mathematics

Article

Comparison of the Average Kappa Coefficients of Two Binary Diagnostic Tests with Missing Data

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 parameter that measures the average beyond-chance agreement between the diagnostic test and the gold standard. This parameter depends on the accuracy of the diagnostic test and also on the disease prevalence. This article studies the comparison of the average kappa coefficients of two binary diagnostic tests when the gold standard is not applied to all individuals in a random sample. In this situation, known as partial disease verification, the disease status of some individuals is a missing piece of data. Assuming that the missing data mechanism is missing at random, the comparison of the average kappa coefficients is solved by applying two computational methods: the EM algorithm and the SEM algorithm. With the EM algorithm the parameters are estimated and with the SEM algorithm their variances-covariances are estimated. Simulation experiments have been carried out to study the sizes and powers of the hypothesis tests studied, obtaining that the proposed method has good asymptotic behavior. A function has been written in R to solve the proposed problem, and the results obtained have been applied to the diagnosis of Alzheimer's disease.

Keywords: EM algorithm; partial verification; SEM algorithm

Citation: Roldán-Nofuentes, J.A.; Regad, S.B. Comparison of the Average Kappa Coefficients of Two Binary Diagnostic Tests with Missing Data. *Mathematics* **2021**, *9*, 2834. https://doi.org/10.3390/math9212834

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

Received: 8 October 2021
Accepted: 5 November 2021
Published: 8 November 2021

1. Introduction

Diagnostic tests are fundamental in the current practice of medicine. A diagnostic test is a medical test that is applied to an individual to determine the presence or absence of a disease [1]. Diagnostic tests can be binary, ordinal or continuous. Binary tests give two possible results: positive or negative. An antigen test for the diagnosis of COVID-19 is an example of a binary diagnostic test. Ordinal tests classify the presence of the disease in different ordinal categories. For example, in the diagnosis of breast cancer, malignant lesions can be classified as "malignant, suspicious, probably benign, benign or normal". With respect to continuous tests, these give rise to continuous values, for example procalcitonin for the diagnosis of infective endocarditis. The efficacy of a diagnostic test is evaluated against a gold standard. A gold standard (GS) is a medical test that objectively determines whether or not an individual has the disease. For example, a biopsy for the diagnosis of cancer. This article focuses on binary diagnostic tests.

The fundamental measures to evaluate the effectiveness of a binary diagnostic test (BDT) are sensitivity and specificity. Sensitivity is the probability that the test result is positive when the individual has the disease, and specificity is the probability that the test result is negative when the individual does not have the disease. The sensitivity and specificity of a BDT depend on the physical, chemical or biological bases with which the test has been developed. When evaluating the effectiveness of a BDT considering the losses associated with misclassification with the BDT, the parameter used is the weighted kappa coefficient [1,2]. The weighted kappa coefficient is a parameter that measures the beyond chance agreement between BDT and GS [1,2], and depends on the sensitivity and specificity of BDT, on the disease prevalence and on the weighting index. The weighting index is a measure of the relative importance between false positives and false negatives. In practice, the weighting index c is set by the clinician depending on the clinical use of the BDT (for example, confirmatory test or screening test) and the clinician's knowledge of the importance of a false positive and a false negative. If the BDT is to be used as a confirmatory test, then the weighting index takes a value between 0 and 0.5. If the BDT is to be used as a screening test, then the weighting index takes a value between 0.5 and 1. The problem with the weighted kappa coefficient is the assignment of values to the weighting index c, since the clinician does not always have a knowledge that allows him to decide how important a false positive is compared to a false negative. Even in the same problem, two clinicians can assign different values to the weighting index. Roldán-Nofuentes and Olvera-Porcel [3] have defined and studied a new measure to evaluate the effectiveness of a BDT: the average kappa coefficient. The average kappa coefficient depends only on the intrinsic accuracy (sensitivity and specificity) of the BDT and on the disease prevalence, and is a parameter that does not depend on the weighting index. Therefore, the average kappa coefficient is a parameter that solves the problem of assigning values to the weighting index. Average kappa coefficient is a measure of the average beyond-chance agreement between the BDT and the GS [3].

Comparison of the effectiveness of two BDTs is a topic of special interest in the study of statistical methods for the diagnosis of diseases. The most frequent type of sampling to compare two BDTs is the paired design, which consists of applying the two BDTs to all individuals in a random sample whose disease status is known by applying a GS. Bloch [4] has studied the comparison of the weighted kappa coefficients of two BDTs under a paired design, and Roldán-Nofuentes and Luna [5] have extended the study of Bloch to the situation in which the weighted kappa coefficients of more than two BDTs are compared. Roldán-Nofuentes and Olvera-Porcel [6] has studied the comparison of the average kappa coefficients of two BDTs under a paired design. However, in clinical practice the GS is not always applied to all individuals in the sample. Consequently, the disease state is unknown for a subset of individuals in the sample. This problem is known as partial verification of disease [7,8]. Zhou [9] has studied a hypothesis test to compare the sensitivities (specificities) of two BDTs in the presence of partial verification, applying the maximum likelihood method. If in this situation the two sensitivities (specificities) are compared, eliminating the individuals whose disease status is unknown, the estimates obtained are biased (the estimators are affected by the so-called verification bias [7]) and the results may be incorrect [9]. Harel and Zhou [10] have compared the sensitivities (specificities) of two BDTs using confidence intervals applying multiple imputation, and Roldán-Nofuentes and Luna [11] have compared the sensitivities (specificities) by applying the EM and the SEM algorithms. Roldán-Nofuentes and Luna [12] have studied a hypothesis test to compare the weighted kappa coefficients of two BDTs in the presence of partial verification of the disease, applying the maximum likelihood method. Regarding the average kappa coefficient, Roldán-Nofuentes and Regad [13] have studied the estimation of this parameter when only a single BDT is evaluated in the presence of partial verification, applying the maximum likelihood method and multiple imputation. The comparison of the average kappa coefficients of two BDTs has never been studied in the presence of partial verification. In this situation, if the weighted kappa coefficients are compared, eliminating the unverified individuals with the GS, then the estimators of the weighted kappa coefficients are biased [12], and therefore the estimators of the average kappa coefficients, and the conclusions can also be incorrect. Consequently, the method of Roldán-Nofuentes and Olvera-Porcel [6] cannot be applied in the presence of partial verification.

In this article, the comparison of the average kappa coefficients of two BDTs in the presence of partial verification of the disease is studied. Therefore, the objective of our manuscript is to study a hypothesis test to compare the average kappa coefficients of two BDTs in the presence of partial verification, a topic that has never been studied. This article is an extension of the article by Roldán-Nofuentes and Olvera-Porcel [6] to the situation in which the GS does not apply to all the individuals in the sample, and is also an extension of the article by Roldán-Nofuentes and Regad [13] to the situation where two BDTs are compared in the presence of partial verification. The article is structured as follows. In Section 2 the average kappa coefficient and its properties are presented. In Section 3 we study the comparison of the weighted kappa coefficients of two BDTs in the presence of partial verification of the disease, applying two computational methods: the EM algorithm and the SEM algorithm. In Section 4, a function written in R is presented to solve the problem and simulation experiments are carried out to study the size and power of the method to solve the hypothesis test for the comparison of the two average kappa coefficients. In Section 5 the results are applied to the diagnosis of Alzheimer disease, and in Section 6 the results obtained are discussed.

2. Average Kappa Coefficient

Let us consider two BDTs, Test 1 and Test 2, whose performances are compared with respect to the same GS. Let L (L') be the loss that occurs when a BDT gives a negative (positive) result for a diseased (non-diseased) patient. Loss L is associated with a false negative and loss L' is associated with a false positive [1,2]. Losses are assumed to be zero when a BDT correctly classifies a diseased patient or a non-diseased patient [1,2]. For example, let us consider the diagnosis of renal cell carcinoma using the MOC 31. If the MOC 31 is positive for an individual without the renal carcinoma (false positive), the individual will undergo a renal biopsy which will be negative. Loss L' is determined by the economic costs of the diagnosis and also by the risk, stress, etc, caused to the individual. If the MOC 31 is negative for an individual with renal carcinoma (false negative), the individual will be diagnosed later, but the cancer will progress and get worse, decreasing the chance that treatment will be successful. Loss L is determined from this situation. Therefore, losses L and L' are measured in terms of economic costs and in terms of risks, stress, etc [1,2], so in clinical practice it is not possible to know L and L'. Let *T* be the binary random variable that models the result of the BDT, in such a way that $T=1$ when the result is positive and $T=0$ when the result is negative. Let *D* be the binary random variable that models the result of the GS, in such a way that $D=1$ when the individual has the disease and $D=0$ when the individual does not have the disease. In Table 1, we show the losses and probabilities associated with the assessment of a BDT in relation to a GS, where *Se* is the sensitivity, *Sp* the specificity and *p* the disease prevalence.

Table 1. Losses and observed frequencies associated with the assessment of a BDT in relation to a GS.

	Losses		
	$T=1$	$T=0$	Total
$D=1$	0	L	L
$D=0$	L'	0	L'
Total	L'	L	$L+L'$
	Probabilities		
	$T=1$	$T=0$	Total
$D=1$	pSe	$p(1-Se)$	p
$D=0$	$(1-p)(1-Sp)$	$(1-p)Sp$	$1-p$
Total	$Q=pSe+(1-p)(1-Sp)$	$1-Q=p(1-Se)+(1-p)Sp$	1

In terms of the losses and probabilities in Table 1, the expected loss [4] is $p(1-Se)\mathrm{L}+q(1-Sp)\mathrm{L}'$ and the random loss [4] is

$p\left\{p\left(1-Se\right)+qSp\right\}L+q\left\{pSe+q\left(1-Sp\right)\right\}L'$, with $q=1-p$. The expected loss is the loss that occurs when erroneously classifying a diseased or non-diseased individual with the BDT. The expected loss varies between zero and infinity. The random loss is the loss that occurs when the BDT and the GS are independent, i.e., when $P\left(T=i\middle|D=j\right)=P\left(T=i\right)$. In terms of these losses, the weighted kappa coefficient is defined as [1,2,4]

$$\kappa=\frac{\text{Random loss}-\text{Expected loss}}{\text{Random loss}-\min\left(\text{Expected loss}\right)}=\frac{\text{Random loss}-\text{Expected loss}}{\text{Random loss}},$$

since $\min\left(\text{Expected loss}\right)=0$. Performing algebraic operations, the weighted kappa coefficient is written as [1,2,4]

$$\kappa_h\left(c\right)=\frac{pqY_h}{pc\left(1-Q_h\right)+q\left(1-c\right)Q_h},\ \ 0\leq c\leq 1,\ h=1,2,$$

where $Y_h=Se_h+Sp_h-1$ is the Youden index [14] of the *h*th Test, $Q_h=pSe_h+q\left(1-Sp_h\right)$ is the probability that the *h*th Test is positive and $c=L/\left(L'+L\right)$ is the weighting index. The weighted kappa coefficient of the *h*th Test can also be written as

$$\kappa_h\left(c\right)=\frac{\kappa_h\left(0\right)\kappa_h\left(1\right)}{c\kappa_h\left(0\right)+\left(1-c\right)\kappa_h\left(1\right)},\ \ 0\leq c\leq 1, \tag{1}$$

where

$$\kappa_h\left(0\right)=\frac{Sp_h-\left(1-Q_h\right)}{Q_h}\ \text{ and }\ \kappa_h\left(1\right)=\frac{Se_h-Q_h}{1-Q_h}.$$

As L and L' are unknown, the clinician sets the value of the weighting index based on the relative importance between false positives and false negatives [1,2]. If the clinician considers that false positives are more important than false negatives, as is the situation 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 $0\leq c<0.5$. For example, if a false positive is four times more important than a false negative then $L'=4L$ and $c=1/\left(1+4\right)=1/5$. If the clinician considers that false negatives are more important than false positives, as is the situation in which the BDT is used as a screening test, then $L>L'$ and $0.5<c\leq 1$. For example, if a false negative is three times more important than a false positive then $L=3L'$ and $c=3/\left(3+1\right)=3/4$. Value $c=0.5$ is used when false positives and false negatives have the same importance, being $\kappa\left(0.5\right)$ the Cohen kappa coefficient. The weighted kappa coefficient has the following properties [1,2,4]:

1. If $Se_h=Sp_h=1$ then $\kappa\left(c\right)=1$, and the agreement between Test and GS is perfect.
2. If $Se_h=1-Sp_h$ then $\kappa_h\left(c\right)=0$, and the Test and the GS are independent.
3. 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 $Q=p$.

The weighted kappa coefficient can be classified in the following scale of values [15]: $0-0.20$, slight; $0.21-0.40$, fair; $0.41-0.60$, moderate; $0.61-0.80$, substantial; and $0.81-1$, almost perfect. Another scale based on levels of clinical significance is [16]: <0.40, poor; $0.40-0.59$, fair; $0.60-0.74$, good; and $0.75-1$, excellent.

Roldán-Nofuentes and Olvera-Porcel [3] have proposed a new measure to evaluate and to compare BDTs: the average kappa coefficient. If $L'>L$, and therefore $0\leq c<0.5$, the average kappa coefficient of the *h*th Test is [3]

$$\kappa_{h1} = \frac{1}{0.5}\int_0^{0.5} \kappa_h(c)\mathrm{d}c = \begin{cases} \dfrac{2\kappa_h(0)\kappa_h(1)}{\kappa_h(0)-\kappa_h(1)}\ln\left\{\dfrac{\kappa_h(0)+\kappa_h(1)}{2\kappa_h(1)}\right\}, & p \neq Q_h \\ Y_h, & p = Q_h, \end{cases} \tag{2}$$

i.e., the average kappa coefficient is the average value of $\kappa_h(c)$ when $0 \leq c < 0.5$. If $L > L'$ and therefore $0.5 < c \leq 1$, the average kappa coefficient of the *h*th Test is [3]

$$\kappa_{h2} = \frac{1}{0.5}\int_{0.5}^{1} \kappa_h(c)\mathrm{d}c = \begin{cases} \dfrac{2\kappa_h(0)\kappa_h(1)}{\kappa_h(0)-\kappa_h(1)}\ln\left\{\dfrac{2\kappa_h(0)}{\kappa_h(0)+\kappa_h(1)}\right\}, & p \neq Q_h \\ Y_h, & p = Q_h, \end{cases} \tag{3}$$

i.e., the average kappa coefficient is the average value of $\kappa_h(c)$ when $0.5 < c \leq 1$. As the weighted kappa coefficient is a measure of the beyond-chance agreement between a BDT and the GS, the average kappa coefficient is a measure of the average beyond-chance agreement between a BDT and a GS [3], and does not depend on the weighting index c. As $\kappa_h(0)$ and $\kappa_h(1)$ depend on Se_h, Sp_h and p, then κ_{h1} and κ_{h2} also depend on these same parameters. The values of the average kappa coefficient can be classified on the same scales [15,16] as the values of the weighted kappa coefficient [3]. The average kappa coefficients κ_{h1} and κ_{h2} have the following properties [3]:

1. If $Se_h = Sp_h = 1$ then $\kappa_{h1} = \kappa_{h2} = 1$, and if $Se_h = 1 - Sp_h$ then $\kappa_{h1} = \kappa_{h2} = 0$. Therefore $0 \leq \kappa_{hi} \leq 1$, $i = 1,2$.
2. $\kappa_{h1} > \kappa_{h2}$ if $p > Q_h$ and $\kappa_{h1} < \kappa_{h2}$ if $Q_h > p$.
3. κ_{h1} minimizes $2\int_0^{0.5}\left\{\kappa_h(c) - x\right\}^2 dc$ and κ_{h2} minimizes $2\int_{0.5}^{1}\left\{\kappa_h(c) - x\right\}^2 dc$. Therefore, when $x = \kappa_{h1}$ ($x = \kappa_{h2}$) the first (second) expression is the variance of $\kappa_h(c)$ around κ_{h1} (κ_{h2}).
4. For fixed values of $\kappa_h(0)$ and $\kappa_h(1)$, the weighted kappa coefficient $\kappa_h(c)$ is a function of c which is continuous in the interval $[0,1]$. Therefore, the average kappa coefficient κ_{hi} is equal to a value of $\kappa_h(c)$ in the interval $[0,1]$. This value of $\kappa_h(c)$ has a value of weighting index c. So, as $\kappa_{hi} = \kappa_h(c)$ for some value of c, from Equation (1) and for a specific sample it is possible to calculate the value of c associated to the estimated of κ_{hi} . Therefore, the estimation of κ_{hi} allows estimating how much greater (or less) the loss L is than the loss L'.

