University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2019 *Paper* 125

A simulation study of diagnostics for bias in non-probability samples

Philip S. Boonstra[*] Roderick JA Little[†] Brady T. West[‡]

Rebecca R. Andridge[**] Fernanda Alvarado-Leiton[††]

[*]University Of Michigan, philb@umich.edu

[†]University of Michigan, rlittle@umich.edu

[‡]University of Michigan, bwest@umich.edu

[**]The Ohio State University, andridge.1@osu.edu

[††]University of Michigan, mleiton@umich.edu

A simulation study of diagnostics for bias in non-probability samples

Philip S. Boonstra, Roderick JA Little, Brady T. West, Rebecca R. Andridge, and Fernanda Alvarado-Leiton

Abstract

A non-probability sampling mechanism 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’, i.e. dependent upon the unobserved outcome of interest, since it is then undetectable and thus cannot be ameliorated. We extend a simulation study by Nishimura et al. [*International Statistical Review*, 84, 43–62 (2016)], adding a recently published statistic, the so-called ’standardized measure of unadjusted bias’, which explicitly quantifies the extent of bias under the assumption that a specified amount of non-ignorable selection exists. Our findings suggest that this new sensitivity diagnostic is considerably 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.

A simulation study of diagnostics for bias in non-probability samples

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

March 11, 2019

Summary

A non-probability sampling mechanism 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', i.e. dependent upon the unobserved outcome of interest, since it is then undetectable and thus cannot be ameliorated. We extend a simulation study by Nishimura et al. [*International Statistical Review*, 84, 43–62 (2016)], adding a recently published statistic, the so-called 'standardized measure of unadjusted bias', which explicitly quantifies the extent of bias under the assumption that a specified amount of non-ignorable selection exists. Our findings suggest that this new sensitivity diagnostic is considerably 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.

[*] 1415 Washington Heights, Ann Arbor, Michigan, USA, 48109-2029; Tel: +1 734 615 1580; philb@umich.edu

Key words: Non-Ignorable Selection Bias; Survey Non-Response; Multiple Imputation

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, etc.; Brick and Williams, 2013; Williams and Brick, 2018), and the costs of 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 be treated as probability samples 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 researchers 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 non-probability sampling 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. (2019) 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 paper 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 paper is organized as follows. Section 2 presents notation and a brief description of the index of selection bias proposed in Little et al. (2019). Section 3 describes the other diagnostics we consider here, which were also evaluated in Nishimura et al. (2016). Sections 4 and 5 describe and present the results from the simulation study, respectively. And 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 $i = 1, \dots, N$, 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 $S = \{S_1, \dots, S_N\}$ and $Z = \{Z_1, \dots, Z_N\}$ are fully observed, and the vector $Y = \{Y_1, \dots, Y_N\}$ 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 or, adapting the terminology of Little and Rubin (2002), '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 $\theta_{\mathrm{unsel}} = \theta_{\mathrm{sel}} \equiv \theta$ and $\{\theta, \pi\}$ and $\{\alpha, \beta\}$ share a 1-1 correspondence (Little, 1994). Further, all parameters become identified in this special case. However, models (1) and (2) will not generally coincide under SAR for any distributional choices. Although the decomposition in (1) is more intuitive by directly capturing the data-generating mechanism, the usefulness in focusing on (2) is that the non-identified parameters are isolated to a single submodel: $[Y_{\text{unsel}}, Z_{\text{unsel}}|\theta_{\mathrm{unsel}}]$. In the pattern-mixture framework, the estimand of interest, μ_y, is equal to $\pi E[Y_{\text{sel}}|\theta_{\mathrm{sel}}] + (1-\pi)E[Y_{\text{unsel}}|\theta_{\mathrm{unsel}}]$. The latter mean, $E[Y_{\text{unsel}}|\theta_{\mathrm{unsel}}]$, is not identified without making further assumptions.

Specifically, for the factorization in (2), assume that $[Z_{\text{sel}}, Y_{\text{sel}}|\theta_{\mathrm{sel}}]$ and $[Z_{\text{unsel}}, Y_{\text{unsel}}|\theta_{\mathrm{unsel}}]$ 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

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

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 (Andridge and Little, 2011; Little, 1994), and a maximum likelihood estimate (MLE) of μ_y as a function of ϕ is given by

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

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. See Andridge and Little (2011) for the derivation of this estimator.

Remark: Little et al. (2019) show that the estimator (4) remains unbiased for its estimand under a more general class of functions than that given in (3), namely

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

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 (4) is a function of the parameter ϕ, which in turn controls the extent to which sampling depends upon the outcome Y, with larger values indicating greater dependence. When $\phi=0$, the selection mechanism is SAR, and the resulting diagnostic is closely related to the measure H1 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, Little, et al. propose varying this parameter in a sensitivity analysis. Subtracting $\bar{y}_{\text{sel}}$ from both sides, we obtain a direct estimate of the 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., 2019):

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

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. The simulation study in Nishimura et al. (2016), prior to the proposal of the estimator in (6), found that “none of the indicators [evaluated] fully depict the impact of non-response in survey estimates.” (p.43) We consider here whether the SMUB index addresses this deficiency. Note that (6) is based on a normal pattern-mixture model, and as such is less well suited to non-normal outcomes. Modifications of (6) for a categorical outcomes are discussed in Andridge and Little (2009) but are not considered in this article.

