{\{1+(r_R-1)P_C\}^2} + \frac{W_C^2(r_R-1)^2 P_C^2(1-P_C)^2}{\{1+(r_R-1)P_C\}^2}$. Noting that from $P_C = \frac{1}{W_C}(P_U - W_B P_B)$ and $P_B = r_B P_U\{1+(r_B-1)P_U\}^{-1}$, we can also express $MSE(\hat{p}_H)$ in terms of P_U.

4. Which estimator is better under what conditions?

The choice of estimators is determined by comparing their MSEs. The following provides sufficient conditions for one estimator to be better than another. It is easily seen that $MSE(P_B) \leqslant MSE(\hat{p}_{A_R})$ provided that $\frac{|(r_B-1)|}{1+(r_B-1)P_U} \leqslant \frac{|(r_R-1)|}{1+(r_R-1)P_U}$, i.e. the absolute bias of P_B is smaller than that of $\hat{p}_{A_R}$ as the latter also has a sampling variance in its MSE. Otherwise, we have to assess the sign of: $MSE(P_B) - MSE(\hat{p}_{A_R}) = \left[\frac{(r_B-1)^2 P_U^2(1-P_U)^2}{\{1+(r_B-1)P_U\}^2} - \frac{(r_R-1)^2 P_U^2(1-P_U)^2}{\{1+(r_R-1)P_U\}^2}\right] - \frac{r_R P_U(1-P_U)}{n_{A_R}\{1+(r_R-1)P_U\}^2}$ which is data dependent and can only resolved numerically.

Provided that $W_C = (1-W_B)$ is sufficiently small such that $W_C^2 \approx 0$, we have $MSE(\hat{p}_H) \approx 0 \leqslant MSE(P_B)$ or $MSE(\hat{p}_{A_R})$ and the hybrid estimates is always preferred. Otherwise, for W_B not close to one, the choice will have to be determined numerically by comparing their MSEs.

5. How to determine the representivity ratios?

Let $\theta_A = \frac{\sum_{i \in A}(1-Y_i)}{\sum_{i \in A} Y_i}, \theta_B = \frac{\sum_{i \in B}(1-Y_i)}{\sum_{i \in B} Y_i}$ and $\theta_{A_R} = \frac{\sum_{i \in A_R}(1-Y_i)}{\sum_{i \in A_R} Y_i}$, then it can be shown using the Bayes Theorem ([6]) that: $\hat{r}_B = \frac{\theta_A}{\theta_B}$ and $\hat{r}_R = \frac{\theta_A}{\theta_{A_R}}$. As both θ_B and θ_{A_R} are observed, we only need θ_A which can be estimated by

well-established non-response adjustments methodology used in statistical offices [7,8], e.g. $\hat{\theta}_A = \frac{\sum_{i\in A_R}(1-Y_i)/\hat{\rho}_i}{\sum_{i\in A_R} Y_i//\hat{\rho}_i}$, where ρ_i is the propensity of the i^{th} sampling unit to respond to the survey, with ρ_i to be estimated using e.g. a logistic regression model. For the application described in Section 7, we used the two-step adjustment of [9] to determine the non-response adjustment weights. This is to ensure that the initial weight $\frac{1}{\hat{\rho}_i}$ is further adjusted to make the weighted auxiliary variables in the sample equal to their benchmarks, subject to the constraint that the weighted (by $\hat{\rho}_i$) sum of the squared differences between the initial and final weights is a minimum. Note that we can use the plug-ins $\hat{P}_U = (1+\hat{\theta}_A)^{-1}$ and $\hat{P}_C = \frac{1}{W_C}(\hat{P}_U - W_B P_B)$ to estimate the MSEs of the different proportion estimators.

6. What if the big data is subject to measurement errors?

We model the random variable, Z, to be Y observed with errors, such that the false negative rate, i.e. observing a "0" when the actual value should be "1", $\Pr(Z=0|Y=1)$, is α_0; and the false positive rate, $\Pr(Z=1|Y=0)$, is α_1. We assume that random variables are mutually statistically independent, i.e., $Z_i \perp Z_j, Y_i \perp Y_j, Y_i \perp Z_j$ where $i \neq j$. Let $\bar{Z}_B = \frac{\sum_{i=1}^{N_B} Z_i}{N_B}$ be used to estimate P_U. Because there are false positive and false negative values in Z, there will be increased uncertainty in using $\bar{Z}_B$ to estimate P_U.

Heuristically, the number of false negatives in the observed counts would be $\alpha_0 N_B P_B$ and the number of false positives would be $\alpha_1 N_B(1-P_B)$. Hence the average net error in counting the number of $Y=1$ cases in B would be $\frac{1}{N_B}|\alpha_1 N_B(1-P_B) - \alpha_0 N_B P_B| = |\alpha_1(1-P_B) - \alpha_0 P_B|$.

