sciendo

Journal of Official Statistics, Vol. 37, No. 3, 2021, pp. 751–769, http://dx.doi.org/10.2478/JOS-2021-0033

A Simulation Study of Diagnostics for Selection Bias

Philip S. Boonstra[1], *Roderick J.A. Little*[1], *Brady T. West*[2], *Rebecca R. Andridge*[3], *and Fernanda Alvarado-Leiton*[4]

A non-probability sampling mechanism arising from nonresponse or non-selection is likely to bias estimates of parameters with respect to a target population of interest. This bias poses a unique challenge when selection is 'non-ignorable', that is, dependent on the unobserved outcome of interest, since it is then undetectable and thus cannot be ameliorated. We extend a simulation study by Nishimura et al. (2016) adding two recently published statistics: the 'standardized measure of unadjusted bias' (SMUB) and 'standardized measure of adjusted bias' (SMAB), which explicitly quantify the extent of bias (in the case of SMUB) or nonignorable bias (in the case of SMAB) under the assumption that a specified amount of non-ignorable selection exists. Our findings suggest that this new sensitivity diagnostic is more correlated with, and more predictive of, the true, unknown extent of selection bias than other diagnostics, even when the underlying assumed level of non-ignorability is incorrect.

Key words: Non-ignorable selection bias; survey nonresponse; multiple imputation; pattern mixture model.

1. Introduction

Classical methods of scientific probability sampling and corresponding design-based frameworks for making statistical inferences about populations have long been used to advance scientific knowledge in various fields. The random selection of elements from a population of interest into a probability sample, where all population elements have a known non-zero probability of selection, ensures that elements included in the sample mirror the population in expectation. That is, for all variables of interest, the mechanism of selection of a subset of elements into the sample is ignorable, following the theoretical framework for missing-data mechanisms originally introduced by Rubin (1976).

The modern survey research environment poses significant challenges to these "tried and true" methodologies: it has become increasingly difficult to contact sampled units, survey response rates continue to decline in all modes of administration (face-to-face, telephone, and so on; Brick and Williams 2013; Williams and Brick 2018), and the costs of

[1] University of Michigan, Department of Biostatistics, 1415 Washington Heights, Ann Arbor, MI 48109-2029, Michigan, 48105, U.S.A. Emails: philb@umich.edu, and rlittle@umich.edu
[2] University of Michigan, Institute for Social Research, Survey Methodology Program, 426 Thompson Street, Ann Arbor, Michigan 48106, U.S.A. Email: bwest@umich.edu
[3] The Ohio State University, 1841 Neil Avenue, 242 Cunz Hall, Columbus, Ohio 43210, U.S.A. Email: andridge.1@osu.edu
[4] University of Michigan, Survey Methodology Progam, 4134 ISR-Thompson 426 Thompson St Ann Arbor, Michigan, U.S.A. Email: mleiton@umich.edu
Acknowledgments: This work was supported by an R21 grant from the National Institutes of Health (1R21HD090366-01A1). The authors thank Dr. Raphael Nishimura for sharing R scripts to calculate the Fraction of Missing Information and Mr. Chen Chen for his initial work on this simulation study.

collecting and maintaining scientific probability samples are steadily rising (Presser and McCulloch 2011). These problems raise the question of whether, and to what extent, samples can still produce reliable estimates when only a small fraction has responded, such that the response mechanism may in fact not be ignorable?

Given the difficulties of collecting data from probability samples, researchers are also turning to non-probability samples, which have the potential to yield large amounts of data at low cost. These may also be prone to non-ignorable selection bias, as the researcher no longer has control over the mechanism that ultimately yields the final sample. Given this trend in research methodology, indicators of the potential non-ignorable selection bias in non-probability samples and probability samples with low response rates are required.

Nishimura et al. (2016) investigated the suitability of various statistics for use as diagnostics for selection bias due to nonresponse mechanisms, of both the 'ignorable' or 'non-ignorable' type (Rubin 1976). They noted that none of the diagnostics they considered were intended to directly quantify selection bias. Moreover, their simulation study found that none of them were suitable as potential diagnostics, leaving the door open for other candidates. A statistic recently proposed in Little et al. (2020) explicitly estimates this bias based on an assumed level of non-ignorability and therefore is potentially appropriate for use as a diagnostic. The primary contribution of this article is the inclusion of this statistic in this comparison of diagnostics. We also extend Nishimura et al. (2016) by simulating two auxiliary variables that are differentially associated with the survey variable and selection, which we argue is an important additional factor when evaluating the diagnostics.

The remainder of this article is organized as follows. Section 2 presents notation and a brief description of the index of selection bias proposed in Little et al. (2020). Section 3 describes the other diagnostics we consider here, which were also evaluated in Nishimura et al. (2016). An important contextual difference between this article and that of Nishimura et al. is that we consider the generic non-selection scenario, of which survey nonresponse – the scenario of interest in Nishimura et al. – is a special case. The practical implication of this difference is that those indices that are dependent upon selection probabilities may not be calculable if they cannot be estimated. Sections 4 and 5 describe and present the results from the simulation study, respectively. Section 6 concludes with a discussion of all of the diagnostics considered in light of our results.

2. An Index of Selection Bias

For a target population of size N, with $i = 1, \ldots, N$, let $S_i \in \{0, 1\}$ indicate the selection of the ith subject into the sample, Y_i be the continuous outcome of interest, and Z_i be an observed auxiliary variable that is relevant due to its association with Y_i. The vectors $S = \{S_1, \ldots, S_N\}$ and $Z = \{Z_1, \ldots, Z_N\}$ are fully observed, and the vector $Y = \{Y_1, \ldots, Y_N\}$ is separated into selected (observed) and unselected (missing) sub-vectors, respectively: $Y_{\text{sel}} = \{Y_i : S_i = 1\}$ and $Y_{\text{unsel}} = \{Y_i : S_i = 0\}$. When needed, we will also use this same convention to separate Z into selected and unselected subvectors, Z_{sel} and Z_{unsel}, although in contrast to Y both subvectors of Z are always assumed to be fully observed. The primary estimand of interest is the average outcome in the target population: $E[Y_i] = \mu_y$.

Two forms of models for the joint distribution of $\{Y, Z, S\}$ are often considered. Selection models (Little and Rubin 2002) factorize the joint distribution as

$$[Y, Z, S|\alpha, \beta] = [Y, Z|\alpha] \Pr(S|Y, Z, \beta) \tag{1}$$

with parameters $\{\alpha, \beta\}$, where α and/or β may themselves be vectors. A model for $\Pr(S|Y, Z, \beta)$ describes the missingness mechanism for Y_{unsel}, since Y_i is not observed when $S_i = 0$. Thus, the strongest possible assumption to make regarding $\Pr(S|Y, Z, \beta)$ is that S and $\{Y, Z\}$ are jointly independent. Modifying the 'missing completely at random' terminology of Little and Rubin (2002), we call this 'selection completely at random' (SCAR). In this case β corresponds to the average selection rate. A weaker assumption is 'selection at random' (SAR), which assumes that S and Y are conditionally independent given Z. The weakest assumption is 'selection not at random' (SNAR), and elements of both α and β are not identified in this case.

The second decomposition is the class of 'pattern-mixture models' (Andridge and Little 2011; Little 1994), which describe outcome models that are specific to the selected and unselected populations:

$$\begin{aligned}[Y, Z, S|\theta_{\text{unsel}}, \theta_{\text{sel}}, \pi] &= [Y, Z|S, \theta_{\text{unsel}}, \theta_{\text{sel}}] \Pr(S|\pi) \\ &= [Y_{\text{unsel}}, Z_{\text{unsel}}|\theta_{\text{unsel}}][Y_{\text{sel}}, Z_{\text{sel}}|\theta_{\text{sel}}] \Pr(S|\pi),\end{aligned} \tag{2}$$

with parameters $\{\theta_{\text{unsel}}, \theta_{\text{sel}}, \pi\}$, where θ_{unsel} and θ_{sel} may be vectors and π is a scalar equal to the probability of selection. Both the selection and pattern-mixture decompositions are statistically valid, and in the special case of a SCAR mechanism, the models coincide, meaning that $\theta_{\text{unsel}} = \theta_{\text{sel}} \equiv \theta$ and $\{\theta, \pi\}$ and $\{\alpha, \beta\}$ share a 1-1 correspondence (Little 1994). Further, all parameters become identified in this special case. However, Equations (1) and (2) will not generally coincide under SAR for any distributional choices. Although the decomposition in Equation (1) is more intuitive by directly capturing the data-generating mechanism, the usefulness in focusing on Equation (2) is that the non-identified parameters are isolated to a single submodel: $[Y_{\text{unsel}}, Z_{\text{unsel}}|\theta_{\text{unsel}}]$. In the pattern-mixture framework, the estimand of interest, μ_y, is equal to $\pi E[Y_{\text{sel}}|\theta_{\text{sel}}] + (1 - \pi)E[Y_{\text{unsel}}|\theta_{\text{unsel}}]$. The latter mean, $E[Y_{\text{unsel}}|\theta_{\text{unsel}}]$, is not identified without making further assumptions.

