mathematics

MDPI

Article

Estimation of the Average Kappa Coefficient of a Binary Diagnostic Test in the Presence of Partial Verification

José Antonio Roldán-Nofuentes [1,*] and Saad Bouh Regad [2]

[1] Department of Statistics, School of Medicine, University of Granada, 18016 Granada, Spain
[2] Epidemiology and Public Health Research Unit and URMCD, School of Medicine, University of Nouakchott Alaasriya, Nouakchott BP 880, Mauritania; regad@una.mr
* Correspondence: jaroldan@ugr.es

Abstract: The average kappa coefficient of a binary diagnostic test is a measure of the beyond-chance average agreement between the binary diagnostic test and the gold standard, and it depends on the sensitivity and specificity of the diagnostic test and on disease prevalence. In this manuscript the estimation of the average kappa coefficient of a diagnostic test in the presence of verification bias is studied. Confidence intervals for the average kappa coefficient are studied applying the methods of maximum likelihood and multiple imputation by chained equations. Simulation experiments have been carried out to study the asymptotic behaviors of the proposed intervals, given some application rules. The results obtained in our simulation experiments have shown that the multiple imputation by chained equations method provides better results than the maximum likelihood method. A function has been written in R to estimate the average kappa coefficient by applying multiple imputation. The results have been applied to the diagnosis of liver disease.

Keywords: average kappa coefficient; missing data; multiple imputation by chained equations; partial verification

Citation: Roldán-Nofuentes, J.A.; Regad, S.B. Estimation of the Average Kappa Coefficient of a Binary Diagnostic Test in the Presence of Partial Verification. *Mathematics* **2021**, *9*, 1694. https://doi.org/10.3390/math9141694

Academic Editors: María Del Mar Rueda and Andrés Cabrera-León

Received: 22 June 2021
Accepted: 14 July 2021
Published: 19 July 2021

1. Introduction

A binary diagnostic test (BDT) is a medical test used to determine whether or not a patient has a certain disease. Scintigraphy for the diagnosis of liver disease is an example of BDT. Sensitivity and specificity are the fundamental parameters to assess the effectiveness of a BDT. Sensitivity (Se) is the probability of a positive result for the BDT when the patient has the disease, and specificity (Sp) is the probability of a negative result for the BDT when the patient does not have the disease. When considering the losses associated with a misclassification with the BDT, the effectiveness of a BDT is measured with the weighted kappa coefficient [1,2], which depends on Se and Sp of the BDT, on the disease's prevalence and on a weighting index, which is a measure of the relative loss between the false positives and the false negatives and it is a value set by the clinician and takes a value between 0.5 and 1 when the BDT is used as a screening test, and the weighting index takes a value between 0 and 0.5 when the BDT is used as a confirmatory test. Therefore, the investigator must assign a value to the weighting index according to the utility of the BDT (screening test or confirmatory test). Roldán-Nofuentes and Olvera-Porcel [3] have defined a new measure to evaluate the effectiveness of a BDT based on the weighted kappa coefficient: the average kappa coefficient. The average kappa coefficient depends on the Se and Sp of the BDT and on the disease prevalence, but it does not depend on the weighting index; the average kappa coefficient solves the problem of assigning values to the weighting index.

In order to obtain unbiased estimators of the parameters of a BDT it is necessary to know the disease status of each patient in a random sample. The medical test through which the disease status of a patient is known is called gold standard (GS), and therefore the effectiveness of a BDT is assessed in relation to a GS. A biopsy for the diagnosis of liver disease is an example of GS. The most common sampling design to evaluate the

effectiveness of a BDT is cross-sectional sampling. This design consists of applying the BDT and the GS to all patients in a random sample. In this situation the true disease state (disease present or disease absent) is known for all patients in the sample. In the cross-sectional sample there are no missing data and therefore corresponds to a complete data situation.

In clinical practice, it is common that when evaluating a BDT, the GS is not applied to all patients in the sample, giving rise to a problem called partial verification of the disease [4]. If the GS is an expensive medical test or a medical test that involves risks for the patient, then the GS is not applied to all the patients in the sample. In this situation, if Se and Sp are estimated without considering the patients for whom the GS is unknown, the estimators are affected by so-called verification bias [4,5]. Begg and Greenes [4] deduced the maximum likelihood estimators of Se and Sp when the missing data mechanism is missing at random (MAR). The MAR assumption holds that the selection of a patient to verify their disease status with the GS depends only on the result of the BDT. Therefore, the true disease state (disease present or disease absent) is unknown for a subset of patients; the missing information is the true disease status for this subset of patients in the sample. Harel and Zhou [6] have studied the estimation of Se and Sp of a BDT through multiple imputation, assuming the MAR assumption, and they have shown through simulation experiments that multiple imputation provides better results than the method of Begg and Greenes [4]. A review of the impact of verification bias in estimating the accuracy of a BDT (and a continuous test) can be seen in Alonzo [7]. Roldán-Nofuentes and Luna [8] have studied the estimation of the weighted kappa coefficient in the presence of partial disease verification.

In this manuscript we study the estimation of the average kappa coefficient in the presence of verification bias. The manuscript is structured as follows: in Section 2, the weighted kappa coefficient and the average kappa coefficient of a BDT are presented. In Section 3 we study the estimation of the average kappa coefficient with complete data. In Section 4 we study the estimation of the average kappa coefficient when there are missing data, applying the maximum likelihood method and the multiple imputation by chained equations method. In Section 5, simulation experiments are carried out to study the asymptotic behaviors of the confidence intervals proposed in Section 3. In Section 6, we present a function written in R to estimate the average kappa coefficient in the presence of missing data. In Section 7, the results obtained are applied to an example on the diagnosis of liver disease, and in Section 8 the results are discussed.

2. Weighted Kappa Coefficient and Average Kappa Coefficient

Let us consider a BDT whose performance is assessed in relation to a GS. Let L be the loss that occurs when the BDT gives a negative result for a diseased patient, and let L′ be the loss that occurs when the BDT gives a positive result for a non-diseased patient. Losses are assumed to be zero when BDT correctly classifies a diseased patient or a non-diseased patient. Loss L is associated with a false negative and loss L′ with false positive. For example, let us consider the diagnosis of liver disease using scintigraphy as a diagnostic test. If the scintigraphy is positive for a non-disease patient (false positive), the patient will undergo a biopsy which will finally be negative. Loss L′ will be determined from the economic costs of the diagnosis, taking into account the risks, stress and anxiety caused for the patient. If the scintigraphy is negative for a disease patient (false negative), the patient may be diagnosed later. In this situation the disease can progress or get worse, decreasing the chance of successful treatment for the disease. Loss L will be determined from these considerations. Therefore, losses L and L′ are not only measured in economic terms but also in terms of risk, stress, anxiety, etc. Therefore, in practice it is not possible to determine the values of the losses L and L′. Finally, we examine the weighted kappa coefficient and the average kappa coefficient.

2.1. Weighted Kappa Coefficient

The weighted kappa coefficient $\kappa(c)$ is a measure of the beyond chance agreement between the BDT and the GS, and it is expressed [1,2] as

$$\kappa(c) = \frac{pqY}{pc(1-Q)+q(1-c)Q}, 0 \le c \le 1$$

where p is the disease prevalence, $q = 1-p$, $Y = Se + Sp - 1$ is the Youden index [9], $Q = pSe + q(1-Sp) = P(T=1)$, and $c = L/(L'+L)$ is the weighting index. The weighted kappa coefficient can also be written as

$$\kappa(c) = \frac{\kappa(0)\kappa(1)}{c\kappa(0)+(1-c)\kappa(1)}, 0 \le c \le 1. \quad (1)$$

The value of the weighting index is assumed depending on the clinician's knowledge about false positives and false negatives [1,2]. If the clinician is more concerned about false positives, as is the case in which the BDT is used as a confirmatory test prior to the application of a risk treatment (for example a surgical operation), then $L' > L$ and $0 \le c < 0.5$. For example, if the clinician decides that a false positive is three times more important than a false negative then $L' = 3L$ and $c = 1/(1+3) = 0.25$. If the clinician is more concerned about false negatives, as is the case in which the BDT is used as a screening test, then $L > L'$ and $0.5 < c \le 1$. For example, if the clinician decides that a false negative is five times more important than a false positive then $L = 5L'$ and $c = 5/(5+1) = 5/6$. Value $c = 0.5$ is used for a simple diagnosis (false positives and false negatives have the same importance), being $\kappa(0.5)$ the Cohen kappa coefficient.

The weighted kappa coefficient can be classified in the following scale of values [10]: 0–0.20, the agreement is slight; 0.21–0.40, the agreement is fair; 0.41–0.60, the agreement is moderate; 0.61–0.80, the agreement is substantial; and 0.81–1, the agreement is almost perfect. Another scale based on levels of clinical significance is [11]: <0.40, poor; 0.40–0.59, fair; 0.60–0.74, good; and 0.75–1, excellent. The weighted kappa coefficient has the following properties: (a) if $c = 0$ then $\kappa(0) = \{Sp-(1-Q)\}/Q$ and if $c = 1$ then $\kappa(1) = (Se-Q)/(1-Q)$; (b) if $Se = Sp = 1$ then $\kappa(c) = 1$, and the agreement between BDT and GS is perfect; (c) if the sensitivity and the specificity are complementary $(Se = 1-Sp)$ then $\kappa(c) = 0$, and the BDT and the GS are independent (the BDT is random and therefore not informative); (d) the weighted kappa coefficient is a function of the index c, which is increasing if $Q > p$, decreasing if $Q < p$, or equal to the Youden index if $Q = p$.

