University of Granada

Doctoral Program in Mathematical and Applied Statistics

Department of Statistics and Operations Research

Doctoral Thesis

Estimation and hypothesis test for parameters of binary diagnostic tests

Saad Bouh Sidaty Regad

Thesis supervised by

José Antonio Roldán Nofuentes

Granada, September, 2020

Editor: Universidad de Granada. Tesis Doctorales
Autor: Saad Bouh Sidaty Regad
ISBN: 978-84-1306-722-3
URI: http://hdl.handle.net/10481/65375

This Doctoral Thesis is a compendium of the following manuscripts:

- Roldán-Nofuentes, J.A., Sidaty-Regad S.B. (2019). Recommended methods to compare the accuracy of two binary diagnostic tests subject to a paired design. *Journal of Statistical Computation and Simulation*, 89, 2621-2644. DOI: 10.1080/00949655.2019.1628234

- Roldán-Nofuentes, J.A., Sidaty-Regad S.B. (2020). Comparison of the likelihood ratios of two diagnostic tests subject to a paired design: confidence intervals and sample size. *REVSTAT-Statistical Journal*. Accepted, in press.

- Roldán-Nofuentes, J.A., Sidaty-Regad S.B. (2020). Asymptotic confidence intervals for the difference and the ratio of the weighted kappa coefficients of two diagnostic tests subject to a paired design. *REVSTAT-Statistical Journal*. Accepted, in press.

- Roldán-Nofuentes, J.A., Sidaty-Regad S.B. (2020). *EM* and *SEM* algorithms to compare the weighted kappa coefficients of two diagnostic tests in the presence of partial verification and discrete covariates. *Journal of Statistical Computation and Simulation.* Accepted, in press. DOI: 10.1080/00949655.2020.1804903.

The investigations carried out in this Doctoral Thesis have been supported by the Spanish Ministry of Economy, Grant Number MTM2016-76938-P.

ACKNOWLEDGEMENTS

I am profoundly grateful to my dear Professor José Antonio Roldán Nofuentes, Professor in the Department of Statistics at the University of Granada, for his generous efforts, his unending care, unlimited support, good guidance and patience during all stages of thesis preparation. Without this pledge, this work would not have come to light.

I want to go further to say that the idea of the thesis itself stems from my admiration for his valuable and systematic lectures that I followed during the Master's period at the University of Nouakchott when he was a visiting professor.

I also thank my great professor Rafael Rodríguez-Contreras Pelayo and the other distinguished professors whom I have benefited from their knowledge.

As well as, I want to thank the distinguished jury for accepting the evaluation of this humble work.

Thanks also to the Spanish cooperation for its great interest and support for scientific research in my country and for my great university, University of Granada, which I have the honor to affiliate with.

Great think to the President of the University of Nouakchott Alaasriya Pr. Ahmedou Haouba for his constant support, encouragement and motivation.

I cannot finish without thanking my family for their support, their standing, and their constant encouragement.

INDEX

SUMMARY

The continual advances in the diagnosis of diseases have led Statistics to develop new methods to solve the problems that have been posed in this field. This Doctoral Thesis seeks to make a contribution to research into new statistical methods in the field of the statistical methods in diagnostic medicine. This Doctoral Thesis is focused on the study of the estimation and the comparison of parameters of two binary diagnostic tests subject to paired design and subject to partial verification of the disease. In the first situation, the problem leads to the analysis of 2×4 table and in the second situation it leads to the analysis of a 3×4 table. As tangible results of this doctoral thesis, one article has been published, two more are currently accepted for publication and a fourth article is being reviewed. These articles are included in their entirety in Appendices I - IV.

Traditionally, the comparison of the sensitivities and the specificities of two binary diagnostic tests subject to a paired design was made comparing the two sensitivities and the two specificities independently, applying a comparison test with two binomial proportions paired to an α error. An alternative method consists of comparing the two sensitivities and the two specificities simultaneously, solving a global hypothesis test to an α error. This PhD has analysed in depth the study of the global test, proposing two test statistics, one obtained applying the Rao score test and the other with the Wald test. Moreover, based on the results of the simulation experiments carried out, some rules of application are given for the different methods studied.

The likelihood ratios of a binary diagnostic test are very widely used parameters to compare the effectiveness of two binary diagnostic tests, as they are technically

equivalent to a relative risk. This Doctoral Thesis proposes new confidence intervals to compare the positive (negative) likelihood ratios of two binary diagnostic tests subject to a paired design, analysing the problem both from a frequentist perspective and a Bayesian one. The simulation experiments carried out to compare the asymptotic behaviour of the confidence intervals have allowed us to give some general rules of application for the intervals studied. Furthermore, a method is proposed to calculate the sample size to compare the positive (negative) likelihood ratios of two binary diagnostic tests through a confidence interval.

When considering the losses of a misclassification with a diagnostic test, the parameter of effectiveness of the diagnostic test is the weighted kappa coefficient. This parameter depends on the sensitivity and specificity of the diagnostic tests, on the prevalence of the disease and on the relative importance between a false positive and a false negative. This Doctoral Thesis has related the comparison of the two weighted kappa coefficients with the relative true (false) positive fraction between the two binary diagnostic tests, and has studied the comparison of the weighted kappa coefficients of two diagnostic tests through confidence intervals for the difference and for the ratio between the two weighted kappa coefficients, analysing the problem from frequentist and Bayesian perspectives. Simulation experiments have been carried out to study the asymptotic behaviour of the confidence intervals proposed, and some general rules of application have been given. Moreover, a method has been proposed to calculate the simple size to compare the two weighted kappa coefficients through a confidence interval.

When comparing the effectiveness of two binary diagnostic tests in the presence of partial disease verification, the selection of an individual to verify his or her disease status may depend on discrete covariates which are related to the disease. In this situation, a hypothesis test has been studied to compare the weighted kappa coefficients of both diagnostic tests. This problem has been solved applying two computational methods: the *EM* algorithm and the *SEM* algorithm. Simulation experiments have been carried out to study the behaviour of the test statistic proposed.

Granada, September, 2020

CHAPTER 1

INTRODUCTION

The application of a diagnostic test for the diagnosis of a determined disease is a fundamental stage in medical practice, prior to the phases of treatment and prognosis. The diagnosis of the disease is not only important for the doctor or the clinician, since it conditions the treatment and the prognosis of the disease, but also for the individual concerned, as it eliminates the level of uncertainty that the individual has about his or her disease status. A diagnostic test is a medical test that is applied to an individual in order to determine the presence or the absence of a certain disease. There are different types of diagnostic tests:

a). Binary diagnostic tests are those which lead to two results: positive (indicating the disease presence) or negative (indicating the absence of the disease). Examples would include a mammography for breast cancer or an echocardiography for the diagnosis of coronary disease.

b). Quantitative diagnostic tests are those which lead to numerical values e.g. the concentration of *PCR* in cerebrospinal fluid for the diagnosis of meningitis.

c). Ordinal diagnostic tests are those which lead to different values with a hierarchical structure e.g. the classification of the disease presence in ‘definitely yes’, ‘probably yes’, ‘probably no’, and ‘definitely no’.

The most common diagnostic tests are binary diagnostic tests and these are the tests which are studied in this doctoral thesis. The application of a diagnostic test has various purposes (McNeil and Adelstein, 1976; Sox et al., 1989; Pepe, 2003; Zhou et al, 2011):

a). To provide reliable information about the disease status of an individual (diseased or non-diseased).

b). To intervene in the planning of the treatment of an individual.

c). To investigate the mechanism and the nature of the disease.

Moreover, the result of a diagnostic test depends on several factors:

a). On the intrinsic accuracy of the test itself to distinguish between diseased and non-diseased individuals (discriminatory accuracy).

b). On external factors (e.g. taking medication, consumption of alcohol, etc.).

c). On the characteristics of each individual (e.g. the sex or the abnormal physiological conditions of an individual may interfere in the measurement of the test, etc.).

The application of a diagnostic test may lead to mistakes, and therefore its accuracy is measured in terms of probabilities of functions of probabilities. In order to obtain unbiased estimators of those probabilities or of their functions it is necessary to assess the diagnostic test in relation to a gold standard. A gold standard (*GS*) is a medical test that objectively determines whether or not an individual has a disease. A biopsy for the diagnosis of breast cancer and a coronary angiography for the diagnosis of coronary disease are examples of *GS*.

This doctoral thesis studies binary diagnostic tests (*BDTs*), the estimation of their parameters and the comparison of parameters of two *BDTs* subject to different types of sampling.

Let us consider a disease that may or may not be present among the individuals in a population. Let D be a 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 it. The probability of an individual in the population chosen at random having the disease is called disease prevalence (p), i.e. $p=P(D=1)$. Let us consider a *BDT*

whose effectiveness is assessed in relation to a *GS*. Let T be the random variable that models the result of the *BDT* in such a way that $T=1$ when the result of the diagnostic test is positive and $T=0$ when it is negative. We then study different measurements of the effectiveness of a *BDT* whose inference has been studied in this doctoral thesis, as well as two types of situation in which it is possible to compare the effectiveness of two *BDTs*: complete disease verification and partial disease verification.

1.1. Parameters of a *BDT*

The parameters which have been subject to study in this doctoral thesis were: sensitivity and specificity, likelihood ratios and weighted kappa coefficient.

1.1.1 Sensitivity and specificity

The sensitivity (Se) is the probability of the result of the *BDT* being positive when the individual has the disease, i.e.

$$Se = P(T=1 \mid D=1).$$

The specificity (Sp) is the probability of the result of the *BDT* being negative when the individual does not have the disease, i.e.

$$Sp = P(T=0 \mid D=0).$$

The sensitivity and the specificity represent the measurements of the accuracy of a *BDT*, as they only depend on the intrinsic ability of the test to distinguish diseased and non-diseased individuals i.e. they depend on the physical, chemical or biological bases with which the *BDT* has been developed. A *BDT* with a high sensitivity is useful to rule out the presence of the disease and a *BDT* with a high specificity is useful to confirm the disease. We must demand of a BDT that its Youden Index (Youden, 1950), defined as

$$Y = Se + Sp - 1,$$

be greater than zero $(Y>0)$. The Youden Index takes values between -1 and 1, and verifies the following properties:

a). If the sensitivity and the specificity are complementary $\left(Se = 1 - Sp\right)$ then $Y = 0$ and the *BDT* is non-informative. In this situation, the *BDT* is not related to the disease and the diagnosis of the disease can be made by tossing a coin whose probability of being either heads or tails is equal to 0.5.

b). If $Y < 0$ then $T = 1$ must be a negative result and $T = 0$ a positive result. Therefore, if $Y < 0$, the results of the *BDT* must be exchanged.

1.1.2. Likelihood ratios

Likelihood ratios (*LRs*) are other parameters which are used to assess the effectiveness of a *BDT*. Each likelihood ratio is a quotient of two probabilities defined as

$$LR = \frac{P\left(T = i \mid D = 1\right)}{P\left(T = i \mid D = 0\right)}, \; i = 0,1.$$

The likelihood ratio represents the quotient between the probability of a postive or a negative result of the *BDT* in diseased individuals and the probability of the same result in non-diseased individuals. When the result of the *BDT* is positive, the *LR*, called the positive *LR*, is the quotient between the sensitivity and one minus the specificity, i.e.

$$LR^{+} = \frac{P\left(T = 1 \mid D = 1\right)}{P\left(T = 1 \mid D = 0\right)} = \frac{Se}{1 - Sp}.$$

When the result of the *BDT* is negative, the *LR*, called the negative *LR*, is the quotient between one minus the sensitivity and the specificity, i.e.

$$LR^{-} = \frac{P\left(T = 0 \mid D = 1\right)}{P\left(T = 0 \mid D = 0\right)} = \frac{1 - Se}{Sp}.$$

The *LRs* vary between 0 and infinity, and have the following properties:

a). If the *BDT* and the *GS* are independent then $LR^{+} = LR^{-} = 1$.

b). If the *BDT* correctly classifies all of the individuals then $LR^{+} = \infty$ and $LR^{-} = 0$.

c). If $LR^{+} > 1$ then a positive result in the *BDT* is more probable for an individual who has the disease than for an individual who does not.

d). If $LR^- < 1$ then a negative result in the *BDT* is more probable for an individual who does not have the disease than for an individual who does.

e). The *LRs* quantify the increase in knowledge of the presence of the disease through the application of the *BDT*. Before applying the test, the odds of an individual having the disease are $\text{pre-test odds} = p/(1-p)$, where p is the disease prevalence. After applying the *BDT*, the odds are $\text{post-test odds} = \dfrac{P(D=1|T=i)}{P(D=0|T=i)}$, $i = 0,1$. The *LRs* relate the pre-test odds and the post-test odds:

$$\text{post test odds } (T=1) = LR^+ \times \text{pre test odds}$$
$$\text{post test odds } (T=0) = LR^- \times \text{pre test odds.}$$

Therefore, the likelihood ratios quantify the change in the odds of the disease obtained by knowledge of the application of the *BDT*.

1.1.3. Weighted kappa coefficient

Let us consider a *BDT* that is assessed in relation to a *GS*. Let L (L') the loss which occurs when for a diseased (non-diseased) individual the *BDT* gives a negative (positive) result. Therefore, the loss L (L') is associated with a false negative (positive). If an individual (with or without the disease) is correctly diagnosed by the *BDT* then $L = L' = 0$. Let $p = P(D=1)$ be the prevalence of the disease and $q = 1-p$. Table 1.1 shows the losses and the probabilities associated with the assessment of a *BDT* in relation to a *GS*, and the probabilities when the *BDT* and the *GS* are independent, i.e. when $P(T=i|D=j) = P(T=i)$. Multiplying each loss in the 2×2 table by its corresponding probability and adding up all the terms, we find

$$p(1-Se)L + q(1-Sp)L',$$

a term that is defined as expected loss. Therefore, the expected loss is the loss that occurs when misclassifying with the *BDT* an individual with or without the disease. Moreover, if the *BDT* and the *GS* are independent, multiplying each loss by its

corresponding probability (subject to the independence between the *BDT* and the GS) and adding up all of the terms we find

$$p\left[\, p\times(1-Se)+q\times Sp\,\right]L+q\left[\, p\times Se+q\times(1-Sp)\,\right]L',$$

a term that is defined as random loss. Therefore, the random loss is the loss that occurs when the *BDT* and the *GS* are independent. The independence between the *BDT* and the *GS* is equivalent to the Youden index of the *BDT* being equal to zero $(Y=0)$ and is also equivalent to the expected loss being equal to the random loss. In terms of expected and random losses, the weighted kappa coefficient of a *BDT* is defined as

$$\kappa=\frac{\text{Random loss}-\text{Expected loss}}{\text{Random loss}}.$$

Substituting in this equation each loss with its expression, the weighted kappa coefficient of a *BDT* is expressed (Kraemer et al, 1990; Kraemer, 1992; Kraemer et al, 2002) as

$$\kappa(c)=\frac{pqY}{p(1-Q)c+qQ(1-c)},$$

where $Y=Se+Sp-1$ is the Youden index, $Q=pSe+q(1-Sp)$ is the probability that the *BDT* result is positive, and $c=L/(L'+L)$ is the weighting index.

The weighted kappa coefficient of a binary test has the following properties:

a). If the classificatory agreement between the *BDT* and the *GS* is perfect $(Se=Sp=1)$ then $\kappa(c)=1$.

b). If the sensitivity and the specificity are complementary $(Se=1-Sp)$ then $\kappa(c)=0$.

c). If the random expected loss is greater than the expected loss then $\kappa(c)>0$,

d). If the expected loss is greater than the random expected loss then $\kappa(c)<0$ and the results of the diagnosis are interchanged, $T=1$ should be a negative result and

$T=0$ should be a positive result; and the analysis should be limited only to the positive values of the weighted kappa coefficient.

e). The weighted kappa coefficient is a function of the weighting index *c* which may be increasing (if $Q > p$), decreasing (if $Q < p$) or it can be a constant function which is equal to the Youden index if $Q = p$.

Table 1.1. Losses and probabilities.

	Losses (Probabilities)		
	$T=1$	$T=0$	Total
$D=1$	0 $(p\times Se)$	L $(p\times(1-Se))$	L (p)
$D=0$	L' $(q\times(1-Sp))$	0 $(q\times Sp)$	L' (q)
Total	L' $(Q=p\times Se+q\times(1-Sp))$	L $(1-Q=p\times(1-Se)+q\times Sp)$	$L+L'$ $(p+q=1)$
Probabilities when the *BDT* and the *GS* are independent			
	$T=1$	$T=0$	Total
$D=1$	$p\times Q$	$p\times(1-Q)$	p
$D=0$	$q\times Q$	$q\times(1-Q)$	q
Total	Q	$1-Q$	1

The weighting index *c* is a measure of the relative importance between the false positives and the false negatives. For example, let us consider the diagnosis of breast cancer using as a diagnostic mammography test. If the mammography test is positive in a woman that does not have cancer (false positive), the woman will be given a biopsy that will give a negative result. The loss L' is determined from the economic costs of the diagnosis and also from the risk, stress, anxiety, etc., caused to the woman. If the mammography test is negative in a woman who has breast cancer (false negative), the woman may be diagnosed at a later stage, but the cancer may spread, and the possibility of the treatment being successful will have diminished. The loss L is determined from these considerations. The losses L and L' are measured in terms of economic costs and also from risks, stress, etc., which is why in practice their values cannot be determined.

Therefore, as loss L cannot be determined, L is substituted by the importance that a false positive has for the clinician; in the same way, as loss L' cannot be determined, then L' is substituted by the importance that a false negative has for the clinician. The value of the weighting index c will depend therefore on the relative importance between a false positive and a false negative. If the clinician has greater concerns about false positives, as it is the situation in which the *BDT* is used as a definitive test prior to a treatment that involves a risk for the individual (e.g., a definitive test prior to a surgical operation), then $0 \le c < 0.5$. If the clinician is more concerned about false negatives, as in a screening test, then $0.5 < c \le 1$. The index c is equal to 0.5 when the clinician considers that the false negatives and the false positives have the same importance, in which case $\kappa(0.5)$ is the Cohen kappa coefficient. Weighting index c quantifies the relative importance between a false positive and a false negative, but it is not a measure that quantifies how much bigger the proportion of false positives is compared to the false negatives. If $c = 0$ then

$$\kappa(0) = \frac{Sp - (1-Q)}{Q} = \frac{p(1-FNF-FPF)}{p(1-FNF)+qFPF},$$

which is the chance-corrected specificity according to the kappa model. If $c = 1$ then

$$\kappa(1) = \frac{Se - Q}{1-Q} = \frac{q(1-FNF-FPF)}{pFNF - q(1-FPF)},$$

which is the chance-corrected sensitivity according to the kappa model. A low (high) value of $\kappa(1)$ will indicate that the value of *FNF* is high (low), and a low (high) value of $\kappa(0)$ will indicate that the value of *FPF* is high (low). The weighted kappa coefficient can be written as

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

which is a weighted average of $\kappa(0)$ and $\kappa(1)$. Therefore, the weighted kappa coefficient is a measure that considers the proportion of false negatives (*FNF*) and the proportion of false positives (*FPF*). Moreover, for a set value of the c index and of the accuracy (*Se* and *Sp*) of the *BDT*, the weighted kappa coefficient strongly depends on

the disease prevalence among the population being studied, and its value increases when the disease prevalence increases. The weighted kappa coefficient is a measure of the beyond-chance agreement between the *BDT* and the *GS*. The properties of the kappa coefficient can be seen in the manuscript of Roldán-Nofuentes and Amro (2018).

The weighted kappa coefficient is a valid parameter to assess and compare the performance of *BDTs* (Kraemer et al, 1990; Kraemer, 1992; Kraemer et al, 2002; Bloch, 1997; Roldán-Nofuentes et al, 2009; Roldán-Nofuentes and Amro, 2018).

1.2. Complete verification and partial verification

The comparison of the effectiveness of two *BDTs* can be made in two types of situations depending on whether or not the disease status of all of the individuals is known or not. When the disease status of all the individuals is known through the application of a *GS*, the situation is called complete disease verification, since all of the individuals have had their disease status verified (present or absent). If the disease status of some individuals in unknown, the situation is called partial disease verification, since for a subset of individuals it is not known if they have the disease or not. Each one of these situations will now be explained.

1.2.1. Complete verification

Complete disease verification corresponds to the situation in which the disease status is known for all of the individuals in the study. In this situation, when comparing two *BDTs* the most frequent type of sample design is the paired design (Pepe, 2003; Zhou et al, 2011). This type of design consists of applying the two *BDTs* and the *GS* to all of the individuals in a random sample sized *n*. Table 1.2 shows the frequencies that are obtained when comparing two *BDTs* in relation to a *GS* subject to a paired design, where T_h is the variable that models the result of the *h*th *BDT*, in such a way that $T_h = 1$ when its result is positive and $T_h = 0$ when it is negative, and *D* models the result of the *GS*, in such a way that $D = 1$ when the individual has the disease and disease and $D = 0$ when this is not the case.

Table 1.2. Frequencies subject to a paired design.

	$T_1=1$		$T_1=0$		
	$T_2=1$	$T_2=0$	$T_2=1$	$T_2=0$	Total
$D=1$	s_{11}	s_{10}	s_{01}	s_{00}	s
$D=0$	r_{11}	r_{10}	r_{01}	r_{00}	r
Total	$s_{11}+r_{11}$	$s_{10}+r_{10}$	$s_{01}+r_{01}$	$s_{00}+r_{00}$	n

1.2.2. Partial verification

In clinical practice, when assessing a single *BDT* it is common for the *GS* not to be applied to all of the individuals in the sample. Therefore, if the *GS* consists of a costly test or it means an important risk for an individual, the *GS* is not applied to that individual and, therefore, his or her disease status is unknown (present or absent). In this situation, the result of the *BDT* is known for all of the individuals in the sample, but the disease status (i.e. the result of the *GS*) is only known for a subset of them (and consequently it is unknown for a subset composed of the rest). This situation is known as partial disease verification (Begg and Greenes, 1983). If the sensitivity and the specificity (and the likelihood ratios and the weighted kappa coefficient) of a *BDT* are estimated excluding those individuals who have not been verified with the *GS*, the estimators obtained are affected by what is known as verification bias (Begg and Greenes, 1983; Roldán-Nofuentes and Luna, 2008; Montero-Alonso and Roldán-Nofuentes, 2019).

The problem of partial disease verification can also appear when comparing the effectiveness of two *BDTs*. In this situation, we obtain the frequencies given in Table 1.3, where the variable V models the verification process, in such a way that $V=1$ when the individual is verified with the *GS* and $V=0$ when the individual is not verified with the *GS*. Assuming that the verification process is missing at random (MAR), there are several different studies that have been carried out to compare parameters of two *BDTs*. Zhou (1998) studied a hypothesis test to compare the sensitivities (specificities) of two *BDTs* applying the method of maximum likelihood. Harel and Zhou (2007) applied multiple imputation to compare the two sensitivities (specificities) through confidence intervals. Roldán-Nofuentes and Luna (2006) studied

a hypothesis test to compare the weighted kappa coefficients of two *BDTs* applying the method of maximum likelihood. Roldán Nofuentes and Luna (2008, 2009) studied hypothesis tests to compare the sensitivities (specificities) of two *BDTs* applying the *EM* and *SEM* algorithms.

Table 1.3. Cross-classification of test results by verification status and disease status.

	$T_1 = 1$		$T_1 = 0$	
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$
$V = 1$				
$D = 1$	s_{11}	s_{10}	s_{01}	s_{00}
$D = 0$	r_{11}	r_{10}	r_{01}	r_{00}
$V = 0$	u_{11}	u_{10}	u_{01}	u_{00}
Total	n_{11}	n_{10}	n_{01}	n_{00}

CHAPTER 2

OBJECTIVES

2.1. Recommended methods to compare the accuracy of two binary diagnostic tests subject to a paired design

Traditionally, the comparison of the accuracy of two *BDTs* subject to paired design consists of solving the two hypothesis tests

$$H_0 : Se_1 = Se_2 \text{ vs } H_1 : Se_1 \neq Se_2$$

and

$$H_0 : Sp_1 = Sp_2 \text{ vs } H_1 : Sp_1 \neq Sp_2 ,$$

each one of them to an α error applying a comparison test with two paired binomial proportions e.g. applying the well-known McNemar test (Zhou, 2013) or another method (Fagerland et al, 2014). Therefore, the classic method consists of comparing the two sensitivities and the two specificities independently, solving each hypothesis test to an α error. An alternative to the classic method consists of contrasting the equality of the two sensitivities and of the two specificities simultaneously (Lachenbruch and Lynch, 1998), i.e. solving the global test

$$H_0 : \left(Se_1 = Se_2 \text{ and } Sp_1 = Sp_2\right) \text{ vs } H_1 : \left(Se_1 \neq Se_2 \text{ and/or } Sp_1 \neq Sp_2\right)$$

to an α error. The objectives to be studied are: 1) to obtain new test statistics to solve the global test; and 2) to compare the asymptotic behaviour of the previous test statistics with others based on the individual hypothesis tests or with others based on the individual hypothesis tests to an α error and applying methods of multiple comparison. The study of these objectives will allow us to determine the optimal asymptotic methods to compare the accuracy of two *BDTs* subject to paired design, giving some general rules of application.

2.2. Comparison of the likelihood ratios of two diagnostic tests subject to a paired design: confidence intervals and sample size

The comparison of the *LRs* of two *BDTs* through hypothesis tests subject to a paired design has been the subject of several studies. Leisenring and Pepe (1998) studied the estimation of the *LRs* of a *BDT* through a regression model, and Pepe (2003) adapted this model to compare the *LRs* of two *BDTs*. Biggerstaff (2000) proposed a graphical method to compare the *LRs* of two (or more) *BDTs*. Nevertheless, this method is non-inferential and can only be applied to the estimators. Roldán-Nofuentes and Luna (2007) studied hypothesis tests to compare the *LRs* individually and also simultaneously. Dolgun et al (2012) extended the method of Leisenring and Pepe (1998) to compare the *LRs* simultaneously. Nevertheless, the comparison of the *LRs* through confidence intervals has not been studied in depth. From the studies by Pepe (2003) and by Roldán-Nofuentes and Luna (2007), confidence intervals are obtained for the ratio of the positive (negative) *LRs*. Therefore, the following objectives are posed: 1) to obtain new confidence intervals for the ratio of the positive (negative) *LRs*; 2) to compare the asymptotic behaviour of the confidence intervals studied; and 3) to study a method to calculate the sample size necessary to compare the *LRs* through a confidence interval.

2.3. Asymptotic confidence intervals for the difference and the ratio of the weighted kappa coefficients of two diagnostic tests subject to a paired design

Bloch (1997) studied the comparison of the weighed kappa coefficients of two *BDTs* subject to a paired design deducing a Wald type statistic. Based on this study, inverting the test statistic we obtain a confidence interval for the difference between the two weighted kappa coefficients. The objectives that are posed are: 1) to study new confidence intervals for the difference and the ratio of the weighted kappa coefficients of two *BDTs*; 2) to compare the asymptotic behaviour of the confidence intervals studied, giving some general rules of application; and 3) to study a method to calculate the sample size necessary to compare the two weighted kappa coefficients through a confidence interval.

2.4. *EM* and *SEM* algorithms to compare the weighted kappa coefficients of two diagnostic tests in the presence of partial verification and discrete covariates

In the presence of partial disease verification, the selection of an individual to verify his or her disease status with the *GS* may also depend on the discrete covariates that are related to the disease. Zhou (1998) studied a hypothesis test to compare the sensitivities (specificities) of two *BDTs* when in the presence of partial disease verification, discrete covariates are observed in all the individuals. The objective that is posed is to study a hypothesis test to compare the global weighted kappa coefficients of two *BDTs* when in the presence of partial disease verification a discrete covariate is observed in all the individuals.

CHAPTER 3

METHODOLOGY

3.1. Recommended methods to compare the accuracy of two binary diagnostic tests subject to a paired design

Traditionally, the comparison of the accuracy of two *BDTs* subject to a paired design consisted of solving the two hypothesis tests

$$H_0 : Se_1 = Se_2 \text{ vs } H_1 : Se_1 \neq Se_2$$

and

$$H_0 : Sp_1 = Sp_2 \text{ vs } H_1 : Sp_1 \neq Sp_2 ,$$

each one of them to an α error applying a comparison test with two paired binomial proportions. Subject to a paired design these hypothesis tests are equivalent to the tests

$$H_0 : p_{10} = p_{01} \text{ vs } H_1 : p_{10} \neq p_{01}$$

and

$$H_0 : q_{01} = q_{10} \text{ vs } H_1 : q_{01} \neq q_{10} ,$$

respectively. Another alternative method is to solve the global test

$$H_0 : \left(Se_1 = Se_2 \text{ and } Sp_1 = Sp_2\right) \text{ vs } H_1 : \left(Se_1 \neq Se_2 \text{ and/or } Sp_1 \neq Sp_2\right)$$

We will now summarize the methods that exist to solve the individual tests and the global test.

3.1.1. Individual tests

The comparison of the sensitivities (specificities) of two *BDTs* subject to a paired design consists of the comparison of two paired binomial proportions. We will now summarize some different methods to solve this problem.

3.1.1.1. Conditional exact test

In hypothesis test $H_0 : p_{10} = p_{01}$ the proportions p_{11} and p_{00} do not appear, and so it is possible to discard these proportions and, consequently, also discard the frequencies s_{11} and s_{00}. Conditioning on the sum of the discordant frequencies, i.e. conditioning on $s_{10} + s_{01}$, it is verified that $p_{10} + p_{01} = 1$, and it is also verified that s_{10} is the product of a binomial distribution of parameters $s_{10} + s_{01}$ and p_{10}, i.e. $Bin\left(s_{10} + s_{01}, p_{10}\right)$. If the null hypothesis is true then $p_{10} = p_{01} = 1/2$, and, therefore, both the hypothesis test $H_0 : q_{01} = q_{10}$ vs $H_1 : p_{10} \neq p_{01}$ is also equivalent to test $H_0 : p_{10} = 1/2$ vs $H_1 : p_{10} \neq 1/2$. Finally, the two-sided exact p-value for the comparison test of the two sensitivities is

$$\text{two-sided exact p-value} = 2 \times \sum_{j=0}^{Min(s_{10}, s_{01})} \binom{s_{10} + s_{01}}{j} \left(\frac{1}{2}\right)^{s_{10}+s_{01}} .$$

If $s_{10} = s_{01}$ then the two-sided exact p-value equals one. In a similar way, the two-sided exact p-value to compare the two specificities is

$$\text{two-sided exact p-value} = 2 \times \sum_{j=0}^{Min(r_{10}, r_{01})} \binom{r_{10} + r_{01}}{j} \left(\frac{1}{2}\right)^{r_{10}+r_{01}} .$$

3.1.1.2. Conditional mid-p test

The conditional mid-p test (Lancaster, 1961) is a modification of the exact conditional test. This method consists of subtracting the probability of the observed outcome s_{10}

from the two-sided exact p-value. Thus, the mid-p values to compare the two sensitivities and the two specificities are

$$\text{mid-p value} = \text{two-sided exact p-value} - \binom{s_{10}+s_{01}}{s_{10}}\left(\frac{1}{2}\right)^{s_{10}+s_{01}}$$

and

$$\text{mid-p value} = \text{two-sided exact p-value} - \binom{r_{10}+r_{01}}{r_{10}}\left(\frac{1}{2}\right)^{r_{10}+r_{01}},$$

respectively. The conditional mid-p test is also referred to as quasi-exact test.

3.1.1.3. McNemar test

The McNemar test (McNemar, 1947) is the asymptotic version of the conditional exact test. Conditioning on the sum of discordant frequencies and applying the Central Limit Theorem, the statistic for hypothesis test $H_0 : p_{10} = p_{01}$ is

$$z = \frac{\hat{p}_{10} - \hat{p}_{01}}{\sqrt{Var\left(\hat{p}_{10} - \hat{p}_{01}\right)}},$$

which is distributed according to a standard normal distribution, where

$$Var\left(\hat{p}_{10} - \hat{p}_{01}\right) = \frac{p_{10} + p_{01} - \left(p_{10} - p_{01}\right)^2}{s}.$$

If the null hypothesis is true, then

$$Var_0\left(\hat{p}_{10} - \hat{p}_{01}\right) = \frac{p_{10} + p_{01}}{s}.$$

Finally, the test statistic for the McNemar test is

$$z_M = \left(s_{10} - s_{01}\right) \Big/ \sqrt{s_{10} + s_{01}}\ .$$

It is very common to express this statistic in terms of the chi-square distribution, i.e.

$$\chi^2_M = \frac{\left(s_{10} - s_{01}\right)^2}{s_{10} + s_{01}},$$

which is distributed asymptotically according to a chi-square distribution with one degree of freedom. In a similar way, the test statistic of the McNemar test is obtained to compare the two specificities:

$$\chi^2_M = \frac{\left(r_{10} - r_{01}\right)^2}{r_{10} + r_{01}}.$$

3.1.1.4. McNemar test with continuity correction

In the McNemar test the binomial distribution is approximated through the normal distribution. In this situation, it is common to apply continuity correction (*cc*). Edwards (1948) proposed the following continuity correction version of the McNemar test,

$$z_{Mcc} = \frac{\left|\hat{p}_{10} - \hat{p}_{01}\right| - \frac{1}{s}}{\sqrt{Var\left(\hat{p}_{10} - \hat{p}_{01}\right)}}.$$

Performing the same algebraic operations in the previous section, it is obtained that the statistics of the McNemar test with *cc* are

$$\chi^2_{Mcc} = \frac{\left(\left|s_{10} - s_{01}\right| - 1\right)^2}{s_{10} + s_{01}} \quad \text{and} \quad \chi^2_{Mcc} = \frac{\left(\left|r_{10} - r_{01}\right| - 1\right)^2}{r_{10} + r_{01}},$$

respectively.

3.1.1.5. Modified McNemar test

Bennett and Underwood (1970) proposed a modification of the statistic of the McNemar test adding 0.5 to the observed frequencies. This correction improves the approximation to the chi-square distribution. The statistics of *MMT* are:

$$\chi^2_{MM} = \frac{\left(s_{10} - s_{01}\right)^2}{s_{10} + s_{01} + 1} \quad \text{and} \quad \chi^2_{MM} = \frac{\left(r_{10} - r_{01}\right)^2}{r_{10} + r_{01} + 1}.$$

3.1.1.6. Wald test

The comparison of the two sensitivities (specificities) can also be made applying the Wald test (1943). The Wald test statistic to compare the sensitivities is

$$\chi^2_W = \frac{s\left(s_{10}-s_{01}\right)^2}{4s_{10}s_{01}+\left(s_{11}+s_{00}\right)\left(s_{10}+s_{01}\right)},$$

which is distributed asymptotically according to a chi-square distribution with one degree of freedom. To compare the two specificities, the Wald test statistics is

$$\chi^2_W = \frac{r\left(r_{10}-r_{01}\right)^2}{4r_{10}r_{01}+\left(r_{11}+r_{00}\right)\left(r_{10}+r_{01}\right)}.$$

3.1.1.7. Modified Wald test

As the *WT* tends to reject too often under the null hypothesis when the sample size is small or moderate, May and Johnson (1997) proposed a modification of the Wald statistic adding 0.5 to each one of the discordant frequencies, i.e.

$$\chi^2_{MW} = \frac{\left(s_{10}-s_{01}\right)^2}{\left(s_{10}+s_{01}+1\right)-\frac{\left(s_{10}-s_{01}\right)^2}{s}} \quad \text{and} \quad \chi^2_{MW} = \frac{\left(r_{10}-r_{01}\right)^2}{\left(r_{10}+r_{01}+1\right)-\frac{\left(r_{10}-r_{01}\right)^2}{s}}.$$

This modification reduces the size of the Wald statistic, and for $n \le 50$ the size of the test is close to the nominal error.

3.1.1.8. Likelihood ratio test

Conditioning in the sum of the discordant frequencies, if the null hypothesis $H_0 : p_{10} = p_{01}$ is true then it is verified that $\hat{p}_{10} = \hat{p}_{01} = \left(s_{10}+s_{01}\right)/\left(2s\right)$. The likelihood ratio statistic to compare the sensitivities is

$$\chi^2_{LR} = 2\left[s_{10}\ln\left(\frac{2s_{10}}{s_{10}+s_{01}}\right) + s_{01}\ln\left(\frac{2s_{01}}{s_{10}+s_{01}}\right)\right],$$

and in a similar way, the likelihood ratio statistic to compare the specificities is

$$\chi^2_{LR} = 2\left[r_{10} \ln\left(\frac{2r_{10}}{r_{10}+r_{01}} \right) + r_{01} \ln\left(\frac{2r_{01}}{r_{10}+r_{01}} \right) \right],$$

whose distributions are asymptotically a chi-square with one degree of freedom.

3.1.1.9. Unconditional exact test

The conditional exact test and the mid-p test are based on the conditioning on the sum of the discordant frequencies. Suissa and Shuster (1991) proposed, based on the statistic of the McNemar test, an exact test which uses all the frequencies in the sample and, therefore, does not condition in the sum of the discordant frequencies. When we compare the two sensitivities, the power function of the test is

$$P(p_{10}, p_{01}) = \sum_C \binom{s}{s_{10}\quad s_{01}\quad s-m} p_{10}^{s_{10}} p_{01}^{s_{01}} (1-p_{10}-p_{01})^{s-m},$$

where $m = s_{10} + s_{01}$ and $C = \{(s_{10}, m):\ s_{10} \geq h(m);\ s_{10} = 0,1,...,m;\ m = 0,1,...,s\}$, with $h(m) = \left(z_M \sqrt{m} + m\right)/2$ and z_M the calculated value of the McNemar statistic. If the null hypothesis is true, then the distribution of $(s_{10}, m, s-m)$ is a trinomial distribution with parameters *s* and probability vector is $(\delta/2, \delta/2, 1-\delta)^T$, i.e.

$$P(\delta) = \sum_C \binom{s}{s_{10}\quad s_{01}\quad s-m} \left(\frac{\delta}{2}\right)^m (1-\delta)^{s-m},$$

and where $\delta = p_{10} + p_{01}$ is the nuisance parameter. The nuisance parameter is eliminated by maximizing this function over the range of δ. The function $P(\delta)$ is simplified as

$$P(\delta) = \sum_{j=k}^{s} \binom{s}{j} \delta^j (1-\delta)^{s-j} F_j(j - i_j - 1),$$

where $k = \text{int}\left[z_m^2 + 1\right]$, $i_j = \text{int}\left[h(j)\right]$, $\text{int}[.]$ is the integer function and F_j is the cumulative binomial distribution function with parameters *j* and $1/2$. Finally, the two-sided exact p-value is calculated as

$$\text{two sided exact p-value} = 2 \times \sup_{0<\delta<1} \left\{ P(\delta) \right\}.$$

The two-sided exact p-value to compare the two specificities is calculated in a similar way, substituting "s" with "*r*" and "*p*" with "*q*".

3.1.1.10. Unconditional McNemar test

Lu (2010) proposed a statistic for the McNemar test that considers all the frequencies in the sample, and which therefore does not condition in the sum of the discordant frequencies. The hypothesis test $H_0 : p_{10} = p_{01}$ vs $H_1 : p_{10} \neq p_{01}$ is equivalent to the hypothesis test

$$H_0 : \frac{p_{10}}{p_{10} + p_{01}} = \frac{p_{01}}{p_{10} + p_{01}} \text{ vs } H_1 : \frac{p_{10}}{p_{10} + p_{01}} \neq \frac{p_{01}}{p_{10} + p_{01}}.$$

Subject to the null hypothesis, the frequency s_{10} (or s_{01}) is the product of binomial distribution of parameters *s* and $\delta = \left(p_{10} + p_{01} \right) / 2$. The estimators of the average and of the variance of the binomial distribution are $s\hat{\delta} = \left(s_{10} + s_{01} \right) / 2$ and $s\hat{\delta}\left(1 - \hat{\delta}\right) = \frac{s_{10} + s_{01}}{2} - \frac{\left(s_{10} + s_{01} \right)^2}{4s}$. Approximating to the normal distribution and applying the Central Limit Theorem, the statistic for the hypothesis test of equality of the two sensitivities is

$$z_{UM} = \frac{s_{10} - s\hat{\delta}}{\sqrt{s\hat{\delta}\left(1 - \hat{\delta}\right)}} = \frac{s_{10} - s_{01}}{\sqrt{\frac{\left(s_{10} + s_{01} \right)\left(s + s_{11} + s_{00} \right)}{s}}},$$

or in terms of the chi-square distribution

$$\chi^2_{UM} = \frac{s\left(s_{10} - s_{01} \right)^2}{\left(s_{10} + s_{01} \right)\left(s + s_{11} + s_{00} \right)},$$

whose distribution is asymptotically a chi-square with one degree of freedom. In a similar way, we obtain the statistic to compare the two specificities:

$$\chi^2_{UM} = \frac{r\left(r_{10} - r_{01}\right)^2}{\left(r_{10} + r_{01}\right)\left(r + r_{11} + r_{00}\right)}.$$

3.1.1.11. Unconditional likelihood ratio test

Lu (2011) also proposed a likelihood ratio test statistic to compare two paired binomial proportions without discarding the concordant frequencies. The likelihood ratio test statistic is obtained in two phases: in the first phase we obtain the likelihood ratio test statistic when the four frequencies s_{ij} are combined in two, s_{10} and $s_{11} + s_{01} + s_{00}$; and in the second phase we obtain the likelihood ratio test statistic when the four frequencies s_{ij} are combined in another two, s_{01} and $s_{11} + s_{10} + s_{00}$. Finally, the likelihood ratio test statistic is calculated as an average of the two likelihood ratio test statistics. In the context studied here, the likelihood ratio test statistics are

$$\chi^2_{ULR} = s_{10}\ln\left(\frac{2s_{10}}{s_{10} + s_{01}}\right) + s_{01}\ln\left(\frac{2s_{01}}{s_{10} + s_{01}}\right) +$$
$$\left(s - s_{10}\right)\ln\left[\frac{2\left(s - s_{10}\right)}{2s - s_{10} - s_{01}}\right] + \left(s - s_{01}\right)\ln\left[\frac{2\left(s - s_{01}\right)}{2s - s_{10} - s_{01}}\right]$$

and

$$\chi^2_{ULR} = r_{10}\ln\left(\frac{2r_{10}}{r_{10} + r_{01}}\right) + r_{01}\ln\left(\frac{2r_{01}}{r_{10} + r_{01}}\right) +$$
$$\left(r - r_{10}\right)\ln\left[\frac{2\left(r - r_{10}\right)}{2r - r_{10} - r_{01}}\right] + \left(r - r_{01}\right)\ln\left[\frac{2\left(r - r_{01}\right)}{2r - r_{10} - r_{01}}\right],$$

which in both cases asymptotically follow a chi-square with one degree of freedom.

3.1.2. Global test

Lachenbruch and Lynch (1998) studied the simultaneous comparison of the sensitivities and of the specificities, i.e. solving the global test

$$H_0 : \left(Se_1 = Se_2 \text{ and } Sp_1 = Sp_2\right) \text{ vs } H_1 : \left(Se_1 \neq Se_2 \text{ and/or } Sp_1 \neq Sp_2\right)$$

to an α error. Therefore, these authors have proposed these test statistics:

$$\chi^2_{LR} = 2\left[s_{10} \ln\left(\frac{2s_{10}}{s_{10}+s_{01}} \right) + s_{01} \ln\left(\frac{2s_{01}}{s_{10}+s_{01}} \right) + r_{10} \ln\left(\frac{2r_{10}}{r_{10}+r_{01}} \right) + r_{01} \ln\left(\frac{2r_{01}}{r_{10}+r_{01}} \right) \right]$$

and

$$\chi^2_R = \frac{\left(s_{10}-s_{01}\right)^2}{s_{10}+s_{01}} + \frac{\left(r_{10}-r_{01}\right)^2}{r_{10}+r_{01}}.$$

The first one is obtained applying the likelihood ratio test and the second one adding the test statistics of the individual tests of the McNemar test.

3.1.3. Methodology

We have studied in greater depth the global test applying the Wald Method (1943) and the Rao Method (1948), in order to obtain new asymptotic solutions for this hypothesis test. Moreover, Monte Carlo simulation experiments were carried out to compare the different methods based on the individual tests (to an α error and applying the multiple comparison methods of Bonferroni (1936) and Holm (1979)) and in the global test, and in this way we were able to determine which are the methods with the best asymptotic behaviour to compare the sensitivities and specificities of two *BDTs* subject to a paired design.

3.2. Comparison of the likelihood ratios of two diagnostic tests subject to a paired design: confidence intervals and sample size

Pepe (2003) and Roldán-Nofuentes and Luna proposed confidence intervals for the ratio of the *LRs* of two *BDTs* subject to a paired design.

3.2.1. Regression model

Leisenring and Pepe (1998) studied the estimation of the *LRs* of a *BDT* in presence of covariates through a regression model. For the positive *LR*, the regression model with *p* covariables is $\ln\left(LR^+\left(X_1\right)\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_{1p}$, where β_i are the parameters of the model

and $X_1 = \left(X_{11},...,X_{1p}\right)$ is the matrix of covariates. This model can be used to compare two *BDTs* (Pepe, 2003), i.e. $\ln\left[LR^+\left(X_T\right)\right] = \beta_0 + \beta_1 X_T$, where X_T is a variable dummy to compare a *BDT* in relation to another. The regression model to compare the two negative *LRs* is $\ln\left[LR^+\left(X_T\right)\right] = \alpha_0 + \alpha_1 X_T$. In these models, the ratio $\omega^+ = LR_1^+ / LR_2^+$ is estimated as $e^{\hat{\beta}_1}$ and the ratio $\omega^- = LR_1^- / LR_2^-$ as $e^{\hat{\alpha}_1}$. The confidence interval for ω^+ is

$$\hat{\omega}^+ \times \exp\left\{\pm z_{1-\alpha/2}\sqrt{\hat{V}ar_0\left[\ln\left(\hat{\omega}^+\right)\right]}\right\},$$

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

$$\hat{V}ar_0\left[\ln\left(\hat{\omega}^+\right)\right] = \frac{1-\hat{S}e_1}{s\hat{S}e_1} + \frac{\hat{S}p_1}{r\left(1-\hat{S}p_1\right)} + \frac{1-\hat{S}e_2}{s\hat{S}e_2} + \frac{\hat{S}p_2}{r\left(1-\hat{S}p_2\right)}$$

is the estimated variance of $\hat{\omega}^+$ subject to the null hypothesis $H_0 : LR_1^+ = LR_2^+$. The confidence interval for ω^- s similar to the previous one, where

$$\hat{V}ar_0\left[\ln\left(\hat{\omega}^-\right)\right] = \frac{\hat{S}e_1}{s\left(1-\hat{S}e_1\right)} + \frac{1-\hat{S}p_1}{r\hat{S}p_1} + \frac{\hat{S}e_1}{s\left(1-\hat{S}e_1\right)} + \frac{1-\hat{S}p_1}{r\hat{S}p_1}.$$

3.2.2. Logarithmic interval

Roldán-Nofuentes and Luna (2007) studied a hypothesis test to compare the *LRs* applying the method of maximum likelihood and the delta method to solve the hypothesis test $H_0 : \ln(\omega) = 0$ vs $H_1 : \ln(\omega) \neq 0$, where ω is ω^+ or ω^-. The test statistic is $\ln(\hat{\omega}) / \sqrt{\hat{V}ar(\hat{\omega})}$, and its distribution is asymptotically normal. Inverting the test, the logarithmic *CI* for ω is

$$\hat{\omega} \times \exp\left\{\pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left[\ln\left(\hat{\omega}\right)\right]}\right\},$$

where $\hat{V}ar\left[\ln\left(\hat{\omega}\right)\right]$ is obtained applying the delta method.

3.2.3. Methodology

New confidence intervals were studied for the ratio of the positive (negative) *LRs* applying the Wald Method (1943), the Fieller method (1940), Bootstrap (Efron and Tibshirani, 1993) and the Bayesian Monte Carlo method (Boos and Stefanski, 2013). Monte Carlo simulation experiments were carried out to study the asymptotic coverage and the average width of the confidence intervals studied. Moreover, a method has been proposed to calculate the sample size to compare the positive (negative) *LRs* through a confidence interval, applying for this purpose an iterative method of calculation of the sample size based on a pilot sample.

3.3. Asymptotic confidence intervals for the difference and the ratio of the weighted kappa coefficients of two diagnostic tests subject to a paired design

The comparison of the weighted kappa coefficients of two *BDTs* subject to a paired design was studied by Bloch (1997).

3.3.1. Bloch method

Bloch (1997) studied the comparison of the weighted kappa coefficients of two *BDTs* subject to a paired design, i.e.

$$H_0 : \kappa_1(c) = \kappa_2(c) \text{ vs } H_1 : \kappa_1(c) \neq \kappa_2(c).$$

The test statistic for this hypothesis test is

$$z = \frac{\hat{\kappa}_1(c) - \hat{\kappa}_2(c)}{\sqrt{\hat{V}ar\left[\hat{\kappa}_1(c)\right] + \hat{V}ar\left[\hat{\kappa}_2(c)\right] - 2\hat{C}ov\left[\hat{\kappa}_1(c), \hat{\kappa}_2(c)\right]}} \xrightarrow[n\to\infty]{} N(0,1),$$

where the variances-covariances are obtained applying the delta method. Inverting the test statistic, a Wald type confidence interval for $\kappa_1(c) - \kappa_2(c)$ is

$$\hat{\kappa}_1(c) - \hat{\kappa}_2(c) \pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left[\hat{\kappa}_1(c)\right] + \hat{V}ar\left[\hat{\kappa}_2(c)\right] - 2\hat{C}ov\left[\hat{\kappa}_1(c), \hat{\kappa}_2(c)\right]}.$$

3.3.2. Methodology

New confidence intervals were obtained for the difference and for the ratio of the weighted kappa coefficients of two *BDTs*, thus extending the study by Bloch (1997). Therefore, we used the Wald method (1943), the Fieller method (1940), Bootstrap (Efron and Tibshirani, 1993) and the Bayesian Monte Carlo method (Boos and Stefanski, 2013). Monte Carlo simulation experiments were carried out to study the asymptotic coverage and the average width of the confidence intervals studied, giving some general rules of application for the intervals for the difference and for the ratio. Furthermore, a method has been proposed to calculate the sample size to compare the two weighted kappa coefficients through a confidence interval, applying for this purpose a methodology similar to that applied in Section 3.2.

3.4. *EM* and *SEM* algorithms to compare the weighted kappa coefficients of two diagnostic tests in the presence of partial verification and discrete covariates

In the presence of partial disease verification, the comparison of the weighted kappa coefficients of two *BDTs* when a discrete covariate is observed in all of the individuals is a question that has not been previously studied. Therefore, in this situation it is considerd that in all of the n individuals in the sample we observe a vector $X=\left(x_1,x_2,...,x_M\right)$ of a discrete covariate, where x_m $\left(m=1,...,M\right)$ is each one of the different values or patterns that the covariate can take. For the mth pattern of the covariate $\left(X=x_m\right)$ we obtain the frequencies from Table 3.1. The total sample of n individuals can be seen as a mixture of M multinomial independent 3×4 tables.

The objective is to study a hypothesis test to compare the two weighted kappa coefficients in this situation.

Table 3.1. Observed frequencies in the presence of partial verification for $X = x_m$.

	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$V = 1$					
$D = 1$	s_{11m}	s_{10m}	s_{01m}	s_{00m}	s_m
$D = 0$	r_{11m}	r_{10m}	r_{01m}	r_{00m}	r_m
$V = 0$	u_{11m}	u_{10m}	u_{01m}	u_{00m}	u_m
Total	n_{11m}	n_{10m}	n_{01m}	n_{00m}	n_m

3.4.1. Methodology

This problem was solved applying two methods of computation: the *EM* algorithm and the *SEM* algorithm.

3.4.1.1. EM algorithm

The *EM* algorithm (Dempster et al, 1977) is a very well known method in Statistics to impute the estimators of parameters in presence of missing data, and it requires that missing data to be missing at random (*MAR*). For the *m*th table $\left(X = x_m\right)$ let us suppose that from each frequency u_{ijm} of non-verified individuals, d_{ijm} have the disease and $u_{ijm} - d_{ijm}$ do not have the disease, with $i, j = 0,1$. Then each one of the M 3×4 tables can be expressed in the form of a 2×4 table with frequencies $s_{ijm} + d_{ijm}$ for $D = 1$ and $r_{ijm} + u_{ijm} - d_{ijm}$ for $D = 0$. Table 3.2 shows the frequencies of the complete data for $X = x_m$. For each one of the multinomial 3×4 tables the missing information is the true disease status of the individuals not verified with the *GS*. This information is reconstructed in the *E* step of the algorithm and in the *M* step the values of the maximum likelihood estimators are imputed. The application of the *EM* algorithm allows us to obtain the estimations of the weighted kappa coefficients of the *BDTs*.

Table 3.2. Complete data for $X = x_m$.

	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$D = 1$	$s_{11m} + d_{11m}$	$s_{10m} + d_{10m}$	$s_{01m} + d_{01m}$	$s_{00m} + d_{00m}$	$s_m + d_m$
$D = 0$	$r_{11m} + u_{11m}$	$r_{10m} + u_{10m}$	$r_{01m} + u_{01m}$	$r_{00m} + u_{00m}$	$r_m + u_m$
	$-d_{11m}$	$-d_{10m}$	$-d_{01m}$	$-d_{00m}$	$-d_m$
Total	n_{11m}	n_{10m}	n_{01m}	n_{00m}	n_m

3.4.1.2. SEM algorithm

The estimation of the matrix of variances-covariances of the estimators was obtained through the application of the *SEM* algorithm (Meng and Rubin, 1991). The *SEM* algorithm (Supplemental *EM*) is a computational method that allows us to estimate the variance-covariance matrix of a vector of estimators using the calculations made in the application of the *EM* algorithm. Let $\sum_{\hat{\boldsymbol{\theta}}}$ be the variance-covariance matrix, and Dempster et al (1977) demonstrated that

$$\sum_{\hat{\boldsymbol{\theta}}} = I_{oc}^{-1}\left(I - DM\right)^{-1}$$

where I is the identity matrix and $DM = I_{mis}I_{oc}^{-1}$, when I_{oc} is the Fisher information matrix of the complete data and I_{mis} is the Fisher information matrix of the missing data. The *SEM* algorithm consists of three phases: (1) to assess the matrix I_{oc}^{-1}, (2) to assess the *DM* matrix, and (3) to assess the variance-covariance matrix $\sum_{\hat{\boldsymbol{\theta}}}$. The main objective of the *SEM* algorithm is to calculate the elements of the DM matrix, which is done through an algorithm that uses calculations made by the *EM* algorithm that runs iterations of the *EM* algorithm. Therefore, the application of the *SEM* algorithm allows us to estimate the variances-covariances of the weighted kappa coefficients of the *BDTs*.

3.4.1.3. Hypothesis Test

Once we have obtained the estimators of the weighted kappa coefficients and their variances-covariances applying the previous algorithms, the test statistic for the hypothesis test

$$H_0 : \kappa_1(c) = \kappa_2(c) \text{ vs } H_1 : \kappa_1(c) \neq \kappa_2(c)$$

is

$$z = \frac{\hat{\kappa}_1(c) - \hat{\kappa}_2(c)}{\sqrt{\hat{V}ar\left[\hat{\kappa}_1(c)\right] + \hat{V}ar\left[\hat{\kappa}_2(c)\right] - 2Cov\left[\hat{\kappa}_1(c), \hat{\kappa}_2(c)\right]}},$$

which is distributed according to a normal distribution when the sample size *n* is large.

Once the hypothesis test is solved, Monte Carlo simulation experiments were carried out to study its type I error and its power.

CHAPTER 4

RESULTS

This Doctoral Thesis has obtained different results related to the comparison of parameters of two *BDTs*. We will summarize the results obtained for each one of the objectives. The complete results can be seen in the Appendices.

4.1. Recommended methods to compare the accuracy of two binary diagnostic tests subject to a paired design

- Applying the Rao score test, a test statistic was obtained for the global hypothesis test:

$$\chi_R^2 = \frac{(s_{10} - s_{01})^2}{s_{10} + s_{01}} + \frac{(r_{10} - r_{01})^2}{r_{10} + r_{01}},$$

  whose distribution is a chi-square with two degrees of freedom when the sample size n is large.

- Applying the Wald method, a test statistic was obtained for the global hypothesis test:

$$\chi_W^2 = \frac{s(s_{10} - s_{01})^2}{4s_{10}s_{01} + (s_{11} + s_{00})(s_{10} + s_{01})} + \frac{r(r_{10} - r_{01})^2}{4r_{10}r_{01} + (r_{11} + r_{00})(r_{10} + r_{01})},$$

whose distribution is a chi-square with two degrees of freedom when the sample size *n* is large.

- Based on the results obtained in the simulation experiments, the following general rules of application can be established:

  1). When the prevalence is small $(p=10\%)$ or very small $(p=5\%)$ and the sample is small $(n=50)$ or moderate $(n=100)$, solve the tests $H_0: Se_1 = Se_2$ and $H_0: Sp_1 = Sp_2$ individually applying the Wald test or the Likelihood ratio test along with the Bonferroni (or Holm) method to an error $\alpha = 5\%$.

  2). In any other situation, solve the global test $H_0: (Se_1 = Se_2 \text{ and } Sp_1 = Sp_2)$ to an error $\alpha = 5\%$ applying the likelihood ratio test or the Wald test. In this situation, if the global test is not significant then the equality of the accuracy of both BDTs is not rejected, and if the global is significant then the causes of the significance will be investigated: a) testing the tests $H_0: Se_1 = Se_2$ and $H_0: Sp_1 = Sp_2$ individually applying the Wald test or the likelihood ratio test along with the Bonferroni or Holm method to an error $\alpha = 5\%$ if the sample size is small or moderate $(n \le 100)$ or if the sample size is very large $(n \ge 1000)$; or b) testing the tests $H_0: Se_1 = Se_2$ and $H_0: Sp_1 = Sp_2$ individually applying the McNemar test with continuity correction to an error $\alpha = 5\%$ if the sample size is large $(200 \le n \le 500)$.