Specifically, for the factorization in Equation (2), assume that $[Z_{\text{sel}}, Y_{\text{sel}}|\theta_{\text{sel}}]$ and $[Z_{\text{unsel}}, Y_{\text{unsel}}|\theta_{\text{unsel}}]$ are both bivariate normal, with θ_{sel} and θ_{unsel} each denoting five parameters (two means, two variances, and a covariance). Additionally, assume that the marginal distribution $\Pr(S|\pi)$ is coherent with some true conditional distribution of S given Z and Y that takes the form

$$\Pr(S = 1|Y, Z, \phi) = g(\phi Y + (1 - \phi)Z), \tag{3}$$

for some invertible function $g(t)$ having range in the interval $(0, 1)$ but otherwise unspecified, and for some scalar parameter $\phi \in [0, 1]$. The population mean μ_y becomes identified under these assumptions (Little 1994) and Andridge and Little (2011) derived a maximum likelihood estimate (MLE) of μ_y as a function of ϕ, given by

$$\hat{\mu}_y(\phi) = \bar{y}_{\text{sel}} + \frac{\phi + (1 - \phi)r_{\text{sel}}}{\phi r_{\text{sel}} + (1 - \phi)} \sqrt{\frac{\hat{\sigma}^2_{y_{\text{sel}}}}{\hat{\sigma}^2_{z_{\text{sel}}}}} (\bar{z}_{\text{sel}} - \bar{z}) \tag{4}$$

(these authors actually used the alternative parameterization $\psi = \phi/(1-\phi)$). Here, $\bar{y}_{\text{sel}}$, $\bar{z}_{\text{sel}}$, and $\bar{z}$ are the sample means of Y_{sel}, Z_{sel}, and Z, respectively; r_{sel} is the sample Pearson correlation between Y_{sel} and Z_{sel}; and $\hat{\sigma}^2_{y_{\text{sel}}}$ and $\hat{\sigma}^2_{z_{\text{sel}}}$ are the sample variances of Y_{sel} and Z_{sel}, respectively. For $\phi = 0$, i.e., when selection depends on Z alone, this estimator reduces to the regression estimator obtained from regressing Y on Z for the selected cases (Andridge and Little 2011).

Remark 1: Little et al. (2020) show that the Estimator (4) remains unbiased for its estimand under a more general class of functions than that given in Equation (3), namely

$$\Pr(S = 1|Y, Z, W, \phi) = g(\phi Y + (1-\phi)Z, W), \tag{5}$$

where W is uncorrelated with Z. This generalization will be important for explaining a key result in our simulation study.

This estimate of μ_y in Equation (4) is a function of the parameter ϕ, which in turn controls the extent to which sampling depends on the outcome Y, with larger values indicating greater dependence. When $\phi = 0$, the selection mechanism is SAR, and the resulting statistic is closely related to the measure H_1 in Särndal and Lundström (2010). When $\phi > 0$, the sampling mechanism is 'non-ignorable' (Rubin 1976), meaning that the sampled population cannot yield unbiased estimates of the target population parameter without knowledge of the true value of ϕ (Little et al. 2019). However, in any non-probability sample, ϕ is, by definition, not estimable, and Little et al. propose varying this parameter in a sensitivity analysis. Subtracting $\bar{y}_{\text{sel}}$ from both sides of Equation (4) and scaling by $\sqrt{\hat{\sigma}^2_{y_{\text{sel}}}}$ to standardize the resulting difference, we obtain a direct estimate of the standardized bias that would arise in using $\bar{y}_{\text{sel}}$ to estimate μ_y for a particular true value of ϕ. The resulting expression is the recently proposed *Standardized Measure of Unadjusted Bias* (SMUB) (Little et al. 2020):

$$\text{SMUB}(\phi) \equiv \frac{\hat{\mu}_y(\phi) - \bar{y}_{\text{sel}}}{\sqrt{\hat{\sigma}^2_{y_{\text{sel}}}}} = \frac{\phi + (1-\phi)r_{\text{sel}}}{\phi r_{\text{sel}} + (1-\phi)} \frac{(\bar{z}_{\text{sel}} - \bar{z})}{\sqrt{\hat{\sigma}^2_{z_{\text{sel}}}}}. \tag{6}$$

This measure quantifies the sensitivity of estimates based upon the selected sample due to increasing levels of non-ignorability, represented by the value of ϕ. As discussed in Little et al. (2020), in addition to a small value of the non-ignorability parameter ϕ decreasing the standardized bias, other characteristics that tend to do so include having an auxiliary variable that is a strong correlate for the outcome, that is, r_{sel} is close to 1, and/or obtaining a large sampled fraction, since $\bar{z}_{\text{sel}} - \bar{z} = (1-\pi)(\bar{z}_{\text{sel}} - \bar{z}_{\text{unsel}})$, where π is the selection probability in Equation (2).

Little et al. (2020) also proposed a *Standardized Measure of Adjusted Bias* (SMAB), defined as

$$\text{SMAB}(\phi) \equiv \text{SMUB}(\phi) - \text{SMUB}(0) = \frac{\phi(1-r_{\text{sel}})^2}{\phi r_{\text{sel}} + (1-\phi)} \frac{(\bar{z}_{\text{sel}} - \bar{z})}{\sqrt{\hat{\sigma}^2_{z_{\text{sel}}}}}. \tag{7}$$

Whereas SMUB measures the summative bias arising from both ignorable and non-ignorable mechanisms, SMAB measures only the excess bias after adjusting for ignorable

bias. As Little et al. (2020) caution, its utility is predicated on the underlying assumptions of the normal pattern-mixture model.

The simulation study in Nishimura et al. (2016, 43), prior to the proposal of the estimator in Equation (6), found that "none of the indicators [evaluated] fully depict the impact of nonresponse in survey estimates." We consider here whether the SMUB or SMAB indices address this deficiency. Note that Equation (6) is based on a normal pattern-mixture model, and as such is less well suited to non-normal outcomes. Modifications of Equation (6) for a categorical outcomes are discussed in Andridge and Little (2020) but are not considered in this article.

3. Other Diagnostics Evaluated

Nishimura et al. (2016) grouped the diagnostics they compared based on whether $\{S, Z\}$ or $\{S, Y_{\text{sel}}, Z\}$ are required to calculate them. Except for the sample mean of the selection indicator, these other diagnostics require at least some individual-level data from the non-sampled population (or some other means of accurately assessing the selection propensity). This situation is exceedingly rare in practice and makes these diagnostics difficult, if not impossible, to compute for non-probability samples. It also provides motivation for the additional diagnostic measures we evaluate in this article, which do not have this same requirement. We return to this important limitation in the Discussion (Section 6).

The simplest diagnostic is $\bar{s}$, that is, the sample mean of the selection indicator, or the selection rate. Small values of $\bar{s}$ increase the upper bound for potential bias due to non-ignorable sampling since a larger fraction of the data are missing (Nishimura et al. 2016) but do not necessarily indicate greater selection bias, for example Bootsma-van der Wiel et al. (2002). Since our focus is on how well measures reflect bias characteristics beyond the selection rate, we choose to include the selection rate as a design factor in our simulation study, rather than as a diagnostic for bias. In this section, we provide a brief rationale for the use of each of these diagnostics in the non-probability sampling setting; Nishimura et al. provide additional justification for each diagnostic in the special case of nonresponse conditional on being sampled.