3 Other Diagnostics Evaluated

Nishimura et al. (2016) grouped the diagnostics they compared based upon whether $\{S, Z\}$ or $\{S, Y_{\text{sel}}, Z\}$ are required to calculate them.

The simplest diagnostic is $\bar{s}$, i.e. 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, e.g. 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.

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, i.e. 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

$$\eta_i \equiv \text{logit}^{-1}(\hat{\gamma}_0 + \hat{\gamma}_z Z_i). \tag{7}$$

The R-Indicator (Schouten et al., 2009), where R stands for 'response', which is 'selection' in our notation, is the following linear transformation of the sample standard deviation of η_i across both the selected and un-selected samples:

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

It theoretically ranges from 0 to 1, where smaller values correspond to greater variability in the selection propensities and, consequently, more potential for selection bias. However, the smallest possible value $\hat{R}=0$, i.e. 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 $\eta_i = 0$, i.e. S can be completely separated by Z, in the sense of Albert and Anderson (1984). In practice, $\hat{R}=1$ 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:

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

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.

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 common '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:

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

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

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

3.2 Diagnostics using $\{S, Y_{\textbf{sel}}, Z\}$

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:

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

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 $\{Y_{\mathrm{sel}}, Z_{\mathrm{sel}}\}$, M sets of unselected outcomes, denoted by $Y_{\mathrm{unsel}}^{(m)}$ are imputed. Each of the M completed datasets, $\{Y_{\mathrm{sel}}, Y_{\mathrm{unsel}}^{(m)}\}$ 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

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

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 so-called "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 so-called "within-imputation variance", and the sum of the between- and within-imputation

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

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.

4 Simulation Study: Description

The purpose of this simulation study is to characterize the association between the true bias in a sampled dataset (only observable in a simulation framework) and each of the aforementioned candidate diagnostics, including the new SMUB diagnostic from Little et al. (2019). 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 extended the simulations in Nishimura et al. (2016) by generating 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.

In more detail, at each iteration, a super-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, i.e. $\kappa < 1$, both 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, i.e. $[Y|X_1] = N(\rho X_1, \sqrt{1-\rho^2})$; 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 $\Pr(S=1|Y,X_2) = \text{logit}^{-1}(\beta_0 + \beta_y Y + \beta_x X_2)$. In total, there are five parameters governing this distribution: κ, ρ, β_0, β_x, and β_y.

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 $\rho \in \{0.25, 0.75\}$. 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. In total, Table 2 gives 36 unique sets of β_x and β_y.

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

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

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) $W = \beta_x\sqrt{1-\kappa^2}\epsilon/(\kappa\beta_x + \beta_y)$, the relaxed assumption (5) is satisfied for any $\kappa \in [0,1]$. In contrast, the more restrictive assumption (3) is only satisfied for $\kappa = 1$, i.e. $W \equiv 0$. Under $\kappa = 1$, the third column in Table 2 give the implied true value of ϕ, which is

Table 2: Values 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 selection mechanism is 'selected completely at random' (SCAR) in the first row, 'selected at random' (SAR) in the second row, and 'selected not at random' in the remaining rows. The implied true value of the non-ignorability parameter ϕ is calculated by the expression $\phi_{\text{true}} = \beta_y/(\kappa\beta_x + \beta_y)$.

		ϕ_{true}		
Label	$\{\beta_x, \beta_y\}$	$\kappa = 1$	$\kappa = 0.5$	$\kappa = 0$
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
$-X_2 + Y$	$\{-c/2, c/2\}$	$-^\dagger$	$-^\dagger$	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 $\kappa > 0$.

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 final two rows of Table 2, for which $\beta_x > 0$ and $\beta_y < 0$ (or vice versa), there is no value of $\phi_{\text{true}} \in [0, 1]$ satisfying (5) except for the case that $\kappa = 0$, and this is noted as such in the table.

Finally, 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 $E_{Y,X_2}\left[\text{logit}^{-1}(\beta_0 + \beta_x X_2 + \beta_y Y)\right]$, can then be numerically solved for β_0. Our choice of a 5% selection rate is a fairly large selection rate for a 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 (6): $\phi \in \{0, 0.5, 1.0\}$. When ϕ is close to the unknown ϕ_{true}, the SMUB estimate will closely match the actual observed bias. And for FMI, we imputed $M = 30$ vectors of the unselected outcomes

Y_{unsel} to estimate μ_y.

We simulated 2000 independent datasets for each of the $2 \times 3 \times 36 = 216$ combinations of ρ, κ, and $\{\beta_x, \beta_y\}$ pair taken from Table 2. 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. To assess performance, we calculated for each dataset the 'standardized estimated bias' (SEB) in using $\bar{y}_{\text{sel}}$ to estimate μ_y, which is given by

$$\text{SEB} = \frac{\bar{y}_{\text{sel}} - \mu_y}{\sigma_y}. \tag{8}$$

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 SEB 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 SEB, and both the qualitative and quantitative nature of this association should be similar for all types of selection mechanisms, i.e. 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 SEB and each diagnostic across all 2000 datasets from each scenario. 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`.

5 Simulation Study: Results

Figures 1, 2, and 3 plot the relationship between the median value of SEB across 2000 simulated datasets from a given scenario against the median of each diagnostic, separately for $\kappa = 1$, 0.5, and 0, respectively. In addition, Figure S1 in the Supplement gives the contents of Figures 1, 2, and 3 overlaid onto a single plot, using different levels of transparency to distinguish between values of κ and therefore allowing for a more direct assessment of the impact of changing κ.