4.2. Comparison of the likelihood ratios of two diagnostic tests subject to a paired design: confidence intervals and sample size

- Applying the Wald method, a confidence interval was obtained for the ratio of the two positive (negative) likelihood ratios:

$$\frac{\hat{LR}_1}{\hat{LR}_2} \times \left[1 \pm z_{1-\alpha/2} \sqrt{\frac{\hat{Var}\left(\hat{LR}_1\right)}{\hat{LR}_1^2} + \frac{\hat{Var}\left(\hat{LR}_2\right)}{\hat{LR}_2^2} - \frac{2\hat{Cov}\left(\hat{LR}_1, \hat{LR}_2\right)}{\hat{LR}_1 \hat{LR}_2}} \right].$$

- Applying the Fieller method, a confidence interval was obtained for the ratio of the two positive (negative) likelihood ratios:

$$\frac{LR_1}{LR_2} \in \frac{\hat{L}R_1\hat{L}R_2 - \hat{\sigma}_{12}z^2_{1-\alpha/2} \pm \sqrt{\left(\hat{L}R_1\hat{L}R_2 - \hat{\sigma}_{12}z^2_{1-\alpha/2}\right)^2 - \left(\hat{L}R_1^2 - \hat{\sigma}_{11}z^2_{1-\alpha/2}\right)\left(\hat{L}R_2^2 - \hat{\sigma}_{22}z^2_{1-\alpha/2}\right)}}{\hat{L}R_2^2 - \hat{\sigma}_{22}z^2_{1-\alpha/2}},$$

  where $\hat{\sigma}_{ii} = \hat{V}ar\left(\hat{L}R_i\right)$ and $\hat{\sigma}_{12} = \hat{C}ov\left(\hat{L}R_1, \hat{L}R_2\right)$

- The ratio of the positive (negative) likelihood ratios was estimated applying Bootstrap, calculating the bias-corrected interval.
- The ratio of the positive (negative) likelihood ratios was estimated applying the Bayesian Monte Carlo method, calculating a confidence interval based on quantiles.
- From the simulation experiments carried out, the following general rules of application for the intervals were given:

  1). For the ratio of the positive likelihood ratios, use the logarithmic confidence interval, whatever the sample size may be, although when $n \geq 200$ we can also use the Wald, the Fieller and the Bootstrap intervals.

  2). For the ratio of the negative likelihood ratios, use the Wald *CI,* whatever the sample size may be.

- A method was proposed to calculate the sample size necessary to estimate the ratio of the positive (negative) likelihood ratios through the Wald confidence interval. The method requires the estimation of all the parameters starting from a pilot sample.

4.3. Asymptotic confidence intervals for the difference and the ratio of the weighted kappa coefficients of two diagnostic tests subject to a paired design

- The relation between the two weighted kappa coefficients and the true (false) positive fraction between the two *BDTs* was studied.
- The difference and the ratio between the two weighted kappa coefficients was estimated applying Bootstrap, calculating the bias-corrected interval for the difference and for the ratio.
- The difference and the ratio between the two weighted kappa coefficients was estimated applying the Bayesian Monte Carlo method, calculating a confidence interval based on quantiles for the difference and for the ratio.
- A logarithmic interval was obtained for the ratio of the two weighted kappa coefficients:

$$\frac{\kappa_1(c)}{\kappa_2(c)} \in \frac{\hat{\kappa}_1(c)}{\hat{\kappa}_2(c)} \times \exp\left\{ \pm z_{1-\alpha/2} \sqrt{\hat{V}ar\left[\ln\left\{ \frac{\hat{\kappa}_1(c)}{\hat{\kappa}_2(c)} \right\} \right]} \right\},$$

where

$$\hat{V}ar\left[\ln\left\{ \frac{\hat{\kappa}_1(c)}{\hat{\kappa}_2(c)} \right\} \right] \approx \frac{\hat{V}ar\left[\hat{\kappa}_1(c)\right]}{\hat{\kappa}_1^2(c)} + \frac{\hat{V}ar\left[\hat{\kappa}_2(c)\right]}{\hat{\kappa}_2^2(c)} - \frac{2\hat{C}ov\left[\hat{\kappa}_1(c), \hat{\kappa}_2(c)\right]}{\hat{\kappa}_1(c)\hat{\kappa}_2(c)}.$$

- Applying the Fieller method, a confidence interval was obtained for the ratio of the two weighted kappa coefficients:

$$\frac{\kappa_1(c)}{\kappa_2(c)} \in \frac{\hat{\omega}_{12} \pm \sqrt{\hat{\omega}_{12}^2 - \hat{\omega}_{11}\hat{\omega}_{22}}}{\hat{\omega}_{22}},$$

where $\hat{\omega}_{ij} = \hat{\kappa}_i(c) \times \hat{\kappa}_j(c) - \hat{\sigma}_{ij} z_{1-\alpha/2}^2$, $\hat{\sigma}_{ii} = \hat{V}ar\left[\hat{\kappa}_i(c)\right]$ and $\hat{\sigma}_{12} = \hat{C}ov\left[\hat{\kappa}_1(c), \hat{\kappa}_2(c)\right]$

- Based on the simulation experiments carried out, the following general rules of application for the intervals were given:

1). If n is small $(n<100)$, use the Wald *CI* for the ratio increasing the frequencies s_{ij} and r_{ij} in 0.5.

2). If $100 \leq n \leq 400$, use the Wald *CI* for the ratio without adding 0.5.

3). If $n \geq 500$, use any of the *CIs* (for the difference or for the ratio) without adding 0.5.

- A method was proposed to calculate the sample size necessary to estimate the ration between the two weighted kappa coefficients through the Wald confidence interval. The method requires estimation of all the parameters starting from a pilot sample.
- A programme written in *R* that allows us to calculate all of the confidence intervals studied and also calculate sample size necessary to compare the two weighted kappa coefficients. The programme, called "citwkc" (Confidence Intervals for Two Weighted Kappa Coefficients) is available free of change at the following URL:

https://www.ugr.es/~bioest/software/cmd.php?seccion=mdb

4.4. *EM* and *SEM* algorithms to compare the weighted kappa coefficients of two diagnostic tests in the presence of partial verification and discrete covariates

- Assuming that the missing data mechanism is *MAR*, the estimations of the weighted kappa coefficients are obtained in each pattern of the covariate applying the *EM* algorithm. The estimator of each global weighted kappa coefficient is obtained combining properly the estimators in each pattern of the covariate.
- The variances-covariances of the weighted kappa coefficients were estimated applying the *SEM* algorithm. It has been demonstrated that the elements of the *DM* matrix between two different patterns of the covariate are equal to zero, and therefore the *DM* matrix is expressed as

$$DM = \text{Diag}\{DM_1, DM_2, ..., DM_M\},$$

when DM_m is the *DM* matrix in the *m*th pattern of the covariate.

- A Wald type test statistic was proposed to solve the hypothesis test of equality of the weighted kappa coefficients and simulation experiments were carried out to study its asymptotic behaviour, and it was found that its type I error fluctuates around the nominal error without overwhelming it when the sample size is large and that for the power to be high it is necessary to have a large sample size.

CHAPTER 5

CONCLUSIONS

5.1. Recommended methods to compare the accuracy of two binary diagnostic tests subject to a paired design

Traditionally, the comparison of the accuracy of two *BDTs* subject to a paired design is made conditioning on the individuals with (without) the disease and comparing the two sensitivities (specificities) applying a comparison test with two paired binomial proportions to an α error. Therefore, each one of the tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ are tested independently to an α error. An alternative to this method is to compare the two sensitivities and the two specificities simultaneously, i.e. performing the global test $H_0 : \left(Se_1 = Se_2 \text{ and } Sp_1 = Sp_2 \right)$ vs $H_1 : \left(Se_1 \neq Se_2 \text{ and/or } Sp_1 \neq Sp_2 \right)$. This manuscript studies this global hypothesis test, extending the study by Lachenbruch and Lynch (1998), through the application of Rao's score test and the Wald test. Lachenbruch and Lynch proposed two statistics for the global test, one obtained applying the likelihood ratio test and another one obtained as the sum of the McNemar statistic for the test $H_0 : Se_1 = Se_2$ and of the McNemar statistic for the test $H_0 : Sp_1 = Sp_2$. In this study the same statistic has been derived applying Rao's score test. Another statistic has also been obtained using the Wald test, which is also the sum of the Wald statistics for the individual tests. Another alternative method that has been studied to compare the accuracy of two *BDTs* consisted of testing the two individual

tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ through a comparison test of paired binomial proportions and application of the Bonferroni method or the Holm method. Simulation experiments were carried out to study the asymptotic behaviour of the different methods to compare the sensitivities and specificities, giving some general rules of application for the hypothesis tests studied.

5.2. Comparison of the likelihood ratios of two diagnostic tests subject to a paired design: confidence intervals and sample size

The likelihood ratios are parameters that are used to assess and compare the effectiveness of *BDTs*, and only depend on the sensitivity and specificity of the *BDT*. The comparison of the likelihood ratios of two *BDTs* consists of the comparison of two relative risks. This manuscript has studied this comparison through *CIs* for the ratio of the two positive (negative) likelihood ratios, considering paired design as a type of sampling. Six *CIs* were studied, of which five were frequentist and one was Bayesian. The five frequentist intervals are based on the asymptotic normality of the estimators of the *LRs* or of functions of the *LRs*. Regarding the Bayesian interval, this was obtained applying the Monte Carlo method and considering distribution which a priori are non-informative. The comparison of the asymptotic behaviour of the *CIs* was studied though simulation experiments. The results of these experiments have shown that in order to estimate the ratio of the two positive likelihood ratios, in general terms, the intervals with the best behaviour are the logarithmic interval (for all sample sizes), the Wald interval, the Fieller interval or the Bootstrap (these last three for large or very large samples); on the other hand, in order to estimate the ratio of the two negative likelihood ratios the interval with the best behaviour is the Wald interval (for all sample sizes). A method has also been proposed to determine the sample size necessary to compare the ratio between the two positive (negative) likelihood ratios with a precision and confidence level.

5.3. Asymptotic confidence intervals for the difference and the ratio of the weighted kappa coefficients of two diagnostic tests subject to a paired design

The weighted kappa coefficient of a *BDT* is a measure of the beyond-chance agreement between the *BDT* and the *GS*, and depends on the sensitivity and specificity of the *BDT*, on the disease prevalence and on the weighting index. The weighting index *c* is a measurement of the relative loss between a false positive and a false negative, and it is a value set by the clinician depending on the way that the *BDT* is going to be used. We have studied the comparison of the weighted kappa coefficients of two *BDTs* through confidence intervals subject to a paired design. Three intervals were studied for the difference of the two weighted kappa coefficients and another five intervals for the ratio between the two parameters. All of the intervals studied are asymptotic and simulation experiments were carried out to study their asymptotic behaviour, and based on the results of these experiments some general rules of application have been given. A method has also been proposed to calculate the sample size to estimate the ratio between the two weighted kappa coefficients with a determined accuracy and confidence.

5.4. *EM* and *SEM* algorithms to compare the weighted kappa coefficients of two diagnostic tests in the presence of partial verification and discrete covariates

The weighted kappa coefficient of a *BDT* is used to assess and compare the effectiveness of *BDTs* when considering the losses of a misclassification with the *BDTs*. A hypothesis test has been studied to compare the weighted kappa coefficients of two *BDTs* when in the presence of partial disease verification a discrete covariate is observed in all of the individuals. The hypothesis test proposed is based on the fact that the verification process with the gold standard only depends on the results of the two *BDTs* and on the covariate, and therefore on the fact that the verification process is *MAR*. The solution of the hypothesis test of equality of the two weighted kappa coefficients was achieved applying computational methods: the *EM* algorithm for the calculation of the estimators and the *SEM* algorithm for the calculation of the variances-covariances. The *EM* algorithm is very well known and applied in a multitude of

problems with missing data. Nevertheless, the application of the *SEM* algorithm is not so frequent, even though this is a method which is inherent to the *EM* algorithm since it uses many of its calculations. Applying the *SEM* algorithm, it has been demonstrated that the elements of the *DM* matrix between estimators of two different patterns of the covariate are equal to 0, and therefore the *DM* matrix is expressed as a diagonal matrix of the form $DM = Diag\{DM_1,...,DM_M\}$, when each DM_m matrix is the *DM* matrix in the *m*th pattern of the covariate. This decomposition of the *DM* matrix simplifies the calculations of the variance-covariance matrix. Simulation experiments were carried out to study the asymptotic behaviour of the hypothesis test when the covariate is binary, which is a type of covariate which is frequent in clinical practice.

BIBLIOGRAPHY

Agresti, A. (2002). Categorical data analysis. Wiley, New York.

Batwala, V., Magnussen, P., Nuwaba, F. (2010). Are rapid diagnostic tests more accurate in diagnosis of plasmodium falciparum malaria compared to microscopy at rural health centers? Malaria Journal, 9, 349.

Begg, C.B., Greenes, R.A., (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics, 39, 207-215.

Bennett, B.M., Underwood, R.E. (1970). On McNemar's test for the 2×2 table and its power function. Biometrics, 26, 339-343.

Berry, G., Smith, C., Macaskill, P., Irwig, L. (2002). 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. Statistics in Medicine, 21, 853-862.

Biggerstaff, B.J. (2000). Comparing diagnostic tests: a simple graphic using likelihood ratios. Statistics in Medicine 19, 649-663.

Bloch, D.A. (1997). Comparing two diagnostic tests against the same 'gold standard' in the same sample. Biometrics, 53, 73-85.

Bloch, D.A. (1997). Comparing two diagnostic tests against the same "gold standard" in the same sample. Biometrics, 53, 73-85.

Bloch, D.A., Kraemer, H.C. (1989). 2×2 Kappa coefficients: measures of agreement or association. Biometrics, 45, 269-287.

Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8, 3-62.

Boos, D. D., Stefanski, L. A. (2013). Essential Statistical Inference. Theory and Method. Springer, New York.

Cicchetti, D.V. (2001). The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements, Journal of Clinical and Experimental Neuropsychology, 23, 695-700.

Dempster, A., Laird, N., Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1-38.

Dolgun, N. A, Gozukara, H., Karaagaoglu, E. (2012). Comparing diagnostic tests: test of hypothesis for likelihood ratios. Journal of Statistical Computation and Simulation, 82, 369-381.

Edwards, A.L. (1948). Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. Psychometrika, 13, 185-187.

Efron B., Tibshirani R J. (1993). An Introduction to the Bootstrap. Chapman and Hall, New York.

Fagerland, M.W., Lydersen, S., Laake, P. (2014). Recommended tests and confidence intervals for paired binomial proportions. Statistics in Medicine, 33, 2850-2875.

Fieller, E. C. (1940). The biological standardization of insulin. Journal of the Royal Statistical Society 7, 1-64.

Fieller, E.C. (1940). The biological standardization of insulin. Journal of the Royal Statistical Society, 7 Supplement, 1-64.

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. (1996). A cross-cultural community based study of dementias: methods and performance of survey instrument. International Journal of Methods in Psychiatric Research, 6, 129-142.

Hamdan, M.A., Pirie, W.R., Arnold, J.C. (1975). Simultaneous testing of McNemar's problem for several populations. Psychometrika, 40, 153-161.

Harel, O., Zhou, X.H. (2007). Multiple imputation for the comparison of two screening tests in two-phase Alzheimer studies. Statistics in Medicine, 26, 2370-2388

Holm, S. (1979). A simple sequential rejective multiple testing procedure. Scandinavian Journal of Statistics, 6, 65-70.

Hosmer, D.W., Lemeshow, S. (1989). Applied Logistic Regression. Wiley, New York.

Kraemer, H.C. (1992). Evaluating Medical Tests. Objective and Quantitative Guidelines. Sage Publications, Newbury Park.

Kraemer, H.C., Bloch, D.A. (1990). A Note on case-control sampling to estimate kappa coefficients. Biometrics, 46, 49-59.

Kraemer, H.C., Periyakoil, V.S., Noda, A. (2002). Kappa coefficients in medical research. Statistics in Medicine, 21, 2109-2129.

Lachenbruch, P.A., Lynch, C.J. (1998). Assessing screening tests: extensions of McNemar's test. Statistics in Medicine, 17, 2207-2217.

Lancaster, H.O. (1961). Significance tests in discrete distribution. Journal of American Statistical Association, 56, 223-234

Leisenring, W., Pepe, M.S. (1998). Regression modelling of diagnostic likelihood ratios for the evaluation of medical diagnostic tests. Biometrics, 54, 444-442.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B, 44, 226-233.

Lu, Y. (2010). A revised version of McNemar's test for paired binary data. Communications in Statistics - Theory and Methods, 39, 3525-3539.

Lu, Y. (2011). Considering the concordant observations in likelihood ratio test for paired binary data. Communications in Statistics - Theory and Methods, 39, 4214-4232.

Martín-Andrés, A. and Álvarez-Hernández, M. (2014). Two-tailed approximate confidence intervals for the ratio of proportions. Statistics and Computing, 24, 65-75.

Martín-Andrés, A., Álvarez-Hernández, M. (2014). Two-tailed asymptotic inferences for a proportion. Journal of Applied Statistics, 41, 1516-1529.

May, W.L., Johnson, W.D. (1997). The validity and power of tests for equality of two correlated proportions. Statistics in Medicine, 16, 1081-1096.

McNeil B., Adelstein J., (1976). Determining the value of diagnostic and screening tests. Journal of Nuclear Medicine, 17, 439-448.

McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages. Psychometrika, 12, 153-157.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. Journal of the Royal Statistical Society, Series B, 51, 127-138.

Meng, X., Rubin, D. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. Journal of the American Statistical Association, 86, 899-909.

Montero-Alonso, M.A., Roldán-Nofuentes, J.A. (2019). Approximate confidence intervals for the likelihood ratios of a binary diagnostic test in the presence of partial disease verification. Journal of Biopharmaceutical Statistics 29: 56-81.

Pepe, M. S. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, New York.

Price, R.M., Bonett, D.G. (2004). An improved confidence interval for a linear function of binomial proportions. Computational Statistics and Data Analysis, 45, 449-456.

Rao, C.R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. Proceedings of the Cambridge Philosophical Society, 44, 50-57.

Roldán-Nofuentes, J.A., Amro, R. (2017). Approximate confidence intervals for the weighted kappa coefficient of a binary diagnostic test subject to a case-control design. Journal of Statistical Computation and Simulation, 87, 530-545.

Roldán-Nofuentes, J.A., Amro, R. (2018). Combination of the weighted kappa coefficients of two binary diagnostic tests. Journal of Biopharmaceutical Statistics, 28, 909-926.

Roldán-Nofuentes, J.A., Luna del Castillo, J.D. (2006). Comparing two binary diagnostic tests in the presence of verification bias. Computational Statistics and Data Analysis, 50, 1551-1564.

Roldán-Nofuentes, J.A., Luna del Castillo, J.D. (2007). Comparison of the likelihood ratios of two binary diagnostic tests in paired designs. Statistics in Medicine, 26, 4179-4201.

Roldán-Nofuentes, J.A., Luna del Castillo, J.D. (2008). EM algorithm for comparing two binary diagnostic tests when not all the patients are verified. Journal of Statistical Computation and Simulation, 78, 19-35.

Roldán-Nofuentes, J.A., Luna del Castillo, J.D., Femia-Marzo, P. (2009). Computational methods for comparing two binary diagnostic tests in the presence of partial verification of the disease. Computational Statistics, 24, 695-718.

Roldán-Nofuentes, J.A., Luna del Castillo, J.D., Montero-Alonso, M.A. (2009). Confidence intervals of weighted kappa coefficient of a binary diagnostic test. Communications in Statistics - Simulation and Computation, 38, 1562-1578.

Roldán-Nofuentes, J.A., Luna del Castillo, J.D. and Montero-Alonso, M.A. (2012). Global hypothesis test to simultaneously compare the predictive values of two binary diagnostic tests. Computational Statistics and Data Analysis, 56, 1161-1173.

Roldán-Nofuentes, J.A., Sidaty-Regad, S.B. (2019). Recommended methods to compare the accuracy of two binary diagnostic tests subject to a paired design. Journal of Statistical Computational and Simulation, 89, 2621-2644.

Roldán-Nofuentes, J.A., Sidaty-Regad S.B. (2020). Comparison of the likelihood ratios of two diagnostic tests subject to a paired design: confidence intervals and sample size. Revstat Statistical Journal. Accepted, in press.

Roldán-Nofuentes, J.A., Sidaty-Regad S.B. (2020). Asymptotic confidence intervals for the difference and the ratio of the weighted kappa coefficients of two diagnostic tests subject to a paired design. Revstat Statistical Journal. Accepted, in press.

Rubin, D. (1976). Inference and missing data. Biometrika, 4, 73-89.

Shao, J. Tu, D. (1995). The Jackknife and Bootstrap. Springer, New York.

Sox H.C., Blatt M.A., Higgins M.C., Marton K.I., (1989). Medical decision making. Butterworths-Heinemann, Boston.

Suissa, S., Shuster, J.J. (1991). The 2×2 matched-pairs trial: exact unconditional design and analysis. Biometrics, 47, 361-372.

Vacek, P.M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics, 41, 959-968.

Vacek, P.M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics, 41, 959-968.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society, 5, 426-482.

Weiner, D. A., Ryan, T. J., McCabe, C. H., Kennedy, J. W., Schloss, M., Tristani, F., Chaitman, B. R., Fisher, L. D. (1979). Correlations among history of angina, ST-segment and prevalence of coronary artery disease in the coronary artery surgery study (CASS). New England Journal of Medicine, 301, 230-235.

Youden, W.J. (1950). Index for rating diagnostic tests. Cancer, 3, 32-35.

Zhou, X.H. (1998). Comparing accuracies of two screening tests in a two-phase study for dementia. Journal of the Royal Statistical Society, Series C Applied Statistics, 47, 135-147.

Zhou, X.H., Obuchowski, N.A., McClish, D.K. (2011). Statistical Methods in Diagnostic Medicine (Second Edition). John Wiley & Sons, New Jersey.

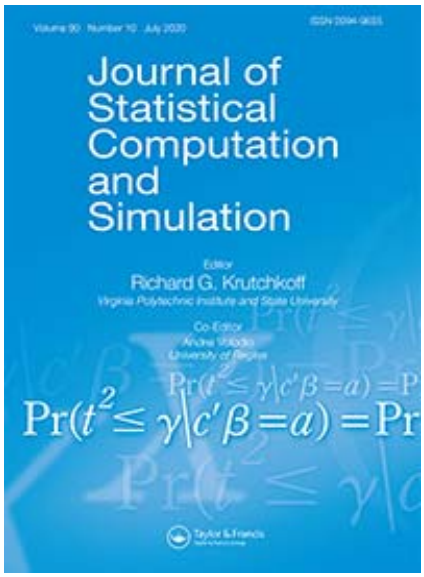
APPENDICES

APPENDIX I

Recommended methods to compare the accuracy of two binary diagnostic tests subject to a paired design

Roldán-Nofuentes, J.A., Sidaty-Regad S.B. (2019). Recommended methods to compare the accuracy of two binary diagnostic tests subject to a paired design. Journal of Statistical Computation and Simulation, 89, 2621-2644. DOI: 10.1080/00949655.2019.1628234.

Category: Statistics and Probability. JCR 2019: 0.918. Rank: 75/124. Quartile: Q3.



Abstract

This manuscript assesses different methods to compare the sensitivities and the specificities of two diagnostic tests. It studies the comparison conditioning on the disease status and testing each hypothesis test individually to an α error, a global hypothesis test to simultaneously compare the parameters, and the individual comparison of the parameters along with the multiple comparison methods of Bonferroni and Holm. Simulation experiments were carried out to study the global type I errors and the global powers of the different methods, providing some general rules of application. When the prevalence is small and the sample size is small or moderate, it is recommended to compare the parameters individually along with the method of Bonferroni or Holm; in any other situation, it is recommended to compare the parameters simultaneously through a global test. The results were applied to an example of the diagnosis of coronary disease.

Keywords: Binary diagnostic test, global hypothesis test, sensitivity, specificity.

Mathematics Subject Classification: 62P10, 6207.

1. Introduction

A diagnostic test is a medical test that is applied to an individual to determine the presence or absence of a disease. When the result of a diagnostic test is positive or negative, the diagnostic test is called a binary diagnostic test (*BDT*). The mammography for the diagnosis of breast cancer and the stress test for the diagnosis of coronary disease are two examples of *BDTs*. The accuracy of a *BDT* is assessed in relation to a gold standard (*GS*), which is a medical test used to objectively diagnose the presence or absence of the disease. From this point on, it is assumed that the *GS* is an error-free test, and therefore it objectively determines if an individual does or does not have the disease. A biopsy for breast cancer and an angiography for coronary disease are two examples of the *GS*. The accuracy of a *BDT* is measured in terms of two fundamental parameters: the sensitivity and the specificity. The sensitivity (*Se*) is the probability of the *BDT* result being positive when the individual has the disease (positive *GS*), and the specificity (*Sp*) is the probability of the *BDT* result being negative when the individual does not have the disease (negative *GS*). Both parameters only depend on the intrinsic ability of the *BDT* to distinguish between individuals with and without the disease.

The comparison of the accuracy of two *BDTs* in relation to a *GS* is an important topic in the study of statistical methods for clinical diagnosis, and consists of comparing the two sensitivities and the two specificities. This comparison can be made subject to two types of sample designs: unpaired design and paired design [1, 2]. Unpaired design consists of applying a *BDT* to a sample of n_1 individuals and the other *BDT* to a sample of n_2 individuals. The paired design consists of applying the two *BDTs* to all the individuals in a random sample sized *n*. In both types of designs, the disease status is known for all the individuals through the application of a *GS*. In unpaired design, the comparison of the two sensitivities (specificities) consists of solving the hypothesis test $H_0 : Se_1 = Se_2$ vs $H_1 : Se_1 \neq Se_2$ ($H_0 : Sp_1 = Sp_2$ vs $H_1 : Sp_1 \neq Sp_2$) to an α error applying a method to compare two independent binomial proportions. In paired design, the problem is traditionally solved conditioning on the disease status and applying a comparison test of two paired binomial proportions. Thus, the comparison of the two sensitivities (specificities) is made conditioning on the individuals with (without) the disease and solving the test $H_0 : Se_1 = Se_2$ vs $H_1 : Se_1 \neq Se_2$ ($H_0 : Sp_1 = Sp_2$ vs $H_1 : Sp_1 \neq Sp_2$) to an α error applying a comparison test of two paired binomial

proportions, e.g. the McNemar test. Therefore, the two sensitivities and the two specificities are compared independently solving each test, $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$, to an α error.

When the two *BDTs* and the *GS* are applied to all the individuals in a sample sized *n*, an alternative method to the traditional one consists of comparing the two sensitivities and the specificities simultaneously, i.e. performing the global hypothesis test $H_0 : (Se_1 = Se_2 \text{ and } Sp_1 = Sp_2)$ vs $H_1 : (Se_1 \neq Se_2 \text{ and/or } Sp_1 \neq Sp_2)$. Lachenbruch and Lynch [3] studied this hypothesis test succinctly, proposing two statistics for the test, one based on the likelihood ratio test and another on the sum of the McNemar tests of each individual test.

This manuscript is motivated by a study [4] in which the accuracy of two *BDTs* (dobutamine echocardiography and myocardial perfusion scintigraphy) is compared, using the coronary angiography as the *GS*. Conditioning on the individuals with the disease and testing the test $H_0 : Se_1 = Se_2$ vs $H_1 : Se_1 \neq Se_2$ to an error $\alpha = 5\%$ applying the McNemar tests, the equality of the two sensitivities is rejected. Testing the test $H_0 : Sp_1 = Sp_2$ vs $H_1 : Sp_1 \neq Sp_2$ to the same error applying the same method the equality of the two specificities is not rejected. Nevertheless, when performing the global test $H_0 : (Se_1 = Se_2 \text{ and } Sp_1 = Sp_2)$ vs $H_1 : (Se_1 \neq Se_2 \text{ and/or } Sp_1 \neq Sp_2)$ to an error $\alpha = 5\%$ applying the results obtained by Lachenbruch and Lynch [3], the null hypothesis of equality of the two sensitivities and of the two specificities is not rejected. In Section 4, this example is discussed. The conclusions obtained (which are partly contradictory) in this example have motivated to study the comparison of the sensitivities and the specificities of two *BDTs* subject to a paired design using the classic method (the individual tests) and the alternative one (the global test), as well as an attempt to determine which methods should be applied and under what conditions.

Section 2 reviews and proposes different methods to compare the two sensitivities and the two specificities individually and simultaneously. In Section 3, simulation experiments are carried out to study the global type I errors and the global powers of the different methods described in Section 2, and some general rules of application are given for the methods. These rules of application are based on the disease prevalence and the sample size. In general terms, when the prevalence is very small or small and

the sample size is small or moderate, it is recommendable to compare the two sensitivities and the two specificities individually along with the Bonferroni or Holm method, i.e. solving the tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ individual along with the Bonferroni (or Holm) method to an α error; in any other situation, it is recommendable to compare the two sensitivities and the two specificities simultaneously, i.e. solving the global test $H_0 : \left(Se_1 = Se_2 \text{ and } Sp_1 = Sp_2\right)$ to an α error. In Section 4, the results obtained are applied to the example of the diagnosis of coronary artery disease, and in Section 5 the results obtained are discussed.

2. Methods to compare the accuracy

Let us consider two *BDTs* and one *GS* that are applied to all of the individuals in a random sample of *n* individuals. Let T_h be the random variable that models the result of the *h*th *BDT*, in such a way that $T_h = 1$ when the result of the *BDT* is positive and $T_h = 0$ when it is negative. Let the random variable *D* that models the result of the *GS*: $D = 1$ when the individual has the disease and $D = 0$ when this is not the case. The application of the two *BDTs* and of the *GS* to all the individuals in the sample leads to the frequencies in Table 1. In this Section, we present different methods to compare the two sensitivities and the two specificities, considering that the hypothesis tests are always two-tailed.

2.1. Individual hypothesis tests

The sensitivities are compared in the individuals with the disease $\left(D = 1\right)$ and testing the hypothesis test $H_0 : Se_1 = Se_2$ vs $H_1 : Se_1 \neq Se_2$ to an α error. By conditioning on $D = 1$ the sample $\left(s_{11}, s_{10}, s_{01}, s_{00}\right)$ is the product of a multinomial distribution with probabilities $p_{ij} = P\left(T_1 = i, T_2 = j \middle| D = 1\right)$, with $\sum_{ij} p_{ij} = 1$. Using the conditional dependence model of Vacek [5], the probabilities p_{ij} are written as

$$p_{ij} = Se_1^i \left(1 - Se_1\right)^{1-i} Se_2^j \left(1 - Se_2\right)^{1-j} + \lambda_{ij}\varepsilon_1 ,$$

where $\lambda_{ij}=1$ if $i=j$ and $\lambda_{ij}=-1$ if $i\neq j$, and ε_1 is the covariance between the two *BDTs* when $D=1$, verifying that $0\leq\varepsilon_1\leq Min\{Se_1(1-Se_2),Se_2(1-Se_1)\}$. If $\varepsilon_1=0$ then the two *BDTs* are conditionally independent when $D=1$, an assumption that is not realistic, and therefore in practice $\varepsilon_1>0$. In terms of the probabilities p_{ij}, the sensitivities are expressed as $Se_1=p_{11}+p_{10}$ and $Se_2=p_{11}+p_{01}$, with $\hat{p}_{ij}=s_{ij}/s$. In this situation, the hypothesis test of equality of the sensitivities is equivalent to test

$$H_0: p_{10}=p_{01} \text{ vs } H_1: p_{10}\neq p_{01}. \quad (1)$$

Table 1. Observed frequencies and probabilities when two *BDTs* are compared in relation to a *GS* subject to a paired design.

Observed frequencies					
	$T_1=1$		$T_1=0$		
	$T_2=1$	$T_2=0$	$T_2=1$	$T_2=0$	Total
$D=1$	s_{11}	s_{10}	s_{01}	s_{00}	s
$D=0$	r_{11}	r_{10}	r_{01}	r_{00}	r
Total	$s_{11}+r_{11}$	$s_{10}+r_{10}$	$s_{01}+r_{01}$	$s_{00}+r_{00}$	n

In a similar way, the specificities are compared conditioning on the individuals without the disease $(D=0)$ and testing the test $H_0: Sp_1=Sp_2$ vs $H_1: Sp_1\neq Sp_2$ to an α error. Conditioning on $D=0$ the sample $(r_{11},r_{10},r_{01},r_{00})$ is the product of a multinomial distribution with probabilities $q_{ij}=P(T_1=i,T_2=j|D=0)$, with $\sum_{ij}q_{ij}=1$. Using the model of Vacek [5] again it is obtained that

$$q_{ij}=Sp_1^{1-i}(1-Sp_1)^i Sp_2^{1-j}(1-Sp_2)^j+\lambda_{ij}\varepsilon_0,$$

where ε_0 is the covariance between the two *BDTs* when $D=0$, verifying that $0\leq\varepsilon_0\leq Min\{Sp_1(1-Sp_2),Sp_2(1-Sp_1)\}$. The same as for ε_1, in practice $\varepsilon_0>0$. In terms of q_{ij}, the specificities are expressed as $Sp_1=q_{01}+q_{00}$ and $Sp_2=q_{10}+q_{00}$, with $\hat{q}_{ij}=r_{ij}/r$. Therefore, the comparison test of the two specificities is equivalent to the test

$$H_0: q_{01}=q_{10} \text{ vs } H_1: q_{01}\neq q_{10}. \quad (2)$$

The hypothesis tests (1) and (2) are solved applying a test to compare two paired binomial proportions. Then eleven methods (two exact methods, one quasi-exact and eight asymptotic ones) are described to compare two paired binomial proportions, in the context studied here (sensitivities and specificities) and using the notation in Table 1. These methods are:

1) Conditional exact test (*CET*)

2) Conditional Mid-p test (*Midp*)

3) McNemar test (*MT*)

4) McNemar test with continuity correction (*MTcc*)

5) Modified McNemar test (*MMT*)

6) Wald test (*WT*)

7) Modified Wald test (*MWT*)

8) Likelihood ratio test (*LRT*)

9) Unconditional exact Test (*UET*)

10) Unconditional McNemar test (*UMT*)

11) Unconditional likelihood ratio test (*ULRT*)

In Appendix *A* there is a detailed explanation of each method. The comparison of the asymptotic behaviour (in terms of type I error and power) of methods to compare two paired binomial proportions has been the subject of several studies. May and Johnson [6] compared the methods *CET*, *Midp*, *MT*, *MTcc*, *MMT*, *WT*, *MWT* and *LRT*, and they recommended using the *Midp*, the *MT* or the *MWT* when the sum of discordant frequencies is lower than 40. Fagerland et al [7] compared the methods *CET*, *Midp*, *MT*, *MTcc* and *UET*, and they recommended using the *MT* and the *Midp*.

2.2. Global hypothesis tests

Another way of comparing the accuracy of two *BDTs* consist of comparing the two sensitivities and the two specificities simultaneously, i.e. performing the global hypothesis test

$$H_0 : \left(Se_1 = Se_2 \text{ and } Sp_1 = Sp_2\right) \text{ vs } H_1 : \left(Se_1 \neq Se_2 \text{ and/or } Sp_1 \neq Sp_2\right). \qquad (3)$$

The solution of this hypothesis test is achieved without conditioning on the disease status. Therefore, this hypothesis test is solved based on the sample design used, in which the researcher has only set the size (n) of the only sample in the study. The observed frequencies given in Table 1, $\mathbf{x}=\left(s_{11},s_{10},s_{01},s_{00},r_{11},r_{10},r_{01},r_{00}\right)^T$, are the product of a multinomial distribution. Let $\boldsymbol{\omega}=\left(\phi_{11},\phi_{10},\phi_{01},\phi_{00},\varphi_{11},\varphi_{10},\varphi_{01},\varphi_{00}\right)^T$ be the vector of probabilities of the multinomial distribution, where $\phi_{ij}=P\left(D=1,T_1=i,T_2=j\right)$ and $\varphi_{ij}=P\left(D=0,T_1=i,T_2=j\right)$ with $i,j=0,1$. It is verified that

$$\phi_{ij}=\pi\left[Se_1^i\left(1-Se_1\right)^{1-i}Se_2^j\left(1-Se_2\right)^{1-j}+\lambda_{ij}\varepsilon_1\right] \tag{4}$$

and

$$\varphi_{ij}=\overline{\pi}\left[Sp_1^{1-i}\left(1-Sp_1\right)^{i}Sp_2^{1-j}\left(1-Sp_2\right)^{j}+\lambda_{ij}\varepsilon_0\right], \tag{5}$$

where $\pi=P\left(D=1\right)=\sum_{ij}\phi_{ij}$ is the disease prevalence and $\overline{\pi}=1-\pi=\sum_{ij}\varphi_{ij}$. In terms of the probabilities ϕ_{ij} and φ_{ij}, the sensitivities and the specificities are expressed as $Se_1=\left(\phi_{11}+\phi_{10}\right)/\pi$, $Se_2=\left(\phi_{11}+\phi_{01}\right)/\pi$, $Sp_1=\left(\varphi_{01}+\varphi_{00}\right)/\overline{\pi}$ and $Sp_2=\left(\varphi_{10}+\varphi_{00}\right)/\overline{\pi}$. As $\hat{\phi}_{ij}=s_{ij}/n$ and $\hat{\varphi}_{ij}=r_{ij}/n$, the estimators of the sensitivities and of the specificities are $\hat{Se}_1=\left(s_{11}+s_{10}\right)/s$, $\hat{Se}_2=\left(s_{11}+s_{01}\right)/n$, $\hat{Sp}_1=\left(r_{01}+r_{00}\right)/r$ and $\hat{Sp}_2=\left(r_{10}+r_{00}\right)/r$, and the estimator of the prevalence is $\hat{\pi}=s/n$. Applying the delta method, its corresponding estimated variances are $\hat{V}ar\left(\hat{Se}_h\right)=\hat{Se}_h\left(1-\hat{Se}_h\right)/\left(n\hat{\pi}\right)$, $\hat{V}ar\left(\hat{Sp}_h\right)=\hat{Sp}_h\left(1-\hat{Sp}_h\right)/\left(n\hat{\overline{\pi}}\right)$ and $\hat{V}ar\left(\hat{\pi}\right)=\hat{\pi}\hat{\overline{\pi}}/n$, with $h=1,2$. The global hypothesis test (3) is equivalent to the test

$$H_0:\left(\phi_{10}=\phi_{01}\ \text{ and }\ \varphi_{01}=\varphi_{10}\right)\ \text{ vs }\ H_1:\left(\phi_{10}\neq\phi_{01}\ \text{ and/or }\ \varphi_{01}=\varphi_{10}\right). \tag{6}$$

Subject to the null hypothesis it is verified that $\phi_{10}=\phi_{01}=\left(\phi_{10}+\phi_{01}\right)/2$ and that $\varphi_{01}=\varphi_{10}=\left(\varphi_{01}+\varphi_{10}\right)/2$, and its estimators are $\hat{\phi}_{10}=\hat{\phi}_{01}=\left(s_{10}+s_{01}\right)/\left(2n\right)$ and $\hat{\varphi}_{01}=\hat{\varphi}_{10}=\left(r_{01}+r_{10}\right)/\left(2n\right)$. It is obvious that the hypothesis test (6) is also equivalent to the test $H_0:\left(p_{10}=p_{01}\ \text{ and }\ q_{01}=q_{10}\right)$ vs $H_1:\left(p_{10}\neq p_{01}\ \text{ and/or }\ q_{01}\neq q_{10}\right)$, since $\phi_{ij}=\pi p_{ij}$ and $\varphi_{ij}=\overline{\pi}q_{ij}$. Nevertheless, in the inconditional model the probabilities

involved are ϕ_{ij} and φ_{ij}, and not the probabilities p_{ij} and q_{ij} obtained under conditioning.

Lachenbruch and Lynch [3] briefly studied the global hypothesis test, proposing two statistics: one obtained through the likelihood ratio test and another obtained by adding the two McNemar statistics (the sum of equations (19) and (20), see Appendix *A*). Three statistics are then presented to solve the global hypothesis test. Firstly, we present the likelihood ratio test statistics of Lachenbruch and Lynch. Secondly, we obtain a statistic applying the Rao's score test [8], demonstrating that this statistic is the same as the second statistic (the sum of the two McNemar statistics) obtained by Lachenbruch and Lynch. Finally, the global test is solved by applying the Wald test [9].

2.2.1. Likelihood ratio test (*LRT*)

Lachenbruch and Lynch [3] solved the global hypothesis test (6) by applying the likelihood ratio test. The likelihood function of the data is $L(\boldsymbol{\omega};\mathbf{x}) = k\phi_{11}^{s_{11}}\phi_{10}^{s_{10}}\phi_{01}^{s_{01}}\phi_{00}^{s_{00}}\varphi_{11}^{r_{11}}\varphi_{10}^{r_{10}}\varphi_{01}^{r_{01}}\varphi_{00}^{r_{00}}$, where $k = n!\Big/\left[\left(\prod_{ij} s_{ij}!\right)\left(\prod_{ij} r_{ij}!\right)\right]$. Subject to the null hypothesis, this function is $L_0(\boldsymbol{\omega};\mathbf{x}) = k\phi_{11}^{s_{11}}\phi_{10}^{s_{10}+s_{01}}\phi_{00}^{s_{00}}\varphi_{11}^{r_{11}}\varphi_{01}^{r_{10}+r_{01}}\varphi_{00}^{r_{00}}$. Applying the likelihood ratio test and performing algebraic operations, it is obtained that the likelihood ratio statistic is

$$\chi^2_{LR} = 2\left[s_{10}\ln\left(\frac{2s_{10}}{s_{10}+s_{01}}\right) + s_{01}\ln\left(\frac{2s_{01}}{s_{10}+s_{01}}\right) + r_{10}\ln\left(\frac{2r_{10}}{r_{10}+r_{01}}\right) + r_{01}\ln\left(\frac{2r_{01}}{r_{10}+r_{01}}\right)\right], \quad (7)$$

which is distributed asymptotically according to a chi-square distribution with two degrees of freedom when the null hypothesis is true. This statistic is the sum of the statistics obtained by applying the likelihood ratio test to compare the two sensitivities and the two specificities independently, i.e. the sum of expressions (27) and (28) (see Appendix *A*), and in which only the frequencies of the discordant pairs intervene.

2.2.2. Rao's score test (*RST*)

The global hypothesis test can be solved applying Rao's score test [8]. The log-likelihood function of the data in Table 1 is $l(\boldsymbol{\omega};\mathbf{x}) = \ln k + \sum_{ij} s_{ij} \ln(\phi_{ij}) + \sum_{ij} r_{ij} \ln(\varphi_{ij})$. Let $\mathbf{U}(\boldsymbol{\omega})$ be a vector whose components are the derivates of $l(\boldsymbol{\omega};\mathbf{x})$ with respect to $\boldsymbol{\omega}$ i.e.

$$\mathbf{U}(\boldsymbol{\omega}) = \left(\frac{s_{11}}{\phi_{11}}, \frac{s_{10}}{\phi_{10}}, \frac{s_{01}}{\phi_{01}}, \frac{s_{00}}{\phi_{00}}, \frac{r_{11}}{\varphi_{11}}, \frac{r_{10}}{\varphi_{10}}, \frac{r_{01}}{\varphi_{01}}, \frac{r_{00}}{\varphi_{00}} \right)^T .$$

As $\boldsymbol{\omega}$ is the probability vector of a multinomial distribution, the variance-covariance matrix of $\hat{\boldsymbol{\omega}}$ is $\Sigma_{\hat{\omega}} = \left(diag(\boldsymbol{\omega}) - \boldsymbol{\omega}\boldsymbol{\omega}^T \right) / n$. The estimator of $\boldsymbol{\omega}$ subject to the null hypothesis is

$$\hat{\boldsymbol{\omega}}_0 = \left(\frac{s_{11}}{n}, \frac{s_{10}+s_{01}}{2n}, \frac{s_{10}+s_{01}}{2n}, \frac{s_{00}}{n}, \frac{r_{11}}{n}, \frac{r_{10}+r_{01}}{2n}, \frac{r_{10}+r_{01}}{2n}, \frac{r_{00}}{n} \right)^T .$$

Substituting in $\mathbf{U}(\boldsymbol{\omega})$ each parameter ϕ_{ij} and φ_{ij} with its corresponding estimator subject to the null hypothesis given in $\hat{\boldsymbol{\omega}}_0$ it is obtained that

$$\mathbf{U}(\hat{\boldsymbol{\omega}}_0) = \left(n, \frac{2s_{10}n}{s_{10}+s_{01}}, \frac{2s_{01}n}{s_{10}+s_{01}}, n, n, \frac{2r_{10}n}{r_{10}+r_{01}}, \frac{2r_{01}n}{r_{10}+r_{01}}, n \right)^T .$$

Finally, the Rao's score test statistic for the global hypothesis test is

$$\chi^2_R = \mathbf{U}^T(\hat{\boldsymbol{\omega}}_0) \hat{\Sigma}_{\hat{\omega}_0} \mathbf{U}(\hat{\boldsymbol{\omega}}_0),$$

which is distributed asymptotically according to a chi-square distribution with two degrees of freedom when the null hypothesis is true, and where $\hat{\Sigma}_{\hat{\omega}_0}$ is the matrix $\Sigma_{\hat{\omega}}$ assessed in $\hat{\boldsymbol{\omega}}_0$. Performing algebraic operations, it is obtained that the Rao's score test statistic is

$$\chi^2_R = \frac{(s_{10}-s_{01})^2}{s_{10}+s_{01}} + \frac{(r_{10}-r_{01})^2}{r_{10}+r_{01}}. \tag{8}$$

This statistic is the sum of the statistics obtained applying the McNemar test to compare the two sensitivities and the two specificities independently, i.e. the sum of expressions (19) and (20) (see Appendix *A*), and therefore it is in line with the one proposed by

Lachenbruch and Lynch [3] based on the construction of two separate 2×2 tables obtained from Table 1 using the results of Hamdan et al [10].

2.2.3. Wald test (*WT*)

Another way of performing the global test is through the classic Wald test [9]. Let $\boldsymbol{\theta}=\left(Se_1,Se_2,Sp_1,Sp_2\right)^T$ be the vector whose components are the sensitivities and the specificities. As the sensitivities and the specificities are written in terms of the probabilities of $\boldsymbol{\omega}$, applying the delta method, the variance-covariance matrix of vector $\hat{\boldsymbol{\theta}}$ is $\sum_{\hat{\boldsymbol{\theta}}}=\left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\pi}}\right)\sum_{\hat{\boldsymbol{\pi}}}\left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\pi}}\right)^T$. The hypothesis test (3) is equivalent to

$$H_0:\boldsymbol{\psi}\boldsymbol{\theta}=0 \text{ vs } H_1:\boldsymbol{\psi}\boldsymbol{\theta}\neq 0. \tag{9}$$

where $\boldsymbol{\psi}$ is a complete range matrix sized 2×4 defined as

$$\boldsymbol{\psi}=\begin{pmatrix}1 & -1 & 0 & 0\\ 0 & 0 & 1 & -1\end{pmatrix}.$$

Through the Multivariate Central Limit Theorem, it is verified that $\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\xrightarrow[n\to\infty]{} N\left(\mathbf{0},\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)$. Finally, the Wald test statistic for the hypothesis test (9) is

$$\chi^2_W=\hat{\boldsymbol{\theta}}^T\boldsymbol{\psi}^T\left(\boldsymbol{\psi}\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}\boldsymbol{\psi}^T\right)^{-1}\boldsymbol{\psi}\hat{\boldsymbol{\theta}}, \tag{10}$$

which is distributed asymptotically according to a chi-square distribution with two degrees of freedom when the null hypothesis is true. Performing algebraic operations, it is obtained that the Wald test statistic is

$$\chi^2_W=\frac{s\left(s_{10}-s_{01}\right)^2}{4s_{10}s_{01}+\left(s_{11}+s_{00}\right)\left(s_{10}+s_{01}\right)}+\frac{r\left(r_{10}-r_{01}\right)^2}{4r_{10}r_{01}+\left(r_{11}+r_{00}\right)\left(r_{10}+r_{01}\right)}, \tag{11}$$

which is also the sum of the statistics obtained applying the Wald test to compare the two sensitivities and the two specificities independently, i.e. the sum of expressions (24) and (25) (see Appendix *A*). In this statistic, all the observed frequencies are used.

2.3. Other alternative methods

The methods described in Section 2.1 compare the two sensitivities and the two specificities individually, testing each one of the hypothesis tests, $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$, to an α error, and the methods described in Section 2.2 compare the parameters simultaneously to an α error. Another alternative consists of comparing the sensitivities and the specificities individually (through the methods described in 2.1) and then applying a multiple comparison method to an α error. As multiple comparison methods, we propose the application of the Bonferroni method [11] and the Holm method [12]. The Bonferroni method is a classic one in post-hoc comparisons and, in the situation studied here, it consists of testing each individual test to an $\alpha/2$ error to thereby control the global α error. The Holm method is a less conservative post-hoc method than the Bonferroni one. Let p_1 and p_2 be the p-values obtained in each individual hypothesis test and let us suppose that $p_1 \leq p_2$, the Holm method consists of the following two steps:

1) If $p_1 > \alpha/2$ then none of the two null hypothesis $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ are rejected. In the opposite case $\left(p_1 \leq \alpha/2\right)$, the null hypothesis corresponding to that hypothesis test is rejected and we go on to the next step.

2) If $p_2 > \alpha$ then the null hypothesis corresponding to that hypothesis test is not rejected. In the opposite case $\left(p_2 \leq \alpha\right)$ that null hypothesis is rejected and the process ends.

3. Simulation experiments

Monte Carlo simulation experiments were carried out to study the global type I errors and the global powers of the methods described in Section 2. For the methods described in Sections 2.1 and 2.3, the objective is to study the global type I error and the global power of each method when comparing the two sensitivities and the two specificities. Thus, the global type I error is the one made when we reject the hypothesis $H_0 : Se_1 = Se_2$ and/or the hypothesis $H_0 : Sp_1 = Sp_2$ when both are true, whether or not each test is to an α error or applying the Bonferroni method (or the Holm method). The

argument for the global power is similar to that of the global type I error. Therefore, the objective is not to study the type I error and the power of each individual hypothesis test, $H_0 : Se_1 = Se_2$ or $H_0 : Sp_1 = Sp_2$, which is a question that has been studied by other authors [6, 7]. For the methods described in Section 2.2, the objective is to study the type I error and the power of the global test $H_0 : \left(Se_1 = Se_2 \text{ and } Sp_1 = Sp_2\right)$. The experiments consisted of generating $N = 10,000$ random samples with multinomial distributions of different sizes, $n = \{50,100,200,300,400,500,1000,2000\}$, and whose probabilities were calculated from equations (4) and (5). As sensitivity and specificity, the values $\{0.70, 0.75, ..., 0.95\}$ were taken for each *BDT*, which are frequent values in clinical practice. As disease prevalence, the values $\pi = \{5\%, 10\%, 25\%, 50\%\}$ were taken, which can be considered to be a very small (5%), small (10%), moderate (25%) and high value (50%) respectively. For the covariances ε_1 and ε_0 we took for both 25% (a low value), 50% (intermediate value), 75% (high value) and 90% (very high value) of its maximum value, i.e. $\varepsilon_1 = f \times Min\{Se_1(1-Se_2), Se_2(1-Se_1)\}$ and $\varepsilon_0 = f \times Min\{Sp_1(1-Sp_2), Sp_2(1-Sp_1)\}$ where $f = \{0.25, 0.50, 0.75, 0.90\}$. The N multinomial random samples were generated independently, and in such a way that in all of them it was possible to apply all the methods described in Section 2. Thus, for example, if a sample has a discordant frequency equal to zero, then it is not possible to apply the likelihood ratio test, and in this situation that sample has been discarded and another one has been generated instead until the completion of *N* samples. This situation has mainly occurred when the sample size has been small $(n = 50)$ or moderate $(n = 100)$. As α error, the value 5% has been taken. For the *N* samples generated from the same multinomial distribution, we calculated the global type I error or the global power, according to the situation of each one of the different methods to compare the two sensitivities and the two specificities, i.e.: 1) $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ each one to an error $\alpha = 5\%$, 2) $H_0 : \left(Se_1 = Se_2 \text{ and } Sp_1 = Sp_2\right)$ to an error $\alpha = 5\%$, 3) $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ and application of the Bonferroni method to an error $\alpha = 5\%$, and 4) $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ and application of the Holm method to an error $\alpha = 5\%$.

3.1. Global type I errors

For each one of the previous methods (1, 2, 3 and 4), the global type I errors were calculated in the different scenarios considered. For Method 1, a count was made of the number of samples in which $H_1 : Se_1 \neq Se_2$ and/or $H_1 : Sp_1 \neq Sp_2$ were accepted (when $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ are true), each one to an error $\alpha = 5\%$, and then we calculated the global type I error dividing the value obtained by *N*. For Methods 3 and 4, the global type I errors were calculated in a similar way but applying the Bonferroni method and the Holm method respectively. Regarding Method 2, a count was made of the number of samples in which $H_1 : \left(Se_1 \neq Se_2 \text{ and/or } Sp_1 \neq Sp_2\right)$ was accepted when $H_0 : \left(Se_1 = Se_2 \text{ and } Sp_1 = Sp_2\right)$ was true, and that value was divided by *N*. The comparison of the global type I errors of the different methods was made using the following criteria. For those methods based on exact tests it was considered that the method performs well when its global type I error is $\leq 5\%$; if the global type I error is $> 5\%$ then the global type I error goes too far above the nominal error. For those methods based on approximate tests, it was considered that the method shows good global type I error performance when the global type I error fluctuates around the nominal error of 5% without going too far above it. Here it has been considered that the global type I error goes too far above the nominal error when the global type I error is $\geq 7\%$. In Appendix *B* these criteria are justified.

In Tables 2, 3 and 4, we show some of the results obtained for methods 1, 2 and 3 respectively, and for different values of the parameters, indicating in bold that the global type I goes too far above the nominal error. The results for the Holm method are not shown, as they are practically identical to those obtained applying the Bonferroni method. The Monte Carlo standard error of a nominal error of 5% based on 10,000 samples is 0.22%, and the overall average Monte Carlo error obtained is 0.21% (a value very close to 0.22%). Prevalence *p* and the covariances ε_i have an important effect on the type I global error of the methods proposed: the increase in prevalence involves (if the other parameters remain constant) an increase in the global type I error, whereas the increase in ε_1 and/or ε_0 involves (if the other parameters remain constant) a decrease in the global type I error. From the results the following conclusions are obtained:

a) Method 1: $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ each one to an error $\alpha = 5\%$. In general terms, the global type I errors of the methods *CET*, *Midp* and *UET* go too far above the nominal error (global type I errors $> 5\%$) when the sample size is large $(200 \le n \le 500)$ or very large $(n \ge 1000)$ depending on the prevalence and on the covariances. The asymptotic tests *MT*, *MMT*, *WT*, *MWT* and *LRT* clearly go too far above the nominal error (global type I errors $\ge 7\%$), and can even double it, when the sample size is large (the *LRT* can goes too far above the nominal error even with $n = 100$) or very large. The *MTcc* test only goes too far above the nominal error when the sample size is very large. The *UMT* and *ULRT* tests never exceed the nominal error, and their global type I errors are very small (they are not usually over 3%), and, therefore, they are excessively conservative tests (even more than the exact tests when they do not exceed the nominal error). In general terms, when comparing the two sensitivities and the two specificities testing the tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ individually each one to an error $\alpha = 5\%$, the method with the best behaviour when $n \le 500$ is the McNemar test with *cc* (*MTcc*), although for $n = 1000 - 2000$ its global type I error can be clearly $\ge 7\%$ depending on the prevalence and on the covariances ε_i. If the prevalence is very small $(p = 5\%)$, the global type I error of the *MTcc* fluctuates around the nominal error without going too far above it; if the prevalence is higher $(p \ge 10\%)$ the global type I error can easily be above the nominal error. As for the effect of the covariances ε_i, the global type I error of the *MTcc* fluctuates around the nominal error without going too far above it when the covariances take very high values (even when $n \ge 500$); for other values of the covariances, the global type I error of the *MTcc* can go well above the nominal error especially when *n* is very large.

b) Method 2: $H_0 : (Se_1 = Se_2$ and $Sp_1 = Sp_2)$ to an error $\alpha = 5\%$. The global type I errors of *LRT*, *RST* and *WT* methods do not go too far above the nominal error. From a sample size between 200 and 500, depending on the prevalence and on the covariances, the global type I error of each test fluctuates around the nominal error. In general terms, although there is no important difference between their type I errors, the *LRT* presents a global type I error closest to 5%, followed by the *WT* and the *RST*, especially when $n \le 500$. For $n \ge 1000$ the three methods have a very similar global type I error, as these methods are asymptotically first order equivalent. Moreover, these three methods have a

global type I error which is slightly lower than the *MTcc* to an error $\alpha = 5\%$ when the prevalence is very small $(p = 5\%)$; for $p \geq 10\%$, the global type I errors of these three methods are nearer to 5% than the global type I error of the del *MTcc* (Table 1: *MTcc*) with $\alpha = 5\%$ (when the latter does not go too far above the nominal error). Regarding the effect of the covariances ε_i, their increase involves (if the other parameters remain constant) a decrease in the global type I error, especially when $n \leq 500$. When the covariances are very high, the global tests *LRT* and *WT* have a global type I error whose fluctuations are slightly better around the nominal error than the *MTcc* with $\alpha = 5\%$.

c) Method 3: $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ and application of the Bonferroni method to an error $\alpha = 5\%$. The global type I errors of the methods *Midp* and *UET* can be $> 5\%$ when the sample size is very large $(n \geq 1000)$, in the rest of the situations their corresponding global type I errors are $\leq 5\%$. The rest of the methods have global type I errors that do not exceed the nominal error. As in case a), the *UMT* and *ULRT* methods are very conservative. In general terms, applying the Bonferroni method, the *MT*, *WT* and *LRT* methods have a global type I error closer to 5% than the rest of the methods. Furthermore, in very general terms, there is no important difference between the global type I errors of the *WT* and of the *LRT* methods along with Bonferroni and the global type I error of the method based on the McNemar test with *cc* to an error $\alpha = 5\%$ (when this method has a global type I error $< 7\%$). The *MT* along with Bonferroni has a global type I error which is slightly lower than that of the McNemar test with *cc* to an error $\alpha = 5\%$. Moreover, although there is no important difference between the global type I errors of the individual methods (*MT*, *WT* and *LRT*) along with Bonferroni and the global type I errors of the methods to solve the global test $H_0 : (Se_1 = Se_2$ and $Sp_1 = Sp_2)$, in very general terms when $n \leq 100$ the global tests are slightly more conservative than the individual tests along with Bonferroni.

d) Method 4: $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ and application of the Holm method to an error $\alpha = 5\%$. The results are practically identical to those obtained with Method 3.

Table 2. Global type I errors (%) of the individual tests with $\alpha = 5\%$.