Next, the comparison of the average kappa coefficients of two BDTs in the presence of partial verification of the disease is studied.

3. Comparison of Average Kappa Coefficients

The objective of this manuscript is to study the hypothesis tests

$$\mathrm{H}_0: \kappa_{11} = \kappa_{21} \text{ vs } \mathrm{H}_1: \kappa_{11} \neq \kappa_{21} \tag{4}$$

and

$$\mathrm{H}_0: \kappa_{12} = \kappa_{22} \text{ vs } \mathrm{H}_1: \kappa_{12} \neq \kappa_{22} \tag{5}$$

when not all patients in a random sample are verified with the GS. The first hypothesis test is used when the clinician considers that $L' > L$ ($0 \leq c < 0.5$) and the second hypothesis test is used when the clinician considers that $L > L'$ ($0.5 < c \leq 1$). Both hypothesis tests will be solved by applying two computational methods: the EM algorithm and the SEM algorithm. The EM algorithm [17] is a classic method to estimate parameters with missing data, and the SEM (Supplemented EM) algorithm [18] is a method that allows estimating the variances-covariances of a vector of parameters from the results obtained by applying the EM algorithm.

In the problem posed here, the sample design is as follows: two BDTs are applied to all individuals of a random sample sized *n* and the GS is applied only to a subset of the *n* individuals. This situation gives rise to Table 2, where T_h is the binary random variable that models the result of the *h*th Test ($T_h = 1$ when the Test is positive and $T_h = 0$ when it is negative), *V* is the binary random variable that models the verification process ($V = 1$ when the disease status of an individual is verified with the GS and $V = 0$ when the disease status of an individual is not verified with the GS), and *D* is the binary random variable that models the GS ($D=1$ when the individual verified with the GS has the disease and $D = 0$ when the individual verified with the GS does not have the disease). In this table, each frequency s_{ij} (r_{ij}) is the number of diseased (non-diseased) individuals in which $T_1 = i$ and $T_2 = j$ ($i, j = 0,1$), each frequency u_{ij} is the number of individuals not verified with the GS in which and $T_1 = i$ and $T_2 = j$, $s = \sum_{i,j=0}^{1} s_{ij}$, $r = \sum_{i,j=0}^{1} r_{ij}$, $u = \sum_{i,j=0}^{1} u_{ij}$, $n_{ij} = s_{ij} + r_{ij} + u_{ij}$ and $n = s + r + u = \sum_{i,j=0}^{1} n_{ij}$.

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

Observed Frequencies					
	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$V = 1$					
$D = 1$	s_{11}	s_{10}	s_{01}	s_{00}	s
$D = 0$	r_{11}	r_{10}	r_{01}	r_{00}	r
$V = 0$	u_{11}	u_{10}	u_{01}	u_{00}	u
Total	n_{11}	n_{10}	n_{01}	n_{00}	n

Let $Se_h = P(T_h = 1 | D = 1)$ and $Sp_h = P(T_h = 0 | D = 0)$ be the sensitivity and the specificity of the *h*th Test, let $p = P(D = 1)$ be the disease prevalence, and let $\lambda_{ijk} = P(V = 1 | T_1 = i, T_2 = j, D = k)$ be the probability of verifying with the GS an individual with results $T_1 = i$, $T_2 = j$ and $D = k$, with $h = 1,2$ and $i, j, k = 0,1$. Assuming that the verification process is missing at random (MAR) [19], i.e., that the probability of verifying with the GS the disease status of an individual only conditionally depends on the results of both BDTs, then $\lambda_{ijk} = \lambda_{ij} = P(V = 1 | T_1 = i, T_2 = j)$. If the disease status of an individual is not verified with the GS, this individual can be considered as a missing value of the disease status, and then missing data analysis methods can be used to compare two BDTs in the presence of partial verification of the disease. The MAR assumption has been widely used in this context to compare parameters of two BDTs [9–12]. Assuming the MAR assumption, the frequencies in Table 1 are the product of a multinomial distribution sized *n*, whose probabilities are:

$$\begin{gathered}
\xi_{ij} = P\left(V=1, D=1, T_1=i, T_2=j\right) = \\
p\lambda_{ij}\left[Se_1^i\left(1-Se_1\right)^{1-i} Se_2^j\left(1-Se_2\right)^{1-j} + \delta_{ij}Se_1Se_2\left(\alpha_1-1\right)\right], \\
\psi_{ij} = P\left(V=1, D=0, T_1=i, T_2=j\right) = \\
q\lambda_{ij}\left[Sp_1^{1-i}\left(1-Sp_1\right)^i Sp_2^{1-j}\left(1-Sp_2\right)^j + \delta_{ij}\left(1-Sp_1\right)\left(1-Sp_2\right)\left(\alpha_0-1\right)\right], \\
\zeta_{ij} = P\left(V=0, T_1=i, T_2=j\right) = \frac{1-\lambda_{ij}}{\lambda_{ij}}\left(\xi_{ijm}+\psi_{ijm}\right),
\end{gathered} \tag{6}$$

where $q = 1-p$, $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = -1$ if $i \neq j$, α_1 (α_0) is the covariance [20] between the two BDTs when $D=1$ ($D=0$), verifying that

$$1 \le \alpha_1 \le \frac{1}{\max\left\{Se_1, Se_2\right\}} \text{ and } 1 \le \alpha_0 \le \frac{1}{\max\left\{1-Sp_1, 1-Sp_2\right\}}, \tag{7}$$

and $\sum_{i,j=0}^{1}\xi_{ij} + \sum_{i,j=0}^{1}\psi_{ij} + \sum_{i,j=0}^{1}\zeta_{ij} = 1$. If $\alpha_1 = \alpha_0 = 1$ then the two BDTs are conditionally independent on the disease, a situation which is not realistic in practice so that $\alpha_1 > 1$ and/or $\alpha_0 > 1$. Solving the system of equations $\kappa_h\left(0\right) = \left\{Sp_h - \left(1-Q_h\right)\right\}/Q_h$ and $\kappa_h\left(1\right) = \left(Se_h - Q_h\right)/\left(1-Q_h\right)$, with $h = 1,2$, it is obtained that

$$Se_h = \frac{p\kappa_h\left(1\right) + q\kappa_h\left(0\right)\kappa_h\left(1\right)}{q\kappa_h\left(0\right) + p\kappa_h\left(1\right)} \text{ and } Sp_h = \frac{q\kappa_h\left(0\right) + p\kappa_h\left(0\right)\kappa_h\left(1\right)}{q\kappa_h\left(0\right) + p\kappa_h\left(1\right)}, \tag{8}$$

and substituting these expressions in Equation (6), the probabilities of the multinomial distribution are obtained in terms of the weighted kappa coefficients. Next we apply the EM algorithm to obtain the estimates of the parameters.

The maximum likelihood (ML) estimates of the parameters are obtained by applying the EM algorithm [17]. The EM algorithm is a computational method that allows estimating parameters in the presence of missing data, and it is a method widely used in statistics to solve estimation problems in different areas, for example in industrial engineering [21] and in epidemiology [22]. Next, we carry out a reparametrization of the EM algorithm that allows us to estimate the weighted kappa coefficients of the two BDTs (and therefore the average kappa coefficients), the covariances and the disease prevalence. In Table 2 the missing data is the true disease status of the individuals who are not verified with the GS; this information is reconstructed in the E step of the EM algorithm. In the M step the ML estimates are imputed. Let us assume that that among the u_{ij} individuals not verified with the GS, y_{ij} have the disease and $u_{ij} - y_{ij}$ do not have the disease. Then the data can be expressed in the form of a 2×4 table with frequencies $s_{ij} + y_{ij}$ for $D=1$ and $r_{ij} + u_{ij} - y_{ij}$, with $i,j = 0,1$. Let $\boldsymbol{\theta} = \left(\kappa_1\left(0\right), \kappa_1\left(1\right), \kappa_2\left(0\right), \kappa_2\left(1\right), p, \alpha_1, \alpha_0\right)^T$ be the vector of parameters. From the complete data, the log-likelihood function based on *n* individuals is

$$l\left(\boldsymbol{\theta}\right) = \sum_{i,j=0}^{1}\left(s_{ij} + y_{ij}\right)\ln\left(\phi_{ij}\right) + \sum_{i,j=0}^{1}\left(r_{ij} + u_{ij} - y_{ij}\right)\ln\left(\varphi_{ij}\right), \tag{9}$$

where

$$\phi_{ij} = P\left(T_1 = i, T_2 = i, D = 1\right) =$$
$$p\left[Se_1^i\left(1-Se_1\right)^{1-i} Se_2^j\left(1-Se_2\right)^{1-j} + \delta_{ij}Se_1Se_2\left(\alpha_1 - 1\right)\right],$$
$$\varphi_{ij} = P\left(T_1 = i, T_2 = i, D = 0\right) =$$
$$q\left[Sp_1^{1-i}\left(1-Sp_1\right)^i Sp_2^{1-j}\left(1-Sp_2\right)^j + \delta_{ij}\left(1-Sp_1\right)\left(1-Sp_2\right)\left(\alpha_0 - 1\right)\right].$$

In these probabilities, covariances α_1 and α_0 verify Equation (7), Se_h and Sp_h are given by Equation (8), and it is verified that $\sum_{i,j=0}^{1}\phi_{ij} + \sum_{i,j=0}^{1}\varphi_{ij} = 1$. The vector $\boldsymbol{\theta}$ is estimated by applying the EM algorithm. Let $y_{ij}^{(m)}$ be the value of y_{ij} in the *m*th iteration of the EM algorithm and $y^{(m)} = \sum_{i,j=0}^{1} y_{ij}^{(m)}$. ML estimate of $\boldsymbol{\theta}$ in the *m*th iteration, $\hat{\boldsymbol{\theta}}^{(m)}$, is:

$$\hat{\kappa}_1^{(m)}\left(0\right) = \frac{\sum_{j=0}^{1}\left(s_{1j} + y_{1j}^{(m)}\right) \times \sum_{j=0}^{1}\left(r_{0j} + u_{0j} - y_{0j}^{(m)}\right) - \sum_{j=0}^{1}\left(s_{0j} + y_{0j}^{(m)}\right) \times \sum_{j=0}^{1}\left(r_{1j} + u_{1j} - y_{1j}^{(m)}\right)}{\left(r + u - y^{(m)}\right)\left(n_{10} + n_{11}\right)},$$

$$\hat{\kappa}_1^{(m)}\left(1\right) = \frac{\sum_{j=0}^{1}\left(s_{1j} + y_{1j}^{(m)}\right) \times \sum_{j=0}^{1}\left(r_{0j} + u_{0j} - y_{0j}^{(m)}\right) - \sum_{j=0}^{1}\left(s_{0j} + y_{0j}^{(m)}\right) \times \sum_{j=0}^{1}\left(r_{1j} + u_{1j} - y_{1j}^{(m)}\right)}{\left(s + y^{(m)}\right)\left(n_{00} + n_{01}\right)},$$

$$\hat{\kappa}_2^{(m)}\left(0\right) = \frac{\sum_{i=0}^{1}\left(s_{i1} + y_{i1}^{(m)}\right) \times \sum_{i=0}^{1}\left(r_{i0} + u_{i0} - y_{i0}^{(m)}\right) - \sum_{i=0}^{1}\left(s_{i0} + y_{i0}^{(m)}\right) \times \sum_{i=0}^{1}\left(r_{i1} + u_{i1} - y_{i1}^{(m)}\right)}{\left(s + r - y^{(m)}\right)\left(n_{01} + n_{11}\right)},$$

$$\hat{\kappa}_2^{(m)}\left(1\right) = \frac{\sum_{i=0}^{1}\left(s_{i1} + y_{i1}^{(m)}\right) \times \sum_{i=0}^{1}\left(r_{i0} + u_{i0} - y_{i0}^{(m)}\right) - \sum_{i=0}^{1}\left(s_{i0} + y_{i0}^{(m)}\right) \times \sum_{i=0}^{1}\left(r_{i1} + u_{i1} - y_{i1}^{(m)}\right)}{\left(s + x^{(m)}\right)\left(n_{00} + n_{10}\right)},$$

$$\hat{p}^{(m)} = \frac{s + y^{(m)}}{n},$$

$$\hat{\alpha}_1^{(m)} = \frac{\left(s + y^{(m)}\right)\left(s_{11} + y_{11}^{(m)}\right)}{\left[\sum_{i=0}^{1}\left(s_{i1} + y_{i1}^{(m)}\right)\right]\left[\sum_{j=0}^{1}\left(s_{1j} + y_{1j}^{(m)}\right)\right]},$$

$$\hat{\alpha}_0^{(m)} = \frac{\left(r + u - y^{(m)}\right)\left(r_{11} + u_{11} - y_{11}^{(m)}\right)}{\left[\sum_{i=0}^{1}\left(r_{i1} + u_{i1} - y_{i1}^{(m)}\right)\right]\left[\left\{\sum_{j=0}^{1}\left(r_{1j} + u_{1j} - y_{1j}^{(m)}\right)\right\}\right]}.$$

The ML estimate of $\boldsymbol{\theta}$ in the $\left(m+1\right)$th iteration, $\hat{\boldsymbol{\theta}}^{(m+1)}$, is calculated applying the previous equations substituting m with $m+1$, where

$$y_{ij}^{(m+1)} = u_{ij}\frac{\hat{\phi}_{ij}^{(k)}}{\hat{\phi}_{ij}^{(k)} + \hat{\varphi}_{ij}^{(k)}}, \quad i,j = 0,1,$$

and where $\hat{\phi}_{ij}^{(m)}$ ($\hat{\varphi}_{ij}^{(m)}$) is the estimate of ϕ_{ij} (φ_{ij}) in the *m*th iteration and it is obtained substituting in ϕ_{ij} (φ_{ij}) the parameters with their respective estimates obtained in the *m*th iteration of the algorithm. As initial value $y_{ij}^{(0)}$ one can take any value $0 \le y_{ij}^{(0)} \le u_{ij}$,

$i, j = 0,1$. The EM algorithm stops when the difference between the values of the log-likelihood functions of two consecutive iterations is equal to or less than a value δ, for example $\delta = 10^{-12}$. If the EM algorithm converges in M iterations, $\hat{\boldsymbol{\theta}} = \left(\hat{\kappa}_1(0), \hat{\kappa}_1(1), \hat{\kappa}_2(0), \hat{\kappa}_2(1), \hat{p}, \hat{\alpha}_1, \hat{\alpha}_0\right)^T$ is the final estimate obtained. The estimates of the weighted kappa coefficients obtained by applying the EM algorithm converge to the ML estimates (proof can be seen in Appendix A). Figure 1 shows the flowchart of the EM algorithm to estimate $\boldsymbol{\theta}$.

