Partial Verification Bias Correction Using Scaled Inverse Probability Resampling for Binary Diagnostic Tests

Wan Nor Arifin[1,2*], Umi Kalsom Yusof[2]

1 Biostatistics and Research Methodology Unit, School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia
2 School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia

* wnarifin@usm.my

Abstract

Diagnostic accuracy studies are crucial for evaluating new tests before their clinical application. These tests are compared to gold standard tests to obtain relevant accuracy measures, most often sensitivity (Sn) and specificity (Sp) for binary diagnostic tests. However, these studies are often biased due to selective verification of patients, known as partial verification bias (PVB). PVB eventually leads to biased accuracy estimates in such studies. Among methods developed for PVB correction under the missing at random assumption for binary diagnostic tests, a bootstrap-based method known as the inverse probability bootstrap (IPB) was proposed. Despite showing low bias for estimating Sn and Sp, the IPB method exhibited higher standard errors than other PVB correction methods. This paper introduces two new methods: scaled inverse probability weighted resampling (SIPW) and scaled inverse probability weighted balanced resampling (SIPW-B), which build upon the IPB approach. Through simulations and clinical data, SIPW and SIPW-B were compared against IPB and other methods. Results demonstrated that the new methods outperformed IPB by showing lower bias and standard errors in Sn and Sp estimation. Specifically, SIPW-B outperformed IPB in Sn estimation, while SIPW performed better in Sp estimation, particularly when disease prevalence is low. These methods offer advantages such as complete data restoration and calculations independent of disease prevalence. Although computationally demanding, this limitation becomes less significant with the increasing power of modern computing resources.

Introduction

Diagnostic tests are crucial in medical care, so ensuring their clinical validity through diagnostic accuracy studies is essential [1, 2]. These studies involve comparing a new test with an established gold standard to evaluate its performance using accuracy measures like sensitivity (Sn) and specificity (Sp) for binary tests [1, 3–5]. However, verifying disease status with the gold standard can be expensive, time-consuming, and invasive [1, 5–8]. This verification challenge often leads to partial verification bias (PVB), where patients with positive test results are predominantly selected for gold standard verification, while fewer patients with negative test results are verified [1, 6, 8, 9]. This gives rise to a missing at random (MAR) missing data mechanism, as the decision to verify depends on the result of the diagnostic test [5, 6].

PVB is known to cause biased accuracy measures [1, 6, 10], so it is crucial to correct for this bias in analysis. Methods are available for PVB correction, depending on the

scale of the diagnostic test and gold standard test, and missing data mechanism. A recent review extensively covered all these methods [2], while a specific review on binary diagnostic and gold standard tests with practical implementation was covered in another article [11]. This study focused on PVB correction for binary diagnostic and gold standard tests under MAR missing data mechanism.

For the binary diagnostic test and disease status (as verified by the gold standard test) under MAR assumption, the available PVB correction methods can be roughly divided into Begg and Greene's (BG)-based methods, propensity score (PS)-based methods, and multiple imputation (MI) method. BG-based and MI methods rely on estimating the probability of disease status given test result as an intermediate step before correcting the Sn and Sp estimates. This approach works because this probability, often calculated as positive and negative predictive values [3], is unbiased under MAR assumption [12]. PS-based methods estimate the probability of verification given test result, before correcting for the bias by a weighting method [13, 14]. By estimating the verification probability, PS demonstrates clear and direct relationship with the verification problem, in this case, the PVB problem. Recent implementations of PVB correction methods can be seen in studies evaluating MRI and ultrasound in prostate cancer [15], serum pepsinogens in gastric cancer [16] and fine needle aspiration cytology in breast cancer [17], where the studies utilized BG-based and MI methods.

It is important to correct the bias during the analysis as PVB leads to biased accuracy measures in diagnostic tests [1, 6, 10]. PVB correction methods have been comprehensively reviewed by [2] and [11]. For binary tests under the MAR assumption, PVB correction methods can be categorized into Begg and Greene's (BG)-based, propensity score (PS)-based, and multiple imputation (MI) methods. BG-based and MI methods adjust accuracy measures by calculating probability of disease status given test result, which are unbiased under MAR assumptions [3, 12]. In contrast, PS-based methods estimate the probability of verification given test result and use weighting to correct the bias [13, 14]. Several medical studies have applied these correction methods in their research, signifying the importance of the bias correction [15–19].

In a separate field of study, inverse probability bootstrap (IPB) sampling was introduced to address sampling bias in model-based analyses [20]. While traditionally used for estimating standard errors, IPB leverages the bootstrap technique to achieve unbiased parameter estimates by creating weighted samples. This method corrects the sample distribution without requiring extensive modifications or new methods [20]. Because it is a bootstrap approach, IPB simplifies the estimation of standard errors and enables the calculation of confidence intervals for statistical inference [20].

The PS-based method for PVB correction and IPB share a common approach, where they both begin by estimating the selection probability, or verification probability in the context of PVB, and then use this probability to correct bias through weighting methods. The IPB approach offers an attractive way to address bias due to its reliance on the bootstrap technique, which comes with several advantages. Therefore, IPB was adapted in the context of PVB correction in a study by Arifin and Yusof [21]. In the study, IPB demonstrated low bias for estimation of Sn and Sp. However, it showed relatively higher SE than other PVB correction methods and it only corrects the distribution of the verified portion of the PVB data. This study proposed two methods namely scaled inverse probability weighted resampling (SIPW) and scaled inverse probability weighted balanced resampling (SIPW-B) to overcome these limitations with IPB in the context of PVB correction under the MAR assumption for binary diagnostic tests.

Materials and methods

This section describes the simulated and clinical data sets used in this study, the proposed methods for PVB correction based on IPB, the metrics for performance evaluation, the selected methods for comparison, and the experimental setup of this study. Important notations used are T = test result, D = disease status, V = verification status, n_1 = verified observations, n_0 = unverified observations, and n = all observations, or $n_1 + n_0$.

Data sets

Simulated and clinical data sets were used in this study for performance evaluation and comparison between the methods. The use of simulated data sets allow performance evaluation in comparison to known parameter values [20, 22]. The use of real data sets allow comparison between the methods using reference data sets, following the practice of previous research in PVB correction [23–26].

Simulated data sets

The simulated data sets were generated using the settings described in [21]. The settings were adapted from [24], [25] and [26]. The settings are described as follows:

1. True disease prevalence (p) or $P(D = 1)$: moderate = 0.40 and low = 0.10.
2. True sensitivity (Sn) $P(T = 1|D = 1)$: moderate = 0.6, high = 0.9
3. True specificity (Sp) $P(T = 0|D = 0)$: moderate = 0.6, high = 0.9.
4. Verification probabilities: When the verification depends only on test result, this is a MAR missingness mechanism. Fixed verification probabilities given test result $P(V = 1|T = t)$ were set at $P(V = 1|T = 1) = 0.8$ and $P(V = 1|T = 0) = 0.4$ [24]. In words, when test results are positive, patients are more likely to be verified with probability of 0.8. On the other hand, when test results are negative, patients are less likely to be verified with probability of 0.4.
5. Sample sizes, n: 200 and 1000.