It can be shown (refer to the Appendix) that: $\mathit{MSE}(\bar{Z}_B) = E(\bar{Z}_B - P_U)^2 \doteq \{\alpha_1(1-P_B) - \alpha_0 P_B\}^2 + \mathit{MSE}(P_B)$. This result shows that the additional term added to the *MSE* is the square of the net counting error, and the approximation is correct to $O(N_B^{-1})$.

Alternatively, a better approach (see Section 7 below) will be to adjust the big data for measurement errors first before constructing the hybrid estimates, as follows: $\hat{p}_{H_{adj}} = W_B\{\hat{P}_B\} + W_C \hat{p}_{A_R \cap C}$, where $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are consistent estimators of α_0 and α_1 respectively. To compute $\hat{P}_B$, we use the following iterative approach:

Step 1. Compute $\hat{P}_B^{(1)} = \bar{Z}_B + \hat{\alpha}_0 \bar{Z}_B - \hat{\alpha}_1(1-\bar{Z}_B)$
Step 2. Compute $\hat{P}_B^{(i+1)} = \bar{Z}_B + \hat{\alpha}_0 \hat{P}_B^{(i)} - \hat{\alpha}_1(1-\hat{P}_B^{(i)})$ for $i \geqslant 2$ until $\hat{P}_B^{(i+1)}$ converges.

We call $\hat{p}_{H_{adj}}$ the measurement bias-adjusted hybrid estimator. Using equation A1 of Appendix, it can be seen that:

$$
\begin{aligned}
E(\hat{P}_B) &= E\left\{ \frac{\sum_{i=1}^{N_B} Z_i}{N_B} + \hat{\alpha}_0 P_B - \hat{\alpha}_1(1-P_B) \right\} \\
&\doteq \frac{\sum_{i=1}^{N_B} E(Z_i)}{N_B} + \alpha_0 P_B - \alpha_1(1-P_B) \\
&= P_B.
\end{aligned}
$$

Hence

$$
\begin{aligned}
\mathit{MSE}(\hat{P}_B) &= E(\hat{P}_B - P_U)^2 \\
&= E\{\hat{P}_B - E(\hat{P}_B) + E(\hat{P}_B) - P_U\}^2 \\
&\doteq E(\hat{P}_B - P_B)^2 + E(P_B - P_U)^2 \\
&= E\{\bar{Z}_B - E(\bar{Z}_B)\}^2 + E\{(\hat{\alpha}_0 - \alpha_0)P_B - (\hat{\alpha}_1 - \alpha_1)(1-P_B)\}^2 + \mathit{MSE}(P_B) \\
&= \mathit{Var}(\hat{\alpha}_1)(1-P_B)^2 + \mathit{Var}(\hat{\alpha}_0)P_B^2 + \mathit{MSE}(P_B) + O(N_B^{-1})
\end{aligned}
$$

$$= Var(\hat{\alpha}_1)(1-P_B)^2 + Var(\hat{\alpha}_0)P_B^2 + \left\{\frac{(r_B-1)P_U(1-P_U)}{1+(r_B-1)P_U}\right\}^2 + O(N_B^{-1}).$$

Table 2
2x2 of observed vs actual value

		Actual value	
		$Y=1$	$Y=0$
Observed value in B	$Z=1$	n_{11}	n_{10}
	$Z=0$	n_{01}	n_{00}

For simple random sampling and ignoring the finite population correction, we have

$$\begin{aligned} MSE(\hat{P}_B) \doteq\ & \frac{1}{n_{1+}^2}\left\{Var(n_{10}) + \alpha_1^2 Var(n_{1+}) - 2\alpha_1 Cov(n_{10}, n_{1+})\right\}(1-P_B)^2 \\ & + \frac{1}{n_{+1}^2}\left\{Var(n_{01}) + \alpha_0^2 Var(n_{+1}) - 2\alpha_0 Cov(n_{01}, n_{+1})\right\}P_B^2 \\ & + \left\{\frac{(r_B-1)P_U(1-P_U)}{1+(r_B-1)P_U}\right\}^2 + O(N_B^{-1}). \end{aligned}$$

Hence,

$$\begin{aligned} MSE(\hat{p}_{H_{adj}}) &= Var(\hat{p}_{H_{adj}}) + E(\hat{p}_{H_{adj}} - P_U)^2 = W_B^2\{Var(\hat{\alpha}_1)(1-P_B)^2 + Var(\hat{\alpha}_0)P_B^2\} + W_C^2 MSE\{\hat{p}_{A_R\cap C}\} \\ &= W_B^2\{Var(\hat{\alpha}_1)(1-P_B)^2 + Var(\hat{\alpha}_0)P_B^2\} \\ &\quad + \frac{W_C^2}{n_{A_R\cap C}}\frac{r_R P_C(1-P_C)}{\{1+(r_R-1)P_C\}^2} + \frac{W_C^2(r_R-1)^2P_C^2(1-P_C)^2}{\{1+(r_R-1)P_C\}^2} + O(N_B^{-1}). \end{aligned}$$

Furthermore, to estimate α_o and α_1, one can take a random sample of B, compare the actual values with the observed values and use the information to tally the "confusion" matrix as in Table 2.