$Se_1 = Se_2 = 0.80$, $Sp_1 = Sp_2 = 0.90$, $p = 5\%$, $\varepsilon_1 = 0.04$, $\varepsilon_0 = 0.0225$ (25% of the maximum value)

n	*CET*	*Midp*	*MT*	*MTcc*	*MMT*	*WT*	*MWT*	*LRT*	*UET*	*UMT*	*ULRT*
50	0.47	1.32	1.32	0.47	1.30	3.06	1.32	3.04	1.32	0.05	0.15
100	1.67	2.97	3.06	1.46	2.48	3.72	3.01	4.09	3.06	0.40	0.48
200	3.03	4.49	5.39	2.95	4.45	5.59	4.75	5.58	**5.39**	0.62	0.77
300	3.33	4.90	5.11	3.23	4.70	5.83	5.09	6.01	**5.11**	0.57	0.70
400	3.81	5.43	5.44	3.71	5.37	6.79	5.52	6.80	**5.44**	0.71	0.81
500	4.24	6.20	6.26	4.10	6.05	**8.32**	6.29	**8.32**	**6.26**	0.92	1.06
1000	**6.14**	**8.84**	**8.85**	5.99	**8.49**	**10.84**	**8.86**	**10.93**	**8.85**	1.28	1.67
2000	**7.33**	**9.42**	**10.25**	**7.21**	**9.29**	**10.29**	**9.79**	**10.29**	**10.25**	1.72	1.88

$Se_1 = Se_2 = 0.80$, $Sp_1 = Sp_2 = 0.90$, $p = 10\%$, $\varepsilon_1 = 0.08$, $\varepsilon_0 = 0.045$ (50% of the maximum value)

n	*CET*	*Midp*	*MT*	*MTcc*	*MMT*	*WT*	*MWT*	*LRT*	*UET*	*UMT*	*ULRT*
50	0.11	0.42	0.42	0.11	0.42	1.50	0.43	1.48	0.42	0.01	0.02
100	0.77	1.77	1.77	0.75	1.71	2.16	1.80	3.63	1.77	0.14	0.24
200	2.86	4.78	5.04	2.48	4.25	6.02	4.78	6.23	**5.04**	0.41	0.73
300	3.15	5.09	5.87	3.08	4.96	**7.35**	4.97	**7.36**	**5.87**	0.55	0.84
400	3.58	5.89	6.48	3.54	5.83	**8.48**	5.89	**8.74**	**6.48**	0.63	0.93
500	4.51	**7.28**	**7.36**	4.36	**7.08**	**9.16**	**7.28**	**10.13**	**7.36**	0.77	1.03
1000	**6.38**	**9.19**	**9.60**	6.03	**8.56**	**10.20**	**9.18**	**10.52**	**9.60**	1.33	1.45
2000	**7.38**	**9.60**	**10.20**	**7.35**	**9.52**	**10.44**	**9.96**	**10.44**	**10.20**	1.37	1.45

$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.90$, $p = 25\%$, $\varepsilon_1 = 0.045$, $\varepsilon_0 = 0.045$ (50% of the maximum value)

n	*CET*	*Midp*	*MT*	*MTcc*	*MMT*	*WT*	*MWT*	*LRT*	*UET*	*UMT*	*ULRT*
50	0	0.10	0.10	0	0.18	0.48	0.10	0.48	0.10	0	0
100	0.82	1.72	1.72	0.82	2.20	2.00	1.72	4.20	1.72	0.12	0.26
200	2.68	4.82	4.88	2.40	6.42	6.12	4.82	**7.72**	4.84	0.44	1.02
300	3.62	**6.30**	6.92	3.26	**8.46**	**7.12**	5.96	**9.52**	**6.84**	0.58	1.80
400	4.72	**7.46**	**8.24**	4.68	**8.18**	**8.68**	**7.46**	**11.30**	**7.90**	0.92	1.30
500	**5.18**	**8.18**	**8.54**	5.06	**8.70**	**9.26**	**8.18**	**10.94**	**8.44**	0.86	1.44
1000	**6.04**	**8.66**	**9.60**	5.58	**9.10**	**9.60**	**8.32**	**9.80**	**9.60**	1.06	1.54
2000	**7.36**	**9.40**	**9.40**	**7.22**	**9.10**	**9.54**	**9.40**	**9.66**	**9.40**	1.26	1.60

$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.90$, $p = 50\%$, $\varepsilon_1 = 0.0675$, $\varepsilon_0 = 0.0675$ (75% of the max.)

n	*CET*	*Midp*	*MT*	*MTcc*	*MMT*	*WT*	*MWT*	*LRT*	*UET*	*UMT*	*ULRT*
50	0	0	0	0	0	0	0	0	0	0	0
100	0.04	0.08	0.08	0.04	0.36	0.18	0.08	0.38	0.08	0.02	0.04
200	0.28	1.18	1.18	0.28	3.36	1.22	1.18	3.42	1.18	0.02	0.30
300	1.58	3.70	3.70	1.58	4.90	3.92	3.70	**8.26**	3.70	0.08	0.74
400	3.30	**6.28**	6.28	3.26	6.58	6.66	6.12	**11.02**	6.28	0.30	0.96
500	3.82	**7.20**	**7.20**	3.66	**7.42**	**7.20**	6.80	**11.82**	**7.20**	0.48	1.56
1000	**5.74**	**8.58**	**10.14**	5.40	**8.34**	**10.14**	**8.10**	**10.26**	**10.14**	1.06	1.26
2000	**7.54**	**10.30**	**10.32**	**7.20**	**8.86**	**10.32**	**10.30**	**10.62**	**10.32**	0.98	1.10

$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.80$, $p = 25\%$, $\varepsilon_1 = 0.0225$, $\varepsilon_0 = 0.04$ (25% of the maximum value)

n	*CET*	*Midp*	*MT*	*MTcc*	*MMT*	*WT*	*MWT*	*LRT*	*UET*	*UMT*	*ULRT*
50	1.14	2.80	2.80	1.12	2.70	5.42	3.04	5.42	2.80	0.16	0.62
100	3.40	5.00	5.28	2.88	4.34	6.26	5.30	6.42	**5.04**	1.00	1.00
200	3.76	**6.44**	6.74	3.74	6.16	**8.76**	6.64	**9.68**	**6.66**	0.86	0.96
300	**5.36**	**8.38**	**8.40**	5.20	**8.32**	**9.20**	**8.40**	**11.36**	**8.38**	1.28	1.52
400	**6.18**	**8.96**	**9.12**	5.80	**8.24**	**9.92**	**9.08**	**10.76**	**9.10**	1.34	1.66
500	**6.20**	**8.70**	**9.44**	5.64	**8.76**	**9.88**	**8.92**	**10.04**	**9.28**	1.28	1.34
1000	**7.68**	**9.72**	**10.16**	**7.62**	**9.38**	**10.40**	**9.82**	**10.40**	**10.15**	1.72	1.68
2000	**8.12**	**9.56**	**9.64**	**8.02**	**9.38**	**9.90**	**9.56**	**9.90**	**9.64**	1.34	1.54

$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.80$, $p = 50\%$, $\varepsilon_1 = 0.081$, $\varepsilon_0 = 0.144$ (90% of the maximum value)

n	*CET*	*Midp*	*MT*	*MTcc*	*MMT*	*WT*	*MWT*	*LRT*	*UET*	*UMT*	*ULRT*
50	0	0	0	0	0	0	0	0	0	0	0
100	0.01	0.02	0.02	0.01	0.02	0.06	0.02	0.14	0.02	0	0
200	0.03	0.16	0.16	0.03	0.16	0.16	0.16	0.72	0.16	0	0
300	0.19	0.71	0.71	0.19	0.71	0.73	0.71	1.46	0.71	0	0.01
400	0.66	1.85	1.87	0.66	1.87	1.88	1.87	3.89	1.87	0.01	0.06
500	1.22	2.93	2.94	1.22	2.91	2.93	2.91	6.23	2.93	0.04	0.15
1000	4.05	**7.35**	**7.56**	3.72	6.75	**7.56**	6.76	**10.03**	**7.56**	0.41	0.89
2000	**5.90**	**8.83**	**9.62**	5.52	**8.35**	**9.62**	**8.38**	**9.84**	**9.82**	0.90	1.34

CET: Conditional exact test. *Midp*: Mid-p test. *MT*: McNemar test. *MTcc*. McNemar test with cc. *MMT*: Modified McNemar test. *WT*: Wald test. *MWT*: Modified Wald test. *LRT*: Likelihood ratio test. *UET*: Unconditional exact test. *UMT*: Unconditional McNemar test. *ULRT*: Unconditional likelihood ratio test.

Table 3. Global type I errors (%) of the global tests with $\alpha = 5\%$.

$Se_1 = Se_2 = 0.80$, $Sp_1 = Sp_2 = 0.90$						
	$p = 5\%$, $\varepsilon_1 = 0.04$, $\varepsilon_0 = 0.0225$ (25% of the max.)			$p = 10\%$, $\varepsilon_1 = 0.08$, $\varepsilon_0 = 0.045$ (50% of the max.)		
n	*LRT*	*RST*	*WT*	*LRT*	*RST*	*WT*
50	0.45	0.18	0.56	0.13	0.03	0.15
100	1.23	0.94	1.26	0.82	0.53	0.75
200	2.07	1.83	2.13	2.65	1.93	2.59
300	2.61	2.16	2.57	3.12	2.49	3.06
400	3.03	2.71	3.10	3.48	2.79	3.42
500	3.82	2.99	3.91	4.41	3.79	4.08
1000	5.01	4.48	5.10	5.19	4.81	5.12
2000	5.14	4.95	5.23	5.08	5.02	5.11
$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.90$						
	$p = 25\%$, $\varepsilon_1 = 0.045$, $\varepsilon_0 = 0.045$ (50% of the max.)			$p = 50\%$, $\varepsilon_1 = 0.0675$, $\varepsilon_0 = 0.0675$ (75% of the max.)		
n	*LRT*	*RST*	*WT*	*LRT*	*RST*	*WT*
50	0	0	0	0	0	0
100	0.84	0.50	0.72	0.08	0.06	0.06
200	2.86	2.10	2.36	0.58	0.32	0.40
300	3.92	3.10	3.46	2.74	1.74	1.84
400	5.00	4.16	4.52	4.44	3.22	3.38
500	5.08	4.40	4.64	4.94	3.94	4.06
1000	4.82	4.72	4.80	5.22	5.15	5.18
2000	4.92	4.85	4.89	5.10	4.88	4.96
$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.80$						
	$p = 25\%$, $\varepsilon_1 = 0.0225$, $\varepsilon_0 = 0.04$ (25% of the max.)			$p = 50\%$, $\varepsilon_1 = 0.045$, $\varepsilon_0 = 0.08$ (90% of the max.)		
n	*LRT*	*RST*	*WT*	*LRT*	*RST*	*WT*
50	1.16	0.46	1.32	0	0	0
100	2.82	2.30	2.86	0.01	0	0.01
200	3.70	3.10	3.78	0.06	0.02	0.02
300	5.12	4.46	4.94	0.34	0.20	0.20
400	4.86	4.28	4.66	1.04	0.55	0.77
500	5.04	4.56	4.98	1.86	1.14	1.36
1000	5.52	5.44	5.52	4.73	3.71	3.96
2000	5.22	5.18	5.20	5.14	4.84	4.91

LRT: Likelihood ratio test. *RST*: Rao's score test. *WT*: Wald test.

3.2. Powers

The global powers were calculated in a similar way to the global type I errors. Table 5 shows some of the results obtained for the methods that have a global type I error with better behaviour, i.e. $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ through the McNemar test with *cc* to an error $\alpha = 5\%$ (Table 5: individual *MTcc*), the three global tests (Table 5: global *LRT*, global *RST* and global *WT*), and $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ through the *MT*, *WT* and *LRT* along with the Bonferroni method (Table 5: Bonferroni *MT*, *WT* and *LRT*). The choice of McNemar test with *cc* to an error $\alpha = 5\%$ is justified because its global type I error has a good behavior when $n \leq 500$. The global powers of the rest of the methods are not shown, since these methods have global type I errors that very frequently go too far above the nominal error. Nor do we show the results obtained with

the Holm method, as they are practically the same as those obtained with the Bonferroni method. From the results, the following conclusions are obtained.

The prevalence of the disease and the covariances have an important effect on the global powers of the methods. The increase in the prevalence (if the other parameters remain constant) means an increase in the powers, whereas the increase in the covariances ε_i means (if the other parameters remain constant) a decrease in the powers. These results are to be expected, since an increase in the prevalence means an increase in the global type I errors, and an increase in the covariances means a decrease in the global type I errors. In general terms, the *LRT* and the *WT* for the global hypothesis test are a little more powerful than the *RST*, especially when $n \leq 300-400$ depending on the differences between the two sensitivities (specificities). Moreover, regarding the global test, there is no important difference between the power of the *LRT* and of the *WT*.

Comparing the power of the *LRT* (*WT*) for the global test and the power of the McNemar test with *cc* to an error $\alpha = 5\%$, in very general terms, the global tests are a little less powerful, between 1% and 5% approximately, when the prevalence is small or very small$\left(p \leq 10\%\right)$ and $n \leq 200$, and the powers are very similar when $n \geq 300$. When $n \leq 500$ and the prevalence is moderate (25%) or large (50%), the global tests are more powerful, between 2% and 10% approximately depending on the sample size and on the prevalence. When $n \geq 1000$ the powers are practically equal.

Regarding the methods based on the individual tests along with the Bonferroni method, there is no important difference between the powers of the three methods, especially when $n \geq 200$. When the sample size is small or moderate, the *WT* and the *LRT* along with Bonferroni are slightly more powerful than the *MT* along with Bonferroni. Comparing the powers of these three methods with the power of the *MTcc* to an error $\alpha = 5\%$, there is no method that is clearly more powerful than another. In very general terms,: when the sample size is small $\left(n = 50\right)$ or moderate $\left(n = 100\right)$ the *WT* and the *LRT* along with Bonferroni are usually slightly more powerful; when the sample size is large $\left(200 \leq n \leq 500\right)$ the individual *MTcc* to an error $\alpha = 5\%$ is usually slightly more powerful, and when the sample size is very large $\left(n \geq 1000\right)$ the powers are practically equal.

Table 4. Global type I errors (%) of the individual tests with Bonferroni method

$Se_1 = Se_2 = 0.80$, $Sp_1 = Sp_2 = 0.90$, $p = 5\%$, $\varepsilon_1 = 0.04$, $\varepsilon_0 = 0.0225$ (25% of the maximum value)

n	*CET*	*Midp*	*MT*	*MTcc*	*MMT*	*WT*	*MWT*	*LRT*	*UET*	*UMT*	*ULRT*
50	0.15	0.47	0.47	0.06	0.15	0.56	0.47	0.99	0.47	0.01	0.02
100	0.80	1.39	1.39	0.60	1.19	1.57	1.42	1.51	1.39	0.07	0.14
200	1.53	1.99	1.99	1.39	2.21	2.05	2.26	2.01	1.99	0.19	0.24
300	1.56	2.33	2.33	1.44	2.11	2.59	2.15	2.58	2.33	0.16	0.21
400	1.73	2.52	2.52	1.58	2.21	3.06	2.61	3.03	2.52	0.16	0.18
500	2.09	3.02	3.02	1.82	2.53	3.78	3.04	3.64	3.02	0.22	0.28
1000	3.00	4.51	4.53	2.64	3.72	5.02	4.53	5.03	4.53	0.34	0.47
2000	3.62	4.87	4.88	3.50	4.67	5.45	4.92	5.09	4.88	0.59	0.70

$Se_1 = Se_2 = 0.80$, $Sp_1 = Sp_2 = 0.90$, $p = 10\%$, $\varepsilon_1 = 0.08$, $\varepsilon_0 = 0.045$ (50% of the maximum value)

n	*CET*	*Midp*	*MT*	*MTcc*	*MMT*	*WT*	*MWT*	*LRT*	*UET*	*UMT*	*ULRT*
50	0.02	0.11	0.11	0.01	0.02	0.25	0.11	0.38	0.11	0.00	0.00
100	0.34	0.75	0.75	0.17	0.38	0.82	0.79	1.13	0.75	0.01	0.03
200	1.32	2.28	2.28	1.12	2.00	2.57	2.29	2.60	2.28	0.14	0.22
300	1.48	2.48	2.48	1.32	2.15	3.07	2.46	3.05	2.48	0.16	0.23
400	1.83	2.89	2.89	1.45	2.24	3.45	2.89	3.72	2.89	0.16	0.20
500	2.07	3.42	3.42	1.72	2.67	3.67	3.40	4.38	3.42	0.19	0.26
1000	3.13	4.66	4.67	2.81	4.26	4.85	4.64	4.91	4.67	0.27	0.43
2000	3.60	4.91	4.95	3.34	4.58	4.94	4.86	4.97	**5.04**	0.39	0.46

$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.90$, $p = 25\%$, $\varepsilon_1 = 0.045$, $\varepsilon_0 = 0.045$ (50% of the maximum value)

n	*CET*	*Midp*	*MT*	*MTcc*	*MMT*	*WT*	*MWT*	*LRT*	*UET*	*UMT*	*ULRT*
50	0	0	0	0	0	0.10	0	0.10	0	0	0
100	0.28	0.82	0.82	0.14	0.28	0.84	0.82	1.48	0.82	0	0.06
200	1.32	2.36	2.36	0.94	2.18	2.40	2.36	2.96	2.36	0.06	0.34
300	1.62	2.92	2.92	1.46	3.86	3.08	2.92	3.98	2.92	0.14	0.54
400	2.52	3.98	3.98	2.14	4.02	4.10	3.98	4.92	3.98	0.14	0.34
500	2.40	4.26	4.26	1.94	4.22	4.28	4.26	4.94	4.26	0.34	0.40
1000	2.98	4.48	4.48	2.80	4.62	4.70	4.42	4.70	4.48	0.28	0.52
2000	3.88	4.92	4.98	3.78	4.86	5.10	4.88	5.12	**5.06**	0.30	0.50

$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.90$, $p = 50\%$, $\varepsilon_1 = 0.0675$, $\varepsilon_0 = 0.0675$ (75% of the maximum value)

n	*CET*	*Midp*	*MT*	*MTcc*	*MMT*	*WT*	*MWT*	*LRT*	*UET*	*UMT*	*ULRT*
50	0	0	0	0	0	0	0	0	0	0	0
100	0.04	0.04	0.04	0.02	0.04	0.04	0.04	0.08	0.04	0	0
200	0.06	0.28	0.28	0.02	0.54	0.28	0.28	1.14	0.28	0	0.04
300	0.62	1.58	1.58	0.16	1.64	1.58	1.58	3.06	1.58	0	0.26
400	1.64	3.26	3.26	0.68	2.44	3.26	2.78	4.78	3.26	0.08	0.24
500	1.92	3.66	3.66	1.14	3.42	3.66	2.62	4.50	3.66	0.08	0.46
1000	2.78	4.34	4.34	2.72	3.86	4.88	4.04	4.88	4.34	0.28	0.42
2000	3.46	**5.06**	5.06	3.22	4.38	5.34	4.88	5.36	**5.06**	0.20	0.24

$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.80$, $p = 25\%$, $\varepsilon_1 = 0.0225$, $\varepsilon_0 = 0.04$ (25% of the maximum value)

n	*CET*	*Midp*	*MT*	*MTcc*	*MMT*	*WT*	*MWT*	*LRT*	*UET*	*UMT*	*ULRT*
50	0.40	1.12	1.12	0.16	0.70	2.60	1.14	2.00	1.10	0	0.08
100	1.64	2.56	2.56	1.48	2.22	3.50	2.80	2.90	2.56	0.18	0.26
200	1.86	2.88	2.88	1.56	2.26	3.08	2.94	3.88	2.88	0.22	0.26
300	2.70	4.14	4.14	2.14	3.28	4.28	4.16	5.00	4.14	0.24	0.38
400	2.86	4.36	4.36	2.46	3.72	4.74	4.40	4.76	4.36	0.30	0.56
500	2.78	4.16	4.16	2.58	4.14	4.44	4.16	4.44	4.16	0.26	0.56
1000	4.16	**5.36**	5.38	3.94	4.92	5.38	5.38	5.38	**5.38**	0.52	0.52
2000	3.96	**5.08**	4.92	3.82	4.30	4.96	4.92	4.94	**5.10**	0.34	0.42

$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.80$, $p = 50\%$, $\varepsilon_1 = 0.081$, $\varepsilon_0 = 0.144$ (90% of the maximum value)

n	*CET*	*Midp*	*MT*	*MTcc*	*MMT*	*WT*	*MWT*	*LRT*	*UET*	*UMT*	*ULRT*
50	0	0	0	0	0	0	0	0	0	0	0
100	0	0.01	0.01	0	0	0.01	0.01	0.02	0.01	0	0
200	0	0.03	0.03	0	0	0.03	0.03	0.10	0.03	0	0
300	0.08	0.19	0.19	0.01	0.08	0.19	0.19	0.43	0.19	0	0
400	0.18	0.66	0.66	0.06	0.18	0.66	0.42	1.17	0.66	0.01	0.01
500	0.42	1.22	1.22	0.17	0.45	1.22	0.49	2.16	1.22	0.01	0.02
1000	1.79	3.52	3.52	1.29	2.47	3.70	2.47	4.11	3.52	0.12	0.21
2000	2.91	4.53	4.53	2.59	4.02	4.84	4.02	4.66	4.53	0.23	0.36

CET: Conditional exact test. *Midp*: Mid-p test. *MT*: McNemar test. *MTcc*. McNemar test with cc. *MMT*: Modified McNemar test. *WT*: Wald test. *MWT*: Modified Wald test. *LRT*: Likelihood ratio test. *UET*: Unconditional exact test. *UMT*: Unconditional McNemar test. *ULRT*: Unconditional likelihood ratio test.

Table 5. Global powers (%) of the methods.

$Se_1 = 0.90$, $Se_2 = 0.80$, $Sp_1 = 0.80$, $Sp_2 = 0.70$, $p = 5\%$, $\varepsilon_1 = 0.04$, $\varepsilon_0 = 0.07$ (25% of the maximum value)

n	*Individual MTcc*	*Global LRT*	*Global RST*	*Global WT*	*Bonferroni MT*	*Bonferroni WT*	*Bonferroni LRT*
50	9.01	7.11	6.10	8.69	8.48	11.96	8.99
100	29.67	25.30	24.27	26.90	27.31	30.33	27.41
200	60.89	56.12	54.86	56.00	57.88	59.66	58.16
300	82.48	78.76	78.15	79.37	80.14	80.95	80.46
400	92.96	90.05	89.74	90.25	90.14	90.44	90.26
500	97.54	96.35	96.24	96.38	96.27	96.43	96.38
1000	99.99	99.98	99.98	99.98	99.98	99.99	99.98
2000	100	100	100	100	100	100	100

$Se_1 = 0.90$, $Se_2 = 0.80$, $Sp_1 = 0.80$, $Sp_2 = 0.70$, $p = 10\%$, $\varepsilon_1 = 0.072$, $\varepsilon_0 = 0.126$ (90% of the maximum value)

n	*Individual MTcc*	*Global LRT*	*Global RST*	*Global WT*	*Bonferroni MT*	*Bonferroni WT*	*Bonferroni LRT*
50	12.36	12.55	8.62	13.41	12.36	16.49	20.61
100	60.07	58.64	55.37	57.56	59.91	60.65	61.73
200	96.39	96.51	95.84	96.27	95.55	96.17	96.17
300	100	100	100	100	100	100	100
400	100	100	100	100	100	100	100
500	100	100	100	100	100	100	100
1000	100	100	100	100	100	100	100
2000	100	100	100	100	100	100	100

$Se_1 = 0.95$, $Se_2 = 0.85$, $Sp_1 = 0.90$, $Sp_2 = 0.80$, $p = 25\%$, $\varepsilon_1 = 0.0213$, $\varepsilon_0 = 0.04$ (50% of the maximum value)

n	*Individual MTcc*	*Global LRT*	*Global RST*	*Global WT*	*Bonferroni MT*	*Bonferroni WT*	*Bonferroni LRT*
50	7.0	8.4	4.6	9.7	7.0	13.2	12.1
100	40.7	46.4	41.8	46.1	40.3	43.3	43.4
200	83.6	89.3	88.3	89.2	82.0	82.5	84.0
300	96.7	98.0	97.7	98.0	95.8	95.9	96.3
400	99.6	99.9	99.8	99.9	99.4	99.4	99.5
500	99.9	100	100	100	99.9	99.9	99.9
1000	100	100	100	100	100	100	100
2000	100	100	100	100	100	100	100

$Se_1 = 0.95$, $Se_2 = 0.85$, $Sp_1 = 0.90$, $Sp_2 = 0.80$, $p = 25\%$, $\varepsilon_1 = 0.0319$, $\varepsilon_0 = 0.06$ (75% of the maximum value)

n	*Individual MTcc*	*Global LRT*	*Global RST*	*Global WT*	*Bonferroni MT*	*Bonferroni WT*	*Bonferroni LRT*
50	4.7	5.6	2.6	6.3	4.7	10.0	4.7
100	45.3	51.7	45.9	49.6	45.3	46.2	45.3
200	91.4	95.2	94.4	94.9	90.3	91.2	91.4
300	99.3	99.7	99.6	99.6	99.2	99.2	99.3
400	99.9	99.9	99.9	99.9	99.9	99.9	99.9
500	100	100	100	100	100	100	100
1000	100	100	100	100	100	100	100
2000	100	100	100	100	100	100	100

$Se_1 = 0.85$, $Se_2 = 0.80$, $Sp_1 = 0.95$, $Sp_2 = 0.90$, $p = 50\%$, $\varepsilon_1 = 0.06$, $\varepsilon_0 = 0.0425$ (50% of the maximum value)

n	*Individual MTcc*	*Global LRT*	*Global RST*	*Global WT*	*Bonferroni MT*	*Bonferroni WT*	*Bonferroni LRT*
50	0.4	0.8	0.6	1.2	0.4	1.3	1.2
100	8.0	12.5	9.7	12.4	7.9	8.4	12.3
200	28.2	35.2	32.6	34.6	27.5	28.5	28.8
300	46.8	51.5	49.6	51.1	43.6	45.2	45.1
400	58.6	64.4	63.4	64.2	56.8	57.8	57.9
500	71.5	75.2	74.4	75.0	67.5	69.9	68.9
1000	95.8	96.9	96.8	96.9	93.9	95.0	95.0
2000	100	100	100	100	99.9	99.9	99.9

$Se_1 = 0.85$, $Se_2 = 0.80$, $Sp_1 = 0.95$, $Sp_2 = 0.90$, $p = 50\%$, $\varepsilon_1 = 0.09$, $\varepsilon_0 = 0.0634$ (75% of the maximum value)

n	*Individual MTcc*	*Global LRT*	*Global RST*	*Global WT*	*Bonferroni MT*	*Bonferroni WT*	*Bonferroni LRT*
50	0.1	0.2	0.1	0.2	0.1	0.4	0.4
100	4.3	8.2	6.1	7.3	4.3	4.5	8.9
200	34.4	46.8	42.2	44.5	34.3	34.4	39.3
300	59.8	70.5	67.9	69.1	58.5	59.5	60.3
400	78.3	85.6	84.5	85.0	75.3	77.3	77.4
500	86.6	90.7	90.1	90.4	84.3	85.6	85.6
1000	99.6	99.7	99.7	99.7	99.4	99.4	99.4
2000	100	100	100	100	100	100	100

Finally, comparing the powers of the global tests with those of the *MT*, *WT* and *LRT* methods along with Bonferroni, in general terms, the *MT*, *WT* and *LRT* methods along with Bonferroni are usually a little more powerful than the three global tests when $p \leq 10\%$ and $n \leq 100$, and the powers are very similar when $n \geq 200$. When $200 \leq n \leq 500$ and the prevalence is moderate or large $(p \geq 25\%)$, the global tests are more powerful, between 2% and almost 10% approximately, depending on the sample size.

3.3. Rules of application

Based on the results obtained in the simulation experiments, the following general rules of application can be established:

1). When the prevalence is small $(p = 10\%)$ or very small $(p = 5\%)$ and the sample is small $(n = 50)$ or moderate $(n = 100)$, solve the tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ individually applying the *WT* or the *LRT* along with the Bonferroni (or Holm) method to an error $\alpha = 5\%$.

2). In any other situation, solve the global test $H_0 : (Se_1 = Se_2$ and $Sp_1 = Sp_2)$ to an error $\alpha = 5\%$ applying the *LRT* or the *WT*. In this situation, if the global test is not significant then the equality of the accuracy of both *BDTs* is not rejected, and if the global is significant then the causes of the significance will be investigated: a) testing the tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ individually applying the *WT* or the *LRT* along with the Bonferroni or Holm method to an error $\alpha = 5\%$ if the sample size is small or moderate $(n \leq 100)$ or if the sample size is very large $(n \geq 1000)$; or b) testing the tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ individually applying the McNemar test with *cc* to an error $\alpha = 5\%$ if the sample size is large $(200 \leq n \leq 500)$.

4. Example

The results obtained were applied to the diagnosis of coronary artery disease [4], using dobutamine echocardiography (*DE*) and myocardial perfusion scintigraphy (*MPS*) as diagnostic tests and coronary angiography (*CA*) as the *GS*. Table 6 shows the

frequencies obtained applying the two *BDTs* and the *GS* to a sample of 548 men, and where the variable T_1 models the result of the *DE*, T_2 models the result of the *MPS* and D the result of the *CA*. This table also shows the estimation (estimation $\pm$ standard error) of the accuracy of each *BDT* and of the disease prevalence, as well as the results obtained performing the global hypothesis test applying the three methods studied (the likelihood ratio test, Rao's score test and the Wald test), and those obtained testing the individual hypothesis tests applying the McNemar test, the McNemar test with *cc*, the Wald test and the likelihood ratio test. With the three statistics for the global hypothesis test the same conclusion is obtained: the homogeneity of the two sensitivities and of the specificities is not rejected. The same conclusion is reached applying the individual hypothesis tests along with the Bonferroni (or Holm) method. Nevertheless, if each one of the individual hypothesis tests is solved to an error $\alpha = 5\%$ applying the McNemar test (Wald test or likelihood ratio test), we conclude that the two sensitivities are different (the sensitivity of the *DE* test is significantly greater than that of the *MPS*) and the equality of the two specificities is not rejected.

Through the general rules given from the simulation experiments, this example must be solved applying the global test; incorrect conclusion would be obtained if the individual hypothesis tests (*MT*, *WT* and *LRT*) are applied to an error $\alpha = 5\%$. If the two individual tests are solved to an error $\alpha = 5\%$ applying the McNemar test with *cc* the same conclusions are obtained as applying the global hypothesis test. Nevertheless, the simulation experiments have shown that this method can overcome in excess the nominal error when the sample size is very large and, therefore, it should not be used in this situation.

Table 6. Study of coronary disease.

Observed frequencies					
	$T_1=1$		$T_1=0$		
	$T_2=1$	$T_2=0$	$T_2=1$	$T_2=0$	Total
$D=1$	152	17	7	36	212
$D=0$	25	10	11	290	336
Total	177	27	18	326	548

Results				
Dobutamine ecocardiography		Myocardial perfusion scintigraphy		Prevalence
$\hat{Se}_1 \pm SE$	$\hat{Sp}_1 \pm SE$	$\hat{Se}_2 \pm SE$	$\hat{Sp}_2 \pm SE$	$\hat{p} \pm SE$
0.797 ± 0.028	0.896 ± 0.017	0.750 ± 0.030	0.893 ± 0.017	0.387 ± 0.021

Global hypothesis test:

$H_0:(Se_1 = Se_2$ and $Sp_1 = Sp_2)$ vs $H_0:(Se_1 \neq Se_2$ and/or $Sp_1 \neq Sp_2)$

Likelihood ratio test (*LRT*)	Rao score test (*RST*)	Wald test (*WT*)
$\chi^2 = 4.344$, $p-value = 0.114$	$\chi^2 = 4.214$, $p-value = 0.122$	$\chi^2 = 4.298$, $p-value = 0.117$

Individual test: $H_0: Se_1 = Se_2$ vs $H_1: Se_1 \neq Se_2$

McNemar test	McNemar test with *cc*	Wald test	Likelihood ratio test
$\chi^2 = 4.167$	$\chi^2 = 3.375$	$\chi^2 = 4.250$	$\chi^2 = 4.296$
$p-value = 0.041$	$p-value = 0.066$	$p-value = 0.039$	$p-value = 0.038$

Individual test: $H_0: Sp_1 = Sp_2$ vs $H_1: Sp_1 \neq Sp_2$

McNemar test	McNemar test with *cc*	Wald test	Likelihood ratio test
$\chi^2 = 0.048$	$\chi^2 = 0$	$\chi^2 = 0.048$	$\chi^2 = 0.048$
$p-value = 0.827$	$p-value = 1$	$p-value = 0.827$	$p-value = 0.827$

5. Discussion

Traditionally, the comparison of the accuracy of two *BDTs* subject to a paired design is made conditioning on the individuals with (without) the disease and comparing the two sensitivities (specificities) applying a comparison test with two paired binomial proportions to an α error. Therefore, each one of the tests $H_0: Se_1 = Se_2$ and $H_0: Sp_1 = Sp_2$ are tested independently to an α error. An alternative to this method is to compare the two sensitivities and the two specificities simultaneously, i.e. performing the global test $H_0:(Se_1 = Se_2$ and $Sp_1 = Sp_2)$ vs $H_1:(Se_1 \neq Se_2$ and/or $Sp_1 \neq Sp_2)$. This article studies this global hypothesis test, extending the study by Lachenbruch and Lynch [3], through the application of Rao's score test and the Wald test. Lachenbruch and Lynch proposed two statistics for the global test, one obtained applying the likelihood ratio test and another one obtained as the sum of the McNemar statistic for the test $H_0: Se_1 = Se_2$ and of the McNemar statistic for the test $H_0: Sp_1 = Sp_2$. This last

statistic is obtained considering two independent 2×2 tables, one made up of individuals with the disease and another one made up of individuals without the disease and applying the results of Hamdan et al [10]. That is to say, it is considered that each table is extracted from a population, and that the two populations considered differ in their disease status. In Section 2.2.2 the same statistic has been derived applying Rao's score test, assuming that there is a single sample extracted from a population that has a determined disease prevalence. Another statistic has also been obtained using the Wald test, which is also the sum of the Wald statistics for the individual tests. Another alternative method that has been studied to compare the accuracy of two *BDTs* consisted of testing the two individual tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ through a comparison test of paired binomial proportions and application of the Bonferroni method or the Holm method.

Simulation experiments were carried out to study the global type I errors and global powers of the methods to compare the accuracy of the two *BDTs*. To study the sizes of the exact tests the criterion that was followed was that their global type I errors would not be over 5%. For approximate tests, the criterion that was followed was that their global type I errors fluctuate around 5% without exceeding it too much (global type I errors are not $\geq 7\%$). This difference in criteria between both types of methods is based on the idea that we should not consider that both types of tests (exact and asymptotic) must be have the same type I error behaviour: an exact test should not give false significances whereas an asymptotic test may exceed the nominal error without giving too many false significances. The 7% is an acceptable value, very close to 5%, and does not lead to too many false significances. Fagerland et al [7] compared exact and asymptotic tests subject to the same criterion of 5%. If this criterion is used to interpret the results of the simulation experiments, in general terms the conclusions are the same regarding the global hypothesis test and the individual hypothesis tests along with Bonferroni or Holm, although on some occasions $(n \geq 500)$ the global type I error may very slightly greater than 5%. Regarding the individual tests to an error $\alpha = 5\%$, the asymptotic tests exceed the 5% with a sample size which is smaller than with the criterion of 7%, which further invalidates these methods for their practical application. Taking into account the "approximate" nature of an asymptotic test, the criterion of 7% for the asymptotic tests is more flexible than the one used by Fagerland et al.

Finally, when the sample is small or moderate, it may happen that there are too frequencies equal to zero, and therefore the tests given in the application rules can not be used. In this case, one solution is to add the value 0.5 to each of the observed frequencies, which is a solution that is widely used in the analysis of contingency tables. Simulation experiments carried out, similar to those in Section 3, have shown that this does not have an important effect either on the type I error or on the power of the tests.

Acknowledgements

We thank the Editor, the Associate Editor and the Referee for their helpful comments that improved the quality of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by the Spanish Ministry of Economy [Grant Number MTM2016-76938-P].

References

1. Zhou, X.H., Obuchowski, N.A. and McClish, D.K. (2011). *Statistical Methods in Diagnostic Medicine (Second Edition)*. New Jersey: John Wiley & Sons.

2. Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* New York: Oxford University Press.

3. Lachenbruch, P.A. and Lynch, C.J. (1998). Assessing screening tests: extensions of McNemar's test. Statistics in Medicine, 17, 2207-2217.

4. Roldán-Nofuentes, J.A., Luna del Castillo, J.D. and Montero-Alonso, M.A. (2012). Global hypothesis test to simultaneously compare the predictive values of two binary diagnostic tests. Computational Statistics and Data Analysis, 56, 1161-1173.

5. Vacek, P.M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics, 41, 959-968.

6. May, W.L. and Johnson, W.D. (1997). The validity and power of tests for equality of two correlated proportions. Statistics in Medicine, 16, 1081-1096.

7. Fagerland, M.W., Lydersen, S. and Laake, P. (2014). Recommended tests and confidence intervals for paired binomial proportions. Statistics in Medicine, 33, 2850-2875.

8. Rao, C.R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. Proceedings of the Cambridge Philosophical Society, 44, 50-57.

9. Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of the American Mathematical Society, 5, 426-482.

10. Hamdan, M.A., Pirie, W.R. and Arnold, J.C. (1975). Simultaneous testing of McNemar's problem for several populations. Psychometrika, 40, 153-161.

11. Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8, 3-62.

12. Holm, S. (1979). A simple sequential rejective multiple testing procedure. Scandinavian Journal of Statistics, 6, 65-70.

13. Lancaster, H.O. (1961). Significance tests in discrete distribution. Journal of American Statistical Association, 56, 223-234

14. McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages. Psychometrika, 12, 153-157.

15. Edwards, A.L. (1948). Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. Psychometrika, 13, 185-187.

16. Bennett, B.M. and Underwood, R.E. (1970). On McNemar's test for the 2×2 table and its power function. Biometrics, 26, 339-343.

17. Suissa, S. and Shuster, J.J. (1991). The 2×2 matched-pairs trial: exact unconditional design and analysis. Biometrics, 47, 361-372.

18. Lu, Y. (2010). A revised version of McNemar's test for paired binary data. Communications in Statistics - Theory and Methods, 39, 3525-3539.

19. Lu, Y. (2011). Considering the concordant observations in likelihood ratio test for paired binary data. Communications in Statistics - Theory and Methods, 39, 4214-4232.

20. Price, R.M. and Bonett, D.G. (2004). An improved confidence interval for a linear function of binomial proportions. Computational Statistics and Data Analysis, 45, 449-456.

21. Martín-Andrés, A. and Álvarez-Hernández, M. (2014). Two-tailed asymptotic inferences for a proportion. Journal of Applied Statistics, 41, 1516-1529.

22. Martín-Andrés, A. and Álvarez-Hernández, M. (2014). Two-tailed approximate confidence intervals for the ratio of proportions. Statistics and Computing, 24, 65-75.

23. Roldán-Nofuentes, J.A., Amro, R. (2017). Approximate confidence intervals for the weighted kappa coefficient of a binary diagnostic test subject to a case-control design. Journal of Statistical Computation and Simulation, 87, 530-545.

Appendix *A*

1. Conditional Exact Test (CET)

In hypothesis test (1) the proportions p_{11} and p_{00} do not appear, and so it is possible to discard these proportions and, consequently, also discard the frequencies s_{11} and s_{00}. Conditioning on the sum of the discordant frequencies, i.e. conditioning on $s_{10}+s_{01}$, it is verified that $p_{10}+p_{01}=1$, and it is also verified that s_{10} is the product of a binomial distribution of parameters $s_{10}+s_{01}$ and p_{10}, i.e. $Bin\left(s_{10}+s_{01},p_{10}\right)$. If the null hypothesis is true then $p_{10}=p_{01}=1/2$, and, therefore, both the hypothesis test (1) is also equivalent to test $H_0:p_{10}=1/2$ vs $H_1:p_{10}\neq 1/2$. Finally, the two-sided exact p-value for the comparison test of the two sensitivities is

$$\text{two-sided exact p-value} = 2 \times \sum_{j=0}^{Min(s_{10}, s_{01})} \binom{s_{10} + s_{01}}{j} \left(\frac{1}{2}\right)^{s_{10}+s_{01}} . \quad (12)$$

If $s_{10} = s_{01}$ then the two-sided exact p-value equals one. In a similar way, the two-sided exact p-value to compare the two specificities is

$$\text{two-sided exact p-value} = 2 \times \sum_{j=0}^{Min(r_{10}, r_{01})} \binom{r_{10} + r_{01}}{j} \left(\frac{1}{2}\right)^{r_{10}+r_{01}} . \quad (13)$$

2. Conditional Mid-p Test (*Midp*)

The conditional mid-p test [13] is a modification of the exact conditional test. This method consists of subtracting the probability of the observed outcome s_{10} from (12). Thus, the mid-p values to compare the two sensitivities and the two specificities are

$$\text{mid-p value} = \text{two-sided exact p-value} - \binom{s_{10} + s_{01}}{s_{10}} \left(\frac{1}{2}\right)^{s_{10}+s_{01}} \quad (14)$$

and

$$\text{mid-p value} = \text{two-sided exact p-value} - \binom{r_{10} + r_{01}}{r_{10}} \left(\frac{1}{2}\right)^{r_{10}+r_{01}} \quad (15)$$

respectively. The condicional mid-p test is also referred to as quasi-exact test.

3. McNemar Test (*MT*)

The McNemar [14] test is the asymptotic version of the conditional exact test. Conditioning on the sum of discordant frequencies and applying the Central Limit Theorem, the statistic for hypothesis test (1) is

$$z = \frac{\hat{p}_{10} - \hat{p}_{01}}{\sqrt{Var\left(\hat{p}_{10} - \hat{p}_{01}\right)}}, \quad (16)$$

which is distributed according to a standard normal distribution, where

$$Var\left(\hat{p}_{10} - \hat{p}_{01}\right) = \frac{p_{10} + p_{01} - \left(p_{10} - p_{01}\right)^2}{s}. \quad (17)$$

If the null hypothesis is true, then

$$Var_0\left(\hat{p}_{10}-\hat{p}_{01}\right)=\frac{p_{10}+p_{01}}{s}. \tag{18}$$

Substituting in equation (18) the parameters with their estimators, and substituting the expression obtained in equation (16), it is obtained that the statistic for the McNemar test $z_M=\left(s_{10}-s_{01}\right)\Big/\sqrt{s_{10}+s_{01}}$. It is very common to express this statistic in terms of the chi-square distribution, i.e.

$$\chi^2_M=\frac{\left(s_{10}-s_{01}\right)^2}{s_{10}+s_{01}}, \tag{19}$$

which is distributed asymptotically according to a chi-square distribution with one degree of freedom. In a similar way, the statistic of the McNemar test is obtained to compare the two specificities:

$$\chi^2_M=\frac{\left(r_{10}-r_{01}\right)^2}{r_{10}+r_{01}}. \tag{20}$$

4. McNemar Test with continuity correction (*MTcc*)

In the McNemar test the binomial distribution is approximated through the normal distribution. In this situation, it is common to apply continuity correction (*cc*). Edwards [15] proposed the following continuity correction version of the McNemar test,

$$z_{Mcc}=\frac{\left|\hat{p}_{10}-\hat{p}_{01}\right|-\dfrac{1}{s}}{\sqrt{Var\left(\hat{p}_{10}-\hat{p}_{01}\right)}}. \tag{21}$$

Performing the same algebraic operations in the previous section, it is obtained that the statistics of the McNemar test with *cc* are

$$\chi^2_{Mcc}=\frac{\left(\left|s_{10}-s_{01}\right|-1\right)^2}{s_{10}+s_{01}} \quad \text{and} \quad \chi^2_{Mcc}=\frac{\left(\left|r_{10}-r_{01}\right|-1\right)^2}{r_{10}+r_{01}}, \tag{22}$$

respectively.

5. Modified McNemar Test (*MMT*)

Bennett and Underwood [16] proposed a modification of the statistic of the McNemar test adding 0.5 to the observed frequencies. This correction improves the approximation to the chi-square distribution. The statistics of *MMT* are:

$$\chi^2_{MM} = \frac{\left(s_{10} - s_{01}\right)^2}{s_{10} + s_{01} + 1} \quad \text{and} \quad \chi^2_{MM} = \frac{\left(r_{10} - r_{01}\right)^2}{r_{10} + r_{01} + 1}. \tag{23}$$

6. Wald Test (*WT*)

The comparison of the two sensitivities (specificities) can also be made applying the Wald test [9]. The statistic of the McNemar test is obtained substituting in equation (16) $Var\left(\hat{p}_{10} - \hat{p}_{01}\right)$ with its expression subject to the null hypothesis, i.e. $Var_0\left(\hat{p}_{10} - \hat{p}_{01}\right)$ (equation (7)). Substituting in (16) the $Var\left(\hat{p}_{10} - \hat{p}_{01}\right)$ with its expression given in equation (17), and substituting the parameters with their estimators, we obtain the Wald test statistic to compare the sensitivities, i.e.

$$\chi^2_W = \frac{s\left(s_{10} - s_{01}\right)^2}{4s_{10}s_{01} + \left(s_{11} + s_{00}\right)\left(s_{10} + s_{01}\right)}, \tag{24}$$

which is distributed asymptotically according to a chi-square distribution with one degree of freedom. To compare the two specificities, the Wald test statistics is

$$\chi^2_W = \frac{r\left(r_{10} - r_{01}\right)^2}{4r_{10}r_{01} + \left(r_{11} + r_{00}\right)\left(r_{10} + r_{01}\right)}. \tag{25}$$

7. Modified Wald Test (*MWT*)

As the *WT* tends to reject too often under the null hypothesis when the sample size is small or moderate, May and Johnson [6] proposed a modification of the Wald statistic adding 0.5 to each one of the discordant frequencies, i.e.

$$\chi^2_{MW} = \frac{\left(s_{10} - s_{01}\right)^2}{\left(s_{10} + s_{01} + 1\right) - \frac{\left(s_{10} - s_{01}\right)^2}{s}} \quad \text{and} \quad \chi^2_{MW} = \frac{\left(r_{10} - r_{01}\right)^2}{\left(r_{10} + r_{01} + 1\right) - \frac{\left(r_{10} - r_{01}\right)^2}{s}}. \tag{26}$$

This modification reduces the size of the Wald statistic, and for $n \leq 50$ the size of the test is close to the nominal error.

8. Likelihood Ratio Test (*LRT*)

Conditioning on the sum of the discordant frequencies, if the null hypothesis (1) is true then it is verified that $\hat{p}_{10} = \hat{p}_{01} = (s_{10} + s_{01})/(2s)$. It is easy to prove that the likelihood ratio statistic to compare the sensitivities is

$$\chi^2_{LR} = 2\left[s_{10} \ln\left(\frac{2s_{10}}{s_{10} + s_{01}} \right) + s_{01} \ln\left(\frac{2s_{01}}{s_{10} + s_{01}} \right) \right], \tag{27}$$

and in a similar way, the likelihood ratio statistic to compare the specificities is

$$\chi^2_{LR} = 2\left[r_{10} \ln\left(\frac{2r_{10}}{r_{10} + r_{01}} \right) + r_{01} \ln\left(\frac{2r_{01}}{r_{10} + r_{01}} \right) \right], \tag{28}$$

whose distributions are asymptotically a chi-square with one degree of freedom.

2.1.9. Unconditional Exact Test (*UET*)

The conditional exact test and the mid-p test are based on the conditioning on the sum of the discordant frequencies. Suissa and Shuster [17] proposed, based on the statistic of the McNemar test, an exact test which uses all the frequencies in the sample and, therefore, does not condition in the sum of the discordant frequencies. When we compare the two sensitivities, the power function of the test is

$$P(p_{10}, p_{01}) = \sum_C \binom{s}{s_{10} \quad s_{01} \quad s-m} p_{10}^{s_{10}} p_{01}^{s_{01}} (1 - p_{10} - p_{01})^{s-m},$$

where $m = s_{10} + s_{01}$ and $C = \{(s_{10}, m):\ s_{10} \geq h(m);\ s_{10} = 0,1,...,m;\ m = 0,1,...,s\}$, with $h(m) = \left(z_M \sqrt{m} + m\right)/2$ and z_M the calculated value of the McNemar statistic. If the null hypothesis is true, then the distribution of $(s_{10}, m, s-m)$ is a trinomial distribution with parameters *s* and probability vector is $(\delta/2, \delta/2, 1-\delta)^T$, i.e.

$$P(\delta)=\sum_{C}\binom{s}{s_{10}\quad s_{01}\quad s-m}\left(\frac{\delta}{2}\right)^{m}(1-\delta)^{s-m},$$

and where $\delta = p_{10} + p_{01}$ is the nuisance parameter. The nuisance parameter is eliminated by maximizing this function over the range of δ. The function $P(\delta)$ is simplified as

$$P(\delta)=\sum_{j=k}^{s}\binom{s}{j}\delta^{j}(1-\delta)^{s-j}F_{j}(j-i_{j}-1),$$

where $k=\text{int}\left[z_{m}^{2}+1\right]$, $i_{j}=\text{int}\left[h(j)\right]$, $\text{int}[.]$ is the integer function and F_j is the cumulative binomial distribution function with parameters j and $1/2$. Finally, the two-sided exact p-value is calculated as

$$\text{two sided exact p-value}=2\times\sup_{0<\delta<1}\left\{P(\delta)\right\}. \tag{29}$$

The two-sided exact p-value to compare the two specificities is calculated in a similar way, substituting "*s*" with "*r*" and "*p*" with "*q*".

2.1.10. Unconditional McNemar Test (*UMT*)

Lu [18] proposed a statistic for the McNemar test that considers all the frequencies in the sample, and which therefore does not condition in the sum of the discordant frequencies. The hypothesis test (1) is equivalent to the hypothesis test

$$H_0:\frac{p_{10}}{p_{10}+p_{01}}=\frac{p_{01}}{p_{10}+p_{01}}\ \text{ vs }\ H_1:\frac{p_{10}}{p_{10}+p_{01}}\neq\frac{p_{01}}{p_{10}+p_{01}}.$$

Subject to the null hypothesis, the frequency s_{10} (or s_{01}) is the product of binomial distribution of parameters s and $\delta=(p_{10}+p_{01})/2$. The estimators of the average and of the variance of the binomial distribution are $s\hat{\delta}=(s_{10}+s_{01})/2$ and $s\hat{\delta}(1-\hat{\delta})=\frac{s_{10}+s_{01}}{2}-\frac{(s_{10}+s_{01})^{2}}{4s}$. Approximating to the normal distribution and applying the Central Limit Theorem, the statistic for the hypothesis test of equality of the two sensitivities is

$$z_{UM} = \frac{s_{10} - s\hat{\delta}}{\sqrt{s\hat{\delta}\left(1-\hat{\delta}\right)}} = \frac{s_{10} - s_{01}}{\sqrt{\frac{\left(s_{10}+s_{01}\right)\left(s+s_{11}+s_{00}\right)}{s}}},$$

or in terms of the chi-square distribution

$$\chi^2_{UM} = \frac{s\left(s_{10} - s_{01}\right)^2}{\left(s_{10}+s_{01}\right)\left(s+s_{11}+s_{00}\right)}, \tag{30}$$

whose distribution is asymptotically a chi-square with one degree of freedom. In a similar way, we obtain the statistic to compare the two specificities:

$$\chi^2_{UM} = \frac{r\left(r_{10} - r_{01}\right)^2}{\left(r_{10}+r_{01}\right)\left(r+r_{11}+r_{00}\right)}. \tag{31}$$

2.1.11. Unconditional Likelihood Ratio Test (*ULRT*)

Lu [19] also proposed a likelihood ratio test statistic to compare two paired binomial proportions without discarding the concordant frequencies. The likelihood ratio test statistic is obtained in two phases: in the first phase we obtain the likelihood ratio test statistic when the four frequencies s_{ij} are combined in two, s_{10} and $s_{11}+s_{01}+s_{00}$; and in the second phase we obtain the likelihood ratio test statistic when the four frequencies s_{ij} are combined in another two, s_{01} and $s_{11}+s_{10}+s_{00}$. Finally, the likelihood ratio test statistic is calculated as an average of the two likelihood ratio test statistics. In the context studied here, the likelihood ratio test statistics are

$$\begin{aligned}\chi^2_{ULR} = {} & s_{10}\ln\left(\frac{2s_{10}}{s_{10}+s_{01}}\right) + s_{01}\ln\left(\frac{2s_{01}}{s_{10}+s_{01}}\right) + \\ & \left(s-s_{10}\right)\ln\left[\frac{2\left(s-s_{10}\right)}{2s-s_{10}-s_{01}}\right] + \left(s-s_{01}\right)\ln\left[\frac{2\left(s-s_{01}\right)}{2s-s_{10}-s_{01}}\right]\end{aligned} \tag{32}$$

and

$$\begin{aligned}\chi^2_{ULR} = {} & r_{10}\ln\left(\frac{2r_{10}}{r_{10}+r_{01}}\right) + r_{01}\ln\left(\frac{2r_{01}}{r_{10}+r_{01}}\right) + \\ & \left(r-r_{10}\right)\ln\left[\frac{2\left(r-r_{10}\right)}{2r-r_{10}-r_{01}}\right] + \left(r-r_{01}\right)\ln\left[\frac{2\left(r-r_{01}\right)}{2r-r_{10}-r_{01}}\right],\end{aligned} \tag{33}$$

which in both cases asymptotically follow a chi-square with one degree of freedom.

Appendix *B*

The choice of the method with the best performance of the global type I error was made comparing the fluctuations of this error around the nominal error $\alpha = 5\%$. For the methods based on exact tests, it was considered that the method has a good type I error performance when the type I error is $\leq 5\%$. For the methods based on asymptotic tests, it was considered that the method has a good global type I error performance when the global type I error fluctuates around 5% without going too far above it, a situation considered when the global type I error is $\geq 7\%$. This difference in criteria between exact methods and asymptotic methods is based on the idea that we should not consider that both types of methods (exact and asymptotic) must have the same type I error performance: an exact method must not go above the nominal error (it must not give false significances) whereas an asymptotic method may go above the nominal error without giving false significances.

For asymptotic methods, the cut-off point of 7% is due to the relation that exists between the asymptotic hypothesis tests and their confidence intervals, and it has been used by different authors to compare the asymptotic performance of approximate intervals [20, 21, 22, 23]. Therefore, these authors have established the following criterion for the choice of an optimum confidence interval (to 95% confidence): the probability of the coverage of the confidence interval being higher than 93%, in which case it is said that the interval does not fail. We define $\Delta\alpha = \alpha - \alpha^* = \gamma^* - \gamma$, where $\gamma = 1 - \alpha = 0.95$ is the nominal confidence of the confidence interval and γ^* the coverage probability calculated. A confidence interval fails if its coverage probability is $\leq 93\%$, i.e. if $\Delta\alpha \leq -2$. In this situation, the type I error of the corresponding two-tailed hypothesis test is $\geq 7\%$, and therefore it is an excessively liberal hypothesis test and one which gives false significances. If $\Delta\alpha > 2\%$, i.e. the coverage probability is higher than 97%, then the corresponding hypothesis test is very conservative (its type I error is very small, $< 3\%$).

APPENDIX II

Comparison of the likelihood ratios of two diagnostic tests subject to a paired design: confidence intervals and sample size

Roldán-Nofuentes, J.A., Sidaty-Regad S.B. (2020). Comparison of the likelihood ratios of two diagnostic tests subject to a paired design: confidence intervals and sample size. REVSTAT-Statistical Journal. Accepted, in press.

Category: Statistics and Probability. JCR 2019 (last published): 0.667. Rank: 101/124. Quartile: Q4.

Abstract

Positive and negative likelihood ratios are parameters which are used to assess and compare the effectiveness of binary diagnostic tests. Both parameters only depend on the sensitivity and specificity of the diagnostic test and are equivalent to a relative risk. This article studies the comparison of the likelihood ratios of two binary diagnostic tests subject to a paired design through confidence intervals. Six approximate confidence intervals are presented for the ratio of the likelihood ratios, and simulation experiments are carried out to study the coverage probabilities and the average lengths of the intervals considered, and some general rules of application are proposed. A method is also proposed to determine the sample size necessary to estimate the ratio between the likelihood ratios with a determined precision. The results were applied to two real examples.

Keywords: Likelihood ratios, binary diagnostic test, sample size.

Mathematics Subject Classification: 62P10, 6207.

1. Introduction

A diagnostic test is a medical test that is applied to an individual in order to determine the presence or absence of a disease. When the result of a diagnostic test is positive or negative, the diagnostic test is called a binary diagnostic test (*BDT*). A stress test for the diagnosis of coronary disease is an example of *BDT*. The effectiveness of a *BDT* is measured in terms of two fundamental parameters: sensitivity and specificity. The sensitivity (*Se*) is the probability of the *BDT* being positive when the individual has the disease, and the specificity (*Sp*) is the probability of the *BDT* being negative when the individual does not have it. The *Se* and the *Sp* of a *BDT* are estimated in relation to a gold standard (*GS*), which is a medical test which objectively determines whether or not an individual has the disease or not. An angiography for coronary disease is an example of *GS*. Other parameters that are used to assess the effectiveness of a *BDT* are the likelihood ratios (*LRs*) (Pepe, 2003; Zhou et al, 2011). When the *BDT* is positive, the likelihood ratio, called the positive likelihood ratio $\left(LR^{+}\right)$, is the ratio between the probability of correctly classifying an individual with the disease and the probability of incorrectly classifying an individual who does not have it. When the *BDT* is negative, the likelihood ratio, called the negative likelihood ratio $\left(LR^{-}\right)$, is the ratio between the probability of incorrectly classifying an individual who has the disease and the probability of correctly classifying an individual who does not have it. The *LRs* only depend on the sensitivity and the specificity of the *BDT* and do not depend on the disease prevalence, and therefore the *LRs* are superior parameters of the accuracy of a *BDT* (Zhou et al, 2011).

The comparison of the parameters of two *BDTs* has been the subject of numerous studies in Statistical literature. When the two *BDTs* and the *GS* are applied to all of the individuals in a random sample sized *n* (paired design), the comparison of the two sensitivities (specificities) is made by applying a comparison test of two paired binomial proportions. Subject to this same sample design, the comparison of the *LRs* of two *BDTs* is more complex. Leisenring and Pepe (1998) studied the estimation of the *LRs* of a *BDT* through a regression model. Pepe (2003) adapted this model to compare the *LRs* of two *BDTs*, for which in the regression model a variable dummy is considered to compare a *BDT* in relation to another. Moreover, Pepe proposed a confidence interval for the ratio of the two positive (negative) *LRs* estimating the variance of the ratios

subject to the null hypothesis of equality of the two *LRs*. Section 3.1 summarizes the method of Pepe (2003). Biggerstaff (2000) proposed a graphical method to compare the *LRs* of two (or more) *BDTs*. Nevertheless, this method is not inferential and can only be applied to the estimators. Roldán-Nofuentes and Luna (2007) studied hypothesis tests to compare the *LRs* individually and simultaneously, and they also studied the same problem for the case of ordinal diagnostic tests. The hypothesis tests proposed by Roldán-Nofuentes and Luna (2007) are based on the logarithmic transformation of the ratio of the positive (negative) *LRs*, and therefore by inverting the test statistics of the individual tests, confidence intervals are obtained for the ratio of the two *LRs* (in Section 3.2 we summarize this method). Dolgun et al (2012) extended the method of Leisenring and Pepe (1998) to compare the *LRs* simultaneously.

Comparing the sensitivities (specificities) of two *BDTs*, we compare the intrinsic accuracy of both *BDTs*, and we determined which *BDT* is more accurate for an individual who has the disease (which *BDT* has the greatest sensitivity) or for an individual who does not have the disease (which *BDT* has the greatest specificity). Comparing the positive (negative) *LRs* of two *BDTs* it is possible to quantify with which *BDT* it is more likely to obtain a positive (negative) result for the *BDT* for an individual who has the disease than for an individual who does not.