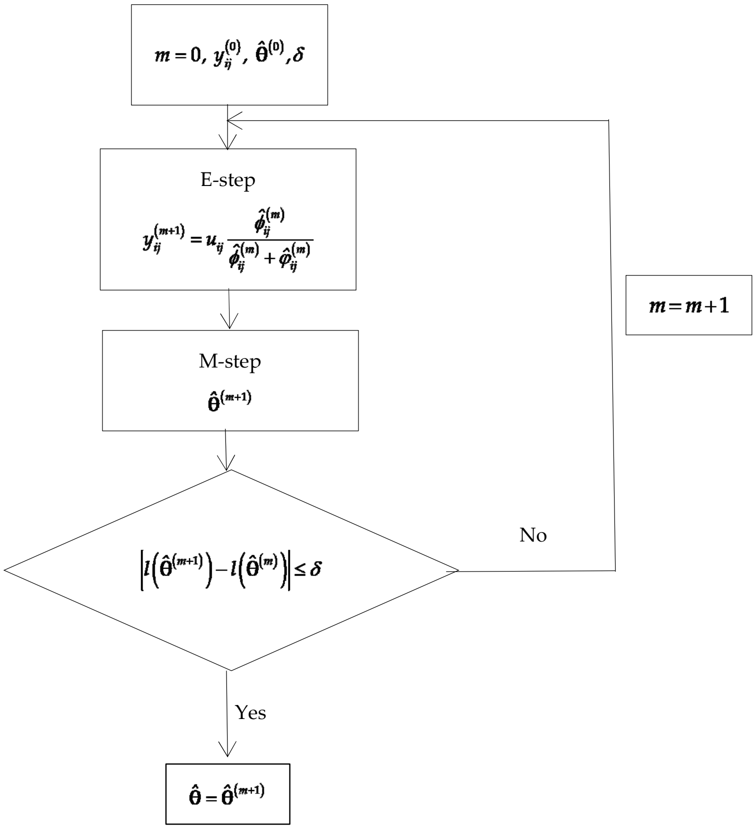
$m = 0,\ y_{ij}^{(0)},\ \hat{\boldsymbol{\theta}}^{(0)}, \delta$
E-step
$y_{ij}^{(m+1)} = u_{ij} \frac{\hat{\phi}_{ij}^{(m)}}{\hat{\phi}_{ij}^{(m)} + \hat{\varphi}_{ij}^{(m)}}$
M-step
$\hat{\boldsymbol{\theta}}^{(m+1)}$
$\left|l\left(\hat{\boldsymbol{\theta}}^{(m+1)}\right) - l\left(\hat{\boldsymbol{\theta}}^{(m)}\right)\right| \leq \delta$
No
$m = m + 1$
Yes
$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{(m+1)}$

Figure 1. Flowchart of the EM algorithm.

Once the value of $\hat{\kappa}_h(1)$ and $\hat{\kappa}_h(0)$ have been imputed, the estimates of average kappa coefficients are easily calculated by applying Equations (2) and (3), i.e.,

$$\hat{\kappa}_{h1} = \begin{cases} \frac{2\hat{\kappa}_h(0)\hat{\kappa}_h(1)}{\hat{\kappa}_h(0) - \hat{\kappa}_h(1)} \ln\left\{\frac{\hat{\kappa}_h(0) + \hat{\kappa}_h(1)}{2\hat{\kappa}_h(1)}\right\}, & \hat{p} \neq \hat{Q}_h \\ \hat{Y}_h, & \hat{p} = \hat{Q}_h, \end{cases}$$

and

$$\hat{\kappa}_{h2} = \begin{cases} \frac{2\hat{\kappa}_h(0)\hat{\kappa}_h(1)}{\hat{\kappa}_h(0) - \hat{\kappa}_h(1)} \ln\left\{\frac{2\hat{\kappa}_h(0)}{\hat{\kappa}_h(0) + \hat{\kappa}_h(1)}\right\}, & \hat{p} \neq \hat{Q}_h \\ \hat{Y}_h, & \hat{p} = \hat{Q}_h. \end{cases}$$

The estimates of Se_h and Sp_h are calculated as:

$$\hat{S}e_h = \frac{\hat{p}\hat{\kappa}_h(1) + \hat{q}\hat{\kappa}_h(0)\hat{\kappa}_h(1)}{\hat{q}\hat{\kappa}_h(0) + \hat{p}\hat{\kappa}_h(1)} \text{ and } \hat{S}p_h = \frac{\hat{q}\hat{\kappa}_h(0) + \hat{p}\hat{\kappa}_h(0)\hat{\kappa}_h(1)}{\hat{q}\hat{\kappa}_h(0) + \hat{p}\hat{\kappa}_h(1)}, \quad h = 1, 2,$$

where $\hat{q} = 1 - \hat{p}$. Once the ML estimates have been obtained, it is necessary to estimate their variances-covariances. For this we apply the Supplemented EM algorithm.

The variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is estimated by applying the supplemented EM (SEM) algorithm [18]. The SEM algorithm is a computational method which estimates the variances-covariances matrix from the calculations obtained by applying the EM algorithm. Dempster et al. [17] have shown that the matrix of variance-covariance of $\hat{\boldsymbol{\theta}}$ is expressed as

$$\hat{\Sigma}_{\hat{\boldsymbol{\theta}}} = I_{oc}^{-1}\left(I - DM\right)^{-1} \tag{10}$$

where I is the identity matrix, $DM = I_{mis}I_{oc}^{-1}$, I_{oc} is the Fisher information matrix of complete data and I_{mis} is the Fisher information matrix of missing data. The application of the SEM algorithm consists of three steps [18]: (1) calculate the matrix I_{oc}^{-1}, (2) calculate the DM matrix, and (3) calculate $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$. The main step is to calculate the DM matrix.

The first step consists of calculating I_{oc}^{-1}. This matrix is the inverse of the Fisher information matrix of the complete data, i.e., $I_{oc} = -\partial^2 l(\boldsymbol{\theta}) / \partial\theta_i\partial\theta_j$, where $l(\boldsymbol{\theta})$ is the function 9 and each θ_i is one of the parameters of $\boldsymbol{\theta}$. This matrix is calculated from the last 2×4 table obtained by applying the EM algorithm. Therefore, if the EM algorithm has converged in M iterations, then the frequencies of this table are $s_{ij} + x_{ij}^{(M)}$ for the diseased individuals and $r_{ij} + u_{ij} - x_{ij}^{(M)}$ for the non-diseased individuals.

The second step of the SEM algorithm consists of calculating the *DM* matrix. The elements (β_{ij} , $i, j = 1, \ldots, 7$) of this matrix are calculated by applying the following algorithm:

Input: $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^{(t)} = \left(\kappa_1^{(t)}(0), \kappa_1^{(t)}(1), \kappa_2^{(t)}(0), \kappa_2^{(t)}(1), p^{(t)}, \alpha_1^{(t)}, \alpha_0^{(t)}\right)^T$.

1. Calculate $\boldsymbol{\theta}^{(t+1)} = \left(\kappa_1^{(t+1)}(0), \kappa_1^{(t+1)}(1), \kappa_2^{(t+1)}(0), \kappa_2^{(t+1)}(1), p^{(t+1)}, \alpha_1^{(t+1)}, \alpha_0^{(t+1)}\right)^T$ applying the EM algorithm.
2. Obtain the vectors

$$\boldsymbol{\theta}_1^{(t)} = \left(\kappa_1^{(t)}(0), \hat{\kappa}_1(1), \hat{\kappa}_2(0), \hat{\kappa}_2(1), \hat{p}, \hat{\alpha}_1, \hat{\alpha}_0\right)^T$$

$$\boldsymbol{\theta}_2^{(t)} = \left(\hat{\kappa}_1(0), \kappa_2^{(t)}(1), \hat{\kappa}_2(0), \hat{\kappa}_2(1), \hat{p}, \hat{\alpha}_1, \hat{\alpha}_0\right)^T$$

$$\boldsymbol{\theta}_3^{(t)} = \left(\hat{\kappa}_1(0), \hat{\kappa}_1(1), \kappa_2^{(t)}(0), \hat{\kappa}_2(1), \hat{p}, \hat{\alpha}_1, \hat{\alpha}_0\right)^T$$

$$\boldsymbol{\theta}_4^{(t)} = \left(\hat{\kappa}_1(0), \hat{\kappa}_1(1), \hat{\kappa}_2(0), \kappa_2^{(t)}(1), \hat{p}, \hat{\alpha}_1, \hat{\alpha}_0\right)^T$$

$$\boldsymbol{\theta}_5^{(t)} = \left(\hat{\kappa}_1(0), \hat{\kappa}_1(1), \hat{\kappa}_2(0), \hat{\kappa}_2(0), \hat{p}^{(t)}, \hat{\alpha}_1, \hat{\alpha}_0\right)^T$$

$$\boldsymbol{\theta}_6^{(t)} = \left(\hat{\kappa}_1(0), \hat{\kappa}_1(1), \hat{\kappa}_2(0), \hat{\kappa}_2(0), \hat{p}, \hat{\alpha}_1^{(t)}, \hat{\alpha}_0\right)^T$$

$$\boldsymbol{\theta}_7^{(t)} = \left(\hat{\kappa}_1(0), \hat{\kappa}_1(1), \hat{\kappa}_2(0), \hat{\kappa}_2(0), \hat{p}, \hat{\alpha}_1, \hat{\alpha}_0^{(t)}\right)^T$$

and for each one of these vectors run the first iteration of the EM algorithm taking $\hat{\boldsymbol{\theta}}_i^{(t)}$ as the initial value of $\boldsymbol{\theta}$ and obtain the vectors $\hat{\hat{\boldsymbol{\theta}}}_1^{(t+1)},\ldots,\ \hat{\hat{\boldsymbol{\theta}}}_7^{(t+1)}$.

3. Calculate

$$\beta_{ij}^{(t)} = \frac{\hat{\hat{\theta}}_{ij}^{(t+1)} - \hat{\theta}_j}{\theta_i^{(t)} - \hat{\theta}_i},\ \ i,j = 1,\ldots,7\ ,$$

where $\hat{\hat{\theta}}_{ij}^{(t+1)}$ is the *j*th component of $\hat{\hat{\boldsymbol{\theta}}}_i^{(t+1)}$, $\theta_i^{(t)}$ is the *i*th component of $\boldsymbol{\theta}^{(t)}$ and $\hat{\theta}_i$ is the *i*th component of $\hat{\boldsymbol{\theta}}$.

Output: $\hat{\boldsymbol{\theta}}^{(t+1)}$ and $\beta_{ij}^{(t)},\ i,j = 1,\ldots,7$.

This algorithm is repeated until $\left|\beta_{ij}^{(t+1)} - \beta_{ij}^{(t)}\right| \le \sqrt{\delta}$ [18], where δ is the stop criterion of the EM algorithm. Figure 2 shows the flowchart of the SEM algorithm to calculate the *DM* matrix.

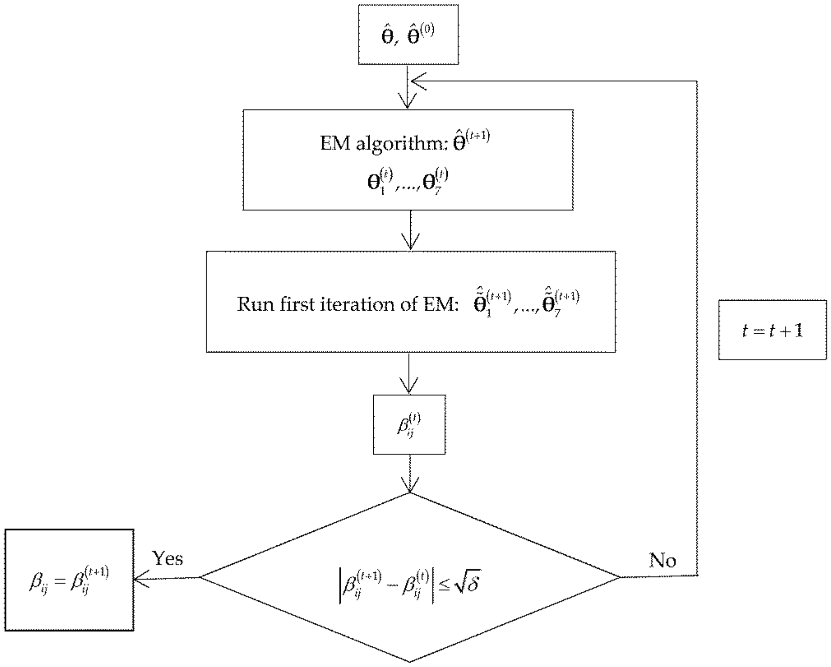
$\hat{\boldsymbol{\theta}},\ \hat{\boldsymbol{\theta}}^{(0)}$
EM algorithm: $\hat{\boldsymbol{\theta}}^{(t+1)}$
$\boldsymbol{\theta}_1^{(t)},\ldots,\boldsymbol{\theta}_7^{(t)}$
Run first iteration of EM: $\hat{\hat{\boldsymbol{\theta}}}_1^{(t+1)},\ldots,\hat{\hat{\boldsymbol{\theta}}}_7^{(t+1)}$
$\beta_{ij}^{(t)}$
$\left|\beta_{ij}^{(t+1)} - \beta_{ij}^{(t)}\right| \le \sqrt{\delta}$
Yes
No
$\beta_{ij} = \beta_{ij}^{(t+1)}$
$t = t+1$

Figure 2. Flowchart of the second step of the SEM algorithm.