Then $\hat{\alpha}_0 = \frac{n_{01}}{n_{11}+n_{01}}$ and $\hat{\alpha}_1 = \frac{n_{10}}{n_{10}+n_{00}}$. In addition, $E(\hat{\alpha}_0) \doteq \alpha_0$, and $E(\hat{\alpha}_1) \doteq \alpha_1$. As a special case, one can use the linked A_R and B sample used to help compute the hybrid estimate to estimate $\hat{\alpha}_0$ and $\hat{\alpha}_1$. This is an efficient approach provided that the cost of verifying the observed value against the actual value for the linked sample is less than total cost of selecting a smaller sample and verification. We used the linked sample to compute $\hat{\alpha}_0$ and $\hat{\alpha}_1$ in the case study below.

Finally, as $\theta_B = \frac{\sum_{i\in B}(1-Y_i)}{\sum_{i\in B} Y_i}$ is not observed, we can estimate it by the following "plug-in" estimator: $\frac{N_B - \sum_{i=1}^{N_B} Z_i - N_B\{\hat{\alpha}_0 P_B - \hat{\alpha}_1(1-P_B)\}}{\sum_{i=1}^{N_B} Z_i + N_B\{\hat{\alpha}_0 P_B - \hat{\alpha}_1(1-P_B)\}}$ because, from (A1) of the appendix, we have

$$\begin{aligned} & E\left\{\sum_{i=1}^{N_B} Z_i + N_B\{\hat{\alpha}_0 P_B - \hat{\alpha}_1(1-P_B)\}\right\} \doteq \sum_{i=1}^{N_B} E(Z_i) + N_B\{\alpha_0 P_B - \alpha_1(1-P_B)\} \\ &= N_B\{(1-\alpha_0)P_B + \alpha_1(1-P_B)\} + N_B\{\alpha_0 P_B - \alpha_1(1-P_B)\} \\ &= N_B P_B \\ &= \sum_{i=1}^{N_B} Y_i. \end{aligned}$$

Table 3
Values of α_0 and α_1 for the with-measurement-error dataset

	Data item			
	Beef	Sheep	Lettuce	Wheat
α_0	0.1296	0.1299	0.1780	0.2391
α_1	0.1186	0.0858	0.0010	0.0613

7. Case study

We tested the methods outlined in the previous section through an empirical study, using data from the Australian Bureau of Statistics (ABS) 2015/2016 Agriculture Census (Ag Census) and 2014/2015 Rural Environment and Agricultural Commodities Survey (REACS). The set of responding businesses on the 2015/2016 Ag Census is considered as the big data B while the 2014/2015 REACS is our probability sample A. For the purposes of the empirical study, the agriculture population in 2014/2015 was deemed to be the population of interest.

There were a number of changes in the agriculture population between 2014/2015 and 2015/2016. These had an impact on the coverage of the 2015/2016 Ag Census relative to the population of interest. The changes included:

1. Some farms operating in 2014/2015 were no longer operating in 2015/2016
2. There were new farms in the 2015/2016 population that were not in existence in 2014/2015
3. There was a population scope change made in 2015/2016 to exclude the smallest farms from the scope of the Ag Census
4. Some farms that produced a commodity in one year might not do so in the next year (and vice versa)

Points 1 and 3, combined with non-response in the Ag Census, introduce some under-coverage error relative to the 2014/2015 population. We end up with an overall coverage rate for the Ag Census of about 53% relative to the 2014/2015 population.

Point 2 introduces some over-coverage error in the Ag Census dataset – we addressed this by removing the new farms from the Ag Census dataset to produce estimates.

Point 4, reflecting the changes in farming activity that may occur from year to year, introduces some measurement error in the 2015/2016 Ag Census data relative to the 'true' 2014/2015 population. This will have an impact on the proportion estimates and the resulting MSEs produced from the Ag Census data.

There is some non-response error present in the 2014/2015 REACS data. For that collection, the response rate was around 80%. Given that it is a sample survey rather than a census of the population, there are also sample errors that will arise from using this dataset to form estimates.

For this case study, linkage error is not a concern, as there is a consistent ABS unit identifier across the two datasets that we were able to use to form the hybrid estimator.

In terms of the Error Frameworks outlined in Section 2, the most important errors for the big data were coverage, missing units (because of non-response) and measurement. The hybrid estimate mitigates the first two error sources but inherits sampling error and potentially non-response error. The third error source, measurement error, can be corrected for by using the bias-adjusted hybrid estimator.