3.1. Diagnostics Using {S, Z}

This category characterizes the associations between the fully observed auxiliary variable Z and the selection indicator S. The underlying rationale for doing so is that a selection rate dependent upon Z, which is itself a surrogate for Y, is suggestive of a selection rate dependent upon Y, that is, selection bias. Nishimura et al. (2016) consider three measures of this type, which are described below.

Consider first the selection model conditioning on Z alone:

$\Pr(S = 1|Z, \gamma_0, \gamma_z) = \text{logit}^{-1}(\gamma_0 + \gamma_z Z)$. This is fit to the data $\{S, Z\}$ from both the selected and unselected populations. Let the fitted probability, or propensity, of selection for the ith observation be given by

$$\eta_i \equiv \text{logit}^{-1}(\hat{\gamma}_0 + \hat{\gamma}_z Z_i). \tag{8}$$

The R-indicator (Schouten et al. 2009), where R stands for 'representativity', is the following linear transformation of the sample standard deviation of η_i across both the selected and un-selected samples:

$$\hat{R} = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\eta_i - \sum_{j=1}^{N}\eta_j/N\right)^2}.$$

Schouten proposed the R-indicator in the context of response propensities, and thus it is computed across all elements in the population and requires data sufficient to estimate the response/selection propensities η_i.

$\hat{R}$ theoretically ranges from 0 to 1, where smaller values correspond to greater variability in the selection propensities and, consequently, greater potential for selection bias. However, the smallest possible value of $\hat{R} = 0$, that is, when the sample standard deviation of the η_i's is 0.5, occurs only under two strong conditions. First, the average fitted selection propensity, $\sum_{j=1}^{N}\eta_j/N$, must be 0.5. Second, each individual propensity must either be $\eta_i = 1$ or $\eta_i = 0$, that is, S can be completely separated by Z, in the sense of Albert and Anderson (1984). In practice, $\hat{R}$ generally ranges between 0.5 and 1.

The *coefficient of variation* of the selection propensities is the ratio of the same standard deviation used in the R-indicator and the mean selection propensity:

$$\mathrm{CV}_S = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\eta_i - \sum_{j=1}^{N}\eta_j/N\right)^2}}{\sum_{j=1}^{N}\eta_j/N}.$$

The theoretical range of CV_S is the set of non-negative numbers. The rationale for using the coefficient of variation is that both variability in selection probabilities (the numerator) and smaller selection rates (the denominator) contribute to the potential for selection bias. As with the other indices, however, the challenge is that this relationship does not always hold, nor is the converse true: selection bias may exist even in the presence of a "small" CV_S.

Highly variable non-selection weights may also indicate greater potential for selection bias, depending on the extent to which the variables used to create the non-selection weights are associated with the outcome of interest. Thus, the *variability in non-selection weights* focuses on the inverse of the estimated selection probabilities, $1/\eta_i$. Nishimura et al. (2016) consider the sample variance of $1/\eta_i$ evaluated in the selected sample:

$$\mathrm{Var}(\eta^{-1}) = \frac{1}{(N\bar{s}) - 1}\sum_{i:S_i=1}\left(1/\eta_i - \left[\sum_{j:S_j=1}1/\eta_j\right]/[N\bar{s}]\right)^2.$$

Two other approaches limited to these same data assess the overall performance of the selection model $\Pr(S = 1|Z, \gamma_0, \gamma_z) = \mathrm{logit}^{-1}(\gamma_0 + \gamma_z Z)$ in distinguishing between selected and non-selected observations. One is the 'Area Under the receiver-operating characteristic Curve' (AUC), an assessment of discriminatory ability. The corresponding estimate counts the proportion of all possible selected-unselected pairs, the selection

propensities of which are correctly ordered:

$$\hat{\text{AUC}} = \frac{\sum\sum_{i,j:s_i=1,s_j=0} 1_{[\eta_i > \eta_j]}}{\sum\sum_{i,j:s_i=1,s_j=0}}.$$

The pseudo-R^2 seeks to generalize the linear model's R^2 metric, or proportion of variation explained, to a logistic framework (Nagelkerke et al. 1991). It is given by

$$\text{ps}R^2 = \frac{1 - \left(\frac{\bar{s}^{(N\bar{s})}[1-\bar{s}]^{(N[1-\bar{s}])}}{\sum_{i=1}^{N} \eta_i^{S_i}[1-\eta_i]^{1-S_i}}\right)^{2/N}}{1 - \left(\bar{s}^{(N\bar{s})}[1-\bar{s}]^{(N[1-\bar{s}])}\right)^{2/N}}.$$

Both $\hat{\text{AUC}}$ and $\text{ps}R^2$ quantify the strength of the model used to create the selection propensities. A better (stronger) relationship between auxiliary variables and selection could indicate a higher risk for selection bias, depending on the strength of the relationship between the auxiliary variables and the outcome of interest.

3.2. *Diagnostics Using {S, Y_{sel}, Z}*

The two diagnostics in this section make use of all available data and are therefore potentially more sensitive to detecting selection bias. The first is the Pearson correlation between the outcome Y and the inverse of the selection propensity η:

$$\text{Cor}(Y_{\text{sel}}, \eta^{-1}) = \frac{\sum_{i:S_i=1}\left(1/\eta_i - \left[\sum_{j:S_j=1} 1/\eta_j\right]/[N\bar{s}]\right)\left(Y_i - \left[\sum_{j:S_j=1} Y_i\right]/[N\bar{s}]\right)}{\sqrt{\sum_{i:S_i=1}\left(1/\eta_i - \left[\sum_{j:S_j=1} 1/\eta_j\right]/[N\bar{s}]\right)^2 \sum_{i:S_i=1}\left(Y_i - \left[\sum_{j:S_{j=1}} Y_i\right]/[N\bar{s}]\right)^2}}.$$

This correlation serves as a measure of the association between the survey variable and the set of auxiliary variables used to create the selection propensities. The stronger this relationship, the more potential there is to adjust for selection bias.

The second diagnostic is called the 'Fraction of Missing Information' (FMI), a statistic borrowed from the literature on multiple imputation (Rubin 2004). Given a posited model for the conditional distribution of the outcome Y given the auxiliary variable Z fit to the observed data $\{Y_{\text{sel}}, Z_{\text{sel}}\}$, M sets of unselected outcomes, denoted by $Y_{\text{unsel}}^{(m)}$ are imputed. Each of the M completed data sets, $\{Y_{\text{sel}}, Y_{\text{unsel}}^{(m)}\}$ are used to construct estimates of μ_y, say, $\hat{\mu}_y^{(m)}$, $m = 1, \ldots, M$. After some simplification, the FMI statistic can be written as

$$\text{FMI}(\mu_y) = \left(\frac{M+1}{M-1}\right)\left(\frac{\sum_{m-1}^{M}\left(\hat{\mu}_y^{(m)} - \frac{1}{M}\sum_{m'=1}^{M}\hat{\mu}_y^{(m')}\right)^2}{\sum_{m=1}^{M}\text{Var}\left(\hat{\mu}_y^{(m)}\right) + \sum_{m=1}^{M}\left(\hat{\mu}_y^{(m)} - \frac{1}{M}\sum_{m'=1}^{M}\hat{\mu}_y^{(m')}\right)^2}\right).$$

There are three contributing elements to this expression. The first element, $\sum_{m=1}^{M}\left(\hat{\mu}_y^{(m)} - \frac{1}{M}\sum_{m'=1}^{M}\hat{\mu}_y^{(m')}\right)^2$, appears in both the numerator and denominator and is the sum of the squared deviations between each imputation-specific estimate and the overall mean. It is proportional to the "between-imputation variance", capturing uncertainty in the estimate across replications of the imputation procedure. The second element, $\sum_{m=1}^{M}\mathrm{Var}\left(\hat{\mu}_y^{(m)}\right)$, is only in the denominator and is the sum of each imputation-specific variance estimate of $\hat{\mu}_y^{(m)}$. This is proportional to the "within-imputation variance", and the sum of the between- and within-imputation variances is thus the total variance. The third element, $(M+1)/(M-1) > 1$, multiplicatively inflates the between-over-total fraction and captures the loss of information due to taking a finite number of imputations. It approaches 1 from above as M is increased. Ranging between 0 and 1, larger values of FMI indicates greater uncertainty about the imputed values (larger between-imputation variance), which could indicate a greater potential for selection bias.