2.2. Average Kappa Coefficient

From the weighted kappa coefficient, Roldán-Nofuentes and Olvera-Porcel [3] have defined a new measure to evaluate the performance of a BDT with respect to a GS: the average kappa coefficient. For fixed values of Se, Sp and p, the weighted kappa coefficient is a continuous function of the index c. If the clinician considers that $L' > L$, and therefore $0 \le c < 0.5$, the average kappa coefficient is [3]

$$\kappa_1 = \frac{1}{0.5}\int_0^{0.5}\kappa(c)dc = \begin{cases} \frac{2\kappa(0)\kappa(1)}{\kappa(0)-\kappa(1)}\ln\left\{\frac{\kappa(0)+\kappa(1)}{2\kappa(1)}\right\}, & p \neq Q \\ Y, & p = Q, \end{cases} \quad (2)$$

i.e., the average kappa coefficient (κ_1) is the average value of $\kappa(c)$ when $0 \le c < 0.5$. If the clinician considers that $L > L'$, and therefore $0.5 < c \le 1$, the average kappa coefficient is [3]

$$\kappa_2 = \frac{1}{0.5}\int_{0.5}^{1}\kappa(c)dc = \begin{cases} \frac{2\kappa(0)\kappa(1)}{\kappa(0)-\kappa(1)}\ln\left\{\frac{2\kappa(0)}{\kappa(0)+\kappa(1)}\right\}, & p \neq Q \\ Y, & p = Q, \end{cases} \quad (3)$$

i.e., the average kappa coefficient (κ_2) is the average value of $\kappa(c)$ when $0.5 < c \leq 1$, where

$$\kappa(0) = \frac{Sp - (1 - Q)}{Q} \text{ and } \kappa(1) = \frac{Se - Q}{1 - Q}.$$

As the weighted kappa coefficient is a measure of the beyond-chance agreement between the BDT and the GS, the average kappa coefficient is a measure of the beyond-chance average agreement between the BDT and the GS, and does not depend on the weighting index c. The values of the average kappa coefficient can be classified on the same scales [10,11] as the values of the weighted kappa coefficient. The average kappa coefficients κ_1 and κ_2 have the following properties [3]:

If $Se = Sp = 1$ then $\kappa_1 = \kappa_2 = 1$, and if $Se = 1 - Sp$ then $\kappa_1 = \kappa_2 = 0$. Therefore $0 \leq \kappa_i \leq 1$, $i = 1, 2$.
Coefficient κ_1 is greater than κ_2 if $p > Q$, and κ_1 is lower than κ_2 if $Q > p$.
κ_1 minimizes the expression $2\int_0^{0.5} \{\kappa(c) - x\}^2 dc$ and κ_2 minimizes the expression $2\int_{0.5}^1 \{\kappa(c) - x\}^2 dc$. Therefore, when $x = \kappa_1$ $(x = \kappa_2)$ the first (second) expression is the variance of the weighted kappa coefficients around κ_1 (κ_2).
For fixed values of $\kappa(0)$ and $\kappa(1)$ (or Se, Sp and p), the weighted kappa coefficient is a function of c which is continuous in the interval $[0, 1]$. Therefore, the average kappa coefficient κ_i coincides with a value of the weighted kappa coefficient in the interval $[0, 1]$. This value of the weighted coefficient kappa has a value of weighting index c. So, as $\kappa_i = \kappa(c)$ for some value of c, from Equation (1) and for a specific sample it is possible to calculate a value of the weighting index c associated to the estimated average kappa coefficient. Thus, the estimation of the average kappa coefficient allows us to estimate how much greater (or smaller) the loss due to the false negatives is than the loss due to the false positives.

3. Estimation with Complete Data

When the BDT and the GS are applied to all patients in a random sample sized m, the observed frequencies in Table 1 are obtained, where the variable T models the result of the BDT ($T = 1$ when the result is positive and $T = 0$ when it is negative) and the variable D models the result of the GS ($D = 1$ when the patient has the disease and $D = 0$ when the patient does not have the disease). In Table 1, each observed frequency x_i (y_i) is the number of diseased (non-diseased) patients in which $T = i$, $x = x_1 + x_0$, $y = y_1 + y_0$, $m_i = x_i + y_i$ and $n = x + y = m_1 + m_0$, with $i = 0, 1$. In this situation the disease status (disease present or disease absent) of all patients is verified by applying the GS, and it corresponds to a cross-sectional sampling.

Table 1. Observed frequencies in the presence of complete data.

Observed Frequencies of the 2 × 2 Table			
	$T = 1$	$T = 0$	**Total**
$D = 1$	x_1	x_0	x
$D = 0$	y_1	y_0	y
Total	m_1	m_0	m

In this situation, the maximum likelihood estimator (MLE) of the weighted kappa coefficient [1,2] is

$$\hat{\kappa}(c) = \frac{x_1 y_0 - x_0 y_1}{m_0 x c + m_1 y (1 - c)}, \; 0 \leq c \leq 1,$$

and that the MLEs of $\kappa(0)$ and $\kappa(1)$ are

$$\hat{\kappa}(0) = \frac{x_1 y_0 - x_0 y_1}{m_1 y} \text{ and } \hat{\kappa}(1) = \frac{x_1 y_0 - x_0 y_1}{m_0 x}.$$

Finally, the MLEs of the average kappa coefficients κ_1 and κ_2 are [3]

$$\hat{\kappa}_1 = \begin{cases} \frac{2(x_1y_0 - x_0y_1)}{m_0x - m_1y} \ln\left\{\frac{m_1y + m_0x}{2m_1y}\right\}, & x_0 \neq y_1 \\ \frac{x_1y_0 - x_0y_1}{xy}, & x_0 = y_1, \end{cases}$$

and

$$\hat{\kappa}_2 = \begin{cases} \frac{2(x_1y_0 - x_0y_1)}{m_0x - m_1y} \ln\left\{\frac{2m_0x}{m_1y + m_0x}\right\}, & x_0 \neq y_1 \\ \frac{x_1y_0 - x_0y_1}{xy}, & x_0 = y_1, \end{cases}$$

respectively.

If $x_0 = y_1 = 0$ then κ_i cannot be estimated. If $x_1y_0 = x_0y_1$ then $\hat{\kappa}_i = 0$. If $x_1y_0 < x_0y_1$, or if $x_1 = 0$ or $y_0 = 0$, then $\hat{Y} < 0$ and it is necessary to interchange the results of the BDT (the positive result should be $T = 0$ and the negative result should be $T = 1$). A fundamental analysis in inference statistics is formign a confidence interval (CI) for an unknown parameter. In this context and with respect to the average kappa coefficient, Roldán-Nofuentes and Olvera-Porcel [3] have studied various CIs for κ_1 and κ_2. These CIs are approximate and their asymptotic behaviors have been studied through simulation experiments. Following this work, two confidence intervals (CIs) for κ_1 and κ_2 studied by Roldán-Nofuentes and Olvera Porcel (Wald CI and logit CI) are summarized and a new CI (arcsine CI) is also presented.

3.1. Wald CI

Based on the asymptotic normality of $(\hat{\kappa}_i - \kappa_i)/\sqrt{\hat{V}ar(\hat{\kappa}_i)}$, i.e., $(\hat{\kappa}_i - \kappa_i)/\sqrt{\hat{V}ar(\hat{\kappa}_i)} \to N(0,1)$ when m is large, the $100(1-\alpha)\%$ Wald CI for κ_i is [3]

$$\kappa_i \in \hat{\kappa}_i \pm z_{1-\alpha/2}\sqrt{\hat{V}ar(\hat{\kappa}_i)},\ i = 1,2,$$

where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the normal standard distribution. Expressions of the estimated variances are shown in Appendix A.

3.2. Logit CI

Based on the logit transformation of $\hat{\kappa}_i$, $\ln[\hat{\kappa}_i/(1-\hat{\kappa}_i)]$, is closer to a normal distribution with mean $\ln[\kappa_i/(1-\kappa_i)]$, the $100(1-\alpha)\%$ CI for the logit of κ_i is

$$logit(\hat{\kappa}_i) \pm z_{1-\alpha/2}\sqrt{\hat{V}ar[logit(\hat{\kappa}_i)]},\ i = 1,2,$$

Taking exponential in this expression, the $100(1-\alpha)\%$ logit CI for κ_i is [3]

$$\kappa_i \in \left(\frac{\exp\left\{logit(\hat{\kappa}_i) - z_{1-\alpha/2}\sqrt{\hat{V}ar(logit(\hat{\kappa}_i))}\right\}}{1+\exp\left\{logit(\hat{\kappa}_i) - z_{1-\alpha/2}\sqrt{\hat{V}ar(logit(\hat{\kappa}_i))}\right\}},\ \frac{\exp\left\{logit(\hat{\kappa}_i) + z_{1-\alpha/2}\sqrt{\hat{V}ar(logit(\hat{\kappa}_i))}\right\}}{1+\exp\left\{logit(\hat{\kappa}_i) + z_{1-\alpha/2}\sqrt{\hat{V}ar(logit(\hat{\kappa}_i))}\right\}} \right),\ i = 1,2,$$

where the estimated variance is obtained by applying the delta method, i.e.,

$$\hat{V}ar(logit(\hat{\kappa}_i)) = \frac{\hat{V}ar(\hat{\kappa}_i)}{\hat{\kappa}_i^2(1-\hat{\kappa}_i)^2},\ i = 1,2.$$