For the case study, two versions of the Ag Census dataset were prepared. The first version replaced the original Ag Census data with reported values from the 2014/2015 REACS where we could. This was possible for farms that were part of the responding REACS sample, A_R, and were also in B. This dataset represented a "no-measurement-error" scenario for the Big Data.

The "no-measurement-error" dataset was amended by artificially introducing some measurement error (false negatives and false positives) into the reported data for four data items of interest:

- Farm with beef cattle
- Farm with sheep
- Farm producing lettuce
- Farm producing wheat

This second dataset represented the "with-measurement-error" scenario. Table 3 contains the simulated rate of false negatives and false positives for the dataset.

Estimates and MSEs were produced for four proportions of interest using the approaches discussed in Sections 3 to 6:

Table 4
Estimates of proportions for selected data items – no measurement error case

	Estimation approach			
Commodity	Big data	Survey (A_R only)*	Survey ($A_{R\cap C}$ only)	Hybrid estimate
Beef	41.9%	43.7%	41.0%	41.5%
Sheep	30.5%	24.9%	17.1%	24.2%
Lettuce	0.24%	0.12%	0.04%	0.15%
Wheat	19.3%	13.5%	2.9%	11.6%

*Note: Applying the two-step non-response adjustment to A_R, the non-response adjusted proportions become 41.6% for Beef, 24.2% for Sheep, 0.15% for Lettuce, and 11.6% for Wheat.

Table 5
Representivity ratios for big data and responding sample

	RR from Ag census	RR from REACS
Beef	1.01	1.09
Sheep	1.37	1.04
Lettuce	1.62	0.82
Wheat	1.82	1.19

- Proportion of beef cattle farms in the 2014/2015 Agriculture farm population
- Proportion of sheep farms in the 2014/2015 Agriculture farm population
- Proportion of lettuce farms in the 2014/2015 Agriculture farm population
- Proportion of wheat farms in the 2014/2015 Agriculture farm population

7.1. No measurement error scenario

Table 4 shows the Australia-level estimates for the 4 proportions. By using the big data only, we over-estimate the proportion of Sheep, Lettuce and Wheat farms relative to the responding sample adjusted for non-response. The non-response adjusted estimates produced using the responding units to the survey are close to the hybrid estimates for these three commodities.

In Table 4, the A_R and $A_{R\cap C}$ estimates do not have any adjustment for non-response. In order to produce non-response adjusted estimates for the estimates from A_R and $A_{R\cap C}$, we applied a two-step weighting adjustment [9]. The first step in the weighting adjustment involved developing a response propensity model using logistic regression; the inverse of the resulting propensities was used to adjust the sample design weights for non-response. In the second step, the response-propensity adjusted weights were adjusted to meet a set of population benchmarks, subject to the constraint that the weighted (by the response propensity) sum of the squared differences between the response-propensity adjusted and final weights is a minimum. The non-response adjusted estimates for A_R are provided in the footnote to Table 4.

In order to produce MSEs for the various estimates, we first estimated $\hat{\theta}_A$. This was accomplished using the two-step non-response adjusted weights calculated for A_R, as described above.

The estimated $\hat{\theta}_A$ was combined with the observed θ_B and θ_{A_R} to estimate the representivity ratios for the big data and the responding survey sample respectively. Table 5 contains the estimated representivity ratios for each of the four commodities. Apart from Beef, the big data representivity ratios tend to be quite far from 1. This reflects an imbalance in the big dataset of the number of farms that are producing the commodities of interest, relative to the population. The representivity ratios for the responding sample are comparatively much closer to 1 for Sheep, Lettuce and Wheat, suggesting there is less non-response bias in the responding sample compared with the big data sample for those items. For beef, the representivity ratio shows that the big data sample is more "representative" than the survey responding sample!

Table 6 shows the MSEs for each of the estimates in Table 4, alongside the corresponding representational bias-squared and variance components. For three of the four proportions produced, the big data-only estimates have the largest MSE compared with the other two estimation approaches due to the large bias associated with that dataset. For Beef, the MSE of the big data estimator is the lowest of the three estimators – reflecting the fact that its

Table 6
MSEs of proportions for selected data items – no measurement error

Data item	Estimator from	$^{(1)(2)(3)}$Representational bias^2 (x10^{-6})	$^{(4)(5)}$Var (x10^{-6})	MSE (x10^{-6})
Beef	Big Data only	11.34	0.00	11.34
	Responding sample only	444.87	2.00	446.87
	Hybrid estimate	97.94	0.95	98.89
Sheep	Big Data only	3947.55	0.00	3947.55
	Responding sample only	41.79	1.52	43.32
	Hybrid estimate	5.57	0.56	6.13
Lettuce	Big Data only	0.85	0.00	0.85
	Responding sample only	0.07	0.0099	0.0792
	Hybrid estimate	0.0014	0.0014	0.0028
Wheat	Big Data only	5919.20	0.00	5919.20
	Responding sample only	356.45	0.95	357.40
	Hybrid estimate	6.30	0.13	6.43