4. Simulation Study: Description

The purpose of this simulation study is to characterize the association between the true bias in a sampled data set (only observable in a simulation framework) and each of the aforementioned candidate diagnostics, including the new SMUB and SMAB diagnostics from Little et al. (2020). The data were generated according to the 'selection model' decomposition described in Equation (1). However, recognizing that, in practice, there may be more than one auxiliary variable having different associations with selection and the survey variable, we used two auxiliary variables, X_1 and X_2, in place of Z. In truth, S and X_1 are conditionally independent given X_2 and Y, and, similarly, Y and X_2 are conditionally independent given X_1.

Remark 2: Nishimura et al. (2016) also use two auxiliary variables but for different purposes. One of their auxiliary variables is latent and used only to control the extent of response/selection not-at-random, whereas we are simulating non-response directly (see Remark 3 below). These approaches are distributionally equivalent. Nishimura's other variable is an observed explanatory variable that is assumed to correlate with both the response/selection indicator and the outcome and thus serves the role of our two joint auxiliary variables, X_1 and X_2.

In more detail, at each iteration, a finite population of size $N = 10^4$ was simulated, wherein each observation consisted of the random vector $\{Y, X_1, X_2, S\}$ drawn from the true models in the second column in Table 1. X_1 and X_2 are bivariate normal with mean 0, variance 1, and correlation κ. When X_1 and X_2 are not identically equal, that is, $\kappa < 1$, both

Table 1. Description of generating models used in the simulation study in Section 4. Five parameters fully specify the generating distribution of the data: κ, ρ, β_0, β_x, and β_y.

Variable	Generating model
Auxiliary	$[X_1, X_2 \mid \kappa] = N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \kappa\\ \kappa & 1\end{pmatrix}\right)$
Outcome	$[Y_i \mid X_1, \rho] = N\left(\rho X_1, \sqrt{1-\rho^2}\right)$
Selection	$\Pr(S=1 \mid Y, X_2, \beta_0, \beta_y, \beta_x) = \mathrm{logit}^{-1}(\beta_0 + \beta_y Y + \beta_x X_2)$

X_1 and X_2 are conditioned on in fitting the outcome and selection models, to emulate what would be done in practice. The scalar parameter ρ is the Pearson correlation between Y and X_1, that is, $[Y|X_1, \rho] = N\left(\rho X_1, \sqrt{1-\rho^2}\right)$; Y and X_2 are conditionally independent given X_1. Finally, the selection probability is controlled by parameters β_0, β_x, and β_y in a logistic framework, with $\Pr(S = 1|Y, X_2, \beta_0, \beta_y, \beta_x) = \text{logit}^{-1}(\beta_0 + \beta_y Y + \beta_x X_2)$. In total, there are five parameters governing this distribution: κ, ρ, β_0, β_x, and β_y.

Remark 3: An equivalent model for inducing correlation between S and Y would be achieved by the introduction of a latent variable into the generating distribution of each, as in Heckman (1979), where $S = 1$ if the latent variable crosses some threshold.

We considered $\kappa \in \{0, 0.5, 1\}$, with the last scenario corresponding to $X_1 \equiv X_2 \equiv Z$, in which case we are in the 'single auxiliary variable' scenario, and one would not condition on both X_1 and X_2. The correlation between the outcome Y and its best predictor X_1 was $\rho \in \{0.10, 0.25, 0.75\}$. Values of β_x and β_y, the log-odds ratios for selection, were taken from one of the scenarios listed in Table 2. The first row, for which $\beta_x = \beta_y = 0$, corresponds to a SCAR mechanism. The second row, for which $\beta_x \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and $\beta_y = 0$, corresponds to five different SAR mechanisms. The remaining rows in the table, for which $\beta_y \neq 0$ and $|\beta_x| + |\beta_y| \equiv c \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, all correspond to different SNAR mechanisms, ranging from mild non-ignorability (third row: $\{\beta_x, \beta_y\} = \{3c/4, c/4\}$) to extreme non-ignorability (sixth row: $\{\beta_x, \beta_y\} = \{0, c\}$) in which selection depends entirely on Y. In total, Table 2 gives 31 unique sets of β_x and β_y.

Under this generating model, the assumption in Equation (5) holds for any $\kappa \in [0, 1]$. To see this, express X_2 as $X_2 = \kappa X_1 + \sqrt{1-\kappa^2}\varepsilon$, where $\varepsilon \sim N(0, 1)$ is independent of X_1 and Y.

Substituting this result into the selection model, we rewrite the selection probability as

$$\begin{aligned}
\Pr(S = 1|Y, X_2) &= \Pr(S = 1|Y, X_1, \varepsilon) \\
&= \text{logit}^{-1}\left(\beta_0 + \beta_y Y + \beta_x\left[\kappa X_1 + \sqrt{1-\kappa^2}\varepsilon\right]\right) \\
&= \text{logit}^{-1}\left(\beta_0 + \beta_y Y + \kappa\beta_x X_1 + \beta_x\sqrt{1-\kappa^2}\varepsilon\right).
\end{aligned}$$

Table 2. Values of ϕ_{true} for the pair of log-odds ratios in the true selection mechanism of the simulation study grouped by the relative relationship of β_x to β_y, where, except for the first row, $|\beta_x| + |\beta_y| \equiv c \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. The implied true value of the non-ignorability parameter ϕ is calculated by the expression $\phi_{\text{true}} = \beta_y/(\kappa\beta_x + \beta_y)$.

Label	$\{\beta_x, \beta_y\}$	ϕ_{true}		
		$\kappa = 1$	$\kappa = 0.5$	$\kappa = 0$
SCAR	$\{0, 0\}$	0*	0*	0*
SAR	$\{c, 0\}$	0	0	0
$3X_2 + Y$	$\{3c/4, c/4\}$	0.25	0.4	1
$X_2 + Y$	$\{c/2, c/2\}$	0.5	0.66	1
$X_2 + 3Y$	$\{c/4, 3c/4\}$	0.75	0.86	1
Y	$\{0, c\}$	1	1	1
$X_2 - Y$	$\{c/2, -c/2\}$	$-^{\dagger}$	$-^{\dagger}$	1

*Mathematically, ϕ_{true} is undefined when $\beta_x = \beta_y = 0$, but we use 0 here to indicate that this is an ignorable sampling mechanism. †There is no value of $\phi_{\text{true}} \in [0, 1]$ satisfying the assumptions required for the SMUB indices when β_x or β_y are negative and $\kappa > 0$.

Now, letting (i) $g(t_1, t_2) = \text{logit}^{-1}(\beta_0 + [\kappa\beta_x + \beta_y]t_1 + t_2)$, (ii) $\phi = \beta_y/(\kappa\beta_x + \beta_y)$, (iii) $Z = X_1$, and (iv) $W = \beta_x\sqrt{1-\kappa^2}\varepsilon/(\kappa\beta_x + \beta_y)$, the relaxed assumption (5) is satisfied for any $\kappa \in [0, 1]$. In contrast, the more restrictive assumption (3) is only satisfied for $\kappa = 1$, that is, $W \equiv 0$. Under $\kappa = 1$, the third column in Table 2 gives the implied true value of ϕ, which is common to all $\{\beta_x, \beta_y\}$ pairs in each row and which we denote as ϕ_{true} to distinguish it from the closely related tuning parameter ϕ used by SMUB. The last two columns give the value of ϕ_{true} for $\kappa = 0.5$ and $\kappa = 0$, respectively. In the last row of Table 2, for which $\beta_x > 0$ and $\beta_y < 0$, there is no value of $\phi_{\text{true}} \in [0,1]$ satisfying Equation (5) except for the case that $\kappa = 0$, and this is noted as such in the table.