3.3. Arcsine CI

The arcsine is a transformation that has been used to estimate parameters, for example, see the work of Martín-Andrés et al. [12] on the estimation of a binomial proportion. A new CI for κ_i can be obtained by applying this transformation. Based on the asymptotic normal-

ity of $\left(\sin^{-1}\sqrt{\hat{\kappa}_i} - \sin^{-1}\sqrt{\kappa_i}\right) / \sqrt{\hat{V}ar\left(\sin^{-1}\sqrt{\hat{\kappa}_i}\right)}$, i.e., $\left(\sin^{-1}\sqrt{\hat{\kappa}_i} - \sin^{-1}\sqrt{\kappa_i}\right) / \sqrt{\hat{V}ar\left(\sin^{-1}\sqrt{\hat{\kappa}_i}\right)} \to N(0,1)$ when m is large, the $100(1-\alpha)\%$ CI for $\sin^{-1}\sqrt{\kappa_i}$ is

$$\sin^{-1}\sqrt{\kappa_i} \in \sin^{-1}\sqrt{\hat{\kappa}_i} \pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\sin^{-1}\sqrt{\hat{\kappa}_i}\right)},\ i = 1,2,$$

where the variance $\hat{V}ar\left(\sin^{-1}\sqrt{\hat{\kappa}_i}\right)$ is easily obtained by applying the delta method, i.e.,

$$\hat{V}ar\left(\sin^{-1}\sqrt{\hat{\kappa}_i}\right) = \frac{\hat{V}ar(\hat{\kappa}_i)}{4\hat{\kappa}_i(1-\hat{\kappa}_i)},\ i = 1,2.$$

Finally, undoing the transformation, the $100(1-\alpha)\%$ arcsine CI for κ_i is

$$\kappa_i \in \sin^2\left\{\sin^{-1}\sqrt{\hat{\kappa}_i} \pm \frac{z_{1-\alpha/2}}{2\sqrt{\hat{\kappa}_i(1-\hat{\kappa}_i)}}\sqrt{\hat{V}ar(\hat{\kappa}_i)}\right\},\ i = 1,2.$$

4. Estimation in the Presence of Partial Verification

The evaluation of a BDT in the presence of partial verification gives the frequencies in Table 2, where the variables T and D are the same as in Section 3, and the variable V models the verification process, i.e., $V = 1$ when the disease status of a patient is verified with the GS and $V = 0$ when it is not.

Table 2. Observed frequencies in the presence of partial verification.

Observed Frequencies of the 3 × 2 Table			
	$T = 1$	$T = 0$	**Total**
$V = 1$			
$D = 1$	s_1	s_0	s
$D = 0$	r_1	r_0	r
$V = 0$	u_1	u_0	u
Total	n_1	n_0	n

Let λ_{ij} be the probability of verifying the disease status of a patient with the GS in which $T = i$ and $D = j$, i.e.,

$$\lambda_{ij} = P(V = 1 | T = i, D = j),\ i,j = 0,1.$$

Assuming that the missing data mechanism is missing at random (MAR) [13], then

$$\lambda_{ij} = \lambda_i = P(V = 1 | T = i),\ i,j = 0,1.$$

The MAR assumption takes that the verification process only depends on the result of the BDT and not the GS. This circumstance obtains in two-phase studies: in the first phase, the BDT is applied to all patients in the sample; in the second phase, the GS is applied to only a subset of patients in the sample, depending only on the result of the BDT. Subject to the MAR assumption, the observed frequencies $(s_1, r_1, u_1, s_0, r_0, u_0)$ are the product of a multinomial distribution whose probabilities are:

$$\begin{aligned} \xi_i &= P(V = 1, D = 1, T = i) = p\lambda_i Se^i(1-Se)^{1-i} \\ \psi_i &= P(V = 1, D = 0, T = i) = q\lambda_i Sp^{1-i}(1-Sp)^i \\ \zeta_i &= P(V = 0, T = i) = \tfrac{1-\lambda_i}{\lambda_i}(\xi_i + \psi_i). \end{aligned} \tag{4}$$

Next, estimation of the average kappa coefficient applying the maximum likelihood (ML) method and applying multiple imputation (MI) is studied.

4.1. Maximum Likelihood

Assuming that the missing data mechanism is MAR the MLEs of sensitivity, specificity and prevalence in the presence of partial verification are [4,5]

$$\hat{Se}_{pv} = \frac{s_1 n_1/(s_1+r_1)}{s_1 n_1/(s_1+r_1) + s_0 n_0/(s_0+r_0)},\ \hat{Sp}_{pv} = \frac{r_0 n_0/(s_0+r_0)}{r_1 n_1/(s_1+r_1) + r_0 n_0/(s_0+r_0)},$$

and

$$\hat{p}_{pv} = \frac{s_1 n_1/(s_1+r_1) + s_0 n_0/(s_0+r_0)}{n}.$$

Substituting in Equations (2) and (3) parameters with their MLEs in the presence of partial verification, the MLEs of κ_1 and κ_2 in the presence of partial verification are

$$\hat{\kappa}_{1pv} = \frac{2\hat{p}_{pv}\hat{q}_{pv}\hat{Y}_{pv}}{\hat{p}_{pv}-\hat{Q}_{pv}} \ln\left(\frac{\hat{p}_{pv}+\hat{Q}_{pv}-2\hat{p}_{pv}\hat{Q}_{pv}}{2\hat{q}_{pv}\hat{Q}_{pv}}\right) \text{ and } \hat{\kappa}_{2pv} = \frac{2\hat{p}_{pv}\hat{q}_{pv}\hat{Y}_{pv}}{\hat{p}_{pv}-\hat{Q}_{pv}} \ln\left(\frac{2\hat{p}_{pv}\hat{Q}_{pv}}{\hat{p}_{pv}+\hat{Q}_{pv}-2\hat{p}_{pv}\hat{Q}_{pv}}\right)$$

when $\hat{p}_{pv} \neq \hat{Q}_{pv}$, and

$$\hat{\kappa}_{1pv} = \hat{\kappa}_{2pv} = \hat{Y}_{pv} = \frac{n_1 n_0 (s_1+r_1)(s_0+r_0)(s_1 r_0 - s_0 r_1)}{\{n_1 r_1 (s_0+r_0) + n_0 r_0 (s_1+r_1)\}\{n_1 s_1 (s_0+r_0) + n_0 s_0 (s_1+r_1)\}}$$

when $\hat{p}_{pv} = \hat{Q}_{pv}$, where $\hat{Q}_{pv} = n_1/n$. The expressions of the estimators $\hat{\kappa}_{1pv}$ and $\hat{\kappa}_{2pv}$ are long and complicated when $\hat{p}_{pv} \neq \hat{Q}$, so statistical software is necessary to calculate them (see Section 6). Next, three asymptotic CIs for κ_i in the presence of partial verification are proposed.

4.1.1. Wald CI

Based on the asymptotic normality of $(\hat{\kappa}_{ipv} - \kappa_i)/\sqrt{\hat{V}ar(\hat{\kappa}_{ipv})}$, the $100(1-\alpha)\%$ Wald CI for κ_i is

$$\kappa_i \in \hat{\kappa}_{ipv} \pm z_{1-\alpha/2}\sqrt{\hat{V}ar(\hat{\kappa}_{ipv})},\ i = 1, 2.$$

The expressions of the estimated variances are shown in Appendix B. These expressions are long and complicated, so it is necessary to use a statistical program to calculate them (see Section 6).

4.1.2. Logit CI

The logit CI is based on the asymptotic normality of the logit of $\{logit(\hat{\kappa}_{ipv}) - logit(\kappa_i)\}/\sqrt{\hat{V}ar\{logit(\hat{\kappa}_{ipv})\}}$. The logit CI for κ_i has a general expression similar to that obtained in Section 3.2, although the expressions for the estimators and the variances are different. The expressions of the variances are shown in Appendix B, and it is necessary to use a statistical program to calculate them.

4.1.3. Arcsine CI

The arcsine CI is also based on the asymptotic normality of $\left(\sin^{-1}\sqrt{\hat{\kappa}_{ipv}} - \sin^{-1}\sqrt{\kappa_i}\right)/\sqrt{\hat{V}ar\left(\sin^{-1}\sqrt{\hat{\kappa}_{ipv}}\right)}$ and its general expression is similar to that given in Section 3.3, where the variances are shown in Appendix B.

4.2. Multiple Imputation

Multiple imputation (MI) [14–17] is a computational method used to solve estimation problems with missing data. MI consists of constructing M complete data sets, obtained by replacing the missing data with M independent imputed sets. In each complete data set, the estimators of the parameters and their standard errors are calculated, and these are combined appropriately to calculate the global estimators, their standard errors and their confidence intervals. Harel and Zhou [6] have applied MI to estimate the sensitivity

(specificity) of a BDT in the presence of partial verification and have shown that this method provides CIs with better asymptotic behavior than the CIs obtained by applying the ML method. Montero-Alonso and Roldán-Nofuentes [18] have studied the estimation of the likelihood ratios of two BDTs in the presence of partial verification using the MI by chained equations (MICE) method and have also shown that this method provides CIs with better asymptotic behavior.