For the complete data, the probability of counts in the 2×2 cross-tabulated table of T versus D are distributed as a multinomial distribution [24, 26]. Based on pre-specified $\text{Sn} = P(T = 1|D = 1)$, $\text{Sp} = P(T = 0|D = 0)$ and $p = P(D = 1) = \pi$, the probabilities of counts are distributed as $M(\pi_1, \pi_2, \pi_3, \pi_4)$, where

$$
\begin{aligned}
\pi_1 &= P(T = 1, D = 1) = P(T = 1|D = 1)P(D = 1), \\
\pi_2 &= P(T = 0, D = 1) = P(T = 0|D = 1)P(D = 1) = [1 - P(T = 1|D = 1)]P(D = 1), \\
\pi_3 &= P(T = 1, D = 0) = P(T = 1|D = 0)P(D = 0) = [1 - P(T = 0|D = 0)]P(D = 0), \\
\pi_4 &= P(T = 0, D = 0) = P(T = 0|D = 0)P(D = 0).
\end{aligned}
$$

Then, for each sample size setting n, the steps to generate a simulated PVB data set for MAR are as follows:

1. A complete data set of size n distributed as multinomial distribution, $M(\pi_1, \pi_2, \pi_3, \pi_4)$ was generated. This process generated numerical values from 1 to 4 based on the probability values.

2. The values were converted into realizations of $T = t$ and $D = d$ variables, where the numbers were mapped as: $1 \to (T = 1, D = 1)$, $2 \to (T = 0, D = 1)$, $3 \to (T = 1, D = 0)$ and $4 \to (T = 0, D = 0)$.

3. Under the MAR assumption, the PVB data set was generated by adding $V = \{1, 0\}$ with verification probability of $P(V = 1|T = 1) = 0.8$ and $P(V = 1|T = 0) = 0.4$, where V follows binomial distribution. D values for $V = 0$ observations were set as NA to create missing values.

The simulated data layout for the full data and PVB data sets is shown in Figure 1.

Full data

i	T	D
1	t_1	d_1
2	t_2	d_2
$\vdots$	$\vdots$	$\vdots$
n-1	t_{n-1}	d_{n-1}
n	t_n	d_n

$\Rightarrow$

PVB data

i	T	D	V
1	t_1	d_1	1
2	t_2	NA	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n-1	t_{n-1}	NA	0
n	t_n	d_n	1

Fig 1. Simulated data layout for full and PVB data.

Clinical data sets

This study utilized two commonly used clinical data sets from [24, 26–30] to show and compare the implementation of the PVB correction methods in real-world situations. The original data obtained from the cited studies were converted to an analysis-ready format (.csv). These data sets are described as follows:

1. **Hepatic scintigraphy test**
   The data set is about hepatic scintigraphy test for detection of liver cancer [27]. Hepatic scintigraphy is an imaging method (diagnostic test) to detect liver cancer. The test was performed on 650 patients, where 344 patients were verified by liver pathological examination (gold standard test). The percentage of unverified patients is 47.1%. The data set contains the following variables:

   (a) **Liver cancer**, *disease*: Binary, 1 = Yes, 0 = No
   (b) **Hepatic scintigraphy**, *test*: Binary, 1 = Positive, 0 = Negative
   (c) **Verified**, *verified*: Binary, 1 = Yes, 0 = No

2. **Diaphanography test**
   The data set is about diaphanography test for detection of breast cancer [28]. Diaphanography is a noninvasive method (diagnostic test) of breast examination by transillumination using visible or infrared light to detect the presence of breast cancer. The test was performed on 900 patients. Only 88 patients were verified by breast tissue biopsy for histological examination (gold standard test). The percentage of unverified patients is 90.2%. The data set contains the following variables:

   (a) **Breast cancer**, *disease*: Binary, 1 = Yes, 0 = No
   (b) **Diaphanography**, *test*: Binary, 1 = Positive, 0 = Negative
   (c) **Verified**, *verified*: Binary, 1 = Yes, 0 = No

Cross-tabulations of these data sets are given in Figure 2.

Hepatic scintigraphy test data

	Liver cancer status			
Test	Cancer	No cancer	Unverified	Total
Positive	231	32	166	429
Negative	27	54	140	221
Total	258	86	306	650

Diaphanography test data

	Breast cancer status			
Test	Cancer	No cancer	Unverified	Total
Positive	26	11	30	67
Negative	7	44	782	833
Total	33	55	812	900

Fig 2. Cross-tabulation of hepatic scintigraphy and diaphanography data sets.

Proposed methods

Scaled inverse probability weighted resampling

Based on the IPB method proposed by Arifin and Yusof [21], the scaled inverse probability weighted resampling (SIPW) method is proposed to overcome the shortcomings of IPB for correcting PVB. The proposed SIPW algorithm is given in Algorithm 1. IPB defines n as the observed sample size. In contrast, SIPW defines n as the full sample size that included both verified and unverified samples. The verified sample size is denoted as n_1, while the unverified sample size is denoted as n_0. A sample of size n is sampled with replacement for b times from n_1, resulting in b samples. Because SIPW does not perform sampling with replacement to obtain the original n_1 observed sample size, this is no longer a bootstrap sample as seen in IPB. Thus, the general term, *resampling*, is used in the name of the method.

Scaled inverse probability weighted balanced resampling

During the course of this research, it was observed that subgroup sample sizes, $n_{D=1}$ and $n_{D=0}$ for diseased and non-diseased observations respectively, directly affect the precision (i.e. SE) of Sn or Sp as a smaller sample size results in lower precision (higher SE). The SE, in particular for Sn, can increase even further when the disease prevalence gets lower as the subgroup size of $n_{D=1}$ will also get smaller. In order to reduce the SE, a possible solution was inspired by how a diagnostic accuracy study is designed.

A diagnostic accuracy study is ideally designed as a cohort study [3, 4, 31]. Whenever the disease is very rare, a case-control study is more feasible, where those with the disease (case group) are compared to those without the disease (control group), often using a 1:1 to 1:3 ratio [31, 32]. This approach is similar to data-level resampling methods commonly used to address the class imbalance problem in machine learning, where the class distribution is rebalanced through over- or under-sampling methods [33].