In this manuscript we study the comparison of the *LRs* of two *BDTs* through confidence intervals (*CIs*), making the following contributions: a) four intervals to compare the *LRs*, and b) a method to calculate the sample size to compare the *LRs* through *CIs*. Section 2 presents the *LRs* and their properties. Section 3 presents the *CIs* studied by Pepe (2003), by Roldán-Nofuentes and Luna (2007), and four new *CIs* are proposed: a Wald type interval, an interval based on the Fieller method, a bootstrap interval based on the bias-corrected interval, and a Bayesian interval based on non-informative beta distributions and on the application of the Monte Carlo method. In Section 4, simulation experiments are carried out to study the coverage probabilities and the average lengths of the *CIs* presented in Section 3. Section 5 presents a method to calculate the sample size to compare the *LRs* through *CIs*. In Section 6, the results are applied to two real examples, and in Section 7 the results obtained are discussed.

2. Likelihood ratios

Let us consider a *BDT* that is assessed in relation to a *GS*. Let *T* be the variable that models the result of the *BDT*: $T=1$ when the *BDT* is positive and $T=0$ when it is negative. Let *D* be the variable that models the result of the *GS*: $D=1$ when the individual has the disease and $D=0$ when this is not the case. Let $\pi = P(D=1)$ be the disease prevalence in the population studied, and $\bar{\pi} = 1-\pi$. The positive *LR* (Pepe, 2003; Zhou et al, 2011) is defined as

$$LR^{+} = \frac{P(T=1|D=1)}{P(T=1|D=0)} = \frac{Se}{1-Sp}, \tag{1}$$

and the negative *LR* as

$$LR^{-} = \frac{P(T=0|D=1)}{P(T=0|D=0)} = \frac{1-Se}{Sp}. \tag{2}$$

The *LR*s vary between 0 and infinity, and have the following properties:

a) If the *BDT* and the *GS* are independent then $LR^{+} = LR^{-} = 1$.

b) If the *BDT* correctly classifies all of the individuals then $LR^{+} = \infty$ and $LR^{-} = 0$.

c) If $LR^{+} > 1$ then a positive result in the *BDT* is more probable for an individual who has the disease than for an individual who does not.

d) If $LR^{-} < 1$ then a negative result in the *BDT* is more probable for an individual who does not have the disease than for an individual who does.

e) The *LR*s quantify the increase in knowledge of the presence of the disease through the application of the *BDT*. Before applying the test, the odds of an individual having the disease are $\text{pre-test odds} = \pi/(1-\pi)$, where π is the disease prevalence. After applying the *BDT*, the odds are $\text{post-test odds} = \frac{P(D=1|T=i)}{P(D=0|T=i)}$, $i=0,1$. The *LR*s relate the pre-test odds and the post-test odds:

$$\text{post test odds } (T=1) = LR^{+} \times \text{pre test odds}$$
$$\text{post test odds } (T=0) = LR^{-} \times \text{pre test odds.}$$

Therefore, the likelihood ratios quantify the change in the odds of the disease obtained by knowledge of the application of the *BDT*.

We then study the comparison of the *LRs* of two *BDTs* subject to a paired design through *CIs*.

3. Confidence intervals

Let us consider two *BDTs* that are assessed in relation to the same *GS*. Let T_h be the variable that models the result of the *h*th *BDT*, with $h = 1,2$, defined in a similar way to the variable *T* given in Section 2. Let Se_h and Sp_h be the sensitivity and the specificity of the *h*th *BDT*, and LR_h^+ and LR_h^- the positive and negative likelihood ratios respectively. Table 1 shows the frequencies and the theoretical probabilities obtained when comparing two *BDTs* in relation to a *GS* subject to a paired design. In the observed frequencies given in Table 1, the only value set by the researcher is the sample size *n*.

Table 1. Frequencies and probabilities subject to a paired design.

Frequencies					
	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$D = 1$	s_{11}	s_{10}	s_{01}	s_{00}	s
$D = 0$	r_{11}	r_{10}	r_{01}	r_{00}	r
Total	$s_{11} + r_{11}$	$s_{10} + r_{10}$	$s_{01} + r_{01}$	$s_{00} + r_{00}$	n
Probabilities					
	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$D = 1$	p_{11}	p_{10}	p_{01}	p_{00}	π
$D = 0$	q_{11}	q_{10}	q_{01}	q_{00}	$\overline{\pi}$
Total	$p_{11} + q_{11}$	$p_{10} + q_{10}$	$p_{01} + q_{01}$	$p_{00} + q_{00}$	1

Applying the model of conditional dependence of Vacek (1985), the theoretical probabilities are expressed as

$$p_{ij} = \pi\left[Se_1^i \left(1 - Se_1\right)^{1-i} Se_2^j \left(1 - Se_2\right)^{1-j} + \delta_{ij}\varepsilon_1 \right] \quad (1)$$

and

$$q_{ij} = \overline{\pi}\left[Sp_1^{1-i}\left(1-Sp_1\right)^{i} Sp_2^{1-j}\left(1-Sp_2\right)^{j} + \delta_{ij}\varepsilon_0 \right], \tag{2}$$

when $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = -1$ if $i \neq j$, with $i, j = 0,1$, and verifying that $\pi = \sum_{ij} p_{ij}$ and $\overline{\pi} = \sum_{ij} q_{ij}$. The parameters ε_1 and ε_0 are the dependence factors between the two *BDTs* when $D = 1$ and when $D = 0$ respectively, verifying that $0 \le \varepsilon_1 \le Min\left\{Se_1\left(1-Se_2\right), Se_2\left(1-Se_1\right)\right\}$ and $0 \le \varepsilon_0 \le Min\left\{Sp_1\left(1-Sp_2\right), Sp_2\left(1-Sp_1\right)\right\}$. If $\varepsilon_1 = \varepsilon_0 = 0$ then the two *BDTs* are conditionally independent from the disease, which is not normally a realistic one. In practice, the *BDTs* are conditionally dependent on the disease, so that $\varepsilon_1 > 0$ and/or $\varepsilon_0 > 0$. The frequencies of Table 1 are the product of a multinomial distribution whose vector of probabilities is $\boldsymbol{\psi} = \left(p_{11}, p_{10}, p_{01}, p_{00}, q_{11}, q_{10}, q_{01}, q_{00}\right)^T$. The maximum likelihood estimators of these probabilities are $\hat{p}_{ij} = s_{ij}/n$ and $\hat{q}_{ij} = r_{ij}/n$, those of π and $\overline{\pi}$ are $\hat{\pi} = s/n$ and $\hat{\overline{\pi}} = r/n$, and the variance-covariance matrix of $\hat{\boldsymbol{\psi}}$ is $\Sigma_{\hat{\psi}} = \left\{\mathrm{diag}\left(\boldsymbol{\psi}\right) - \boldsymbol{\psi}\boldsymbol{\psi}^T\right\}/n$.

In terms of the probabilities of the vector $\boldsymbol{\psi}$, the sensitivity and the specificity of each *BDT* are written as $Se_1 = \left(p_{10} + p_{11}\right)/\pi$, $Sp_1 = \left(q_{00} + q_{01}\right)/\overline{\pi}$, $Se_2 = \left(p_{01} + p_{11}\right)/\pi$ and $Sp_2 = \left(q_{00} + q_{10}\right)/\overline{\pi}$. The estimators of the sensitivities and the specificities are $\hat{Se}_1 = \frac{s_{11} + s_{10}}{s}$, $\hat{Se}_2 = \frac{s_{11} + s_{01}}{s}$, $\hat{Sp}_1 = \frac{r_{01} + r_{00}}{r}$ and $\hat{Sp}_2 = \frac{r_{10} + r_{00}}{r}$, and those of the dependence factors are $\hat{\varepsilon}_1 = \frac{\hat{p}_{11}}{\hat{\pi}} - \hat{Se}_1\hat{Se}_2 = \frac{s_{11}s_{00} - s_{10}s_{01}}{s}$ and $\hat{\varepsilon}_0 = \frac{\hat{q}_{00}}{\hat{\overline{\pi}}} - \hat{Sp}_1\hat{Sp}_2 = \frac{r_{11}r_{00} - r_{10}r_{01}}{r}$. Applying the delta method, it holds that the variances-covariances of $\hat{Se}_h$ and $\hat{Sp}_h$ are

$$\begin{gathered} Var\left(\hat{Se}_h\right) = \frac{Se_h\left(1-Se_h\right)}{n\pi},\ Var\left(\hat{Sp}_h\right) = \frac{Sp_h\left(1-Sp_h\right)}{n\overline{\pi}}, \\ Cov\left(\hat{Se}_1, \hat{Se}_2\right) = \frac{\varepsilon_1}{n\pi},\ Cov\left(\hat{Sp}_1, \hat{Sp}_2\right) = \frac{\varepsilon_0}{n\overline{\pi}}. \end{gathered} \tag{5}$$

The rest of the covariances are zero. Regarding the *LRs*, applying the delta method again, their variances-covariances (the proof can be seen in Appendix A) are

$$
\begin{aligned}
Var\left(\hat{LR}_h^+\right) &\approx \frac{Se_h^2 Var\left(\hat{Sp}_h\right) + \left(1-Sp_h\right)^2 Var\left(\hat{Se}_h\right)}{\left(1-Sp_h\right)^4}, \\
Var\left(\hat{LR}_h^-\right) &\approx \frac{\left(1-Se_h\right)^2 Var\left(Sp_h\right) + Sp_h^2 Var\left(\hat{Se}_h\right)}{Sp_h^4}, \\
Cov\left(\hat{LR}_1^+, \hat{LR}_1^+\right) &\approx \frac{Se_1 Se_2 Cov\left(\hat{Sp}_1, \hat{Sp}_2\right) + \left(1-Sp_1\right)\left(1-Sp_2\right) Cov\left(\hat{Se}_1, \hat{Se}_2\right)}{\left(1-Sp_1\right)^2\left(1-Sp_2\right)^2}, \\
Cov\left(\hat{LR}_1^-, \hat{LR}_1^-\right) &\approx \frac{\left(1-Se_1\right)\left(1-Se_2\right) Cov\left(\hat{Sp}_1, \hat{Sp}_2\right) + Sp_1 Sp_2 Cov\left(\hat{Se}_1, \hat{Se}_2\right)}{Sp_1^2 Sp_2^2}.
\end{aligned}
\tag{6}
$$

Substituting in the previous expressions the parameters with their estimators, we obtain the expressions of the estimators of the variances-covariances. Pepe (2003) studied the comparison of the *LRs* considering the ratio between them, i.e. $\omega^+ = LR_1^+ / LR_2^+$ and $\omega^- = LR_1^- / LR_2^-$. Roldán-Nofuentes and Luna (2007) considered the Napierian logarithm of ω. In this study, we are going to follow the same criteria as Pepe, and therefore we are going to compare the *LRs* through *CIs* for ω^+ and ω^-. From here onwards, we are going to consider that LR_h is LR_h^+ or LR_h^-, and that ω is ω^+ or ω^-, depending on whether we compare the positive *LRs* or the negative *LRs*. If the *CI* for ω contains the value one, then we do not reject the equality of the *LRs* of both *BDTs*; in the opposite case, the *LR* of a *BDT* is significantly higher than that of the other *BDT*. Applying the delta method (see Appendix A), the variance of $\hat{\omega}$ is

$$
Var\left(\hat{\omega}\right) \approx \omega^2 \left[\frac{Var\left(\hat{LR}_1\right)}{LR_1^2} + \frac{Var\left(\hat{LR}_2\right)}{LR_2^2} - \frac{2Cov\left(\hat{LR}_1, \hat{LR}_2\right)}{LR_1 LR_2} \right]. \tag{7}
$$

Then six *CIs* are presented for each ratio ω^+ and ω^-. The first interval was proposed by Pepe (2003), the second is deduced from the study by Roldán-Nofuentes and Luna (2007), and the rest of the intervals are contributions made by this manuscript.

3.1. Regression model

Leisenring and Pepe (1998) studied the estimation of the *LRs* of a *BDT* in presence of covariates through a regression model. For the positive *LR*, the regression model with *p*

covariates is $\ln\left(LR^{+}\left(X_1\right)\right)=\beta_0+\sum_{i=1}^{p}\beta_i X_{1p}$, where β_i are the parameters of the model and $X_1=\left(X_{11},...,X_{1p}\right)$ is the matrix of covariates. This model can be used to compare two *BDTs* (Pepe, 2003), i.e. $\ln\left[LR^{+}\left(X_T\right)\right]=\beta_0+\beta_1 X_T$, where X_T is a variable dummy to compare a *BDT* in relation to another. The regression model to compare the two negative *LRs* is $\ln\left[LR^{+}\left(X_T\right)\right]=\alpha_0+\alpha_1 X_T$. In these models, the ratio ω^{+} is estimated as $e^{\hat{\beta}_1}$ and the ratio ω^{-} as $e^{\hat{\alpha}_1}$. The confidence interval for ω^{+} is

$$\hat{\omega}^{+}\times\exp\left\{\pm z_{1-\alpha/2}\sqrt{\hat{V}ar_0\left[\ln\left(\hat{\omega}^{+}\right)\right]}\right\}, \tag{8}$$

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

$$\hat{V}ar_0\left[\ln\left(\hat{\omega}^{+}\right)\right]\approx\frac{1-\hat{S}e_1}{s\hat{S}e_1}+\frac{\hat{S}p_1}{r\left(1-\hat{S}p_1\right)}+\frac{1-\hat{S}e_2}{s\hat{S}e_2}+\frac{\hat{S}p_2}{r\left(1-\hat{S}p_2\right)}$$

is the estimated variance of $\hat{\omega}^{+}$ subject to the null hypothesis $H_0: LR_1^{+}=LR_2^{+}$. The confidence interval for ω^{-} s similar to the previous one, where

$$\hat{V}ar_0\left[\ln\left(\hat{\omega}^{-}\right)\right]\approx\frac{\hat{S}e_1}{s\left(1-\hat{S}e_1\right)}+\frac{1-\hat{S}p_1}{r\hat{S}p_1}+\frac{\hat{S}e_1}{s\left(1-\hat{S}e_1\right)}+\frac{1-\hat{S}p_1}{r\hat{S}p_1}.$$

The book by Pepe (2003) discusses the confidence interval obtained from the regression model.

3.2. Logarithmic interval

Roldán-Nofuentes and Luna (2007) studied a hypothesis test to compare the positive (negative) *LRs* of two *BDTs* subject to a paired design. These hypothesis tests are based on the transformation of the Napierian logarithm of the ratio between the two positive (negative) *LRs*, i.e., $H_0:\ln(\omega)=0$ vs $H_1:\ln(\omega)\neq 0$, where ω is $\omega^{+}=LR_1^{+}/LR_2^{+}$ or $\omega^{-}=LR_1^{-}/LR_2^{-}$, and the test statistic is

$$\frac{\ln\left(\hat{\omega}\right)}{\sqrt{\hat{V}ar\left[\ln\left(\hat{\omega}\right)\right]}}\rightarrow N(0,1), \tag{9}$$

where $\hat{Var}\left[\ln\left(\hat{\omega}\right)\right]$ is an unrestricted estimator of the variance and is calculated applying the delta method (see Appendix A), i.e.

$$Var\left[\ln\left(\hat{\omega}\right)\right] \approx \frac{Var\left(\hat{LR}_1\right)}{LR_1^2} + \frac{Var\left(\hat{LR}_2\right)}{LR_2^2} - \frac{2Cov\left(\hat{LR}_1, \hat{LR}_2\right)}{LR_1 LR_2}, \tag{10}$$

and substituting in this expression each parameter with its estimator. Inverting the test statistic (9), it holds that the *CI* for $\ln\left(\omega\right)$ is $\ln\left(\hat{\omega}\right) \pm z_{1-\alpha/2}\sqrt{\hat{Var}\left[\ln\left(\hat{\omega}\right)\right]}$. Finally, the logarithmic *CI* for ω is

$$\hat{\omega} \times \exp\left\{\pm z_{1-\alpha/2}\sqrt{\hat{Var}\left[\ln\left(\hat{\omega}\right)\right]}\right\}. \tag{11}$$

Roldán-Nofuentes and Luna studied the size (and the power) of the test $H_0 : \ln\left(\omega\right) = 0$ through simulation experiments. As the logarithmic interval (11) is obtained by inverting the test statistic (9), the coverage probability of this interval is equal to 1 minus the type I error obtained in the simulations carried out by Roldán-Nofuentes and Luna, and therefore the results are equivalent.

3.3. Wald CI

The Wald interval (Wald, 1943) is a classic interval for a parameter. Assuming the asymptotic normality of $\hat{\omega}$, i.e. $\hat{\omega} \xrightarrow[n\to\infty]{} N\left[\omega, \mathrm{Var}\left(\omega\right)\right]$, the Wald *CI* for ω is

$$\hat{\omega}\left[1 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{Var}\left(\hat{LR}_1\right)}{\hat{LR}_1^2} + \frac{\hat{Var}\left(\hat{LR}_2\right)}{\hat{LR}_2^2} - \frac{2\hat{Cov}\left(\hat{LR}_1, \hat{LR}_2\right)}{\hat{LR}_1 \hat{LR}_2}}\right]. \tag{12}$$

3.4. Fieller CI

The Fieller method (1940) is a classic method used to calculate a *CI* for the ratio of two parameters, and requires us to assume that the estimators are distributed according to a normal bivariate distribution. Therefore, assuming the bivariant normality, i.e. $\left(\hat{LR}_1, \hat{LR}_2\right)^T \xrightarrow[n\to\infty]{} N\left[\left(LR_1, LR_2\right)^T, \Sigma_{\mathbf{LR}}\right]$, where

$$\Sigma_{\mathbf{LR}} = \begin{pmatrix} Var(LR_1) & Cov(LR_1, LR_2) \\ Cov(LR_1, LR_2) & Var(LR_2) \end{pmatrix},$$

and applying the Fieller method, it is verified that

$$\hat{LR}_1 - \omega\hat{LR}_2 \xrightarrow[n\to\infty]{} N\left(0, Var(LR_1) - 2\omega Cov(LR_1, LR_2) + \omega^2 Var(LR_2)\right).$$

The Fieller *CI* is obtained by searching for the set of values for ω that satisfy the inequality

$$\frac{\left(\hat{LR}_1 - \omega\hat{LR}_2\right)^2}{\hat{Var}\left(\hat{LR}_1\right) - 2\omega\hat{Cov}\left(\hat{LR}_1, \hat{LR}_2\right) + \omega^2\hat{Var}\left(\hat{LR}_2\right)} < z^2_{1-\alpha/2}.$$

Solving this inequation, the Fieller *CI* for ω is

$$\frac{\hat{LR}_1\hat{LR}_2 - \hat{\sigma}_{12}z^2_{1-\alpha/2} \pm \sqrt{\left(\hat{LR}_1\hat{LR}_2 - \hat{\sigma}_{12}z^2_{1-\alpha/2}\right)^2 - \left(\hat{LR}_1^2 - \hat{\sigma}_{11}z^2_{1-\alpha/2}\right)\left(\hat{LR}_2^2 - \hat{\sigma}_{22}z^2_{1-\alpha/2}\right)}}{\left(\hat{LR}_2^2 - \hat{\sigma}_{22}z^2_{1-\alpha/2}\right)}, \quad (13)$$

where $\hat{\sigma}_{ii} = \hat{Var}\left(\hat{LR}_i\right)$ and $\hat{\sigma}_{12} = \hat{Cov}\left(\hat{LR}_1, \hat{LR}_2\right)$. This interval is valid when $\left(\hat{LR}_1\hat{LR}_2 - \hat{\sigma}_{12}z^2_{1-\alpha/2}\right)^2 > \left(\hat{LR}_1^2 - \hat{\sigma}_{11}z^2_{1-\alpha/2}\right)\left(\hat{LR}_2^2 - \hat{\sigma}_{22}z^2_{1-\alpha/2}\right)$ and $\hat{LR}_2^2 - \hat{\sigma}_{22}z^2_{1-\alpha/2} \neq 0$.

3.5. Bootstrap CI

The Bootstrap method is one which is widely used for the estimation of parameters. The Bootstrap *CI* is calculated generating *B* random samples with replacement from the sample sized *n*, and then a *CI* is calculated. For the interval, we considered the bias-corrected Bootstrap *CI* (Efron and Tibshirani, 1993). For each one of the *B* samples with replacement, we calculate the estimators of the *LR*s and of ω, i.e. $\hat{LR}_{1Bi}$, $\hat{LR}_{2Bi}$ and $\hat{\omega}_{Bi}$, with $i = 1,...,B$. The parameter ω is estimated as the average of the *B* Bootstrap estimations, i.e. $\hat{\bar{\omega}}_B = \frac{1}{B}\sum_{i=1}^{B}\hat{\omega}_{Bi}$. Let $A = \#\left(\hat{\omega}_{Bi} < \hat{\omega}\right)$ be the number of samples in which the Bootstrap estimator $\hat{\omega}_{Bi}$ is lower than the maximum likelihood estimator $\hat{\omega}$. Let $\hat{z}_0 = \Phi^{-1}\left(A/B\right)$, where $\Phi^{-1}(\cdot)$ is the inverse function of the standard

normal cumulative distribution function. Let $q_1 = \Phi\left(2\hat{z}_0 - z_{1-\alpha/2}\right)$ and $q_2 = \Phi\left(2\hat{z}_0 + z_{1-\alpha/2}\right)$, then the bias-corrected Bootstrap *CI* is

$$\left(\hat{\omega}_B^{(q_1)}\ ,\ \hat{\omega}_B^{(q_2)}\right) \tag{14}$$

where $\hat{\omega}_B^{(q)}$ is the *q*th quantile of the distribution of the *B* Bootstrap estimations of ω. The bias-corrected bootstrap *CI* is consistent, as it verifies (Shao and Tu, 1995) that $P\left[\sqrt{n}\left(\hat{\omega}_n - \omega\right) \le x\right] - P_B\left[\sqrt{n}\left(\hat{\omega}_{B,n} - \hat{\omega}_n\right) \le x\right]$ converges in probability to zero when the sample size is very large $\left(n \to \infty\right)$ for every value *x*, where P_B is the bootstrap distribution and $\hat{\omega}_{B,n}$ is the is the upper (lower) limit of the bootstrap *CI*.

3.6. Bayesian CI

The previous *CIs* are all frequentists, the problem can also be addressed from a Bayesian perspective. Conditioning on $D=1$, i.e. on the individuals who have the disease, it is verified that $s_{11} + s_{10} \to B\left(s, Se_1\right)$ and that $s_{11} + s_{01} \to B\left(s, Se_2\right)$. Conditioning on $D=0$ it is verified that $r_{01} + r_{00} \to B\left(r, Sp_1\right)$ and that $r_{10} + r_{00} \to B\left(r, Sp_2\right)$. Considering the distribution of the *BDT* 1, the estimators of its sensitivity and specificity are $\hat{Se}_1 = \dfrac{s_{11} + s_{10}}{s}$ and $\hat{Sp}_1 = \dfrac{r_{01} + r_{00}}{r}$, which are estimators of binomial proportions. In a similar way, the estimators $\hat{Se}_2 = \dfrac{s_{11} + s_{01}}{s}$ and $\hat{Sp}_2 = \dfrac{r_{10} + r_{00}}{r}$ are also estimators of binomial proportions. Therefore, for these estimators, conjugate beta prior distributions are proposed, i.e.

$$\hat{Se}_h \to Beta\left(\alpha_{Se_h}, \beta_{Se_h}\right) \text{ and } \hat{Sp}_h \to Beta\left(\alpha_{Sp_h}, \beta_{Sp_h}\right), \tag{15}$$

with $h = 1,2$. Let $\mathbf{n} = \left(s_{11}, s_{10}, s_{01}, s_{00}, r_{11}, r_{10}, r_{01}, r_{00}\right)$ be the vector of observed frequencies, then the posteriori distributions for the estimators of the sensitivity and the specificity of the *BDT* 1 are

$$\hat{Se}_1 \left| \mathbf{n} \to Beta\left(s_{11} + s_{10} + \alpha_{Se_1}, s_{01} + s_{00} + \beta_{Se_1}\right)\right. \tag{16}$$

and

$$\hat{Sp}_1 \mid \mathbf{n} \rightarrow Beta\left(r_{01} + r_{00} + \alpha_{Sp_1}, r_{11} + r_{10} + \beta_{Sp_1}\right). \tag{17}$$

In a similar way, the posteriori distributions for the estimators of the sensitivity and the specificity of the *BDT* 2 are

$$\hat{Se}_2 \mid \mathbf{n} \rightarrow Beta\left(s_{11} + s_{01} + \alpha_{Se_2}, s_{10} + s_{00} + \beta_{Se_2}\right) \tag{18}$$

and

$$\hat{Sp}_2 \mid \mathbf{n} \rightarrow Beta\left(r_{10} + r_{00} + \alpha_{Sp_2}, r_{11} + r_{01} + \beta_{Sp_2}\right). \tag{19}$$

Once all the distributions have been defined, the posteriori distribution for the *LR*s of each *BDT*, and for ω^+ and ω^-, can be approximated by applying the Monte Carlo method (Boos and Stefanski, 2013). This method consists of generating *M* random values of the posteriori distributions given in equations (16) to (19). In each interaction the generated values of sensitivities $\left(\hat{Se}_{hi}\right)$ and specificities $\left(\hat{Sp}_{hi}\right)$ are plugged in the equations $\hat{LR}^+_{hi} = \dfrac{\hat{Se}_{hi}}{1-\hat{Sp}_{hi}}$ and $\hat{LR}^-_{hi} = \dfrac{1-\hat{Se}_{hi}}{\hat{Sp}_{hi}}$, and from these each ratio $\hat{\omega}_i$ is calculated. As an estimator of each ratio the average of the *M* Bayesian estimations is calculated, i.e. $\hat{\bar{\omega}}_{Ba} = \dfrac{1}{M}\sum_{i=1}^{M}\hat{\omega}_i$. Finally, from the *M* values $\hat{\omega}_i$ a *CI* based on the quantiles is calculated, i.e. the $100\times(1-\alpha)\%$ *CI* for ω is

$$\left(\hat{\omega}_{Ba}^{(\alpha/2)}\ ,\ \hat{\omega}_{Ba}^{(1-\alpha/2)}\right), \tag{20}$$

where $\hat{\omega}_{Ba}^{(q)}$ is the qth quantile of the distribution of the *M* Bayesian estimations $\hat{\omega}_i$.

All of the *CI*s presented are for $\omega = LR_1/LR_2$. If we want to calculate the *CI* for LR_2/LR_1 $\left(=\omega' = 1/\omega\right)$, the regression, logarithmic, Fieller, Bootstrap and Bayesian intervals are obtained by calculating the inverse of each boundary of the corresponding interval for ω. Nevertheless, the Wald *CI* for ω' is obtained from the Wald *CI* for ω dividing each boundary by $\hat{\omega}^2$, i.e. if $\left(L_\omega, U_\omega\right)$ is the Wald *CI* for ω then the Wald *CI* for $\omega' = 1/\omega$ is $\left(L_\omega/\hat{\omega}^2, U_\omega/\hat{\omega}^2\right)$.

4. Simulation experiments

Monte Carlo simulation experiments were carried out to study the coverage probability (*CP*) and the average length (*AL*) of each one of the *CIs* presented in the Section 3. For this purpose, $N = 10,000$ random samples of multinomial distributions with sizes $n = \{50,100,200,300,400,500,1000\}$ were generated, and their probabilities were calculated from equations (1) and (2). As the sensitivity and the specificity of each *BDT*, the values $Se_h, Sp_h = \{0.70, 0.75, ..., 0.90, 0.95\}$ were taken, which are realistic values in clinical practice, and the *LRs* were calculated with the equations $LR_h^+ = Se_h / (1 - Sp_h)$ and $LR_h^- = (1 - Se_h) / Sp_h$ with $h = 1,2$. For the disease prevalence, $\pi = \{10\%, 25\%, 50\%\}$ was considered, and for the dependence factors ε_1 and ε_0 intermediate values (50% of the maximum value of each ε_i) and high values (80% of the maximum value of each ε_i) were taken, i.e.

$$\varepsilon_1 = k \times Min\{Se_1(1 - Se_2), Se_2(1 - Se_1)\} \text{ and } \varepsilon_0 = k \times Min\{Sp_1(1 - Sp_2), Sp_2(1 - Sp_1)\},$$

where $k = \{0.50, 0.80\}$. Once the value of the parameters in each scenario was set, the probabilities of each multinomial distribution were calculated by substituting the value of the parameters in equations (3) and (4).

For the Bootstrap interval, for each one of the *N* random samples generated, $B = 2,000$ replacement samples were generated in turn, and from the *B* replacement samples the bias-corrected bootstrap *CI* was calculated through the method described in Section 3.5.

Regarding the Bayesian *CI*, for the estimators of the two sensitivities and of the two specificities, the $Beta(1,1)$ distribution was considered as prior distribution. The choice of this distribution is justified by the fact that it is a non-informative distribution, which is flat for every possible value of the sensitivities and the specificities, and it has a minimum impact on the posteriori distributions. Moreover, for each one of the *N* generated random samples, $M = 10,000$ random samples were generated in turn, and from the *M* samples the Bayesian *CI* was been calculated by applying the method described in Section 3.6.

The simulation experiments were designed so that in every random sample generated, it is possible to estimate all the parameters and their variances-covariances. Therefore, if a parameter could not be estimated in a sample (for example, $\hat{Se}_h = 0$) then that sample was discarded and another one was generated in its place. This problem mainly occurred in the samples with a size of 50. In each one of the scenarios considered (values set for Se_h, Sp_h, π, ε_1 and ε_0) the coverage probability (*CP*) and the average length (*AL*) were calculated for each one of the six *CIs* for ω^+ and ω^-. The *CP* of each *CI* was calculated as the quotient between the number of intervals that contained the parameter (ω^+ or ω^-, depending on the case) and the number of samples generated *N*, and the *AL* was calculated adding the length of the *N* intervals and dividing this number by *N*. As the confidence level we took 95%.

The comparison of the asymptotic behaviour of the *CIs* was made following the criterion based on whether the *CI* "fails" or "does not fail" for a confidence of 95%. This criterion, which has been used by other authors (Price and Bonett, 2004; Martín-Andrés and Álvarez-Hernández, 2014a, 2014b; Montero-Alonso and Roldán-Nofuentes, 2018), establishes that a *CI* fails (or does not fail) if its coverage probability is $\leq 93\%$ $(>93\%)$. The selection of the *CI* with the best asymptotic behaviour was made through the following steps: 1) Choose the *CIs* with the fewest failures, and 2) Choose the *CIs* which are the most accurate, i.e. those with least *AL*, and among these those which have a *CP* closest to 95%. This method is justified in Appendix *B*.

4.1. Positive LRs

Tables 2 and 3 show some of the results obtained for the intervals of ω^+, considering two different scenarios of sensitivities and specificities. In these tables, failures are indicated in bold type. From the results of the experiments, the following conclusions are reached:

a) Regression method. The *CI* obtained applying the regression method does not fail, and it has a *CP* of 100% or very close to this value. In general terms, its *AL* is larger than that of the rest of the intervals.

b) Logarithmic CI. The logarithmic *CI* does not fail. In very general terms, when the sample size is small $(n=50)$ or moderate $(n=100)$ its *CP* is 100% or very near to this

value. When the sample size is large $(n = 200 - 400)$ or very large $(n \geq 500)$ its *CP* fluctuates around 95%. The *AL* of this interval is lower than that of the interval calculated through regression.

c) Wald CI. When $\omega^{+} \neq 1$, this interval may fail if $n \leq 100$ and the prevalence is moderate $(\pi = 25\%)$ or large $(\pi = 50\%)$, whereas if $n \geq 200$ the interval does not fail. When $\omega^{+} = 1$ the interval does not fail. In situations in which the Wald *CI* does not fail, its *CP* and *AL* are very similar to those of the logarithmic *CI*.

c) Fieller CI. The Fieller *CI* does not fail. In general terms, its *CP* is 100% or very close to this value when $n \leq 100$. When $n \geq 200$ its *CP* behaves in a very similar way to the *CP* of the logarithmic and Wald intervals (and the *ALs* are very similar). Therefore, when $n \geq 200$, the behaviour of the Fieller *CI* is very similar to the logarithmic and Wald intervals.

d) Bootstrap CI. In very general terms, when $n \leq 100$ this interval may fail if $\omega^{+} \neq 1$ or its *CP* is equal (or very near) to 100% if $\omega^{+} = 1$. When $n \geq 200$, the Bootstrap *CI* does not fail, its *CP* fluctuates around 95% and its *AL* is very similar to that of the logarithmic, Wald and Fieller intervals. Therefore, when $n \geq 200$ the Bootstrap interval has an asymptotic behaviour which is very similar to that of logarithmic, Wald and Fieller intervals.

Table 2. Coverage probabilities and average lengths of the *CIs* for the ratio of the two positive *LRs* (I).

$LR_1^+ = 9.5$ $LR_2^+ = 4.5$ $LR_1^- = 0.056$ $LR_2^- = 0.125$ $\omega^+ = 2.111$ $\omega^- = 0.444$

$Se_1 = 0.95$ $Sp_1 = 0.90$ $Se_2 = 0.90$ $Sp_2 = 0.80$

$\pi = 10\%$ $\varepsilon_1 = 0.0225$ $\varepsilon_0 = 0.0400$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	99.95	7.06	99.40	5.72	97.20	4.53	100	8.93	98.30	3.09	99.90	5.90
100	99.25	5.73	97.90	4.75	97.40	4.16	99.80	5.64	98.50	3.69	99.10	5.42
200	99.40	3.04	96.85	2.49	96.60	2.38	97.90	2.61	96.90	2.52	99.30	3.04
300	98.90	2.26	96.15	1.86	96.10	1.81	96.85	1.90	95.60	1.89	99.00	2.27
400	99.10	1.86	95.90	1.53	95.85	1.50	96.10	1.55	95.80	1.55	99.15	1.86
500	98.50	1.61	95.55	1.33	95.45	1.31	95.90	1.35	95.05	1.34	98.35	1.62
1000	98.20	1.07	95.45	0.89	95.30	0.88	95.65	0.89	95.35	0.90	98.20	1.08

$\pi = 10\%$ $\varepsilon_1 = 0.0360$ $\varepsilon_0 = 0.0640$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	99.95	6.54	99.10	4.78	95.50	3.93	99.95	7.74	91.80	2.51	99.95	5.48
100	99.90	5.15	98.60	3.76	96.55	3.39	99.45	4.57	95.60	2.72	99.90	4.91
200	99.60	2.93	96.90	2.09	96.00	2.01	98.15	2.19	96.35	1.95	99.55	2.93
300	99.65	2.21	96.30	1.57	95.90	1.53	97.25	1.61	96.00	1.53	99.60	2.22
400	99.80	1.82	95.90	1.30	95.95	1.28	97.10	1.32	96.30	1.28	99.85	1.83
500	99.75	1.59	95.80	1.13	95.75	1.12	96.35	1.15	95.65	1.13	99.80	1.60
1000	99.55	1.07	95.45	0.76	95.35	0.76	95.70	0.77	95.50	0.76	99.60	1.08

$\pi = 25\%$ $\varepsilon_1 = 0.0225$ $\varepsilon_0 = 0.0400$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	99.85	6.04	97.80	4.89	**91.30**	3.95	99.90	6.38	93.60	2.72	99.65	5.49
100	99.50	5.19	97.90	4.28	95.05	3.74	99.40	4.52	97.45	3.28	99.35	4.90
200	98.45	2.96	95.60	2.44	94.75	2.32	97.30	2.50	95.90	2.62	98.40	2.91
300	98.45	2.28	95.45	1.88	95.25	1.83	97.05	1.91	94.95	2.03	98.30	2.25
400	99.00	1.91	96.10	1.59	95.95	1.55	96.65	1.60	95.60	1.68	98.85	1.90
500	98.55	1.65	95.60	1.37	95.25	1.35	96.15	1.38	95.55	1.43	98.55	1.65
1000	98.30	1.14	95.15	0.95	94.90	0.94	95.30	0.95	94.65	0.97	98.35	1.14

$\pi = 25\%$ $\varepsilon_1 = 0.0360$ $\varepsilon_0 = 0.0640$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	5.77	96.80	4.21	**91.50**	3.50	99.65	5.56	**83.55**	2.20	100	5.25
100	99.85	4.45	95.40	3.19	**91.85**	2.88	97.15	3.45	**89.15**	2.31	99.80	4.25
200	99.60	2.85	96.15	2.02	94.00	1.95	96.40	2.08	94.85	1.93	99.60	2.80
300	99.40	2.23	94.15	1.59	94.10	1.55	95.15	1.62	94.10	1.60	99.40	2.21
400	99.55	1.87	94.95	1.32	94.85	1.30	95.15	1.34	94.65	1.35	99.50	1.85
500	99.15	1.66	94.85	1.18	94.75	1.16	95.70	1.19	95.05	1.21	99.15	1.65
1000	99.50	1.14	95.00	0.81	95.15	0.81	95.70	0.82	94.90	0.83	99.30	1.14

$\pi = 50\%$ $\varepsilon_1 = 0.0225$ $\varepsilon_0 = 0.0400$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	99.75	5.98	96.75	4.88	**89.35**	3.80	99.75	6.11	**86.45**	2.31	99.55	5.39
100	99.60	5.91	96.35	4.87	**92.20**	3.97	98.90	5.22	94.45	2.81	99.40	5.38
200	98.85	3.78	95.90	3.13	94.15	2.89	97.70	3.21	96.85	3.10	98.70	3.65
300	98.50	2.87	95.00	2.38	94.70	2.26	96.40	2.41	95.40	2.61	98.30	2.82
400	98.50	2.40	95.35	1.99	95.05	1.92	96.80	2.02	94.65	2.20	98.25	2.37
500	98.35	2.08	95.80	1.72	95.45	1.68	95.25	1.74	95.25	1.88	98.20	2.06
1000	97.50	1.41	94.55	1.17	94.80	1.15	95.50	1.17	93.80	1.22	97.60	1.40

$\pi = 50\%$ $\varepsilon_1 = 0.0360$ $\varepsilon_0 = 0.0640$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	99.90	5.47	94.15	4.03	**88.70**	3.28	99.20	5.26	**67.35**	1.89	99.80	4.97
100	99.85	5.20	93.80	3.80	**91.40**	3.22	96.65	4.24	**78.55**	2.13	99.75	4.79
200	99.70	3.45	93.75	2.47	93.65	2.32	93.70	2.56	**89.75**	2.15	99.45	3.34
300	99.55	2.72	94.65	1.93	94.45	1.86	94.65	1.98	94.10	1.90	99.55	2.67
400	99.65	2.33	95.15	1.66	94.90	1.62	95.45	1.69	95.35	1.69	99.65	2.31
500	99.45	2.06	95.55	1.46	95.15	1.43	95.25	1.48	96.00	1.51	99.20	2.04
1000	99.20	1.40	94.75	1.00	94.80	0.99	94.85	1.00	94.80	1.03	99.25	1.40

Table 3. Coverage probabilities and average lengths of the *CIs* for the ratio of the two positive *LRs* (II).

$LR_1^+ = 6$ $LR_2^+ = 6$ $LR_1^- = 0.118$ $LR_2^- = 0.118$ $\omega^+ = 1$ $\omega^- = 1$

$Se_1 = 0.90$ $Sp_1 = 0.85$ $Se_2 = 0.90$ $Sp_2 = 0.85$

$\pi = 10\%$ $\varepsilon_1 = 0.0450$ $\varepsilon_0 = 0.0638$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	99.95	3.61	99.50	2.51	99.85	2.18	100	4.67	100	1.96	99.95	3.16
100	99.80	2.38	97.75	1.65	97.90	1.52	98.85	2.37	98.60	1.51	99.75	2.33
200	99.65	1.33	96.40	0.92	96.90	0.89	97.65	1.02	97.00	0.91	99.60	1.35
300	99.65	1.00	96.25	0.70	96.45	0.68	97.90	0.74	96.75	0.69	99.70	1.01
400	99.65	0.84	95.60	0.58	96.00	0.58	96.95	0.61	96.10	0.58	99.65	0.84
500	99.50	0.72	95.30	0.51	95.70	0.50	96.35	0.52	95.70	0.51	99.60	0.73
1000	99.25	0.48	94.65	0.34	94.30	0.34	95.15	0.35	94.80	0.34	99.25	0.49

$\pi = 10\%$ $\varepsilon_1 = 0.0720$ $\varepsilon_0 = 0.1020$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	3.18	100	1.79	99.90	1.62	100	3.65	100	1.43	100	2.77
100	100	2.19	99.85	1.11	99.75	1.06	100	1.58	99.95	0.99	100	2.15
200	100	1.28	98.15	0.60	98.20	0.59	98.75	0.67	98.55	0.57	100	1.29
300	100	0.98	97.05	0.45	97.15	0.45	97.45	0.48	97.95	0.43	100	0.98
400	100	0.82	96.85	0.37	96.90	0.37	97.05	0.39	97.15	0.37	100	0.82
500	100	0.71	96.30	0.33	96.40	0.32	96.80	0.34	96.65	0.32	100	0.72
1000	100	0.49	95.80	0.22	95.80	0.22	96.15	0.22	96.32	0.22	100	0.49

$\pi = 25\%$ $\varepsilon_1 = 0.0450$ $\varepsilon_0 = 0.0638$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	99.90	3.24	99.35	2.25	99.55	1.97	100	3.58	99.95	1.81	99.85	3.06
100	99.65	2.05	96.95	1.39	96.95	1.30	100	1.78	99.15	1.38	99.75	2.00
200	99.30	1.24	95.00	0.86	94.85	0.84	98.45	0.94	95.00	0.90	99.15	1.23
300	99.70	0.97	94.45	0.68	94.10	0.66	97.35	0.71	94.20	0.70	99.65	0.96
400	99.45	0.82	95.55	0.57	94.85	0.57	97.10	0.60	95.05	0.59	99.35	0.82
500	99.45	0.73	94.70	0.51	94.15	0.50	96.15	0.53	94.25	0.52	99.40	0.72
1000	99.60	0.51	95.45	0.36	95.25	0.36	95.85	0.36	95.15	0.36	99.50	0.51

$\pi = 25\%$ $\varepsilon_1 = 0.0720$ $\varepsilon_0 = 0.1020$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	2.80	100	1.49	99.85	1.38	100	2.51	100	1.27	100	2.66
100	100	1.93	99.30	0.89	99.25	0.86	100	1.15	100	0.82	100	1.89
200	100	1.21	96.95	0.53	96.50	0.53	98.70	0.59	98.30	0.53	100	1.20
300	100	0.96	95.85	0.42	95.65	0.42	96.75	0.45	97.65	0.42	100	0.95
400	100	0.82	95.35	0.36	94.95	0.36	96.30	0.38	96.35	0.37	100	0.82
500	100	0.73	95.25	0.32	95.25	0.32	95.90	0.33	95.80	0.33	100	0.73
1000	100	0.50	95.25	0.22	95.25	0.22	95.70	0.23	95.40	0.23	100	0.50

$\pi = 50\%$ $\varepsilon_1 = 0.0450$ $\varepsilon_0 = 0.0638$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	99.95	3.27	99.95	2.27	99.60	1.97	100	3.54	100	1.67	99.95	3.06
100	100	2.51	98.90	1.69	97.65	1.52	100	2.39	99.85	1.50	99.85	2.39
200	99.55	1.54	95.60	1.06	94.30	1.01	98.80	1.22	96.45	1.12	99.35	1.51
300	99.35	1.20	96.00	0.83	95.10	0.81	97.70	0.90	95.65	0.86	99.25	1.19
400	99.55	1.02	95.40	0.71	95.40	0.69	96.10	0.75	95.55	0.74	99.50	1.01
500	99.55	0.89	95.20	0.62	94.75	0.61	96.20	0.65	94.15	0.64	99.50	0.89
1000	99.55	0.61	94.40	0.43	94.75	0.43	95.75	0.44	94.25	0.44	99.50	0.61

$\pi = 50\%$ $\varepsilon_1 = 0.0720$ $\varepsilon_0 = 0.1020$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	2.81	100	1.50	99.95	1.38	100	2.58	100	1.24	100	2.66
100	100	2.25	99.90	1.05	99.70	1.00	100	1.51	100	0.94	100	2.16
200	100	1.49	99.20	0.66	98.45	0.65	99.95	0.77	99.95	0.64	100	1.47
300	100	1.17	97.70	0.51	97.05	0.50	99.50	0.56	99.45	0.51	100	1.16
400	100	1.00	96.50	0.43	96.40	0.43	98.55	0.46	97.95	0.44	100	0.99
500	100	0.89	95.75	0.39	95.35	0.38	97.55	0.40	96.80	0.39	100	0.88
1000	99.95	0.61	95.55	0.27	95.25	0.27	96.65	0.28	95.60	0.27	99.95	0.61

e) Bayesian CI. The Bayesian *CI* does not fail and has a *CP* and an *AL* which are very similar to those of the interval obtained by regression. The *CP* and the *AL* of the Bayesian interval are almost always higher than those of the logarithmic, Wald, Fieller and Bootstrap intervals.

4.2. Negative LRs

Tables 4 and 5 show some of the results obtained for ω^- considering the same scenarios as for ω^+. Failures are indicated in bold type. From the results, the following conclusions are obtained:

a) Regression method. This interval has an asymptotic behaviour which is very similar to that of the same interval for ω^+.

b) Logarithmic CI. In general terms, this interval can fail when $\omega^+ \neq 1$ and the dependence factors are high, whatever the sample size may be. This interval does not fail when $\omega^+ = 1$, and its *CP* is 100% or very near to this value when $n \leq 100$, and even with $n \geq 200$ if the prevalence is small. When this interval does not fail, its *AL* is lower than that of the interval obtained through regression.

c) Wald CI. The Wald *CI* does not fail, and its *CP* is 100% (or very near) when $n \leq 100$, and its *CP* fluctuates around 95% when $n \geq 200$. The *AL* of the Wald *CI* is slightly lower than that of the logarithmic *CI* (when this does not fail), and its *CP* shows better fluctuations around 95% than that of the logarithmic interval.

c) Fieller CI. This interval does not show any failures. In very general terms, the Fieller *CI* has a very similar *CP* to that of the Wald *CI* when $\omega^+ \neq 1$. When $\omega^+ = 1$, the *CP* of the Fieller *CI* is 100% (or near) when $n \leq 100$, and fluctuates around 95% if $n \geq 200$. Its *AL* is greater than that of the Wald *CI,* especially when $n \leq 500$.

d) Bootstrap CI. This interval has many failures when $\omega^+ \neq 1$, especially when the prevalence is small or moderate, and regardless of the sample size. When $\omega^+ = 1$, the interval does not fail, and its *CP* is greater than that of the Wald *CI* or the logarithmic *CI,* especially when the prevalence is small or moderate. Regarding the Fieller *CI*, the *CP* of the Bootstrap interval is very similar to that of the Fieller interval, and its *AL* is slightly lower than that of the Fieller *CI*, especially for $n \leq 500$.

e) Bayesian CI. The same as for ω^+, the Bayesian *CI* for ω^- does not fail and has a *CP* and an *AL* which are very similar to those of the interval obtained through regression. The same as for ω^+, the *CP* and the *AL* of the Bayesian interval are higher than those of the logarithmic, Wald, Fieller and Bootstrap intervals.

4.3. Rules of application

Considering the asymptotic behaviour of each one of the *CIs* studied, it is possible to give some general rules of application for the *CIs* studied. These rules of application are for the different scenarios considered in the simulation experiments, scenarios that correspond to realistic values of prevalence, sensitivities and specificities in clinical practice. Based on the sample size, which in practice is the only parameter set by the researcher, the rules are the following:

a) For the ratio ω^+, use the logarithmic *CI*, whatever the sample size may be, although when $n \geq 200$ we can also use the Wald, the Fieller and the Bootstrap intervals.

b) For the ratio ω^-, use the Wald *CI,* whatever the sample size may be.

Table 4. Coverage probabilities (%) and average lengths of the *CIs* for the ratio of the two negative *LRs* (I).

$LR_1^+ = 9.5$ $LR_2^+ = 4.5$ $LR_1^- = 0.056$ $LR_2^- = 0.125$ $\omega^+ = 2.111$ $\omega^- = 0.444$

$Se_1 = 0.95$ $Sp_1 = 0.90$ $Se_2 = 0.90$ $Sp_2 = 0.80$

$\pi = 10\%$ $\varepsilon_1 = 0.0225$ $\varepsilon_0 = 0.0400$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	99.95	2.07	97.30	1.65	96.85	1.27	99.50	2.71	**14.80**	1.79	99.75	1.90
100	99.90	2.02	96.60	1.59	96.05	1.17	99.60	2.49	**35.50**	1.83	99.85	1.81
200	99.95	1.99	96.15	1.42	95.90	1.09	99.55	2.40	**53.70**	1.68	99.85	1.79
300	99.85	1.81	95.45	1.30	95.15	1.03	99.05	1.99	**75.95**	1.59	99.75	1.65
400	99.85	1.67	96.55	1.23	95.55	0.97	99.10	1.75	**86.05**	1.55	99.75	1.54
500	99.80	1.62	96.95	1.20	95.95	0.96	98.80	1.70	**88.80**	1.48	99.60	1.50
1000	99.55	1.22	96.90	0.93	95.90	0.81	97.85	1.16	95.80	1.05	99.45	1.16

$\pi = 10\%$ $\varepsilon_1 = 0.0360$ $\varepsilon_0 = 0.0640$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	2.18	**92.60**	1.63	99.95	1.31	99.50	2.83	**5.50**	1.66	100	2.01
100	100	2.11	**90.85**	1.53	98.90	1.19	99.25	2.48	**17.25**	1.70	100	1.91
200	100	2.16	**91.15**	1.38	98.35	1.12	99.25	2.57	**33.00**	1.53	100	1.96
300	99.95	1.94	**90.20**	1.21	97.60	1.01	98.10	2.02	**54.45**	1.43	99.90	1.78
400	99.95	1.76	**92.40**	1.13	97.10	0.95	97.65	1.64	**65.25**	1.39	99.90	1.63
500	99.90	1.68	**92.80**	1.09	96.10	0.91	97.85	1.55	**70.45**	1.35	99.85	1.56
1000	99.90	1.22	93.40	0.79	95.60	0.71	97.45	0.97	**84.65**	0.93	99.80	1.16

$\pi = 25\%$ $\varepsilon_1 = 0.0225$ $\varepsilon_0 = 0.0400$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	2.06	97.80	1.56	96.35	1.18	99.30	2.66	**34.05**	1.86	99.75	1.87
100	100	1.87	96.20	1.34	95.95	1.04	99.65	2.13	**64.85**	1.67	99.80	1.70
200	99.65	1.64	96.00	1.22	95.80	0.98	98.00	1.77	**89.30**	1.50	99.60	1.52
300	99.50	1.44	95.95	1.07	95.60	0.90	97.40	1.46	93.15	1.28	99.45	1.35
400	99.10	1.21	95.75	0.93	95.40	0.81	96.55	1.16	95.35	1.05	98.90	1.15
500	99.50	1.06	95.55	0.82	95.45	0.73	96.00	0.97	95.60	0.89	99.20	1.01
1000	98.60	0.65	95.20	0.52	95.15	0.50	94.65	0.55	95.55	0.52	98.45	0.64

$\pi = 25\%$ $\varepsilon_1 = 0.0360$ $\varepsilon_0 = 0.0640$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	2.13	**91.90**	1.48	99.90	1.19	99.30	2.60	**18.35**	1.71	99.95	1.95
100	100	2.07	**90.35**	1.29	99.00	1.08	98.45	2.31	**37.80**	1.53	99.95	1.89
200	99.85	1.71	**91.65**	1.09	96.55	0.92	97.40	1.58	**67.35**	1.35	99.80	1.59
300	99.85	1.48	**92.25**	0.95	96.35	0.82	97.15	1.28	**77.20**	1.14	99.75	1.39
400	99.85	1.26	**91.90**	0.81	95.90	0.72	96.85	1.02	**82.05**	0.94	99.85	1.20
500	99.85	1.06	**92.70**	0.69	95.70	0.63	96.35	0.80	**87.20**	0.77	99.65	1.02
1000	99.50	0.65	94.45	0.43	95.35	0.42	96.20	0.45	94.40	0.44	99.55	0.64

$\pi = 50\%$ $\varepsilon_1 = 0.0225$ $\varepsilon_0 = 0.0400$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	99.90	1.82	97.65	1.35	99.90	1.07	99.60	2.13	**71.70**	1.76	99.85	1.69
100	99.85	1.67	96.35	1.23	99.30	0.98	99.05	1.82	**84.60**	1.56	99.80	1.55
200	99.70	1.23	97.10	0.94	96.95	0.81	98.00	1.19	96.05	1.07	99.60	1.17
300	98.75	0.92	96.25	0.73	94.40	0.66	95.60	0.81	97.25	0.76	98.50	0.89
400	98.55	0.75	95.45	0.60	94.45	0.56	95.25	0.64	96.80	0.61	98.60	0.73
500	98.15	0.66	94.35	0.53	94.40	0.50	94.10	0.55	95.05	0.53	97.80	0.65
1000	98.65	0.44	95.20	0.35	95.20	0.35	94.80	0.36	94.35	0.36	98.45	0.43

$\pi = 50\%$ $\varepsilon_1 = 0.0360$ $\varepsilon_0 = 0.0640$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	1.90	**92.35**	1.25	99.30	1.04	98.35	2.01	**47.90**	1.60	99.95	1.77
100	100	1.74	**92.05**	1.11	97.80	0.93	97.80	1.63	**60.20**	1.43	99.95	1.62
200	100	1.26	93.55	0.82	96.30	0.73	97.45	1.02	**81.85**	0.97	99.90	1.20
300	99.65	0.94	94.70	0.62	95.15	0.58	96.65	0.70	**90.15**	0.67	99.50	0.91
400	99.70	0.77	94.55	0.51	95.30	0.48	95.95	0.54	93.10	0.52	99.50	0.75
500	99.75	0.65	95.30	0.44	95.20	0.42	95.85	0.46	94.80	0.44	99.55	0.64
1000	99.65	0.43	95.75	0.30	94.80	0.29	95.40	0.30	96.30	0.29	99.55	0.43

Table 5. Coverage probabilities (%) and average lengths of the *CIs* for the ratio of the two negative *LRs* (II).