In our context, from the 3×2 table given in Table 2, M 2×2 tables are imputed (as in Table 1), and from each one of these M tables the estimator of κ_i, its standard error and the CIs given in Section 3 are calculated. The M results are then combined by applying the Rubin rules [14] and, in this way, the CI for κ_i is calculated. Regarding the imputation of missing data, MICE method was used. MICE method requires the MAR assumption and can be used with different types of variables. In the problem posed in this article there are two binary random variables: variable T and variable D. The work by White et al. [19] explains in detail the imputation of binary variables using the MICE method. For variable T there are no missing data since BDT is applied to all patients. However, variable D is not observed in all patients and therefore this variable has missing data. Firstly, all missing values are filled in at random. Variable D is then regressed on the variable T through a logistic regression. The estimation is thus restricted to individuals with observed T. Missing values in D are then replaced by simulated draws from the posterior predictive distribution of variable D. This process is called a cycle, and in order to stabilize the results the process is repeated for a determined number of cycles in order to obtain a set of imputed data. Applying multiple imputation, the estimator of κ_i is the mean of the estimators obtained in M complete data sets, and their standard errors are calculated by applying the Rubin rules [14]. In the situation studied in this article, the application MICE requires that $s_i > 0$ and $r_i > 0$.

5. Simulation Experiments

Monte Carlo simulation experiments have been carried out to study the asymptotic behavior (coverage probability and average length) of the CIs studied in Section 4. The relative biases of the estimators of the average kappa coefficients obtained through ML and through MI have also been studied. These experiments consisted of the generation of 10,000 random samples of multinomial distributions sized $n = \{50, 100, 200, 500, 1000\}$, and whose probabilities have been calculated from equations. These probabilities have been calculated in the following way: with respect to verification probabilities, we have taken two sets of values, $(\lambda_1 = 0.70, \lambda_0 = 0.25)$ and $(\lambda_1 = 0.95, \lambda_0 = 0.40)$, which can be considered low and high verification probability values. As values of disease prevalence we took the values $p = \{10\%, 30\%, 50\%, 70\%\}$ and as values of κ_1 and κ_2 we took the values $\{0.20, 0.40, 0.60, 0.80\}$. Once we have set the values of κ_1 and κ_2, the values of $\kappa(0)$ and $\kappa(1)$ are obtained solving with the Newton-Raphson method the system made by Equations (1) and (2), only considering those values whose solutions are between 0 and 1. Once we have obtained the values of $\kappa(0)$ and $\kappa(1)$, as the value prevalence p has been set previously, the values of Se and Sp are calculated solving the system made by equations $\kappa(0) = \{Sp - (1 - Q)\}/Q$ and $\kappa(1) = (Se - Q)/(1 - Q)$, and then the probabilities of the multinomial distributions are calculated. Therefore, the samples have been generated by fixing κ_1 and κ_2. The random samples have been generated in such a way that κ_1 and κ_2 and their standard errors can be estimated in all of them, and also verifying that $\hat{\kappa}_i > 0$ (and, in this way, to be able to calculate all CIs). For example, if, in a sample, a frequency s_i or r_i is equal to 0, then MICE cannot be applied; in this situation this sample has been ruled out and another one has been generated instead until we have obtained 10,000 samples. The simulation experiments have been carried out using the R program [20] and the "mice" library [21]. Regarding MICE, this has been carried out using $M = 20$ data sets and performing 100 cycles. The $M = 20$ complete data sets are generated in such a way that κ_1 and κ_2 (and their standard errors) can be estimated in all of them. Thus, for example, if, in a complete data set $\hat{\kappa}_i < 0$, then that complete data set is neglected and

another is generated in its place, and so on until obtaining 20 complete data sets. In a first phase of these experiments, we have considered $M = 20$ and $M = 50$ complete data sets and we have also considered 100 and 200 cycles in each case, obtaining very similar results. Therefore, we have considered $M = 20$ and 100 cycles to save computation time and stabilize the results. These 20 complete data sets have been generated in such a way that κ_1 and κ_2 and their standard errors can be estimated in all of them, verifying that each estimate of κ_i is greater than 0. In each sample generated, we have calculated the three CIs (95% confidence) given in Section 3 along with the MICE method and the three CIs given in Section 4.1. Finally, we have calculated the coverage probabilities and the average lengths of the CIs in each scenario. The relative biases of the estimators of κ_1 and κ_2 obtained through ML and through MICE have also been calculated.

Tables 3 and 4 show some of the results obtained for $\kappa_1 = \{0.2, 0.4, 0.6, 0.8\}$, indicating in each case the values of Se, Sp and p.

Table 3. Coverage probabilities and average lengths of CIs for $\kappa_1 = \{0.2, 0.4\}$.

κ_1 = 0.2 Se = 0.7773 Sp = 0.7308 p = 10%														
						$\lambda_1 = 0.70$ $\lambda_0 = 0.25$								
		Maximum Likelihood Method							MICE Method					
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−13.6	0.998	0.492	0.994	0.721	0.875	0.527	−16.7	0.998	0.481	0.999	0.831	0.946	0.577
100	−11.4	0.982	0.372	0.989	0.486	0.973	0.394	−14.3	0.982	0.372	0.996	0.649	0.979	0.437
200	−7.5	0.960	0.276	0.988	0.302	0.985	0.277	−10.6	0.954	0.293	0.996	0.413	0.993	0.315
500	−3.5	0.942	0.173	0.972	0.174	0.957	0.172	−6.1	0.948	0.187	0.978	0.199	0.965	0.189
1000	−1.8	0.948	0.121	0.963	0.122	0.956	0.121	−2.4	0.948	0.128	0.971	0.131	0.962	0.129
						$\lambda_1 = 0.95$ $\lambda_0 = 0.40$								
		Maximum Likelihood Method							MICE Method					
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−13.5	0.988	0.425	0.993	0.602	0.946	0.460	−15.3	0.988	0.420	0.995	0.751	0.952	0.503
100	−9.2	0.960	0.322	0.984	0.380	0.979	0.330	−11.7	0.953	0.329	0.991	0.509	0.980	0.367
200	−5.6	0.953	0.232	0.988	0.239	0.981	0.230	−7.8	0.948	0.242	0.993	0.279	0.987	0.246
500	−2.7	0.951	0.146	0.962	0.146	0.954	0.145	−3.9	0.950	0.151	0.971	0.153	0.962	0.150
1000	−0.5	0.947	0.102	0.956	0.102	0.953	0.102	−1.1	0.951	0.104	0.956	0.105	0.953	0.104
κ_1 = 0.4 Se = 0.7413 Sp = 0.7441 p = 30%														
						$\lambda_1 = 0.70$ $\lambda_0 = 0.25$								
		Maximum Likelihood Method							MICE Method					
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−18.1	0.979	0.630	0.996	0.655	0.973	0.618	−20.9	0.979	0.624	0.997	0.740	0.965	0.653
100	−10.1	0.963	0.476	0.994	0.470	0.988	0.464	−13.8	0.952	0.499	0.995	0.554	0.973	0.509
200	−4.5	0.961	0.340	0.989	0.330	0.977	0.333	−6.2	0.947	0.365	0.984	0.373	0.970	0.364
500	−1.5	0.952	0.213	0.961	0.210	0.956	0.211	−2.6	0.949	0.225	0.960	0.224	0.955	0.224
1000	−1.1	0.954	0.150	0.959	0.149	0.958	0.149	−1.5	0.950	0.158	0.955	0.158	0.951	0.158
						$\lambda_1 = 0.95$ $\lambda_0 = 0.40$								
		Maximum Likelihood Method							MICE Method					
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−15.5	0.960	0.550	0.993	0.561	0.982	0.539	−18.4	0.955	0.559	0.996	0.638	0.989	0.575
100	−8.2	0.955	0.405	0.985	0.393	0.980	0.395	−10.9	0.950	0.421	0.991	0.431	0.985	0.418
200	−3.4	0.956	0.283	0.974	0.277	0.967	0.279	−5.1	0.956	0.294	0.980	0.290	0.967	0.290
500	−1.1	0.947	0.178	0.957	0.176	0.952	0.177	−1.3	0.950	0.182	0.963	0.181	0.957	0.181
1000	−0.6	0.955	0.125	0.958	0.125	0.957	0.125	−0.7	0.951	0.128	0.958	0.128	0.955	0.128

CP: coverage probability. AL: average length.

Table 4. Coverage probabilities and average lengths of CIs for $\kappa_1 = \{0.6, 0.8\}$.