With regard to the intercept β_0, we did not directly set its value but rather fixed a desired overall selection probability $\Pr(S = 1) = 0.05$ (marginally over all other random variables), which, when set equal to $E_{Y,X_2}[\text{logit}^{-1}(\beta_0 + \beta_x X_2 + \beta_y Y)]$, can then be numerically solved for β_0. Our choice of a 5% selection rate is a fairly large selection rate for non-probability samples.

Two of the diagnostics have input values that the user must select. For SMUB, we inspected three choices of the non-ignorability tuning parameter in Equation (6): $\phi \in \{0, 0.5, 1.0\}$. When ϕ is close to the unknown ϕ_{true}, the SMUB statistic will estimate well the unadjusted bias, as defined below. For SMAB, we used $\phi \in \{0.5, 1.0\}$, since SMAB($\phi = 0$) is always equal to 0. As with SMUB, when ϕ is close to the unknown ϕ_{true}, the SMAB statistic will be close to its estimand, namely adjusted bias. For FMI, we estimate μ_y by imputing $M = 30$ vectors of the unselected outcomes Y_{unsel} conditional on the auxiliary variables X_1 and X_2 within a Bayesian linear regression model framework.

For each of the $3 \times 3 \times 31 = 279$ combinations of ρ, κ, and $\{\beta_x, \beta_y\}$ pairs taken from Table 2, we simulated 2,000 independent populations of size 10^4 and, from each, sampled a data set according to the corresponding parameters. The available data were always $\{S, X_1, X_2, Y_{\text{sel}}\}$, although not all diagnostics make use of all data, as noted in the previous sections. For those diagnostics depending on the sampling probability η, we regressed S against the auxiliary covariates $X1$ and $X2$ in the entire population data.

To assess performance, we calculated for each data set the 'standardized error measure' (SEM) in using $\bar{y}_{\text{sel}}$ to estimate μ_y, which is given by

$$\text{SEM} = \frac{\bar{y}_{\text{sel}} - \mu_y}{\sigma_y}. \tag{9}$$

In words, this is the difference between the empiric mean of the outcome in the selected observations and the target population mean, divided by the true standard deviation of the outcome. We plot the median value of SEM against the median value of each diagnostic to visualize the systematic relationship between these two quantities. A diagnostic that is sensitive to selection bias should be associated with SEM, and both the qualitative and quantitative nature of this association should be similar for all types of selection mechanisms, that is, values of ϕ_{true}. Also important is the pairwise relationship due to sampling variability, or "chance bias". To that end, we also calculate the Spearman correlation between the value of SEM and each diagnostic across all 2,000 data sets from each scenario.

Because calibration is often used in practice to adjust for the potential selection bias in non-probability samples, we also calculated a secondary error measure using a calibrated estimator of the average outcome. Specifically, we separately categorized X_1 and X_2 into

groups defined by the marginal quartiles in the population data, yielding 16 bivariate categories, and then weighted each observation in the sampled data by the ratio of its corresponding category's relative frequency in the population data versus its relative frequency in the sampled data. The calibrated estimator is the weighted mean of the outcome in the sampled data, denoted by $\bar{y}_{\text{sel}}^{\text{cal}}$. Then, the 'standardized adjusted error measure (SAEM)' is defined as

$$\text{SAEM} = \frac{\bar{y}_{\text{sel}}^{\text{cal}} - \mu_y}{\sigma_y}. \tag{10}$$

Results corresponding to SEM are given in the main text, and those corresponding to SAEM are in the Online Supplemental Material. All analyses were conducted in the R statistical environment (R Core Team 2018; Van Buuren and Groothuis-Oudshoorn 2011; Wickham 2017). Code for the simulation study is available here: https://github.com/bradytwest/IndicesOfNISB/tree/master/SelectionBiasDiagnostics.

5. Simulation Study: Results

Figures 1, 2, and 3 plot the relationship between the median value of SEM across 2,000 simulated data sets from a given scenario against the median of each diagnostic, separately for $\kappa = 1$, 0.5, and 0, respectively. Figures S1, S2, and S3 in the Online Supplemental Material give these analogous results using the alternative metric SAEM.

Points in which the underlying selection mechanism share their row in Table 2 in common are connected. Generally speaking, a diagnostic is good at *detecting* bias if its value (on the *x*-axis) changes at a similar rate with the observed bias (on the *y*-axis) across all of the different selection mechanisms, that is, each plotted segment has a similar sized slope. It is useful for *estimating* bias if its value changes at the same rate as the observed bias across the selection mechanisms, that is, each plotted segment is close to the line $y = x$ (which is given by a solid black line but is not visible in all panels due to the scale of each diagnostic). There is no information in the data to determine the extent to which selection depends on Y, as represented by the different lines in the figures. If, for a single value of a diagnostic on the *x*-axis, there are many different values of SEM on the *y*-axis across different selection mechanisms, this is evidence against it being a good diagnostic. The set of candidate diagnostics are separated into two groups in each figure, with the set of six in the top three rows (one row each for $\rho = 0.75$, $\rho = 0.25$, and $\rho = 0.10$) roughly corresponding to the best performing diagnostics, and the set in the bottom three rows corresponding to the worst performing diagnostics.

Considering first the diagnostics in the bottom rows of Figure 1, $\text{Cor}(Y_{\text{sel}}, \eta^{-1})$ and $\text{FMI}(\mu_y)$ are not notably sensitive to changes in SEM, as indicated by the steep vertical segments. The $\text{Var}(\eta^{-1})$ diagnostic changes with SEM, but the range of its *x*-axis is very wide, potentially limiting interpretability as to what constitutes an extreme value. The $\hat{R}$, $\hat{\text{AUC}}$, and $\text{ps}R^2$ diagnostics are also sensitive to SEM and have a narrower range along the *x*-axis than $\text{Var}(\eta^{-1})$. Considering the better-performing diagnostics in the top pair of rows in Figure 1, they are all visually similar to one another. Interestingly, the behavior of $\text{CV}(\eta)$ very closely resembles SMUB(0.5) and relatively closely aligns with the value of SEM, as exhibited by the segments' close proximity to the $y = x$ line. The SMUB indices