$LR_1^+ = 6 \;\; LR_2^+ = 6 \;\; LR_1^- = 0.118 \;\; LR_2^- = 0.118 \;\; \omega^+ = 1 \;\; \omega^- = 1$

$Se_1 = 0.90 \;\; Sp_1 = 0.85 \;\; Se_2 = 0.90 \;\; Sp_2 = 0.85$

$\pi = 10\% \;\; \varepsilon_1 = 0.0450 \;\; \varepsilon_0 = 0.0638$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	2.55	100	1.84	99.50	1.55	100	3.35	100	1.72	100	2.34
100	100	2.54	100	1.74	98.85	1.44	99.95	3.04	100	1.65	100	2.33
200	100	2.52	100	1.58	95.90	1.36	99.90	3.01	100	1.56	100	2.32
300	100	2.48	100	1.52	93.85	1.34	99.60	2.70	100	1.52	100	2.31
400	100	2.39	99.65	1.51	93.15	1.32	99.20	2.53	100	1.51	99.90	2.26
500	100	2.35	99.65	1.43	94.35	1.31	99.05	2.45	100	1.50	100	2.25
1000	99.85	1.98	97.15	1.33	93.95	1.24	96.85	1.86	98.70	1.38	99.85	1.91

$\pi = 10\% \;\; \varepsilon_1 = 0.0720 \;\; \varepsilon_0 = 0.1020$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	2.73	100	1.76	99.90	1.56	100	3.45	100	1.39	100	2.51
100	100	2.68	100	1.56	99.80	1.40	100	2.84	100	1.34	100	2.49
200	100	2.65	100	1.42	99.80	1.30	100	2.78	100	1.23	100	2.40
300	100	2.61	100	1.28	98.60	1.19	99.95	2.62	100	1.12	100	2.33
400	100	2.52	100	1.19	97.70	1.11	98.90	2.05	100	1.07	100	2.27
500	100	2.42	100	1.13	97.10	1.07	97.95	1.85	100	1.03	100	2.18
1000	100	1.91	99.80	0.85	96.80	0.82	97.15	1.16	100	0.80	100	1.85

$\pi = 25\% \;\; \varepsilon_1 = 0.0450 \;\; \varepsilon_0 = 0.0638$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	2.56	100	1.72	98.20	1.46	100	3.23	100	1.67	100	2.40
100	100	2.51	100	1.53	95.45	1.35	99.85	2.91	100	1.55	99.95	2.35
200	99.95	2.40	99.50	1.50	93.90	1.31	98.90	2.57	99.95	1.53	99.90	2.20
300	99.85	2.26	98.55	1.48	94.65	1.25	98.00	2.35	99.75	1.47	99.80	2.15
400	99.70	1.98	96.95	1.33	93.05	1.19	96.20	1.85	98.20	1.37	99.55	1.92
500	99.55	1.74	95.20	1.18	92.35	1.10	94.55	1.50	96.40	1.24	99.40	1.70
1000	99.25	1.15	94.80	0.79	94.25	0.75	94.40	0.86	94.15	0.84	99.20	1.13

$\pi = 25\% \;\; \varepsilon_1 = 0.0720 \;\; \varepsilon_0 = 0.1020$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	2.86	100	1.57	99.80	1.41	100	3.11	100	1.40	100	2.65
100	100	2.81	100	1.37	98.90	1.26	100	2.95	100	1.22	100	2.54
200	100	2.45	100	1.14	97.75	1.07	100	1.86	100	1.04	100	2.30
300	100	2.22	99.90	1.01	97.20	0.97	99.80	1.54	100	0.93	100	2.10
400	100	1.92	96.95	0.86	96.80	0.83	98.55	1.17	99.95	0.80	100	1.85
500	100	1.69	96.55	0.74	96.45	0.72	98.15	0.94	99.90	0.71	100	1.65
1000	100	1.13	96.05	0.49	95.95	0.48	96.50	0.53	98.45	0.49	100	1.10

$\pi = 50\% \;\; \varepsilon_1 = 0.0450 \;\; \varepsilon_0 = 0.0638$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	2.45	100	1.59	95.50	1.40	99.90	2.80	100	1.65	99.95	2.31
100	99.95	2.42	99.25	1.56	94.70	1.35	99.00	2.69	99.95	1.55	99.90	2.25
200	99.80	2.01	96.90	1.34	93.30	1.25	96.20	1.89	98.85	1.37	99.70	1.94
300	99.65	1.57	96.75	1.08	94.30	1.05	96.30	1.29	97.20	1.15	99.60	1.54
400	99.70	1.32	95.40	0.91	94.65	0.88	95.20	1.02	95.20	0.97	99.70	1.30
500	99.70	1.17	95.10	0.81	94.90	0.78	94.70	0.88	94.20	0.85	99.65	1.15
1000	99.40	0.78	95.20	0.54	94.60	0.54	95.05	0.56	94.75	0.56	99.35	0.77

$\pi = 50\% \;\; \varepsilon_1 = 0.0720 \;\; \varepsilon_0 = 0.1020$

	Regression		Logarithmic		Wald		Fieller		Bootstrap		Bayesian	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
50	100	2.67	99.95	1.36	99.50	1.26	99.95	2.64	100	1.30	100	2.51
100	100	2.49	100	1.16	98.45	1.09	100	1.95	100	1.09	100	2.36
200	100	1.94	99.55	0.86	97.30	0.83	99.40	1.18	100	0.81	100	1.88
300	100	1.55	98.80	0.67	97.00	0.66	98.55	0.80	99.75	0.65	100	1.51
400	100	1.30	96.95	0.56	96.90	0.55	97.80	0.63	99.60	0.55	100	1.28
500	100	1.14	96.25	0.50	96.25	0.49	96.05	0.54	98.20	0.50	100	1.13
1000	100	0.78	95.35	0.34	95.10	0.34	95.35	0.35	95.30	0.35	100	0.77

5. Sample size

An important question when comparing two parameters is the calculation of the sample size necessary to compare the parameters with a determined error and power. In the context of the comparison of the *LR*s, Roldán-Nofuentes and Luna (2007) proposed a method to calculate the sample size to solve the hypothesis test $H_0:\ln(\omega)=0$ vs $H_1:\ln(\omega)\neq 0$. We then study the same problem but from the perspective of the *CI*s. Therefore, we study the problem of calculating the sample size necessary to estimate the ratio between the two *LR*s with a precision δ and a confidence $100(1-\alpha)\%$. As in the previous sections, we consider that ω is ω^+ or ω^-. Let us first consider the Wald *CI*, which can be applied both to estimate ω^+ (with $n\geq 200$) and ω^- (for any sample size). Based on the asymptotic normality of the estimator of ω, it is verified that $\hat{\omega}\in\omega\pm z_{1-\alpha/2}\sqrt{Var(\hat{\omega})}$, i.e. the probability of obtaining an estimator $\hat{\omega}$ is in this interval with a probability $100(1-\alpha)\%$. Let us consider that $LR_2 > LR_1$ and, therefore, that $\omega<1$ (the Wald interval will be lower than one) and let δ be the precision set by the researcher. As it has been assumed that $\omega<1$, then δ must be lower than one, and if we want to have a high level of precision then δ must be a small value. The sample size *n* is calculated from the expression

$$\delta = z_{1-\alpha/2}\omega\sqrt{\frac{Var\left(\hat{LR}_1\right)}{LR_1^2}+\frac{Var\left(\hat{LR}_2\right)}{LR_2^2}-\frac{2Cov\left(\hat{LR}_1,\hat{LR}_2\right)}{LR_1LR_2}}\,. \tag{21}$$

This equation is obtained from the Wald *CI* (equation (12)). Substituting the variances and the covariance with their respective expressions given in equations (6) and clearing *n* we obtain the expression of the sample size to estimate ω with a precision δ and a confidence $100(1-\alpha)\%$. For ω^+ the equation of the sample size is

$$n=\left(\frac{z_{1-\alpha/2}\omega^+}{\delta}\right)^2\left[\sum_{h=1}^{2}\left(\frac{1-Se_h}{\pi Se_h}+\frac{Sp_h}{\bar{\pi}(1-Sp_h)}\right)-\frac{2\varepsilon_1}{\pi Se_1Se_2}-\frac{2\varepsilon_0}{\bar{\pi}(1-Sp_1)(1-Sp_2)}\right], \tag{22}$$

and for ω^- is

$$n=\left(\frac{z_{1-\alpha/2}\omega^-}{\delta}\right)^2\left[\sum_{h=1}^{2}\left(\frac{Se_h}{\pi(1-Se_h)}+\frac{1-Sp_h}{\bar{\pi}Sp_h}\right)-\frac{2\varepsilon_1}{\pi(1-Se_1)(1-Se_2)}-\frac{2\varepsilon_0}{\bar{\pi}Sp_1Sp_2}\right]. \tag{23}$$

If it is considered that $\omega > 1$ (and consequently the Wald *CI* is higher than one) the *BDTs* can always be permuted and ω will then be lower than one. Another alternative consists of setting a value for a precision δ', in a similar way to the previous situation when $\omega < 1$, and then apply equation (22) or (23) considering $\delta = \hat{\omega}^2\delta'$. As is explained at the end of Section 3, this is due to the fact that if (L_ω, U_ω) is the Wald *CI* for $\omega = LR_1/LR_2 < 1$ then the Wald *CI* for $\omega' = 1/\omega = LR_2/LR_1$ is $\left(\frac{L_\omega}{\hat{\omega}^2}, \frac{U_\omega}{\hat{\omega}^2}\right)$. It is easy to check that the calculated value of the sample size *n* is the same both if $\omega < 1$ (with a precision δ) and if $\omega > 1$ (with precision $\delta = \hat{\omega}^2\delta'$).

In order to be able to apply the previous equations, it is necessary to know the sensitivities, the specificities (and therefore the *LRs,* ω^+ and ω^-), the dependence factors between the two *BDTs* (ε_i) and the prevalence (π). In practice, these values can be estimated from a pilot sample or can be obtained from another similar study. Therefore, the method to calculate the sample size requires us to know some estimations of the accuracy (*Se* and *Sp*) of each *BDT*, of the dependence factors between the *BDTs* and of the disease prevalence, obtained for example from a pilot study or from other previous studies. The method to calculate the size of the sample consists of the following steps:

Step 1. Take a pilot sample sized n_0 (in general terms, $n_0 \geq 200$ if ω^+ is estimated to then be able to calculate the Wald *CI*), and with this sample we calculate $\hat{Se}_h$, $\hat{Sp}_h$ (and therefore $\hat{LR}_h$, $\hat{\omega}^+$ and $\hat{\omega}^-$), $\hat{\varepsilon}_i$ and $\hat{\pi}$. The Wald *CI* for ω is then calculated, and if this interval has a precision δ, i.e. $z_{1-\alpha/2}\sqrt{\hat{Var}(\hat{\omega})} \leq \delta$, then the required precision has been reached; if not, go to the following step.

Step 2. Based on the estimations obtained in Step 1, calculate the sample size *n* applying equation (22) or (23).

Step 3. Take the sample of *n* individuals (add $n - n_0$ individuals to the initial pilot sample), and from this new sample we calculate $\hat{Se}_h$, $\hat{Sp}_h$, $\hat{\varepsilon}_i$, $\hat{\pi}$ and the Wald *CI*. If the Wald *CI* has a precision δ, then the set precision has been achieved; if not, consider the new sample to be a pilot sample $(n_0 = n)$ and go back to Step 1.

This proposed procedure to calculate the sample size is iterative, and therefore it does not guarantee that with the sample size calculated we can then estimate the parameter ω with the required precision. Moreover, if the researcher sets a precision δ^+ to estimate ω^+ and also sets a precision δ^- to estimate ω^-, once both sample sizes have been calculated through the previous method, the researcher must take a sample size of at least the maximum of the two sample sizes, to thus guarantee the precision in both estimations. In general, the calculation of the sample size makes sense when the confidence interval for ω does not contain the value one, since in this situation (the interval contains the value one) the equality of both *LRs* is not rejected and it does not make sense to determine how much larger one *LR* is compared to the other. Nevertheless, if the pilot sample is small (for example to estimate ω^-) and the Wald *CI* for ω^- contains the value 1, it may be useful to calculate the sample size to estimate the ω^-. In this situation, the Wald *CI* for ω^- will be very wide (as the pilot sample is small) and may contain the value 1 even if LR_1^- and LR_2^- are different.

The calculation of the sample size depends on the estimations obtained from an initial pilot sample. In order to study the effect that this sample has on the calculation of the sample size, simulation experiments were carried out which were similar to those carried out in Section 4. From the values of the parameters, we calculated the sample size *n* applying equation (22) or (23) depending on the case, taking a precision equal to 0.10, and we then generated $N = 10{,}000$ random samples with multinomial distributions sized n. In each one of the *N* random samples, we calculated the sample size n_i' from the estimators calculated with the random sample, and then calculated the average sample size $\overline{n} = \sum n_i' / N$ and the relative bias $RB(n') = (\overline{n} - n)/n$. Table 6 shows the results obtained for the scenarios considered in Tables 5 and 6 ($\omega \neq 1$). From the results, it holds that that the dependence factors ε_i have an important effect on the calculation of the sample size, and the sample size is smaller when the dependence factors are larger. Moreover, the increase in the prevalence means an increase (decrease) in the sample size to estimate ω^+ (ω^-). The relative biases obtained are very small, and therefore the sample sizes calculated from equations (22) and (23) are robust. Consequently, the initial pilot sample does not have an important effect on the determination of the sample size to estimate ω.

Table 6. Sample size to estimate ω.

$LR_1^+ = 9.5$ $LR_2^+ = 4.5$ $LR_1^- = 0.056$ $LR_2^- = 0.125$ $\omega^+ = 2.111$ $\omega^- = 0.444$			
$Se_1 = 0.95$ $Sp_1 = 0.90$ $Se_2 = 0.90$ $Sp_2 = 0.80$			
Sample size for ω^+			
$\varepsilon_1 = 0.0225$ $\varepsilon_0 = 0.0400$			
	$\pi = 10\%$	$\pi = 25\%$	$\pi = 50\%$
Sample size	958	1,073	1,571
Average sample size	981	1,084	1,597
Relative bias (%)	2.40	1.03	1.66
$\varepsilon_1 = 0.0360$ $\varepsilon_0 = 0.0640$			
	$\pi = 10\%$	$\pi = 25\%$	$\pi = 50\%$
Sample size	701	786	1,152
Average sample size	734	796	1,160
Relative bias (%)	4.71	1.27	0.69
Sample size for ω^-			
$\varepsilon_1 = 0.0225$ $\varepsilon_0 = 0.0400$			
	$\pi = 10\%$	$\pi = 25\%$	$\pi = 50\%$
Sample size	14,439	5,793	2,922
Average sample size	14,715	5,896	2,966
Relative bias (%)	1.91	1.78	1.51
$\varepsilon_1 = 0.0360$ $\varepsilon_0 = 0.0640$			
	$\pi = 10\%$	$\pi = 25\%$	$\pi = 50\%$
Sample size	10,336	4,147	2,092
Average sample size	10,482	4,186	2,118
Relative bias (%)	1.41	0.94	1.24

If the initial pilot sample has a small or moderate size, then in order to estimate ω^+ we use the logarithmic *CI*. In this situation, the process is similar to the previous one, and the sample size is calculated from the equation $\ln(\delta) = z_{1-\alpha/2}\sqrt{Var\left[\ln\left(\hat{\omega}^+\right)\right]}$, where the expression of $Var\left[\ln\left(\hat{\omega}^+\right)\right]$ is given in equation (10). Following a similar process to the previous one, it holds that

$$n = \left(\frac{z_{1-\alpha/2}}{\ln(\delta)}\right)^2 \left[\sum_{h=1}^{2}\left(\frac{1-Se_h}{\pi Se_h} + \frac{Sp_h}{\overline{\pi}\left(1-Sp_h\right)}\right) - \frac{2\varepsilon_1}{\pi Se_1 Se_2} - \frac{2\varepsilon_0}{\overline{\pi}\left(1-Sp_1\right)\left(1-Sp_2\right)}\right]. \quad (24)$$

6. Applications

The results obtained were applied to two real examples: a) a study of the diagnosis of coronary disease, and another study of the diagnosis of colorectal cancer.

6.1. Diagnosis of coronary disease

The results obtained were applied to the study by Weiner et al (1979) on the diagnosis of coronary disease, which is a widely used study to illustrate statistical methods for the estimation and comparison of parameters of *BDTs*. Weiner et al studied the diagnosis of coronary artery disease using as diagnostic tests the exercise test and the resting *EKG*, and the coronary arteriography as a *GS*. Table 7 shows the frequencies obtained by applying three medical tests to a sample of 1,465 males, where T_1 models the result of the exercise test, T_2 models the result of the resting *EKG* and *D* the result of the *GS*. Table 7 also shows the estimations of the *LRs* (ω) and their standard errors, as well as the *CIs* for ω^+ and ω^-.

For ω^+, from any of the six *CIs* (all of them are greater than one) it holds that the positive *LR* of the exercise test is significantly larger than the positive *LR* of the resting *EKG*, i.e. a positive result in the exercise test is more indicative of the presence of the disease than a positive result in the resting *EKG*. Interpreting the results of the logarithmic *CI*, the positive *LR* of the exercise test is (with a confidence of 95%) a value between 1.632 and 2.713 times larger than the positive *LR* of the resting *EKG*.

Regarding ω^-, all of the *CIs* intervals (all are less than one) we reject the equality of the two negative *LRs*, and it holds that a negative result for the resting *EKG* is more indicative of the absence of the disease than a negative result of the exercise test. Interpreting the Wald *CI*, the negative *LR* of the resting *EKG* is (with a confidence of 95%) a value between 2.872 $\left(=0.262/0.302^2\right)$ and 3.783 $\left(=0.345/0.302^2\right)$ times larger than the negative *LR* of the exercise test.

Moreover, in order to illustrate the method to calculate the sample size, we are going to consider that the researcher wants to estimate ω^+ with a precision equal to 0.10, which can be considered to be a high precision. The Wald *CI* for ω^+ is $(1.569\ ,\ 2.639)$, and therefore multiplying this interval by $1/\left(\hat{\omega}^+\right)^2 = 1/2.109^2$ it holds that the 95% Wald *CI* for $\omega'^+ = LR_2^+/LR_1^+$ is $(0.353\ ,\ 0.593)$, and the precision is 0.12. As 0.12 is higher than 0.10, it is necessary to increase the sample size to estimate ω^+ with the required precision. Setting the confidence at 95% and taking

$\delta = \left(\hat{\omega}^{+}\right)^{2} \delta' = 2.109^{2} \times 0.10 \approx 0.445$, applying equation (22) it holds that $n = 2{,}146$. Consequently, it is necessary to add 681 new individuals to the initial sample of 1,465 individuals, and once the data are obtained it is necessary to check that the required precision has been achieved.

Table 7. Diagnosis of coronary disease.

	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$D = 1$	224	591	32	176	1,023
$D = 0$	35	80	41	286	442
Total	259	671	73	462	1,465
Results					
	Se	*Sp*	LR^{+}	LR^{-}	
Exercise test	0.797 ± 0.013	0.740 ± 0.021	3.065 ± 0.250	0.274 ± 0.019	
Resting EKG	0.250 ± 0.014	0.828 ± 0.018	1.453 ± 0.171	0.906 ± 0.026	
p	ε_1	ε_0	$\omega^{+} = LR_1^{+} / LR_2^{+}$	$\omega^{-} = LR_1^{-} / LR_2^{-}$	
0.698	0.020	0.034	2.109 ± 0.273	0.302 ± 0.021	
*CI*s for $\omega^{+} = LR_1^{+} / LR_2^{+}$					
Regression *CI*		Logarithmic *CI*		Wald *CI*	
$(1.589 , 2.786)$		$(1.632 , 2.713)$		$(1.569 , 2.639)$	
Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
$(1.647 , 2.765)$		$(1.501 , 2.612)$		$(1.668 , 2.567)$	
*CI*s for $\omega^{-} = LR_1^{-} / LR_2^{-}$					
Regression *CI*		Logarithmic *CI*		Wald *CI*	
$(0.263 , 0.351)$		$(0.265 , 0.348)$		$(0.262 , 0.345)$	
Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
$(0.262 , 0.346)$		$(0.280 , 0.348)$		$(0.264 , 0.343)$	

6.2. Diagnosis of colorectal cancer

The results obtained were applied to a study of the diagnosis of colorectal cancer, using as diagnostic tests *Fecal Occult Blood Testing* (*FOBT*) and *Fecal Immunochemical Testing* (*FIT*), and the biopsy as the *GS*. Table 8 shows the results obtained by applying the three tests to a sample of 168 adult men with suspicious symptoms of the disease, where the variable T_1 models the result of the *FOBT*, T_2 models the result of the *FIT* and *D* models the result of the biopsy. This data came from a study carried out at the University Hospital of Granada in Spain. Table 8 also shows the estimations of the *LR*s, their standard errors and the confidence intervals for ω^{+} and ω^{-}.

Applying the rule given in Section 4.3, as $n = 168 < 200$ the logarithmic *CI* for ω^+ must be used in addition to the Wald *CI* for ω^-. For ω^+, the logarithmic *CI* contains the value one, and therefore we do not reject the equality of both positive *LRs*. Regarding ω^-, the Wald *CI* does not contain the value one, and therefore we reject the equality of both negative *LRs*. Thus, a negative result for the *FOBT* is more indicative of the presence of colorectal cancer than a negative result for the *FIT*. The negative *LR* of the *FOBT* is (with a confidence of 95%) a value between 1.321 and 3.183 times larger than the negative *LR* of the *FIT*. The Wald *CI* for $1/\omega^-$ is $(0.260\ ,\ 0.628)$, calculated as $\left(1.321/2.252^2\ ,\ 3.183/2.252^2\right)$.

Table 8. Diagnosis of colorectal cancer.

	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$D = 1$	68	1	18	13	100
$D = 0$	4	2	1	61	68
Total	72	3	19	74	168
Results					
	Se	*Sp*	LR^+	LR^-	
FOBT	0.690 ± 0.046	0.912 ± 0.034	7.841 ± 3.093	0.340 ± 0.052	
FIT	0.860 ± 0.035	0.926 ± 0.032	11.622 ± 5.057	0.151 ± 0.038	
p	ε_1	ε_0	$\omega^+ = LR_1^+ / LR_2^+$	$\omega^- = LR_1^- / LR_2^-$	
0.595	0.087	0.052	0.675 ± 0.215	2.252 ± 0.475	
*CI*s for $\omega^+ = LR_1^+ / LR_2^+$					
Regression *CI*		Logarithmic *CI*		Wald *CI*	
$(0.212\ ,\ 2.108)$		$(0.356\ ,\ 1.255)$		$(0.254\ ,\ 1.096)$	
Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
$(0.278\ ,\ 2.277)$		$(0.281\ ,\ 1.283)$		$(0.222\ ,\ 2.057)$	
*CI*s for $\omega^- = LR_1^- / LR_2^-$					
Regression *CI*		Logarithmic *CI*		Wald *CI*	
$(1.265\ ,\ 4.001)$		$(1.488\ ,\ 3.403)$		$(1.321\ ,\ 3.183)$	
Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
$(1.556\ ,\ 3.894)$		$(1.553\ ,\ 3.778)$		$(1.281\ ,\ 4.006)$	

In order to illustrate in this example the method of sample size calculation, let us suppose that the researchers want to estimate $1/\omega^-$ with a precision equal to 0.10, or in other words, to estimate ω^- with a precision of $0.10 \times \left(\hat{\omega}^-\right)^2 = 0.10 \times 2.252^2 \approx 0.50$. As with the sample of 168 individuals the precision obtained with the Wald *CI* for ω^- is

$0.931 > 0.50$, or rather a precision equal to 0.184 (> 0.10) with the Wald *CI* for $1/\omega^{-}$, then it is necessary to calculate the sample size. Considering the sample of 168 individuals to be a pilot sample, applying equation (23) it holds that $n = 561$. Therefore, 561 individuals are needed (we have to add 393 to the sample of 198) in order to estimate ω^{-} $\left(1/\omega^{-}\right)$ with a precision equal to 0.50 (0.10) with a confidence of 95%.

7. Discussion

The *LRs* are parameters that are used to assess and compare the effectiveness of *BDTs*, and only depend on the accuracy (sensitivity and specificity) of the *BDT*. The comparison of the positive (negative) *LRs* of two *BDTs* subject to a paired design is a topic which has not been widely studied in Statistical literature and consists of the comparison of two relative risks subject to the same type of design. The previous studies (Leisenring and Pepe (1998) and Pepe (2003), Roldán-Nofuentes and Luna (2007), Dolgun et al (2012) focused mainly on the study of hypothesis tests to compare the positive (negative) *LRs* of the two *BDTs*. The comparison of the positive (negative) *LRs* through *CIs* has been the object of the very little research, and the studies that have been published by Pepe (2003) and Roldán-Nofuentes and Luna (2007) have focused on proposing *CIs* without dealing with this question in more depth. In this article, we extend the scope of these previous studies, proposing four new intervals: three of which are frequentist (Wald, Fieller and Bootstrap) and one which is Bayesian. The Wald and Fieller intervals are based on the asymptotic normality of the ratio of the *LRs*, and the Bootstrap interval is based on the fact that the bootstrap estimator of the ratio of the *LRs* can be transformed to a normal distribution. Regarding the Bayesian Interval, this was obtained by applying the Monte Carlo method considering a priori non-informative distributions. The importance of the study of the *CIs* for the ratio of the positive (negative) *LRs* does not only lie in the fact that these *CIs* allow us to compare the two positive (negative) *LRs*, but also that it allows us to determine (when the equality of both *LRs* is rejected) how much bigger one *LR* than the other, which means an advantage over the hypothesis tests.

The comparison of the asymptotic behaviour of the six *CIs* was studied through simulation experiments. The results of these experiments has shown that, in the

scenarios considered, in order to estimate the ratio $\omega^+ = LR_1^+ / LR_2^+$, in general terms, the intervals with the best behaviour are the logarithmic one (for all the sample sizes), the Wald, Fieller or Bootstrap intervals (these last three for large or very large samples); whereas in order to estimate $\omega^- = LR_1^- / LR_2^-$ the interval with the best behaviour is the Wald interval (for all of the samples sizes). The use of different *CIs* for ω^+ and for ω^- may be due to the convergence to the normal distribution of the estimators. For an informative *BDT*, i.e. for a *BDT* whose Youden index is higher than 0 $\left(Y = Se + Sp - 1 > 0\right)$, it must be verified that $LR^+ > 1$ and that $LR^- < 1$. Then, considering that the two *BDTs* are informative (as should be the case in clinical practice), ω^+ is the ratio between two values greater than 1 and ω^- is the ratio between two values lower than 1. For ω^+, $\ln \hat{\omega}^+$ converges better to the normal distribution than $\hat{\omega}^+$ for $n < 200$, but when $n \geq 200$ both ($\hat{\omega}^+$ and $\ln \hat{\omega}^+$) has a good approximation to the normal distribution. The Wald *CI* for ω^- has a better asymptotic behaviour than the logarithmic *CI* for ω^-, which must be due to the fact that $\hat{\omega}^-$ converges more quickly to the normal distribution (even with large samples) than $\ln \hat{\omega}^-$.

An important question when comparing parameters of two *BDTs* is the calculation of the sample size necessary to compare the parameters based on certain specifications. When a hypothesis test is carried out, the sample size is calculated based on an α error, a θ power and a difference (or ratio) to be detected among the parameters. Roldán-Nofuentes and Luna (2007) proposed a method to calculate sample size to solve the hypothesis test ($H_0 : \ln \omega = 0$) of equality of the positive (negative) *LRs*. This article proposes, as a complement to the study of the *CIs*, a method to determine the sample size necessary to estimate the ratio between the *LRs* with a previously set precision. This is a topic that has never been studied and, therefore, represents a contribution to Statistical literature on the subject analysed in this article. The method, which is based on the Wald (logarithmic) *CI*, requires knowledge of the estimations of the sensitivities, specificities, dependence factors and disease prevalence. These estimations can be obtained from a pilot sample or another similar study and, therefore, as it depends on the pilot sample selected. Therefore, the method does not guarantee that with the calculated sample size the parameter ω can be estimated with the set precision, and it is necessary to check this precision.

The intervals studied in this article can also be applied when the sample design is case-control. In this type of design, the two *BDTs* are applied to all of the individuals in two random samples, one of n_1 individuals with the disease and another one of n_2 individuals without the disease. If thus type of sampling is used, two multinomial distributions are involved, one for the case sample, whose probabilities are $p_{ij} = Se_1^i\left(1-Se_1\right)^{1-i} Se_2^j\left(1-Se_2\right)^{1-j} + \delta_{ij}\varepsilon_1$ with $\sum p_{ij} = 1$, and the other for the control sample, whose probabilities are $q_{ij} = Sp_1^{1-i}\left(1-Sp_1\right)^i Sp_2^{1-j}\left(1-Sp_2\right)^j + \delta_{ij}\varepsilon_0$ with $\sum q_{ij} = 1$. Here, the variances-covariances of the sensitivities and specificities are

$$Var\left(\hat{Se}_h\right) = \frac{Se_h\left(1-Se_h\right)}{n_1},\ Var\left(\hat{Sp}_h\right) = \frac{Sp_h\left(1-Sp_h\right)}{n_2},$$

$$Cov\left(\hat{Se}_1, \hat{Se}_2\right) = \frac{\varepsilon_1}{n_1},\ Cov\left(\hat{Sp}_1, \hat{Sp}_2\right) = \frac{\varepsilon_0}{n_2}.$$

The equations of the estimators and of the variances-covariances given in the regression, logarithmic, Wald and Fieller intervals are valid substituting *s* with n_1 and *r* with n_2. Regarding the Bootstrap interval, it is necessary to generate *B* samples with replacement from the case sample and another *B* samples with replacement from the control sample, and the process is the same as the one described in Section 3.5. Regarding the Bayesian interval, the process is similar substituting *s* with n_1 and *r* with n_2.

The methodology used in this article, both to obtain the *CIs* and to calculate the sample size, can be used to compare other parameters of *BDTs*, e.g. the odds ratios. The odds ratio of a *BDT* is defined as $OR = SeSp/\left[\left(1-Se\right)\left(1-Sp\right)\right]$ and is a measure of the association between the *BDT* and the *GS*. It is easy to check that the ratio of the odds ratios of two *BDTs* is $LR_1^+ LR_2^- / \left(LR_1^- LR_2^+\right)$, and therefore from this expression it is possible to deduce *CIs* similar to those given in Section 3 and can also be applied to the same procedure as in Section 5 to determine the sample size necessary to compare the odds ratios of two *BDTs* through a *CI*.

In this manuscript we studied the comparison of the *LRs* of two binary diagnostic tests. When the diagnostic test is quantitative, its accuracy is measured by the area under the *ROC* curve. The *LRs* are related to the equation of the *ROC* curve. Thus, for a single

quantitative diagnostic test, for each one of the cut off points c of the estimated *ROC* curve a value for $\hat{Se}$ and a value $1-\hat{Sp}$ are obtained, and therefore a value for $\hat{LR}^{+}$ (and another one for $\hat{LR}^{-}$). For $\hat{LR}^{+}$, its numerator $\hat{Se}$ is the "*y*" coordinate of the estimated *ROC* curve, and the denominator $1-\hat{Sp}$ is the "*x*" coordinate of the estimated *ROC* curve. The estimator of *LR* for an interval (c_1, c_2) of test values corresponds to the slope of the line segment between c_1 and c_2 on the estimated *ROC* curve. In the case of two quantitative diagnostic test, for each cut off point of each estimated *ROC* curve, we obtain a value for $\hat{\omega}^{+}$ and another one for $\hat{\omega}^{-}$, and therefore it is possible to calculate the *CIs* studied in Section 3.

Appendix *A*

The variances-covariances of all of the parameters were obtained applying the delta method. Let $\boldsymbol{\theta} = (Se_1, Sp_1, Se_2, Sp_2)^T$ be a vector whose components are the sensitivities and the specificities, let $\mathbf{LR} = \left(LR_1^{+}, LR_2^{+}, LR_1^{-}, LR_2^{-}\right)^T$. be a vector whose components are the positive *LRs* and the negative *LRs*, and $\boldsymbol{\omega} = \left(\omega^{+}, \omega^{-}\right)^T$. The matrix of variances-covariances of $\hat{\boldsymbol{\theta}}$ is

$$\Sigma_{\hat{\theta}} = \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}}\right) \Sigma_{\hat{\psi}} \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}}\right)^T .$$

Regarding the *LRs*, the matrix of variances-covariances of $\hat{\mathbf{LR}}$ is

$$\Sigma_{\hat{\mathbf{LR}}} = \left(\frac{\partial \mathbf{LR}}{\partial \boldsymbol{\theta}}\right) \Sigma_{\hat{\theta}} \left(\frac{\partial \mathbf{LR}}{\partial \boldsymbol{\theta}}\right)^T .$$

Finally, the matrix of variances-covariances of $\hat{\boldsymbol{\omega}}$ is

$$\Sigma_{\hat{\omega}} = \left(\frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\theta}}\right) \Sigma_{\hat{\theta}} \left(\frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\theta}}\right)^T . \tag{25}$$

The matrix of variances-covariances of $\ln(\hat{\boldsymbol{\omega}})$ is calculate in a similar way, i.e.

$$\Sigma_{\ln(\hat{\omega})} = \left(\frac{\partial \ln(\boldsymbol{\omega})}{\partial \boldsymbol{\theta}}\right) \Sigma_{\hat{\theta}} \left(\frac{\partial \ln(\boldsymbol{\omega})}{\partial \boldsymbol{\theta}}\right)^T .$$

Performing the algebraic operations in each one of the previous expressions and substituting each parameter with its estimator, we obtain the asymptotic variances-covariances given in the equations (5), (6), (7) and (10) respectively.

Appendix *B*

The selection of the *CI* with the best asymptotic behaviour was made through the following steps: 1) Choose the *CIs* with the least failures ($CP > 93\%$), 2) Choose the *CIs* which are the most precise (lowest *AL*) and among these those which have a *CP* closest to 95%. The first step in this method establishes that the *CI* does not fail when $CP > 93\%$. The confidence level was set at 95%, i.e. $\gamma = 1 - \alpha = 0.95$ was set as the nominal confidence and, therefore, a nominal error $\alpha = 5\%$. Let γ^* be the calculated *CP*, then $\Delta\alpha = \gamma^* - \gamma = \alpha - \alpha^*$, where α^* is the type I error.

Furthermore, the hypothesis test to check the equality of the two LRs is $H_0 : LR_1 = LR_2$ vs $H_1 : LR_1 \neq LR_2$, which is equivalent to checking $H_0 : \omega = 1$ vs $H_0 : \omega \neq 1$. In Step 1, a *CI* fails if $CP \leq 93\%$, i.e. if $\Delta\alpha \leq -2$. In this situation, the type I error of the hypothesis test is $\geq 7\%$, and therefore it is a very liberal hypothesis test and can give false significances. If $\Delta\alpha > 2\%$, i.e. $CP > 97\%$, then the hypothesis test is very conservative (its type I error is very small, $< 3\%$), but does not give false significances. Therefore, the choice of the *CI* is linked to the decisions of the hypothesis test, and it is preferable to choose a conservative test rather than a very liberal one (as there will be no false significances due to the fact that its type I error is lower than the nominal one).

Acknowledgements

This research was supported by the Spanish Ministry of Economy, Grant Number MTM2016-76938-P.

References

Biggerstaff, B. J. (2000). Comparing diagnostic tests: a simple graphic using likelihood ratios. *Statistics in Medicine* **19**, 649-663.

Boos, D. D., Stefanski, L. A. (2013). *Essential Statistical Inference. Theory and Method*. New York: Springer.

Dolgun, N. A, Gozukara, H., Karaagaoglu, E. (2012). Comparing diagnostic tests: test of hypothesis for likelihood ratios. *Journal of Statistical Computation and Simulation* **82**, 369-381.

Efron B., Tibshirani R J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Fieller, E. C. (1940). The biological standardization of insulin. *Journal of the Royal Statistical Society* **7**, 1-64.

Leisenring, W., Pepe, M. S. (1998). Regression modelling of diagnostic likelihood ratios for the evaluation of medical diagnostic tests. *Biometrics* **54**, 444-442.

Martín-Andrés, A., Álvarez-Hernández, M. (2014a). Two-tailed asymptotic inferences for a proportion. *Journal of Applied Statistics* **41**, 1516-1529.

Martín-Andrés, A., Álvarez-Hernández, M. (2014b). Two-tailed approximate confidence intervals for the ratio of proportions. Statistics and Computing **24**, 65-75.

Montero-Alonso, M. A., Roldán-Nofuentes, J. A. (2018). Approximate confidence intervals for the likelihood ratios of a binary diagnostic test in the presence of partial disease verification. Journal of Biopharmaceutical Statistics, in press, DOI: 10.1080/10543406.2018.1452025.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* New York: Oxford University Press.

Price, R. M., Bonett, D. G. (2004). An improved confidence interval for a linear function of binomial proportions. *Computational Statistics and Data Analysis* **45**, 449-456.

Shao, J. Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer.

Roldán-Nofuentes, J. A., Luna del Castillo, J. D. (2007). Comparison of the likelihood ratios of two binary diagnostic tests in paired designs. *Statistics in Medicine* **26**, 4179-4201.

Vacek, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* **41**, 959-968.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* **5**, 426-482.

Weiner, D. A., Ryan, T. J., McCabe, C. H., Kennedy, J. W., Schloss, M., Tristani, F., Chaitman, B. R., Fisher, L. D. (1979). Correlations among history of angina, ST-segment and prevalence of coronary artery disease in the coronary artery surgery study (CASS). *New England Journal of Medicine* **301**, 230-235.

Zhou, X. H., Obuchowski, N.A., McClish, D. K. (2011). *Statistical Methods in Diagnostic Medicine, Second Edition*. New York: Wiley.

APPENDIX III

Asymptotic confidence intervals for the difference and the ratio of the weighted kappa coefficients of two diagnostic tests subject to a paired design

Roldán-Nofuentes, J.A., Sidaty-Regad S.B. (2020). Asymptotic confidence intervals for the difference and the ratio of the weighted kappa coefficients of two diagnostic tests subject to a paired design. REVSTAT-Statistical Journal. Accepted, in press.

Category: Statistics and Probability. JCR 2019 (last published): 0.667. Rank: 101/124. Quartile: Q4.

Abstract

The weighted kappa coefficient of a binary diagnostic test is a measure of the beyond-chance agreement between the diagnostic test and the gold standard, and depends on the sensitivity and specificity of the diagnostic test, on the disease prevalence and on the relative importance between the false positives and the false negatives. This article studies the comparison of the weighted kappa coefficients of two binary diagnostic tests subject to a paired design through confidence intervals. Three asymptotic confidence intervals are studied for the difference between the parameters and five other intervals for the ratio. Simulation experiments were carried out to study the coverage probabilities and the average lengths of the intervals, giving some general rules for application. A method is also proposed to calculate the sample size necessary to compare the two weighted kappa coefficients through a confidence interval. A program in R has been written to solve the problem studied and it is available as supplementary material. The results were applied to a real example of the diagnosis of malaria.

Keywords: weighted kappa coefficient, paired design, binary diagnostic test.

Mathematics Subject Classification: 62P10, 6207.

1. Introduction

A diagnostic test is medical test that is applied to an individual in order to determine the presence or absence of a disease. When the result of a diagnostic test is positive (indicating the presence of the disease) or negative (indicating its absence), the diagnostic test is called a binary diagnostic test (*BDT*) and its accuracy is measured in terms of two fundamental parameters: sensitivity and specificity. Sensitivity (*Se*) is the probability of the *BDT* result being positive when the individual has the disease, and specificity (*Sp*) is the probability of the *BDT* result being negative when the individual does not have the disease. Sensitivity is also called true positive fraction (*TPF*) and specificity is also called true negative fraction (*TNF*), verifying that $TPF = 1 - FNF$ and that $TNF = 1 - FPF$, where *FNF* (*FPF*) is the false negative (positive) fraction. The accuracy of a *BDT* is assessed in relation to a gold standard (*GS*), which is a medical test that objectively determines whether or not an individual has the disease. When considering the losses of an erroneous classification with the *BDT*, the performance of the *BDT* is measured in terms of the weighted kappa coefficient (Kraemer et al, 1990; Kraemer, 1992; Kraemer et al, 2002). The weighted kappa coefficient depends on the *Se* and *Sp* of the *BDT*, on the disease prevalence (*p*) and on the relative importance between the false positives and the false negatives (weighting index *c*). The weighted kappa coefficient is a measure of the beyond-chance agreement between the *BDT* and the *GS*.

Furthermore, the comparison of the performance of two *BDTs* is an important topic in the study of Statistical Methods for Diagnosis in Medicine. The comparison of two *BDTs* can be made subject to two types of sample designs: unpaired design and paired design. In the book by Pepe (2003) we can see a broad discussion about both types of sample designs. Summing up, subject to an unpaired design each individual is tested with a single *BDT*, whereas subject to a paired design each individual is tested with the two *BDT*s. Consequently, unpaired design consists of applying a *BDT* to a sample of n_1 individuals and the other *BDT* to another sample of n_2 individuals; paired design consists of applying both *BDTs* to all of the individuals of a sample sized *n*. The comparative studies based on a paired design are more efficient from a statistical point of view than the studies based on an unpaired design, since it minimizes the impact of the between-individual variability. Therefore, in this article we focus on paired design. Subject to this type of design, Bloch (1997) has studied an asymptotic hypothesis test to

compare the weighted kappa coefficients of two *BDTs*. Nevertheless, if the hypothesis test is significant, this method does not allow us to assess how much bigger one weighted kappa coefficient is compared to another one, and it is necessary to estimate this effect through confidence intervals (*CIs*). Thus, the objective of our study is to compare the weighted kappa coefficients of two *BDTs* through *CIs*. Frequentist and Bayesian *CIs* have been studied for the difference and for the ratio of the two weighted kappa coefficients. If a *CI* for the difference (ratio) does not contain the zero (one) value, then we reject the equality between the two weighted kappa coefficients and we estimate how much bigger one coefficient is than another one. Consequently, our study is an extension of the Bloch method to the situation of the *CIs*. We have also dealt with the problem of calculating the sample size to compare the two parameters through a *CI*.

The manuscript is structured in the following way. In Section 2, we explain the weighted kappa coefficient of a *BDT* and we relate the comparison of the weighted kappa coefficients of two *BDTs* with the relative true (false) positive fraction of the two *BDTs*. Section 3 summarizes the Bloch method and we propose *CIs* for the difference and the ratio of the weighted kappa coefficients of two *BDTs* subject to a paired design. In Section 4, simulation experiments are carried out to study the asymptotic behaviour of the proposed *CIs*, and some general rules of application are given. In Section 5, we propose a method to calculate the sample size necessary to compare the two weighted kappa coefficients through a *CI*. In Section 6, a programme written in *R* is presented to solve the problems posed in this manuscript. In Section 7, the results were applied to a real example on the diagnosis of malaria, and in Section 8 the results are discussed.

2. Weighted kappa coefficient

Let us consider a *BDT* that is assessed in relation to a *GS*. Let L $\left(L'\right)$ the loss which occurs when for a diseased (non-diseased) individual the *BDT* gives a negative (positive) result. Therefore, the loss L $\left(L'\right)$ is associated with a false negative (positive). If an individual (with or without the disease) is correctly diagnosed by the *BDT* then $L = L' = 0$. Let D be the variable that models the result of the *GS*: $D = 1$ when an individual has the disease and $D = 0$ when this is not the case. Let $p = P\left(D = 1\right)$ be the prevalence of the disease and $q = 1 - p$. Let T be the random variable that models the

result of the *BDT*: $T=1$ when the result of the *BDT* is positive and $T=0$ when the result is negative. Table 1 shows the losses and the probabilities associated with the assessment of a *BDT* in relation to a *GS*, and the probabilities when the *BDT* and the *GS* are independent, i.e. when $P\left(T=i\middle|D=j\right)=P\left(T=i\right)$. Multiplying each loss in the 2×2 table by its corresponding probability and adding up all the terms, we find $p\left(1-Se\right)L+q\left(1-Sp\right)L'$, a term that is defined as expected loss. Therefore, the expected loss is the loss that occurs when erroneously classifying with the *BDT* an individual with or without the disease. Moreover, if the *BDT* and the *GS* are independent, multiplying each loss by its corresponding probability (subject to the independence between the *BDT* and the GS) and adding up all of the terms we find $p\left[\,p\times\left(1-Se\right)+q\times Sp\,\right]L+q\left[\,p\times Se+q\times\left(1-Sp\right)\right]L'$, a term that is defined as random loss. Therefore, the random loss is the loss that occurs when the *BDT* and the *GS* are independent. The independence between the *BDT* and the *GS* is equivalent to the Youden index of the *BDT* being equal to zero i.e. $Se+Sp-1=0$, and is also equivalent to the expected loss being equal to the random loss. In terms of expected and random losses, the weighted kappa coefficient of a *BDT* is defined as

$$\kappa=\frac{\text{Random loss}-\text{Expected loss}}{\text{Random loss}}.$$

Substituting in this equation each loss with its expression, the weighted kappa coefficient of a *BDT* is expressed (Kraemer et al, 1990; Kraemer, 1992; Kraemer et al, 2002) as

$$\kappa(c)=\frac{pqY}{p\left(1-Q\right)c+qQ\left(1-c\right)},\qquad(1)$$

where $Y=Se+Sp-1$ is the Youden index, $Q=pSe+q\left(1-Sp\right)$ is the probability that the *BDT* result is positive, and $c=L/\left(L'+L\right)$ is the weighting index. The weighting index *c* is a measure of the relative importance between the false positives and the false negatives. For example, let us consider the diagnosis of breast cancer using as a diagnostic mammography test. If the mammography test is positive in a woman that does not have cancer (false positive), the woman will be given a biopsy that will give a negative result. The loss L' is determined from the economic costs of the diagnosis and also from the risk, stress, anxiety, etc., caused to the woman. If the mammography test

is negative in a woman who has breast cancer (false negative), the woman may be diagnosed at a later stage, but the cancer may spread, and the possibility of the treatment being successful will have diminished. The loss L is determined from these considerations. The losses L and L' are measured in terms of economic costs and also from risks, stress, etc., which is why in practice their values cannot be determined. Therefore, as loss L cannot be determined, L is substituted by the importance that a false positive has for the clinician; in the same way, as loss L' cannot be determined, then L' is substituted by the importance that a false negative has for the clinician. The value of the weighting index c will depend therefore on the relative importance between a false positive and a false negative. If the clinician has greater concerns about false positives, as it is the situation in which the *BDT* is used as a definitive test prior to a treatment that involves a risk for the individual (e.g., a definitive test prior to a surgical operation), then $0 \le c < 0.5$. If the clinician is more concerned about false negatives, as in a screening test, then $0.5 < c \le 1$. The index c is equal to 0.5 when the clinician considers that the false negatives and the false positives have the same importance, in which case $\kappa(0.5)$ is the Cohen kappa coefficient. Weighting index c quantifies the relative importance between a false positive and a false negative, but it is not a measure that quantifies how much bigger the proportion of false positives is compared to the false negatives. If $c = 0$ then

$$\kappa(0) = \frac{Sp - (1-Q)}{Q} = \frac{p(1 - FNF - FPF)}{p(1 - FNF) + qFPF}, \tag{2}$$

which is the chance corrected specificity according to the kappa model. If $c = 1$ then

$$\kappa(1) = \frac{Se - Q}{1 - Q} = \frac{q(1 - FNF - FPF)}{pFNF - q(1 - FPF)}, \tag{3}$$

which is the chance corrected sensitivity according to the kappa model. A low (high) value of $\kappa(1)$ will indicate that the value of *FNF* is high (low), and a low (high) value of $\kappa(0)$ will indicate that the value of *FPF* is high (low). The weighted kappa coefficient can be written as

$$\kappa(c) = \frac{pc(1-Q)\kappa(1) + q(1-c)Q\kappa(0)}{pc(1-Q) + q(1-c)Q}, \tag{4}$$

which is a weighted average of $\kappa(0)$ and $\kappa(1)$. Therefore, the weighted kappa coefficient is a measure that considers the proportion of false negatives (*FNF*) and the proportion of false positives (*FPF*). Moreover, for a set value of the *c* index and of the accuracy (*Se* and *Sp*) of the *BDT*, the weighted kappa coefficient strongly depends on the disease prevalence among the population being studied, and its value increases when the disease prevalence increases. The weighted kappa coefficient is a measure of the beyond-chance agreement between the *BDT* and the *GS*. The properties of the kappa coefficient can be seen in the manuscript of Roldán-Nofuentes and Amro (2018).

Table 1. Losses and probabilities.

Losses (Probabilities)			
	$T=1$	$T=0$	Total
$D=1$	$0\ (p\times Se)$	$L\ (p\times(1-Se))$	$L\ (p)$
$D=0$	$L'\ (q\times(1-Sp))$	$0\ (q\times Sp)$	$L'\ (q)$
Total	L' $(Q=p\times Se+q\times(1-Sp))$	L $(1-Q=p\times(1-Se)+q\times Sp)$	$L+L'\ (1)$
Probabilities when the *BDT* and the *GS* are independent			
	$T=1$	$T=0$	Total
$D=1$	$p\times Q$	$p\times(1-Q)$	p
$D=0$	$q\times Q$	$q\times(1-Q)$	q
Total	Q	$1-Q$	1

The weighted kappa coefficient is a valid parameter to assess and compare the performance of *BDTs* (Kraemer et al, 1990; Kraemer, 1992; Kraemer et al, 2002; Bloch, 1997; Roldán-Nofuentes et al, 2009; Roldán-Nofuentes and Amro, 2018).

When comparing the accuracies of two *BDTs*, Pepe (2003) recommends using the parameters $rTPF_{12}=\frac{Se_1}{Se_2}$ and $rFPF_{12}=\frac{FPF_1}{FPF_2}$, where $FPF_h=1-Sp_h$, with $h=1,2$. If $rTPF_{12}>1$ then the sensitivity of Test 1 is greater than that of Test 2, and if $rFPF_{12}>1$ then the *FPF* of Test 1 is greater than that of Test 2 (the specificity of Test 2 is greater than that of Test 1). The comparison of the weighted kappa coefficients of two *BDTs* can be related to the previous measures, and these have an important effect on the comparison of $\kappa_1(c)$ and $\kappa_2(c)$. From now onwards, it is considered that $0<Se_h<1$, $0<Sp_h<1$ and $0<p<1$, with $h=1,2$. Let us consider the subindexes *i* and *j*, in such a

way that if $i=1$ $(i=2)$ then $j=2$ $(j=1)$. It is obvious that if $rTPF_{ij}=rFPF_{ij}=1$ then $Se_1=Se_2$ and $Sp_1=Sp_2$, and that therefore $\kappa_1(c)=\kappa_2(c)$ with $0\le c\le 1$. Let

$$c'=\frac{(1-p)\left[Se_2(1-Sp_1)-Se_1(1-Sp_2)\right]}{p(Se_1-Se_2)+(1-Sp_1)(Se_2-p)-(1-Sp_2)(Se_1-p)}. \quad (5)$$

In terms of $rTPF_{ij}$ and $rFPF_{12}$ the following rules are verified to compare $\kappa_1(c)$ and $\kappa_2(c)$:

a) If $rTPF_{ij}\ge 1$ and $rFPF_{ij}<1$, or $rTPF_{ij}>1$ and $rFPF_{ij}\le 1$, then $\kappa_i(c)>\kappa_j(c)$ for $0\le c\le 1$.

b). If $rTPF_{ij}>1$ and $rFPF_{ij}>1$, then:

b.1) $\kappa_i(c)>\kappa_j(c)$ if $0<c'<c\le 1$

b.2) $\kappa_i(c)<\kappa_j(c)$ if $0\le c<c'<1$

b.3) $\kappa_1(c)=\kappa_2(c)$ if $c=c'$, with $0<c'<1$

b.4) $\kappa_i(c)>\kappa_j(c)$ for $0\le c\le 1$ if $c'<0$ (or $c'>1$) and $rTPF_{ij}>rFPF_{ij}>1$

b.5) $\kappa_i(c)<\kappa_j(c)$ for $0\le c\le 1$ if $c'<0$ (or $c'>1$) and $rFPF_{ij}>rTPF_{ij}>1$

c) If $rTPF_{ij}<1$ and $rFPF_{ij}<1$, then:

c.1) $\kappa_i(c)>\kappa_j(c)$ if $0\le c<c'<1$

c.2) $\kappa_i(c)<\kappa_j(c)$ if $0<c'<c\le 1$

c.3) $\kappa_2(c)=\kappa_1(c)$ if $c=c'$, with $0<c'<1$

c.4) $\kappa_i(c)>\kappa_j(c)$ for $0\le c\le 1$ if $c'<0$ (or $c'>1$) and $rTPF_{ij}>rFPF_{ij}>1$

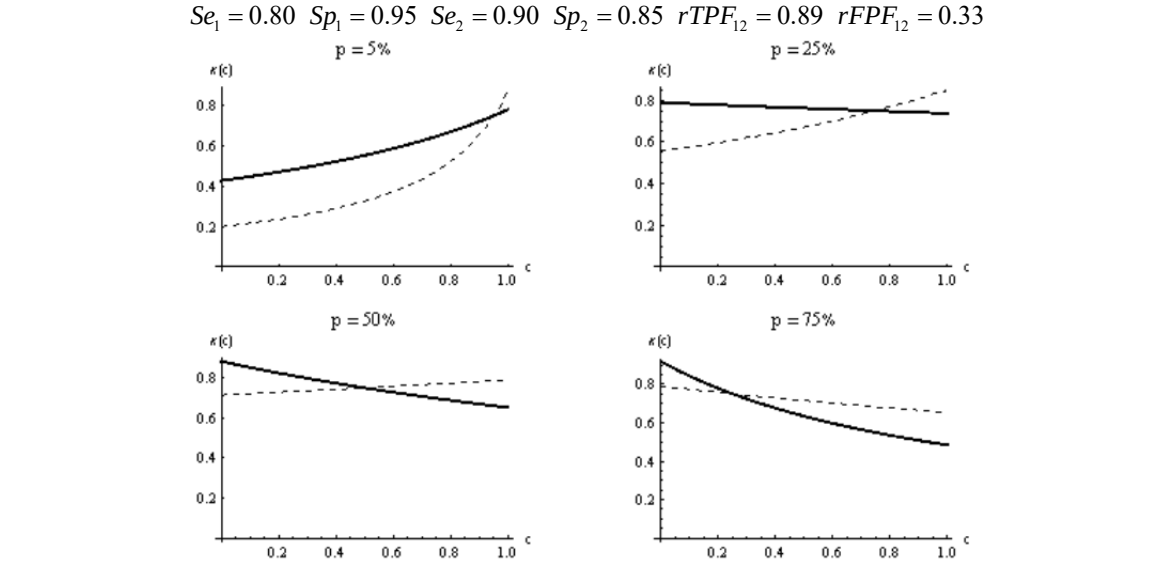
c.5) $\kappa_i(c)<\kappa_j(c)$ for $0\le c\le 1$ if $c'<0$ (or $c'>1$) and $rFPF_{ij}>rTPF_{ij}>1$

The demonstrations can be seen in the Appendix *A* of the supplementary material. Regarding c', this is obtained solving the equation $\kappa_1(c)-\kappa_2(c)=0$ in *c*. The graphs in Figure 1 show how $\kappa_1(c)$ (on a continuous line) and $\kappa_2(c)$ (on a dotted line) vary

depending on the weighting index *c*, taking as prevalence $p = \{5\%, 25\%, 50\%, 75\%\}$, for $Se_1 = 0.80$, $Sp_1 = 0.95$, $Se_2 = 0.90$ and $Sp_2 = 0.85$. These graphs correspond to the case in which $rTPF_{12} < 1$ and $rFPF_{12} < 1$, and therefore $\kappa_1(c) > \kappa_2(c)$ when $c < c'$, and $\kappa_2(c) > \kappa_1(c)$ when $c > c'$, and c' is equal to 0.95 when $p = 5\%$, 0.75 when $p = 25\%$, 0.50 when $p = 50\%$ and 0.25 when $p = 75\%$. If the clinician considers that a false positive is 1.5 times more important than a false negative, then $c = 0.4$ and $\kappa_1(c) > \kappa_2(c)$ in the population with $p = \{5\%, 25\%, 50\%\}$ and $\kappa_2(c) > \kappa_1(c)$ in the population with $p = 75\%$. If in the population with $p = 75\%$ the clinician has a greater concern about a false positive than a false negative $(0 \le c < 0.5)$, then $\kappa_1(c) > \kappa_2(c)$ if $0 \le c < 0.25$ and $\kappa_2(c) > \kappa_1(c)$ if $0.25 < c < 0.5$; in the populations with $p = \{5\%, 25\%, 50\%\}$, $\kappa_1(c) > \kappa_2(c)$ when $0 \le c < 0.5$.

We will now study the comparison of the weighted kappa coefficients of two *BDTs* through *CIs* subject to a paired design.

Figure 1. Weighted kappa coefficients with $rTPF_{12} < 1$ and $rFPF_{12} < 1$.



3. Confidence intervals

Let us consider two *BDTs* which are assessed in relation to the same *GS*. Let T_1 and T_2 be the random binary variables that model the results of each *BDT* respectively. Let Se_h and Sp_h be the sensitivity and specificity of the *h*th *BDT*, with $h=1,2$. Table 2 (Observed frequencies) shows the frequencies that are obtained when both *BDTs* and the *GS* are applied to all the individuals in a random sample sized *n*. The frequencies s_{ij} and r_{ij} are the product of a multinomial distribution whose probabilities are also shown in Table 1 (Theoretical probabilities), where $p_{ij}=P\left(D=1,T_1=i,T_2=j\right)$ and $q_{ij}=P\left(D=0,T_1=i,T_2=j\right)$, with $i,j=0,1$. The probability of the two *BDTs* being positive when an individual has the disease is $Se_1Se_2+\varepsilon_1$, where ε_1 is the covariance or dependence factor between the two *BDTs* when $D=1$; and the probability of the two *BDTs* being negative when an individual does not have the disease is $\left(1-Sp_1\right)\left(1-Sp_2\right)+\varepsilon_0$, where ε_0 is the covariance or dependence factor between the two *BDTs* when $D=0$. This model is known as the Vacek (1985) conditional dependence model. Applying this model, the probabilities p_{ij} and q_{ij} are written as

$$p_{ij}=p\left[Se_1^i\left(1-Se_1\right)^{1-i}Se_2^j\left(1-Se_2\right)^{1-j}+\delta_{ij}\varepsilon_1\right] \tag{6}$$

and

$$q_{ij}=q\left[Sp_1^{1-i}\left(1-Sp_1\right)^i Sp_2^{1-j}\left(1-Sp_2\right)^j+\delta_{ij}\varepsilon_0\right], \tag{7}$$

where $\delta_{ij}=1$ if $i=j$ and $\delta_{ij}=-1$ if $i\neq j$, with $i,j=0,1$. It is verified that $0\le\varepsilon_1\le Min\left\{Se_1\left(1-Se_2\right),Se_2\left(1-Se_1\right)\right\}$ and $0\le\varepsilon_0\le Min\left\{Sp_1\left(1-Sp_2\right),Sp_2\left(1-Sp_1\right)\right\}$. If $\varepsilon_1=\varepsilon_0=0$ then the two *BDTs* are conditionally independent on the disease. In practice, the assumption of conditional independence is not realistic, and so $\varepsilon_1>0$ and/or $\varepsilon_0>0$. Let $\boldsymbol{\pi}=\left(p_{11},p_{10},p_{01},p_{00},q_{11},q_{10},q_{01},q_{00}\right)^T$ be the vector of probabilities of the multinomial distribution, and it is verified that $p=\sum_{i,j=0}^{1}p_{ij}$ and $q=1-p=\sum_{i,j=0}^{1}q_{ij}$. The maximum likelihood estimators of the probabilities are $\hat{p}_{ij}=s_{ij}/n$ and $\hat{q}_{ij}=r_{ij}/n$.

The rules given in Section 2 about the effect of *rTPF* and *rFPF* on the comparison of $\kappa_1(c)$ and $\kappa_2(c)$ are theoretical rules that can be applied to the estimators, but they cannot guarantee that one weighted kappa coefficient will be higher than another. This question should be studied through hypothesis tests and confidence intervals. The Bloch method to compare the weighted kappa coefficients of two *BDTs* subject to a paired design is summarized below, and different *CIs* are proposed to compare these parameters subject to the same type of sample design.

Table 2. Observed frequencies and theoretical probabilities when two *BDTs* are compared in relation to a *GS* subject to a paired design.

Observed frequencies (Theoretical probabilities)					
	$T_1=1$		$T_1=0$		
	$T_2=1$	$T_2=0$	$T_2=1$	$T_2=0$	Total
$D=1$	$s_{11}\ (p_{11})$	$s_{10}\ (p_{10})$	$s_{01}\ (p_{01})$	$s_{00}\ (p_{00})$	$s\ (p)$
$D=0$	$r_{11}\ (q_{11})$	$r_{10}\ (q_{10})$	$r_{01}\ (q_{01})$	$r_{00}\ (q_{00})$	$r\ (q)$
Total	$s_{11}+r_{11}$ $(p_{11}+q_{11})$	$s_{10}+r_{10}$ $(p_{10}+q_{10})$	$s_{01}+r_{01}$ $(p_{01}+q_{01})$	$s_{00}+r_{00}$ $(p_{00}+q_{00})$	$n\ (1)$

3.1. Hypothesis test

Bloch (1997) studied the comparison of the weighted kappa coefficients of two *BDTs* subject to a paired design. In terms of probabilities (6) and (7), the weighted kappa coefficient of *BDT* 1 is

$$\kappa_1(c)=\frac{(p_{11}+p_{10})(q_{01}+q_{00})-(p_{01}+p_{00})(q_{10}+q_{11})}{pc\sum_{k=0}^{1}(p_{0k}+q_{0k})+q(1-c)\sum_{k=0}^{1}(p_{1k}+q_{1k})},$$

and that of *BDT* 2 is

$$\kappa_2(c)=\frac{(p_{11}+p_{01})(q_{10}+q_{00})-(p_{10}+p_{00})(q_{01}+q_{11})}{pc\sum_{k=0}^{1}(p_{k0}+q_{k0})+q(1-c)\sum_{k=0}^{1}(p_{k1}+q_{k1})}.$$

Substituting in the previous expressions the parameters by their estimators, the estimators of the weighted kappa coefficients are

$$\hat{\kappa}_1(c)=\frac{(s_{11}+s_{10})(r_{01}+r_{00})-(s_{01}+s_{00})(r_{10}+r_{11})}{sc\sum_{k=0}^{1}(s_{0k}+r_{0k})+r(1-c)\sum_{k=0}^{1}(s_{1k}+r_{1k})} \tag{8}$$

and

$$\hat{\kappa}_2(c)=\frac{(s_{11}+s_{01})(r_{10}+r_{00})-(s_{10}+s_{00})(r_{01}+r_{11})}{sc\sum_{k=0}^{1}(s_{k0}+r_{k0})+r(1-c)\sum_{k=0}^{1}(s_{k1}+r_{k1})}. \tag{9}$$

Their variances-covariance are obtained applying the delta method (see the Appendix *B* of the supplementary material). Subject to paired design the covariance between the two sensitivities and between the two specificities are given by $Cov\left[\hat{Se}_1,\hat{Se}_2\right]=\frac{\varepsilon_1}{np}$ and $Cov\left[\hat{Sp}_1,\hat{Sp}_2\right]=\frac{\varepsilon_0}{nq}$ respectively (Appendix *B* of the supplementary material), where ε_1 and ε_0 are the covariances between the two *BDTs* when $D=1$ and $D=0$ respectively. These covariances also affect the covariances between the two weighted kappa coefficients, just as can be seen in the expressions given in the Appendix *B* of the supplementary material. Finally, the statistic for the hypothesis test $H_0:\kappa_1(c)=\kappa_2(c)$ vs $H_1:\kappa_1(c)\neq\kappa_2(c)$ is

$$z=\frac{\hat{\kappa}_1(c)-\hat{\kappa}_2(c)}{\sqrt{\hat{Var}\left[\hat{\kappa}_1(c)\right]+\hat{Var}\left[\hat{\kappa}_2(c)\right]-2\hat{Cov}\left[\hat{\kappa}_1(c),\hat{\kappa}_2(c)\right]}}\xrightarrow[n\to\infty]{}N(0,1). \tag{10}$$

3.2. Confidence intervals

When two parameters are compared, the interest is generally focused on studying the difference or the ratio between them. We then compare the weighted kappa coefficients of two *BDTs* through *CIs* for the difference $\delta=\kappa_1(c)-\kappa_2(c)$ and for the ratio $\theta=\kappa_1(c)/\kappa_2(c)$. Through the *CIs*: a) the two weighted kappa coefficients are compared, in such a way that if a *CI* for the difference (ratio) does not contain the zero (one) value, then we reject the equality between the weighted kappa coefficients; and b) we estimate (if the two weighted kappa coefficients are different) how much bigger one weighted kappa coefficient is than the other. Firstly, three *CIs* are proposed for the

difference of the two weighted kappa coefficients, and secondly five *CIs* are proposed for the ratio.