κ_1 = 0.6 *Se* = 0.6816 *Sp* = 0.8624 *p* = 50%														
$\lambda_1 = 0.70$ $\lambda_0 = 0.25$														
	Maximum Likelihood Method							MICE Method						
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−17.4	0.984	0.701	1	0.638	0.997	0.658	−20.3	0.989	0.714	1	0.694	0.977	0.692
100	−8.9	0.969	0.508	0.994	0.471	0.987	0.485	−11.5	0.963	0.539	0.993	0.514	0.981	0.521
200	−4.9	0.963	0.358	0.974	0.343	0.968	0.349	−6.3	0.955	0.384	0.973	0.371	0.964	0.376
500	−2.0	0.946	0.224	0.952	0.221	0.950	0.222	−2.7	0.950	0.238	0.956	0.235	0.954	0.237
1000	−0.6	0.953	0.157	0.954	0.156	0.954	0.156	−0.8	0.951	0.165	0.953	0.165	0.953	0.166
$\lambda_1 = 0.95$ $\lambda_0 = 0.40$														
	Maximum Likelihood Method							MICE Method						
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−13.5	0.973	0.600	1	0.555	0.991	0.568	−15.2	0.973	0.608	1	0.587	0.971	0.593
100	−6.7	0.958	0.420	0.980	0.398	0.967	0.407	−7.2	0.952	0.433	0.986	0.414	0.968	0.421
200	−3.1	0.960	0.293	0.968	0.285	0.962	0.289	−3.6	0.954	0.303	0.968	0.295	0.963	0.299
500	−1.5	0.954	0.184	0.958	0.182	0.956	0.183	−1.7	0.950	0.187	0.951	0.187	0.950	0.188
1000	−0.4	0.952	0.130	0.953	0.130	0.953	0.130	−0.5	0.950	0.133	0.953	0.133	0.953	0.133
κ_1 = 0.8 *Se* = 0.7969 *Sp* = 0.9707 *p* = 70%														
$\lambda_1 = 0.70$ $\lambda_0 = 0.25$														
	Maximum Likelihood Method							MICE Method						
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−17.9	0.990	0.646	0.987	0.596	0.994	0.612	−20.2	0.987	0.682	0.978	0.640	0.976	0.652
100	−9.2	0.979	0.434	0.948	0.418	0.970	0.417	−10.7	0.978	0.471	0.947	0.455	0.964	0.453
200	−4.7	0.969	0.291	0.949	0.288	0.959	0.285	−5.8	0.971	0.322	0.952	0.316	0.961	0.313
500	−1.8	0.961	0.179	0.964	0.180	0.964	0.180	−2.1	0.961	0.186	0.954	0.187	0.957	0.187
1000	−0.6	0.959	0.123	0.952	0.122	0.956	0.122	−0.7	0.957	0.134	0.951	0.134	0.954	0.133
$\lambda_1 = 0.95$ $\lambda_0 = 0.40$														
	Maximum Likelihood Method							MICE Method						
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−13.2	0.997	0.535	0.949	0.504	0.965	0.509	−14.8	0.971	0.551	0.957	0.523	0.968	0.527
100	−6.9	0.973	0.349	0.946	0.341	0.961	0.339	−7.5	0.968	0.363	0.944	0.355	0.963	0.352
200	−3.3	0.965	0.231	0.949	0.230	0.961	0.227	−3.6	0.967	0.240	0.953	0.241	0.959	0.239
500	−1.4	0.961	0.141	0.952	0.141	0.959	0.140	−1.5	0.956	0.148	0.945	0.147	0.950	0.147
1000	−0.6	0.953	0.099	0.951	0.099	0.952	0.099	−0.6	0.950	0.102	0.945	0.102	0.946	0.102

CP: coverage probability. AL: average length.

From the results of these experiments we reach the following conclusions:

(a) With respect to ML, the verification probabilities do not have a clear effect on the coverage probabilities (CPs) of the CIs. With respect to the CIs, in general terms their CPs far exceed 95% when the sample size is small ($n = 50$) or moderate (n = 100–200), fluctuating around 95% when the sample size is large (n = 500–1000). The Wald CI has a CP that fluctuates around 95% when the sample size is moderate or large. The logit CI has a higher CP than that of the Wald CI, especially when the sample size is small or moderate. The arcsine CI can have a CP of less than 90% when the sample size is small and κ_1 is small ($\kappa_1 = 0.2$) and fluctuates around 95% when the sample size is large. In general terms, the Wald CI is the interval with the best performance when the sample size is small or moderate, while all three CIs have a very similar asymptotic behavior when the sample size is large.

(b) With respect to MICE, the verification probabilities do not have a clear effect on the CPs of the CIs. The Wald CI has a coverage probability that exceeds 95% when the sample size is small or moderate and the value of κ_1 is small ($\kappa_1 = 0.2$), fluctuating around 95% in the other situations and sample sizes. The logit CI has a CP that is slightly higher than that of the Wald CI, especially when the sample size is small or moderate. The arcsine CI has a CP closer to 95% when the sample size is small, and in the rest of sample size its CP is slightly higher than that of the Wald CI.

(c) Comparing the CIs obtained by ML and those obtained by MICE, MICE along with the Wald CI presents, in general terms, better fluctuations around 95% than any of the CIs obtained by ML; once MICE, along with the Wald CI, reaches a CP of 95%, it fluctuates very slightly around 95%. Furthermore, in general terms, MICE along with the Wald CI begins to fluctuate around 95% with a sample size smaller than the CIs by ML. Regarding the average lengths, the CIs obtained by ML have an average length slightly less than that of the CIs obtained by applying MICE when the sample size is small or moderate, although the latter show better fluctuations around 95%. The average lengths are very similar when the sample size is large.

(d) Regarding the comparison of the estimators obtained by ML and MICE, relative biases are very similar. Difference (in absolute value) is small (less than 5%) when the sample size is small, and the difference is very small (less than 1%) when the sample size is large. Therefore, ML and MICE provide estimators of κ_1 that are, on average, very similar.

Tables 5 and 6 show some of the results obtained for $\kappa_2 = \{0.2, 0.4, 0.6, 0.8\}$. In very general terms, very similar conclusions are obtained to those obtained for the ICs of κ_1.

Table 5. Coverage probabilities and average lengths of CIs for $\kappa_2 = \{0.2, 0.4\}$.

κ_2 = 0.2 *Se* = 0.5904 *Sp* = 0.6901 *p* = 70%														
$\lambda_1 = 0.70$ $\lambda_0 = 0.25$														
	Maximum Likelihood Method							MICE Method						
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−15.8	0.984	0.597	0.976	0.744	0.883	0.610	−19.4	0.981	0.616	0.978	0.795	0.947	0.662
100	−7.3	0.985	0.465	0.971	0.583	0.943	0.486	−10.7	0.971	0.490	0.976	0.663	0.972	0.529
200	−2.9	0.958	0.357	0.960	0.418	0.967	0.365	−5.1	0.953	0.381	0.967	0.500	0.967	0.401
500	−1.7	0.945	0.241	0.960	0.248	0.963	0.239	−1.9	0.949	0.260	0.963	0.281	0.962	0.261
1000	−0.7	0.949	0.172	0.968	0.173	0.955	0.171	−0.8	0.950	0.180	0.960	0.185	0.957	0.182
$\lambda_1 = 0.95$ $\lambda_0 = 0.40$														
	Maximum Likelihood Method							MICE Method						
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−11.1	0.979	0.511	0.970	0.662	0.907	0.530	−13.6	0.980	0.530	0.973	0.728	0.953	0.577
100	−6.1	0.971	0.395	0.968	0.485	0.960	0.409	−8.2	0.966	0.412	0.972	0.560	0.967	0.440
200	−2.3	0.955	0.299	0.961	0.326	0.974	0.301	−4.1	0.953	0.312	0.971	0.375	0.977	0.321
500	−1.1	0.937	0.196	0.961	0.197	0.951	0.194	−1.4	0.947	0.206	0.965	0.214	0.962	0.206
1000	−0.5	0.956	0.138	0.962	0.139	0.959	0.138	−0.6	0.951	0.147	0.959	0.148	0.957	0.146
κ_2 = 0.4 *Se* = 0.7773 *Sp* = 0.7308 *p* = 10%														
$\lambda_1 = 0.70$ $\lambda_0 = 0.25$														
	Maximum Likelihood Method							MICE Method						
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−36.5	0.999	0.743	1	0.842	0.891	0.701	−38.8	0.999	0.683	1	0.892	0.967	0.738
100	−26.1	0.981	0.630	1	0.680	0.983	0.631	−29.3	0.963	0.598	1	0.778	0.991	0.649
200	−16.3	0.964	0.496	0.998	0.493	0.995	0.486	−19.2	0.943	0.516	0.995	0.597	0.997	0.534
500	−7.2	0.954	0.320	0.984	0.312	0.969	0.315	−9.5	0.949	0.356	0.989	0.361	0.968	0.354
1000	−3.9	0.957	0.226	0.966	0.223	0.962	0.224	−4.4	0.951	0.247	0.971	0.247	0.958	0.247
$\lambda_1 = 0.95$ $\lambda_0 = 0.40$														
	Maximum Likelihood Method							MICE Method						
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−31.5	0.999	0.684	0.999	0.762	0.984	0.678	−35.3	0.999	0.652	1.000	0.847	0.976	0.702
100	−19.7	0.969	0.556	0.999	0.571	0.992	0.548	−23.2	0.951	0.559	1.000	0.674	0.992	0.586
200	−10.9	0.964	0.415	0.996	0.402	0.994	0.404	−13.4	0.951	0.437	0.999	0.458	0.992	0.437
500	−5.3	0.954	0.261	0.970	0.256	0.962	0.258	−7.6	0.952	0.278	0.978	0.277	0.966	0.277
1000	−1.9	0.945	0.184	0.954	0.182	0.949	0.183	−2.3	0.951	0.193	0.961	0.191	0.959	0.193

CP: coverage probability. AL: average length.

Table 6. Coverage probabilities and average lengths of CIs for $\kappa_2 = \{0.6, 0.8\}$.