(1) Representational bias for big data calculated by $\left\{\frac{(r_B-1)P_U(1-P_U)}{1+(r_B-1)P_U}\right\}^2$. (2) Representational bias for responding sample calculated by $\left\{\frac{(r_R-1)P_U(1-P_U)}{1+(r_R-1)P_U}\right\}^2$. (3) Representational bias hybrid estimate calculated by $\frac{W_C^2(r_R-1)^2P_C^2(1-P_C)^2}{\{1+(r_R-1)P_C\}^2}$. (4) Variance for responding sample calculated by $\frac{1}{n_{A_R}}\frac{r_RP_U(1-P_U)}{\{1+(r_R-1)P_U\}^2}$. (5) Variance for hybrid estimate calculated by $\frac{W_C^2}{n_{A_R\cap C}}\frac{r_RP_C(1-P_C)}{\{1+(r_R-1)P_C\}^2}$.

representivity ratio is very close to 1 thus leading to a small bias-squared component. This result shows that where the representational bias from the big data is relatively small, it can be an effective estimator.

The responding sample estimator tends to have moderate MSEs, largely due to a moderately large bias-squared component from non-response. The hybrid estimator has the lowest MSE for all commodities except for Beef, where we might expect it to perform worse than just the big data estimator due to the larger representivity ratio in the responding sample. For the beef item, as the hybrid estimator is trying to correct for representational bias by using the survey data which in reality suffers more bias, it therefore not surprisingly does not perform as well as the big data estimator. This demonstrates the importance of checking the representativity ratios before deciding whether to apply the hybrid estimator, rather than applying the hybrid estimator blindly in the expectation that it will always provide the best MSE. The result shows that it is counter-productive to use the hybrid estimator if the representivity ratio of the big data is smaller than that of the responding sample data.

7.2. Measurement error scenario

The presence of measurement error in the big data will lead to an increase in the MSE of the estimators that make use of the big data. By linking the reported values from the 2014/2015 REACS responding sample onto the 2015/2016 Ag Census, this allowed us to obtain an idea of the nature of the measurement error and estimate α_0, the false negative rate, and α_1, the false positive rate. However, the linkage will not be complete since REACS has non-responding businesses.

Several options can be used to estimate α_0 and α_1. We tested out four approaches as part of the case study. All the approaches begin by linking the REACS responding records onto the Ag Census. What to do with the unlinked Ag Census records then varies:

1. Use the linked data directly to form α_0 and α_1. This assumes that the linked records provide a good representation of the measurement error for the overall big dataset.
2. Assume that the REACS value for all unlinked records was 1
3. Assume that the REACS value for all unlinked records was 0
4. Treat the unlinked records as 'non-response' on the Big Dataset. Account for the non-response through a response propensity weight adjustment.

Approach 1 may be reasonable if we believe that the unlinked records are 'not linking' completely at random, i.e. the missingness is not related to the target variable in the survey data sets. Approach 4 attempts to address possible 'non-linking bias' that may be present in Approach 1 by applying a response propensity adjustment for the unlinked

Table 7
Estimated $\hat{\alpha}_0$ and $\hat{\alpha}_1$

	Data item			
	Beef	Sheep	Lettuce	Wheat
$\hat{\alpha}_0$	0.1347	0.1319	0.2169	0.2434
$\hat{\alpha}_1$	0.1203	0.0859	0.0009	0.0622

Table 8
Estimates of proportions for selected data items – with measurement error

	Without measurement error correction		With measurement error correction	
Commodity	Big data	Hybrid estimate	Big data	Hybrid estimate
Beef	43.4%	42.3%	42.1%	41.6%
Sheep	32.5%	25.2%	30.6%	24.2%
Lettuce	0.30%	0.18%	0.27%	0.16%
Wheat	19.6%	11.8%	19.3%	11.6%