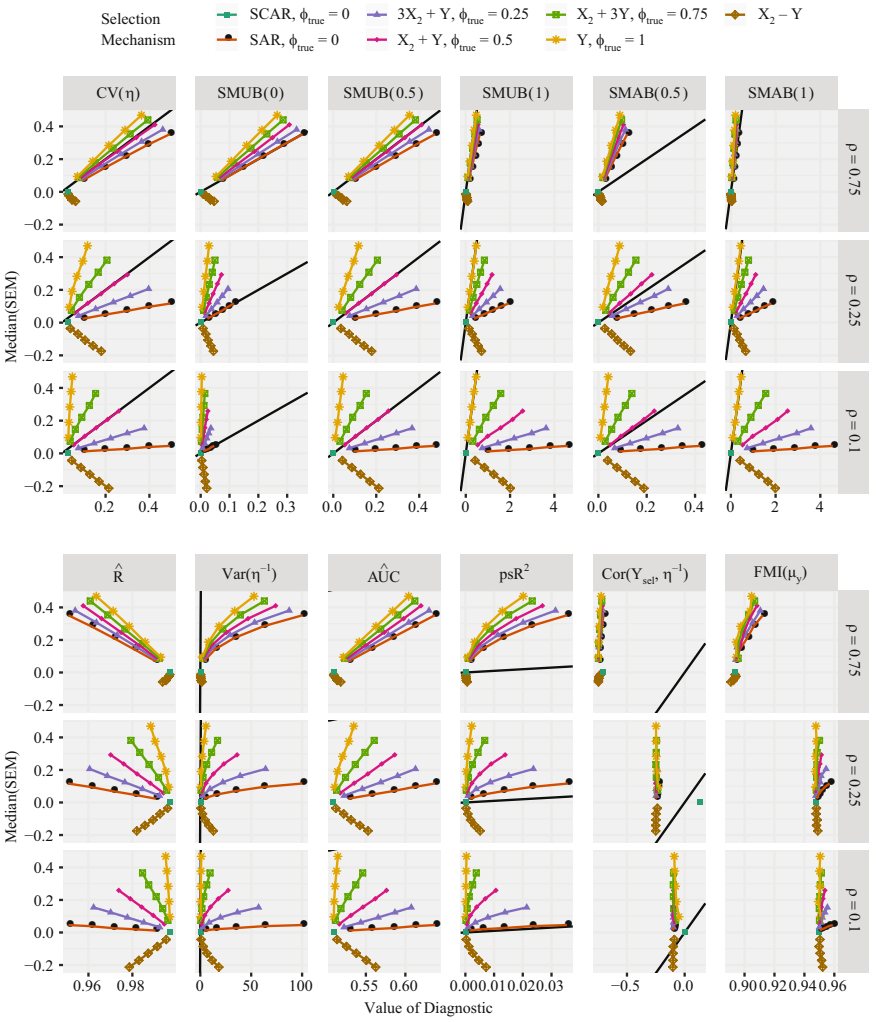

Fig. 1. Median standardized error measure (SEM, y-axes) against value of diagnostic (x-axes) for twelve candidate diagnostics (columns), three values of $\rho = Cor(X_1, Y)$ (rows) using the median of 2000 simulated data sets. $\kappa \equiv Cor(X_1, X_2)$ is fixed at 1 (Figures 2 and 3 give the same results for $\kappa = 0.5$ and $\kappa = 0$, respectively) For reference, the $y = x$ line is plotted in black. Shape and grayscale indicate different true selection mechanisms from Table 2, and connected segments represent different values of $\{\beta_x, \beta_y\}$ corresponding to the same selection mechanism.

generally increase with SEM in the $\rho = 0.75$ scenarios and, furthermore, are often nearly in 1-1 correspondence with SEM.

The extent to which this last statement is true depends upon the proximity between ϕ and ϕ_{true}, as the development of these estimators would suggest. The SMAB indices, which estimate the excess bias after adjusting for ignorable bias, vary little when $\rho = 0.75$, since in this case X_1 is actually a relatively good surrogate for Y, and are therefore less sensitive to SEM. When $\rho = 0.10$, most of the bias is non-ignorable, and so the SMUB and SMAB indices nearly correspond. For the third and sixth rows of Figure 1, the auxiliary

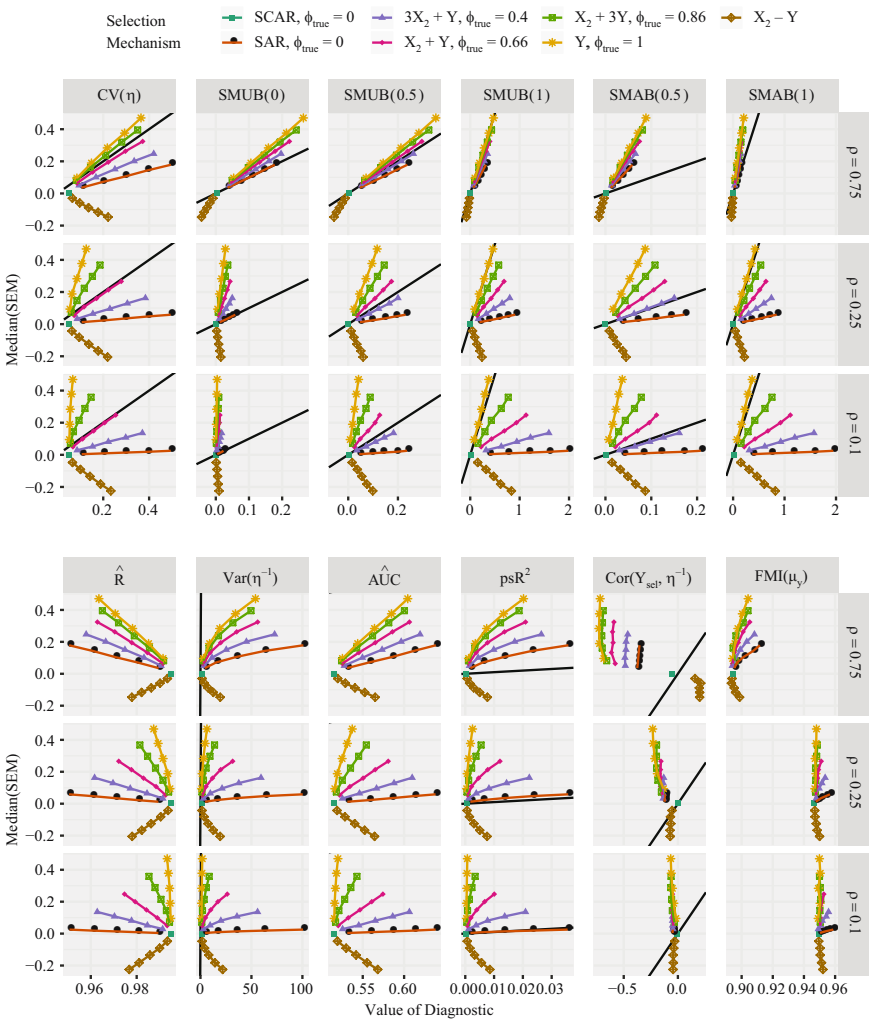

Fig. 2. Median standardized error measure (SEM, y-axes) against value of diagnostic (x-axes) for ten candidate diagnostics (columns), two values of $\rho = Cor(X_1, Y)$ (rows) using the median of 2000 simulated data sets. $\kappa = Cor(X_1, X_2)$ is fixed at 0.5 (Figures 1 and 3 give the same results for $\kappa = 1$ and $\kappa = 0$, respectively). For reference, the $y = x$ line is plotted in black. Shape and grayscale indicate different true selection mechanisms from Table 2, and connected segments represent different values of $\{\beta_x, \beta_y\}$ corresponding to the same selection mechanism.

variable is an especially poor predictor of the survey outcome ($\rho = 0.10$). In this setting, all the diagnostics show a wide scatter of values across the different selection mechanisms, suggesting that none of them are of much use in predicting the bias. This finding supports the statement in Little et al. (2019) that having an auxiliary variable that is a good predictor of the survey outcome is a key requirement for detecting bias.

Figures 2 and 3 illustrate how these diagnostics change when $\kappa < 1$, that is, when the auxiliary variable for the outcome and the auxiliary variable for selection differ. As expected, diagnostics that are based solely on the propensity, that is, CV(η), $\hat{AUC}$, $\hat{R}$, Var(η^{-1}), and