κ_2 = 0.6 *Se* = 0.8864 *Sp* = 0.6746 *p* = 30%														
							$\lambda_1 = 0.70$ $\lambda_0 = 0.25$							
		Maximum likelihood method							MICE Method					
n	Relative	Wald CI		logit CI		Arcsine CI		Relative	Wald CI		logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−32.9	0.97	0.799	1	0.74	0.987	0.742	−35.6	0.973	0.767	1	0.794	0.968	0.762
100	−18.1	0.971	0.61	1	0.555	0.996	0.576	−21.9	0.944	0.649	0.997	0.629	0.972	0.629
200	−9.8	0.970	0.417	0.984	0.394	0.976	0.404	−12.3	0.955	0.470	0.981	0.450	0.966	0.458
500	−3.8	0.960	0.254	0.966	0.248	0.964	0.251	−4.9	0.956	0.278	0.960	0.278	0.958	0.281
1000	−2.2	0.945	0.177	0.949	0.176	0.949	0.176	−2.9	0.948	0.187	0.951	0.187	0.949	0.187
							$\lambda_1 = 0.95$ $\lambda_0 = 0.40$							
		Maximum likelihood method							MICE Method					
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−22.9	0.978	0.694	1	0.633	0.996	0.652	−26.4	0.967	0.709	1	0.701	0.973	0.692
100	−12.6	0.966	0.487	0.996	0.454	0.978	0.468	−16.1	0.956	0.531	0.995	0.507	0.969	0.514
200	−6.2	0.967	0.331	0.976	0.319	0.969	0.324	−8.8	0.959	0.360	0.972	0.348	0.971	0.350
500	−2.4	0.956	0.203	0.960	0.200	0.959	0.201	−3.3	0.955	0.216	0.956	0.212	0.957	0.215
1000	−1.3	0.960	0.142	0.957	0.142	0.956	0.142	−1.5	0.956	0.150	0.957	0.150	0.957	0.150
κ_2 = 0.8 *Se* = 0.8644 *Sp* = 0.9817 *p* = 50%														
							$\lambda_1 = 0.70$ $\lambda_0 = 0.25$							
		Maximum likelihood method							MICE Method					
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−20.6	0.938	0.688	0.933	0.644	0.931	0.662	−23.8	0.935	0.711	0.933	0.672	0.942	0.695
100	−10.6	0.956	0.495	0.912	0.492	0.942	0.485	−13.1	0.949	0.531	0.922	0.525	0.941	0.519
200	−5.5	0.958	0.356	0.936	0.356	0.954	0.348	−7.2	0.951	0.392	0.942	0.391	0.954	0.382
500	−2.3	0.960	0.228	0.954	0.228	0.959	0.228	−3.0	0.953	0.235	0.946	0.234	0.950	0.233
1000	−1.1	0.953	0.158	0.956	0.158	0.953	0.157	−1.5	0.949	0.175	0.950	0.175	0.949	0.174
							$\lambda_1 = 0.95$ $\lambda_0 = 0.40$							
		Maximum likelihood method							MICE Method					
n	Relative	Wald CI		Logit CI		Arcsine CI		Relative	Wald CI		Logit CI		Arcsine CI	
	Bias (%)	CP	AL	CP	AL	CP	AL	Bias (%)	CP	AL	CP	AL	CP	AL
50	−13.8	0.957	0.567	0.915	0.553	0.938	0.553	−16.1	0.965	0.59	0.924	0.537	0.943	0.573
100	−6.7	0.963	0.399	0.933	0.401	0.958	0.392	−7.9	0.952	0.422	0.933	0.423	0.945	0.418
200	−3.6	0.943	0.285	0.935	0.285	0.942	0.279	−4.4	0.947	0.301	0.937	0.302	0.943	0.296
500	−1.3	0.954	0.178	0.944	0.179	0.949	0.177	−1.6	0.950	0.190	0.945	0.191	0.946	0.188
1000	−0.7	0.949	0.126	0.949	0.126	0.947	0.126	−0.8	0.950	0.132	0.953	0.132	0.950	0.132

CP: coverage probability. AL: average length.

6. Function Eakcpv

We have written a function in R [20], called "eakcpv" (Estimation of the Average Kappa Coefficient in the presence of Partial Verification), to estimate the average kappa coefficient of a BDT in the presence of partial disease verification. The command to run the "eakcpv" function is "eakcpv$(s_1, r_1, u_1, s_0, r_0, u_0, conf, imp, cycl)$", where $(s_1, r_1, u_1, s_0, r_0, u_0)$ are the observed frequencies, "conf" is the confidence level, "imp" is the number of complete data sets and "cycl" is the number of cycles. The complete data sets are generated in such a way that κ_1 and κ_2 (and their standard errors) can be estimated in all of them. Thus, for example, if, in a complete data, set $\hat{\kappa}_i < 0$, then that complete data set is neglected and another is generated in its place, and so on until obtaining "imp" complete data sets. The function always checks that the values are valid and that the analysis can be performed (e.g., no frequency s_i or r_i is equal to 0, etc.). The function estimates κ_1 and κ_2 applying MICE, along with the Wald and arcsine CIs. The function estimates the relative loss between false positives and false negatives, and also estimates how much greater (or less) the loss associated with a false positive is than the loss associated with a false negative. The results obtained are recorded in a file called "results_eakcpv.txt" in the same folder from which the function is run. The function "eakcpv" is available as Supplementary Materials of this manuscript.

7. Example

The results obtained have been applied to the study of Drum and Christacopoulos [22] on the diagnosis of liver disease. Drum and Christacopoulos [22] have studied the diagnosis of liver disease using a hepatic scintigraphy as BDT and a biopsy as GS. In Table 7, we show the observed frequencies, where variable T models the result of the hepatic scintigraphy, variable V models the verification process and variable D models the result of the biopsy.

Table 7. Diagnosis of liver disease.

Observed Frequencies of the Study of Drum and Christacopoulos		
	$T=1$	$T=0$
$V=1$		
$D=1$	231	27
$D=0$	32	54
$V=0$	166	140
Total	429	221

Running the "eakcpv" function with the command

$$\text{eakcpv}(231, 32, 166, 27, 54, 140, 0.95, 20, 100),$$

it is obtained that $\hat{\kappa}_{pv}(0) = 0.597$ and $\hat{\kappa}_{pv}(1) = 0.507$. With respect to κ_1, it is obtained that $\hat{\kappa}_{1mice} = 0.572$, its standard error is 0.059 and the 95% Wald CI for κ_1 is $(0.452\ ,\ 0.691)$. The estimated relative loss between the false positives and the false negatives is $\hat{c} = 0.252$, and the loss associated with the false positives (L') is 2.97 times greater than the loss associated with the false negatives (L). With respect to κ_2, it is obtained that $\hat{\kappa}_{2mice} = 0.526$, its standard error is 0.066 and the 95% Wald CI for κ_2 is $(0.393\ ,\ 0.660)$. Estimated relative loss between the false positives and the false negatives is $\hat{c} = 0.752$, and the loss associated with the false negatives (L) is 3.03 times greater than the loss associated with the false positives (L').

When hepatic scintigraphy is to be used as a confirmatory test prior to risky treatment ($L' > L$ and $0 \leq c < 0.5$), the beyond-chance average agreement between the hepatic scintigraphy and the biopsy is moderate ($\hat{\kappa}_{1mice} = 0.572$), and in terms of the Wald CI, the beyond-chance average agreement between the hepatic scintigraphy and the biopsy is a value between moderate and substantial (95% confidence). Estimated relative loss between the false positives and the false negatives is 0.252. As $c = L/(L + L') = (L/L')/\{1 + (L/L')\}$, it is possible to calculate which loss (L or L') is greater. Loss associated with the false positives (L') is 2.97 times greater than the loss associated with the false negatives (L). Therefore, if the clinician considers that $L' > L$, then the beyond-chance average agreement between the hepatic scintigraphy and the biopsy is moderate $(\hat{\kappa}_1 = 0.572)$, and the loss that occurs when erroneously classifying a non-diseased patient with the hepatic scintigraphy is 2.97 times greater than the loss that occurs when erroneously classifying a diseased patient with the hepatic scintigraphy.

When hepatic scintigraphy is to be used as a screening test ($L > L'$ and $0.5 < c \leq 1$), the beyond-chance average agreement between the hepatic scintigraphy and the biopsy is moderate $(\hat{\kappa}_{2mice} = 0.526)$. In terms of the Wald CI, the beyond-chance average agreement between the hepatic scintigraphy and the biopsy is a value between fair and substantial (95% confidence). Estimated relative loss between the false positives and the false negatives is 0.752, so that the loss associated with the false negatives (L) is 3.03 times greater than the loss associated with the false positives (L'). Therefore, if the clinician considers that $L > L'$, then the loss that occurs when erroneously classifying a diseased patient with the hepatic scintigraphy is 3.03 times greater than the loss committed when erroneously classifying a non-diseased patient with the hepatic scintigraphy.

8. Discussion

The average kappa coefficient is a measure of the beyond-chance average agreement between the BDT and the GS, and depends only on the Se and Sp of the BDT and on disease prevalence. The average kappa coefficient solves the problem of assigning values to the weighting index of the weighted kappa coefficient. In this manuscript we study the estimation of the average kappa coefficient when the gold standard is not applied to all patients in a sample. We study the estimation of the average kappa coefficient when the gold standard is not applied to all patients in a sample, a situation known as partial verification of the disease. The estimation of the average kappa coefficient has been carried out by applying two methods: the maximum likelihood method and the MICE method. As both methods require that the verification process be MAR, it therefore follows the verification process does not depend on disease status.

We have carried out simulation experiments to study the asymptotic behavior of the proposed ICs, both using the maximum likelihood approach and MICE. The relative biases of the two estimators (maximum likelihood and MICE) of the average kappa coefficient have also been calculated. MICE method along with the arcsine CI is the interval that has been shown to have a better coverage probability when the sample size is small, while MICE method along the Wald CI has shown to have a better coverage probability when the sample size is moderate or large. Regarding the relative biases, the difference between the relative biases of both types of estimators is small, such that both methods give rise to estimators that on average are very similar. Therefore, we recommend using MICE instead of the maximum likelihood method.