Therefore, the scaled inverse probability weighted balanced resampling (SIPW-B) method is proposed to mimic the case-control study design. SIPW-B balances the effect of subgroup sample sizes by resizing the subgroup sizes to a predefined ratio of $n_{D=0}$:$n_{D=1}$, while keeping the original sample size $n = n_{D=0} + n_{D=1}$. SIPW-B algorithm is given in Algorithm 2. In the algorithm, the target control:case ratio or $n_{D=0} : n_{D=1}$ is achieved through the following steps:

Algorithm 1 Scaled Inverse Probability Weighted Resampling

Data, notations and definitions:

Test status T, where $T = \{\text{Positive: 1, Negative: 0}\}$
Disease status D, where $D = \{\text{Yes: 1, No: 0, Unknown: } NA\}$
Verification status V, where $V = \{\text{Yes: 1, No: 0}\}$
PVB data $Y = \{T, D, V\}$ of size $n \times 3$
Individual patient instance is denoted with subscript $i = 1, ..., n$
D consists of verified ($V = 1$) and unverified ($V = 0$) sets, where $D = \{D^1, D^0\}$
Unverified set of data Y^0 is of size n_0, consisting of instances with $V_i = 0$
Verified set of data Y^1 is of size $n_1 = n - n_0$, consisting of instances with $V_i = 1$
Propensity score for ith patient PS_i is $\hat{P}(V_i = 1|T_i)$
Inverse probability weight for ith patient IPW_i is $\frac{V_i}{PS_i} + \frac{1-V_i}{1-PS_i}$
Scaled inverse probability weight for ith patient $SIPW_i$ is $\frac{IPW_i}{\sum_{i=1}^{n_1} IPW_i}$, such that $\sum_{i=1}^{n_1} SIPW_i = 1$
Sensitivity Sn is $P(T = 1|D = 1)$
Specificity Sp is $P(T = 0|D = 0)$

Procedure:

Estimate $\widehat{PS}_i$, using Y by a logistic regression
Calculate $\widehat{IPW}_i \leftarrow \frac{1}{\widehat{PS}_i}$ for instances with $V_i = 1$
Calculate $\widehat{SIPW}_i \leftarrow \frac{\widehat{IPW}_i}{\sum_{i=1}^{n_1} \widehat{IPW}_i}$ for instances with $V_i = 1$
/*—Begin resampling block—*/
$Y^{List} \leftarrow$ Empty list of size b samples
$i \leftarrow 1$
while $i < b + 1$ **do**
    $Y_i \leftarrow$ Sample n instances from Y^1 of n_1 instances with probability $SIPW_i$
/*Ensure valid Y_i, where $T \times D$ cross-classification table dimension $TblDim$ is 4*/
    if $TblDim < 4$ **then**
        Discard Y_i
        Repeat sampling, $i \leftarrow i$
    else if $TblDim = 4$ **then**
        Save ith sample, $Y_i^{List} \leftarrow Y_i$
        Continue with next i, $i \leftarrow i + 1$
    end if
end while
/*—End resampling block—*/
$Sn^{List} \leftarrow$ Empty list of size b samples
$Sp^{List} \leftarrow$ Empty list of size b samples
for $i = 1$ **to** b in Y^{List} **do**
    $Sn_i^{List} \leftarrow$ Estimate $\widehat{Sn}_i$ from $T \times D$ cross-classification table
    $Sp_i^{List} \leftarrow$ Estimate $\widehat{Sp}_i$ from $T \times D$ cross-classification table
end for
Estimate $\widehat{Sn} \leftarrow \frac{1}{b} \sum_{i=1}^{b} \widehat{Sn}_i$ from Sn^{List}
Estimate $\widehat{Sp} \leftarrow \frac{1}{b} \sum_{i=1}^{b} \widehat{Sn}_i$ from Sp^{List}

1. Set the desired relative size or ratio of control:case.
2. Calculate the initial relative size of $n_{D=0}$ to $n_{D=1}$ in the PVB sample.
3. Update IPW_i for instances $D_i = 1$ by multiplying the value with the initial relative size.

4. Calculate k as the sum of IPW_i for instances $D_i = 0$, divided by sum of IPW_i for instances $D_i = 1$.

5. Update IPW_i for instances $D_i = 1$ by multiplying the value with $\frac{k}{\text{Desired relative size}}$

6. Calculate $SIPW_i$ as IPW_i divided by the sum of IPW_i.

For the experiments in this study, the target relative size of control:case was set at 1 or a ratio 1:1.

Performance evaluation

Performance metrics

The performance evaluation was based on the metrics that measure the difference between an estimate and its true value [22, 34, 35]. The selected performance metrics, bias and standard error are defined below. For a finite number of simulations B, these are calculated as follows:

1. **Bias**
   Bias of a point estimator $\hat{\theta}$ is the difference between the expected value of $\hat{\theta}$ and the true value of a parameter θ [35]. Bias is calculated as follows:

$$Bias = E[\hat{\theta}] - \theta = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i - \theta. \tag{1}$$

2. **Standard error**
   Standard error (SE) is the square root of the variance, calculated as follows:

$$SE = \sqrt{Var(\hat{\theta})} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i - \bar{\theta})^2}, \tag{2}$$

   where $\bar{\theta}$ is the mean of $\hat{\theta}_i$ across repetitions.

Bias is often the main metric of interest [22], where it indicates the accuracy of a method [35] and whether, on average, the method targets the parameter θ [22]. SE shows the precision of the method [22, 35], where a smaller SE indicates better precision [35].

Methods for comparison

The proposed SIPW and SIPW-B were compared with selected existing PVB correction methods, which are the BG method (representing the BG-based methods), the inverse probability weighting estimator (IPWE) and IPB methods (representing the PS-based methods) and the MI method (representing the imputation-based methods) were selected to represent different approaches for PVB correction. In addition, two additional methods were also compared; full data analysis (FDA) as the ideal method whenever full and unbiased data are available [3], and complete case analysis (CCA) as the uncorrected method that is biased in the presence of PVB [36]. The details of Sn and Sp calculation for these comparison methods were already described elsewhere [21, 37].

For the simulated data sets, the methods are compared by the mean of the estimates and performance metrics (bias and SE), arranged by the sample sizes and Sn-Sp

Algorithm 2 Scaled Inverse Probability Weighted Balanced Resampling

Data, notations and definitions:

Test status T, where $T = \{\text{Positive: 1, Negative: 0}\}$
Disease status D, where $D = \{\text{Yes: 1, No: 0, Unknown: } NA\}$
Verification status V, where $V = \{\text{Yes: 1, No: 0}\}$
PVB data $Y = \{T, D, V\}$ of size $n \times 3$
Individual patient instance is denoted with subscript $i = 1, ..., n$
D consists of verified ($V = 1$) and unverified ($V = 0$) sets, where $D = \{D^1, D^0\}$
Unverified set of data Y^0 is of size n_0, consisting of instances with $V_i = 0$
Verified set of data Y^1 is of size $n_0 = n - n_0$, consisting of instances with $V_i = 1$
Propensity score for ith patient PS_i is $\hat{P}(V_i = 1|T_i)$
Inverse probability weight for ith patient IPW_i is $\frac{V_i}{PS_i} + \frac{1-V_i}{1-PS_i}$
Desired relative size of $n_{D=0}$:$n_{D=1}$, $RelSize$ i.e. control:case size
Scaled inverse probability weight for ith patient $SIPW_i$ is $\frac{IPW_i}{\sum_{i=1}^{n_{D=1}} IPW_i}$, such that $\sum_{i=1}^{n_{D=1}} SIPW_i = 1$
Sensitivity Sn is $P(T = 1|D = 1)$ and Specificity Sp is $P(T = 0|D = 0)$