3.2.1. CIs for the difference

For the difference of the two weighted kappa coefficients we propose the Wald, bootstrap and Bayesian *CIs*.

Wald CI. Based on the asymptotic normality of the estimator of $\delta = \kappa_1(c) - \kappa_2(c)$, i.e. $\hat{\delta} \to N\left[\delta, Var(\delta)\right]$ when the sample size *n* is large, the Wald *CI* for the difference δ is very easy to obtain inverting the test statistic proposed by Bloch (1997), therefore

$$\delta \in \hat{\kappa}_1(c) - \hat{\kappa}_2(c) \pm z_{1-\alpha/2}\sqrt{\hat{Var}\left[\hat{\kappa}_1(c)\right] + \hat{Var}\left[\hat{\kappa}_2(c)\right] - 2\hat{Cov}\left[\hat{\kappa}_1(c), \hat{\kappa}_2(c)\right]}, \quad (11)$$

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

Bootstrap CI. The bootstrap *CI* is calculated generating *B* random samples with replacement from the sample of *n* individuals. In each sample with replacement, we calculate the estimators of the weighted kappa coefficients and the difference between them, i.e. $\hat{\kappa}_{i1B}(c)$, $\hat{\kappa}_{i2B}(c)$ and $\hat{\delta}_{iB} = \hat{\kappa}_{i1B}(c) - \hat{\kappa}_{i2B}(c)$, with $i = 1,...,B$. Then, based on the *B* differences calculated, the average difference is estimated as $\hat{\bar{\delta}}_B = \frac{1}{B}\sum_{i=1}^{B}\hat{\delta}_{iB}$. Assuming that the bootstrap statistic $\hat{\bar{\delta}}_B$ can be transformed to a normal distribution, the bias-corrected bootstrap *CI* (Efron and Tibshirani, 1993) for δ is calculated in the following way. Let $A = \#\left(\hat{\delta}_{iB} < \hat{\delta}\right)$ be the number of bootstrap estimators $\hat{\delta}_{iB}$ that are lower than the maximum likelihood estimator $\hat{\delta} = \hat{\kappa}_1(c) - \hat{\kappa}_2(c)$, and let $\hat{z}_0 = \Phi^{-1}(A/B)$, where $\Phi^{-1}(\cdot)$ is the inverse function of the standard normal cumulative distribution function. Let $\alpha_1 = \Phi\left(2\hat{z}_0 - z_{1-\alpha/2}\right)$ and $\alpha_2 = \Phi\left(2\hat{z}_0 + z_{1-\alpha/2}\right)$, then the bias-corrected bootstrap *CI* is $\left(\hat{\delta}_B^{(\alpha_1)}, \hat{\delta}_B^{(\alpha_2)}\right)$, where $\hat{\delta}_B^{(\alpha_j)}$ is the *j*th quantile of the distribution of the *B* bootstrap estimations of δ.

Bayesian CI. The problem is now approached from a Bayesian perspective. The number of individuals with the disease (*s*) is the product of a binomial distribution with

parameters *n* and *p*, i.e. $s \to B(n, p)$. Conditioning on the individuals with the disease, i.e. conditioning on $D = 1$, it is verified that

$$s_{11} + s_{10} \to B(s, Se_1) \text{ and } s_{11} + s_{01} \to B(s, Se_2). \quad (12)$$

The number of individuals without the disease (*r*) is the product of a binomial distribution with parameters *n* and *q*, i.e. $s \to B(n, q)$, with $q = 1 - p$. Conditioning on the individuals without the disease $(D = 0)$, it is verified that

$$r_{01} + r_{00} \to B(r, Sp_1) \text{ and } r_{10} + r_{00} \to B(r, Sp_2). \quad (13)$$

Considering the marginal distributions of each *BDT*, the estimators of the sensitivity and the specificity of the *BDT* 1, $\hat{Se}_1 = \frac{s_{11} + s_{10}}{s}$ and $\hat{Sp}_1 = \frac{r_{01} + r_{00}}{r}$, and of the *BDT* 2, $\hat{Se}_2 = \frac{s_{11} + s_{01}}{s}$ and $\hat{Sp}_2 = \frac{r_{10} + r_{00}}{r}$, are estimators of binomial proportions. In a similar way, considering the marginal distribution of the *GS*, the estimator of the disease prevalence, $\hat{p} = \frac{s}{n}$, is also the estimator of a binomial proportion. Therefore, for these estimators we propose conjugate beta prior distributions, which are the appropriate distributions for the binomial distributions involved, i.e.

$$\hat{Se}_h \to Beta(\alpha_{Se_h}, \beta_{Se_h}),\ \hat{Sp}_h \to Beta(\alpha_{Sp_h}, \beta_{Sp_h}) \text{ and } \hat{p} \to Beta(\alpha_p, \beta_p). \quad (14)$$

Let $\mathbf{v} = (s_{11}, s_{10}, s_{01}, s, r_{11}, r_{10}, r_{01}, n - s)$ be the vector of observed frequencies, with $s_{00} = s - s_{11} - s_{10} - s_{01}$, $r = n - s$ and $r_{00} = n - s - r_{11} - r_{10} - r_{01}$. Then the posteriori distributions for the estimators of the sensitivities, of the specificities and of the prevalence are:

$$\begin{aligned}
\hat{Se}_1 | \mathbf{v} &\to Beta(s_{11} + s_{10} + \alpha_{Se_1}, s - s_{11} - s_{10} + \beta_{Se_1}), \\
\hat{Se}_2 | \mathbf{v} &\to Beta(s_{11} + s_{01} + \alpha_{Se_2}, s - s_{11} - s_{01} + \beta_{Se_2}), \\
\hat{Sp}_1 | \mathbf{v} &\to Beta(r_{01} + r_{00} + \alpha_{Sp_1}, n - s - r_{01} - r_{00} + \beta_{Sp_1}), \\
\hat{Sp}_2 | \mathbf{v} &\to Beta(r_{10} + r_{00} + \alpha_{Sp_2}, n - s - r_{10} - r_{00} + \beta_{Sp_2}), \\
\hat{p} | \mathbf{v} &\to Beta(s + \alpha_p, n - s + \beta_p).
\end{aligned} \quad (15)$$

Once we have defined all distributions, the posteriori distribution for the weighted kappa coefficient of each *BDT*, and for the difference between them, can be approximated applying the Monte Carlo method. This method consists of generating *M* values of the posteriori distributions given in equations (15). In the *i*th iteration, the values generated for sensitivities $\left(\hat{Se}_h^{(i)}\right)$ and specificities $\left(\hat{Sp}_h^{(i)}\right)$ of each *BDT*, and for the prevalence $\left(\hat{p}^{(i)}\right)$, are plugged in the equations

$$\hat{\kappa}_h^{(i)}(c) = \frac{\hat{p}^{(i)}\hat{q}^{(i)}\left(\hat{Se}_h^{(i)} + \hat{Sp}_h^{(i)} - 1\right)}{\hat{p}^{(i)}\left(1-\hat{Q}_h^{(i)}\right)c + \hat{q}^{(i)}\hat{Q}_h^{(i)}\left(1-c\right)}, \quad h = 1,2, \tag{16}$$

where $\hat{Q}_h^{(i)} = \hat{p}^{(i)}\hat{Se}_h^{(i)} + \hat{q}^{(i)}\left(1-\hat{Sp}_h^{(i)}\right)$. We then calculate the difference between the two weighted kappa coefficients in the *i*th iteration: $\hat{\delta}^{(i)} = \hat{\kappa}_1^{(i)}\left(c\right) - \hat{\kappa}_2^{(i)}\left(c\right)$. As the estimator of the average difference of the weighted kappa coefficients, we calculate the average of the *M* estimations of difference, i.e. $\hat{\bar{\delta}} = \frac{1}{M}\sum_{i=1}^{M}\hat{\delta}^{(i)}$. Once the Monte Carlo method is applied, based on the *M* values $\hat{\delta}^{(i)}$ we propose the calculation of a *CI* based on quantiles, i.e. the $100\times\left(1-\alpha\right)\%$ *CI* for δ is

$$\left(q_{\alpha/2}, q_{1-\alpha/2}\right), \tag{17}$$

where q_γ is the γth quantile of the distribution of the *M* values $\hat{\delta}^{(i)}$.

3.2.2. CIs for the ratio

We propose five *CIs* for the ratio of the two weighted kappa coefficients: Wald, logarithmic, Fieller, bootstrap and Bayesian *CIs*.

Wald CI. Assuming the asymptotic normality of the estimator of $\theta = \kappa_1\left(c\right)/\kappa_2\left(c\right)$, i.e. $\hat{\theta} \to N\left[\theta, \mathrm{Var}\left(\theta\right)\right]$ when the sample size *n* is large, the Wald *CI* for θ is

$$\theta \in \hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{Var}\left(\hat{\theta}\right)}, \tag{18}$$

where $\hat{Var}\left(\hat{\theta}\right)$ is obtained applying the delta method (Agresti, 2002), and whose expression (see Appendix B) is

$$\hat{Var}\left(\hat{\theta}\right) \approx \frac{\hat{\kappa}_2^2(c)\hat{Var}\left[\hat{\kappa}_1(c)\right] + \hat{\kappa}_1^2(c)\hat{Var}\left[\hat{\kappa}_2(c)\right] - 2\hat{\kappa}_1(c)\hat{\kappa}_2(c)\hat{Cov}\left[\hat{\kappa}_1(c), \hat{\kappa}_2(c)\right]}{\hat{\kappa}_2^4(c)}.$$

Expressions of the variances-covariance can be seen in Appendix B.

Logarithmic CI. Assuming the asymptotic normality of the Napierian logarithm of the $\hat{\theta}$, i.e. $\ln\left(\hat{\theta}\right) \to N\left(\ln(\theta), \text{Var}\left[\ln(\theta)\right]\right)$ when the sample size *n* is large, an asymptotic *CI* for $\ln(\theta)$ is

$$\ln(\theta) \in \ln\left(\hat{\theta}\right) \pm z_{1-\alpha/2}\sqrt{\hat{Var}\left[\ln\left(\hat{\theta}\right)\right]}.$$

Taking exponential, the logarithmic *CI* for θ is

$$\theta \in \hat{\theta} \times \exp\left\{\pm z_{1-\alpha/2}\sqrt{\hat{Var}\left[\ln\left(\hat{\theta}\right)\right]}\right\}, \tag{19}$$

where $\hat{Var}\left[\ln\left(\hat{\theta}\right)\right]$ is obtained applying the delta method (see Appendix B), i.e.

$$\hat{Var}\left[\ln\left(\hat{\theta}\right)\right] \approx \frac{\hat{Var}\left[\hat{\kappa}_1(c)\right]}{\hat{\kappa}_1^2(c)} + \frac{\hat{Var}\left[\hat{\kappa}_2(c)\right]}{\hat{\kappa}_2^2(c)} - \frac{2\hat{Cov}\left[\hat{\kappa}_1(c), \hat{\kappa}_2(c)\right]}{\hat{\kappa}_1(c)\hat{\kappa}_2(c)}.$$

Fieller CI. The Fieller method (1940) is a classic method to obtain a *CI* for the ratio of two parameters. This method requires us to assume that the estimators are distributed according to a normal bivariate distribution, i.e. $\left(\hat{\kappa}_1(c), \hat{\kappa}_2(c)\right)^T \to N\left[\boldsymbol{\kappa}(c), \Sigma_{\boldsymbol{\kappa}(c)}\right]$ when the sample size *n* is large, where $\boldsymbol{\kappa}(c) = \left(\kappa_1(c), \kappa_2(c)\right)^T$ and

$$\Sigma_{\boldsymbol{\kappa}(c)} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} Var\left[\kappa_1(c)\right] & Cov\left[\kappa_1(c), \kappa_2(c)\right] \\ Cov\left[\kappa_1(c), \kappa_2(c)\right] & Var\left[\kappa_2(c)\right] \end{pmatrix}.$$

Applying the Fieller method it is verified that $\hat{\kappa}_1(c) - \theta\hat{\kappa}_2(c) \xrightarrow[n\to\infty]{} N\left(0, \sigma_{11} - 2\theta\sigma_{12} + \theta^2\sigma_{22}\right)$. The Fieller *CI* is obtained by searching for the set of values for θ that satisfy the inequality

$$\frac{\left[\hat{\kappa}_1(c)-\theta\hat{\kappa}_2(c)\right]^2}{\hat{\sigma}_{11}-2\theta\hat{\sigma}_{12}+\theta^2\hat{\sigma}_{22}} < z^2_{1-\alpha/2}\,.$$

Finally, the Fieller *CI* for $\theta=\kappa_1(c)/\kappa_2(c)$ is

$$\theta \in \frac{\hat{\omega}_{12} \pm \sqrt{\hat{\omega}_{12}^2-\hat{\omega}_{11}\hat{\omega}_{22}}}{\hat{\omega}_{22}}, \tag{20}$$

where $\hat{\omega}_{ij}=\hat{\kappa}_i(c)\times\hat{\kappa}_j(c)-\hat{\sigma}_{ij}z^2_{1-\alpha/2}$ with $i,j=1,2$, and verifying that $\hat{\omega}_{12}=\hat{\omega}_{21}$. This interval is valid when $\hat{\omega}_{12}^2>\hat{\omega}_{11}\hat{\omega}_{22}$ and $\hat{\omega}_{22}\neq 0$.

Bootstrap CI. The bootstrap *CI* for θ is calculated in a similar way to that of the bootstrap interval explained in Section 3.1 but considering θ instead of δ. In each sample with replacement obtained we calculate the estimators of the weighted kappa coefficients and the ratio between them, i.e. $\hat{\kappa}_{i1B}(c)$, $\hat{\kappa}_{i2B}(c)$ and $\hat{\theta}_{iB}=\hat{\kappa}_{i1B}(c)/\hat{\kappa}_{i2B}(c)$, with $i=1,...,B$. Then, based on the *B* ratios calculated we estimate the average ratio as $\hat{\bar{\theta}}_B=\frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_{iB}$. Assuming that the statistic $\hat{\bar{\theta}}_B$ can be transformed to a normal distribution, the bias-corrected bootstrap *CI* (Efron and Tibshirani, 1993) for θ is obtained in a similar way to how the bootstrap *CI* for δ is calculated, considering now that $A=\#\left(\hat{\theta}_{iB}<\hat{\theta}\right)$. Finally, the bias-corrected bootstrap *CI* is $\left(\hat{\theta}_B^{(\alpha_1)},\hat{\theta}_B^{(\alpha_2)}\right)$, where $\hat{\theta}_B^{(\alpha_j)}$ is the *j*th quantile of the distribution of the *B* bootstrap estimations of θ.

Bayesian CI. The Bayesian *CI* for θ is also calculated in a similar way to that of the bayesian *CI* presented in Section 3.1. Considering the same distributions given in equations (14) and (15), in the *i*th iteration of the Monte Carlo method we calculate the ratio $\hat{\theta}^{(i)}=\hat{\kappa}_1^{(i)}(c)/\hat{\kappa}_2^{(i)}(c)$ and as an estimator we calculate $\hat{\bar{\theta}}=\frac{1}{M}\sum_{i=1}^{M}\hat{\theta}^{(i)}$. Finally, based on the *M* values $\hat{\theta}^{(i)}$ we calculate the *CI* based on quantiles.

The five previous *CIs* are for the ratio $\theta=\kappa_1(c)/\kappa_2(c)$. If we want to calculate the *CI* for the ratio $\kappa_2(c)/\kappa_1(c)$ $\left(=\theta'=1/\theta\right)$, then the logarithmic, Fieller, bootstrap and Bayesian *CIs* are obtained by calculating the inverse of each boundary of the

corresponding *CI* for $\theta = \kappa_1(c)/\kappa_2(c)$. Nevertheless, the Wald *CI* for θ' is obtained from the Wald *CI* for θ dividing each boundary by $\hat{\theta}^2$, i.e. if (L_θ, U_θ) is the Wald *CI* for $\theta = \kappa_1(c)/\kappa_2(c)$ then the Wald *CI* for $\theta' = \kappa_2(c)/\kappa_1(c)$ is $\left(L_\theta/\hat{\theta}^2, U_\theta/\hat{\theta}^2\right)$.

4. Simulation experiments

Monte Carlo simulation experiments were carried out to study the coverage probability (*CP*) and the average length (*AL*) of each of the *CIs* presented in Section 3.2. For this purpose, we generated $N = 10{,}000$ random samples with multinomial distribution sized $n = \{25, 50, 100, 200, 300, 400, 500, 1000\}$. The random samples were generated setting the values of the weighted kappa coefficients, following these steps:

1. For the disease prevalence, we took the values $p = \{5\%, 10\%, 25\%, 50\%\}$.

2. For the weighting index, we took a small, intermediate and high value: $c = \{0.1, 0.5, 0.9\}$.

3. As values of the weighted kappa coefficients with $c = 0$ and $c = 1$, we took the following values: $\kappa_h(0), \kappa_h(1) = \{0.01, 0.02, ..., 0.98, 0.99\}$.

4. Next, using all of the values set previously, we calculated the sensitivity and the specificity of each diagnostic test solving the equations

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

considering, quite logically, only those cases in which the Youden index is higher than 0, i.e. $Y_h = Se_h + Sp_h - 1 > 0$.

5. The values of $\kappa_h(c)$ were calculated applying the equation

$$\kappa_h(c) = \frac{pc(1-Q_h)\kappa_h(1) + q(1-c)Q_h\kappa_h(0)}{pc(1-Q_h) + q(1-c)Q_h},$$

where $Q_h = pSe_h + q(1 - Sp_h)$.

6. As values of the weighted kappa coefficients we considered $\kappa_h(c)=\{0.2,0.4,0.6,0.8\}$, and from these we calculated δ and θ. In order to be able to compare the coverage probabilities of the *CIs* for δ and for θ, $\kappa_1(c)$ and $\kappa_2(c)$ must be the same for δ and θ.

Following the idea of Cicchetti (2001), simulations were carried out for values of $\kappa_h(c)$ with different levels of significance: poor $(\kappa_h(c)<0.40)$, fair $(0.40\le\kappa_h(c)\le 0.59)$, good $(0.60\le\kappa_h(c)\le 0.74)$ and excellent $(0.75\le\kappa_h(c)\le 1)$. As values of the dependence factors ε_1 and ε_0 we took intermediate values (50% of the maximum value of each ε_i) and high values (80% of the maximum value of each ε_i), i.e. $\varepsilon_1 = f\times Min\{Se_1(1-Se_2), Se_2(1-Se_1)\}$ and $\varepsilon_0 = f\times Min\{Sp_1(1-Sp_2), Sp_2(1-Sp_1)\}$ where $f=\{0.50,\ 0.80\}$. Probabilities of the multinomial distributions, equations (6) and (7), were calculated from values of the weighted kappa coefficients, and not setting the values of the sensitivities and specificities. In each scenario considered, for each one of the *N* random samples we calculated all the *CIs* proposed in Section 3.2. For the bayesian *CIs* we considered as prior distribution a $Beta(1,1)$ distribution for all of the estimators (sensitivities, specificities and prevalence). This distribution is a non-informative distribution and is flat for all possible values of each sensitivity, specificity and prevalence, and has a minimum impact on each posteriori distribution. For the bootstrap method, for each one of the *N* random samples we also generated $B=2,000$ samples with replacement; and for the Bayesian method, for each one of the *N* random samples we also generated another $M=10,000$. Moreover, the simulation experiments were designed in such a way that in all of the random samples generated we can estimate the weighted kappa coefficients and their variances-covariance, in order to be able to calculate all of the intervals proposed in Section 3.2. As the confidence level, we took 95%.

The comparison of the asymptotic behaviour of the *CIs* was made following a similar procedure to that used by other authors (Price and Bonett, 2004; Martín-Andrés and Álvarez-Hernández, 2014a, 2014b; Montero-Alonso and Roldán-Nofuentes, 2019). This procedure consists of determining if the *CI* "fails" for a confidence of 95%, which happens if the *CI* has a $CP\le 93\%$. The selection of the *CI* with the best asymptotic

behaviour (for the difference and for the ratio) was made following the following steps: 1) Choose the *CIs* with the least failures ($CP > 93\%$), and 2) Choose the *CIs* which are the most accurate, i.e. those which have the lowest *AL*. In the Appendix *C* of the supplementary material this method is justified.

4.1. CIs for the difference δ

Tables 3 and 4 show some of the results obtained (*CPs* and *ALs*) for $\delta = \{-0.6, -0.4, -0.2, 0\}$, indicating in each case the scenarios ($\kappa_h(c)$, Se_h, Sp_h and p) in which these values were obtained, and for intermediate values of the dependence factors ε_1 and ε_0. These Tables indicate the failures in bold type and it was considered that $\kappa_1(c) \leq \kappa_2(c)$. If it is considered that $\kappa_1(c) > \kappa_2(c)$, the *CPs* are the same and the conclusions too. From the results, the following conclusions are obtained:

a) Wald *CI*. For $\delta = \{-0.6, -0.4\}$ the Wald *CI* fails for a small $(n \leq 50)$ and a moderate $(n = 100)$ sample size, and for a large sample size $(n \geq 200)$ the Wald *CI* does not fail. For $\delta = \{-0.2, 0\}$ the Wald *CI* does not fail.

b) Bootstrap *CI*. In very general terms, for $\delta = \{-0.6, -0.4\}$ this *CI* fails when $n \leq 100$, and for $n \geq 200$ this interval does not fail. For $\delta = -0.2$ this *CI* fails for almost all the sample sizes, and for $\delta = 0$ does not fail. When this *CI* does not fail, the *AL* is slightly lower than the Wald *CI* for $\delta = \{-0.2, 0\}$, and slightly higher for $\delta = \{-0.6, -0.4\}$ and $n \geq 200$.

c) Bayesian *CI*. In very general terms, for $\delta = \{-0.6, -0.4\}$ this *CI* fails when $n \leq 50$, whereas for $n \geq 100$ this *CI* does not fail. For $\delta = \{-0.2, 0\}$ this *CI* does not fail. Regarding the *AL*, in the situations in which it does not fail, the *AL* is slightly higher than the *ALs* of the Wald *CI* and of the bootstrap *CI*.

Table 3. Coverage probabilities (*CP*s) and average lengths (*AL*s) of the *CI*s for the difference δ of the two weighted kappa coefficients (I).

$\kappa_1(0.1)=0.2\ \ \kappa_2(0.1)=0.8\ \ \delta=-0.6$

$Se_1=0.484\ \ Sp_1=0.684\ \ Se_2=0.852\ \ Sp_2=0.911\ \ \varepsilon_1=0.0359\ \ \varepsilon_0=0.0306\ \ p=50\%$

	Wald *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	**0.335**	0.866	**0**	0.643	**0.287**	0.923
50	**0.737**	0.646	**0.038**	0.589	**0.762**	0.690
100	**0.912**	0.470	**0.750**	0.473	0.937	0.501
200	0.958	0.337	0.952	0.354	0.968	0.364
300	0.972	0.276	0.980	0.295	0.982	0.301
400	0.960	0.239	0.969	0.258	0.971	0.262
500	0.955	0.214	0.972	0.231	0.975	0.236
1000	0.937	0.152	0.963	0.164	0.965	0.168

$\kappa_1(0.9)=0.2\ \ \kappa_2(0.9)=0.8\ \ \delta=-0.6$

$Se_1=0.28\ \ Sp_1=0.92\ \ Se_2=0.82\ \ Sp_2=0.98\ \ \varepsilon_1=0.0252\ \ \varepsilon_0=0.0092\ \ p=10\%$

	Wald *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	**0.114**	0.999	**0**	0.651	**0.033**	0.987
50	**0.566**	0.863	**0**	0.640	**0.280**	0.838
100	**0.760**	0.682	**0.031**	0.614	**0.600**	0.667
200	**0.885**	0.503	**0.487**	0.490	**0.815**	0.503
300	0.934	0.411	**0.733**	0.402	**0.886**	0.418
400	0.935	0.354	**0.823**	0.347	**0.903**	0.365
500	0.947	0.314	**0.892**	0.309	0.937	0.326
1000	0.947	0.220	0.938	0.218	0.947	0.233

$\kappa_1(0.1)=0.4\ \ \kappa_2(0.1)=0.8\ \ \delta=-0.4$

$Se_1=0.804\ \ Sp_1=0.887\ \ Se_2=0.82\ \ Sp_2=0.98\ \ \varepsilon_1=0.0723\ \ \varepsilon_0=0.0089\ \ p=10\%$

	Wald *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	**0.847**	0.812	**0.473**	0.671	**0.920**	0.899
50	**0.856**	0.715	**0.602**	0.608	**0.910**	0.764
100	**0.924**	0.534	**0.847**	0.528	0.953	0.580
200	0.968	0.373	0.955	0.423	0.978	0.426
300	0.957	0.302	0.986	0.367	0.976	0.369
400	0.951	0.261	0.992	0.313	0.978	0.315
500	0.955	0.232	0.994	0.259	0.979	0.262
1000	0.941	0.164	0.994	0.202	0.967	0.204

$\kappa_1(0.5)=0.4\ \ \kappa_2(0.5)=0.8\ \ \delta=-0.4$

$Se_1=0.76\ \ Sp_1=0.72\ \ Se_2=0.85\ \ Sp_2=0.95\ \ \varepsilon_1=0.0570\ \ \varepsilon_0=0.0180\ \ p=25\%$

	Wald *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	**0.894**	0.810	**0.004**	0.613	0.962	0.858
50	0.935	0.580	**0.516**	0.516	0.961	0.641
100	0.945	0.397	**0.824**	0.379	0.970	0.458
200	0.946	0.275	**0.928**	0.271	0.971	0.320
300	0.952	0.221	0.934	0.220	0.974	0.259
400	0.940	0.191	0.938	0.192	0.963	0.224
500	0.948	0.171	0.942	0.170	0.979	0.200
1000	0.945	0.120	0.944	0.119	0.979	0.140

Table 4. Coverage probabilities (*CPs*) and average lengths (*ALs*) of the *CIs* for the difference δ of the two weighted kappa coefficients (II).

$\kappa_1(0.9)=0.6 \;\; \kappa_2(0.9)=0.8 \;\; \delta=-0.2$

$Se_1=0.62 \;\; Sp_1=0.98 \;\; Se_2=0.911 \;\; Sp_2=0.937 \;\; \varepsilon_1=0.0277 \;\; \varepsilon_0=0.0094 \;\; p=5\%$

	Wald *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	1	1.009	**0.757**	0.724	1	1.018
50	0.996	0.913	**0.829**	0.659	0.999	0.916
100	0.993	0.823	**0.928**	0.580	0.998	0.801
200	0.934	0.642	**0.763**	0.535	0.986	0.649
300	0.922	0.533	**0.745**	0.483	0.964	0.551
400	0.941	0.456	**0.794**	0.434	0.971	0.481
500	0.933	0.404	**0.799**	0.393	0.962	0.430
1000	0.948	0.282	**0.913**	0.282	0.967	0.305

$\kappa_1(0.1)=0.6 \;\; \kappa_2(0.1)=0.8 \;\; \delta=-0.2$

$Se_1=0.195 \;\; Sp_1=0.995 \;\; Se_2=0.477 \;\; Sp_2=0.987 \;\; \varepsilon_1=0.0509 \;\; \varepsilon_0=0.0026 \;\; p=25\%$

	Wald *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	1	0.928	1	0.644	1	0.981
50	0.999	0.787	1	0.613	1	0.866
100	0.994	0.604	0.999	0.581	0.999	0.692
200	0.985	0.429	0.997	0.464	0.998	0.505
300	0.981	0.347	0.991	0.393	0.994	0.411
400	0.973	0.297	0.986	0.346	0.992	0.352
500	0.967	0.263	0.984	0.311	0.989	0.311
1000	0.957	0.182	0.988	0.222	0.987	0.213

$\kappa_1(0.5)=0.4 \;\; \kappa_2(0.5)=0.4 \;\; \delta=0$

$Se_1=0.76 \;\; Sp_1=0.72 \;\; Se_2=0.40 \;\; Sp_2=0.943 \;\; \varepsilon_1=0.0480 \;\; \varepsilon_0=0.0206 \;\; p=25\%$

	Wald *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	0.990	0.811	0.988	0.624	0.999	0.826
50	0.978	0.683	0.998	0.598	0.994	0.691
100	0.962	0.499	0.967	0.466	0.985	0.522
200	0.955	0.353	0.963	0.340	0.981	0.381
300	0.944	0.288	0.943	0.280	0.965	0.314
400	0.960	0.250	0.962	0.244	0.980	0.274
500	0.946	0.223	0.945	0.219	0.966	0.246
1000	0.951	0.158	0.951	0.155	0.972	0.175

$\kappa_1(0.9)=0.4 \;\; \kappa_2(0.9)=0.4 \;\; \delta=0$

$Se_1=0.943 \;\; Sp_1=0.229 \;\; Se_2=0.70 \;\; Sp_2=0.70 \;\; \varepsilon_1=0.0200 \;\; \varepsilon_0=0.0343 \;\; p=50\%$

	Wald *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	1	0.936	1	0.735	1	0.950
50	0.997	0.788	0.997	0.717	1	0.786
100	0.992	0.602	0.982	0.578	0.997	0.617
200	0.980	0.435	0.981	0.432	0.990	0.461
300	0.959	0.356	0.965	0.358	0.973	0.382
400	0.951	0.307	0.958	0.311	0.972	0.332
500	0.956	0.274	0.958	0.278	0.969	0.297
1000	0.956	0.193	0.958	0.196	0.970	0.210

Similar conclusions are obtained when the dependence factors take high values. Therefore, regarding the effect of the dependence factors ε_i on the asymptotic behaviour of the *CIs*, in general terms they do not have a clear effect on the *CPs* of the *CIs*.

4.2. CIs for the ratio θ

Tables 5 and 6 show some of the results obtained for $\theta = \{0.25, 0.50, 0.75, 1\}$, considering the same scenarios as in Tables 3 and 4. As in the case of the previous *CIs*, it was considered that $\kappa_1(c) \leq \kappa_2(c)$, and the same conclusions are obtained if $\kappa_1(c) > \kappa_2(c)$. From the results, the following conclusions are obtained:

a) Wald *CI*. The Wald *CI* fails when $\theta = 0.25$ and the sample size is small $(n \leq 50)$ or moderate $(n = 100)$, and this *CI* does not fail for the rest of the values of θ and sample sizes.

b) Logarithmic *CI*. This *CI* fails when $\theta = \{0.25, 0.50\}$ and $n \leq 200 - 300$ depending on the value of θ. For $\theta = 0.75$ this *CI* fails for some large sample sizes, and for $\theta = 1$ it does not fail. This *CI* fails more than the Wald *CI*, and in the situations in which it does not fail, its *AL* is slightly higher than that of the Wald *CI*.

c) Fieller *CI*. This *CI* fails when $\theta = \{0.25, 0.5\}$ and $n \leq 50$, and it does not fail for the rest of the values of θ and sample sizes. In general terms, when there are no failures, its *AL* is similar to that of the Wald and Logarithmic *CIs*.

d) Bootstrap *CI*. This *CI* has numerous failures when $\theta = \{0.25, 0.50, 0.75\}$, whereas for $\theta = 1$ it does not fail. When $\theta = 1$, its *AL* is greater than that of the Wald and Logarithmic *CIs*, especially when $n \leq 400$, and its *AL* is also slightly lower than that of the Fieller *CI*.

e) Bayesian *CI*. This *CI* only fails when $\theta = 0.25$ and $n \leq 50$. When this *CI* does not fail, its *AL* is, in general terms, somewhat larger than that of the rest of the *CIs*.

Table 5. Coverage probabilities (*CPs*) and average lengths (*ALs*) of the *CIs* for the ratio θ of the two weighted kappa coefficients (I).

$\kappa_1(0.1)=0.2 \quad \kappa_2(0.1)=0.8 \quad \theta=0.25$

$Se_1=0.484 \quad Sp_1=0.684 \quad Se_2=0.852 \quad Sp_2=0.911 \quad \varepsilon_1=0.0359 \quad \varepsilon_0=0.0306 \quad p=50\%$

	Wald *CI*		Logarit. *CI*		Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	**0.823**	1.351	**0.088**	1.517	**0.700**	1.950	**0.368**	2.260	**0.884**	2.704
50	**0.837**	0.803	**0.532**	0.886	**0.828**	0.851	**0.634**	0.882	**0.905**	0.965
100	0.931	0.551	**0.832**	0.608	0.942	0.565	**0.889**	0.569	0.954	0.585
200	0.957	0.389	**0.920**	0.422	0.962	0.392	0.952	0.388	0.970	0.402
300	0.970	0.318	0.933	0.340	0.974	0.319	0.969	0.316	0.984	0.328
400	0.960	0.277	0.936	0.293	0.967	0.278	0.962	0.276	0.976	0.285
500	0.957	0.248	0.944	0.260	0.967	0.248	0.969	0.247	0.975	0.256
1000	0.945	0.175	0.963	0.179	0.944	0.176	0.943	0.175	0.953	0.182

$\kappa_1(0.9)=0.2 \quad \kappa_2(0.9)=0.8 \quad \theta=0.25$

$Se_1=0.28 \quad Sp_1=0.92 \quad Se_2=0.82 \quad Sp_2=0.98 \quad \varepsilon_1=0.0252 \quad \varepsilon_0=0.00092 \quad p=10\%$

	Wald *CI*		Logarit. *CI*		Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	**0.885**	1.760	**0.002**	2.029	**0.566**	3.567	**0.011**	3.175	**0.866**	3.851
50	**0.916**	1.249	**0.259**	1.415	**0.765**	1.660	**0.040**	1.722	**0.767**	1.816
100	0.936	0.846	**0.636**	0.947	**0.884**	0.939	**0.363**	1.048	**0.843**	0.986
200	0.958	0.560	**0.835**	0.617	0.945	0.581	**0.807**	0.607	0.932	0.594
300	0.967	0.440	**0.900**	0.479	0.960	0.450	**0.902**	0.456	0.948	0.459
400	0.965	0.373	0.931	0.402	0.959	0.379	0.932	0.380	0.943	0.387
500	0.971	0.327	0.936	0.349	0.971	0.331	0.942	0.330	0.960	0.339
1000	0.950	0.227	0.941	0.235	0.950	0.228	0.949	0.227	0.955	0.234

$\kappa_1(0.1)=0.4 \quad \kappa_2(0.1)=0.8 \quad \theta=0.5$

$Se_1=0.804 \quad Sp_1=0.887 \quad Se_2=0.82 \quad Sp_2=0.98 \quad p=10\%$

$\varepsilon_1=0.0723 \quad \varepsilon_0=0.0089$

	Wald *CI*		Logarit. *CI*		Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	**0.918**	1.141	**0.835**	1.259	**0.893**	2.824	**0.543**	1.157	**0.906**	2.310
50	0.959	1.021	**0.859**	1.119	0.939	1.518	**0.897**	1.140	0.978	1.710
100	0.961	0.619	**0.922**	0.655	0.949	0.693	**0.880**	0.670	0.975	0.828
200	0.962	0.395	0.947	0.406	0.959	0.409	**0.914**	0.400	0.977	0.470
300	0.955	0.315	0.951	0.320	0.956	0.321	**0.928**	0.312	0.976	0.363
400	0.953	0.271	0.949	0.274	0.952	0.274	0.935	0.265	0.975	0.308
500	0.951	0.240	0.950	0.242	0.953	0.242	0.932	0.234	0.971	0.271
1000	0.939	0.169	0.943	0.170	0.939	0.170	0.934	0.163	0.963	0.189

$\kappa_1(0.5)=0.4 \quad \kappa_2(0.5)=0.8 \quad \theta=0.5$

$Se_1=0.76 \quad Sp_1=0.72 \quad Se_2=0.85 \quad Sp_2=0.95 \quad \varepsilon_1=0.0570 \quad \varepsilon_0=0.0180 \quad p=25\%$

	Wald *CI*		Logarit. *CI*		Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	0.997	1.328	**0.918**	1.493	0.966	2.222	**0.901**	2.463	0.999	2.825
50	0.983	0.780	**0.924**	0.848	0.966	0.855	**0.925**	0.894	0.995	1.057
100	0.977	0.488	0.957	0.510	0.969	0.501	0.952	0.498	0.990	0.586
200	0.958	0.323	0.956	0.329	0.957	0.327	0.940	0.320	0.981	0.372
300	0.958	0.257	0.954	0.260	0.957	0.259	0.945	0.252	0.978	0.292
400	0.948	0.221	0.947	0.222	0.948	0.221	0.936	0.215	0.966	0.249
500	0.954	0.196	0.953	0.197	0.954	0.196	0.943	0.190	0.972	0.220
1000	0.944	0.137	0.951	0.137	0.945	0.137	0.933	0.132	0.968	0.152

Table 6. Coverage probabilities (*CPs*) and average lengths (*ALs*) of the *CIs* for the ratio θ of the two weighted kappa coefficients (II).

$\kappa_1(0.9)=0.6 \;\; \kappa_2(0.9)=0.8 \;\; \theta=0.75$

$Se_1=0.62 \;\; Sp_1=0.98 \;\; Se_2=0.911 \;\; Sp_2=0.936 \;\; \varepsilon_1=0.0277 \;\; \varepsilon_0=0.0094 \;\; p=5\%$

	Wald *CI*		Logarit. *CI*		Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	1	1.514	1	1.679	1	2.689	0.999	2.578	1	3.538
50	0.999	1.409	0.994	1.487	0.993	1.972	0.979	2.311	1	2.392
100	0.999	1.323	0.993	1.451	0.993	1.899	0.975	1.425	1	1.980
200	0.971	0.909	0.933	0.965	0.940	1.037	0.965	0.998	0.991	1.173
300	0.946	0.709	0.916	0.738	0.939	0.767	0.958	0.784	0.973	0.854
400	0.955	0.583	0.933	0.599	0.944	0.601	0.959	0.620	0.977	0.679
500	0.943	0.506	0.925	0.516	0.931	0.516	0.961	0.551	0.969	0.579
1000	0.947	0.341	0.945	0.344	0.943	0.344	0.969	0.375	0.969	0.377

$\kappa_1(0.1)=0.6 \;\; \kappa_2(0.1)=0.8 \;\; \theta=0.75$

$Se_1=0.195 \;\; Sp_1=0.995 \;\; Se_2=0.477 \;\; Sp_2=0.987 \;\; \varepsilon_1=0.0509 \;\; \varepsilon_0=0.0026 \;\; p=25\%$

	Wald *CI*		Logarit. *CI*		Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	1	1.687	1	1.924	1	4.747	1	2.676	1	4.561
50	1	1.266	1	1.400	1	2.837	1	1.609	1	2.308
100	0.999	0.865	0.997	0.923	0.997	0.946	0.998	0.945	1	1.188
200	0.992	0.565	0.990	0.583	0.986	0.579	0.975	0.618	0.997	0.700
300	0.971	0.444	0.990	0.452	0.976	0.449	0.958	0.493	0.992	0.536
400	0.971	0.375	0.985	0.380	0.972	0.378	0.960	0.420	0.989	0.448
500	0.966	0.328	0.976	0.331	0.971	0.331	0.964	0.371	0.987	0.390
1000	0.955	0.223	0.965	0.224	0.960	0.224	0.976	0.255	0.986	0.258

$\kappa_1(0.5)=0.4 \;\; \kappa_2(0.5)=0.4 \;\; \theta=1$

$Se_1=0.76 \;\; Sp_1=0.72 \;\; Se_2=0.40 \;\; Sp_2=0.943 \;\; \varepsilon_1=0.0480 \;\; \varepsilon_0=0.0206 \;\; p=25\%$

	Wald *CI*		Logarit. *CI*		Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	0.979	1.627	0.999	1.835	0.990	5.762	0.977	2.244	0.999	3.650
50	0.953	1.525	0.991	1.708	0.977	3.028	0.981	2.173	0.995	2.728
100	0.941	1.350	0.983	1.467	0.962	2.342	0.956	1.703	0.984	2.051
200	0.953	0.972	0.971	1.014	0.955	1.212	0.960	1.091	0.979	1.251
300	0.950	0.770	0.953	0.790	0.944	0.851	0.941	0.825	0.965	0.931
400	0.955	0.658	0.969	0.670	0.960	0.705	0.959	0.694	0.980	0.776
500	0.951	0.582	0.954	0.590	0.947	0.612	0.943	0.607	0.965	0.678
1000	0.952	0.403	0.955	0.406	0.951	0.413	0.950	0.410	0.972	0.458

$\kappa_1(0.9)=0.4 \;\; \kappa_2(0.9)=0.4 \;\; \theta=1$

$Se_1=0.943 \;\; Sp_1=0.229 \;\; Se_2=0.70 \;\; Sp_2=0.70 \;\; \varepsilon_1=0.0200 \;\; \varepsilon_0=0.0343 \;\; p=50\%$

	Wald *CI*		Logarit. *CI*		Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	1	1.857	1	2.233	1	4.483	1	2.595	1	4.216
50	0.999	1.762	0.999	2.134	0.997	3.455	0.979	1.943	1	3.294
100	0.995	1.685	0.997	1.876	0.992	2.338	0.974	1.770	0.997	2.396
200	0.983	1.195	0.988	1.278	0.980	1.345	0.980	1.268	0.990	1.445
300	0.964	0.943	0.982	0.986	0.959	1.003	0.965	0.989	0.971	1.093
400	0.957	0.803	0.976	0.828	0.951	0.838	0.957	0.839	0.971	0.913
500	0.954	0.709	0.970	0.726	0.956	0.733	0.960	0.739	0.970	0.801
1000	0.956	0.491	0.964	0.496	0.956	0.499	0.959	0.505	0.969	0.545

Similar conclusions are obtained when the dependence factors take high values. Therefore, regarding the effect of the dependence factors on the *CIs*, in general terms they do not have a clear effect on the *CPs* of the *CIs*.

4.3. CIs with a small sample

The results of the simulation experiments have shown that the *CIs* may fail when the sample size is small $(n = 25-50)$. A classic solution to this problem is adding the correction 0.5 to each observed frequency, as is frequent in the analysis of 2×2 tables. To assess this procedure, the same simulation experiments as before were carried out for $n = \{25,50,100\}$ adding the value 0.5 to all of the observed frequencies s_{ij} and r_{ij}. Table 7 shows some of the results obtained for the *CIs* for the ratio θ. The results for the difference δ are not shown since, although this method improves the *CP* of the *CIs*, these intervals continue to fail when they failed without adding the correction. The results for $n = 100$ are not shown either, since these are very similar to those obtained without adding the correction.

Table 7. Coverage probabilities (*CPs*) and average lengths (*ALs*) of the *CIs* for θ with small samples.

$\kappa_1(0.9)=0.2\ \ \kappa_2(0.9)=0.8\ \ \theta=0.25$										
$Se_1=0.28\ \ Sp_1=0.92\ \ Se_2=0.82\ \ Sp_2=0.98\ \ \varepsilon_1=0.0252\ \ \varepsilon_0=0.00092\ \ p=10\%$										
	Wald *CI*		Logarit. *CI*		Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	0.999	1.808	0.008	1.960	0.653	3.014	0.145	2.150	0.783	3.531
50	0.940	1.287	0.262	1.464	0.768	1.710	0.556	1.440	0.768	1.813
$\kappa_1(0.5)=0.4\ \ \kappa_2(0.5)=0.8\ \ \theta=0.5$										
$Se_1=0.76\ \ Sp_1=0.72\ \ Se_2=0.85\ \ Sp_2=0.95\ \ \varepsilon_1=0.0570\ \ \varepsilon_0=0.0180\ \ p=25\%$										
	Wald *CI*		Logarit. *CI*		Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	1	1.458	0.961	1.659	0.984	2.332	0.940	1.897	1	3.118
50	0.992	0.836	0.960	0.913	0.982	0.932	0.962	0.869	0.997	1.141
$\kappa_1(0.9)=0.6\ \ \kappa_2(0.9)=0.8\ \ \theta=0.75$										
$Se_1=0.62\ \ Sp_1=0.98\ \ Se_2=0.911\ \ Sp_2=0.936\ \ \varepsilon_1=0.0277\ \ \varepsilon_0=0.0094\ \ p=5\%$										
	Wald *CI*		Logarit. *CI*		Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	1	1.812	1.000	2.073	1	3.554	1	2.425	1	4.053
50	1	1.593	1.000	1.789	1	2.564	0.999	2.067	1	2.682
$\kappa_1(0.9)=0.4\ \ \kappa_2(0.9)=0.4\ \ \theta=1$										
$Se_1=0.943\ \ Sp_1=0.229\ \ Se_2=0.70\ \ Sp_2=0.70\ \ \varepsilon_1=0.0200\ \ \varepsilon_0=0.0343\ \ p=50\%$										
	Wald *CI*		Logarit. *CI*		Fieller *CI*		Bootstrap *CI*		Bayesian *CI*	
n	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*	*CP*	*AL*
25	1	1.896	1	2.140	1	4.727	1	2.571	1	4.234
50	1	1.798	1	1.991	1	3.211	1	2.418	1	3.242

As conclusions, in general terms, it holds that: a) the Wald *CI* for θ does not fail, its *CP* is 100% or very close to 100%, and its *AL* is lower than the rest of the intervals when these do not fail; b) the logarithmic, Fieller, Bootstrap and Bayesian *CIs* may

continue to fail when $\theta = 0.25$. Consequently, when the sample size is small one must use the Wald *CI* for θ adding the value 0.5 to all of the observed frequencies.

4.4. Rules of application

The *CIs* for the difference and for the ratio of the two weighted kappa coefficients compare both parameters, and therefore we can decide which method is preferable to make this comparison. Once we have studied the coverage probabilities and the average lengths of the *CIs* for $\delta = \kappa_1(c) - \kappa_2(c)$ and for $\theta = \kappa_1(c)/\kappa_2(c)$, from the results obtained some general rules of application can be given for the *CIs* in terms of sample size. These rules are based on the failures and on the coverage probabilities, since the average lengths of the *CIs* for the difference and for the ratio cannot be compared as they are different intervals. In terms of sample size *n*:

a) If *n* is small $(n < 100)$, use the Wald *CI* for θ increasing the frequencies s_{ij} and r_{ij} in 0.5.

b) If $100 \leq n \leq 400$, use the Wald *CI* for the ratio θ without adding 0.5.

c) If $n \geq 500$, use any of the *CIs* (for the difference or for the ratio) proposed in Section 3.2 without adding 0.5.

In general terms, if the sample size is small, the Wald *CI* calculated adding 0.5 to each observed frequency does not fail. In this situation, its *AL* increases in relation to the Wald *CI* without adding 0.5, but its *CP* also increases meaning that the interval does not fail. When $100 \leq n \leq 400$ the *CI* that behaves best (fewest failures and its *CP* shows better fluctuations around 95%) is the Wald *CI* for the ratio θ. When the sample size is very large $(n \geq 500)$, there is no important difference between the asymptotic behaviour of the proposed *CIs*, and therefore any one of them can be used. When the sample size is small, $(n \leq 50)$ the *CIs* may fail, especially when the difference between the two weighted kappa coefficients is not small.

5. Sample size

The determination of the sample size to compare parameters of two *BDTs* is a topic of interest. We then propose a method to calculate the sample size to estimate the ratio θ between two weighted kappa coefficients with a precision ϕ and a confidence $100(1-\alpha)\%$. This method is based on the Wald *CI* for θ, which is, in general terms, the interval with the best asymptotic behaviour. Furthermore, this method requires a pilot sample (or another previous study) from which we calculate estimations of all of the parameters (Se_i, Sp_i, ε_i and *p*, and consequently of $\kappa_i(c)$) and the Wald *CI* for θ. If the pilot sample size is not small and the Wald *CI* for θ calculated from this sample contains the value 1, it makes no sense to determine the sample size necessary to estimate how much bigger one weighted kappa coefficient is than the other one, as the equality between both is not rejected. Nevertheless, if the pilot sample is small and the Wald *CI* (adding 0.5) contains the value 1, it may be useful to calculate the sample size to estimate the ratio θ. In this situation, the Wald *CI* (adding 0.5) will be very wide (as the pilot sample is small) and may contain the value 1 even if $\kappa_1(c)$ and $\kappa_2(c)$ are different. Let us considerer that $\kappa_2(c) \geq \kappa_1(c)$ and therefore $\theta \leq 1$, and let ϕ be the precision set by the researcher. As it has been assumed that $\theta \leq 1$, then ϕ must be lower than one, and if we want to have a high level of precision then ϕ must be a small value. On the other and, based on the asymptotic normality of $\hat{\theta} = \hat{\kappa}_1(c)/\hat{\kappa}_2(c)$ it is verified that $\hat{\theta} \in \theta \pm z_{1-\alpha/2}\sqrt{Var\left(\hat{\theta}\right)}$, i.e. the probability of obtaining an estimator $\hat{\theta}$ is in this interval with a probability $100(1-\alpha)\%$. Setting a precision ϕ, we can then calculate the sample size *n* from

$$\phi = z_{1-\alpha/2}\sqrt{Var\left(\hat{\theta}\right)}. \tag{21}$$

where

$$Var\left(\hat{\theta}\right) \approx \frac{\kappa_2^2(c)Var\left[\hat{\kappa}_1(c)\right] + \kappa_1^2(c)Var\left[\hat{\kappa}_2(c)\right] - 2\kappa_1(c)\kappa_2(c)Cov\left[\hat{\kappa}_1(c), \hat{\kappa}_2(c)\right]}{\kappa_2^4(c)}.$$

In the Appendix *B* of the supplementary material, we can see how this expression is obtained. This variance depends on the weighted kappa coefficients and on their

respective variances and covariance. Furthermore, the variances $Var\left[\hat{\kappa}_i(c)\right]$ and the covariance $Cov\left[\hat{\kappa}_1(c),\hat{\kappa}_2(c)\right]$ (their expressions can be seen in the Appendix B of the supplementary material) depend, among other parameters, on the sample size *n*. Consequently, it is possible to use this relation to calculate the sample size to estimate the ratio θ. Substituting in the equation of $Var\left(\hat{\theta}\right)$ the variances and the covariance with its respective expressions, substituting the parameters with their estimators and clearing *n* in equation (21), it is obtained that

$$n = \frac{z_{1-\alpha/2}^2 \hat{\theta}^2}{\phi^2 \hat{p}^3 \hat{q}^3} \times \left\{ \sum_{h=1}^{2} \left[\frac{\hat{a}_{h1}^2 \hat{Se}_h \left(1-\hat{Se}_h\right)\hat{q} + \hat{a}_{h2}^2 \hat{Sp}_h \left(1-\hat{Sp}_h\right)\hat{p} + \hat{a}_{h3}^2 \hat{p}^2 \hat{q}^2}{\hat{Y}_h^2} \right] - \frac{2}{\hat{Y}_1 \hat{Y}_2} \left[\hat{a}_{11}\hat{a}_{21}\hat{\varepsilon}_1 \hat{q} + \hat{a}_{12}\hat{a}_{22}\hat{\varepsilon}_0 \hat{p} + \hat{a}_{13}\hat{a}_{23}\hat{p}^2\hat{q}^2 \right] \right\}, \tag{22}$$

where $\hat{a}_{h1} = \hat{p}\hat{q} - \hat{p}\left(\hat{q}-c\right)\hat{\kappa}_h(c)$, $\hat{a}_{h2} = \hat{a}_{h1} + \left(\hat{q}-c\right)\hat{\kappa}_h(c)$ and $\hat{a}_{h3} = \left(1-2\hat{p}\right)\hat{Y}_h - \left[\left(1-c-2\hat{p}\right)\hat{Y}_h + \hat{Sp}_h + c - 1\right]\hat{\kappa}_h(c)$. This method requires us to know $\hat{Se}_h$, $\hat{Sp}_h$, $\hat{\varepsilon}_i$ and $\hat{p}$ (and therefore $\hat{\kappa}_h(c)$), for example obtained from a pilot sample or from previous studies. The procedure to calculate the sample size consists of the following steps:

1) Take pilot samples sized n' (in general terms, $n' \geq 100$ to be able to calculate the Wald *CI* without adding 0.5 or use the Wald *CI* adding 0.5 to the frequencies if *n* is small), and from this sample calculate $\hat{Se}_h$, $\hat{Sp}_h$, $\hat{\varepsilon}_i$, $\hat{p}$ and $\hat{\kappa}_h(c)$, and a then calculate the Wald *CI* for θ. If the Wald *CI* calculated has a precision ϕ, i.e. if $\frac{\text{Upper limit} - \text{Lower limit}}{2} \leq \phi$, then with the pilot sample the precision has been reached and the process has finished (θ has been estimated with a precision ϕ to a confidence $100(1-\alpha)\%$); if this is not the case, go to the following step.

2) From the estimations obtained in Step 1, calculate the new sample size n applying equation (22).

3) Take the sample of n individuals ($n-n'$ is added to the pilot sample), and from the new sample we calculate $\hat{Se}_h$, $\hat{Sp}_h$, $\hat{\varepsilon}_i$, $\hat{p}$, $\hat{\kappa}_h(c)$ and the Wald *CI* for θ. If the Wald *CI* calculated has a precision ϕ, then with the new sample the precision has been reached and the process has finished. If the Wald *CI* does not have the required precision, then this new sample is considered as a pilot sample and the process starts again at step 1. In this situation, the new sample has a size n calculated in step 2, i.e. we add $n-n'$ individuals to the initial pilot sample (sized n'). Therefore, the process starts again at step 1 considering the new sample as the pilot sample and from this sample we calculate the values of the estimators and the Wald *CI*.

The method to calculate the sample size is an iterative method which depends on the pilot sample and which does not guarantee that θ will be estimated with the required precision. Each time that the previous process (steps 1-3) is repeated, we calculate (starting from an initial sample) the new sample size to estimate θ, i.e. we calculate the number of individuals that must be added to the initial sample to obtain a new sample. Therefore, this process adjusts the size of the initial pilot sample, adding (in each iteration of the process: steps 1-3) the number of individuals necessary to obtain the right sample size to estimate θ with the precision required. The programme in *R* described in the Section 6 allows us to calculate the sample size to estimate θ.

If the Wald *CI* for θ is higher than one, the *BDTs* can always be permuted and θ will then be lower than one. Another alternative consists of setting a value for a precision ϕ', in a similar way to the previous situation when $\theta \leq 1$, and then apply the equation (22) with $\phi = \hat{\theta}^2\phi'$, where $\hat{\theta} = \hat{\kappa}_1(c)/\hat{\kappa}_2(c) \leq 1$. This is due to the fact that if (L_θ, U_θ) is the Wald *CI* for $\theta = \kappa_1(c)/\kappa_2(c) \leq 1$ then the Wald *CI* for $\theta' = 1/\theta = \kappa_2(c)/\kappa_1(c)$ is $\left(L_\theta/\hat{\theta}^2 \ , \ U_\theta/\hat{\theta}^2\right)$. It is easy to check that the calculated value of the sample size n is the same both if $\theta \leq 1$ (with precision ϕ) and if $\theta > 1$ (with precision $\phi = \hat{\theta}^2\phi'$).

Simulation experiments were carried out to study the effect that the pilot sample has on the calculation of the sample size. These experiments consisted of generating $N = 10,000$ random samples of multinomial distributions considering the same scenarios as those given in Tables 5 and 6. The equation of the sample size depends on

the values of the estimators, which in turn depend on the pilot sample. Consequently, the pilot sample may have an effect on the sample size calculated. To study this effect, the simulation experiments consisted of the following steps:

1) Calculate the sample size *n* from the values of the parameters set in the different scenarios considered. Therefore, equation (22) was applied using the values of the parameters (instead of their estimators).

2) Generate the *N* multinomial random samples sized *n* calculating the probabilities from equations (6) and (7), using the values of the previous parameters, and as ε_i we considered low values (25%), intermediate values (50%) and high values (80%). From each one of the *N* random samples, $\hat{Se}_h$, $\hat{Sp}_h$, $\hat{\varepsilon}_i$ and $\hat{p}$ (and therefore $\hat{\kappa}_h(c)$) were calculated, and then we calculated the sample size n_i' applying equation (22).

3) For each scenario, the average sample size and the relative bias were calculated, i.e. $\bar{n} = \sum n_i' / N$ and $RB(n') = (\bar{n} - n)/n$.

Table 7. Effect of the pilot sample on the sample size.

$\kappa_1(0.1) = 0.2$ $\kappa_2(0.1) = 0.8$ $\theta = 0.25$						
$Se_1 = 0.484$ $Sp_1 = 0.684$ $Se_2 = 0.852$ $Sp_2 = 0.911$ $p = 50\%$						
	$\varepsilon_1 = 0.0179$ $\varepsilon_0 = 0.0153$		$\varepsilon_1 = 0.0359$ $\varepsilon_0 = 0.0306$		$\varepsilon_1 = 0.0574$ $\varepsilon_0 = 0.0489$	
	$\phi = 0.05$	$\phi = 0.10$	$\phi = 0.05$	$\phi = 0.10$	$\phi = 0.05$	$\phi = 0.10$
Sample size	3170	793	3066	767	2942	736
Average sample size	3173	795	3068	769	2946	738
Relative bias (%)	0.095	0.252	0.065	0.261	0.136	0.272
$\kappa_1(0.9) = 0.2$ $\kappa_2(0.9) = 0.8$ $\theta = 0.25$						
$Se_1 = 0.28$ $Sp_1 = 0.92$ $Se_2 = 0.82$ $Sp_2 = 0.98$ $p = 10\%$						
	$\varepsilon_1 = 0.0126$ $\varepsilon_0 = 0.0046$		$\varepsilon_1 = 0.0252$ $\varepsilon_0 = 0.0092$		$\varepsilon_1 = 0.0403$ $\varepsilon_0 = 0.0147$	
	$\phi = 0.05$	$\phi = 0.10$	$\phi = 0.05$	$\phi = 0.10$	$\phi = 0.05$	$\phi = 0.10$
Sample size	5104	1276	4947	1237	4758	1190
Average sample size	5113	1287	4948	1246	4759	1218
Relative bias (%)	0.18	0.83	0.02	0.73	0.02	2.35

Table 7 (Effect of the pilot sample) shows some of the results obtained. The relative biases are very small, which indicates that the equation of the calculation of the sample size provides robust values, and therefore the choice of the pilot sample does not have an important effect on the calculation of the sample size.