As in other studies [6,17], multiple imputation has proven to be a good method (and better than the maximum likelihood method) to estimate parameters of a binary diagnostic test in the presence of partial verification of the disease. In the situation studied here, the application of MICE has been carried out by generating 20 data sets. Rubin [14] recommended imputing five complete data sets in order to be able to apply multiple imputation. As our simulations have given stable values with 20 and 50 data sets, we decided, finally, to use 20.

The MICE method is requires the missing data to be MAR, so if the verification process depends on disease status then the MAR assumption is not verified and MICE cannot be applied. Therefore, it is necessary to study other methods of estimating the average kappa coefficient when the MAR assumption is not verified. The application of the method used by Kosinski and Barnhart [23] may be a solution to this problem. Future research should also focus on estimating the average kappa coefficient when covariates are observed in all patients in the sample.

Finally, we have written a function in R to estimate the average kappa coefficient in the situation studied in this manuscript, applying MICE. The function is available as Supplementary Materials.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/math9141694/s1. The function "eakcpv" is a function written in R that allows estimating the average kappa coefficient by applying the MICE method.

Author Contributions: J.A.R.-N. and S.B.R. have collaborated equally in the realization of this work. Both authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank the anonymous referees for their helpful comments that improved the quality of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Roldán-Nofuentes and Olvera-Porcel [3] have deduced (applying the delta method) the expressions of the estimated variances of the estimators of the average kappa coefficients when the BDT and GS are applied to all patients in a sample.

When $\hat{p} \neq \hat{Q}$ (which is equivalent to $x_0 \neq y_1$), the estimated asymptotic variances of $\hat{\kappa}_1$ and $\hat{\kappa}_2$ are [3]

$$\hat{V}ar(\hat{\kappa}_1) = \frac{1}{[\hat{\kappa}(0)+\hat{\kappa}(1)]^2[\hat{\kappa}(0)-\hat{\kappa}(1)]^2} \times$$
$$\left\{ \left\{ \frac{2\hat{\kappa}(0)^2\hat{\kappa}(1)-\hat{\kappa}(1)[\hat{\kappa}(0)+\hat{\kappa}(1)]\hat{\kappa}_1}{\hat{\kappa}(0)} \right\}^2 \frac{\left(1-\hat{S}p\right)^2\hat{Y}^2\hat{V}ar(\hat{p})+\hat{p}^2\left[\left(1-\hat{S}p\right)^2\hat{V}ar\left(\hat{S}e\right)+\hat{S}e^2\hat{V}ar\left(\hat{S}p\right)\right]}{\hat{Q}^4} + \right.$$
$$\left\{ \frac{\hat{\kappa}(0)[(\hat{\kappa}(0)+\hat{\kappa}(1))\hat{\kappa}_1-2\hat{\kappa}(0)\hat{\kappa}(1)]}{\hat{\kappa}(1)} \right\}^2 \frac{\left(1-\hat{S}e\right)^2\hat{Y}^2\hat{V}ar(\hat{p})+\hat{q}^2\left[\hat{S}p^2\hat{V}ar\left(\hat{S}e\right)+\left(1-\hat{S}e\right)^2\hat{V}ar\left(\hat{S}p\right)\right]}{\left(1-\hat{Q}\right)^4} +$$
$$2\left\{ \frac{2\hat{\kappa}(0)^2\hat{\kappa}(1)+\hat{\kappa}(1)[\hat{\kappa}(0)+\hat{\kappa}(1)]\hat{\kappa}_1}{\hat{\kappa}(0)} \right\} \left\{ \frac{\hat{\kappa}(0)[(\hat{\kappa}(0)-\hat{\kappa}(1))\hat{\kappa}_1-2\hat{\kappa}(0)\hat{\kappa}(1)]}{\hat{\kappa}(1)} \right\} \times$$
$$\left. \frac{\hat{p}\hat{q}\left[\left(1-\hat{S}e\right)\hat{S}e\hat{V}ar\left(\hat{S}p\right)+\left(1-\hat{S}p\right)\hat{S}p\hat{V}ar\left(\hat{S}e\right)\right]-\left(1-\hat{S}e\right)\left(1-\hat{S}p\right)\hat{Y}^2\hat{V}ar(\hat{p})}{\hat{Q}^2\left(1-\hat{Q}\right)^2} \right\}$$

and

$$\hat{V}ar(\hat{\kappa}_2) = \frac{1}{[\hat{\kappa}(0)+\hat{\kappa}(1)]^2[\hat{\kappa}(0)-\hat{\kappa}(1)]^2} \times$$
$$\left\{ \left\{ \frac{\hat{\kappa}(1)[2\hat{\kappa}(0)\hat{\kappa}(1)-(\hat{\kappa}(0)+\hat{\kappa}(1))\hat{\kappa}_2]}{\hat{\kappa}(0)} \right\}^2 \frac{\left(1-\hat{S}p\right)^2\hat{Y}^2\hat{V}ar(\hat{p})+\hat{p}^2\left[\left(1-\hat{S}p\right)^2\hat{V}ar\left(\hat{S}e\right)+\hat{S}e^2\hat{V}ar\left(\hat{S}p\right)\right]}{\hat{Q}^4} + \right.$$
$$\left\{ \frac{\hat{\kappa}(0)[\hat{\kappa}(0)+\hat{\kappa}(1)]\hat{\kappa}_2-2\hat{\kappa}(0)\hat{\kappa}(1)^2}{\hat{\kappa}(1)} \right\}^2 \frac{\left(1-\hat{S}e\right)^2\hat{Y}^2\hat{V}ar(\hat{p})+\hat{q}^2\left[\hat{S}p^2\hat{V}ar\left(\hat{S}e\right)+\left(1-\hat{S}e\right)^2\hat{V}ar\left(\hat{S}p\right)\right]}{\left(1-\hat{Q}\right)^4} +$$
$$2\left\{ \frac{\hat{\kappa}(1)[2\hat{\kappa}(0)\hat{\kappa}(1)-(\hat{\kappa}(0)+\hat{\kappa}(1))\hat{\kappa}_2]}{\hat{\kappa}(0)} \right\} \left\{ \frac{\hat{\kappa}(0)[\hat{\kappa}(0)+\hat{\kappa}(1)]\hat{\kappa}_2-2\hat{\kappa}(0)\hat{\kappa}(1)^2}{\hat{\kappa}(1)} \right\} \times$$
$$\left. \frac{\hat{p}\hat{q}\left[\left(1-\hat{S}e\right)\hat{S}e\hat{V}ar\left(\hat{S}p\right)+\left(1-\hat{S}p\right)\hat{S}p\hat{V}ar\left(\hat{S}e\right)\right]-\left(1-\hat{S}e\right)\left(1-\hat{S}p\right)\hat{Y}^2\hat{V}ar(\hat{p})}{\hat{Q}^2\left(1-\hat{Q}\right)^2} \right\}$$

respectively, where $\hat{S}e = x_1/x$, $\hat{S}p = y_0/y$, $\hat{p} = x/m$, $\hat{q} = y/m$, $\hat{Y} = \frac{x_1y_0-x_0y_1}{xy}$ and $\hat{Q} = m_1/m$.

When $\hat{p} = \hat{Q}$ (which is equivalent to $x_0 = y_1$) the estimated asymptotic variances are

$$\hat{V}ar(\hat{\kappa}_1) = \hat{V}ar(\hat{\kappa}_2) = \hat{V}ar(\hat{Y}) = \frac{\hat{S}e(1-\hat{S}e)}{x} + \frac{\hat{S}p(1-\hat{S}p)}{y}.$$

Appendix B

In this appendix, the expressions of the estimated variances of the estimators of the average kappa coefficients in the presence of partial verification are obtained. Applying the delta method, the asymptotic estimated variance of $\hat{\kappa}_{ipv}$, with $i = 1, 2$, is

$$\hat{V}ar(\hat{\kappa}_{ipv}) = \left(\frac{\partial\kappa_{ipv}}{\partial Se}\right)^2_{Se=\hat{S}e_{pv}} \hat{V}ar(\hat{S}e_{pv}) + \left(\frac{\partial\kappa_{ipv}}{\partial Sp}\right)^2_{Sp=\hat{S}p_{pv}} \hat{V}ar(\hat{S}p_{pv}) + \left(\frac{\partial\kappa_{ipv}}{\partial p}\right)^2_{p=\hat{p}_{pv}} \hat{V}ar(\hat{p}_{pv}) +$$
$$2\left(\frac{\partial\kappa_{ipv}}{\partial Se}\right)_{Se=\hat{S}e_{pv}} \left(\frac{\partial\kappa_{ipv}}{\partial Sp}\right)_{Sp=\hat{S}p_{pv}} \hat{C}ov(\hat{S}e_{pv}, \hat{S}p_{pv}) +$$
$$2\left(\frac{\partial\kappa_{ipv}}{\partial Se}\right)_{Se=\hat{S}e_{pv}} \left(\frac{\partial\kappa_{ipv}}{\partial p}\right)_{p=\hat{p}_{pv}} \hat{C}ov(\hat{S}e_{pv}, \hat{p}_{pv}) +$$
$$2\left(\frac{\partial\kappa_{ipv}}{\partial Sp}\right)_{Sp=\hat{S}p_{pv}} \left(\frac{\partial\kappa_{ipv}}{\partial p}\right)_{p=\hat{p}_{pv}} \hat{C}ov(\hat{S}p_{pv}, \hat{p}_{pv}),$$