Procedure:

Estimate $\widehat{PS}_i$, using Y by a logistic regression
Calculate $\widehat{IPW}_i \leftarrow \frac{1}{\widehat{PS}_i}$ for instances with $V_i = 1$
/*Balance $D = 1$ size relative to $D = 0$ size*/
Calculate $RelSizeInit \leftarrow \frac{n_{D=0}}{n_{D=1}}$
Update $\widehat{IPW}_i \leftarrow \widehat{IPW}_i \times RelSizeInit$ for instances with $D_i = 1 \cap V_i = 1$
Calculate constant $k \leftarrow \sum_{i=1}^{n_{D=0}} \widehat{IPW}_i / \sum_{i=1}^{n_{D=1}} \widehat{IPW}_i$
Update $\widehat{IPW}_i \leftarrow \widehat{IPW}_i \times (k/RelSize)$ for instances with $D_i = 1 \cap V_i = 1$
Calculate $\widehat{SIPW}_i \leftarrow \widehat{IPW}_i / \sum_{i=1}^{n_{D=1}} \widehat{IPW}_i$ for instances with $V_i = 1$
/*—Begin resampling block—*/
$Y^{List} \leftarrow$ Empty list of size b samples
$i \leftarrow 1$
while $i < b + 1$ **do**
    $Y_i \leftarrow$ Sample n instances from Y^1 of $n_{V=1}$ instances with probability $SIPW_i$
    /*Ensure valid Y_i, where $T \times D$ cross-classification table dimension $TblDim$ is 4*/
    if $TblDim < 4$ **then**
        Discard Y_i
        Repeat sampling, $i \leftarrow i$
    else if $TblDim = 4$ **then**
        Save ith sample, $Y_i^{List} \leftarrow Y_i$
        Continue with next i, $i \leftarrow i + 1$
    end if
end while
/*—End resampling block—*/
$Sn^{List} \leftarrow$ Empty list of size b samples
$Sp^{List} \leftarrow$ Empty list of size b samples
for $i = 1$ **to** b in Y^{List} **do**
    $Sn_i^{List} \leftarrow$ Estimate $\widehat{Sn}_i$ from $T \times D$ cross-classification table
    $Sp_i^{List} \leftarrow$ Estimate $\widehat{Sp}_i$ from $T \times D$ cross-classification table
end for
Estimate $\widehat{Sn} \leftarrow \frac{1}{b} \sum_{i=1}^{b} \widehat{Sn}_i$ from Sn^{List}
Estimate $\widehat{Sp} \leftarrow \frac{1}{b} \sum_{i=1}^{b} \widehat{Sn}_i$ from Sp^{List}

combinations. We did not consider *coverage*, i.e. the proportion of times the CI includes the estimate [22, 34, 35], as one of the performance metric for comparing the methods for the simulation. This was because, we did not propose a new method to obtain the CI for SIPW and SIPW-B.

For the clinical data sets, point estimates and the respective 95% CIs were estimated for comparison. For CCA, the CIs for Sn and Sp were calculated by using Wald interval, while for the BG method, the calculation step given in the original article was followed [38]. For IPWE, IPB, SIPW and SIPW-B, the CIs were obtained by bootstrap percentile interval method [39]. For MI, the CIs were obtained by Rubin's rule [37, 40].

Experimental setup

R statistical programming language [41] version 3.6.3 was used to run the experiments within RStudio [42] integrated development environment. The choice of the final stable version of the 3.x.x versions of the R language was to ensure reproducibility of the experimental results compared to the currently used 4.x.x versions of the R language which is still in continuous development. *mice* [43] version 3.14.0 and *simstudy* [44] version 0.5.0 R packages were used. The seed number for the random number generator was set to 3209673. For simulated data sets, the settings were the number of simulation runs $B = 500$, samples $b = 1000$ (for IPB, SIPW and SIPW-B) and imputations $m = 100$ [45, 46]. For clinical data sets, the settings were the number of samples $b = 1000$ (for IPB, SIPW and SIPW-B), bootstrap replicates $R = 1000$ (to obtain the CIs for IPWE, SIPW and SIPW-B) and $m =$ the percentage of incomplete cases for real clinical data sets [47–49]. Please note that IPB does not need an additional round of bootstrapping to obtain its CI, as bootstrap is integral to the method itself.

Results

Simulated data sets

The simulation results for the FDA, CCA and PVB correction methods for $p = 0.4$ are displayed in Table 1, comparing sample sizes $n = 200$ and 1000. The results are arranged by Sn = (0.6, 0.9) and Sp = (0.6, 0.9) parameter combinations. The proportions of verification $P(V = 1)$ were 0.59, 0.52 and 0.64 for (0.6, 0.6), (0.6, 0.9) and (0.9, 0.6) for (Sn, Sp) pairs respectively. The best values (i.e. the smallest bias and SE values) achieved by IPB, SIPW, and SIPW-B are marked with an asterisk.

Next, The simulation results for the FDA, CCA and PVB correction methods for $p = 0.1$ are displayed in Table 2 for sample sizes $n = 200$ and 1000. The results are arranged by Sn = (0.6, 0.9) and Sp = (0.6, 0.9) parameter combinations. The proportions of verification $P(V = 1)$ were 0.57, 0.46 and 0.58 for (0.6, 0.6), (0.6, 0.9) and (0.9, 0.6) for (Sn, Sp) pairs respectively. Again, the smallest bias and SE values achieved by IPB, SIPW, and SIPW-B are marked with an asterisk.

As observed in Table 1 and 2, both SIPW and SIPW-B generally performed better than IPB for Sn and Sp estimation as indicated by smaller bias and SE values. SIPW-B performed better (smaller SE) for Sn compared to SIPW by enlarging the size of the case group ($n_{D=1}$). As expected, SIPW showed smaller SE than SIPW-B for Sp estimation because it maintains the original size of the control group ($n_{D=0}$), of which at a disease prevalence of $p = 0.4$, the control group size is larger than the case group (Table 1). This effect was more notable at a lower disease prevalence of $p = 0.1$ (Table 2). Bias values showed mixed results, with marginal differences observed among IPB, SIPW, and SIPW-B. When compared to the existing methods (EBG, IPWE and MI), SIPW-B closely matched these methods for the Sn, while SIPW closely aligned

Table 1. Comparison between IPB and existing PVB correction methods for $p = 0.4$ with $n = 200$ and 1000 under three combinations of Sn and Sp.