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, i.e. 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, i.e. 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). 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 SEB 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 five in the top two rows (one row each for $\rho = 0.75$ and $\rho = 0.25$) roughly corresponding to the best performing diagnostics, and the set in the bottom two rows corresponding to the worst performing diagnostics.

Considering first the diagnostics in the bottom rows of Figure 1, $\mathrm{Cor}(Y_{\mathrm{sel}}, \eta^{-1})$ and $\mathrm{FMI}(\mu_y)$ are not notably sensitive to changes in SEB, as indicated by the steep vertical segments. The $\mathrm{Var}(\eta^{-1})$ diagnostic changes with SEB, but the range of its x-axis is very wide, potentially limiting interpretability as to what constitutes an extreme value. The $\hat{R}$ and $\mathrm{ps}R^2$ diagnostics are also sensitive to SEB and, unlike $\mathrm{Var}(\eta^{-1})$, have a narrow range of the x-axis. 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

$\mathrm{CV}(\eta)$ very closely resembles $\mathrm{SMUB}(0.5)$ and relatively closely aligns with the value of SEB, as exhibited by the segments' close proximity to the $y = x$ line. The SMUB indices, which are in the right-most three columns, generally increase with SEB in the $\rho = 0.75$ scenarios and, furthermore, are often nearly in 1-1 correspondence with SEB. 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. For the second and fourth rows of Figure 1, the auxiliary variable is a relative poor predictor of the survey outcome ($\rho = 0.25$). In this setting, all the diagnostics show a wide scatter of SEB 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.

Figure 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, $\mathrm{CV}(\eta)$, $\mathrm{A\hat{U}C}$, $\hat{R}$, $\mathrm{Var}(\eta^{-1})$, and $\mathrm{ps}R^2$, 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 as long as $\beta_y \neq 0$, causing $\mathrm{SMUB}(0)$ to underestimate SEB more so relative to the corresponding results in Figure 1 and causing $\mathrm{SMUB}(1)$ to be a relatively better estimator of SEB than in 1. In the extreme case of $\kappa = 0$, which is given in Figure 3, the SMUB indices are all nearly collinear. $\mathrm{SMUB}(1)$ 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. Figure S1 (in the Supplement) gives results for all values of κ overlaid on a single panel to illustrate the change in behavior as κ goes from 1 to 0.5 to 0.