6. Programme *citwkc*

A programme has been written in *R* and called citwkc (Confidence Intervals for Two Weighted Kappa Coefficients) which allows us to calculate the *CIs* proposed in Section 3 and the sample size proposed in Section 5. The programme runs with the command

$$\text{“citwkc}\left(s_{11}, s_{10}, s_{01}, s_{00}, r_{11}, r_{10}, r_{01}, r_{00}, \textit{cindex}, \textit{preci} = 0, \textit{conf} = 0.95\right)\text{”},$$

where *cindex* is the weighting index, *preci* is the precision that is needed to calculate the sample size and *conf* is the level of confidence (by default 95%). By default $\textit{preci} = 0$, and the programme does not calculate the sample size, and only calculates it when $\textit{preci} > 0$. In this situation $\left(\textit{preci} > 0\right)$, the programme checks if it is necessary to calculate the sample size. The programme checks that the values of the frequencies and of the parameters are viable (e.g. that there are no negative values, frequencies with decimals, etc,) and also checks that it is possible to estimate all of the parameters and their variances-covariances. For the intervals obtained applying the Bootstrap method, 2,000 samples with replacement are generated, and for the Bayesian intervals 10,000 random samples are generated. The results obtained on running the programme are saved in file called “Results_citwkc.txt” in the same folders from where the programme is run. The program is available for free at URL

“https://www.ugr.es/~bioest/software/cmd.php?seccion=mdb”.

7. Application

The results obtained have been applied to the study by Batwala et al (2010) on the diagnosis of malaria. Batwala et al have applied the *Expert Microscopy Test* and the *HRP2-Based Rapid Diagnostic Test* to a sample of 300 individuals using the *PCR* as the *GS*. The observed frequencies of this study are shown in Table 9, where the T_1 models the result of the *Expert Microscopy Test*, T_2 models the result of the *HRP2-Based Rapid Diagnostic Test* and *D* models the result of the *PCR*. In this example, $\hat{Se}_1 = 46.07\%$, $\hat{Sp}_1 = 97.16\%$, $\hat{Se}_2 = 91.01\%$ and $\hat{Sp}_2 = 86.26\%$, and therefore $\widehat{rTPF_{12}} = 0.506$ and $\widehat{rFPF_{12}} = 0.207$. Applying the equation (5) it holds that $c' = 0.1902$. As $\widehat{rTPF_{12}} < 1$ and $\widehat{rFPF_{12}} < 1$, applying the rule c) given in Section 2, it holds that $\hat{\kappa}_1(c) > \hat{\kappa}_2(c)$ for

$0 \le c < 0.1902$ and that $\hat{\kappa}_1(c) < \hat{\kappa}_2(c)$ for $0.1902 < c \le 1$. Applying the rules given in Section 4, as $n = 300 < 400$ then it is necessary to use the Wald *CI* for the ratio θ. Table 10 shows the values of $\hat{\kappa}_h(c)$, $\hat{\delta}$, $\hat{\theta}$ and the 95% *CIs* when $c = \{0.1, 0.1902, 0.2, ..., 0.8, 0.9\}$. The results were obtained running the programme "*citwkc*" with the command $\text{citwkc}(41, 0, 40, 8, 5, 1, 24, 181, c)$ taking $c = \{0.1, 0.1902, 0.2, ..., 0.8, 0.9\}$.

For $c = \{0.1, 0.1902, 0.2, 0.3\}$, the Wald *CI* for θ contains the value 1, and therefore in these cases we do not reject the equality of the weighted kappa coefficients of the *Expert Microscopy Test* and of the *HRP2-Based Rapid Diagnostic Test*. Therefore, when the clinician considers that a false positive is 9, 4 or 2.33 times more important than a false negative, we do not reject the equality between the weighted kappa coefficients of the *Expert Microscopy Test* and of the *HRP2-Based Rapid Diagnostic Test* in the population studied. The rest of the intervals for θ also contain the value 1.

For $c = \{0.4, 0.5, ..., 0.8, 0.9\}$, the Wald *CI* θ does not contain the value 1, and therefore in all of these cases we do not reject the equality of the weighted kappa coefficients of the *Expert Microscopy Test* and of the *HRP2-Based Rapid Diagnostic Test* in the population studied. Therefore, the clinician considers that a false negative is more important than a false positive (as happens in the situation in which the diagnostic tests are applied as screening tests), the weighted kappa coefficient of the *HRP2-Based Rapid Diagnostic Test* is significantly greater than the weighted kappa coefficient of the *Expert Microscopy Test* in the population studied. The same conclusion is obtained when the clinician considers that a false positive and a false negative have the same importance $(c = 0.5)$. If the clinician considers that a false positive is 1.5 times greater than a false negative (i.e. $c = 0.4$), then the same conclusion is obtained. The rest of the *CIs* for θ do not contain the value 1. For example, considering $c = 0.9$, interpreting the Wald *CI* for the ratio, it is concluded that in the population being studied the between the *HRP2-Based Rapid Diagnostic Test* and the *PCR* is, with a confidence of 95%, a value between 1.72 $(1/0.58 \approx 1.72)$ and 2.94 $(1/0.34 \approx 2.94)$ times greater than the agreement beyond chance between the *Expert Microscopy Test* and the *PCR*.

In order to illustrate the method to calculate the sample size presented in Section 5 we will consider that $c = 0.9$, and therefore that the two *BDTs* are applied as a screening test. In this situation, the 95% Wald *CI* for θ is $(0.34\ ,\ 0.58)$, and the precision is 0.12. As an example, we will consider that the clinician wishes to estimate the ratio between the two weighted kappa coefficients with a precision $\phi = 0.10$. As with the sample of 300 individuals the desired precision $(\phi = 0.10 < 0.12)$ was not achieved, then using this sample as a pilot sample and running the programme *citwkc* with the command $\text{citwkc}(41,0,40,8,5,1,24,181,0.9,0.1)$ it holds that $n = 435$. Therefore, to the sample pilot of 300 individuals we must add 135 more. Once the new sample has been taken, it is necessary to check that the precision $\phi = 0.10$ is verified.

Table 9. Study of Batwala et al and results.

	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$D = 1$	41	0	40	8	89
$D = 0$	5	1	24	181	211
Total	46	1	64	189	300

Table 10. *CIs* for the ratio $\theta = \kappa_1(c)/\kappa_2(c)$.

c	$\hat{\kappa}_1(c)$	$\hat{\kappa}_2(c)$	$\hat{\theta}$	Wald	Logarithmic	Fieller	Bias-corrected	Bayesian
0.1	0.726	0.642	1.131	0.925 , 1.335	0.943 , 1.355	0.940 , 1.357	0.926 , 1.344	0.883 , 1.393
0.1902	0.659	0.659	1	0.811 , 1.189	0.828 , 1.208	0.823 , 1.206	0.817 , 1.204	0.776 , 1.234
0.2	0.653	0.661	0.988	0.800 , 1.174	0.817 , 1.194	0.812 , 1.192	0.808 , 1.192	0.766 , 1.219
0.3	0.593	0.681	0.871	0.695 , 1.046	0.711 , 1.065	0.704 , 1.059	0.701 , 1.065	0.673 , 1.083
0.4	0.543	0.701	0.775	0.609 , 0.939	0.625 , 0.958	0.615 , 0.948	0.615 , 0.952	0.593 , 0.971
0.5	0.501	0.723	0.693	0.537 , 0.847	0.553 , 0.866	0.541 , 0.854	0.541 , 0.857	0.525 , 0.877
0.6	0.464	0.747	0.621	0.476 , 0.768	0.492 , 0.786	0.479 , 0.772	0.481 , 0.776	0.468 , 0.799
0.7	0.433	0.772	0.561	0.425 , 0.698	0.440 , 0.716	0.426 , 0.701	0.430 , 0.707	0.418 , 0.727
0.8	0.406	0.799	0.508	0.380 , 0.637	0.395 , 0.654	0.381 , 0.639	0.384 , 0.644	0.375 , 0.667
0.9	0.382	0.827	0.462	0.341, 0.582	0.356 , 0.599	0.342 , 0.584	0.347 , 0.594	0.339 , 0.611

8. Discussion

The weighted kappa coefficient of a *BDT* is a measure of the beyond-chance agreement between the *BDT* and the *GS*, and depends on the sensitivity and specificity of the *BDT*, on the disease prevalence and on the weighting index. The weighted kappa coefficient is a parameter that is used to assess and compare the performance of *BDTs*. In this article,

we have studied the comparison of the weighted kappa coefficients of two *BDTs* through confidence intervals when the sample design is paired.

Three intervals have been studied for the difference of the two weighted kappa coefficients and five more intervals for the ratio of the two parameters. All the intervals studied are asymptotic and simulation experiments have been carried out to study their coverage probabilities and average lengths subject to different scenarios and for different sample sizes. Based on the results of the simulation experiments, some general rules of application have been given. When the sample size is moderate $(n=100)$ or large $(n=200-400)$ it is preferable to compare the two weighted kappa coefficients through an interval for the ratio, and when the sample size is very large $(n\geq 500)$ the two weighted kappa coefficients can be compared through the difference or the ratio. When the sample size is small $(n\leq 50)$, the interval with the best behaviour is the Wald *CI* for the ratio θ adding 0.5 to all of the observed frequencies. Adding 0.5 to all of the frequencies does not improve the behaviour of the intervals for the difference δ, since these continue to fail when they failed without adding the value 0.5. This question may be due to the fact that the ratio $\hat{\theta}$ converges more quickly to the normal distribution than the difference $\hat{\delta}$. In the simulation experiments, the asymptotic behaviour of the Bayesian *CIs* has been studied using the $Beta(1,1)$ distribution as prior distribution for all of the parameters. The choice of the values of the hyperparameters of the *Beta* distribution will depend on the previous information that the researcher has. If the researcher has some information and wants this information to have some weight in the data, then it is possible to use higher values of α and β, i.e. considering a $Beta(\alpha,\beta)$ distribution with $\alpha,\beta>1$. The increase in α and β adds information and decreases the variance and, therefore, there is less uncertainty about the parameter. If the researcher does not want this information to have a great weight in the posteriori distribution, then the researcher chooses moderate values of α and β which are consistent with the information available, i.e. the average should be compatible with that information. To assess the effect that the *Beta* distribution has on the asymptotic behaviour of the Bayesian interval, we have carried out simulations (in a similar way to those carried out in Section 4) using as prior the distributions $Beta(5,5)$ and $Beta(25,25)$ for the Bayesian interval for $\theta=\kappa_1(c)/\kappa_2(c)$. These two distributions have the same average

as the $Beta(1,1)$ distribution but different variances. The first distribution has a moderate weight in the subsequent distribution, the second has an important weight and the third one has a very important weight. In general terms, the results obtained with the distribution $Beta(5,5)$ are very similar to those obtained with the $Beta(1,1)$ distribution. Regarding the $Beta(25,25)$ distribution, there is no important difference in relation to the *CPs* obtained with the $Beta(1,1)$, although for $\theta = \{0.25, 0.50\}$ the *AL* is slightly lower with the $Beta(25,25)$, and when $\theta = \{0.75,1\}$ the *AL* is slightly higher with the $Beta(25,25)$. In general terms, when the Bayesian interval fails using the $Beta(1,1)$ distribution then it also fails using the $Beta(5,5)$ and the $Beta(25,25)$. Furthermore, the Bayesian *CI* for $\theta = \kappa_1(c)/\kappa_2(c)$ with the $Beta(5,5)$ and $Beta(25,25)$, respectively, does not display a better *CP* than the Wald *CI* (when it does not fail), and therefore the Bayesian *CI* does not improve the asymptotic behaviour of the Wald *CI*.

The application of the *CIs* requires the marginal frequencies *s* and *r* to be higher than zero. If the marginal frequency *s* (or *r*) is equal to zero, then it is not possible to estimate the weighted kappa coefficient of each *BDT*. Moreover, if a marginal frequency $s_{ij} + r_{ij}$ is equal to zero, then it is possible to calculate all of the *CIs* proposed; but not if two of these marginal frequencies are equal to zero. In this last situation, one of the weighted kappa coefficients (or both) is equal to zero, and the variance and the covariance are also equal to zero. If $s_{10} + r_{10} = s_{01} + r_{01} = 0$ then $\hat{\kappa}_1(c) = \hat{\kappa}_2(c)$ and $\hat{Var}(\hat{\kappa}_1(c)) = \hat{Var}(\hat{\kappa}_2(c)) = Cov(\hat{\kappa}_1(c), \hat{\kappa}_2(c))$, and the frequentist intervals cannot be calculated. A solution to this problem is to add 0.5 to each observed frequency.

In this article, we have also proposed a method to calculate the sample size to estimate the ratio between the two weighted kappa coefficients with a determined precision and confidence. This method, based on the Wald *CI* for the ratio, is an iterative method, which starting from a pilot sample adds individuals to the sample until the *CI* has the set precision. From the initial sample we estimate a vector of parameters and in the second stage we calculate the sample size. Furthermore, the simulation experiments carried out to study the robustness of the method to calculate the sample

size have shown that the method has practical validity and the choice of the pilot sample has very little effect on this method.

When the two diagnostic tests are continuous, for each cut off point of each estimated ROC curve there will be a value of $\hat{Se}_h$ and of $\widehat{FPF}_h$ (and therefore of $\hat{Sp}_h = 1 - \widehat{FPF}_h$), with $h = 1, 2$. Once the clinician has set the value of the weighting index, $\hat{\kappa}_1(c)$ and $\hat{\kappa}_2(c)$ are calculated and therefore the confidence intervals studied in Section 3 can be applied.

Supplementary material: Appendices *A*, *B* and *C*

Appendices *A*, *B* and *C* are available as supplementary material of the manuscript in the URL: https://www.ugr.es/~bioest/software/cmd.php?seccion=mdb.

Acknowledgements

This research was supported by the Spanish Ministry of Economy, Grant Number MTM2016-76938-P. We thank the referee, the Associate Editor and the Editor (M. Isabel Fraga Alves) of REVSTAT Statistical Journal for their helpful comments that improved the quality of the paper.

References

Agresti, A. (2002). *Categorical data analysis*. Wiley, New York.

Batwala, V., Magnussen, P., Nuwaba, F. (2010). Are rapid diagnostic tests more accurate in diagnosis of plasmodium falciparum malaria compared to microscopy at rural health centers? *Malaria Journal* 9: 349.

Bloch, D.A. (1997). Comparing two diagnostic tests against the same ‘gold standard’ in the same sample. *Biometrics* 53: 73-85.

Cicchetti, D.V. (2001). The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology* 23: 695-700.

Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Fieller, E.C. (1940). The biological standardization of insulin. *Journal of the Royal Statistical Society* 7 Supplement: 1-64.

Kraemer, H.C., Bloch, D.A. (1990). A Note on case-control sampling to estimate kappa coefficients. *Biometrics* 46: 49-59.

Kraemer, H.C. (1992). *Evaluating medical tests. Objective and quantitative guidelines*. Sage Publications, Newbury Park.

Kraemer, H.C., Periyakoil, V.S., Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine* 2: 2109-2129.

Martín-Andrés, A., Álvarez-Hernández, M. (2014a). Two-tailed asymptotic inferences for a proportion. *Journal of Applied Statistics* 41: 1516-1529.

Martín-Andrés A, Álvarez-Hernández M. (2014b) Two-tailed approximate confidence intervals for the ratio of proportions. *Statistics and Computing* 24: 65–75.

Montero Alonso, M.A., Roldán-Nofuentes, J.A. (2019). Approximate confidence intervals for the likelihood ratios of a binary diagnostic test in the presence of partial disease verification. *Journal of Biopharmaceutical Statistics* 29: 56-81.

Price, R.M., Bonett, D.G. (2004). An improved confidence interval for a linear function of binomial proportions. *Computational Statistics & Data Analysis* 45: 449-456.

Roldán Nofuentes, J.A., Luna del Castillo, J.D., Montero Alonso, M.A. (2009). Confidence intervals of weighted kappa coefficient of a binary diagnostic test. *Communications in Statistics - Simulation and Computation* 38: 1562-1578.

Roldán-Nofuentes, J.A., Amro, R. (2018). Combination of the weighted kappa coefficients of two binary diagnostic tests. *Journal of Biopharmaceutical Statistics* 28: 909-926.

Vacek, P.M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics, 41, 959-968.

Supplementary material of the manuscript:

Asymptotic confidence intervals for the difference and the ratio of the weighted kappa coefficients of two diagnostic tests subject to a paired design

Appendix *A*

From now onwards, we are going to suppose that $0 < Se_h < 1$, $0 < Sp_h < 1$, $0 < p < 1$ and $q = 1 - p$. Performing algebraic operations it is verified that

$$\kappa_1(c) - \kappa_2(c) = \frac{pq}{D_1 D_2}\nu \tag{1}$$

where $D_h = p(1-Q_h)c + qQ_h(1-c)$ is the denominator of $\kappa_h(c)$, with $h = 1, 2$, and

$$\nu = q\Delta_1 - c(\Delta_1 - p\Delta_2) \tag{2}$$

where $\Delta_1 = Se_1(1-Sp_2) - Se_2(1-Sp_1)$ and $\Delta_2 = Y_1 - Y_2 = Se_1 - Se_2 + Sp_1 - Sp_2$. Then $\kappa_1(c) > \kappa_2(c)$ if $\nu > 0$, since $D_h > 0$. Solving equation $\kappa_1(c) - \kappa_2(c) = 0$ in c it holds that

$$c' = c = \frac{q\Delta_1}{\Delta_1 - p\Delta_2}, \tag{3}$$

being c' a real value. From now onwards, the rules so that $\kappa_1(c) > \kappa_2(c)$, $\kappa_2(c) > \kappa_1(c)$ and $\kappa_1(c) = \kappa_2(c)$, considering that $i = 1$ and $j = 2$ (the demonstrations for $i = 2$ and $j = 1$ are analogous).

a) If $rTPF_{12} \geq 1$ and $rFPF_{12} < 1$, or $rTPF_{12} > 1$ and $rFPF_{12} \leq 1$, then $\kappa_1(c) > \kappa_2(c)$ for $0 \leq c \leq 1$.

Let us suppose in the first place that $rTPF_{12}=1$ and that $rFPF_{12}<1$, then $Se_1=Se_2=Se$ and $Sp_1>Sp_2$. Substituting in equation (2) it holds that $\nu=(Sp_1-Sp_2)\left[cp+(q-c)Se\right]$. Here $\nu>0$ if $cp+(q-c)Se>0$, since $Sp_1>Sp_2$. If $c=0$ or $c=1$, then $cp+(q-c)Se>0$ since $qSe>0$, and $p(1-Se)>0$ is verified; and as $Sp_1>Sp_2$, then $\nu>0$ and $\kappa_1(c)>\kappa_2(c)$. Let us suppose that $0<c<1$ and $p\geq Se$, then $cp+(q-c)Se=c(p-Se)+qSe>0$, and it is verified that $\nu>0$ and $\kappa_1(c)>\kappa_2(c)$. If $p<Se$, then $cp+(q-c)Se=(1-c)(Se-p)+(1-Se)p>0$, since $(1-c)(Se-p)>0$ and $(1-Se)p>0$. Therefore, $\nu>0$ and $\kappa_1(c)>\kappa_2(c)$.

Let us now suppose that $rTPF_{12}>1$ and that $rFPF_{12}<1$, then $Se_1>Se_2$ and $Sp_1>Sp_2$. It is easy to check that when $c=0$ or $c=1$ it is verified that $\nu>0$ and, therefore, $\kappa_1(c)>\kappa_2(c)$. Moreover, as $rTPF_{12}>1$ and $rFPF_{12}<1$ then dividing both parameters $(rTPF_{12}/rFPF_{12}>1)$ it holds that $\dfrac{rTPF_{12}}{rFPF_{12}}=\dfrac{Se_1(1-Sp_2)}{Se_2(1-Sp_1)}>1$, verifying that $\Delta_1=Se_1(1-Sp_2)-Se_2(1-Sp_1)>0$. As $Se_1>Se_2$ and $Sp_1>Sp_2$ then $\Delta_2=Se_1-Se_2+Sp_1-Sp_2>0$. Furthermore, as it is verified that $Se_1>Se_2$ then $1-Se_1<1-Se_2$, and $0<\dfrac{1-Se_1}{1-Se_2}<1$. Moreover, as $\dfrac{Sp_1}{Sp_2}>1$ then $\dfrac{Sp_1}{Sp_2}-\dfrac{1-Se_1}{1-Se_2}=\dfrac{\Delta_3}{Sp_2(1-Se_2)}>0$, when $\Delta_3=(1-Se_2)Sp_1-(1-Se_1)Sp_2>0$. It is easy to check that $\Delta_1=\Delta_2-\Delta_3$, so that $\Delta_2>\Delta_1$. Equation (2) can be written as

$$\nu=(q-c)\Delta_1+cp\Delta_2. \tag{4}$$

Let us suppose that $0<c<1$, then if $q\geq c$ it is verified that $\nu>0$ and $\kappa_1(c)>\kappa_2(c)$. Let us now suppose that $q<c$, then $q-c<0$. Equation (4) can be written as

$$\nu=-(c-q)\Delta_1+cp\Delta_2$$

being $c-q>0$. Let us suppose that

$$\nu<0\Rightarrow-(c-q)\Delta_1+cp\Delta_2<0,$$

so that

$$-(c-q)\Delta_1 < -cp\Delta_2 \Rightarrow (c-q)\Delta_1 > cp\Delta_2 \Rightarrow c-q > cp\frac{\Delta_2}{\Delta_1}.$$

As $\Delta_2 > \Delta_1$ then $\frac{\Delta_2}{\Delta_1} > 1$, so that

$$c-q > cp\frac{\Delta_2}{\Delta_1} > cp > 0,$$

from where we obtain

$$c-q-cp > 0. \tag{5}$$

Performing algebraic operations

$$c-q-cp = q(c-1)$$

As $0 < c < 1$, $1-c > 0$ and $c-1 < 0$, then $q(c-1) < 0$, which is contradictory with expression (5). Therefore, if $q < c$ then $\nu > 0$ and $\kappa_1(c) > \kappa_2(c)$.

The demonstrations for $rTPF_{12} > 1$ and $rFPF_{12} \leq 1$ are performed following a similar process to the previous one.

b). If $rTPF_{12} > 1$ and $rFPF_{12} > 1$, then:

b.1) $\kappa_1(c) > \kappa_2(c)$ if $0 < c' < c \leq 1$

b.2) $\kappa_1(c) < \kappa_2(c)$ if $0 \leq c < c' < 1$

b.3) $\kappa_1(c) = \kappa_2(c)$ if $c = c'$, with $0 < c' < 1$

b.4) $\kappa_1(c) > \kappa_2(c)$ for $0 \leq c \leq 1$ if $c' < 0$ (or $c' > 1$) and $rTPF_{12} > rFPF_{12} > 1$

b.5) $\kappa_1(c) < \kappa_2(c)$ for $0 \leq c \leq 1$ if $c' < 0$ (or $c' > 1$) and $rFPF_{12} > rTPF_{12} > 1$

Firstly, we are going to demonstrate that c' cannot be equal to 0 or to 1. As $rTPF > 1$ and $rFPF > 1$, then it is verified that $Se_1 > Se_2$ and $Sp_1 < Sp_2$. If $c' = 0$ then $\Delta_1 = 0$, and it is verified that

$$\frac{Se_1}{Se_2}\times\frac{1-Sp_2}{1-Sp_1}=1,$$

which is incompatible with $rTPF>1$ and $rFPF>1$, since as $\frac{Se_1}{Se_2}>1$ and $0<\frac{1-Sp_2}{1-Sp_1}<1$ then it is verified that $\frac{Se_1}{Se_2}\times\frac{1-Sp_2}{1-Sp_1}\neq 1$. Therefore c' cannot be equal to 0 if $rTPF>1$ and $rFPF>1$. If $c'=1$ then $\Delta_1-\Delta_2=Sp_2\left(1-Se_1\right)-Sp_1\left(1-Se_2\right)=0$, and it is verified that

$$\frac{Sp_2}{Sp_1}\times\frac{1-Se_1}{1-Se_2}=1,$$

which is incompatible with $rTPF>1$ and $rFPF>1$, since as $\frac{Sp_2}{Sp_1}>1$ and $0<\frac{1-Se_1}{1-Se_2}<1$ then it is verified that $\frac{Sp_2}{Sp_1}\times\frac{1-Se_1}{1-Se_2}\neq 1$. Therefore, c' cannot be equal to 1 if $rTPF>1$ and $rFPF>1$.

Let us consider that $0<c'<1$, then we must verify one of the two following: 1) $0<q\Delta_1<\Delta_1-p\Delta_2$, or 2) $\Delta_1-p\Delta_2<q\Delta_1<0$. Condition 1 implies that $\Delta_1>0$ and $\Delta_1>p\Delta_2$, and Condition 2 implies that $\Delta_1<0$ and $\Delta_1<p\Delta_2$.

Moreover, as $Se_1>Se_2$ and $Sp_1<Sp_2$ (which implies that $1-Sp_1>1-Sp_2$) then $Q_1>Q_2$. Furthermore, if $c=c'$ then performing algebraic operations, each weighted kappa coefficient is expressed as

$$\kappa_h\left(c'\right)=\frac{Y_h}{\tau_h},$$

when $\tau_h=\frac{\Delta_1-Q_h\Delta_2}{\Delta_1-p\Delta_2}$, with $h=1,2$. As $Q_1>Q_2$, then $\tau_2-\tau_1>0$ if $\Delta_2>0$, and $\tau_2-\tau_1<0$ if $\Delta_2<0$. If $\Delta_2>0$, then

$$\tau_2-\tau_1=\frac{\Delta_2\left(Q_1-Q_2\right)}{\Delta_1-p\Delta_2}>0\Rightarrow\Delta_1-p\Delta_2>0\Rightarrow\Delta_1>p\Delta_2>0.$$

If $\Delta_2<0$, then

$$\tau_2 - \tau_1 = \frac{\Delta_2 (Q_1 - Q_2)}{\Delta_1 - p\Delta_2} < 0 \Rightarrow \Delta_1 - p\Delta_2 > 0 \Rightarrow \Delta_1 > p\Delta_2 .$$

Therefore, whether $\Delta_2 > 0$ or $\Delta_2 < 0$, it is always verified that $\Delta_1 > p\Delta_2$. This condition is only compatible with Condition 1 obtained by the fact that $0 < c' < 1$, i.e. $0 < q\Delta_1 < \Delta_1 - p\Delta_2$. Therefore, it is always verified that $\Delta_1 > 0$ and $\Delta_1 > p\Delta_2$.

Moreover, from equation (3) it holds that $q\Delta_1 = c'(\Delta_1 - p\Delta_2)$, so that substituting this expression in equation (2) it holds that

$$\nu = (\Delta_1 - p\Delta_2)(c' - c). \tag{6}$$

As $\Delta_1 > p\Delta_2$ then $\Delta_1 - p\Delta_2 > 0$. Based on equation (6), if $0 \le c < c' < 1$ then $\nu > 0$ and $\kappa_1(c) > \kappa_2(c)$. If $0 < c' < c \le 1$ then $\nu < 0$ and $\kappa_1(c) < \kappa_2(c)$, and if $c = c'$ (with $0 < c' < 1$) then $\nu = 0$ and $\kappa_1(c) = \kappa_2(c)$.

If $c' < 0$ then one of the following two conditions must be verified: 1) $0 < q\Delta_1 < \Delta_1 < p\Delta_2 < \Delta_2$, or 2) $\Delta_2 < p\Delta_2 < \Delta_1 < q\Delta_1 < 0$. Condition 1 implies that $\Delta_1 > 0$ and therefore $Se_1(1 - Sp_2) > Se_2(1 - Sp_1)$, and from this inequality it holds that

$$\frac{Se_1}{Se_2} > \frac{1 - Sp_1}{1 - Sp_2} > 1 \Rightarrow rTPF_{12} > rFPF_{12} > 1.$$

As $q\Delta_1 > 0$ and $\Delta_1 - p\Delta_2 < 0$, then applying equation (2) it holds that $\nu > 0$ and therefore $\kappa_1(c) > \kappa_2(c)$. Condition 2 implies that $\Delta_1 < 0$ and therefore $Se_1(1 - Sp_2) < Se_2(1 - Sp_1)$, and it holds that

$$\frac{1 - Sp_1}{1 - Sp_2} > \frac{Se_1}{Se_2} > 1 \Rightarrow rFPF_{12} > rTPF_{12} > 1.$$

As $q\Delta_1 < 0$ and $\Delta_1 - p\Delta_2 > 0$, applying equation (2) again it holds that $\nu < 0$ and therefore $\kappa_1(c) < \kappa_2(c)$. If $c' > 1$, the demonstrations are similar to those of $c' < 0$.

c) If $rTPF_{12} < 1$ and $rFPF_{12} < 1$, then $rTPF_{21} > 1$ and $rFPF_{21} > 1$, and the demonstrations are analogous to case b).

Appendix *B*

Bloch (1997) has deduced the expressions of the variances of $\hat{\kappa}_1(c)$ and $\hat{\kappa}_2(c)$ and of the covariance between them. We then obtain equivalent expressions and we also deduce the variance of the ratio of the two weighted kappa coefficients, an expression which is necessary to apply the method to calculate the sample size explained in Section 5. Let $\boldsymbol{\omega} = (Se_1, Sp_1, Se_2, Sp_2, p)^T$ be the vector of parameters, where $Se_1 = \frac{p_{10} + p_{11}}{p}$, $Sp_1 = \frac{q_{00} + q_{01}}{q}$, $Se_2 = \frac{p_{01} + p_{11}}{p}$ and $Sp_2 = \frac{q_{00} + q_{10}}{q}$, with $q = 1 - p$. Applying the delta method, the matrix of the asymptotic variances-covariances of $\hat{\boldsymbol{\omega}}$ is

$$\Sigma_{\hat{\omega}} = \left(\frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\pi}}\right) \Sigma_{\hat{\pi}} \left(\frac{\partial \boldsymbol{\omega}}{\partial \boldsymbol{\pi}}\right)^T .$$

Performing the algebraic operations it is obtained that

$$Var\left(\hat{Se}_1\right) = \frac{(p_{11} + p_{10})(p_{01} + p_{00})}{np^3} = \frac{Se_1(1 - Se_1)}{np},$$

$$Var\left(\hat{Se}_2\right) = \frac{(p_{11} + p_{01})(p_{10} + p_{00})}{np^3} = \frac{Se_2(1 - Se_2)}{np},$$

$$Var\left(\hat{Sp}_1\right) = \frac{(q_{11} + q_{10})(q_{01} + q_{00})}{nq^3} = \frac{Sp_1(1 - Sp_1)}{nq},$$

$$Var\left(\hat{Sp}_2\right) = \frac{(q_{11} + q_{01})(q_{10} + q_{00})}{nq^3} = \frac{Sp_2(1 - Sp_2)}{nq},\ Var(\hat{p}) = \frac{pq}{n},$$

$$Cov\left[\hat{Se}_1, \hat{Se}_2\right] = \frac{p_{11}p_{00} - p_{10}p_{01}}{np^3} = \frac{\varepsilon_1}{np},\ Cov\left[\hat{Sp}_1, \hat{Sp}_2\right] = \frac{q_{11}q_{00} - q_{10}q_{01}}{nq^3} = \frac{\varepsilon_0}{nq}$$

and

$$Cov\left(\hat{Se}_h, \hat{Sp}_h\right) = Cov\left(\hat{Se}_h, \hat{p}\right) = Cov\left(\hat{Sp}_h, \hat{p}\right) = 0, \text{ with } h = 1, 2.$$

The estimators of the variances-covariances are obtained substituting each parameter with its corresponding estimator, where $\hat{Se}_1 = \frac{s_{11}+s_{10}}{s}$, $\hat{Se}_2 = \frac{s_{11}+s_{01}}{s}$, $\hat{Sp}_1 = \frac{r_{01}+r_{00}}{r}$, $\hat{Sp}_2 = \frac{r_{10}+r_{00}}{r}$, $\hat{p} = \frac{s}{n}$, $\hat{q} = \frac{r}{n}$, $\hat{\varepsilon}_1 = \frac{\hat{p}_{11}}{\hat{p}} - \hat{Se}_1\hat{Se}_2 = \frac{s_{11}s_{00}-s_{10}s_{01}}{s^2}$ and $\hat{\varepsilon}_0 = \frac{\hat{q}_{00}}{\hat{q}} - \hat{Sp}_1\hat{Sp}_2 = \frac{r_{11}r_{00}-r_{10}r_{01}}{r^2}$. Applying the delta method, the variance of $\hat{\kappa}_h(c)$ is

$$Var\left[\hat{\kappa}_h(c)\right] \approx \left[\frac{\partial \kappa_h(c)}{\partial Se_h}\right]^2 Var\left(\hat{Se}_h\right) + \left[\frac{\partial \kappa_h(c)}{\partial Sp_h}\right]^2 Var\left(\hat{Sp}_h\right) + \left[\frac{\partial \kappa_h(c)}{\partial p}\right]^2 Var\left(\hat{p}\right).$$

In this expression the covariances are zero. Performing the algebraic operations, it is obtained that

$$Var\left[\hat{\kappa}_h(c)\right] \approx \left[\frac{\kappa_h(c)}{pqY_h}\right]^2 \times \left[\left\{a_{h1}^2 Var\left(\hat{Se}_h\right) + a_{h2}^2 Var\left(\hat{Sp}_h\right) + a_{h3}^2 Var\left(\hat{p}\right)\right\}\right]$$

with $h = 1, 2$, and where

$$a_{h1} = pq - p(q-c)\kappa_h(c),$$

$$a_{h2} = a_{h1} + (q-c)\kappa_h(c)$$

and

$$a_{h3} = (1-2p)Y_h - \left[(1-c-2p)Y_h + Sp_h + c - 1\right]\kappa_h(c).$$

The expression of $\hat{Var}\left[\hat{\kappa}_h(c)\right]$ is obtained substituting in the previous expressions each parameter with its estimator. Regarding the covariance between $\hat{\kappa}_1(c)$ and $\hat{\kappa}_2(c)$, applying the delta method again it is obtained that

$$Cov\left[\hat{\kappa}_1(c), \hat{\kappa}_2(c)\right] \approx \frac{\partial \kappa_1(c)}{\partial Se_1}\frac{\partial \kappa_2(c)}{\partial Se_2} Cov\left[\hat{Se}_1, \hat{Se}_2\right] + \frac{\partial \kappa_1(c)}{\partial Sp_1}\frac{\partial \kappa_2(c)}{\partial Sp_2} Cov\left[\hat{Sp}_1, \hat{Sp}_2\right] + \frac{\partial \kappa_1(c)}{\partial p}\frac{\partial \kappa_2(c)}{\partial p} Var\left(\hat{p}\right).$$

In this expression, the rest of the covariances are equal to zero. Performing the algebraic operations it is obtained that

$$Cov\left[\hat{\kappa}_1(c),\hat{\kappa}_2(c)\right]\approx$$

$$\frac{\kappa_1(c)\kappa_2(c)}{p^2q^2Y_1Y_2}\times\left[a_{11}a_{21}Cov\left(\hat{Se}_1,\hat{Se}_2\right)+a_{12}a_{22}Cov\left(\hat{Sp}_1,\hat{Sp}_2\right)+a_{13}a_{23}\hat{Var}\left(\hat{p}\right)\right].$$

The expression of $\hat{Cov}\left[\hat{\kappa}_1(c),\hat{\kappa}_2(c)\right]$ is obtained substituting in this equation each parameter with its estimator.

Regarding the ration of the two weighted kappa coefficients, the variance of θ is easily calculated applying the delta method again, i.e.

$$Var\left(\hat{\theta}\right)\approx\sum_{h=1}^{2}\left(\frac{\partial\theta}{\partial\kappa_h(c)}\right)^2 Var\left[\hat{\kappa}_h(c)\right]+2\frac{\partial\theta}{\partial\kappa_1(c)}\frac{\partial\theta}{\partial\kappa_2(c)}Cov\left[\hat{\kappa}_1(c),\hat{\kappa}_2(c)\right].$$

Performing the algebraic operations,

$$Var\left(\hat{\theta}\right)\approx\frac{\kappa_2^2(c)Var\left[\hat{\kappa}_1(c)\right]+\kappa_1^2(c)Var\left[\hat{\kappa}_2(c)\right]-2\kappa_1(c)\kappa_2(c)Cov\left[\hat{\kappa}_1(c),\hat{\kappa}_2(c)\right]}{\kappa_2^4(c)},\quad(7)$$

and substituting in this equation each parameter with its estimator, we obtain the expression of $\hat{Var}\left(\hat{\theta}\right)$. The expression of variance of $\hat{Var}\left[\ln\left(\hat{\theta}\right)\right]$ is calculated in a similar way to in the previous case, but considering $\ln(\theta)$ instead of θ.

Appendix *C*

The selection of the *CI* with the best asymptotic behaviour, both for the difference δ and for the ratio θ, was made taking the following steps: 1) Choose the *CIs* with the least failures ($CP>93\%$), 2) Choose the *CIs* that are the most accurate i.e. those with the lowest *AL*. The first step in this method establishes that the *CI* does not fail when $CP>93\%$. In the simulation experiments the *CIs* were calculated to a 95% confidence i.e. $\gamma=1-\alpha=0.95$ is the nominal confidence and $\alpha=5\%$ is the nominal error. Then $\Delta\alpha=\alpha-\alpha^*=\gamma^*-\gamma$, where γ^* is the *CP* calculated and α^* is the type I error.

Moreover, the hypothesis test to check the equality of the two weighted kappa coefficients is $H_0:\kappa_1(c)=\kappa_2(c)$ vs $H_1:\kappa_1(c)\neq\kappa_2(c)$. Based on the difference of both parameters, this hypothesis test is equivalent to test $H_0:\delta=0$ vs $H_1:\delta\neq0$. This test can be solved through different methods. Applying Bloch's method (1997), the test

statistic is given by equation (equation (10) of the manuscript). The statistics for the bootstrap method and for the Bayesian method are obtained computationally.

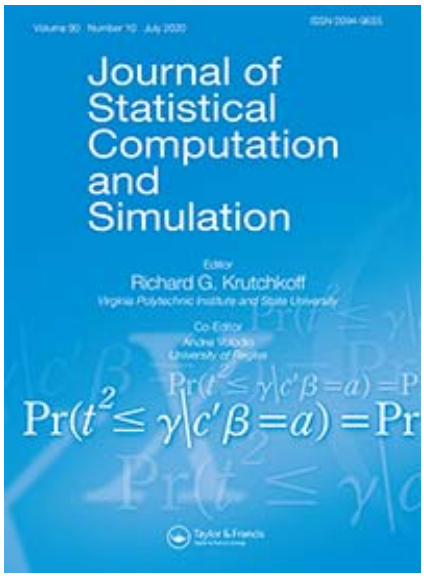
In step 1 of the method, a *CI* has a failure if $CP \leq 93\%$, i.e. if $\Delta\alpha \leq -2$. In this situation, the type I error of the corresponding hypothesis test is $\geq 7\%$, and therefore it is a very liberal hypothesis test and it can give false significances. The criteria of 93% has been used by other authors (Price and Bonett, 2004; Martín-Andrés and Álvarez-Hernández, 2014a, 2014b; Montero-Alonso and Roldán-Nofuentes, 2018). If $\Delta\alpha > 2\%$, i.e. $CP > 97\%$, then the hypothesis test is very conservative (its type I error is very small, $< 3\%$), but it does not give false significances. Consequently, the choose of the optimal *CI* is linked to the decisions of the hypothesis test, and it is preferable to choose a conservative test rather than a very liberal one (as there will be no false significances because its type I error is lower than the nominal one). The method for the *CIs* for the ratio θ is justified in a similar way.

APPENDIX IV

EM and *SEM* algorithms to compare the weighted kappa coefficients of two diagnostic tests in the presence of partial verification and discrete covariates

Roldán-Nofuentes, J.A., Sidaty-Regad S.B. (2020). *EM* and *SEM* algorithms to compare the weighted kappa coefficients of two diagnostic tests in the presence of partial verification and discrete covariates. Journal of Statistical Computation and Simulation. Accepted, in press. DOI: 10.1080/00949655.2020.1804903.

Category: Statistics and Probability. JCR 2019 (last published): 0.918. Rank: 75/124. Quartile: Q3.



Abstract

The weighted kappa coefficient of a binary diagnostic test is a measure of the beyond chance agreement between the diagnostic test and the gold standard, and depends on the sensitivity and the specificity of the diagnostic test, on the disease prevalence and on the relative importance between the false positives and the false negatives. This manuscript studies a hypothesis test to compare the weighted kappa coefficients of two binary diagnostic tests when, in the presence of partial disease verification, a discrete covariate is observed in all individuals. The *EM* algorithm is applied to estimate the weighted kappa coefficients and the *SEM* algorithm is applied to estimate their variances-covariances. Simulation experiments were carried out to study the size and the power of the proposed hypothesis test. The results were applied to a real example on the diagnosis of the Alzheimer's disease.

Key words: Discrete covariate, EM and SEM algorithms, Partial verification, Weighted kappa coefficient.

Mathematics Subject Classification: 62P10, 6207.

1. Introduction

The fundamental parameters of a binary diagnostic test (*BDT*) are sensitivity and specificity. The sensitivity (*Se*) is the probability of the *BDT* result being positive when the individual has the disease, and the specificity (*Sp*) is the probability of the result of the *BDT* being negative when the individual does not have the disease. In order to obtain unbiased estimators of the *Se* and *Sp* of a *BDT* it is necessary to assess the *BDT* in relation to a gold standard (*GS*), which is a medical test that objectively determines if an individual has the disease or not. When we consider the losses or costs associated with an erroneous classification with a *BDT*, the effectiveness of the *BDT* is estimated by the weighted kappa coefficient [1, 2, 3]. The weighted kappa coefficient of a *BDT* is a measure of the beyond chance agreement between the *BDT* and the *GS*, and depends on the *Se* and *Sp* of the *BDT*, on the disease prevalence (*p*) and on the relative importance between the false positives and the false negatives (weighting index).

When comparing the parameters of two *BDTs* in relation to the same *GS*, the most frequent type of sampling is the paired design [4, 5]. This design consists of applying the two *BDTs* to all of the individuals in a random sample sized *n*, where the disease status (whether the disease is present or absent) of all of the *n* individuals is known through the application of the *GS*. Subject to this type of design, Roldán-Nofuentes and Sidaty-Regad [6] studied different methods to compare the sensitivities and the specificities of the two *BDTs*. Subject to this same type of sampling, the comparison of the weighted kappa coefficients of two *BDTs* is made by applying the method of Bloch [7].

Moreover, in clinical practice when comparing the parameters of two (or more) *BDTs* it is common for the *GS* not to be applied to all of the individuals in the sample. Therefore, if the *GS* consists of an expensive test or a test that represents a high risk for the individual, the *GS* is not applied to all of the individuals in the sample. In this situation, the results of the *BDTs* are known for all of the individuals in a sample, but the disease status (i.e. the result of the *GS*) is only known for a subset of them (and therefore is unknown for the remaining subset). This situation is known as partial disease verification. Assuming that the verification process is missing at random (*MAR*) [8], there are several studies that have been carried out to compare two *BDTs*. Zhou [9] studied a hypothesis test to compare the sensitivities (specificities) of two *BDTs* applying the method of maximum likelihood. Harel and Zhou [10] applied multiple

imputation to compare the two sensitivities (specificities) through confidence intervals. Roldán Nofuentes and Luna [11, 12] studied hypothesis tests to compare the sensitivities (specificities) of two *BDTs* applying the *EM* and *SEM* algorithms. Roldán-Nofuentes and Luna [13] studied a hypothesis test to compare the weighted kappa coefficients of two *BDTs* applying the method of maximum likelihood.

In the presence of partial disease verification, the selection of a patient to verify his or her disease status with the *GS* may also depend on discrete covariates which are related to the disease. For example, in the diagnosis of the Alzheimer's disease [14] an advanced age of the patient (≥ 75 years) is a risk factor for this disease. The probability of selecting a patient to perform a clinical assessment (*GS*) conditionally depends on the result of the cognitive test (*BDT*) and on the age of the patient (≥ 75 years or < 75). The age of the patient is modelled using a discrete (binary) variable $(\geq 75$ years or $< 75)$. Zhou [9] studied a hypothesis test to compare the sensitivities (specificities) of two *BDTs* when, in the presence of partial verification of the disease, discrete covariates are observed in all individuals. The probability of selecting a patient to perform a clinical assessment (*GS*) depends on the results of two *BDTs* (a new *BDT* and a cognitive test) and on the age of the patient (≥ 75 years or < 75). Here the age of the patient is modelled through a binary variable. As cognitive deterioration increases in line with the age of the patient, age-adjustment is needed to properly describe the diagnosis effectiveness of each *BDTs*, and consequently to compare parameters of both *BDTs*.

The objective of this manuscript is to study a hypothesis test to compare the weighted kappa coefficients of two *BDTs* when, in the presence of partial disease verification, a discrete covariate is observed in all individuals. The manuscript is structured in the following way. Section 2 explains the weighted kappa coefficient of a *BDT*. In Section 3, an asymptotic hypothesis test is deduced to compare the two weighted kappa coefficients in the situation previously described by applying the *EM* and *SEM* algorithms. In Section 4, simulation experiments are carried out to study the size and the power of the hypothesis test deduced in Section 3 when the covariate is binary. In Section 5, the results are applied to a real example on the diagnosis of the Alzheimer's disease, and in Section 6 the results obtained are discussed.

2. Weighted kappa coefficient

Let us consider a *BDT* whose effectiveness is evaluated with respect to a *GS*. Let L be the loss or cost that is committed when the *BDT* is negative in an individual who has the disease, and let L' be the loss or cost that is committed when the *BDT* is positive in an individual who does not have the disease. The loss L is associated with the false negatives and the loss L' is associated with the false positives, assuming that $L = L' = 0$ when an individual (with or without the disease) is classified correctly with the *BDT*. The weighted kappa coefficient $\kappa(c)$ of a *BDT* is expressed [1, 2, 3] as

$$\kappa(c) = \frac{p\overline{p}Y}{pc\overline{Q} + \overline{p}\overline{c}Q}, \tag{1}$$

where $\overline{p} = 1 - p$, $Y = Se + Sp - 1$ is the Youden's index [15], $Q = pSe + \overline{p}(1 - Sp)$ is the probability of the *BDT* result being positive, $\overline{Q} = 1 - Q$, $c = L/(L' + L)$ is the weighting index and $\overline{c} = 1 - c$. The weighting index c is a measure of the relative importance between the false positives and the false negatives. For example, let us consider the diagnosis of the Alzheimer's disease using a cognitive test as a diagnostic test. If the cognitive test is positive for a patient who does not have this disease (false positive), then a clinical assessment (*GS*) will be performed, which will finally give a negative diagnosis. The loss L' will be determined from the economic costs of the diagnosis and also based on the stress, anxiety, etc., caused to the patient. If the cognitive test is negative for a patient who has this disease (false negative), the patient may be diagnosed some time later. In this situation, the Alzheimer's disease may have advanced and the possibilities of the treatment will help reduce some symptoms and help control some behavioural symptoms will be reduced. The loss L is determined based on these considerations. Consequently, the losses L and L' are not only measured in economic terms but also based on stress, anxiety, risks, etc., and therefore in clinical practice the value of these losses cannot be determined. This is the reason why the relative importance between the losses L and L' is substituted by the relative importance between the false positives and the false negatives. The value of the weighting index can be assumed depending on the considerations made by the clinician about the false positives and the false negatives. If the clinician is more concerned about the false positives, as is the case in which the *BDT* is used as a prior step to a treatment involving some risk (for example a surgical operation), then $0 \le c < 0.5$. If the clinician is more

concerned about the false negatives, as is the case in which the *BDT* is used as a screening test, then $0.5 < c \leq 1$. Index c is 0.5 when the *BDT* is used for a simple diagnosis (the false positives and the false negatives have the same importance), and in this situation $\kappa(0.5)$ is known as Cohen's kappa coefficient. If $c = 0$ then $\kappa(0) = (Sp - \bar{Q})/Q$ ans if $c = 1$ then $\kappa(1) = (Se - Q)/\bar{Q}$. The weighted kappa coefficient can also be written as

$$\kappa(c) = \frac{\kappa(0)\kappa(1)}{c\kappa(0) + \bar{c}\kappa(1)}, \tag{2}$$

with $0 < c < 1$. The weighted kappa coefficient is a measure of the beyond chance agreement between the *BDT* and the *GS*. In the studies by Kraemer et al [3], Roldán-Nofuentes et al [16] and Roldán-Nofuentes and Amro [17, 18], we can see a broad review of the use and the properties of the weighted kappa coefficient.

We will now study the comparison of the weighted kappa coefficients of two *BDT*s when in the presence of partial verification a discrete covariate is observed in all of the individuals.

3. The model

Let us consider two *BDT*s, *Test* 1 and *Test* 2, that are applied to all n individuals in a random sample. Let T_h be the random variable that models the result of the hth *BDT*, so that $T_h = 1$ when the result is positive and $T_h = 0$ when it is negative. Let V be the random variable that models the verification process, so that $V = 1$ when the disease status of an individual is verified with the *GS* and $V = 0$ when it is not. Let D be the random variable that models the result of the *GS*: $D = 1$ when the individual verified has the disease and $D = 0$ when the individual verified does not have the disease. Disease prevalence is $p = P(D = 1)$ and $\bar{p} = 1 - p = P(D = 0)$. Moreover, let us consider that for all of the n individuals of the sample we observe a vector $X = (x_1, x_2, ..., x_M)$ of a discrete covariate, where x_m is each one of the different values or patterns that the covariate can take with $m = 1, ..., M$. Let us suppose that the number of individuals that

verify $X = x_m$ is n_m, and therefore $n = \sum_{m=1}^{M} n_m$. For $X = x_m$ the frequencies obtained are those given in Table 1.

Table 1. Observed frequencies in the presence of partial verification for $X = x_m$.

	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$V = 1$					
$D = 1$	s_{11m}	s_{10m}	s_{01m}	s_{00m}	s_m
$D = 0$	r_{11m}	r_{10m}	r_{01m}	r_{00m}	r_m
$V = 0$	u_{11m}	u_{10m}	u_{01m}	u_{00m}	u_m
Total	n_{11m}	n_{10m}	n_{01m}	n_{00m}	n_m

The sample of *n* individuals can be seen as a sample of a mixture of *M* multinomial independent 3×4 tables. For $X = x_m$, i.e. for the *m*-th table, the sensitivities of the *BDTs* are defined as

$$Se_{1m} = P\left(T_1 = 1 \middle| D = 1, X = x_m\right) \text{ and } Se_{2m} = P\left(T_2 = 1 \middle| D = 1, X = x_m\right),$$

and the specificities as

$$Sp_{1m} = P\left(T_1 = 0 \middle| D = 0, X = x_m\right) \text{ and } Sp_{2m} = P\left(T_1 = 0 \middle| D = 0, X = x_m\right).$$

It is assumed, just as happens in practice, that both *BDTs* are conditionally dependent on the disease, applying the conditional dependence model of Berry et al [19] to each one of the values of the covariate *X*, for $X = x_m$ it is verified that

$$P\left(T_1 = i, T_2 = i \middle| D = 1, X = x_m\right) =$$
$$P\left(T_1 = i \middle| D = 1, X = x_m\right) \times P\left(T_2 = j \middle| D = 1, X = x_m\right) + \Delta_{ij} Se_{1m} Se_{2m} \left(\alpha_{1m} - 1\right)$$

and

$$P\left(T_1 = i, T_2 = i \middle| D = 0, X = x_m\right) =$$
$$P\left(T_1 = i \middle| D = 0, X = x_m\right) \times P\left(T_2 = j \middle| D = 0, X = x_m\right) + \Delta_{ij} \left(1 - Sp_{1m}\right)\left(1 - Sp_{2m}\right)\left(\alpha_{0m} - 1\right),$$

where $\Delta_{ij} = 1$ if $i = j$ and $\Delta_{ij} = -1$ if $i \neq j$, and the parameter α_{1m} $\left(\alpha_{0m}\right)$ is the covariance [19] between both *BDTs* when $D = 1$ $\left(D = 0\right)$ and $X = x_m$, verifying that

$1 \le \alpha_{1m} \le 1/\max\{Se_{1m}, Se_{2m}\}$ and $1 \le \alpha_{0m} \le 1/\max\{(1-Sp_{1m}),(1-Sp_{2m})\}$. If $\alpha_{1m} = \alpha_{0m} = 1$ then both *BDTs* are conditionally independent on the disease when $X = x_m$, an assumption that is not realistic, so in practice $\alpha_{1m} > 1$ and/or $\alpha_{0m} > 1$.

For the *m*-th table the verification probabilities are defined as

$$\lambda_{ijkm} = P\left(V = 1 \middle| T_1 = i, T_2 = j, D = k, X = x_m\right),$$

i.e. λ_{ijkm} is the probability of verifying with the *GS* the disease status of an individual in which $T_1 = i$, $T_2 = j$, $D = k$ and $X = x_m$, with $i, j, k = 0,1$ and $m = 1,...,M$. Assuming that the verification process is *MAR* [8], i.e. that the probability of verifying the disease status of an individual only conditionally depends on the results of both *BDTs* and on the value of the covariate *X*, then

$$\lambda_{ijkm} = \lambda_{ijm} = P\left(V = 1 \middle| T_1 = i, T_2 = j, X = x_m\right). \tag{3}$$

Let $\delta_m = P(X = x_m)$ be the probability that in an individual $X = x_m$. Let $p_m = P\left(D = 1 \middle| X = x_m\right)$ be the disease prevalence for the individuals with $X = x_m$, and $\overline{p}_m = 1 - p_m$. For $X = x_m$ the data s_{ijm}, r_{ijm} and u_{ijm}, with $i, j = 0,1$, are the product of a multinomial distribution sized n_m and whose probabilities under the *MAR* assumption are shown in Appendix *A* (Partial verification: probabilities) of supplementary material of the manuscript.

For the *h*-th *BDT* and $X = x_m$ it holds that $\kappa_{hm}(0) = \left(Sp_{hm} - \overline{Q}_{hm}\right)/Q_{hm}$ and $\kappa_{hm}(1) = \left(Se_{hm} - Q_{hm}\right)/\overline{Q}_{hm}$, with $h = 1,2$. Let $p = \sum_{m=1}^{M} \delta_m p_m$ be the overall disease prevalence and $\overline{p} = \sum_{m=1}^{M} \delta_m \overline{p}_m$. The overall weighted kappa coefficients $\kappa_h(0)$ and $\kappa_h(1)$ are

$$\kappa_h(0) = \frac{\overline{p}\left[\sum_{m=1}^{M} \delta_m p_m \xi_{hm} \kappa_{hm}(1)\right] - p\left[\sum_{m=1}^{M} \delta_m \overline{p}_m \left(1 - \psi_{hm}\kappa_{hm}(0)\right)\right]}{\overline{p}\left\{\sum_{m=1}^{M} \delta_m \left[p_m \xi_{hm} \kappa_{hm}(1) + \overline{p}_m \left(1 - \psi_{hm}\kappa_{hm}(0)\right)\right]\right\}} \tag{4}$$

and

$$\kappa_h(1)=\frac{p\left[\sum_{m=1}^{M}\delta_m\overline{p}_m\psi_{hm}\kappa_{hm}(0)\right]-\overline{p}\left[\sum_{m=1}^{M}\delta_m p_m\left(1-\xi_{hm}\kappa_{hm}(1)\right)\right]}{p\left\{\sum_{m=1}^{M}\delta_m\left[p_m\left(1-\xi_{hm}\kappa_{hm}(1)\right)+\overline{p}_m\psi_{hm}\kappa_{hm}(0)\right]\right\}}, \tag{5}$$

where

$$\xi_{hm}=\frac{\overline{p}_m\kappa_{hm}(0)+p_m}{\overline{p}_m\kappa_{hm}(0)+p_m\kappa_{hm}(1)} \text{ and } \psi_{hm}=\frac{p_m\kappa_{hm}(1)+\overline{p}_m}{\overline{p}_m\kappa_{hm}(0)+p_m\kappa_{hm}(1)}.$$

The proof can be seen in Appendix *A* (Overall weighted kappa coefficients) of supplementary material. Finally, for $0<c<1$, the overall weighted kappa coefficient $\kappa_h(c)$ of the *h*-th *BDT* is

$$\kappa_h(c)=\frac{\kappa_h(0)\kappa_h(1)}{c\kappa_h(0)+\overline{c}\kappa_h(1)}, \tag{6}$$

with $h=1,2$.

The objective of this manuscript is to study the comparison of the weighted kappa coefficients of both *BDTs*, i.e.

$$H_0:\kappa_1(c)=\kappa_2(c) \text{ vs } H_1:\kappa_1(c)\neq\kappa_2(c).$$

If $\hat{\boldsymbol{\kappa}}(c)=\left(\hat{\kappa}_1(c),\hat{\kappa}_2(c)\right)$ is the vector whose components are the estimators of $\kappa_h(c)$ and $\sum_{\hat{\boldsymbol{\kappa}}(c)}$ is the variance-covariance matrix of $\hat{\boldsymbol{\kappa}}(c)$, then based on the asymptotic normality of $\hat{\boldsymbol{\kappa}}(c)$, $\left(\hat{\boldsymbol{\kappa}}(c)-\boldsymbol{\kappa}(c)\right)^T\xrightarrow[n\to\infty]{}N\left(0,\sum_{\hat{\boldsymbol{\kappa}}(c)}\right)$, a Wald type test statistic for the hypothesis test is

$$z=\frac{\hat{\kappa}_1(c)-\hat{\kappa}_2(c)}{\sqrt{\hat{V}ar\left[\hat{\kappa}_1(c)\right]+\hat{V}ar\left[\hat{\kappa}_2(c)\right]-2\hat{C}ov\left[\hat{\kappa}_1(c),\hat{\kappa}_2(c)\right]}}, \tag{7}$$

which is distributed according to a standard normal distribution when the sample size *n* is large. We then obtain the estimators of $\kappa_h(c)$ applying the *EM* algorithm and the estimators of the variances-covariances applying the *SEM* algorithm.

For each one of the *M* multinomial 3×4 tables the missing information is the true disease status of the individuals not verified with the *GS* $(V=0)$. For the *m*-th table $(X=x_m)$ let us suppose that from each frequency u_{ijm} of non-verified individuals, d_{ijm}

have the disease and $u_{ijm} - d_{ijm}$ do not have the disease, with $i, j = 0,1$. Then each one of the *M* tables can be expressed in the form of a 2×4 table with frequencies $s_{ijm} + d_{ijm}$ for $D = 1$ and $r_{ijm} + u_{ijm} - d_{ijm}$ for $D = 0$. Table 2 shows the frequencies of each 2×4 table.