Methods	Mean	Bias	SE	Mean	Bias	SE	Mean	Bias	SE	Mean	Bias	SE
	n = **200**						n = **1000**					
	Sn = 0.6			**Sp = 0.6**			**Sn = 0.6**			**Sp = 0.6**		
FDA	0.603	0.003	0.055	0.602	0.002	0.044	0.602	0.002	0.023	0.600	0.000	0.019
CCA	0.754	0.154	0.060	0.430	-0.170	0.060	0.751	0.151	0.026	0.428	-0.172	0.026
BG	0.607	0.007	0.072	0.602	0.002	0.050	0.602	0.002	0.031	0.600	0.000	0.022
IPWE	0.607	0.007	0.072	0.602	0.002	0.050	0.602	0.002	0.031	0.600	0.000	0.022
MI	0.605	0.005	0.075	0.599	-0.001	0.052	0.601	0.001	0.032	0.599	-0.001	0.022
IPB	0.609	0.009	0.105	0.602	0.002	0.078	0.599	*-0.001	0.044	0.601	0.001	0.034
SIPW	0.608	0.008	0.090	0.603	0.003	*0.065	0.604	0.004	*0.037	0.600	*0.000	*0.029
SIPW-B	0.606	*0.006	*0.084	0.600	*0.000	0.070	0.602	0.002	0.038	0.600	*0.000	0.031
	Sn = 0.6			**Sp = 0.9**			**Sn = 0.6**			**Sp = 0.9**		
FDA	0.603	0.003	0.055	0.902	0.002	0.027	0.602	0.002	0.023	0.900	0.000	0.012
CCA	0.754	0.154	0.061	0.822	-0.078	0.054	0.752	0.152	0.026	0.818	-0.082	0.025
EBG	0.608	0.008	0.075	0.903	0.003	0.030	0.602	0.002	0.031	0.900	0.000	0.014
IPWE	0.608	0.008	0.075	0.903	0.003	0.030	0.602	0.002	0.031	0.900	0.000	0.014
MI	0.605	0.005	0.076	0.901	0.001	0.031	0.602	0.002	0.032	0.900	0.000	0.014
IPB	0.605	*0.005	0.118	0.903	*0.003	0.049	0.599	*-0.001	0.048	0.899	-0.001	0.024
SIPW	0.610	0.010	0.094	0.903	*0.003	*0.040	0.604	0.004	0.040	0.901	0.001	*0.019
SIPW-B	0.608	0.008	*0.086	0.905	0.005	0.042	0.602	0.002	*0.039	0.900	*0.000	0.020
	Sn = 0.9			**Sp = 0.6**			**Sn = 0.9**			**Sp = 0.6**		
FDA	0.899	-0.001	0.033	0.601	0.001	0.044	0.901	0.001	0.015	0.601	0.001	0.020
CCA	0.945	0.045	0.027	0.427	-0.173	0.057	0.948	0.048	0.012	0.429	-0.171	0.027
EBG	0.896	-0.004	0.046	0.600	0.000	0.046	0.901	0.001	0.022	0.601	0.001	0.021
IPWE	0.896	-0.004	0.046	0.600	0.000	0.046	0.901	0.001	0.022	0.601	0.001	0.021
MI	0.889	-0.011	0.046	0.598	-0.002	0.047	0.899	-0.001	0.023	0.600	0.000	0.021
IPB	0.894	-0.006	0.064	0.595	-0.005	0.072	0.901	*0.001	0.028	0.600	*0.000	0.033
SIPW	0.896	-0.004	0.058	0.597	-0.003	*0.064	0.901	*0.001	0.027	0.600	*0.000	*0.029
SIPW-B	0.897	*-0.003	*0.055	0.599	*-0.001	0.069	0.901	*0.001	*0.025	0.603	0.003	0.031

Abbreviations: BG, Begg and Greenes' method; CCA, complete case analysis; FDA, Full data analysis; IPB, inverse probability bootstrap; IPWE, inverse probability weighting estimator; MI, multiple imputation; n, sample size; p, disease prevalence; SE, standard error; SIPW, scaled inverse probability weighted resampling; SIPW-B, scaled inverse probability weighted balanced resampling; Sn, sensitivity; Sp, specificity.

with the results for these methods for Sp estimation. A further observation was made that, across all PVB correction methods, SE decreased with increasing disease prevalence and sample size.

Clinical data sets

The results for the PVB correction methods using the clinical data sets are displayed in Table 3. CCA is displayed to illustrate the results without bias correction. All PVB correction methods showed almost identical results for Sn and Sp point estimates across these data sets, with the exception of MI, which showed a slightly different Sn estimate for the diaphanography data set. In contrast to IPB, for the hepatic data set, SIPW and SIPW-B showed 95% CIs that were more consistent with existing methods. The diaphanography data set further highlighted the limitations of IPB, as its 95% CIs for

Table 2. Comparison between IPB and existing PVB correction methods for $p = 0.1$ with $n = 200$ and 1000 under three combinations of Sn and Sp.