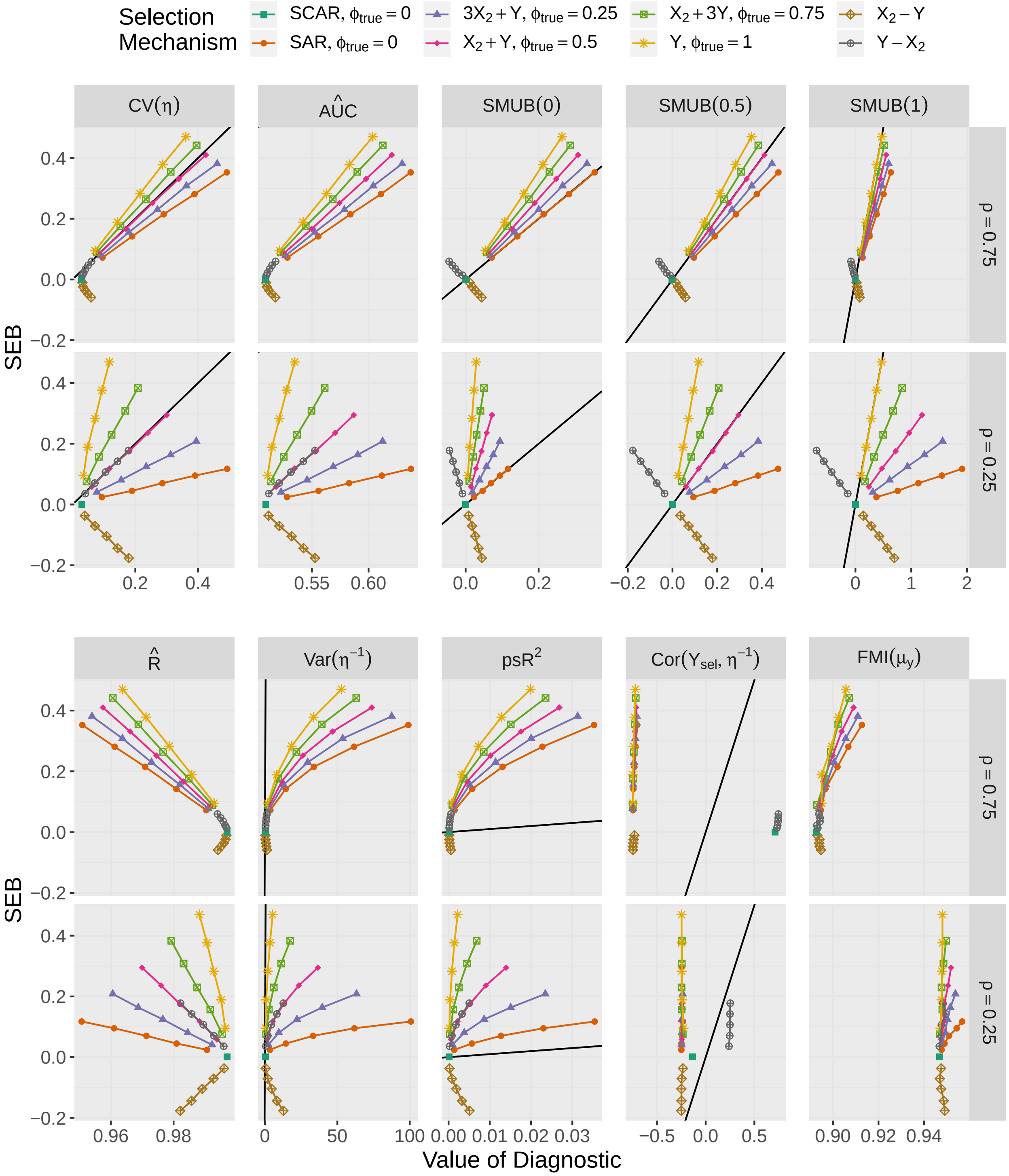

Figure 1: Standardized estimated bias (SEB, y-axes) against value of diagnostic (x-axes) for ten candidate diagnostics (columns), two values of $\rho \equiv \mathrm{Cor}(X_1, Y)$ (rows) using the median of 2000 simulated datasets. $\kappa \equiv \mathrm{Cor}(X_1, X_2)$ is fixed at 1 (Figures 2 and 3 give the same results for $\kappa = 0.5$ and $\kappa = 0$, respectively) For reference, the $y = x$ line is plotted in black. Shape and color 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.

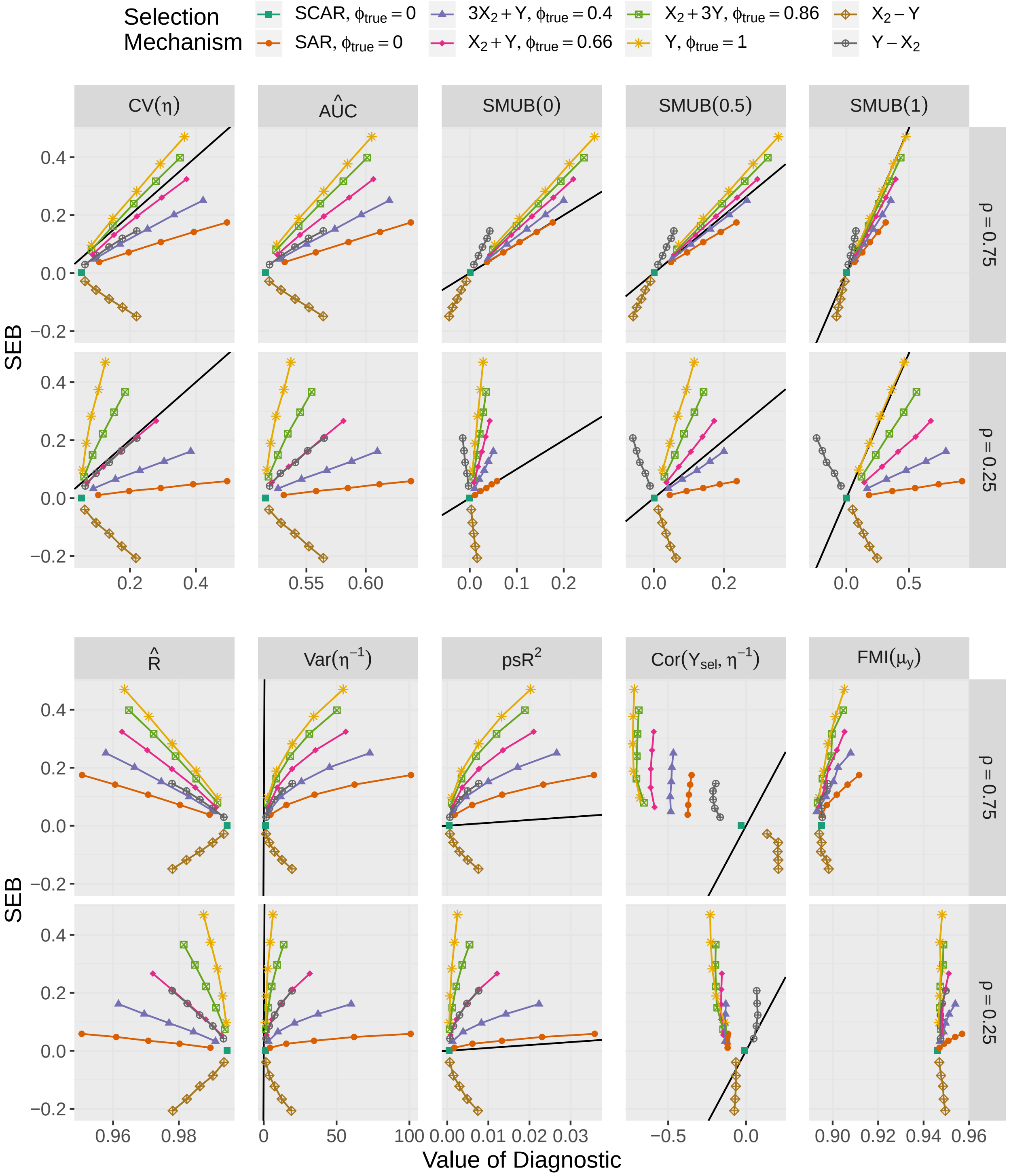

Figure 2: Standardized estimated bias (SEB, y-axes) against value of diagnostic (x-axes) for ten candidate diagnostics (columns), two values of $\rho \equiv \mathrm{Cor}(X_1, Y)$ (rows) using the median of 2000 simulated datasets. $\kappa \equiv \mathrm{Cor}(X_1, X_2)$ is fixed at 0.5 (Figures 1 and 3 give the same results for $\kappa = 1$ and $\kappa = 0$, respectively). For reference, the $y = x$ line is plotted in black. Shape and color 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.