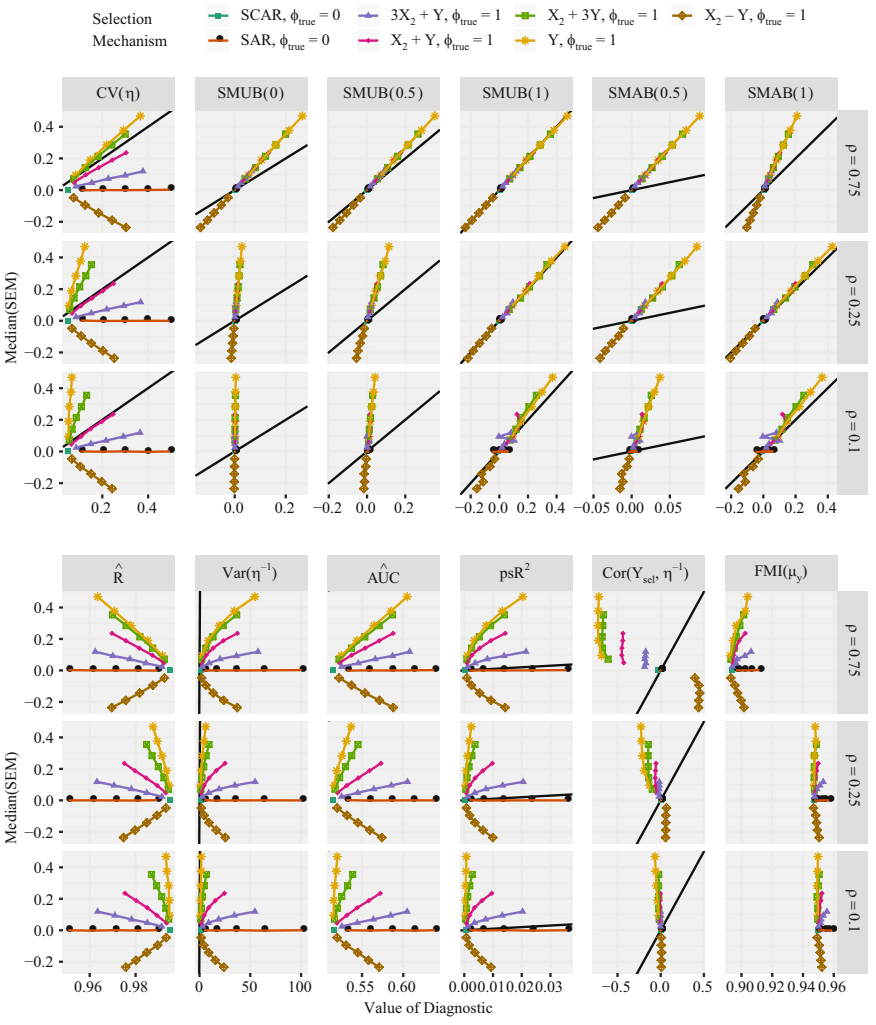

Fig. 3. Median standardized error measure (SEM, y-axes) against value of diagnostic (x-axes) for ten candidate diagnostics (columns), two values of $\rho \equiv Cor(X_1,Y)$ (rows) using the median of 2000 simulated data sets. $\kappa \equiv Cor(X_1,X_2)$ is fixed at 0 (Figures 1 and 2 give the same results for $\kappa = 1$ and $\kappa = 0.5$, respectively). For reference, the $y = x$ line is plotted in black. Shape and grayscale indicate different true selection mechanisms from Table 2, and connected segments represent different values of $\{\beta_x,\beta_y\}$ corresponding to the same selection mechanism.

psR^2, tend to falsely "detect" bias in these scenarios. False detection here means that segments are flat, varying in the x-value without any accompanying variation in the y-value. As noted in Table 2, smaller values of κ will increase the value of ϕ_{true} towards 1 when $\beta_y = 0$, causing SMUB(0) to underestimate SEM more so relative to the corresponding results in Figure 1. In the extreme case of $\kappa = 0$, which is given in Figure 3, the SMUB and SMAB indices are all nearly collinear. SMUB(1) looks most reasonable in this scenario because all selection mechanisms either have $\phi_{\text{true}} = 1$ (when $\beta_y \neq 0$) or $\phi_{\text{true}} = 0$ (when $\beta_y = 0$). In this latter case, all results fall on the origin, and there is no bias to detect.

Figures S1–S3 (Online Supplemental Material) give the analogous results using the alternative bias measure SAEM. Because SAEM is adjusted for ignorable bias, SMUB

now tends to overestimate bias and SMAB is the better-performing estimator. None of the other diagnostics considered perform qualitatively differently.

Figures 1–3 characterize the systematic relationship between SEM and each diagnostic, but there is also sampling variability that occurs within each data set. That is, does the realized value of a diagnostic in a given data set correspondingly change when the realized value of SEM is higher or lower than its mean? Table 3 reports the Spearman correlation (multiplied by 100) between each candidate diagnostic and the SEM value under seven selected sets of $\{\beta_x, \beta_y\}$ taken from Table 2 and three values of κ under $\rho = 0.75$. Those correlations that are within 5% of the largest magnitude correlation (the row-wise maximum absolute value) are in boldface. From Table 3, all of the metrics except $\text{Cor}(Y_{\text{sel}}, \eta^{-1})$ and $\text{FMI}(\mu_y)$ exhibit strong positive or negative correlation with SEM, that is, less than -0.6 or greater than 0.6, when $\kappa = 1$ and when the selection mechanism is not SCAR. However, as κ decreases, the Spearman correlations decrease or even change signs when the signs of β_x and β_y are in opposite directions. This even holds for $\text{CV}(\eta)$, which Figures 1–3 showed to be most sensitive to SEM on a systematic basis from among the existing diagnostics. For example, in the bottom-most three rows of Table 3, $\text{CV}(\eta)$ has a Spearman correlation with SEM of about 0.70 when $\kappa = 1$, but this decreases to -0.46 when $\kappa = 0$. Insofar as one does not know the true value of κ and thus whether to expect a positive or negative correlation with the error, this is problematic. The realized values of the SMUB measures do not exhibit this undesirable behavior, but rather exhibit a consistently high Spearman correlation with the realized values of the SEM.

Tables S1 and S2 in the Online Supplemental Material gives the analogous results under $\rho = 0.25$ and $\rho = 0.10$, respectively. When ρ is small, as in Table S2, none of the diagnostics, including SMUB or SMAB, have high correlation with SEM, highlighting the importance of obtaining auxiliary variables that correlate well with the outcome.

6. Discussion

Nishimura et al. (2016) found that none of their candidate diagnostics for detecting selection bias due to non-ignorable selection mechanisms were suitable for use. Our simulation study showed that the SMUB and SMAB family of measures proposed by Little et al. (2020) outperformed other diagnostics, both in terms of detecting the presence of bias as well as directly estimating its value, and both systematically (Figure 1–3) as well as on the basis of sampling variability (Table 3). The extent of non-ignorable selection is by definition inestimable, but the SMUB family is indexed by a tuning parameter ϕ, which allows the analyst to directly estimate the amount of selection bias by assuming that a specific degree of non-ignorable sampling had occurred. Our simulation study showed that the middle value of $\phi = 0.5$, which minimizes the maximum possible distance from ϕ_{true} and which Little et al. (2020) heuristically suggested for default use, resulted in a diagnostic that most consistently estimated the true amount of selection bias.

A number of additional qualities recommend the SMUB/SMAB family of statistics for the task of diagnosing and estimating selection bias. First, it correlates moderately well with the true measure of selection bias as evidenced in Table 3. Second, our simulation study demonstrates that the difference between the median values of the SMUB statistic and SEM was zero when the tuning parameter ϕ matched the unknown value ϕ_{true}. This

*Table 3. Spearman correlations (two significant digits; × 100) between each candidate diagnostic and the standardized error measure (SEM) for seven exemplar sets of $\{\beta_x,\beta_y\}$ taken from Table 2 and three values of $\kappa \equiv Cor(X_1, X_2)$ with $\rho \equiv Cor(X_1, Y)$ set to 0.75 (Table S1 in the Online Supplemental Material gives the same results with ρ set to 0.25). Those values in **bold** are within 5% of each row-wise maximum (in magnitude).*

$\{\beta_x, \beta_y\}$	κ	$\hat{R}$	Var(η^{-1})	CV(η)	AÛC	psR^2	Cor(Y_{sel}, η^{-1})	FMI(μ_y)	SMUB(0)	SMUB(0.5)	SMUB(1.0)	SMAB(0.5)	SMAB(1.0)
SCAR													
{0, 0}	1.0	-4	4	4	4	4	-51	-1	**73**	**73**	**73**	**73**	**73**
{0, 0}	0.5	-1	2	1	2	1	-67	2	**73**	**73**	**73**	**73**	**73**
{0, 0}	0.0	3	-4	-3	-4	-3	-66	-4	**72**	**72**	**72**	**72**	**72**
SAR													
{0.5, 0}	1.0	-67	62	**74**	**71**	**73**	25	9	70	70	65	51	47
{0.5, 0}	0.5	-35	30	38	37	38	-48	1	**68**	**68**	**67**	62	60
{0.5, 0}	0.0	-2	2	3	3	3	-61	3	**67**	**67**	**67**	**67**	**67**
$3X_2 + Y$													
{0.375, 0.125}	1.0	-68	63	**74**	**72**	**74**	25	5	**71**	**71**	66	53	49
{0.375, 0.125}	0.5	-41	39	45	44	45	-43	4	**69**	**69**	**68**	62	59
{0.375, 0.125}	0.0	-19	16	19	19	19	-60	-1	**67**	**67**	**67**	**66**	**65**
$X_2 + Y$													
{0.25, 0.25}	1.0	-65	62	**72**	**69**	**71**	25	8	**69**	68	63	51	48
{0.25, 0.25}	0.5	-58	55	61	59	61	-30	8	**70**	**70**	**68**	59	57
{0.25, 0.25}	0.0	-39	39	41	40	41	-53	3	**71**	**70**	**69**	65	64
$X_2 + 3Y$													
{0.125, 0.375}	1.0	-68	65	**72**	**70**	**72**	22	4	**70**	**69**	65	54	50
{0.125, 0.375}	0.5	**-66**	61	**69**	66	69	-4	5	**70**	**69**	**67**	58	55
{0.125, 0.375}	0.0	-62	59	65	63	65	-18	5	**71**	**71**	**68**	60	57
Y													
{0, 0.5}	1.0	-66	65	71	69	71	22	9	**69**	**69**	66	56	53
{0, 0.5}	0.5	-69	66	72	70	72	18	7	**71**	**71**	67	57	54
{0, 0.5}	0.0	-67	66	71	69	71	14	12	**69**	**69**	65	54	51
$X_2 - Y$													
{0.25, -0.25}	1.0	**-70**	**69**	**70**	68	**70**	-15	-1	**72**	**72**	**72**	**71**	**71**
{0.25, -0.25}	0.5	19	-19	-19	-19	-19	-67	-4	**72**	**72**	**72**	**72**	**72**
{0.25, -0.25}	0.0	44	-44	-46	-44	-46	-49	-4	**71**	**71**	**70**	66	64