Methods	Mean	Bias	SE	Mean	Bias	SE	Mean	Bias	SE	Mean	Bias	SE
	n = **200**						n = **1000**					
	Sn = 0.6			**Sp = 0.6**			**Sn = 0.6**			**Sp = 0.6**		
FDA	0.596	-0.004	0.112	0.601	0.001	0.037	0.600	0.000	0.048	0.600	0.000	0.017
CCA	0.743	0.143	0.117	0.429	-0.171	0.049	0.754	0.154	0.053	0.429	-0.171	0.022
EBG	0.603	0.003	0.144	0.601	0.001	0.038	0.607	0.007	0.067	0.601	0.001	0.017
IPWE	0.603	0.003	0.144	0.601	0.001	0.038	0.607	0.007	0.067	0.601	0.001	0.017
MI	0.579	-0.021	0.136	0.598	-0.002	0.039	0.601	0.001	0.067	0.600	0.000	0.017
IPB	0.595	-0.005	0.202	0.606	0.006	0.062	0.605	0.005	0.093	0.600	*0.000	0.027
SIPW	0.613	0.013	0.184	0.603	0.003	*0.055	0.604	*0.004	0.084	0.600	*0.000	*0.025
SIPW-B	0.600	*0.000	*0.157	0.601	*0.001	0.062	0.607	0.007	*0.071	0.601	0.001	0.027
	Sn = 0.6			**Sp = 0.9**			**Sn = 0.6**			**Sp = 0.9**		
FDA	0.600	0.000	0.115	0.900	0.000	0.022	0.600	0.000	0.048	0.900	0.000	0.010
CCA	0.738	0.138	0.119	0.818	-0.082	0.043	0.749	0.149	0.052	0.819	-0.081	0.019
EBG	0.598	-0.002	0.143	0.900	0.000	0.023	0.601	0.001	0.065	0.900	0.000	0.010
IPWE	0.598	-0.002	0.143	0.900	0.000	0.023	0.601	0.001	0.065	0.900	0.000	0.010
MI	0.568	-0.032	0.134	0.900	0.000	0.023	0.593	-0.007	0.065	0.900	0.000	0.010
IPB	0.599	-0.001	0.214	0.898	-0.002	0.042	0.604	0.004	0.099	0.901	0.001	0.018
SIPW	0.600	*0.000	0.178	0.899	*-0.001	*0.033	0.602	0.002	0.080	0.901	0.001	*0.014
SIPW-B	0.598	-0.002	*0.152	0.902	0.002	0.038	0.600	*0.000	*0.067	0.900	*0.000	0.017
	Sn = 0.9			**Sp = 0.6**			**Sn = 0.9**			**Sp = 0.6**		
FDA	0.875	-0.025	0.062	0.602	0.002	0.035	0.899	-0.001	0.028	0.600	0.000	0.016
CCA	0.910	0.010	0.042	0.430	-0.170	0.048	0.947	0.047	0.025	0.428	-0.172	0.021
EBG	0.837	-0.063	0.068	0.599	-0.001	0.037	0.900	0.000	0.044	0.600	0.000	0.017
IPWE	0.837	-0.063	0.068	0.599	-0.001	0.037	0.900	0.000	0.044	0.600	0.000	0.017
MI	0.800	-0.100	0.080	0.597	-0.003	0.037	0.889	-0.011	0.046	0.600	0.000	0.017
IPB	0.832	-0.068	0.133	0.605	0.005	0.059	0.897	-0.003	0.060	0.601	0.001	0.027
SIPW	0.837	*-0.063	0.108	0.600	*0.000	*0.050	0.901	0.001	0.052	0.600	*0.000	*0.023
SIPW-B	0.837	*-0.063	*0.077	0.600	*0.000	0.061	0.900	*0.000	*0.045	0.601	0.001	0.028

Abbreviations: BG, Begg and Greenes' method; CCA, complete case analysis; FDA, Full data analysis; IPB, inverse probability bootstrap; IPWE, inverse probability weighting estimator; MI, multiple imputation; n, sample size; p, disease prevalence; SE, standard error; SIPW, scaled inverse probability weighted resampling; SIPW-B, scaled inverse probability weighted balanced resampling; Sn, sensitivity; Sp, specificity.

Sn and Sp were notably wider and diverged from those estimated by SIPW, SIPW-B and existing methods. Based on the small observed sample size of this data set (only 88 verified out of 900 patients), IPB appears to perform poorly with interval estimation in situations with limited data.

Discussion

This study introduced two new methods based on the IPB method: SIPW and SIPW-B. These were compared against IPB and other existing methods across simulated data sets and clinical data sets to evaluate their performance in estimating Sn and Sp. In simulated data sets, both SIPW and SIPW-B demonstrated smaller bias and standard error values than IPB for Sn and Sp estimation, respectively. Specifically, SIPW-B and SIPW outperformed IPB in Sn and Sp estimation, respectively. In clinical data sets, all

Table 3. Sn and Sp estimates of IPB, SIPW, SIPW-B and other methods with the respective 95% CIs using clinical data sets.

Methods	Scintigraphy data set		Diaphanography data set	
	Sn (95% CI)	Sp (95% CI)	Sn (95% CI)	Sp (95% CI)
CCA	0.895 (0.858, 0.933)	0.628 (0.526, 0.730)	0.788 (0.648, 0.927)	0.800 (0.694, 0.906)
BG	0.836 (0.788, 0.884)	0.738 (0.662, 0.815)	0.292 (0.134, 0.449)	0.973 (0.958, 0.988)
IPWE	0.836 (0.784, 0.883)	0.738 (0.651, 0.809)	0.292 (0.165, 0.509)	0.973 (0.955, 0.986)
MI	0.834 (0.782, 0.885)	0.738 (0.661, 0.815)	0.279 (0.124, 0.435)	0.972 (0.957, 0.987)
IPB	0.838 (0.793, 0.882)	0.738 (0.653, 0.822)	0.290 (0.059, 0.520)	0.973 (0.935, 1.000)
SIPW	0.837 (0.785, 0.886)	0.739 (0.655, 0.811)	0.292 (0.176, 0.548)	0.973 (0.957, 0.987)
SIPW-B	0.837 (0.785, 0.885)	0.739 (0.655, 0.812)	0.291 (0.176, 0.548)	0.973 (0.957, 0.987)

Abbreviations: BG, Begg and Greenes' method; CCA, complete case analysis; CI, confidence interval; IPB, inverse probability bootstrap; IPWE, inverse probability weighting estimator; MI, multiple imputation; SIPW, scaled inverse probability weighted resampling; SIPW-B, scaled inverse probability weighted balanced resampling; Sn, sensitivity; Sp, specificity.

methods generally showed consistent results with each other, except for the Diaphanography data set where IPB and MI diverged from others. For practical applications, when disease prevalence is low, SIPW is preferable for estimating Sp, whereas SIPW-B is better suited for Sn estimation.

There are several advantages of the newly proposed methods as compared to IPB, while sharing the other advantages of IPB as discussed in another study [21]. First, SIPW and SIPW-B share the same advantage with IPB and MI as they allow the use of any full data analytical methods. In addition, SIPW and SIPW-B aligns better with MI by restoring full data N, while IPB restores the data containing the complete cases n only. The ability to utilize full data approach is advantageous as shown in [50] for the kappa coefficient in diagnostic accuracy studies. Second, both proposed methods showed lower SEs as compared to IPB. As highlighted in [21], a major disadvantage of IPB is its large SE compared to other existing methods. SIPW and SIPW-B overcome this issue by having comparable SEs to these existing methods. Third, the proposed methods were similar to IPB in that they are less biased than MI at low disease prevalence and comparable to EBG and IPWE in terms of bias [21], while showing smaller SEs than IPB. Fourth, these two methods are easy to implement compared to MI, where they require the estimation of PS values, followed by the weighted resampling procedure. In contrast, in implementing MI for the PVB correction, one must choose suitable imputations methods as the performance depends on the selected imputation method [40, 51]. Finally, for PS-based methods (IPWE, IPB, SIPW and SIPW-B), they share the same advantage of using PS; the probability of verification given test result, $P(V = 1|T = t)$. In contrast, BG-based and MI methods depend on the accurate estimation of the probability of disease status given test result, $P(D = d|T = t)$. PS-based methods, however, utilize the accurate estimation of $P(V = 1|T = t)$ to perform the correction. This makes PS-based methods particularly advantageous when a case-control study design is employed for diagnostic accuracy studies, as $P(D = d|T = t)$ can be inaccurate in this situation [8].