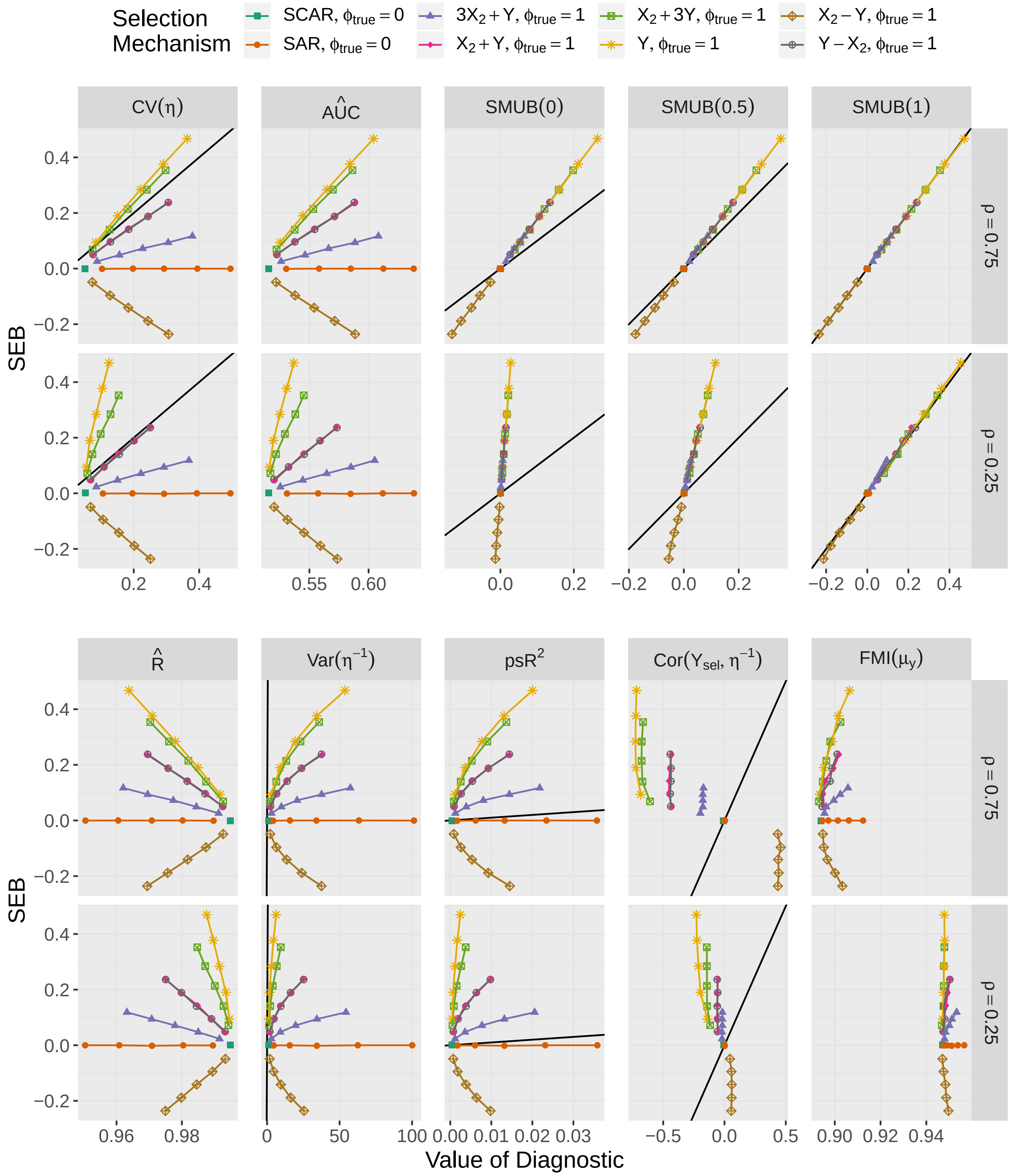

Figure 3: Standardized estimated bias (SEB, y-axes) against value of diagnostic (x-axes) for ten candidate diagnostics (columns), two values of $\rho \equiv \text{Cor}(X_1, Y)$ (rows) using the median of 2000 simulated datasets. $\kappa \equiv \text{Cor}(X_1, X_2)$ is fixed at 0 (Figures 1 and 2 give the same results for $\kappa = 1$ and $\kappa = 0.5$, respectively). For reference, the $y = x$ line is plotted in black. Shape and color 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.