result is consistent with the theoretical derivation of the SMUB. Third, the SMUB calculation does not require individual-level data from the non-sampled data but rather only summary statistics of the auxiliary variables, which makes it especially useful for non-probability samples and which stands in contrast to the other diagnostics evaluated. Fourth and finally, SMUB is specific to an estimand of interest, meaning that it will enable an analyst to order estimates computed from a non-probability sample in terms of their potential selection bias. Among those statistics considered in Nishimura et al. (2016), only the FMI(μ_y) and Cor(Y_{sel}, η^{-1}) statistic have this characteristic. In contrast, the values of all other potential diagnostics considered do not actually vary with the estimand. This fact alone arguably precludes from consideration any of the aforementioned diagnostics, insofar as it is impossible to expect a single statistic to serve as a universal diagnostic for bias with respect to an arbitrary estimand. Moreover, the FMI statistic focuses on variance rather than bias, and the simulation study clearly points to its deficiency as a diagnostic for bias.

Because the actual selection mechanism is unknown in practice, it is not sufficient to have a candidate diagnostic that correlates well with SEM under each selection mechanism. Rather, it must be correlated with SEM in the same way across many different selection mechanisms, since by definition of a non-probability sample, one does not know the true selection mechanism. Furthermore, high correlation between a diagnostic for selection bias and true selection bias is only useful if there is knowledge about the distribution of the diagnostic, or even just its support. For example, although psR^2 was consistently correlated with SEM, the values that we observed in the simulation study were typically limited to a very small interval close to zero, such that it would be difficult to know in practice whether one has encountered an extreme-enough value that would be suggestive of selection bias. The Var(η^{-1}) diagnostic is similarly limited: its range is arguably so extreme as to make it impractical for general use.

With regard to the other candidate diagnostics, our results were largely consistent with those reported in Nishimura et al. (2016). Because the only code from that paper that we used here was the function for calculating FMI(μ_y), our work largely represents an independent validation of their findings. Ironically, we found that the two statistics that make use of the greatest amount of data, Cor(Y_{sel}, η^{-1}) and FMI(μ_y), were actually among the least effective at detecting selection bias. We found that CV(η) generally had a high correlation with the true amount of selection bias, even under non-ignorable settings. Concerning, however, is its variation due to sampling variability, as demonstrated in Table 3.

Finally, a lack of a globally optimal value of the tuning parameter ϕ points to one possible and novel extension of the SMUB statistic. Although the ϕ_{true} is, by definition of a non-probability sample, inestimable, the sampling probabilities could be learned about, for example, with the collection of a small, auxiliary probability sample or via non-response follow-up with a small sample of non-selected cases, the non-ignorable bias could potentially be estimated and accounted for. Or, one might propose a shrinkage-type SMUB statistic that is an adaptive combination of estimates from the large, non-probability sample (high bias/low variance) and the small, probability sample (low bias/high variance), akin to the Empirical Bayes estimator of Mukherjee and Chatterjee (2008).

7. References

Albert, A., and J. Anderson. 1984. "On the existence of maximum likelihood estimates in logistic regression models." *Biometrika* 71: 1–10. DOI: https://doi.org/10.2307/2336390.

Andridge, R.R., and R.J. Little. 2011. "Proxy pattern-mixture analysis for survey nonresponse." *Journal of Official Statistics* 27: 153–180. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/proxy-pattern-mixture-analysis-for-survey-nonresponse.pdf (accessed May 2021).

Andridge, R.R., and R.J. Little. 2020. "Proxy pattern-mixture analysis for a binary variable subject to nonresponse." *Journal of Official Statistics*. DOI: https://doi.org/10.2478/jos-2020-0035.

Bootsma-van der Wiel, A.V., E. Van Exel, A. De Craen, J. Gussekloo, A. Lagaay, D. Knook, and R. Westendorp. 2002. "A high response is not essential to prevent selection bias: results from the leiden 85-plus study." *Journal of Clinical Epidemiology* 55: 1119–1125. DOI: https://doi.org/10.1016/s0895-4356(02)00505-x.

Brick, J.M., and D. Williams. 2013. "Explaining rising nonresponse rates in cross-sectional surveys." *The Annals of the American Academy of Political and Social Science* 645: 36–59. DOI: https://doi.org/10.1177%2F0002716212456834.

Heckman, J.J. 1979. "Sample selection bias as a specification error." *Econometrica* 47: 153–161. DOI: https://doi.org/10.2307/1912352.

Little, R.J. 1994. "A class of pattern-mixture models for normal incomplete data." *Biometrika* 81: 471–483. DOI: https://doi.org/10.2307/2337120.

Little, R.J., and D.B. Rubin. 2002. *Statistical Analysis with Missing Data*. John Wiley & Sons, Hoboken, NJ, 2nd edition.

Little, R.J., B.T. West, P. Boonstra, and J. Hu. 2020. "Measures of the degree of departure from ignorable sample selection." *Journal of Survey Statistics and Methodology* 8: 932–964. DOI:https://doi.org/10.1093/jssam/smz023.

Mukherjee, B., and N. Chatterjee. 2008. "Exploiting gene-environment independence for analysis of case-control studies: An empirical bayes-type shrinkage estimator to trade-off between bias and efficiency." *Biometrics* 64: 685–694. DOI: https://doi.org/10.1111/j.1541-0420.2007.00953.x.

Nagelkerke, N.J. 1991. "A note on a general definition of the coefficient of determination." *Biometrika* 78: 691–692. DOI: https://doi.org/10.1093/biomet/78.3.691.

Nishimura, R., J. Wagner, and M. Elliott. 2016. "Alternative indicators for the risk of nonresponse bias: a simulation study." *International Statistical Review* 84: 43–62. DOI: https://doi.org/10.1111/insr.12100.

Presser, S., and S. McCulloch. 2011. "The growth of survey research in the United States: Government-sponsored surveys, 1984–2004." *Social Science Research* 40: 1019–1024. DOI: https://doi.org/10.1016/j.ssresearch.2011.04.004.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rubin, D.B. 1976. "Inference and missing data." *Biometrika* 63: 581–592. DOI: https://doi.org/10.2307/2335739.

Rubin, D.B. 2004. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Särndal, C.-E., and S. Lundström. 2010. "Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias." *Survey Methodology* 36: 131–144.

Schouten, B., F. Cobben, J. Bethlehem, et al. 2009. "Indicators for the representativeness of survey response." *Survey Methodology* 35: 101–113.

Van Buuren, S., and K. Groothuis-Oudshoorn. 2011. "mice: Multivariate imputation by chained equations in R." *Journal of Statistical Software* 45: 1–67.

Wickham, H. 2017. *tidyverse: Easily install and load the 'tidyverse'*. R package version 1.2.1

Williams, D., and J.M. Brick. 2018. "Trends in US face-to-face household survey nonresponse and level of effort." *Journal of Survey Statistics and Methodology* 6: 186–211. DOI: https://doi.org/10.1093/jssam/smx019.

Received July 2019
Revised June 2020
Accepted November 2020