However, there are several disadvantages of the proposed methods that can be highlighted. SIPW and SIPW-B are computationally intensive methods that rely on repeated resampling. Unlike IPB, CIs cannot be derived directly from these resamples because they no longer constitute valid bootstrap samples. For interval estimation, an additional computationally demanding bootstrapping procedure must be performed on top of the algorithms. Specifically, if 1000 bootstrap replicates ($R = 1000$) are required for interval estimation, the computational burden significantly increases. For instance, if 1000 bootstrap replicates ($R = 1000$) are needed for interval estimation, the time taken

to complete a full SIPW procedure (e.g., 10 seconds with $b = 1000$ SIPW samples) will be multiplied by 1000. In addition, SIPW-B was designed to mimic the case-control study design. SIPW-B balances the effect of subgroup sample sizes by resizing the subgroup sizes to a predefined ratio of $n_{D=0}$:$n_{D=1}$, while keeping the original sample size $n = n_{D=0} + n_{D=1}$. The side effect of the procedure is that it will lower the precision of the larger-sized subgroup, although the precision for the smaller-sized subgroup will increase. As an example, for a sample size $n = 1000$ with $p = 0.2$, the subgroup sizes will be $n_{D=1} = n \times p = 200$ and $n_{D=0} = n \times (1 - p) = 800$. The desired ratio of control:case is 1:1 ratio. Before the balancing procedure, Sn will be calculated from $n_{D=1} = 200$ while Sp from $n_{D=0} = 800$ sample sizes. After the balancing procedure, Sn will be calculated from $n_{D=1} = 500$ while Sp from $n_{D=0} = 500$ sample sizes. Notably, the subgroup size $n_{D=0}$ reduces from 800 to 500 for the calculation of Sp, reducing both the counts for numerator and denominator, thus lowering the precision of Sp. On the positive note, the subgroup size $n_{D=1}$ increases from 200 to 500 for the calculation of Sn, increasing both the counts for numerator and denominator, thus increasing the precision of Sn. Therefore, it is recommended to use SIPW-B only when Sn and Sp are the main estimates to be calculated.

Therefore, we recommend using SIPW when complete data restoration is needed for further analysis involving additional metrics than depends on prevalence like PPV, NPV and accuracy [3, 52], rather than just Sn and Sp. Conversely, SIPW-B should be employed only when the primary aim is to obtain accurate Sn and Sp estimates. After applying SIPW-B, which mimics a case-control study design, statistics that rely on an accurate distribution of $P(D = 1)$ (such as PPV $= P(D = 1|T = 1)$ or NPV $= P(D = 0|T = 0)$) are no longer valid [8]. Although PPV and NPV can be calculated indirectly from Sn and Sp point estimates of SIPW-B by utilizing Bayes' theorem [3, 37], further research must be done to investigate this issue.

Conclusion

PVB correction is crucial to ensure valid accuracy estimates for diagnostic studies affected by PVB. Over the years, various PVB correction methods under the MAR assumption have been proposed, each with its strengths and limitations. This paper discusses the SIPW and SIPW-B methods as solutions to overcome the limitations of the IPB method in the context of PVB correction under the MAR assumption for binary diagnostic tests. The results show that both SIPW and SIPW-B outperformed IPB and were consistent with existing PVB correction methods. Specifically, SIPW excelled in estimating Sp, while SIPW-B excelled in Sn estimation, particularly when disease prevalence is low. The proposed methods offer advantages such as complete data restoration for subsequent analysis and the use of PS that are not influenced by disease prevalence. Although the new methods currently require more computational resources, this should become less of an issue with advancements in computational power.

Author contributions

Conceptualization, W.N.A. and U.K.Y.; methodology, W.N.A. and U.K.Y.; software, W.N.A.; validation, W.N.A. and U.K.Y.; formal analysis, W.N.A.; investigation, W.N.A.; resources, U.K.Y.; data curation, W.N.A.; writing—original draft preparation, W.N.A.; writing—review and editing, W.N.A. and U.K.Y.; supervision, U.K.Y.; project administration, W.N.A.; funding acquisition, W.N.A. and U.K.Y. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We thank our colleagues at the School of Computer Sciences and the School of Medical Sciences, Universiti Sains Malaysia for their comments on the early findings of this study and this article's draft.

Funding

The publication of this article was funded by the Research Creativity and Management Office and the School of Medical Sciences, Universiti Sains Malaysia.

Supporting Information

The data and code presented in this article are available from this GitHub repository: `https://github.com/wnarifin/sipw_in_pvb`.

References

1. O'Sullivan JW, Banerjee A, Heneghan C, Pluddemann A. Verification bias. BMJ Evidence-based Medicine. 2018;23(2):54–55.
2. Umemneku Chikere CM, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard–An update. PLoS ONE. 2019;14(10):e0223832.
3. Zhou XH, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. 2nd ed. Hoboken, New Jersey: John Wiley & Sons; 2011.
4. Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York, USA: Oxford University Press; 2011.
5. Alonzo TA. Verification bias-impact and methods for correction when assessing accuracy of diagnostic tests. Revstat Statistical Journal. 2014;12(1):67–83.
6. de Groot JAH, Bossuyt PMM, Reitsma JB, Rutjes AWS, Dendukuri N, Janssen KJM, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ. 2011;343:d4770.
7. Schmidt RL, Walker BS, Cohen MB. Verification and classification bias interactions in diagnostic test accuracy studies for fine-needle aspiration biopsy. Cancer Cytopathology. 2015;123(3):193–201.
8. Kohn MA. Studies of diagnostic test accuracy: Partial verification bias and test result-based sampling. Journal of Clinical Epidemiology. 2022;145:179–182. doi:10.1016/j.jclinepi.2022.01.022.
9. Schmidt RL, Factor RE. Understanding Sources of Bias in Diagnostic Accuracy Studies. Archives of Pathology & Laboratory Medicine. 2013;137(4):558–565. doi:10.5858/arpa.2012-0198-ra.
10. Rutjes A, Reitsma J, Coomarasamy A, Khan K, Bossuyt P, et al. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technology Assessment. 2007;11(50).

11. Arifin WN, Yusof UK. Correcting for partial verification bias in diagnostic accuracy studies: A tutorial using R. Statistics in Medicine. 2022;41(9):1709–1727.

12. Zhou XH. Effect of verification bias on positive and negative predictive values. Statistics in Medicine. 1994;13(17):1737–1745.

13. Alonzo TA, Pepe MS. Assessing accuracy of a continuous screening test in the presence of verification bias. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2005;54(1):173–190.

14. He H, McDermott MP. A robust method using propensity score stratification for correcting verification bias for binary tests. Biostatistics. 2012;13(1):32–47.