Table 9
MSEs of proportions for selected data items – with measurement error

Data item	Estimator from	[(a)] Repr. Bias^2 (x10^{-6}) (1)	Var (x10^{-6}) (2)	[(b)(c)] Meas. Bias^2 (x10^{-6}) (3)	[(d)(e)] Meas. Err. Corr. Var (x10^{-6}) (4)	MSE uncorrected (x10^{-6}) (1)+(2)+(3)	MSE corrected (x10^{-6}) (1)+(2)+(4)
Beef	Big Data only	27.67	0.00	169.96	2.38	197.63	30.06
	Responding sample only	444.87	2.00	0.00	0.00	446.87	446.87
	Hybrid estimate	97.94	0.95	47.68	0.67	146.58	99.56
Sheep	Big Data only	3895.42	0.00	373.81	2.06	4269.22	3897.47
	Responding sample only	41.79	1.52	0.00	0.00	43.32	43.32
	Hybrid estimate	5.57	0.56	1048.75	0.58	1054.88	6.71
Lettuce	Big Data only	1.87	0.00	0.08	0.04	1.94	1.90
	Responding sample only	0.069	0.0099	0.000	0.000	0.079	0.0792
	Hybrid estimate	0.001	0.0014	0.021	0.010	0.024	0.0130
Wheat	Big Data only	6826.62	0.00	10.28	1.90	6836.91	6828.53
	Responding sample only	356.45	0.95	0.00	0.00	357.40	357.40
	Hybrid estimate	6.30	0.13	2.89	0.53	9.31	6.96

(a) Representational bias-squared term calculated using measurement error corrected estimates of r_B. (b) Measurement error bias-squared for big data calculated by $\{\hat{\alpha}_1(1-P_B)-\hat{\alpha}_0 P_B\}^2$. (c) Measurement error bias-squared for hybrid estimator calculated by $W_B^2\{\hat{\alpha}_1(1-P_B)-\hat{\alpha}_0 P_B\}^2$. (d) Measurement error correction variance for big data calculated by $Var(\hat{\alpha}_1)(1-P_B)^2+Var(\hat{\alpha}_0)P_B^2$. (e) Measurement error correction variance for hybrid estimator calculated by $W_B^2\{Var(\hat{\alpha}_1)(1-P_B)^2+Var(\hat{\alpha}_0)P_B^2\}$.

records due to non-response in the survey data set, assuming 'missing at random'. The effectiveness of this approach depends on the availability of suitable auxiliary information that helps to explain the non-response. For this case study, the propensity model included a measure of the value of agricultural production as well as the area of holding of the farm. Both of the data items were available on the agricultural frame.

Approaches 2 and 3 provide 'worst-case' scenarios for estimating the impact to MSE due to measurement error.

We tested all four approaches using the confusion matrix methodology described in Section 6 above and found that for this case study Approach 1 and Approach 4 produced estimated $\hat{\alpha}_0$ and $\hat{\alpha}_1$ values that were very close to the true α_0 and α_1. Approach 4 produced slightly better estimates for all the data items of interest except Beef farms, and this is the approach that we adopted for the remainder of the case study. Table 7 shows the estimated $\hat{\alpha}_0$ and $\hat{\alpha}_1$ based on Approach 4.

The estimated $\hat{\alpha}_0$ and $\hat{\alpha}_1$ can be used to correct the measurement error in the big data. Table 8 compares the estimated proportions for the four data items with and without measurement error correction, when using the "with-measurement-error" Ag Census dataset. The introduction of measurement error results in big data and hybrid proportions that are a little larger than the corresponding estimates in Table 4, with the hybrid estimates addressing some but not all of the measurement error. Correcting for the measurement error brings the estimates much closer to the Table 4 values.

Table 9 contains the estimated MSE components for the 4 proportions, including the measurement error bias-squared contribution and the measurement error correction variance component. The measurement error contribution from the big data to overall MSE can be large, as seen in the fourth-last column in Table 9. For all four items, the presence of measurement error in the big data erodes the MSE gains from the hybrid estimator (third-last column). For Sheep, unlike the result In Table 6, the hybrid estimator is less efficient than the estimator based on the responding sample only. However, for the other three data items – including beef where the big data estimator in Table 6 shows that it performs better than the hybrid estimator when there is no measurement error – the hybrid estimator is preferred.

Correcting for the measurement error in the estimator is effective in significantly reducing the MSE for the big data and hybrid estimators. The bias-adjusted hybrid estimator has the lowest MSE for all data items except Beef, and for all the items it is preferred over the estimator based on the responding sample only (last column). With the correction for measurement error, the big data estimator is again preferred over the hybrid estimator for Beef.

8. Application

The Case Study illustrates how the REACS data, in conjunction with the data from the Ag Census can be used to develop hybrid estimates, when the scope of the Census was narrower than the survey (e.g. the exclusion of smaller agriculture units). It shows that big data can be extremely inaccurate if the big data is not representative and, for this data set at least, the hybrid estimates are more accurate than those estimates based on the big data or responding sample only, with the exception of estimator for beef because the big data is representative for this commodity. The case study also shows that measurement error in the big data set may impact on the effectiveness of the hybrid estimates and, where this happens, it is advisable to remove these errors before compiling the hybrid estimators. Where this is done, the case study shows that, with the exception of beef, the measurement error bias-adjusted hybrid estimates are still better than survey or big data estimates.