Figures 1–3 characterize the systematic relationship between SEB and each diagnostic, but there is also sampling variability that occurs within each dataset. That is, does the realized value of a diagnostic in a given dataset correspondingly change when the realized value of SEB is higher or lower than its mean? Table 3 reports the Spearman correlation (multiplied by 100) between each candidate diagnostic and the SEB value under eight 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 are in boldface. Table S1 in the Supplement gives the analogous results under $\rho = 0.25$. From Table 3, all of the metrics except $\mathrm{Cor}(Y_{\text{sel}}, \eta^{-1})$ and $\mathrm{FMI}(\mu_y)$ exhibit strong positive or negative correlation with SEB, i.e. less than -0.6 or greater than 0.6, when $\kappa = 1$. 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 $\mathrm{CV}(\eta)$, which Figures 1–3 showed to be most sensitive to SEB on a systematic basis from among the existing diagnostics. For example, in the bottom-most three rows of Table S1, $\mathrm{CV}(\eta)$ has a Spearman correlation with SEB of about -0.72 when $\kappa = 1$, but this increases to 0.44 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 bias, 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 SEB.

Table 3: Spearman correlations (two significant digits; $\times 100$) between each candidate diagnostic and the standardized estimated bias (SEB) for eight exemplar sets of $\{\beta_x, \beta_y\}$ taken from Table 2 and three values of $\kappa \equiv \mathrm{Cor}(X_1, X_2)$ with $\rho \equiv \mathrm{Cor}(X_1, Y)$ set to 0.75 (Table S1 in the Supplement 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	$\hat{R}$	$\mathrm{Var}(\eta^{-1})$	$\mathrm{CV}(\eta)$	$\mathrm{A\hat{U}C}$	$\mathrm{ps}R^2$	$\mathrm{Cor}(Y_{\mathrm{sel}}, \eta^{-1})$	$\mathrm{FMI}(\mu_y)$	SMUB(0)	SMUB(0.5)	SMUB(1.0)
SCAR											
$\{0, 0\}$	1.0	4	-4	-4	-3	-4	-52	0	**72**	**72**	**72**
$\{0, 0\}$	0.5	1	-1	-1	-1	-1	-67	1	**73**	**73**	**73**
$\{0, 0\}$	0.0	-1	1	1	0	1	-66	2	**74**	**74**	**74**
SAR											
$\{0.5, 0\}$	1.0	-66	58	**72**	**70**	**71**	26	11	**69**	**68**	63
$\{0.5, 0\}$	0.5	-33	29	35	34	35	-53	8	**70**	**70**	**68**
$\{0.5, 0\}$	0.0	-1	1	2	2	2	-59	-0	**65**	**65**	**65**
$3X_2 + Y$											
$\{0.375, 0.125\}$	1.0	-65	61	**71**	**68**	**71**	25	7	**68**	**68**	63
$\{0.375, 0.125\}$	0.5	-46	40	49	48	49	-45	5	**70**	**70**	**68**
$\{0.375, 0.125\}$	0.0	-19	17	19	19	20	-62	-0	**70**	**70**	**70**
$X_2 + Y$											
$\{0.25, 0.25\}$	1.0	-67	64	**73**	**70**	**72**	26	6	**70**	**69**	65
$\{0.25, 0.25\}$	0.5	-55	53	59	58	59	-31	7	**71**	**71**	**68**
$\{0.25, 0.25\}$	0.0	-41	41	44	42	43	-52	4	**71**	**71**	**70**
$X_2 + 3Y$											
$\{0.125, 0.375\}$	1.0	-67	64	**72**	**70**	**72**	22	7	**71**	**70**	66
$\{0.125, 0.375\}$	0.5	-66	61	**69**	67	**69**	-7	8	**71**	**70**	67
$\{0.125, 0.375\}$	0.0	-64	62	67	65	67	-21	3	**71**	**71**	**69**
Y											
$\{0, 0.5\}$	1.0	-72	69	**76**	**74**	**76**	20	9	**74**	**74**	70
$\{0, 0.5\}$	0.5	-69	65	**73**	**71**	**73**	11	4	**72**	**71**	68
$\{0, 0.5\}$	0.0	-71	67	**74**	**72**	**74**	16	3	**73**	**72**	69
$X_2 - Y$											
$\{0.25, -0.25\}$	1.0	**-71**	**71**	**71**	68	**71**	-15	4	**73**	**73**	**73**
$\{0.25, -0.25\}$	0.5	24	-25	-25	-25	-25	-64	-3	**70**	**70**	**70**
$\{0.25, -0.25\}$	0.0	42	-40	-43	-42	-43	-53	-5	**72**	**72**	**71**
$Y - X_2$											
$\{-0.25, 0.25\}$	1.0	**71**	**-71**	**-72**	-70	**-72**	-16	1	**74**	**74**	**74**
$\{-0.25, 0.25\}$	0.5	-15	15	16	15	16	-68	-0	**72**	**72**	**72**
$\{-0.25, 0.25\}$	0.0	-42	41	44	43	44	-53	8	**73**	**73**	**72**

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 measure proposed by Little et al. (2019) 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. (2019) 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 family of statistics for the task of diagnosing and estimating selection bias. First, it correlates moderately well with the true measure of selection bias, given by the value of SEB, even when the underlying assumptions about the structural form of non-ignorability were violated, i.e. the last two rows of Table 2. Second, our simulation study demonstrates that the difference between the median values of the SMUB statistic and SEB was zero when the tuning parameter ϕ matched the unknown value ϕ_{true}. This result is consistent with the theoretical derivation of the SMUB. Third 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 statistic has 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 SEB under each selection mechanism. Rather, it must be correlated with SEB 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 $\mathrm{ps}R^2$ was consistently correlated with SEB, 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 $\mathrm{Var}(\eta^{-1})$ 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 $\mathrm{FMI}(\mu_y)$, 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, $\mathrm{Cor}(Y_{\mathrm{sel}}, \eta^{-1})$ and $\mathrm{FMI}(\mu_y)$, were actually among the least effective at detecting selection bias. We found that $\mathrm{CV}(\eta)$, $\hat{\mathrm{AUC}}$, and $\mathrm{ps}R^2$ had generally high correlation with the true amount of selection bias, even under non-ignorable settings. Concerning, however, is the variation of these diagnostics 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, e.g. 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).