15. Day E, Eldred-Evans D, Prevost AT, Ahmed HU, Fiorentino F. Adjusting for verification bias in diagnostic accuracy measures when comparing multiple screening tests - an application to the IP1-PROSTAGRAM study. BMC Medical Research Methodology. 2022;22(1). doi:10.1186/s12874-021-01481-w.

16. Robles C, Rudzite D, Polaka I, Sjomina O, Tzivian L, Kikuste I, et al. Assessment of Serum Pepsinogens with and without Co-Testing with Gastrin-17 in Gastric Cancer Risk Assessment—Results from the GISTAR Pilot Study. Diagnostics. 2022;12(7):1746. doi:10.3390/diagnostics12071746.

17. El Chamieh C, Vielh P, Chevret S. Statistical methods for evaluating the fine needle aspiration cytology procedure in breast cancer diagnosis. BMC Medical Research Methodology. 2022;22(1). doi:10.1186/s12874-022-01506-y.

18. Niknejad F, Ahmadi F, Roudbari M. Verification Bias Correction in Endometrial Abnormalities in Infertile Women Referred to Royan Institute Using Statistical Methods. Medical Journal of The Islamic Republic of Iran. 2023;doi:10.47176/mjiri.37.122.

19. Olthof EP, Bergink-Voorthuis BJ, Wenzel HHB, Mongula J, van der Velden J, Spijkerboer AM, et al. Diagnostic accuracy of MRI, CT, and [18F]FDG-PET-CT in detecting lymph node metastases in clinically early-stage cervical cancer — a nationwide Dutch cohort study. Insights into Imaging. 2024;15(1). doi:10.1186/s13244-023-01589-1.

20. Nahorniak M, Larsen DP, Volk C, Jordan CE. Using Inverse Probability Bootstrap Sampling to Eliminate Sample Induced Bias in Model Based Analysis of Unequal Probability Samples. PLOS ONE. 2015;10(6):e0131765. doi:10.1371/journal.pone.0131765.

21. Arifin WN, Yusof UK. Partial Verification Bias Correction Using Inverse Probability Bootstrap Sampling for Binary Diagnostic Tests. Diagnostics. 2022;12:2839.

22. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019;38(11):2074–2102. doi:10.1002/sim.8086.

23. Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. Biometrics. 2003;59(1):163–171.

24. Harel O, Zhou XH. Multiple imputation for correcting verification bias. Statistics in Medicine. 2006;25(22):3769–3786.

25. Ünal İ, Burgut HR. Verification bias on sensitivity and specificity measurements in diagnostic medicine: a comparison of some approaches used for correction. Journal of Applied Statistics. 2014;41(5):1091–1104.

26. Rochani H, Samawi HM, Vogel RL, Yin J. Correction of Verication Bias using Log-Linear Models for a Single Binaryscale Diagnostic Tests. Journal of Biometrics and Biostatistics. 2015;6(5):266.

27. Drum DE, Christacopoulos JS. Hepatic scintigraphy in clinical decision making. Journal of Nuclear Medicine. 1972;13(12):908–915.

28. Marshall V, Williams DC, Smith KD. Diaphanography as a means of detecting breast cancer. Radiology. 1984;150(2):339–343. doi:10.1148/radiology.150.2.6691086.

29. Greenes R, Begg C. Assessment of diagnostic technologies. Methodology for unbiased estimation from samples of selectively verified patients. Investigative Radiology. 1985;20(7):751–756.

30. Zhou XH. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. Communications in Statistics-Theory and Methods,. 1993;22(11):3177–3198.

31. Linnet K, Bossuyt PMM, Moons KGM, Reitsma JB. Quantifying the Accuracy of a Diagnostic Test or Marker. Clinical Chemistry. 2012;58(9):1292–1301. doi:10.1373/clinchem.2012.182543.

32. Leeflang MMG, Allerberger F. How to: evaluate a diagnostic test. Clinical Microbiology and Infection. 2019;25(1):54–59. doi:10.1016/j.cmi.2018.06.011.

33. Khushi M, Shaukat K, Alam TM, Hameed IA, Uddin S, Luo S, et al. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. IEEE Access. 2021;9:109960–109975. doi:10.1109/access.2021.3102399.

34. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in Medicine. 2006;25(24):4279–4292. doi:10.1002/sim.2673.

35. Casella G, Berger RL. Statistical inference. 2nd ed. Duxbury Advanced Series. Delhi, India: Cengage Learning; 2002.

36. de Groot JAH, Janssen KJM, Zwinderman AH, Bossuyt PMM, Reitsma JB, Moons KGM. Correcting for partial verification bias: a comparison of methods. Annals of Epidemiology. 2011;21(2):139–148.

37. Arifin WN, Yusof UK. Correcting for partial verification bias in diagnostic accuracy studies: a tutorial using R. Statistics in Medicine. 2022;41(9):1709–1727.

38. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983; p. 207–215.

39. Woodward M. Epidemiology: Study Design and Data Analysis. 3rd ed. Chapman Hall/CRC Texts in Statistical Science. Boca Raton, FL: CRC Press; 2014.

40. van Buuren S. Flexible Imputation of Missing Data. 2nd ed. Chapman & Hall/CRC Interdisciplinary Statistics. Boca Raton, FL: CRC Press; 2018.

41. R Core Team. R: A Language and Environment for Statistical Computing; 2024. Available from: `https://www.r-project.org/`.

42. R Studio Team. RStudio: Integrated Development for R. Boston, MA: RStudio, Inc; 2024.

43. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45(3):1–67.

44. Goldfeld K, Wujciak-Jens J. simstudy: Illuminating research methods through data generation. Journal of Open Source Software. 2020;5(54):2763. doi:10.21105/joss.02763.

45. Dong Y, Peng CYJ. Principled missing data methods for researchers. SpringerPlus. 2013;2(1). doi:10.1186/2193-1801-2-222.

46. Royston P, White I. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. Journal of Statistical Software. 2011;45(i04).

47. Bodner TE. What Improves with Increased Missing Data Imputations? Structural Equation Modeling: A Multidisciplinary Journal. 2008;15(4):651–675. doi:10.1080/10705510802339072.

48. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine. 2011;30(4):377–399. doi:10.1002/sim.4067.

49. Pedersen A, Mikkelsen E, Cronin-Fenton D, Kristensen N, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clinical Epidemiology. 2017;Volume 9:157–166. doi:10.2147/clep.s129785.

50. Roldán-Nofuentes JA, Regad SB. Estimation of the Average Kappa Coefficient of a Binary Diagnostic Test in the Presence of Partial Verification. Mathematics. 2021;9(14):1694.

51. Faisal S, Tutz G. Multiple imputation using nearest neighbor methods. Information Sciences. 2021;570:500–516. doi:10.1016/j.ins.2021.04.009.

52. Šimundić AM. Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC. 2009;19(4):203–11.