The smaller δ is, the smaller are the errors that are made when calculating the *DM* matrix, and then smaller are the errors that are committed when calculating the variance-covariance matrix $\sum_{\hat{\boldsymbol{\theta}}}$.

The third and final step of the SEM algorithm consists of estimating the variance-covariance matrix $\Sigma_{\hat{\boldsymbol{\theta}}}$ applying equation 10. This matrix is not normally symmetrical due to the numerical errors made in the calculation of the *DM* matrix [18]. The assessment of $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ is performed calculating the matrix $\Delta\hat{\Sigma}_{\hat{\boldsymbol{\theta}}} = I_{oc}^{-1}DM\left(I - DM\right)^{-1}$ [18], a matrix which represents the increase in the variances-covariances estimated owing to the missing information. The matrix $\Delta\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ is the more symmetric the smaller the value of δ, therefore the asymmetry of $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ is solved taking a value a very small value of δ [18].

Once the matrix $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ has been calculated, the asymptotic variance-covariance matrix of the average kappa coefficients is obtained by applying the delta method. Let $\boldsymbol{\kappa}_1 = \left(\kappa_{11}, \kappa_{21}\right)^T$ and $\boldsymbol{\kappa}_2 = \left(\kappa_{12}, \kappa_{22}\right)^T$ be the vectors whose components are the average kappa coefficients. Let $\boldsymbol{\kappa} = \left(\kappa_1\left(0\right), \kappa_1\left(1\right), \kappa_2\left(0\right), \kappa_2\left(1\right), p\right)^T$ be the vectors whose components are the weighted kappa coefficients and the prevalence, and let $\hat{\Sigma}_{\hat{\boldsymbol{\kappa}}}$ be the estimated asymptotic variance-covariance of $\hat{\boldsymbol{\kappa}}$ (obtained by eliminating the variances and covariances corresponding to $\hat{\alpha}_1$ and $\hat{\alpha}_0$ from the matrix $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$), since the average kappa coefficients do not depend on the covariances α_1 and α_0. Then, applying the delta method, the asymptotic variance-covariance matrices are

$$\hat{\Sigma}_{\hat{\boldsymbol{\kappa}}_i} = \left(\frac{\partial \boldsymbol{\kappa}_i}{\partial \boldsymbol{\kappa}}\right)_{\boldsymbol{\kappa}=\hat{\boldsymbol{\kappa}}} \hat{\Sigma}_{\hat{\boldsymbol{\kappa}}} \left(\frac{\partial \boldsymbol{\kappa}_{i1}}{\partial \boldsymbol{\kappa}}\right)^T_{\boldsymbol{\kappa}=\hat{\boldsymbol{\kappa}}}, \quad i = 1, 2\,. \tag{11}$$

Once the estimates of the average kappa coefficients and their variances-covariances have been calculated, the test statistics for the hypothesis tests

$$\mathrm{H}_0\!:\kappa_{1i} = \kappa_{2i} \text{ vs } \mathrm{H}_1\!:\kappa_{1i} \neq \kappa_{2i}\,, \quad i = 1, 2\,,$$

are

$$z_i = \frac{\hat{\kappa}_{1i} - \hat{\kappa}_{2i}}{\sqrt{\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\kappa}_{1i}\right) + \hat{\mathrm{V}}\mathrm{ar}\left(\hat{\kappa}_{2i}\right) - 2\hat{\mathrm{C}}\mathrm{ov}\left(\hat{\kappa}_{1i}, \hat{\kappa}_{2i}\right)}}\,, \quad i = 1, 2\,,$$

whose distribution is a normal standard distribution when the sample size n is large. Inverting each test statistic, the $100 \times \left(1 - \alpha\right)\%$ Wald-type confidence interval for the difference of the two average kappa coefficients is

$$\kappa_{1i} - \kappa_{2i} \in \hat{\kappa}_{1i} - \hat{\kappa}_{2i} \pm z_{1-\alpha/2} \sqrt{\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\kappa}_{1i}\right) + \hat{\mathrm{V}}\mathrm{ar}\left(\hat{\kappa}_{2i}\right) - 2\hat{\mathrm{C}}\mathrm{ov}\left(\hat{\kappa}_{1i}, \hat{\kappa}_{2i}\right)}\,, \quad i = 1, 2\,,$$

where $z_{1-\alpha/2}$ is the $100 \times \left(1 - \alpha/2\right)$th percentile of the normal standard distribution.

4. Simulation Study

Monte Carlo simulation experiments have been carried out to study the sizes and the powers of the hypothesis tests 4 and 5 solved with the EM-SEM algorithms. These experiments have consisted of generating $N = 10,000$ random samples of multinomial distributions. As sample size we have considered $n = \left\{50, 100, 200, 500, 1000, 2000\right\}$. Probabilities of multinomial distributions have been calculated from equations 6 written in terms of the weighted kappa coefficients. These simulation experiments have been designed from the equations of the average kappa coefficients (Equations (2) and (3)). For the prevalence, the values 5%, 10%, 30% and 50% have been considered, and that it is a sufficient range of values to study the effect of prevalence on the behaviour of the hypothesis tests. Regarding the average kappa coefficients, the values 0.2, 0.4, 0.6 and 0.8 have been considered, values that correspond to different levels of clinical significance [16]. Once the values for the disease prevalence and the average kappa coefficient have been set, the values of $\kappa_h\left(0\right)$ and $\kappa_h\left(1\right)$ are calculated by solving (using the Newton-Raphson method) the system formed by Equations (2) and (3), considering only the solutions that are between 0 and 1. Next, the values of Se_h and Sp_h are calculated by applying equation (8). Once the values for Se_h and Sp_h have been calculated, the maximum values of the covariances α_1 and α_0 have been calculated by applying

Equation (7), considering intermediate values (50% of the maximum value) and high values (90% of the maximum value), i.e.,:

$$\alpha_1 = \frac{f}{\max\left\{Se_1, Se_2\right\}} + 1 - f \text{ and } \alpha_0 = \frac{f}{\max\left\{\left(1 - Sp_1\right), \left(1 - Sp_2\right)\right\}} + 1 - f\,,$$

with $f = \left\{0.50, 0.90\right\}$. As verification probabilities, three scenarios have been considered: $\lambda_{11} = 0.50, \lambda_{10} = \lambda_{01} = 0.30, \lambda_{00} = 0.05$, $\lambda_{11} = 0.95, \lambda_{10} = \lambda_{01} = 0.60, \lambda_{00} = 0.25$ and $\lambda_{11} = \lambda_{10} = \lambda_{01} = \lambda_{00} = 1$. The first scenario corresponds to a situation in which the verification is low, the second corresponds to a situation in which the verification is high and the third scenario corresponds to the situation in which all individuals are verified with the GS (a situation that can be called complete verification). In the last scenario, there is no verification bias and the sample design corresponds to a paired design, and the average kappa coefficients are compared using the method of Roldán-Nofuentes and Olvera-Porcel [6]. Finally, the probabilities of the multinomial distributions have been calculated by applying Equation (6) (in terms of the weighted kappa coefficients). Therefore, the probabilities of the multinomial distributions have been calculated from the values of the average kappa coefficients and not by fixing the sensitivities and specificities of the BDTs.

The Monte Carlo simulation experiments have been designed in such a way that in all of the random samples it is possible to apply the EM-SEM algorithms. For the application of the EM-SEM algorithms, the values $\delta = 10^{-12}$ and $\sqrt{\delta} = 10^{-6}$ have been considered as stop criterion, and $y_{ij}^{(0)} = u_{ij}/2$ as initial values of the EM algorithm. As nominal error, $\alpha = 5\%$ has been considered.

The simulation experiments have been carried out with R [23], and have been made with computers i7-3770 CPU 3.4 GHz. For this, a function, called "cakcmd" (Comparison of Average Kappa Coefficients with Missing Data), has been programmed to solve the hypothesis tests 1 and 2 applying the EM and SEM algorithms. The function runs with the command

$$\text{cakcmd}\left(\text{s11,s10,s01,s00,r11,r10,r01,r00,u11,u10,u01,u00}\right).$$

By default the stop criterion of the EM algorithm is 10^{-12}, the confidence level for the CIs is 95% and $y_{ij}^{(0)} = u_{ij}/2$. The function does not use any R library and the EM and SEM algorithms have been specifically programmed. The function always checks that the problem can be solved by applying the methods described, for example that there are no negative frequencies, that $u > 0$, etc. The function provides all the estimates and their standard errors, all the matrices described in Section 3, the test statistics, the p-values and the CIs for the difference between the two average kappa coefficients. The "cakcmd" function is available as Supplementary Materials to this manuscript.

Table 3 shows the type I error (in %) of the hypothesis test to compare the two average kappa coefficients when $L' > L$ ($0 \leq c < 0.5$) for different scenarios. The verification probabilities and the covariances α_1 and α_0 have an important effect on the type I error of the hypothesis test. For fixed values of the covariances, the increase in the verification probabilities produces an increase in the type I error. For fixed values of the verification probabilities, the increase in the covariances produces a decrease in the type I error. In general terms and depending on the verification probabilities and on the covariances, the type I error is very small (much lower than the nominal error) when the sample size is not very large ($n \leq 500$), and fluctuates around the error nominal (without exceeding it excessively) when the sample size is very large ($n \geq 1000$). Therefore, this hypothesis test is a conservative test (which is preferable to a liberal test) when the sample size is not very large and it has the behaviour of an asymptotic test when the sample size is very large. The hypothesis test does not give too many false significances even when the sample size is very large.

In the complete verification situation ($\lambda_{ij}=1$), the type I error behaves in a very similar way to the type I error obtained in partial verification. Comparing the partial verification scenarios with the complete verification scenario, the partial verification implies a decrease in type I error. Consequently, the presence of missing data implies that the type I error decreases with respect to the situation in which all individuals are verified with the GS.

Table 3. Type I error (in %) of the hypothesis test when $L' > L$ ($0 \le c < 0.5$).

$\kappa_{11}=\kappa_{21}=0.2$ $\kappa_1(0)=0.16$ $\kappa_1(1)=0.67$ $\kappa_2(0)=0.16$ $\kappa_2(1)=0.67$ $p=10\%$						
	$\lambda_{11}=0.50,\lambda_{10}=0.30,\lambda_{01}=0.30,\lambda_{00}=0.05$		$\lambda_{11}=0.95,\lambda_{10}=0.60,\lambda_{01}=0.60,\lambda_{00}=0.25$		$\lambda_{11}=\lambda_{10}=\lambda_{01}=\lambda_{00}=1$	
n	$\alpha_1=1.14$ $\alpha_0=2.37$	$\alpha_1=1.24$ $\alpha_0=3.47$	$\alpha_1=1.14$ $\alpha_0=2.37$	$\alpha_1=1.24$ $\alpha_0=3.47$	$\alpha_1=1.14$ $\alpha_0=2.37$	$\alpha_1=1.24$ $\alpha_0=3.47$
50	0	0	0.05	0	0.50	0
100	0.05	0	0.50	0	1.20	0
200	0.15	0	0.85	0	3.10	0.10
500	1.10	0.10	2.90	0.10	4.40	1.05
1000	1.70	0.20	3.40	0.95	4.75	2.05
2000	3.25	0.55	4.55	2.25	5.50	4.35
$\kappa_{11}=\kappa_{21}=0.4$ $\kappa_1(0)=0.34$ $\kappa_1(1)=0.78$ $\kappa_2(0)=0.34$ $\kappa_2(1)=0.78$ $p=30\%$						
	$\lambda_{11}=0.50,\lambda_{10}=0.30,\lambda_{01}=0.30,\lambda_{00}=0.05$		$\lambda_{11}=0.95,\lambda_{10}=0.60,\lambda_{01}=0.60,\lambda_{00}=0.25$		$\lambda_{11}=\lambda_{10}=\lambda_{01}=\lambda_{00}=1$	
n	$\alpha_1=1.06$ $\alpha_0=2.03$	$\alpha_1=1.11$ $\alpha_0=2.85$	$\alpha_1=1.06$ $\alpha_0=2.03$	$\alpha_1=1.11$ $\alpha_0=2.85$	$\alpha_1=1.06$ $\alpha_0=2.03$	$\alpha_1=1.11$ $\alpha_0=2.85$
50	0	0	0.45	0	2.05	1.10
100	0.30	0	1.50	0	4.50	3.90
200	1.40	0	2.30	0.25	4.90	4.35
500	2.90	0.45	4.15	1.25	4.25	3.55
1000	3.85	1.90	5.15	2.35	5.25	4.70
2000	4.55	2.65	4.75	4.15	4.80	4.40
$\kappa_{11}=\kappa_{21}=0.6$ $\kappa_1(0)=0.77$ $\kappa_1(1)=0.34$ $\kappa_2(0)=0.77$ $\kappa_2(1)=0.34$ $p=5\%$						
	$\lambda_{11}=0.50,\lambda_{10}=0.30,\lambda_{01}=0.30,\lambda_{00}=0.05$		$\lambda_{11}=0.95,\lambda_{10}=0.60,\lambda_{01}=0.60,\lambda_{00}=0.25$		$\lambda_{11}=\lambda_{10}=\lambda_{01}=\lambda_{00}=1$	
n	$\alpha_1=1.91$ $\alpha_0=96.02$	$\alpha_1=2.64$ $\alpha_0=172.39$	$\alpha_1=1.91$ $\alpha_0=96.02$	$\alpha_1=2.64$ $\alpha_0=172.39$	$\alpha_1=1.91$ $\alpha_0=96.02$	$\alpha_1=2.64$ $\alpha_0=172.39$
50	0	0	0	0	0.60	0.10
100	0.05	0	0.05	0	1.25	0.15
200	0.45	0	0.35	0	3.30	1.05
500	0.60	0.05	2.05	0.15	5.35	3.75
1000	1.60	0.25	4.15	0.45	4.95	4.90
2000	3.45	0.65	4.50	1.50	4.55	4.40
$\kappa_{11}=\kappa_{21}=0.8$ $\kappa_1(0)=0.86$ $\kappa_1(1)=0.66$ $\kappa_2(0)=0.86$ $\kappa_2(1)=0.66$ $p=50\%$						
	$\lambda_{11}=0.50,\lambda_{10}=0.30,\lambda_{01}=0.30,\lambda_{00}=0.05$		$\lambda_{11}=0.95,\lambda_{10}=0.60,\lambda_{01}=0.60,\lambda_{00}=0.25$		$\lambda_{11}=\lambda_{10}=\lambda_{01}=\lambda_{00}=1$	
n	$\alpha_1=1.12$ $\alpha_0=8.73$	$\alpha_1=1.21$ $\alpha_0=14.91$	$\alpha_1=1.12$ $\alpha_0=8.73$	$\alpha_1=1.21$ $\alpha_0=14.91$	$\alpha_1=1.12$ $\alpha_0=8.73$	$\alpha_1=1.21$ $\alpha_0=14.91$
50	0	0	0	0	0.30	0.10
100	0.05	0	0.40	0	2.25	0.25
200	0.45	0	1.65	0	4.25	1.05
500	2.40	0.05	2.90	0.55	5.55	3.40
1000	3.65	0.90	4.65	1.15	5.35	4.10
2000	3.75	2.40	5.35	3.35	5.60	5.10

Table 4 shows the type I error (in %) of the hypothesis test to compare the two average kappa coefficients when $L > L'$ ($0.5 < c \le 1$) for different scenarios. The verification probabilities and the covariances also have an important effect on the type I error of this hypothesis test, its effects being the same as in the previous case. The type I error of this test has the same behaviour as that of the previous hypothesis test, and is therefore a conservative test when the sample size is not very large and fluctuates around the nominal error when the sample size is very large. Comparing the partial verification scenarios with the full verification scenario, the same conclusions as those previous are obtained.