Another potential application of hybrid estimates is to obtain more reliable estimates of Covid-19 prevalence. Many countries have a surveillance testing data base which is certainly a big data set. However, it is not representative of the population. It is based on a self-selected sample of persons who have reported for testing because they are concerned for some reason and there are testing facilities they can access. It will also include persons who have been asked to test through Covid-19 surveillance systems e.g. close contacts of someone who has tested positive or required to test to enable travel etc.. This sample will under-represent certain classes of persons especially those who are asymptomatic or mildly symptomatic. The sample will also over-represent those that are more seriously ill including those requiring hospitalisation. Trends can also be misleading as cases will increase as tests increase. Nevertheless, this data base contains a lot of detailed data of potential value. It could be supplemented with a probability survey similar to that conducted successfully by the UK Office of National Statistics (https://www.ons.gov.uk/cis). Hybrid estimates could then be produced that use the strengths of both data sets to provide estimates of population prevalence by population group.

It may help overcome a tendency for an upward bias in forecasts derived from epidemiological models over a long period of time (see [11]). We believe a contributing factor is the non-adjustment for selection bias in the testing data bases that are used to estimate the model parameters.

In this application, the big data includes coverage errors in part induced by self-selection bias. Validity and measurement errors should be low if the testing protocols are satisfactory. The probability survey should not have coverage errors but will have sampling and non-response errors and will have larger measurement errors if testing is of lower quality. It is likely that hybrid estimates will be effective if any non-response bias can be mitigated in the survey (e.g. through weighting) and quantitative information is available on measurement errors that can be used in the estimation process.

Linkage error may be a concern. It should not be so important if a population register is used to identify those persons missing from the scope of the surveillance system. However, linkage error may be important if there is reliance on a question such as “Have you undertaken a Covid-19 test over the last x weeks?” where there would be measurement errors in the survey data set. [10] explores the use of a combination of surveillance testing and probability survey data and propose something like a hybrid estimate.

9. Discussion and conclusions

Hybrid estimates play an important role in potentially making greater use of big data especially when it is subject to coverage error. The first step should be to assess the error structure of the big data. Are the errors sufficiently unimportant for it to be used in isolation to produce official statistics perhaps with the type of adjustments described in the first three options outlined in Section 1.

Alternatively, does the big data need to be supplemented by a survey data source to adjust for the most important error sources? For estimating the finite population proportion, a good indicator is to check the representivity ratio against 1, or the relative size of the off-diagonals in a confusion matrix. Where required, hybrid estimates may be used, for example, to adjust for coverage error or self-selection error, or to adjust for validity error where the data concept available in the big data source is different to the target concept.

The paper provides an error framework that can be used to make decisions among these alternatives. They help to identify the most important potential error sources so consideration can be given to how they might be best mitigated. In this analysis, it is important to understand both the systematic and random components of these errors. It may be necessary to commission some special studies to support this analysis.

This paper also provides a method for helping make these decisions under restrictive (but realistic) assumptions on representational and measurement errors. It assumes the most important sources of error are non-response for surveys and coverage, self-selection bias or measurement errors in big data. It also assumes that survey statistician can determine which of the survey units is included in the big data set and can construct a "validity" sample for measurement error adjustment, where required. Representivity ratios are the key statistic in determining the "representativeness" of the big data and survey data when estimating the finite population proportion and should be used as a guide to determine if the efficiency of the estimation can be improved by hybrid estimates. As well, the confusion matrix can be used to determine the extent of measurement errors and can be used to provide estimates of false positive and false negative rates for measurement error adjustment.

It is also clear in the above exposition that sample surveys are going to stay in the statistician's tool kit in spite of the emergence and popularity of big data.

References

[1] Biemer P, Lyberg L. *Introduction to Survey Quality*, Wiley and sons. 2003.
[2] Amaya A, Biemer P, Kinyon D. Total Error in a Big data World. Adapting the TSE Framework to Big Data. Journal of Survey Statistics and Methodology. 2020; 8: 89–119.
[3] Reid G, Zabala F, Holmberg A. Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ. Journal of Official Statistics. 2017; 37: 477–511.
[4] Kim JK, Tam SM. Data integration by combining big data and survey sample data for finite population inference. International Statistical Review. 2021; 89: 382–401.
[5] Lohr S. Multiple-frame surveys for a multiple-data-source world, *Survey Methodology*, to appear in December 2021 issue. 2021.
[6] Tam SM, Kim JK. Big data ethics and selection bias: an official statistician's perspective. Statistical Journal of the International Association of Official Statistics. 2018; 34: 577–588.
[7] Brick M. Unit nonresponse and weighting adjustments: A critical review. Journal of Official Statistics. 2013; 29: 329–353.
[8] Kim J, Shao J. *Statistical Methods for Handling Incomplete Data*, Chapman and Hall. 2014.
[9] Dupont F. Alternative adjustment where there are several levels of auxiliary information. Survey Methodology. 1995; 21: 125–135.
[10] Alleva G, Arbia G, Falorsi PD, Zuliani A. A sample approach to the estimation of critical parameters of the SARS-CoV-2 epidemics: an operational design with a focus on the Italian health system. University of Rome. 2020. https://web.uniroma1.it/memotef/sites/default/files/Paper.pdf.PDF file.
[11] John PA. Ioannidis, Sally Cripps & Martin A. Tanner. "Forecasting for COVID-19 has failed". 2020. https://forecasters.org/blog/2020/06/14/forecasting-for-covid-19-has-failed/.