when $\hat{p}_{pv} \neq \hat{Q}_{pv}$, and

$$\hat{V}ar(\hat{\kappa}_{ipv}) = \hat{V}ar(\hat{Y}_{pv}) = \hat{S}p_{pv}^2\hat{V}ar(\hat{S}e_{pv}) + \hat{S}e_{pv}^2\hat{V}ar(\hat{S}p_{pv}) + 2\hat{S}e_{pv}\hat{S}p_{pv}\hat{C}ov(\hat{S}e_{pv}, \hat{S}p_{pv})$$

when $\hat{p}_{pv} = \hat{Q}_{pv}$, and where [4]

$$\begin{aligned}\hat{V}ar(\hat{S}e_{pv}) &= \{\hat{S}e_{pv}(1-\hat{S}e_{pv})\}^2\left\{\tfrac{n}{n_1n_0} + \tfrac{r_1}{s_1(s_1+r_1)} + \tfrac{r_0}{s_0(s_0+r_0)}\right\},\\ \hat{V}ar(\hat{S}p_{pv}) &= \{\hat{S}p_{pv}(1-\hat{S}p_{pv})\}^2\left\{\tfrac{n}{n_1n_0} + \tfrac{s_1}{r_1(s_1+r_1)} + \tfrac{s_0}{r_0(s_0+r_0)}\right\}\end{aligned}$$

and [24]

$$\hat{C}ov(\hat{S}e_{pv}, \hat{S}p_{pv}) = \hat{S}e_{pv}\hat{S}p_{pv}(1-\hat{S}e_{pv})(1-\hat{S}p_{pv})\left\{\frac{u_1}{n_1(s_1+r_1)} + \frac{u_0}{n_0(s_0+r_0)}\right\}.$$

The variance $\hat{V}ar(\hat{p})$ and covariances $\hat{C}ov(\hat{p}, \hat{S}e)$ and $\hat{C}ov(\hat{p}, \hat{S}p)$ are obtained by applying the delta method. Let τ be the positive predictive value of the BDT, let v be the negative predictive value of the BDT, let Q be the probability of a positive result of the BDT, and let $\boldsymbol{\psi} = (\tau, v, Q)^T$. Applying the delta method, the variance-covariance matrix of $\boldsymbol{\psi}$ is [25]

$$\sum_{\boldsymbol{\psi}} = Diag\left\{\frac{\tau^2(1-\tau)^2}{s_1(1-\tau)^2 + r_1\tau^2}, \frac{v^2(1-v)^2}{s_0v^2 + r_0(1-\tau)^2}, \frac{Q^2(1-Q)^2}{n_1(1-Q)^2 + n_0Q^2}\right\}$$

MLEs of predictive values in the presence of partial verification are [26] $\hat{\tau}_{pv} = s_1/(s_1+r_1)$ and $\hat{v}_{pv} = r_0/(s_0+r_0)$, and the MLE of Q is $\hat{Q}_{pv} = n_1/n$. Therefore, in the presence of partial verification, the estimators of the predictive values coincide with the naïve estimators (those obtained regardless of the unverified patients) when the MAR hypothesis is assumed [26]. Let $\boldsymbol{\theta} = (Se, Sp, p)^T$ be the vector whose components are the sensitivity, the specificity and the prevalence. As the sensitivity, specificity and prevalence can be written in terms of the predictive values and of Q as $Se = \frac{\tau\{v-(1-p)\}}{p(\tau+v-1)}$, $Sp = \frac{v(\tau-p)}{(1-p)(\tau+v-1)}$ and $p = 1-(1-\tau)Q-(1-Q)v$, then the estimated variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is obtained by applying the delta method, i.e.,

$$\hat{\sum}_{\hat{\boldsymbol{\theta}}} = \left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\psi}}\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{pv}} \hat{\sum}_{\hat{\boldsymbol{\psi}}} \left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\psi}}\right)^T_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{pv}}$$

Carrying out the algebraic operations it is obtained:

$$\begin{aligned}\hat{V}ar(\hat{p}_{pv}) &= \tfrac{\hat{\tau}_{pv}(1-\hat{\tau}_{pv})\hat{Q}_{pv}^2}{s_1+r_1} + \tfrac{\hat{v}_{pv}(1-\hat{v}_{pv})(1-\hat{Q}_{pv})^2}{s_0+r_0} + \tfrac{(s_1r_0-s_0r_1)^2\hat{Q}_{pv}(1-\hat{Q}_{pv})}{n(s_1+r_1)^2(s_0+r_0)^2},\\ \hat{C}ov(\hat{S}e_{pv}, \hat{p}_{pv}) &= \tfrac{n_1n_0s_1s_0(s_1+r_1)(s_0+r_0)}{\{n_1s_1(s_0+r_0)+n_0s_0(s_1+r_1)\}^2}\left\{\tfrac{(1-\hat{\tau}_{pv})\hat{Q}_{pv}}{s_1+r_1} - \tfrac{\hat{v}_{pv}(1-\hat{Q}_{pv})}{s_0+r_0} + \tfrac{s_1r_0-s_0r_1}{n(s_1+r_1)(s_0+r_0)}\right\}\end{aligned}$$

and

$$\hat{C}ov(\hat{S}p_{pv}, \hat{p}_{pv}) = \tfrac{n_1n_0r_1r_0(s_1+r_1)(s_0+r_0)}{\{n_1r_1(s_0+r_0)+n_0r_0(s_1+r_1)\}^2}\left\{\tfrac{\hat{\tau}_{pv}\hat{Q}_{pv}}{s_1+r_1} - \tfrac{(1-\hat{v}_{pv})(1-\hat{Q}_{pv})}{s_0+r_0} - \tfrac{s_1r_0-s_0r_1}{n(s_1+r_1)(s_0+r_0)}\right\}.$$

References

1. Kraemer, H.C. *Evaluating Medical Tests. Objective and Quantitative Guidelines*; Sage Publications: Newbury Park, CA, USA, 1992.
2. Kraemer, H.C.; Periyakoil, V.S.; Noda, A. Kappa coefficients in medical research. *Stat. Med.* **2002**, *21*, 2109–2129. [CrossRef]
3. Roldán-Nofuentes, J.A.; Olvera-Porcel, C. Average kappa coefficient: A new measure to assess a binary test considering the losses associated with an erroneous classification. *J. Stat. Comput. Simul.* **2015**, *85*, 1601–1620. [CrossRef]

4. Begg, C.B.; Greenes, R.A. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* **1983**, *39*, 207–215. [CrossRef]
5. Zhou, X.H. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Comm. Statist. Theory Methods* **1993**, *22*, 3177–3198. [CrossRef]
6. Harel, O.; Zhou, X.H. Multiple imputation for correcting verification bias. *Stat. Med.* **2006**, *25*, 3769–3786. [CrossRef] [PubMed]
7. Alonzo, T.A. Verification bias-impact and methods for correction when assessing accuracy of diagnostic tests. *REVSTAT* **2014**, *12*, 67–83.
8. Roldán-Nofuentes, J.A.; Luna del Castillo, J.D. Risk of error and the kappa coefficient of a binary diagnostic test in the presence of partial verification. *J. Appl. Stat.* **2007**, *34*, 887–898. [CrossRef]
9. Youden, W.J. Index for rating diagnostic tests. *Cancer* **1950**, *3*, 32–35. [CrossRef]
10. Landis, R.; Koch, G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
11. Cicchetti, D.V. The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *J. Clin. Exp. Neuropsychol.* **2001**, *23*, 695–700. [CrossRef]
12. Martín-Andrés, A.; Álvarez-Hernández, M. Two-tailed asymptotic inferences for a proportion. *J. Appl. Stat.* **2014**, *41*, 1516–1529. [CrossRef]
13. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *4*, 73–89. [CrossRef]
14. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; Wiley: New York, NY, USA, 1987.
15. Rubin, D.B. Multiple Imputation after 18+ years. *J. Am. Stat. Assoc.* **1996**, *91*, 473–489. [CrossRef]
16. Schafer, J.L. *Analysis of Incomplete Multivariate Data*; Chapman and Hall: New York, NY, USA, 1997.
17. Harel, O.; Zhou, X.H. Multiple imputation: Review of theory, implementation and software. *Stat. Med.* **2007**, *26*, 3057–3077. [CrossRef]
18. Montero-Alonso, M.A.; Roldán-Nofuentes, J.A. Approximate confidence intervals for the likelihood ratios of a binary diagnostic test in the presence of partial disease verification. *J. Biopharm. Stat.* **2019**, *29*, 56–81. [CrossRef] [PubMed]
19. White, I.R.; Royston, P.; Wood, A.M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **2011**, *30*, 377–399. [CrossRef] [PubMed]
20. R Core Team. *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2016; Available online: https://www.R-project.org/ (accessed on 1 June 2021).
21. van Buuren, S.; Groothuis-Oudshoorn, K. Mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **2011**, *45*, 3. [CrossRef]
22. Drum, D.E.; Christacopoulos, J.S. Hepatic scintigraphy in clinical decision making. *J. Nucl. Med.* **1972**, *13*, 908–915.
23. Kosinski, A.S.; Barnhart, H.X. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics* **2003**, *59*, 163–171. [CrossRef]
24. Kosinski, A.S.; Chenb, Y.; Lylesc, R.H. Sample size calculations for evaluating a diagnostic test when the gold standard is missing at random. *Stat. Med.* **2010**, *29*, 1572–1579. [CrossRef]
25. Zhou, X.H.; Obuchowski, N.; McClish, D. *Statistical Methods in Diagnostic Medicine*, 2nd ed.; Wiley: New York, NY, USA, 2011.
26. Zhou, X.H. Effect of verification bias on positive and negative predictive values. *Stat. Med.* **1994**, *13*, 1737–1745. [CrossRef] [PubMed]