Table 4. Type I error (in %) of the hypothesis test when $L > L'$ ($0.5 < c \le 1$).

			$\kappa_{12} = \kappa_{22} = 0.2$ $\kappa_1(0) = 0.93$ $\kappa_1(1) = 0.16$ $\kappa_2(0) = 0.93$ $\kappa_2(1) = 0.16$ $p = 50\%$			
	$\lambda_{11} = 0.50, \lambda_{10} = 0.30, \lambda_{01} = 0.30, \lambda_{00} = 0.05$		$\lambda_{11} = 0.95, \lambda_{10} = 0.60, \lambda_{01} = 0.60, \lambda_{00} = 0.25$		$\lambda_{11} = \lambda_{10} = \lambda_{01} = \lambda_{00} = 1$	
n	$\alpha_1 = 2.26$ $\alpha_0 = 49.16$	$\alpha_1 = 3.28$ $\alpha_0 = 87.69$	$\alpha_1 = 2.26$ $\alpha_0 = 49.16$	$\alpha_1 = 3.28$ $\alpha_0 = 87.69$	$\alpha_1 = 2.26$ $\alpha_0 = 49.16$	$\alpha_1 = 3.28$ $\alpha_0 = 87.69$
50	0	0	0.40	0	2.55	1.05
100	0.50	0	1.70	0	4.45	2.15
200	1.70	0	2.80	0	4.90	3.30
500	4.30	0.60	3.70	2.20	4.20	4.40
1000	4.10	2.90	4.50	4.35	5.45	4.05
2000	4.30	3.60	5.20	5.05	5.75	5.35
			$\kappa_{12} = \kappa_{22} = 0.4$ $\kappa_1(0) = 0.16$ $\kappa_1(1) = 0.67$ $\kappa_2(0) = 0.16$ $\kappa_2(1) = 0.67$ $p = 10\%$			
	$\lambda_{11} = 0.50, \lambda_{10} = 0.30, \lambda_{01} = 0.30, \lambda_{00} = 0.05$		$\lambda_{11} = 0.95, \lambda_{10} = 0.60, \lambda_{01} = 0.60, \lambda_{00} = 0.25$		$\lambda_{11} = \lambda_{10} = \lambda_{01} = \lambda_{00} = 1$	
n	$\alpha_1 = 1.14$ $\alpha_0 = 2.37$	$\alpha_1 = 1.26$ $\alpha_0 = 3.47$	$\alpha_1 = 1.14$ $\alpha_0 = 2.37$	$\alpha_1 = 1.26$ $\alpha_0 = 3.47$	$\alpha_1 = 1.14$ $\alpha_0 = 2.37$	$\alpha_1 = 1.26$ $\alpha_0 = 3.47$
50	0	0	0	0	0.05	0
100	0	0	0.20	0.05	0.60	0.15
200	0.20	0.10	0.90	0.55	2.25	0.95
500	0.80	0.45	3.00	2.15	4.55	3.35
1000	1.80	1.30	3.60	2.25	5.25	3.55
2000	3.75	2.80	4.25	3.35	5.10	3.80
			$\kappa_{12} = \kappa_{22} = 0.6$ $\kappa_1(0) = 0.34$ $\kappa_1(1) = 0.78$ $\kappa_2(0) = 0.34$ $\kappa_2(1) = 0.78$ $p = 30\%$			
	$\lambda_{11} = 0.50, \lambda_{10} = 0.30, \lambda_{01} = 0.30, \lambda_{00} = 0.05$		$\lambda_{11} = 0.95, \lambda_{10} = 0.60, \lambda_{01} = 0.60, \lambda_{00} = 0.25$		$\lambda_{11} = \lambda_{10} = \lambda_{01} = \lambda_{00} = 1$	
n	$\alpha_1 = 1.06$ $\alpha_0 = 2.03$	$\alpha_1 = 1.11$ $\alpha_0 = 2.85$	$\alpha_1 = 1.06$ $\alpha_0 = 2.03$	$\alpha_1 = 1.11$ $\alpha_0 = 2.85$	$\alpha_1 = 1.06$ $\alpha_0 = 2.03$	$\alpha_1 = 1.11$ $\alpha_0 = 2.85$
50	0	0	0.10	0	0.20	0.05
100	0.30	0.05	0.70	0.10	2.35	0.15
200	1.05	0.10	1.50	0.65	4.35	0.80
500	2.50	0.70	4.50	1.40	5.35	3.15
1000	4.10	1.40	5.10	1.80	4.95	4.80
2000	4.80	2.60	5.15	3.15	5.95	5.50
			$\kappa_{12} = \kappa_{22} = 0.8$ $\kappa_1(0) = 0.88$ $\kappa_1(1) = 0.78$ $\kappa_2(0) = 0.88$ $\kappa_2(1) = 0.78$ $p = 5\%$			
	$\lambda_{11} = 0.50, \lambda_{10} = 0.30, \lambda_{01} = 0.30, \lambda_{00} = 0.05$		$\lambda_{11} = 0.95, \lambda_{10} = 0.60, \lambda_{01} = 0.60, \lambda_{00} = 0.25$		$\lambda_{11} = \lambda_{10} = \lambda_{01} = \lambda_{00} = 1$	
n	$\alpha_1 = 1.13$ $\alpha_0 = 93.98$	$\alpha_1 = 1.24$ $\alpha_0 = 168.37$	$\alpha_1 = 1.13$ $\alpha_0 = 93.98$	$\alpha_1 = 1.24$ $\alpha_0 = 168.37$	$\alpha_1 = 1.13$ $\alpha_0 = 93.98$	$\alpha_1 = 1.24$ $\alpha_0 = 168.37$
50	0	0	0	0	0.05	0.10
100	0.05	0	0.20	0.10	0.45	0.25
200	0.10	0	0.45	0.15	0.40	0.55
500	0.55	0.30	0.90	0.50	2.05	1.45
1000	1.25	0.95	3.05	1.95	3.55	3.25
2000	2.10	1.30	3.85	2.65	4.25	3.05

Table 5 shows the power (in %) of the hypothesis test when $L' > L$ ($0 \leq c < 0.5$) for different values of the average kappa coefficients.

Table 5. Power (in %) of the hypothesis test when $L' > L$ ($0 \leq c < 0.5$).

$\kappa_{11}=0.4$ $\kappa_{21}=0.2$ $\kappa_1(0)=0.34$ $\kappa_1(1)=0.78$ $\kappa_2(0)=0.16$ $\kappa_2(1)=0.67$ $p=10\%$						
	$\lambda_{11}=0.50, \lambda_{10}=0.30, \lambda_{01}=0.30, \lambda_{00}=0.05$		$\lambda_{11}=0.95, \lambda_{10}=0.60, \lambda_{01}=0.60, \lambda_{00}=0.25$		$\lambda_{11}=\lambda_{10}=\lambda_{01}=\lambda_{00}=1$	
n	$\alpha_1=1.11$ $\alpha_0=2.37$	$\alpha_1=1.19$ $\alpha_0=3.47$	$\alpha_1=1.11$ $\alpha_0=2.37$	$\alpha_1=1.19$ $\alpha_0=3.47$	$\alpha_1=1.11$ $\alpha_0=2.37$	$\alpha_1=1.19$ $\alpha_0=3.47$
50	0.15	0.05	1.30	0.85	7.85	22.35
100	3.80	3.00	17.95	21.05	61.15	77.15
200	26.45	36.00	64.15	86.45	93.10	99.65
500	81.90	97.95	99.05	100	100	100
1000	99.15	100	100	100	100	100
2000	100	100	100	100	100	100
$\kappa_{11}=0.6$ $\kappa_{21}=0.4$ $\kappa_1(0)=0.56$ $\kappa_1(1)=0.76$ $\kappa_2(0)=0.34$ $\kappa_2(1)=0.78$ $p=30\%$						
	$\lambda_{11}=0.50, \lambda_{10}=0.30, \lambda_{01}=0.30, \lambda_{00}=0.05$		$\lambda_{11}=0.95, \lambda_{10}=0.60, \lambda_{01}=0.60, \lambda_{00}=0.25$		$\lambda_{11}=\lambda_{10}=\lambda_{01}=\lambda_{00}=1$	
n	$\alpha_1=1.06$ $\alpha_0=2.03$	$\alpha_1=1.11$ $\alpha_0=2.85$	$\alpha_1=1.06$ $\alpha_0=2.03$	$\alpha_1=1.11$ $\alpha_0=2.85$	$\alpha_1=1.06$ $\alpha_0=2.03$	$\alpha_1=1.11$ $\alpha_0=2.85$
50	3.35	3.40	24.45	31.75	29.05	38.55
100	37.80	54.60	83.15	94.30	85.10	82.05
200	87.90	98.10	99.75	100	100	100
500	99.95	100	100	100	100	100
1000	100	100	100	100	100	100
2000	100	100	100	100	100	100
$\kappa_{11}=0.6$ $\kappa_{21}=0.2$ $\kappa_1(0)=0.56$ $\kappa_1(1)=0.76$ $\kappa_2(0)=0.17$ $\kappa_2(1)=0.39$ $p=5\%$						
	$\lambda_{11}=0.50, \lambda_{10}=0.30, \lambda_{01}=0.30, \lambda_{00}=0.05$		$\lambda_{11}=0.95, \lambda_{10}=0.60, \lambda_{01}=0.60, \lambda_{00}=0.25$		$\lambda_{11}=\lambda_{10}=\lambda_{01}=\lambda_{00}=1$	
n	$\alpha_1=1.14$ $\alpha_0=6.09$	$\alpha_1=1.26$ $\alpha_0=10.16$	$\alpha_1=1.14$ $\alpha_0=6.09$	$\alpha_1=1.26$ $\alpha_0=10.16$	$\alpha_1=1.14$ $\alpha_0=6.09$	$\alpha_1=1.26$ $\alpha_0=10.16$
50	0.60	0.25	6.05	3.20	18.75	31.05
100	11.75	13.20	29.50	44.10	97.45	99.05
200	43.40	63.05	69.60	90.60	100	100
500	87.30	98.25	98.60	99.95	100	100
1000	99.55	99.95	100	100	100	100
2000	100	100	100	100	100	100
$\kappa_{11}=0.8$ $\kappa_{21}=0.6$ $\kappa_1(0)=0.90$ $\kappa_1(1)=0.60$ $\kappa_2(0)=0.80$ $\kappa_2(1)=0.33$ $p=50\%$						
	$\lambda_{11}=0.50, \lambda_{10}=0.30, \lambda_{01}=0.30, \lambda_{00}=0.05$		$\lambda_{11}=0.95, \lambda_{10}=0.60, \lambda_{01}=0.60, \lambda_{00}=0.25$		$\lambda_{11}=\lambda_{10}=\lambda_{01}=\lambda_{00}=1$	
n	$\alpha_1=1.16$ $\alpha_0=9.06$	$\alpha_1=1.28$ $\alpha_0=15.51$	$\alpha_1=1.16$ $\alpha_0=9.06$	$\alpha_1=1.28$ $\alpha_0=15.51$	$\alpha_1=1.16$ $\alpha_0=9.06$	$\alpha_1=1.28$ $\alpha_0=15.51$
50	0.10	0.05	0.15	0.10	9.95	23.05
100	0.15	0.10	0.20	0.15	70.05	77.95
200	0.30	0.15	2.75	1.95	96.10	100
500	7.65	5.65	30.10	39.30	100	100
1000	34.45	41.15	69.05	89.15	100	100
2000	70.35	89.55	95.10	99.85	100	100

Verification probabilities and covariances also have an important effect on the power of the hypothesis test. For fixed values of the covariances, increasing the verification probabilities produces an increase in power. With respect to the covariances, for fixed values of verification probabilities, in general terms their increase produces an increase in power (although when the sample is small or moderate, the power may decrease slightly, depending on the difference between the values of the average kappa coefficients). Comparing the partial verification scenarios with the complete verification scenario, the partial verification implies a lower power. A decrease in the verification probabilities implies a decrease in power, with respect to the complete verification situation. In very general terms, the following conclusions are obtained:

(1) When the difference between the two average kappa coefficients is small (0.2), a large ($n = 500$) or very large ($n \geq 1000$) sample is needed, for the power is greater than 80–90%, depending on the verification probabilities and on the covariances.

(2) When the difference between the two average kappa coefficients is moderate or large (≥ 0.4), a sample of moderate size ($n = 100 - 200$) is needed for the power to be greater than 80–90%, depending on the verification probabilities and on the covariances.

Table 6 shows the power (in %) of the hypothesis test when $L > L'$ ($0.5 < c \leq 1$) for different values of the average kappa coefficients. In general terms, the conclusions are the same as those obtained for the previous hypothesis test.

Table 6. Power (in %) of the hypothesis test when $L > L'$ ($0.5 < c \leq 1$).