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 Chen Chen for his initial work on this simulation study.

Disclosure

The authors report no potential conflicts of interest.

References

Albert, A. and Anderson, J. (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika **71,** 1–10.

Andridge, R. R. and Little, R. J. (2009). Extensions of proxy pattern-mixture analysis for survey nonresponse. In Proceedings of the 2009 Joint Statistical Meetings, Section on Survey Research Methods, pages 2468–2482.

Andridge, R. R. and Little, R. J. (2011). Proxy pattern-mixture analysis for survey nonresponse. Journal of Official Statistics **27,** 153–180.

Bootsma-van der Wiel, A. v., Van Exel, E., De Craen, A., Gussekloo, J., Lagaay, A., Knook, D., and Westendorp, R. (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.

Brick, J. M. and Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. The Annals of the American Academy of Political and Social Ccience **645,** 36–59.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. Biometrika **81,** 471–483.

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

Little, R. J. A., West, B. T., Boonstra, P. S., and Hu, J. (2019). Measures of the degree of departure from ignorable sample selection. Journal of Survey Statistics and Methodology **To Appear,**.

Mukherjee, B. and Chatterjee, N. (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.

Nagelkerke, N. J. et al. (1991). A note on a general definition of the coefficient of determination. Biometrika **78,** 691–692.

Nishimura, R., Wagner, J., and Elliott, M. (2016). Alternative indicators for the risk of non-response bias: a simulation study. International Statistical Review **84,** 43–62.

Presser, S. and McCulloch, S. (2011). The growth of survey research in the united states: government-sponsored surveys, 1984–2004. Social Science Research **40,** 1019–1024.

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

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

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

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

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

van Buuren, S. and Groothuis-Oudshoorn, K. (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 Brick, J. M. (2018). Trends in us face-to-face household survey nonresponse and level of effort. Journal of Survey Statistics and Methodology **6,** 186–211.

Supplement (Online Only)

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

$\{\beta_x, \beta_y\}$	κ	$\hat{R}$	$\mathrm{Var}(\eta^{-1})$	$\mathrm{CV}(\eta)$	$\mathrm{A\hat{U}C}$	$\mathrm{ps}R^2$	$\mathrm{Cor}(Y_{\mathrm{sel}}, \eta^{-1})$	$\mathrm{FMI}(\mu_y)$	SMUB(0)	SMUB(0.5)	SMUB(1.0)
SCAR											
$\{0, 0\}$	1.0	-1	2	1	2	1	-17	-2	**25**	**26**	**26**
$\{0, 0\}$	0.5	-1	1	1	3	1	-24	-1	**26**	**27**	**27**
$\{0, 0\}$	0.0	-2	2	2	2	2	-19	-2	**22**	**22**	**22**
SAR											
$\{0.5, 0\}$	1.0	-23	20	**24**	**24**	**24**	-4	4	16	**23**	8
$\{0.5, 0\}$	0.5	-8	10	10	**11**	10	-6	2	10	9	7
$\{0.5, 0\}$	0.0	2	-1	-2	-1	-2	-10	3	**12**	**12**	**11**
$3X_2 + Y$											
$\{0.375, 0.125\}$	1.0	-23	23	**26**	**25**	**25**	3	4	12	**25**	14
$\{0.375, 0.125\}$	0.5	-11	10	11	11	11	-12	1	15	**17**	**16**
$\{0.375, 0.125\}$	0.0	3	-1	-3	-3	-3	**-15**	-0	**15**	**15**	**15**
$X_2 + Y$											
$\{0.25, 0.25\}$	1.0	-17	**19**	**19**	18	**19**	1	3	11	18	12
$\{0.25, 0.25\}$	0.5	-14	12	14	14	14	-10	6	14	**17**	16
$\{0.25, 0.25\}$	0.0	-5	6	6	5	6	**-14**	-3	**14**	**14**	13
$X_2 + 3Y$											
$\{0.125, 0.375\}$	1.0	-28	**30**	**30**	**29**	**30**	1	1	22	**30**	23
$\{0.125, 0.375\}$	0.5	-16	15	16	15	16	-14	2	**22**	**23**	20
$\{0.125, 0.375\}$	0.0	-14	14	14	14	14	-15	0	**20**	**21**	20
Y											
$\{0, 0.5\}$	1.0	**-23**	**24**	**24**	**23**	**23**	-2	1	22	**24**	21
$\{0, 0.5\}$	0.5	**-22**	21	**21**	21	**22**	-8	3	21	**22**	20
$\{0, 0.5\}$	0.0	**-22**	**22**	**23**	**22**	**23**	-2	2	19	22	20
$X_2 - Y$											
$\{0.25, -0.25\}$	1.0	**-24**	**24**	**25**	**24**	**25**	2	2	19	**25**	21
$\{0.25, -0.25\}$	0.5	-5	4	5	4	5	-20	-0	**21**	**22**	**21**
$\{0.25, -0.25\}$	0.0	2	-2	-3	-3	-3	**-18**	1	**18**	**18**	**18**
$Y - X_2$											
$\{-0.25, 0.25\}$	1.0	**24**	**-25**	**-25**	**-24**	**-25**	0	0	20	**24**	20
$\{-0.25, 0.25\}$	0.5	11	-10	-10	-9	-10	-13	2	16	**17**	**17**
$\{-0.25, 0.25\}$	0.0	-5	6	6	5	6	-18	5	**19**	**20**	**19**