Appendix

From $Z_i \perp Z_j, Y_i \perp Y_j, Y_i \perp Z_j$ where $i \neq j$; $\Pr(Z = 0|Y = 1) = \alpha_0$ and $\Pr(Z = 1|Y = 0) = \alpha_1$; and $\bar{Z}_B = \frac{\sum_{i=1}^{n_B} Z_i}{N_B}$, we have

$$
\begin{aligned}
\mathit{MSE}(\bar{Z}_B) &= E(\bar{Z}_B - P_U)^2 \\
&= E(\bar{Z}_B - P_B)^2 + 2E(\bar{Z}_B - P_B)(P_B - P_U) + E(P_B - P_U)^2 \\
&= \frac{1}{N_B^2} E\left\{\sum_{i=1}^{N_B}(Z_i - Y_i)\right\}^2 + \frac{2}{N_B^2} E\left\{\sum_{i=1}^{N_B}(Z_i - Y_i)(Y_i - P_U)\right\} + \mathit{MSE}(P_B) \\
&= \frac{1}{N_B^2}\sum_{i=1}^{N_B}\{E(Z_i) - 2E(Z_iY_i) + E(Y_i)\} + \frac{2}{N_B^2}\sum_{i=1}^{N_B}\{E(Z_iY_i) - E(Y_i) - P_U[E(Z_i) - E(Y_i)]\} \\
&\quad + \frac{1}{N_B^2}\sum_{i\neq j}^{N_B}\{E(Z_i) - E(Y_i)\}\{E(Z_j) - E(Y_j)\} \\
&\quad + \mathit{MSE}(P_B) \\
&= \frac{1}{N_B^2}\sum_{i=1}^{N_B}\{E(Z_i) - P_B - 2P_U[E(Z_i) - P_B]\} + \frac{1}{N_B^2}\sum_{i\neq j}^{N_B}\{E(Z_i) - P_B\}\{E(Z_j) - P_B\} \\
&\quad + \mathit{MSE}(P_B) \\
&= \frac{1}{N_B^2}\sum_{i=1}^{N_B}\{[E(Z_i) - P_B]\}(1 - 2P_U) \\
&\quad + \frac{1}{N_B^2}\sum_{i\neq j}^{N_B}\{E(Z_i) - P_B\}\{E(Z_j) - P_B\} \\
&\quad + \mathit{MSE}(P_B) \\
&= \frac{1}{N_B}\{\alpha_1(1 - P_B) - \alpha_0 P_B\}(1 - 2P_u)\} + \frac{N_B - 1}{N_B}\{\alpha_1(1 - P_B) - \alpha_0 P_B\}^2 + \mathit{MSE}(P_B)
\end{aligned}
$$

noting $Z_i^2 = Z_i$ and $Y_i^2 = Y_i$

$$
\begin{aligned}
E(Z_i) &= \Pr(Z_i = 1) * 1 + \Pr(Z_i = 1) * 0 \\
&= \Pr(Z_i = 1) \\
&= \Pr(Z_i = 1|Y_i = 1)\Pr(Y_i = 1) + \Pr(Z_i = 1|Y_i = 0)\Pr(Y_i = 0) \\
&= (1 - \alpha_0)P_B + \alpha_1(1 - P_B),
\end{aligned}
\tag{A1}
$$

noting $E(Y_i) = \Pr(Y_i = 1) = P_B$. In addition,

$$
\begin{aligned}
E(Z_iY_i) &= \Pr(Z_iY_i = 1) \\
&= \Pr(Z_i = 1, Y_i = 1) \\
&= \Pr(Z_i = 1|Y_i = 1)\Pr(Y_i = 1) \\
&= (1 - \alpha_0)P_B
\end{aligned}
$$

Because $\frac{1}{N_B}\{\alpha_1(1 - P_B) - \alpha_0 P_B\}(1 - 2P_B)\} + \frac{N_B - 1}{N_B}\{\alpha_1(1 - P_B) - \alpha_0 P_B\}^2 = \{\alpha_1(1 - P_B) - \alpha_0 P_B\}^2 + O(N_B^{-1})$, we can say that, because of measurement errors, $\{\alpha_1(1 - P_B) - \alpha_0 P_B\}^2$ is the extra uncertainty added to the MSE and this is correct to $O(N_B^{-1})$.