$\kappa_{12} = 0.4$ $\kappa_{22} = 0.2$ $\kappa_1(0) = 0.27$ $\kappa_1(1) = 0.47$ $\kappa_2(0) = 0.39$ $\kappa_2(1) = 0.17$ $p = 30\%$						
	$\lambda_{11} = 0.50, \lambda_{10} = 0.30, \lambda_{01} = 0.30, \lambda_{00} = 0.05$		$\lambda_{11} = 0.95, \lambda_{10} = 0.60, \lambda_{01} = 0.60, \lambda_{00} = 0.25$		$\lambda_{11} = \lambda_{10} = \lambda_{01} = \lambda_{00} = 1$	
n	$\alpha_1 = 1.22$ $\alpha_0 = 2.10$	$\alpha_1 = 1.39$ $\alpha_0 = 2.99$	$\alpha_1 = 1.22$ $\alpha_0 = 2.10$	$\alpha_1 = 1.39$ $\alpha_0 = 2.99$	$\alpha_1 = 1.22$ $\alpha_0 = 2.10$	$\alpha_1 = 1.39$ $\alpha_0 = 2.99$
50	0.10	0.05	0.05	1.10	9.85	12.40
100	0.90	0.15	9.00	8.20	42.05	45.15
200	4.90	5.60	24.70	28.30	73.80	78.35
500	22.7	26.80	70.60	76.90	97.15	99.40
1000	54.05	59.20	94.10	98.40	100	100
2000	88.30	91.05	100	100	100	100
$\kappa_{12} = 0.6$ $\kappa_{22} = 0.4$ $\kappa_1(0) = 0.46$ $\kappa_1(1) = 0.66$ $\kappa_2(0) = 0.27$ $\kappa_2(1) = 0.47$ $p = 50\%$						
	$\lambda_{11} = 0.50, \lambda_{10} = 0.30, \lambda_{01} = 0.30, \lambda_{00} = 0.05$		$\lambda_{11} = 0.95, \lambda_{10} = 0.60, \lambda_{01} = 0.60, \lambda_{00} = 0.25$		$\lambda_{11} = \lambda_{10} = \lambda_{01} = \lambda_{00} = 1$	
n	$\alpha_1 = 1.22$ $\alpha_0 = 2.10$	$\alpha_1 = 1.39$ $\alpha_0 = 2.99$	$\alpha_1 = 1.22$ $\alpha_0 = 2.10$	$\alpha_1 = 1.39$ $\alpha_0 = 2.99$	$\alpha_1 = 1.22$ $\alpha_0 = 2.10$	$\alpha_1 = 1.39$ $\alpha_0 = 2.99$
50	0.05	0.01	3.30	1.60	13.25	16.75
100	5.60	4.10	19.05	29.90	56.85	74.05
200	25.50	29.20	49.30	78.90	84.15	99.10
500	65.50	90.30	89.40	99.90	100	100
1000	89.55	99.60	99.30	100	100	100
2000	99.80	100	100	100	100	100
$\kappa_{12} = 0.6$ $\kappa_{22} = 0.2$ $\kappa_1(0) = 0.34$ $\kappa_1(1) = 0.78$ $\kappa_2(0) = 0.39$ $\kappa_2(1) = 0.17$ $p = 10\%$						
	$\lambda_{11} = 0.50, \lambda_{10} = 0.30, \lambda_{01} = 0.30, \lambda_{00} = 0.05$		$\lambda_{11} = 0.95, \lambda_{10} = 0.60, \lambda_{01} = 0.60, \lambda_{00} = 0.25$		$\lambda_{11} = \lambda_{10} = \lambda_{01} = \lambda_{00} = 1$	
n	$\alpha_1 = 1.11$ $\alpha_0 = 4.23$	$\alpha_1 = 1.19$ $\alpha_0 = 6.81$	$\alpha_1 = 1.11$ $\alpha_0 = 4.23$	$\alpha_1 = 1.19$ $\alpha_0 = 6.81$	$\alpha_1 = 1.11$ $\alpha_0 = 4.23$	$\alpha_1 = 1.19$ $\alpha_0 = 6.81$
50	0.10	0.05	0.30	1.10	17.10	22.65
100	0.50	0.06	7.10	9.10	35.95	42.15
200	5.30	5.20	39.90	44.10	82.05	84.05
500	44.20	56.20	93.10	94.40	100	100
1000	91.70	94.30	99.80	100	100	100
2000	99.80	100	100	100	100	100
$\kappa_{12} = 0.8$ $\kappa_{22} = 0.6$ $\kappa_1(0) = 0.88$ $\kappa_1(1) = 0.78$ $\kappa_2(0) = 0.95$ $\kappa_2(1) = 0.53$ $p = 5\%$						
	$\lambda_{11} = 0.50, \lambda_{10} = 0.30, \lambda_{01} = 0.30, \lambda_{00} = 0.05$		$\lambda_{11} = 0.95, \lambda_{10} = 0.60, \lambda_{01} = 0.60, \lambda_{00} = 0.25$		$\lambda_{11} = \lambda_{10} = \lambda_{01} = \lambda_{00} = 1$	

n	$\alpha_1 = 1.13$ $\alpha_0 = 93.98$	$\alpha_1 = 1.24$ $\alpha_0 = 168.37$	$\alpha_1 = 1.13$ $\alpha_0 = 93.98$	$\alpha_1 = 1.24$ $\alpha_0 = 168.37$	$\alpha_1 = 1.13$ $\alpha_0 = 93.98$	$\alpha_1 = 1.24$ $\alpha_0 = 168.37$
50	0.05	0.01	0.10	0.05	14.20	17.85
100	0.08	0.03	0.08	0.02	44.85	52.10
200	0.10	0.08	2.60	2.70	89.05	96.95
500	5.30	7.05	20.30	23.10	100	100
1000	21.80	30.01	48.50	53.80	100	100
2000	49.70	63.70	84.05	86.80	100	100

5. Example

The model has been applied to the study by Hall et al. [24] on the diagnosis of Alzheimer's disease. Hall et al. have used two BDTs for the diagnosis of Alzheimer's disease: a new BDT based on a cognitive test applied to the patient (NBDT), and another BDT related to another person who knows the patient and a standard diagnostic test based on a cognitive test (CT). As a GS, a clinical assessment (a neurological exploration, computerized tomography, neuro-psychological and laboratory tests, etc.) has been used. This study corresponds to a two-phase study: in the first phase, two BDTs have been applied to all of the patients, and in the second phase only a subset of patients are verified with the GS, depending on the results of both BDTs [9]. Therefore, it is assumed that the verification process is MAR. Table 7 shows the data obtained by Hall et al. when applying medical tests to a sample of 588 patients, where T_1 models the result of the NBDT, T_2 models the result of the CT, and D models the result of the clinical assessment.

Table 7. Diagnosis of coronary stenosis.

Observed Frequencies				
	$T_1 = 1$		$T_1 = 0$	
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$
$V = 1$				
$D = 1$	31	5	3	1
$D = 0$	25	10	19	55
$V = 0$	22	6	65	346

Executing the "cakcmd" function with the command

$$\text{cakcmd}(31,5,3,1,25,10,19,55,22,6,65,346),$$

the results given in Table 8 are obtained.

Table 8. Results for the example of the diagnosis of Alzheimer's disease.

Comparison of Average Kappa Coefficients of Two Bdts with Missing Data:
Em and Sem Algorithms

Iterations of the EM Algorithm: 217

Inverse Matrix of the Fisher Information Matrix for Complete Data:

	Kappa10	Kappa11	Kappa20	Kappa21	p	a1	a0
Kappa10	2.70×10^{-3}	1.38×10^{-3}	9.91×10^{-4}	3.69×10^{-4}	3.70×10^{-4}	-3.64×10^{-4}	4.35×10^{-3}
Kappa11	1.38×10^{-3}	3.88×10^{-3}	2.19×10^{-4}	1.08×10^{-3}	-4.43×10^{-5}	1.08×10^{-3}	5.07×10^{-4}
Kappa20	9.91×10^{-4}	2.19×10^{-4}	1.13×10^{-3}	1.07×10^{-3}	2.77×10^{-4}	-3.24×10^{-4}	3.87×10^{-3}
Kappa21	3.69×10^{-4}	1.08×10^{-3}	1.07×10^{-3}	4.30×10^{-3}	-4.08×10^{-5}	-1.37×10^{-3}	1.14×10^{-3}
p	3.70×10^{-4}	-4.43×10^{-5}	2.77×10^{-4}	-4.08×10^{-5}	1.77×10^{-4}	-2.17×10^{-19}	-6.58×10^{-19}
a1	-3.64×10^{-4}	-1.08×10^{-3}	-3.24×10^{-4}	-1.37×10^{-3}	1.84×10^{-19}	2.21×10^{-3}	2.74×10^{-18}

a0	4.35×10^{-3}	5.07×10^{-4}	3.87×10^{-3}	1.14×10^{-3}	8.00×10^{-19}	-6.87×10^{-20}	1.15×10^{-1}

DM matrix:

	Kappa10	Kappa11	Kappa20	Kappa21	p	a1	a0
Kappa10	0.25747856	0.22670197	0.04820999	−0.03430295	0.02544226	−0.00081874	0.00737780
Kappa11	0.04018192	0.46969774	−0.15323480	−0.01112308	−0.06342952	0.01088978	−0.68269317
Kappa20	0.07169712	−0.31072616	0.30117681	0.43206262	0.06510425	−0.09497630	−0.06344007
Kappa21	−0.04157550	0.08880363	0.02017579	0.22133723	−0.01974722	−0.11024679	−0.25032953
p	−0.11756844	−1.20091231	−0.02896885	−1.17760749	0.15870433	0.64074585	1.86688820
a1	−0.03532342	−0.11283512	−0.15022054	−0.50246330	0.00919001	0.67340594	−1.33229242
a0	−0.00750592	−0.01428990	−0.00424797	−0.01550081	0.00044486	−0.01171524	0.09379489

Variance-covariance matrix of weighted kappa coefficients, prevalence and covariances:

	Kappa10	Kappa11	Kappa20	Kappa21	p	a1	a0
Kappa10	0.00380263	0.00282528	0.00127952	0.00097428	0.00040832	−0.00110158	0.00480053
Kappa11	0.00283219	0.01558280	−0.00086300	0.00807824	−0.00148826	−0.00801231	−0.00461472
Kappa20	0.00127269	−0.00094112	0.00233169	0.00289762	0.00053473	−0.00191751	0.00794194
Kappa21	0.00096024	0.00784534	0.00296983	0.01611484	−0.00088940	−0.01223615	0.00684820
p	0.00040347	−0.00152204	0.00051321	−0.00098494	0.00041010	0.00090368	0.00090234
a1	−0.00109354	−0.00785617	−0.00198050	−0.01227948	0.00083346	0.01314268	−0.00816490
a0	0.00477417	−0.00487427	0.00791000	0.00651661	0.00096002	−0.00789270	0.14222082

Estimated weighted kappa coefficient K(0) of Test 1 is 0.4410538 and its standard error is 0.06166551
Estimated weighted kappa coefficient K(1) of Test 1 is 0.6692124 and its standard error is 0.1248311
Estimated weighted kappa coefficient K(0) of Test 2 is 0.2446698 and its standard error is 0.04828762
Estimated weighted kappa coefficient K(1) of Test 2 is 0.7152702 and its standard error is 0.1269442
Estimated disease prevalence is 0.1177224 and its standard error is 0.0202509
Estimated covariance a1 is 1.082158
Estimated covariance a0 is 3.365059
COMPARISON OF AVERAGE KAPPA COEFFICIENTS FOR L′ > L (0 < c < 0.5)
Variance-covariance matrix:

	Average kappa11	Average kappa21
Average kappa11	0.003978628	0.001180153
Average kappa21	0.001159865	0.003010255

Estimated average kappa coefficient of Test 1 is 0.4835519 and its standard error is 0.06307636
Estimated average kappa coefficient of Test 2 is 0.2967101 and its standard error is 0.05486579
Test Statistic for the hypothesis test is 2.746314 and the p-value is 0.006026899
95% confidence interval for the difference between the two average kappa coefficients is: 0.05349828; 0.3201853

COMPARISON OF AVERAGE KAPPA COEFFICIENTS FOR L > L′ (0.5 < c < 1)
Variance-covariance matrix:

	Average kappa12	Average kappa22
Average kappa12	0.007956845	0.002206378
Average kappa22	0.002102579	0.006436081

Estimated average kappa coefficient of Test 1 is 0.5951878 and its standard error is 0.08920115
Estimated average kappa coefficient of Test 2 is 0.5011507 and its standard error is 0.08022519
Test Statistic for the hypothesis test is 0.9413048 and the p-value is 0.3465487
95% confidence interval for the difference between the two average kappa coefficients is: −0.1017649; 0.2898391

The EM algorithm has converged in 217 iterations using $\delta = 10^{-12}$ as the stopping criterion. The execution time of the function has been 0.2 s with a computer i7-3770 CPU 3.4 GHz. The estimates of the weighted kappa coefficients, prevalence and covariances are

$$\hat{\boldsymbol{\theta}} = \left(\hat{\kappa}_1(0), \hat{\kappa}_1(1), \hat{\kappa}_2(0), \hat{\kappa}_2(1), \hat{p}, \hat{\alpha}_1, \hat{\alpha}_0\right)^T \approx \left(0.44, 0.67, 0.24, 0.72, 0.12, 1.08, 3.37\right).$$

Applying the SEM algorithm, the variance-covariance matrix of $\hat{\boldsymbol{\theta}} = \left(\hat{\kappa}_1(0), \hat{\kappa}_1(1), \hat{\kappa}_2(0), \hat{\kappa}_2(1), \hat{p}, \hat{\alpha}_1, \hat{\alpha}_0\right)^T$ is obtained (see Table 8). The variance-covariance matrices of the estimates of average kappa coefficients are obtained from the previous matrix by applying the delta method (Equation (11)). All these matrices are not symmetric due to the numerical errors made in the application of the SEM algorithm.

If the clinician considers that false positives are more important than false negatives ($L' > L$ and $0 \le c < 0.5$), then the estimates of the average kappa coefficients are $\hat{\kappa}_{11} \approx 0.48$ and $\hat{\kappa}_{12} \approx 0.30$, and the estimates of the variances and covariance are $\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\kappa}_{11}\right) \approx 0.0040$, $\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\kappa}_{21}\right) \approx 0.0030$ and $\hat{\mathrm{C}}\mathrm{ov}\left(\hat{\kappa}_{11}, \hat{\kappa}_{21}\right) \approx 0.0012$. The value of the test statistic for $\mathrm{H}_0\!:\!\kappa_{11} = \kappa_{21}$ is $z_1 \approx 2.75$ (two sided p-value ≈ 0.0060). Therefore, with $\alpha = 5\%$, the equality of both average kappa coefficients is rejected. The average kappa coefficient of the NBDT is significantly higher than the average kappa coefficient of the CT (95% CI for the difference: 0.0535 to 0.3202). If the clinician considers that false positives are more important than false negatives, the average kappa coefficient of the NBDT is greater than the average kappa coefficient of the CT. Therefore, the average beyond-chance agreement between the new BDT and the clinical assessment is greater than the average beyond-chance agreement between the cognitive test and clinical assessment.

If the clinician considers that false negatives are more important than false positives ($L > L'$ and $0.5 < c \le 1$), then the estimates of the average kappa coefficients are $\hat{\kappa}_{12} \approx 0.60$ and $\hat{\kappa}_{22} \approx 0.50$, and the estimates of the variances and covariance are $\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\kappa}_{12}\right) \approx 0.0080$, $\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\kappa}_{22}\right) \approx 0.0064$ and $\hat{\mathrm{C}}\mathrm{ov}\left(\hat{\kappa}_{12}, \hat{\kappa}_{22}\right) \approx 0.0022$. The value of the test statistic for $\mathrm{H}_0\!:\!\kappa_{12} = \kappa_{22}$ is $z_2 = 0.9413$ (two sided p-value ≈ 0.3465). Therefore, with $\alpha = 5\%$ the equality of both average kappa coefficients is not rejected. With $\alpha = 5\%$, we cannot reject that the average kappa coefficient of the NBDT and CT are equal, and therefore we cannot reject that the average beyond-chance agreement between the NBDT and the clinical assessment is equal to the average beyond-chance agreement between the CT and clinical assessment (95% CI for the difference: −0.1018 to 0.2898).