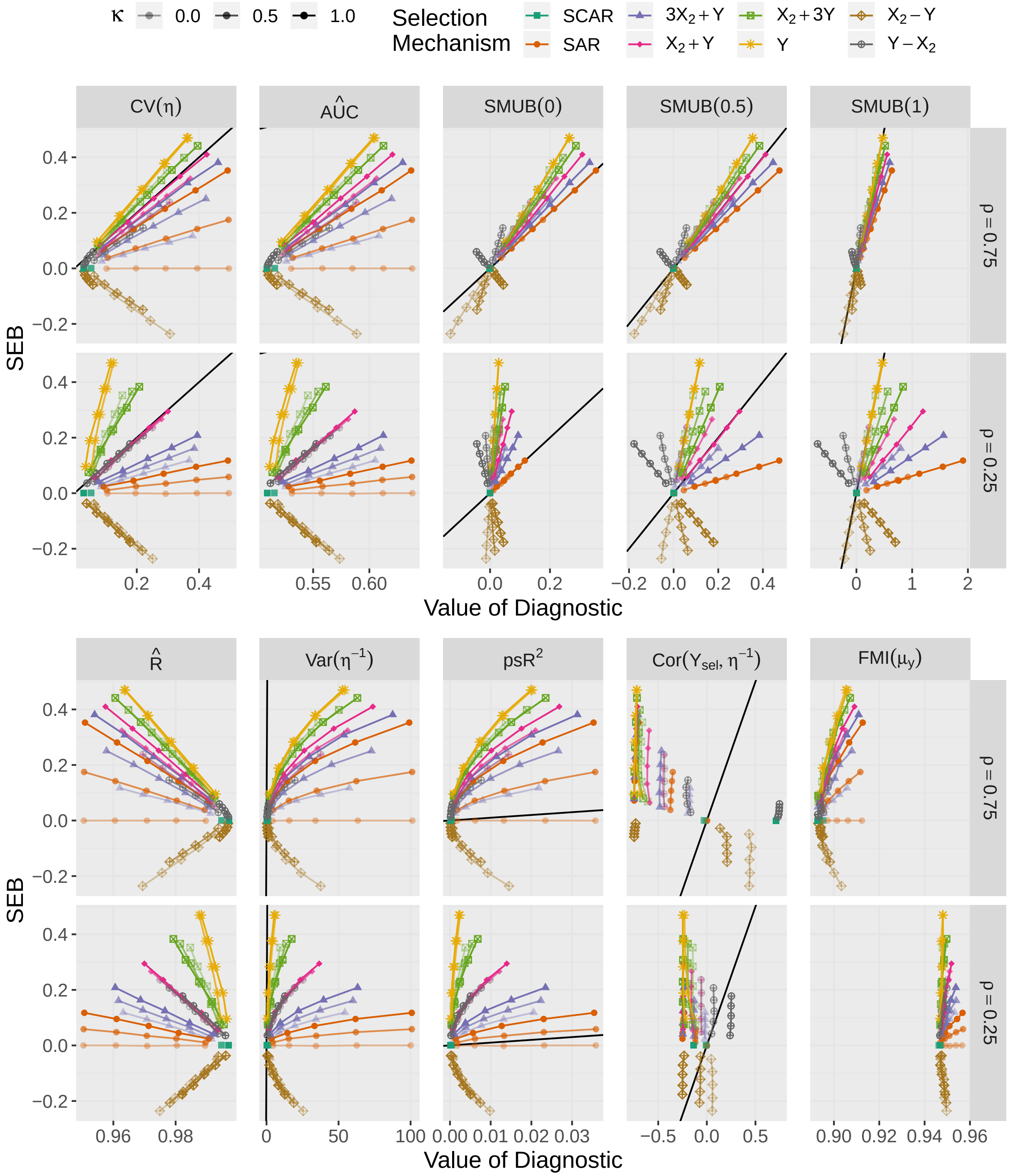

Figure S1: Standardized estimated bias (SEB, y-axes) against value of diagnostic (x-axes) for ten candidate diagnostics (columns), two values of $\rho = \mathrm{Cor}(X_1, Y)$ (rows) using the median of 2000 simulated datasets. Here, $\kappa \equiv \mathrm{Cor}(X_1, X_2)$ is varying with the degree of transparency, and Figures 1–3 in the manuscript give the same results separately for each value of κ. For reference, the $y = x$ line is plotted in black. Shape and color indicate different true selection mechanisms from Table 2 in the manuscript, and connected segments represent different values of $\{\beta_x, \beta_y\}$ corresponding to the same selection mechanism.