6. Discussion and Conclusions

The average kappa coefficient of a BDT is a measure of average beyond-chance agreement between the BDT and the GS, and solves the problem of assigning values to the weighting index of the weighted kappa coefficient. The average kappa coefficient depends solely on the sensitivity and specificity of BDT and the prevalence of the disease, and is therefore a parameter that can be used to evaluate the efficacy of a BDT and to compare the efficacy of two (or more) BDTs. In this manuscript, the comparison of the average kappa coefficients of two BDTs is studied when the GS is not applied to all individuals in a sample. In this situation, the disease state is unknown for a subset of individuals and therefore the missing information is the true disease status for these individuals. The applied methods require the assumption that the missing data is MAR. This assumption is widely used in these types of studies, and establishes that the probability of verifying an individual with GS depends solely on the results of the two BDTs. This situation also corresponds to two-phase studies: in the first phase the two BDTs are applied to all individuals and in the second phase the GS is applied only to a subset of them depending on the results of the two BDTs in the previous phase.

Two hypothesis tests have been studied to compare the two average kappa coefficients: a first hypothesis test when false positives are more important than false negatives and another when false negatives are more important than false positives. For example, the first hypothesis test is applied when the two BDTs are used as confirmatory tests before a risk treatment, and the second hypothesis test is applied when the two BDTs are used as screening tests. Both hypothesis tests have been solved by applying computational methods for the estimation of parameters with missing data: the EM algorithm and the SEM. The EM algorithm allows us to estimate the parameters. The SEM algorithm, which is based on the calculations of the EM algorithm, allows us to estimate the variance-covariance matrix of the parameter vector. The EM algorithm requires assuming the MAR assumption. If the MAR assumption cannot be assumed, then the method proposed in this manuscript cannot be applied. For example, if the probability of verifying with the GS also depends on the disease status, then the MAR assumption is not verified. Future research will focus on studying, through a sensitivity analysis, the behavior of the hypothesis tests applying the EM-SEM algorithms when the MAR assumption is not verified.

Simulation experiments have been carried out to study the size and power of each hypothesis test. The results have shown that both hypothesis tests are conservative when the sample size is small or moderate, and that the type I error fluctuates around the nominal error when the sample size is large or very large. Regarding the power of each hypothesis test, in general terms, a moderate or large sample is necessary (depending on the verification probabilities, covariances, and difference between the values of the two average kappa coefficients) for the power of each hypothesis test to be large. Consequently, the two hypothesis tests have an asymptotic behavior that allows them to be applied in practice.

A function has been written in R to solve the hypothesis tests of comparison of the two average kappa coefficients applying the EM and SEM algorithms. This function allows the researcher to solve the problem in a simple and fast way, providing all the necessary results to carry out a study. This function is available as supplemental material to this manuscript.

Hypothesis tests can also be solved by applying the maximum likelihood method to obtain the estimates of the average kappa coefficients and the delta method to estimate the variances-covariances. For this, the methodology applied in the manuscript of Roldán-Nofuentes and Luna [12] is used. However, the maximum likelihood method cannot be applied when some frequency s_{ij} (or r_{ij}) is equal to zero (since the variances-covariances cannot be estimated). In this situation, the EM and SEM algorithms can be applied. Therefore, this is the advantage of EM-SEM algorithms over the maximum likelihood method.

Another alternative computational method to EM-SEM algorithms is multiple imputation [25–27]. Multiple imputation is a computational method used to solve problems with missing data. Appendix B describes in detail the multiple imputation by chained equations [28] used to solve the hypothesis test for the comparison of the two average kappa coefficients. We have carried out simulation experiments to study the asymptotic behaviour of the hypothesis tests 4 and 5 by applying multiple imputation. The experiments have been designed similarly to those performed in Section 4. The experiments have also been carried out with R and the "mice" library [29] has been used. For multiple imputation, 10 complete data sets have been generated and 100 cycles have been performed. Table 9 shows the results obtained for some of the scenarios given in Tables 3–6. The type I error of the hypothesis test solved by applying multiple imputation is slightly less than that of the hypothesis test solved by applying the EM-SEM algorithms, both having very similar asymptotic behavior. Regarding the power of the test, this is also a little lower than the power of the test solved by applying the EM-SEM algorithms, also having a very similar asymptotic behavior. In very general terms, although the differences between multiple imputation and EM-SEM algorithms are not very important, the hypothesis tests solved with multiple imputation are slightly more conservative (and also slightly less powerful) than the hypothesis tests solved with EM-SEM algorithms. Multiple imputation has the disadvantage that it cannot be applied when some frequency s_{ij} (or r_{ij}) is equal to zero, since logistic regression models cannot be applied to impute missing data.

Table 9. Type I errors (in%) and powers (in%) applying multiple imputation.

Type I error when $L' > L$ $(0 \le c < 0.5)$				
$\kappa_{11} = \kappa_{21} = 0.2$ $\kappa_1(0) = 0.16$ $\kappa_1(1) = 0.67$ $\kappa_2(0) = 0.16$ $\kappa_2(1) = 0.67$ $p = 10\%$				
	$\lambda_{11} = 0.50, \lambda_{10} = \lambda_{01} = 0.30, \lambda_{00} = 0.05$		$\lambda_{11} = 0.95, \lambda_{10} = \lambda_{01} = 0.60, \lambda_{00} = 0.25$	
n	$\alpha_1 = 1.14$ $\alpha_0 = 2.37$	$\alpha_1 = 1.24$ $\alpha_0 = 3.47$	$\alpha_1 = 1.14$ $\alpha_0 = 2.37$	$\alpha_1 = 1.24$ $\alpha_0 = 3.47$
50	0	0	0	0
100	0	0	0.10	0
200	0.05	0	0.55	0
500	0.95	0.05	2.15	0.05
1000	1.20	0.15	3.05	0.90
2000	2.95	0.40	3.80	1.85
Power when $L' > L$ $(0 \le c < 0.5)$				
$\kappa_{11} = 0.4$ $\kappa_{21} = 0.2$ $\kappa_1(0) = 0.34$ $\kappa_1(1) = 0.78$ $\kappa_2(0) = 0.16$ $\kappa_2(1) = 0.67$ $p = 10\%$				
	$\lambda_{11} = 0.50, \lambda_{10} = \lambda_{01} = 0.30, \lambda_{00} = 0.05$		$\lambda_{11} = 0.95, \lambda_{10} = \lambda_{01} = 0.60, \lambda_{00} = 0.25$	
n	$\alpha_1 = 1.11$ $\alpha_0 = 2.37$	$\alpha_1 = 1.19$ $\alpha_0 = 3.47$	$\alpha_1 = 1.11$ $\alpha_0 = 2.37$	$\alpha_1 = 1.19$ $\alpha_0 = 3.47$
50	0.10	0.01	0.55	0.85
100	2.70	3.05	14.90	17.35
200	24.85	34.25	61.80	84.05
500	80.05	95.95	97.75	100
1000	98.20	99.15	100	100
2000	100	100	100	100
Type I error when $L > L'$ $(0.5 < c \le 1)$				
$\kappa_{12} = \kappa_{22} = 0.4$ $\kappa_1(0) = 0.16$ $\kappa_1(1) = 0.67$ $\kappa_2(0) = 0.16$ $\kappa_2(1) = 0.67$ $p = 10\%$				
	$\lambda_{11} = 0.50, \lambda_{10} = \lambda_{01} = 0.30, \lambda_{00} = 0.05$		$\lambda_{11} = 0.95, \lambda_{10} = \lambda_{01} = 0.60, \lambda_{00} = 0.25$	
n	$\alpha_1 = 1.14$ $\alpha_0 = 2.37$	$\alpha_1 = 1.26$ $\alpha_0 = 3.47$	$\alpha_1 = 1.14$ $\alpha_0 = 2.37$	$\alpha_1 = 1.26$ $\alpha_0 = 3.47$
50	0	0	0	0
100	0	0	0.10	0
200	0.15	0.05	0.55	0.25
500	0.65	0.45	2.85	1.95
1000	1.45	1.05	3.25	2.05
2000	3.45	2.40	3.90	3.10
Type I error when $L > L'$ $(0.5 < c \le 1)$				
$\kappa_{12} = 0.6$ $\kappa_{22} = 0.4$ $\kappa_1(0) = 0.46$ $\kappa_1(1) = 0.66$ $\kappa_2(0) = 0.27$ $\kappa_2(1) = 0.47$ $p = 50\%$				
	$\lambda_{11} = 0.50, \lambda_{10} = \lambda_{01} = 0.30, \lambda_{00} = 0.05$		$\lambda_{11} = 0.95, \lambda_{10} = \lambda_{01} = 0.60, \lambda_{00} = 0.25$	
n	$\alpha_1 = 1.22$ $\alpha_0 = 2.10$	$\alpha_1 = 1.39$ $\alpha_0 = 2.99$	$\alpha_1 = 1.22$ $\alpha_0 = 2.10$	$\alpha_1 = 1.39$ $\alpha_0 = 2.99$
50	0.05	0.01	1.25	2.05
100	4.40	3.35	14.35	26.85
200	23.80	26.80	47.25	75.80
500	62.75	84.95	88.15	99.10

1000	86.85	94.45	98.35	100
2000	99.70	100	100	100

Future research should also focus on comparing the two average kappa coefficients through confidence intervals and on extending the hypothesis tests to the situation in which the average kappa coefficients of more than two BDTs are compared. In the first case, multiple imputation can be applied together with confidence intervals for the difference or ratio of two average kappa coefficients, adapting the intervals studied by Roldán-Nofuentes and Regad [30,31]. For the second case, an adaptation of the method used by Regad and Roldán-Nofuentes [32] and Roldán-Nofuentes and Regad [33] can be a solution to the problem.

Supplementary Materials: The following are available online at www.mdpi.com/2227-7390/9/21/2834/s1. The "cakcmd" function is a function written in R to compare the average kappa coefficients of two binary diagnostic tests in the presence of missing data.

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

Funding: This research received no external funding.

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 author declares no conflict of interest.

Appendix A

For simplicity, only $\kappa_1(0)$ is considered. The ML estimator of $\kappa_1(0)$ in the presence of missing data is [12]

$$\hat{\kappa}_1(0) = \frac{\sum_{j=0}^{1} \frac{n_{1j} s_{1j}}{s_{1j} + r_{1j}} - \frac{n_{10} + n_{11}}{n} \sum_{i,j=0}^{1} \frac{n_{ij} s_{ij}}{s_{ij} + r_{ij}}}{\sum_{i,j=0}^{1} \frac{n_{ij} r_{ij}}{s_{ij} + r_{ij}}},$$

From Equation (9) it is obtained that

$$\hat{\phi}_{ij} = \frac{s_{ij} + y_{ij}}{n} \text{ and } \hat{\varphi}_{ij} = \frac{r_{ij} + u_{ij} - y_{ij}}{n},$$

In order to demonstrate that the EM algorithm converges to the ML estimates, we are going to follow the same steps as Little and Rubin [27]. With the *EM* algorithm, the estimator of $\kappa_1(0)$ is

$$\hat{\kappa}_1^{(m+1)}(0) = \frac{\sum_{j=0}^{1}\left(s_{1j} + y_{1j}^{(m+1)}\right) \times \sum_{j=0}^{1}\left(r_{0j} + u_{0j} - y_{0j}^{(m+1)}\right) - \sum_{j=0}^{1}\left(s_{0j} + y_{0j}^{(m+1)}\right) \times \sum_{j=0}^{1}\left(r_{1j} + u_{1j} - y_{1j}^{(m+1)}\right)}{\left(r + u - y^{(m+1)}\right)\left(n_{10} + n_{11}\right)}.$$

Then, taking $\hat{\phi}_{ij}^{(m)} = \hat{\phi}_{ij}^{(m+1)} = \hat{\phi}_{ij} = \frac{s_{1j} + y_{1j}}{n}$, $\hat{\varphi}_{ij}^{(m)} = \hat{\varphi}_{ij}^{(m+1)} = \hat{\varphi}_{ij} = \frac{r_{ij} + c_{ij} - y_{ij}}{n}$ and $y_{ij}^{(m)} = y_{ij}^{(m+1)} = y_{ij} = u_{ij} \frac{\hat{\phi}_{ij}}{\hat{\phi}_{ij} + \hat{\varphi}_{ij}}$, it is obtained that $y_{ij}^{(m)} = y_{ij}^{(m+1)} = y_{ij} = \frac{s_{ij} u_{ij}}{s_{ij} + r_{ij}}$, with $i,j = 0,1$.

Substituting in the expression for $\hat{\kappa}_1^{(m+1)}(0)$ and performing algebraic operations, it is obtained that

$$\hat{\kappa}_1^{(m+1)}(0)=\frac{\sum_{j=0}^{1}\frac{n_{1j}s_{1j}}{s_{1j}+r_{1j}}-\frac{n_{10}+n_{11}}{n}\sum_{i,j=0}^{1}\frac{n_{ij}s_{ij}}{s_{ij}+r_{ij}}}{\sum_{i,j=0}^{1}\frac{n_{ij}r_{ij}}{s_{ij}+r_{ij}}}=\hat{\kappa}_1(0)$$

Therefore, $\hat{\kappa}_1^{(m+1)}(0)$ converges to $\hat{\kappa}_1(0)$. The convergence of the other estimates obtained by applying the EM algorithm is demonstrated in a similar way.

Appendix B

Multiple imputation [25–27] is another computational method used to solve problems with missing data. Multiple imputation consists of constructing $K \geq 2$ sets of complete data obtained replacing the missing data with the sets imputed independently. Parameters are estimated from each of the *K* complete datasets, obtaining *K* estimates of each parameter, and then the *K* estimates of each parameter are combined in an appropriate way, obtaining a global estimate of each parameter and its variance. From these global estimates it is possible to obtain confidence intervals for each parameter and also to solve hypothesis tests.

In this manuscript, the multiple imputation by chained equations has been used for the imputation of the missing data. Multiple imputation by chained equations (MICE), also known as fully conditional specification or sequential regression multivariate imputation, requires us to assume that the missing data are MAR. The MICE method is described in detail in the work by White et al. [28]. In the problem studied here there are three random binary variables: T_1, T_2 and D. Variables T_1 and T_2 have no missing data, because the two BDTs have been applied to all the individuals in the sample. Nevertheless, variable *D* is missing for a subset of individuals, since the GS has not been applied to all the individuals in the sample. First, all missing values are filled in at random and then variable *D* is regressed on the variables T_1 and T_2 through a logistic regression [28]. Next, the missing values in variable *D* (disease status for individuals non-verified with the GS) are replaced by simulated draws from the posterior predictive distribution of variable *D* [28]. This process, called a cycle, is repeated a determined number of times to stabilize the results [28]. Finally a set of imputed data is obtained. Therefore, from the 3×4 table (Table 2) $K \geq 2$ 2×4 tables are imputed, and from each one of these 2×4 tables the estimates of the average kappa coefficients and their variances-covariances are obtained. Frequencies of the *k*th 2×4 table are $a_{ij}^{(k)}$ for individuals with the disease and $b_{ij}^{(k)}$ for individuals without the disease, with $i,j=0,1$ and $k=1,\ldots,K$. Therefore, in each imputed 2×4 table, the average kappa coefficients are estimated by applying the equations deduced by Roldán-Nofuentes and Olvera-Porcel [6], i.e.,:

$$\hat{\kappa}_{11}^{(k)}=\frac{2\left\{\left(a_{10}^{(k)}+a_{11}^{(k)}\right)\left(b_{00}^{(k)}+b_{01}^{(k)}\right)-\left(a_{00}^{(k)}+a_{01}^{(k)}\right)\left(b_{10}^{(k)}+b_{11}^{(k)}\right)\right\}}{n\sum_{j=0}^{1}\left(a_{0j}^{(k)}-b_{1j}^{(k)}\right)}\times\ln\left\{\frac{1}{2}\left(\frac{a^{(k)}\sum_{j=0}^{1}n_{0j}^{(k)}}{b^{(k)}\sum_{j=0}^{1}n_{1j}^{(k)}}+1\right)\right\},\text{ if }\hat{p}^{(k)}\neq\hat{Q}_1^{(k)},$$

$$\hat{\kappa}_{11}^{(k)}=\hat{S}e_1^{(k)}+\hat{S}p_1^{(k)}-1,\text{ if }\hat{p}^{(k)}=\hat{Q}_1^{(k)},$$

$$\hat{\kappa}_{21}^{(k)} = \frac{2\left\{\left(a_{01}^{(k)} + a_{11}^{(k)}\right)\left(b_{00}^{(k)} + b_{10}^{(k)}\right) - \left(a_{00}^{(k)} + a_{10}^{(k)}\right)\left(b_{01}^{(k)} + b_{11}^{(k)}\right)\right\}}{n\sum_{i=0}^{1}\left(a_{i0}^{(k)} - b_{i1}^{(k)}\right)} \times \ln\left\{\frac{1}{2}\left(\frac{a^{(k)}\sum_{i=0}^{1} n_{i0}^{(k)}}{b^{(k)}\sum_{i=0}^{1} n_{i1}^{(k)}} + 1\right)\right\}, \text{ if } \hat{p}^{(k)} \neq \hat{Q}_2^{(k)},$$

$$\hat{\kappa}_{12}^{(k)} = \hat{S}e_2^{(k)} + \hat{S}p_2^{(k)} - 1, \text{ if } \hat{p}^{(k)} = \hat{Q}_2^{(k)}$$

if $L' > L$ $(0 \le c < 0.5)$, and

$$\hat{\kappa}_{12}^{(k)} = \frac{2\left\{\left(a_{10}^{(k)} + a_{11}^{(k)}\right)\left(b_{00}^{(k)} + b_{01}^{(k)}\right) - \left(a_{00}^{(k)} + a_{01}^{(k)}\right)\left(b_{10}^{(k)} + b_{11}^{(k)}\right)\right\}}{n\sum_{j=0}^{1}\left(a_{0j}^{(k)} - b_{1j}^{(k)}\right)} \times \ln\left\{2\frac{a^{(k)}\sum_{j=0}^{1} n_{0j}^{(k)}}{a^{(k)}\sum_{j=0}^{1} n_{0j}^{(k)} + b^{(k)}\sum_{j=0}^{1} n_{1j}^{(k)}}\right\}, \text{ if } \hat{p}^{(k)} \neq \hat{Q}_1^{(k)},$$

$$\hat{\kappa}_{12}^{(k)} = \hat{S}e_1^{(k)} + \hat{S}p_1^{(k)} - 1, \text{ if } \hat{p}^{(k)} = \hat{Q}_1^{(k)},$$

$$\hat{\kappa}_{22}^{(k)} = \frac{2\left\{\left(a_{01}^{(k)} + a_{11}^{(k)}\right)\left(b_{00}^{(k)} + b_{10}^{(k)}\right) - \left(a_{00}^{(k)} + a_{10}^{(k)}\right)\left(b_{01}^{(k)} + b_{11}^{(k)}\right)\right\}}{n\sum_{i=0}^{1}\left(a_{i0}^{(k)} - b_{i1}^{(k)}\right)} \times \ln\left\{2\frac{a^{(k)}\sum_{i=0}^{1} n_{i0}^{(k)}}{a^{(k)}\sum_{i=0}^{1} n_{i0}^{(k)} + b^{(k)}\sum_{i=0}^{1} n_{i1}^{(k)}}\right\}, \text{ if } \hat{p}^{(k)} \neq \hat{Q}_2^{(k)},$$

$$\hat{\kappa}_{22}^{(k)} = \hat{S}e_2^{(k)} + \hat{S}p_2^{(k)} - 1, \text{ if } \hat{p}^{(k)} = \hat{Q}_2^{(k)},$$

if $L' < L$ $(0.5 < c \le 1)$, and where

$$a^{(k)} = \sum_{i,j=0}^{1} a_{ij}^{(k)}, \quad b^{(k)} = \sum_{i,j=0}^{1} b_{ij}^{(k)}, \quad n_{ij}^{(k)} = a_{ij}^{(k)} + b_{ij}^{(k)}, \quad \hat{p}^{(k)} = \frac{a^{(k)}}{n},$$

$$\hat{S}e_1^{(k)} = \frac{a_{10}^{(k)} + a_{11}^{(k)}}{a^{(k)}}, \quad \hat{S}p_1^{(k)} = \frac{b_{01}^{(k)} + b_{00}^{(k)}}{b^{(k)}}, \quad \hat{S}e_2^{(k)} = \frac{a_{01}^{(k)} + a_{11}^{(k)}}{a^{(k)}}, \quad \hat{S}p_2^{(k)} = \frac{b_{10}^{(k)} + b_{00}^{(k)}}{b^{(k)}}$$

and

$$\hat{Q}_h^{(k)} = \hat{S}e_h^{(k)} + \hat{S}p_h^{(k)} - 1.$$

The overall estimates of the average kappa coefficients and their variances-covariances are then calculated using Rubin's rules [25]. Overall estimates of the average kappa coefficients are

$$\bar{\kappa}_{11} = \frac{1}{K}\sum_{k=1}^{K}\hat{\kappa}_{11}^{(k)} \text{ and } \bar{\kappa}_{21} = \frac{1}{K}\sum_{k=1}^{K}\hat{\kappa}_{21}^{(k)}$$

and the overall estimates of the difference is

$$\bar{\kappa}_1 = \bar{\kappa}_{11} - \bar{\kappa}_{21} = \frac{1}{K}\sum_{k=1}^{K}\hat{\kappa}_1^{(k)} \text{ and } \bar{\kappa}_2 = \bar{\kappa}_{12} - \bar{\kappa}_{22} = \frac{1}{K}\sum_{k=1}^{K}\hat{\kappa}_2^{(k)}$$

where $\hat{\kappa}_1^{(k)} = \hat{\kappa}_{11}^{(k)} - \hat{\kappa}_{21}^{(k)}$ and $\hat{\kappa}_2^{(k)} = \hat{\kappa}_{12}^{(k)} - \hat{\kappa}_{22}^{(k)}$. The variance of $\bar{\kappa}_1$ is $\hat{\mathrm{V}}\mathrm{ar}\left(\bar{\kappa}_1\right) = \bar{\mathrm{V}}\mathrm{ar}\left(\hat{\kappa}_1\right) + \frac{1}{K+1}B_1$, where $\bar{\mathrm{V}}\mathrm{ar}\left(\hat{\kappa}_1\right) = \frac{1}{K}\sum_{k=1}^{K}\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\kappa}_{11}^{(k)} - \hat{\kappa}_{21}^{(k)}\right)$ is the within imputation variance (complete equation of this variance can be seen in the article by Roldán-Nofuentes and Olvera-Porcel [6]) and $B_1 = \frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\kappa}_1^{(k)} - \bar{\kappa}_1\right)^2$ is the between imputation variance (the variance of the complete data point estimates) [25]. Similarly,

$\hat{\text{V}}\text{ar}\left(\bar{\kappa}_2\right) = \bar{\text{V}}\text{ar}\left(\hat{\kappa}_2\right) + \frac{1}{K+1}B_2$, where $\bar{\text{V}}\text{ar}\left(\hat{\kappa}_2\right) = \frac{1}{K}\sum_{k=1}^{K}\hat{\text{V}}\text{ar}\left(\hat{\kappa}_{12}^{(k)} - \hat{\kappa}_{22}^{(k)}\right)$ and $B_2 = \frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\kappa}_2^{(k)} - \bar{\kappa}_2\right)^2$. Finally, the test statistic for the hypothesis test

$$H_0\!:\kappa_{1i} = \kappa_{2i} \text{ vs } \text{H}_1\!:\kappa_{1i} \neq \kappa_{2i}\,,\ \ i = 1,2\,,$$

is

$$t_i = \frac{\bar{\kappa}_i}{\sqrt{\hat{\text{V}}\text{ar}\left(\bar{\kappa}_i\right)}}\,,\ \ i = 1,2\,,$$

whose distribution is [25] a Student *t*-distribution with $v_i = \left(K-1\right)\left(1 + \frac{K}{K+1}\frac{\hat{\text{V}}\text{ar}\left(\bar{\kappa}_i\right)}{B_i}\right)$ degrees of freedom. With respect to the confidence intervals for the difference of the two average kappa coefficients, their expressions are

$$\kappa_{1i} - \kappa_{2i} \in \bar{\kappa}_i \pm t_{v_i,1-\alpha/2}\sqrt{\hat{\text{V}}\text{ar}\left(\bar{\kappa}_i\right)}\,,\ \ i = 1,2\,,$$

where $t_{v_i,1-\alpha/2}$ is the $100\times\left(1-\alpha/2\right)$th percentile of the Student *t*-distribution with v_i degrees of freedom.

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.
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.
4. Bloch, D.A. Comparing two diagnostic tests against the same "gold standard" in the same sample. *Biometrics* **1997**, *53*, 73–85.
5. Roldán-Nofuentes, J.A.; Luna del Castillo, J.D. Comparison of weighted kappa coefficients of multiple binary diagnostic tests done on the same subjects. *Stat. Med.* **2010**, *29*, 2149–2165.
6. Roldán-Nofuentes, J.A.; Olvera-Porcel, C. Comparison of the average kappa coefficients of binary diagnostic tests done on the same subjects. *Revstat Stat. J.* **2018**, *16*, 405–428.
7. Begg, C.B.; Greenes, R.A. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* **1983**, *39*, 207–215.
8. Zhou, X.H. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Comm. Statist. Theory Methods* **1993**, *22*, 3177–3198.
9. Zhou, X.H. Comparing accuracies of two screening tests in a two-phase study for dementia. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1998**, *47*, 135–147.
10. Harel, O.; Zhou, X.H. Multiple imputation for the comparison of two screening tests in two-phase Alzheimer studies. *Stat. Med.* **2007**, *26*, 2370–2388.
11. Roldán-Nofuentes, J.A.; Luna del Castillo, J.D. EM algorithm for comparing two binary diagnostic tests when not all the patients are verified. *J. Stat. Comput. Simul.* **2008**, *78*, 19–35.
12. Roldán-Nofuentes, J.A.; Luna del Castillo, J.D. Comparing two binary diagnostic tests in the presence of verification bias. *Comput. Stat. Data Anal.* **2006**, *50*, 1551–1564.
13. 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.
14. Youden, W.J. Index for rating diagnostic tests. *Cancer* **1950**, *3*, 32–35.
15. Landis, R.; Koch, G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174.
16. 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.
17. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM Algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* **1977**, *39*, 1–38.
18. Meng, X.; Rubin, D.B. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *J. Am. Stat. Assoc.* **1991**, *86*, 899–909.
19. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *4*, 73–89.

20. Berry, G.; Smith, C.; Macaskill, P.; Irwig, L. Analytic methods for comparing two dichotomous screening or diagnostic tests applied to two populations of differing disease prevalence when individuals negative on both tests are unverified. *Stat. Med.* **2002**, *21*, 853–862.
21. Tsai, T.-R.; Lio, Y.; Ting, W.C. EM algorithm for mixture distributions model with type-I hybrid censoring scheme. *Mathematics* **2021**, *9*, 2483.
22. Gallardo, D.I.; de Castro, M.; Gómez, H.W. An alternative promotion time cure model with overdispersed number of competing causes: An application to melanoma data. *Mathematics* **2021**, *9*, 1815.
23. R Core Team. A Language and Environment for Statistical Computing. Vienna, Austria. 2016. Available online: https://www.R-project.org/ (accessed on 1 October 2021).
24. Hall, K.S.; Ogunniyi, A.O.; Hendrie, H.C.; Osuntokun, B.O.; Hui, S.L.; Musick, B.; Rodenberg, C.S.; Unverzagt, F.W.; Guerje, O.; Baiyewu, O. A cross-cultural community based study of dementias: Methods and performance of survey instrument. *Int. J. Methods Psychiatr. Res*. **1996**, *6*, 129–142.
25. Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*; Wiley: New York, NY, USA, 1987.
26. Schafer, J.L. *Analysis of Incomplete Multivariate Data*; Chapman and Hall: New York, NY, USA, 1997.
27. Little, R.J.A.; Rubin, D.B. *Statistical analysis with missing data,* 2nd ed.; Wiley: New Jersey, NJ, USA, 2002.
28. White, I.R.; Royston, P.; Wood, A.M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med*. **2011**, *30*, 377–399.
29. van Buuren, S.; Groothuis-Oudshoorn, K. Mice: Multivariate imputation by chained equations in R. *J. Stat. Softw*. **2011**, *45*, 3.
30. Roldán-Nofuentes, J.A.; Regad, S.B. Confidence intervals and sample size to compare the predictive values of two diagnostic tests. *Mathematics* **2021**, *9*, 1462.
31. Roldán-Nofuentes, J.A.; Regad, S.B. Asymptotic confidence intervals for the difference and the ratio of the weighted kappa coefficients of two diagnostic tests subject to a paired design. *Revstat Stat. J.* **2021**, in press.
32. Regad, S.B.; Roldán-Nofuentes, J.A. Global hypothesis test to compare the predictive values of diagnostic tests subject to a case-control design. *Mathematics* **2021**, *9*, 658.
33. Roldán-Nofuentes, J.A.; Regad, S.B. Recommended methods to compare the accuracy of two binary diagnostic tests subject to a paired design. *J. Stat. Comput. Simul.* **2019**, *89*, 2621–2644.