Table 2. Frequencies of the complete data for $X = x_m$.

	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$D = 1$	$s_{11m} + d_{11m}$	$s_{10m} + d_{10m}$	$s_{01m} + d_{01m}$	$s_{00m} + d_{00m}$	$s_m + d_m$
$D = 0$	$r_{11m} + u_{11m}$ $-d_{11m}$	$r_{10m} + u_{10m}$ $-d_{10m}$	$r_{01m} + u_{01m}$ $-d_{01m}$	$r_{00m} + u_{00m}$ $-d_{00m}$	$r_m + u_m$ $-d_m$
Total	n_{11m}	n_{10m}	n_{01m}	n_{00m}	n_m

Let the vectors be $\boldsymbol{\delta} = (\delta_1,...,\delta_M)$, $\boldsymbol{\kappa}_1(0) = (\kappa_{11}(0),...,\kappa_{1M}(0))$, $\boldsymbol{\kappa}_1(1) = (\kappa_{11}(1),...,\kappa_{1M}(1))$, $\boldsymbol{\kappa}_2(0) = (\kappa_{21}(0),...,\kappa_{2M}(0))$, $\boldsymbol{\kappa}_2(1) = (\kappa_{21}(1),...,\kappa_{2M}(1))$, $\mathbf{p} = (p_1,...,p_M)$, $\boldsymbol{\alpha}_1 = (\alpha_{11},...,\alpha_{1M})$ and $\boldsymbol{\alpha}_0 = (\alpha_{01},...,\alpha_{0M})$. Then, subject to the *MAR* assumption (3), the log-likelihood function based on *n* individuals is

$$l(\boldsymbol{\delta}, \boldsymbol{\kappa}_1(0), \boldsymbol{\kappa}_1(1), \boldsymbol{\kappa}_2(0), \boldsymbol{\kappa}_2(1), \mathbf{p}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_0) = \sum_{i,j=0}^{1} \sum_{m=1}^{M} (s_{ijm} + d_{ijm}) \log(\delta_m \phi_{ijm}) + \sum_{i,j=0}^{1} \sum_{m=1}^{M} (r_{ijm} + u_{ijm} - d_{ijm}) \log(\delta_m \varphi_{ijm}), \tag{8}$$

where

$$\phi_{11m} = p_m \alpha_{1m} \xi_{1m} \kappa_{1m}(1) \xi_{2m} \kappa_{2m}(1),\ \phi_{10m} = p_m \xi_{1m} \kappa_{1m}(1)\left[1 - \alpha_{1m} \xi_{2m} \kappa_{2m}(1)\right],$$

$$\phi_{01m} = p_m \xi_{2m} \kappa_{2m}(1)\left[1 - \alpha_{1m} \xi_{1m} \kappa_{1m}(1)\right],$$

$$\phi_{00m} = p_m \left[1 - \xi_{1m} \kappa_{1m}(1) - \xi_{2m} \kappa_{2m}(1) + \alpha_{1m} \xi_{1m} \kappa_{1m}(1) \xi_{2m} \kappa_{2m}(1)\right]$$

$$\varphi_{11m} = \overline{p}_m \alpha_{0m} \left[1 - \psi_{1m} \kappa_{1m}(0)\right]\left[1 - \psi_{2m} \kappa_{2m}(0)\right],$$

$$\varphi_{10m} = \overline{p}_m \left[1 - \psi_{1m} \kappa_{1m}(0)\right]\left[1 - \alpha_{0m} + \alpha_{0m} \psi_{2m} \kappa_{2m}(0)\right],$$

$$\varphi_{01m} = \overline{p}_m \left[1 - \psi_{2m} \kappa_{2m}(0)\right]\left[1 - \alpha_{0m} + \alpha_{0m} \psi_{1m} \kappa_{1m}(0)\right]$$

and

$$\varphi_{00m} = \overline{p}_m \left\{ \psi_{1m}\kappa_{1m}(0) + \psi_{2m}\kappa_{2m}(0) - 1 + \alpha_{0m}\left[1-\psi_{1m}\kappa_{1m}(0)\right]\left[1-\psi_{2m}\kappa_{2m}(0)\right] \right\}.$$

The proof can be seen in Appendix *A* (Complete data: probabilities) of supplementary material. The log-likelihood function is also written as

$$\begin{gathered} l\left(\boldsymbol{\delta},\boldsymbol{\kappa}_1(0),\boldsymbol{\kappa}_1(1),\boldsymbol{\kappa}_2(0),\boldsymbol{\kappa}_2(1),\mathbf{p},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_0\right) = \\ l_1(\boldsymbol{\delta}) + l_2\left(\boldsymbol{\kappa}_1(0),\boldsymbol{\kappa}_1(1),\boldsymbol{\kappa}_2(0),\boldsymbol{\kappa}_2(1),\mathbf{p},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_0\right), \end{gathered} \tag{9}$$

where

$$l_1(\boldsymbol{\delta}) = \sum_{i,j=0}^{1}\sum_{m=1}^{M} n_{ijm}\log\left(\delta_m\right) \tag{10}$$

and

$$\begin{gathered} l_2\left(\boldsymbol{\kappa}_1(0),\boldsymbol{\kappa}_1(1),\boldsymbol{\kappa}_2(0),\boldsymbol{\kappa}_2(1),\mathbf{p},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_0\right) = \\ \sum_{i,j=0}^{1}\sum_{m=1}^{M}\left(s_{ijm}+d_{ijm}\right)\log\left(\phi_{ijm}\right) + \sum_{i,j=0}^{1}\sum_{m=1}^{M}\left(r_{ijm}+u_{ijm}-d_{ijm}\right)\log\left(\varphi_{ijm}\right). \end{gathered} \tag{11}$$

The Fisher information matrix of function (9) is

$$I\left(\boldsymbol{\delta},\boldsymbol{\kappa}_1(0),\boldsymbol{\kappa}_1(1),\boldsymbol{\kappa}_2(0),\boldsymbol{\kappa}_2(1),\mathbf{p},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_0\right) = \mathrm{Diag}\left\{I_1, I_2\right\}, \tag{12}$$

where $I_1 = I(\boldsymbol{\delta})$ and $I_2 = I\left(\boldsymbol{\delta},\boldsymbol{\kappa}_1(0),\boldsymbol{\kappa}_1(1),\boldsymbol{\kappa}_2(0),\boldsymbol{\kappa}_2(1),\mathbf{p},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_0\right)$ are the Fisher information matrixes of functions (10) and (11) respectively, verifying that

$$I^{-1}\left(\boldsymbol{\delta},\boldsymbol{\kappa}_1(0),\boldsymbol{\kappa}_1(1),\boldsymbol{\kappa}_2(0),\boldsymbol{\kappa}_2(1),\mathbf{p},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_0\right) = \mathrm{Diag}\left\{I_1^{-1}, I_2^{-1}\right\}, \tag{13}$$

and consequently the covariances between $\boldsymbol{\delta}$ and the rest of parameters $\left(\boldsymbol{\kappa}_1(0),\boldsymbol{\kappa}_1(1),\boldsymbol{\kappa}_2(0),\boldsymbol{\kappa}_2(1),\mathbf{p},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_0\right)$ are zero.

The estimator of δ_m is easily calculated from the function (10), i.e. $\hat{\delta}_m = n_m/n$. All of the other parameters are going to be estimated from function (11) applying the *EM* algorithm.

3.2. EM algorithm

The missing information (disease status of the individuals who are not verified with the *GS*) is reconstructed in the *E* step of the algorithm and in the *M* step the values of the

maximum likelihood estimators are imputed. Let $d_{ijm}^{(t)}$ be the value of d_{ijm} in the *t*-th iteration of the *EM* algorithm and $d_m^{(t)} = \sum_{i,j=0}^{1} d_{ijm}^{(t)}$. Let $s_m = \sum_{i,j=0}^{1} s_{ijm}$, $r_m = \sum_{i,j=0}^{1} r_{ijm}$, $u_m = \sum_{i,j=0}^{1} u_{ijm}$, $n_{ijm} = s_{ijm} + r_{ijm} + u_{ijm}$ and $n_m = \sum_{i,j=0}^{1} n_{ijm}$. The values of the *MLEs* in the *t*-th iteration are calculated through the following equations:

$$\hat{\kappa}_{1m}^{(t)}(0) = \frac{\left[\sum_{j=0}^{1}\left(s_{1jm} + d_{1jm}^{(h)}\right)\right]\left[\sum_{j=0}^{1}\left(r_{0jm} + u_{0jm} - d_{0jm}^{(t)}\right)\right] - \left[\sum_{j=0}^{1}\left(s_{0jm} + d_{0jm}^{(t)}\right)\right]\left[\sum_{j=0}^{1}\left(r_{1jm} + u_{1jm} - d_{1jm}^{(t)}\right)\right]}{\left(r_m + u_m - d_m^{(t)}\right)\left(\sum_{j=0}^{1} n_{1jm}\right)},$$

$$\hat{\kappa}_{1m}^{(t)}(1) = \frac{\left[\sum_{j=0}^{1}\left(s_{1jm} + d_{1jm}^{(t)}\right)\right]\left[\sum_{j=0}^{1}\left(r_{0jm} + u_{0jm} - d_{0jm}^{(t)}\right)\right] - \left[\sum_{j=0}^{1}\left(s_{0jm} + d_{0jm}^{(t)}\right)\right]\left[\sum_{j=0}^{1}\left(r_{1jm} + u_{1jm} - d_{1jm}^{(t)}\right)\right]}{\left(s_m + d_m^{(t)}\right)\left(\sum_{j=0}^{1} n_{0jm}\right)},$$

$$\hat{\kappa}_{2m}^{(t)}(0) = \frac{\left[\sum_{i=0}^{1}\left(s_{i1m} + d_{i1m}^{(t)}\right)\right]\left[\sum_{i=0}^{1}\left(r_{i0m} + u_{i0m} - d_{i0m}^{(t)}\right)\right] - \left[\sum_{i=0}^{1}\left(s_{i0m} + d_{i0m}^{(t)}\right)\right]\left[\sum_{i=0}^{1}\left(r_{i1m} + u_{i1m} - d_{i1m}^{(t)}\right)\right]}{\left(r_m + u_m - d_m^{(t)}\right)\left(\sum_{i=0}^{1} n_{i1m}\right)},$$

$$\hat{\kappa}_{2m}^{(t)}(1) = \frac{\left[\sum_{i=0}^{1}\left(s_{i1m} + d_{i1m}^{(t)}\right)\right]\left[\sum_{i=0}^{1}\left(r_{i0m} + u_{i0m} - d_{i0m}^{(t)}\right)\right] - \left[\sum_{i=0}^{1}\left(s_{i0m} + d_{i0m}^{(t)}\right)\right]\left[\sum_{i=0}^{1}\left(r_{i1m} + u_{i1m} - d_{i1m}^{(t)}\right)\right]}{\left(s_m + d_m^{(t)}\right)\left(\sum_{i=0}^{1} n_{i0m}\right)},$$

$$\hat{p}_m^{(t)} = \frac{s_m + d_m^{(t)}}{n_m},\ \hat{\alpha}_{1m}^{(t)} = \frac{\left(s_m + d_m^{(t)}\right)\left(s_{11m} + d_{11m}^{(t)}\right)}{\left[\sum_{i=0}^{1}\left(s_{i1m} + d_{i1m}^{(t)}\right)\right]\left[\sum_{j=0}^{1}\left(s_{1jm} + d_{1jm}^{(t)}\right)\right]}$$

and

$$\hat{\alpha}_{0m}^{(t)} = \frac{\left(r_m + u_m - d_m^{(t)}\right)\left(r_{11m} + u_{11m} - d_{11m}^{(t)}\right)}{\left[\sum_{i=0}^{1}\left(r_{i1m} + u_{i1m} - d_{i1m}^{(t)}\right)\right]\left[\left\{\sum_{j=0}^{1}\left(r_{1jm} + u_{1jm} - d_{1jm}^{(t)}\right)\right\}\right]}$$

The proof can be seen in Appendix *A* (*EM* Algorithm: estimators) of supplementary material. The estimators in the $(t+1)$-th iteration of the algorithm are calculated with the same previous equations substituting super index t with $t+1$, where

$$d_{ijm}^{(t+1)} = u_{ijm} \frac{\hat{\phi}_{ijm}^{(t)}}{\hat{\phi}_{ijm}^{(t)} + \hat{\varphi}_{ijm}^{(t)}}, \;\; i, j = 0,1 \;\; , \;\; m = 1,...,M \;\; ,$$

and where $\hat{\phi}_{ijm}^{(t)}$ and $\hat{\varphi}_{ijm}^{(t)}$ are the estimators of probabilities ϕ_{ijm} and φ_{ijm} in the *t*-th iteration of the algorithm, and are calculated substituting in the expressions of ϕ_{ijm} and φ_{ijm} the parameters with their respective estimators obtained in the *t*-th iteration. As an initial value $d_{ijm}^{(0)}$ any value between 0 and u_{ijm} can be taken, i.e. $0 \le d_{ijm}^{(0)} \le u_{ijm}$. The *EM* algorithm stops when the difference between the values of the log-likelihood functions of two consecutive iterations is lower than a sufficiently small γ value, e.g. $\gamma = 10^{-10}$ or $\gamma = 10^{-12}$. From the *EM* algorithm, in the *m*-th frequency table $(X = x_m)$ seven parameters are estimated: $\kappa_{1m}(0)$, $\kappa_{1m}(1)$, $\kappa_{2m}(0)$, $\kappa_{2m}(1)$, p_m, α_{1m} and α_{0m}. If the *EM* algorithm converged in T iterations, we denote as $\hat{\boldsymbol{\theta}}(m) = (\hat{\kappa}_{1m}(0), \hat{\kappa}_{1m}(1), \hat{\kappa}_{2m}(0), \hat{\kappa}_{2m}(1), \hat{p}_m, \hat{\alpha}_{1m}, \hat{\alpha}_{0m})$ the final estimators obtained for $X = x_m$, with $m = 1,...,M$. As the number of values of the covariate *X* is *M*, with the *EM* algorithm $7M$ parameters are estimated in total, to which we have to add the estimation of the components of the vector $\boldsymbol{\delta}$ ($M-1$, since $\sum_{m=1}^{M} \delta_m = 1$). Therefore, in total $8M-1$ parameters are estimated.

Once we have obtained $\hat{\kappa}_{hm}(0)$ and $\hat{\kappa}_{hm}(1)$, with $h = 1,2$ and $m = 1,...,M$, the estimators $\hat{\kappa}_h(0)$ and $\hat{\kappa}_h(1)$ are calculated from equations (4) and (5), and finally $\hat{\kappa}_1(c)$ and $\hat{\kappa}_2(c)$ are calculated applying equations (6).

Then the variances-covariances are estimated by applying the *SEM* algorithm [20].

3.3. SEM algorithm

The estimation of the asymptotic variance-covariance matrix of $\hat{\kappa}_{hm}(0)$, $\hat{\kappa}_{hm}(1)$, $\hat{p}_m$, $\hat{\alpha}_{1m}$ and $\hat{\alpha}_{0m}$, with $h = 1,2$ and $m = 1,...,M$, can be obtained through the application of the *SEM* algorithm [20]. The *SEM* algorithm is a computational method which allows us to estimate the variance-covariance matrix of a vector of estimators using the

calculations made in the application of the *EM* algorithm. Let $\sum_{\hat{\boldsymbol{\theta}}}$ be the matrix of variances-covariances of $\hat{\kappa}_{hm}(0)$, $\hat{\kappa}_{hm}(1)$, $\hat{p}_m$, $\hat{\alpha}_{1m}$ and $\hat{\alpha}_{0m}$, with $h=1,2$ and $m=1,...,M$, sized $7M\times 7M$. Dempster et al [21] demonstrated that

$$\sum_{\hat{\boldsymbol{\theta}}} = I_{oc}^{-1}\left(I - DM\right)^{-1}, \tag{14}$$

where *I* is the matrix identity and $DM = I_{mis}I_{oc}^{-1}$, and where I_{oc} is the Fisher information matrix of the complete data and I_{mis} is the Fisher information matrix of the missing data. The *SEM* algorithm consists of three phases: 1) the evaluation of the matrix I_{oc}^{-1}, 2) the evaluation of the matrix *DM*, and 3) the evaluation of the variance-covariance matrix $\sum_{\hat{\boldsymbol{\theta}}}$. The main objective of the *SEM* algorithm is to calculate the *DM* matrix. We then analyse the three phases of this algorithm in the situation studied here.

The *SEM* algorithm firstly requires the evaluation of the matrix I_{oc}^{-1}. This matrix is the inverse matrix of the Fisher information matrix of the complete data, and is calculated with the log-likelihood function (9) obtained from the last *M* tables after the application of the *EM* algorithm described in Section 2.2. If the *EM* algorithm converges in *T* iterations, then the frequencies of the *m*-th table are $s_{ijm} + d_{ijm}^{(T)}$ for $D=1$ and $r_{ijm} + u_{ijm} - d_{ijm}^{(T)}$ for $D=0$, with $m=1,...,M$. For the calculation of the Fisher information matrix, the parameters are substituted by their corresponding estimations obtained in the last iteration of the *EM* algorithm.

The second phase of the *SEM* algorithm consists of calculating the *DM* matrix. Let the vectors be $\boldsymbol{\theta}(m) = \left(\kappa_{1m}(0), \kappa_{1m}(1), \kappa_{2m}(0), \kappa_{2m}(1), p_m, \alpha_{1m}, \alpha_{0m}\right)$ and $\hat{\boldsymbol{\theta}}(m)$, with $m=1,...,M$. Each vector $\boldsymbol{\theta}(m)$, sized 7, has as components the parameters in $X = x_m$, and $\hat{\boldsymbol{\theta}}(m)$ has as components the final estimators in $X = x_m$ obtained by applying the *EM* algorithm. Let the vectors be

$$\hat{\boldsymbol{\theta}}^{(t)}(m) = \left(\hat{\kappa}_{1m}^{(t)}(0), \hat{\kappa}_{1m}^{(t)}(1), \hat{\kappa}_{2m}^{(t)}(0), \hat{\kappa}_{2m}^{(t)}(1), \hat{p}_m^{(t)}, \hat{\alpha}_{1m}^{(t)}, \hat{\alpha}_{0m}^{(t)}\right),$$

with $m=1,...,M$, which has as components the estimations of the parameters for $X = x_m$ in the *t*-th iteration of the *EM* algorithm. Let the vectors be

$$\boldsymbol{\theta} = \left(\boldsymbol{\theta}(1),...,\boldsymbol{\theta}(M)\right),\ \hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\theta}}(1),...,\hat{\boldsymbol{\theta}}(M)\right) \text{ and } \hat{\boldsymbol{\theta}}^{(t)} = \left(\hat{\boldsymbol{\theta}}^{(t)}(1),...,\hat{\boldsymbol{\theta}}^{(t)}(M)\right),$$

each one of them sized $7M$, obtained by concatenating the *M* respective vectors $\boldsymbol{\theta}(m)$, $\hat{\boldsymbol{\theta}}(m)$ and $\hat{\boldsymbol{\theta}}^{(t)}(m)$. The elements of the *DM* matrix, sized $7M \times 7M$, are obtained by applying the following algorithm:

INPUT: $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\theta}}(1),...,\hat{\boldsymbol{\theta}}(M)\right)$ and $\hat{\boldsymbol{\theta}}^{(t)} = \left(\hat{\boldsymbol{\theta}}^{(t)}(1),...,\hat{\boldsymbol{\theta}}^{(t)}(M)\right)$.

Step 1. Calculate $\hat{\boldsymbol{\theta}}^{(t+1)} = \left(\hat{\boldsymbol{\theta}}^{(t+1)}(1),...,\hat{\boldsymbol{\theta}}^{(t+1)}(M)\right)$ applying the *EM* algorithm.

Step 2. Let the vectors be

$$\hat{\boldsymbol{\tau}}_1^{(t)}(m) = \left(\hat{\kappa}_{1m}^{(t)}(0), \hat{\kappa}_{1m}(1), \hat{\kappa}_{2m}(0), \hat{\kappa}_{2m}(1), \hat{p}_m, \hat{\alpha}_{1m}, \hat{\alpha}_{0m}\right)$$
$$\hat{\boldsymbol{\tau}}_2^{(t)}(m) = \left(\hat{\kappa}_{1m}(0), \hat{\kappa}_{1m}^{(t)}(1), \hat{\kappa}_{2m}(0), \hat{\kappa}_{2m}(1), \hat{p}_m, \hat{\alpha}_{1m}, \hat{\alpha}_{0m}\right)$$
$$\vdots$$
$$\hat{\boldsymbol{\tau}}_7^{t}(m) = \left(\hat{\kappa}_{1m}(0), \hat{\kappa}_{1m}(1), \hat{\kappa}_{2m}(0), \hat{\kappa}_{2m}(1), \hat{p}_m, \hat{\alpha}_{1m}, \hat{\alpha}_{0m}^{(t)}\right),$$

with $m = 1,...,M$. Therefore, each vector $\hat{\boldsymbol{\tau}}_i^{(t)}(m)$ has as the *i*-th component the estimation of the corresponding parameter obtained in the *t*-th iteration of the *EM* algorithm, and the rest of the components are the final estimations obtained by applying the *EM* algorithm. Let the vectors be

$$\hat{\boldsymbol{\upsilon}}_1^{(t)}(1) = \left(\hat{\boldsymbol{\tau}}_1^{t}(1), \hat{\boldsymbol{\theta}}(2),...,\hat{\boldsymbol{\theta}}(M)\right),...,\hat{\boldsymbol{\upsilon}}_7^{(t)}(1) = \left(\hat{\boldsymbol{\upsilon}}_7^{(t)}(1), \hat{\boldsymbol{\theta}}(2),...,\hat{\boldsymbol{\theta}}(M)\right)$$
$$\vdots$$
$$\hat{\boldsymbol{\upsilon}}_1^{t}(m) = \left(\hat{\boldsymbol{\theta}}(1),..,\hat{\boldsymbol{\tau}}_1^{(t)}(m),..,\hat{\boldsymbol{\theta}}(M)\right),...,\hat{\boldsymbol{\upsilon}}_7^{(t)}(m) = \left(\hat{\boldsymbol{\theta}}(1),..,\hat{\boldsymbol{\tau}}_7^{(t)}(m),..,\hat{\boldsymbol{\theta}}(M)\right)$$
$$\vdots$$
$$\hat{\boldsymbol{\upsilon}}_1^{(t)}(M) = \left(\hat{\boldsymbol{\theta}}(1),..,\hat{\boldsymbol{\theta}}(M-1),\hat{\boldsymbol{\tau}}_1^{(t)}(M)\right),...,\hat{\boldsymbol{\upsilon}}_7^{(t)}(M) = \left(\hat{\boldsymbol{\theta}}(1),..,\hat{\boldsymbol{\theta}}(M-1),\hat{\boldsymbol{\tau}}_7^{(t)}(M)\right).$$

For each vector $\hat{\boldsymbol{\upsilon}}_i^{(t)}(m)$, with $i=1,...,7$ and $m=1,...,M$, execute the first iteration of the *EM* algorithm considering $\hat{\boldsymbol{\upsilon}}_i^{(t)}(m)$ to be the initial value of the algorithm, and as a result we obtain the vectors $\hat{\boldsymbol{\omega}}_i^{(t+1)}(m)$, with $i=1,...,7$ and $m=1,...,M$.

Step 3. Calculate the elements of the *DM* matrix as

$$\beta_{ij}^{(t)}(k,l)=\begin{cases}\dfrac{\hat{\boldsymbol{\omega}}_{ij}^{(t+1)}(k)-\hat{\boldsymbol{\theta}}_j(l)}{\hat{\boldsymbol{\upsilon}}_i^{(t)}(k)-\hat{\boldsymbol{\theta}}_i(k)}, & k=l \\ 0, & k\neq l\end{cases} \tag{15}$$

with $i,j=1,...,7$ and $k,l=1,...,M$, and where $\hat{\boldsymbol{\omega}}_{ij}^{(t+1)}(k)$ is the *j*-th component of $\hat{\boldsymbol{\omega}}_i^{(t+1)}(k)$. The proof can be seen in Appendix *B* of supplementary material.

OUTPUT: $\hat{\boldsymbol{\theta}}^{(t+1)}=\left(\hat{\boldsymbol{\theta}}^{(t+1)}(1),...,\hat{\boldsymbol{\theta}}^{(t+1)}(M)\right)$ and $\beta_{ij}^{(t)}(k,l)$, with $i,j=1,...,7M$ and $k,l=1,...,M$.

This algorithm is repeated until

$$\left|\beta_{ij}^{(t+1)}(k,l)-\beta_{ij}^{(t)}(k,l)\right|\leq\gamma', \tag{16}$$

where $\gamma'=\sqrt{\gamma}$ [20]. Therefore, the bigger γ' (or γ) is, the greater the numerical errors of the *DM* matrix, affecting the estimation of the variance-covariance matrix.

For each element of the *DM* matrix, convergence is reached in a number of different iterations. Let us suppose that $T_{ij}(k,l)$ iterations of the previous algorithm are necessary to calculate the element $\beta_{ij}(k,l)$, then it is verified that

$$\beta_{ij}^{(T_{ij})}(m,m)=\frac{\hat{\boldsymbol{\omega}}_{ij}^{(T_{ij}+1)}(m)-\hat{\boldsymbol{\theta}}_j(m)}{\hat{\boldsymbol{\upsilon}}_i^{(T_{ij})}(m)-\hat{\boldsymbol{\theta}}_i(m)},\quad i,j=1,...,7,\quad m=1,...,M. \tag{17}$$

and

$$\beta_{ij}^{(T_{ij})}(k,l)=0,\quad i,j=1,...,7,\quad k,l=1,...,M,\quad k\neq l. \tag{18}$$

Therefore, for the same pattern of the covariate *X*, the elements of the *DM* matrix are calculated from equation (17); whereas the elements of the *DM* matrix are equal to 0

when the β_{ij} elements are calculated between estimators of two different patterns of the covariate X (e.g. $X = x_k$ and $X = x_l$). This simplifies the expression of the *DM* matrix. For $X = x_m$, with $m = 1,...,M$, we define the DM_m matrix (sized 7×7) as

$$DM_m = \left(\beta_{ij}^{(T_{ij})}(m,m) \right) \in \mathrm{R}^{7\times 7}, \tag{19}$$

i.e., the DM_m is a matrix whose elements are the values given by equation (17) for $X = x_m$. Then the *DM* matrix is a diagonal matrix given by

$$DM = \mathrm{Diag}\left\{ DM_1, DM_2, ..., DM_M \right\}. \tag{20}$$

The proof can be seen in Appendix *B* of supplementary material.

Once the *DM* matrix has been imputed, the third phase of the *SEM* algorithm consists of calculating the asymptotic variance-covariance matrix by applying equation (14). The estimated variance-covariance matrix is not normally symmetrical due to the numerical errors committed in the calculation of the *DM* matrix. The assessment of the $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ matrix is made calculating the matrix $\Delta\hat{\Sigma}_{\hat{\boldsymbol{\theta}}} = \hat{I}_{oc}^{-1} DM \left(I - DM \right)^{-1}$, a matrix which represents the increase in the estimated variances-covariances estimated owing to the missing information. The smaller the stopping criterion (γ) of the *EM* algorithm, the more symmetrical the matrix $\Delta\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$, and therefore the more symmetrical $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ will be. Therefore, the problem of the asymmetry of $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ is solved by decreasing the stopping criterion of the *EM* algorithm [20]. Moreover, the $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ matrix may be nearly singular if the $I - DM$ matrix is nearly singular. This situation may occur when the convergence of the *EM* algorithm is extremely slow. A discussion of this problem can be seen in the manuscript of Meng and Rubin [20].

Regarding vector $\boldsymbol{\delta} = (\delta_1,...,\delta_M)$, as it is the vector of probabilities of a multinomial distribution, its variance-covariance matrix is estimated as $\hat{\Sigma}_{\hat{\boldsymbol{\delta}}} = I_1^{-1}\left(\hat{\boldsymbol{\delta}}\right) = \left[\mathrm{Diag}\left(\hat{\boldsymbol{\delta}}\right) - \hat{\boldsymbol{\delta}}^T\hat{\boldsymbol{\delta}} \right] / n$. If the covariate is binary, then $\delta_2 = 1 - \delta_1$ and $\hat{\Sigma}_{\hat{\boldsymbol{\delta}}} = \hat{V}ar\left(\hat{\delta}_1\right) = \hat{\delta}_1\hat{\delta}_2 / n$.

Once the matrixes $\sum_{\hat{\boldsymbol{\theta}}}$ and $\sum_{\hat{\boldsymbol{\delta}}}$ are estimated, the variances-covariances of $\hat{\kappa}_h(0)$ and $\hat{\kappa}_h(1)$, $h=1,2$, are estimated by applying the delta method. Let $\boldsymbol{\kappa}=\left(\kappa_1(0),\kappa_1(1),\kappa_2(0),\kappa_2(1)\right)$, taking into account the fact that from equation (13) it is verified that $\hat{\sum}_{(\hat{\boldsymbol{\delta}},\hat{\boldsymbol{\theta}})}=\text{Diag}\left\{\hat{\sum}_{\hat{\boldsymbol{\delta}}},\hat{\sum}_{\hat{\boldsymbol{\theta}}}\right\}$, the estimated variance-covariance matrix of $\hat{\boldsymbol{\kappa}}$ is

$$\hat{\sum}_{\hat{\boldsymbol{\kappa}}}=\left(\frac{\partial\boldsymbol{\kappa}}{\partial\boldsymbol{\delta}}\right)^T_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}}\hat{\sum}_{\hat{\boldsymbol{\delta}}}\left(\frac{\partial\boldsymbol{\kappa}}{\partial\boldsymbol{\delta}}\right)_{\boldsymbol{\delta}=\hat{\boldsymbol{\delta}}}+\left(\frac{\partial\boldsymbol{\kappa}}{\partial\boldsymbol{\theta}}\right)^T_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\hat{\sum}_{\hat{\boldsymbol{\theta}}}\left(\frac{\partial\boldsymbol{\kappa}}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{21}$$

Finally, the variance-covariance matrix of $\hat{\boldsymbol{\kappa}}(c)=\left(\hat{\kappa}_1(c),\hat{\kappa}_2(c)\right)$ is estimated applying the delta method again, i.e.

$$\hat{\sum}_{\hat{\boldsymbol{\kappa}}(c)}=\left(\frac{\partial\boldsymbol{\kappa}(c)}{\partial\boldsymbol{\kappa}}\right)^T_{\boldsymbol{\kappa}=\hat{\boldsymbol{\kappa}}}\hat{\sum}_{\hat{\boldsymbol{\kappa}}}\left(\frac{\partial\boldsymbol{\kappa}(c)}{\partial\boldsymbol{\kappa}}\right)_{\boldsymbol{\kappa}=\hat{\boldsymbol{\kappa}}}. \tag{22}$$

4. Simulation experiments

Monte Carlo simulation experiments were carried out to study the size and the power of the hypothesis test $H_0:\kappa_1(c)=\kappa_2(c)$ vs $H_1:\kappa_1(c)\neq\kappa_2(c)$. These experiments consisted of generating $N=10{,}000$ random samples with multinomial distributions sized $n=\{100,200,...,500,1000,2000\}$, whose probabilities were calculated from the expressions given in Appendix *A* (Partial verification: probabilities) of supplementary material. It was considered that the covariate *X* is binary $(M=2)$ with patterns x_1 and x_2, such as for example any family history of the disease (Yes or No), sex, etc., and this is a frequent situation in clinical practice. As values for δ_m we considered 0.25, 0.50 and 0.75, and for p_m we considered the values 5%, 25% and 50%, which represent a sufficient range of values to study the effect that these parameters have on the behaviour of the hypothesis test. As values of the weighted kappa coefficients $\kappa_{1m}(0)$, $\kappa_{1m}(1)$, $\kappa_{2m}(0)$ and $\kappa_{2m}(1)$ we took the values $\{0.1,...,0.9\}$. From these values, we calculated the sensitivities and the specificities of the *BDT*s in each pattern of the covariate, i.e. from the system of equations $\kappa_{hm}(0)=\left(Sp_{hm}-\bar{Q}_{hm}\right)/Q_{hm}$ and $\kappa_{hm}(1)=\left(Se_{hm}-Q_{hm}\right)/\bar{Q}_{hm}$ it holds that

$$Se_{hm} = \frac{p_m \kappa_{hm}(1) + \overline{p}_m \kappa_{hm}(0) \kappa_{hm}(1)}{\overline{p}_m \kappa_{hm}(0) + p_m \kappa_{hm}(1)} \tag{23}$$

and

$$Sp_{hm} = \frac{\overline{p}_m \kappa_{hm}(0) + p_m \kappa_{hm}(0) \kappa_{hm}(1)}{\overline{p}_m \kappa_{hm}(0) + p_m \kappa_{hm}(1)}, \tag{24}$$

with $h = 1,2$ and $m = 1,2$. Then from the values Se_{hm} and Sp_{hm} we calculated the maximum values of the factors α_{1m} and α_{0m}. As values of α_{1m} and α_{0m} we took low, intermediate and high values, i.e.

$$\alpha_{1m} = \frac{f}{Max\{Se_{1m}, Se_{2m}\}} + 1 - f$$

and

$$\alpha_{0m} = \frac{f}{Max\{(1 - Sp_{1m}), (1 - Sp_{2m})\}} + 1 - f \ ,$$

with $f = \{0.25, 0.50, 0.75\}$, and then we calculated the covariances applying these equations. As weighting indexes for $\kappa_1(c)$ and $\kappa_2(c)$ we took the values $c = \{0.1, 0.5, 0.9\}$, and then we calculated the weighted kappa coefficients $\kappa_1(c)$ and $\kappa_2(c)$ applying equations (6), considering for $\kappa_1(c)$ and $\kappa_2(c)$ only the values $\{0.2, 0.4, 0.6, 0.8\}$. Therefore, the simulation experiments were designed from the values set for the weighted kappa coefficients. Moreover, following the idea of Cicchetti [22], we considered weighted kappa coefficients with different levels of clinical significance: poor $(\kappa_i(c) < 0.40)$, fair $(0.40 \le \kappa_i(c) \le 0.59)$, good $(0.60 \le \kappa_i(c) \le 0.74)$ and excellent $(0.75 \le \kappa_i(c) \le 1)$. As verification probabilities the following scenarios were considered: I) $\lambda_{11m} = 0.50, \lambda_{10m} = \lambda_{01m} = 0.35, \lambda_{00m} = 0.05$, II) $\lambda_{11m} = 0.75, \lambda_{10m} = \lambda_{01m} = 0.50, \lambda_{00m} = 0.15$, III) $\lambda_{11m} = 0.95, \lambda_{10m} = \lambda_{01m} = 0.65, \lambda_{00m} = 0.25$ and IV) $\lambda_{11m} = \lambda_{10m} = \lambda_{01m} = \lambda_{00m} = 1$. Scenarios I, II and III correspond to situations of partial disease verification in which the verification probabilities are low, intermediate and high, respectively. Scenario IV corresponds to the case in which all of the individuals are verified with the *GS* and, consequently, it is a situation which can be

called complete verification (which is equivalent to a paired design). In this situation, the comparison of the two weighted kappa coefficients is made extending the Bloch method [7] to the case in which in all of the individuals we can observe a discrete covariate. In Appendix *C* of supplementary material, we give a brief description of this method.

The simulation experiments were designed in such a way that if in a sample it is not possible to estimate a parameter (for example if $\hat{Se}_{hm} = 0$) then that sample is discarded and another one is generated in its place until we obtain the *N* samples. As the nominal error we took $\alpha = 5\%$.

4.1. Partial Verification

For scenarios I, II and II, the following conclusions are obtained.

4.1.1 Type I error

Tables 3 shows some of the results obtained for $\kappa_1(c) = \kappa_2(c) = \{0.2, 0.8\}$ and for different values of the rest of the parameters. The covariances α_{1m} and α_{0m} and the verification probabilities have an important effect on the type I error of the hypothesis test, while the rest of the parameters do not have a clear effect upon it. In general terms, for set values of verification probabilities, the increase in the covariances α_{1m} and α_{0m} means a decrease in the type I error, regardless of the sample size. Regarding the verification probabilities, in general terms, their increase (for the same values of the covariances) means an increase in the type I error, especially when $n \leq 500$. The type I error of the test does not normally exceed too much the nominal error of 5%, fluctuating around especially when $n \geq 500 - 1000$, depending on the covariances and the verification probabilities. When the sample size is ≤ 500, the type I error is very small and therefore the test is conservative. In general terms, the hypothesis test is a conservative test for not very large sample sizes, and has a type I error that fluctuates around the nominal error when the sample size is very large, but does not normally exceed too much the nominal error and, therefore the hypothesis test does not give too many false significances. Therefore, the hypothesis test studied has the classic

behaviour of an asymptotic test, its type I error is lower than the nominal error α and from a certain sample size onwards it fluctuates around α .

4.1.2. Power

Table 4 shows some of the results obtained for the power of the hypothesis test for different values of all the parameters. As for the type I error, the covariances α_{1m} and α_{0m} and the verification probabilities have an important effect on the power of the hypothesis test, whereas the rest of the parameters do not have a clear effect on the power. In general terms, for set values of the verification probabilities, the increase in the covariances α_{1m} and α_{0m} means an increase in the power of the test for any sample size. Regarding the verification probabilities, in general terms, an increase (for the same values of the covariances) means an increase in power. In very general terms, when the difference between $\kappa_1(c)$ and $\kappa_2(c)$ is small (e.g. $|\kappa_1(c)-\kappa_2(c)|=0.2$) it is necessary to have a very large sample size, $n \geq 500-1000$ depending on the values of the covariances and the verification probabilities, so that the power is higher than 80% or 90%. When the difference between $\kappa_1(c)$ and $\kappa_2(c)$ is greater, with a moderate sample size, $n \geq 200-300$ (depending on the values of the covariances and the verification probabilities), we obtain a power higher than 80% or 90%.

Table 3. Size (in %) of the hypothesis test in the presence of partial verification.

$\kappa_1(0.5)=\kappa_2(0.5)=0.2$		
$p_1=0.10$, $p_2=0.25$, $\delta_1=0.25$, $\delta_2=0.75$		
$\kappa_{11}(0)=0.2$, $\kappa_{11}(1)=0.6$, $\kappa_{21}(0)=0.2$, $\kappa_{21}(1)=0.6$		
$\kappa_{12}(0)=0.1$, $\kappa_{12}(1)=0.3$, $\kappa_{22}(0)=0.1$, $\kappa_{22}(1)=0.3$		
$\lambda_{111}=0.50$, $\lambda_{011}=\lambda_{101}=0.35$, $\lambda_{001}=0.05$		
$\lambda_{112}=0.75$, $\lambda_{012}=\lambda_{102}=0.50$, $\lambda_{002}=0.15$		
n	$\alpha_{11}=1.11$ $\alpha_{01}=2$ $\alpha_{12}=1.13$ $\alpha_{02}=1.31$	$\alpha_{11}=1.21$ $\alpha_{01}=3$ $\alpha_{12}=1.27$ $\alpha_{02}=1.61$	$\alpha_{11}=1.32$ $\alpha_{01}=4$ $\alpha_{12}=1.40$ $\alpha_{01}=1.92$
100	0.09	0.03	0.01
200	0.54	0.49	0.19
300	1.16	1.07	0.51
400	1.88	1.54	0.88
500	2.84	2.64	2.34
1000	4.37	4.15	4.03
2000	4.85	4.51	4.12
$\lambda_{111}=0.75$, $\lambda_{011}=\lambda_{101}=0.50$, $\lambda_{001}=0.15$		
$\lambda_{112}=0.95$, $\lambda_{012}=\lambda_{102}=0.65$, $\lambda_{002}=0.25$		
n	$\alpha_{11}=1.11$ $\alpha_{01}=2$ $\alpha_{12}=1.13$ $\alpha_{02}=1.31$	$\alpha_{11}=1.21$ $\alpha_{01}=3$ $\alpha_{12}=1.27$ $\alpha_{02}=1.61$	$\alpha_{11}=1.32$ $\alpha_{01}=4$ $\alpha_{12}=1.40$ $\alpha_{01}=1.92$
100	0.13	0.05	0
200	0.99	0.78	0.22
300	1.57	1.34	0.58
400	2.28	2.01	0.94
500	2.94	2.76	2.44
1000	4.01	3.98	3.75
2000	4.59	4.31	4.18
$\kappa_1(0.9)=\kappa_2(0.9)=0.8$		
$p_1=0.10$, $p_2=0.50$, $\delta_1=0.50$, $\delta_2=0.50$		
$\kappa_{11}(0)=0.4$, $\kappa_{11}(1)=0.9$, $\kappa_{21}(0)=0.4$, $\kappa_{21}(1)=0.9$		
$\kappa_{12}(0)=0.2$, $\kappa_{12}(1)=0.8$, $\kappa_{22}(0)=0.2$, $\kappa_{22}(1)=0.8$		
$\lambda_{111}=0.50$, $\lambda_{011}=\lambda_{101}=0.35$, $\lambda_{001}=0.05$		
$\lambda_{112}=0.75$, $\lambda_{012}=\lambda_{102}=0.50$, $\lambda_{002}=0.15$		
n	$\alpha_{11}=1.02$ $\alpha_{01}=2.83$ $\alpha_{12}=1.01$ $\alpha_{02}=1.14$	$\alpha_{11}=1.04$ $\alpha_{01}=4.67$ $\alpha_{12}=1.02$ $\alpha_{02}=1.28$	$\alpha_{11}=1.07$ $\alpha_{01}=6.5$ $\alpha_{12}=1.03$ $\alpha_{01}=1.42$
100	0	0	0
200	0.14	0.05	0.01
300	0.33	0.11	0.07
400	0.57	0.29	0.14
500	1.62	0.30	1.19
1000	3.51	3.36	3.12
2000	4.87	4.52	4.15
$\lambda_{111}=0.75$, $\lambda_{011}=\lambda_{101}=0.50$, $\lambda_{001}=0.15$		
$\lambda_{112}=0.95$, $\lambda_{012}=\lambda_{102}=0.65$, $\lambda_{002}=0.25$		
n	$\alpha_{11}=1.02$ $\alpha_{01}=2.83$ $\alpha_{12}=1.01$ $\alpha_{02}=1.14$	$\alpha_{11}=1.04$ $\alpha_{01}=4.67$ $\alpha_{12}=1.02$ $\alpha_{02}=1.28$	$\alpha_{11}=1.07$ $\alpha_{01}=6.5$ $\alpha_{12}=1.03$ $\alpha_{01}=1.42$
100	0.01	0	0
200	0.18	0.07	0.03
300	0.46	0.17	0.04
400	0.59	0.32	0.19
500	1.85	0.50	0.36
1000	3.88	3.48	3.31
2000	4.79	4.76	4.27

Table 4. Power (in %) of the hypothesis test in the presence of partial verification.

$\kappa_1(0.1)=0.6$, $\kappa_2(0.1)=0.4$			
$p_1=0.05$, $p_2=0.50$, $\delta_1=0.50$, $\delta_2=0.50$			
$\kappa_{11}(0)=0.8$, $\kappa_{11}(1)=0.8$, $\kappa_{21}(0)=0.7$, $\kappa_{21}(1)=0.7$			
$\kappa_{12}(0)=0.4$, $\kappa_{12}(1)=0.4$, $\kappa_{22}(0)=0.1$, $\kappa_{22}(1)=0.1$			
$\lambda_{111}=0.50$, $\lambda_{011}=\lambda_{101}=0.35$, $\lambda_{001}=0.05$			
$\lambda_{112}=0.75$, $\lambda_{012}=\lambda_{102}=0.50$, $\lambda_{002}=0.15$			
n	$\alpha_{11}=1.06$ $\alpha_{01}=17.42$	$\alpha_{11}=1.12$ $\alpha_{01}=33.83$	$\alpha_{11}=1.17$ $\alpha_{01}=50.25$
	$\alpha_{12}=1.11$ $\alpha_{02}=1.31$	$\alpha_{12}=1.21$ $\alpha_{02}=1.61$	$\alpha_{12}=1.32$ $\alpha_{01}=1.92$
100	1.73	1.81	1.91
200	13.96	19.42	25.55
300	33.34	44.09	58.35
400	51.37	63.41	80.13
500	66.55	77.31	90.91
1000	95.42	98.70	99.83
2000	99.97	100	100
$\lambda_{111}=0.75$, $\lambda_{011}=\lambda_{101}=0.50$, $\lambda_{001}=0.15$			
$\lambda_{112}=0.95$, $\lambda_{012}=\lambda_{102}=0.65$, $\lambda_{002}=0.25$			
n	$\alpha_{11}=1.06$ $\alpha_{01}=17.42$	$\alpha_{11}=1.12$ $\alpha_{01}=33.83$	$\alpha_{11}=1.17$ $\alpha_{01}=50.25$
	$\alpha_{12}=1.11$ $\alpha_{02}=1.31$	$\alpha_{12}=1.21$ $\alpha_{02}=1.61$	$\alpha_{12}=1.32$ $\alpha_{01}=1.92$
100	2.53	2.67	2.64
200	19.47	26.72	35.84
300	42.94	53.80	70.98
400	62.11	74.36	87.73
500	74.94	86.04	95.70
1000	98.12	99.43	99.97
2000	99.98	100	100
$\kappa_1(0.9)=0.8$, $\kappa_2(0.9)=0.6$			
$p_1=0.10$, $p_2=0.25$, $\delta_1=0.25$, $\delta_2=0.75$			
$\kappa_{11}(0)=0.3$, $\kappa_{11}(1)=0.3$, $\kappa_{21}(0)=0.4$, $\kappa_{21}(1)=0.9$			
$\kappa_{12}(0)=0.7$, $\kappa_{12}(1)=0.9$, $\kappa_{22}(0)=0.3$, $\kappa_{22}(1)=0.6$			
$\lambda_{111}=0.50$, $\lambda_{011}=\lambda_{101}=0.35$, $\lambda_{001}=0.05$			
$\lambda_{112}=0.75$, $\lambda_{012}=\lambda_{102}=0.50$, $\lambda_{002}=0.15$			
n	$\alpha_{11}=1.02$ $\alpha_{01}=2.83$	$\alpha_{11}=1.04$ $\alpha_{01}=4.67$	$\alpha_{11}=1.07$ $\alpha_{01}=6.5$
	$\alpha_{12}=1.02$ $\alpha_{02}=1.64$	$\alpha_{12}=1.04$ $\alpha_{02}=2.29$	$\alpha_{12}=1.06$ $\alpha_{01}=2.93$
100	2.53	2.61	2.72
200	25.33	26.31	28.44
300	44.68	49.13	53.92
400	59.56	64.47	69.85
500	69.74	74.79	81.07
1000	92.55	95.56	98.06
2000	99.71	99.94	99.98
$\lambda_{111}=0.75$, $\lambda_{011}=\lambda_{101}=0.50$, $\lambda_{001}=0.15$			
$\lambda_{112}=0.95$, $\lambda_{012}=\lambda_{102}=0.65$, $\lambda_{002}=0.25$			
n	$\alpha_{11}=1.02$ $\alpha_{01}=2.83$	$\alpha_{11}=1.04$ $\alpha_{01}=4.67$	$\alpha_{11}=1.07$ $\alpha_{01}=6.5$
	$\alpha_{12}=1.02$ $\alpha_{02}=1.64$	$\alpha_{12}=1.04$ $\alpha_{02}=2.29$	$\alpha_{12}=1.06$ $\alpha_{01}=2.93$
100	4.68	4.22	3.84
200	28.23	30.01	32.80
300	47.85	51.89	57.03
400	61.73	66.63	72.66
500	71.64	77.18	82.88
1000	94.58	97.02	98.67
2000	99.94	99.96	99.99

4.2. Complete verification

Table 5 shows some of the results obtained for Scenario IV (complete verification). For the same sample size and the same values of the covariances, the type I errors obtained subject to complete verification are always greater than those obtained in the presence of partial verification, without exceeding the error $\alpha = 5\%$. Regarding the power subject to complete verification, this is always greater than when subject to partial verification. Subject to complete it is necessary to have a lower sample size to obtain a high power than when subject to partial verification.

In summary, partial verification involves a decrease both in the type I error and the power of the hypothesis test to compare the two weighted kappa coefficients when in all of the individuals we observe a binary covariate.

Table 5. Size and power (in %) of the hypothesis test in the presence of complete verification.

$\kappa_1(0.5) = \kappa_2(0.5) = 0.2$ $p_1 = 0.10$, $p_2 = 0.25$, $\delta_1 = 0.25$, $\delta_2 = 0.75$, $\kappa_{11}(0) = 0.2$, $\kappa_{11}(1) = 0.6$, $\kappa_{21}(0) = 0.2$, $\kappa_{21}(1) = 0.6$ $\kappa_{12}(0) = 0.1$, $\kappa_{12}(1) = 0.3$, $\kappa_{22}(0) = 0.1$, $\kappa_{22}(1) = 0.3$			
n	$\alpha_{11} = 1.11$ $\alpha_{01} = 2$ $\alpha_{12} = 1.13$ $\alpha_{02} = 1.31$	$\alpha_{11} = 1.21$ $\alpha_{01} = 3$ $\alpha_{12} = 1.27$ $\alpha_{02} = 1.61$	$\alpha_{11} = 1.32$ $\alpha_{01} = 4$ $\alpha_{12} = 1.40$ $\alpha_{01} = 1.92$
100	0.75	0.49	0.11
200	2.05	1.66	0.65
300	3.51	2.55	1.56
400	4.16	3.12	2.04
500	3.85	2.36	2.91
1000	4.06	4.35	3.62
2000	4.21	4.47	4.95
$\kappa_1(0.9) = 0.8$, $\kappa_2(0.9) = 0.6$ $p_1 = 0.10$, $p_2 = 0.25$, $\delta_1 = 0.50$, $\delta_2 = 0.50$ $\kappa_{11}(0) = 0.3$, $\kappa_{11}(1) = 0.3$, $\kappa_{21}(0) = 0.4$, $\kappa_{21}(1) = 0.9$ $\kappa_{12}(0) = 0.7$, $\kappa_{12}(1) = 0.9$, $\kappa_{22}(0) = 0.3$, $\kappa_{22}(1) = 0.6$			
n	$\alpha_{11} = 1.02$ $\alpha_{01} = 2.83$ $\alpha_{12} = 1.02$ $\alpha_{02} = 1.64$	$\alpha_{11} = 1.04$ $\alpha_{01} = 4.67$ $\alpha_{12} = 1.04$ $\alpha_{02} = 2.29$	$\alpha_{11} = 1.07$ $\alpha_{01} = 6.5$ $\alpha_{12} = 1.06$ $\alpha_{01} = 2.93$
100	11.65	15.45	18.61
200	51.61	62.12	76.93
300	77.72	85.53	96.31
400	89.25	95.35	99.64
500	95.66	98.68	99.90
1000	100	100	100
2000	100	100	100

5. Example

The model proposed in Section 3 was applied to the study by Hall et al [14] on the diagnosis of the Alzheimer's disease. Hall et al used two diagnostic tests for the diagnosis of the Alzheimer's disease: a new diagnostic test (*NDT*) based on a cognitive test applied to the patient and another test related to another person who knows the patient, and a standard diagnostic test based on a cognitive test (*CT*). As a *GS*, they used a clinical assessment (a neurological exploration, computerized tomography, neuro-psychological and laboratory tests,…). As the advanced age (≥ 75 years) of a patient is considered to be a risk factor for the Alzheimer's disease, the probability of selecting a patient for the clinical assessment was based on the results of two diagnostic tests and on the age of the patient (≥ 75 years or < 75). The study by Hall et al corresponds to a two-phase study: in the first phase, two diagnostic tests were applied to all of the patients, and in the second phase the clinical assessment (*GS*) is only applied to a subset of patients depending on the results of both diagnostic tests and on the age of the patient (covariate). Therefore, it is assumed that the verification process is *MAR*. Table 6 shows the data from the study by Hall et al, where T_1 models the result of the *NDT*, T_2 that of the *CT* and D that of the clinical assessment. In the following $m = 1$ refers to patients whose age is ≥ 75 years and $m = 2$ to the patients whose age is < 75 years.

Table 6. Data from the study of Hall et al.

	Age ≥ 75 years				
	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$V = 1$					
$D = 1$	31	5	3	1	40
$D = 0$	25	10	19	55	109
$V = 0$	22	6	65	346	429
Total	78	21	87	402	588
	Age < 75 years				
	$T_1 = 1$		$T_1 = 0$		
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	Total
$V = 1$					
$D = 1$	7	0	0	0	7
$D = 0$	10	19	6	34	69
$V = 0$	9	11	52	759	831
Total	26	30	58	793	907

In order to illustrate the model proposed in Section 3, it was considered that $c = 0.9$, a situation in which the clinician considers that a false negative is nine times more important than a false positive. Applying the *EM* algorithm taking $d_{ijm}^{(0)} = u_{ijm}/2$ and as a stopping criterion $\gamma = 10^{-12}$, the algorithm converged in 778 iterations. For the patients whose age was ≥ 75 years

$$\hat{\kappa}_{11}(0) = 0.441\ ,\ \hat{\kappa}_{11}(1) = 0.669\ ,\ \hat{\kappa}_{21}(0) = 0.245\ ,\ \hat{\kappa}_{21}(1) = 0.715$$
$$\hat{p}_1 = 0.118\ ,\ \hat{\alpha}_{11} = 1.082\ ,\ \hat{\alpha}_{01} = 3.365,$$

and for patients whose age was < 75 years

$$\hat{\kappa}_{12}(0) = 0.182\ ,\ \hat{\kappa}_{12}(1) = 1\ ,\ \hat{\kappa}_{22}(0) = 0.117\ ,\ \hat{\kappa}_{22}(1) = 1$$
$$\hat{p}_2 = 0.012\ ,\ \hat{\alpha}_{12} = 1\ ,\ \hat{\alpha}_{02} = 4.129.$$

From the data in Table 6, $\hat{\delta}_1 = 0.393$ and $\hat{\delta}_2 = 1 - 0.393 = 0.607$. The overall estimated prevalence is $\hat{p} = 0.054$. Substituting in equations (4) and (5) each parameter with its estimation

$$\hat{\kappa}_1(0) = 0.359 \text{ and } \hat{\kappa}_1(1) = 0.734,$$

and

$$\hat{\kappa}_2(0) = 0.223 \text{ and } \hat{\kappa}_2(1) = 0.787.$$

Finally, applying equation (6)

$$\hat{\kappa}_1(0.9) = 0.665 \text{ and } \hat{\kappa}_2(0.9) = 0.628.$$

Calculating the inverse Fisher information matrix of the complete data (from the last 2×4 table obtained from the application of the *EM* algorithm), applying the *SEM* algorithm taking as the stopping criterion $\gamma' = 10^{-6}$ and applying equation (21), the estimated variance-covariance matrix of $\hat{\boldsymbol{\kappa}} = \left(\hat{\kappa}_1(0), \hat{\kappa}_1(1), \hat{\kappa}_2(0), \hat{\kappa}_2(1)\right)$ is

$$\hat{\Sigma}_{\hat{\boldsymbol{\kappa}}} = \begin{pmatrix} 0.0020 & 0.0015 & 0.0009 & 0.0005 \\ 0.0015 & 0.0111 & -0.0006 & 0.0055 \\ 0.0009 & -0.0007 & 0.0013 & 0.0012 \\ 0.0005 & 0.0054 & 0.0013 & 0.0096 \end{pmatrix},$$

and applying equation (22), the estimated variance-covariance matrix of $\hat{\boldsymbol{\kappa}}(0.9)=\left(\hat{\kappa}_1(0.9),\hat{\kappa}_2(0.9)\right)$ is

$$\hat{\Sigma}_{\hat{\boldsymbol{\kappa}}(0.9)}=\begin{pmatrix}0.0070 & 0.0023\\ 0.0022 & 0.0051\end{pmatrix}.$$

Test statistic for the hypothesis test $H_0:\kappa_1(0.9)=\kappa_2(0.9)$ vs $H_1:\kappa_1(0.9)\neq\kappa_2(0.9)$ is $z=0.43$ and the two-sided p-value is 0.670, then the equality of the two weighted kappa coefficients is not rejected when $c=0.9$. Therefore, when the clinician considers that a false negative is 9 times more important than a false positive $(c=0.9)$, we do not reject the equality between the weighted kappa coefficients of the new diagnostic test and of the cognitive test in the population studied.

Table 7 shows some of the results ($\hat{\kappa}_i(c)$, test statistic and two-sided p-value) when comparing the two weighted kappa coefficients for different values of the weighting index *c*. When both diagnostic tests are going to apply as tests previous to a treatment involving some risk $(0<c<0.5)$, the weighted kappa coefficient of the *NDT* is significantly higher than that of the *CT*. The same conclusion is reached when $c=0.5$. When the diagnostic tests are going to apply as screening tests $(0.5<c<1)$ and the clinician considers that $c=0.6$ (a false negative is 1.5 times more important than a false positive), then the weighted kappa coefficient of the *NDT* is significantly higher than that of the *CT*. For the rest of the situations, screening tests with $c=\{0.7,0.8,0.9\}$, the equality of the weighted kappa coefficients of the *NDT* and the *CT* is not rejected.

Table 7. Results from the example of Hall et al for different values of the weighting index *c*.

c	$\hat{\kappa}_1(c)$	$\hat{\kappa}_1(c)$	Test statistic	Two-sided *p*-value
0.1	0.378	0.240	3.28	0.001
0.2	0.399	0.260	3.13	0.002
0.3	0.424	0.283	2.94	0.003
0.4	0.451	0.312	2.70	0.007
0.5	0.482	0.347	2.40	0.016
0.6	0.517	0.391	2.05	0.041
0.7	0.559	0.447	1.62	0.106
0.8	0.607	0.522	1.09	0.274
0.9	0.665	0.628	0.43	0.670

6. Discussion

The weighted kappa coefficient is a measure of the beyond chance agreement between the *BDT* and the *GS*, and is used to assess and compare the effectiveness of *BDTs* when considering the losses of an erroneous classification with the *BDTs*. In this article, we have studied a hypothesis test to compare the weighted kappa coefficients of two *BDTs* when in the presence of partial disease verification a discrete covariate is observed in all individuals. The hypothesis test proposed is based on the fact that the verification process with the *GS* only depends on the results of the two *BDTs* and on the covariate, and consequently that the verification process is *MAR*.

The solution of the hypothesis test of equality of the two weighted kappa coefficients was carried out by applying computational methods: the *EM* algorithm for the calculation of the estimators and the *SEM* algorithm for the calculation of the variances-covariances. The *EM* algorithm is well known and is applied in many problems with missing data. Nevertheless, the application of the *SEM* algorithm is not so frequent, and this is a method which is inherent to the *EM* algorithm as it uses many of its calculations. When applying the *SEM* algorithm to the situation analysed here with a discrete covariate, it is demonstrated that the elements of the *DM* matrix between estimators of two different patterns of the covariate are equal to 0, i.e. $\beta_{ij}(k,l)=0$, which leads to expressing the *DM* matrix as a diagonal matrix, $DM = Diag\{DM_1,...,DM_M\}$, where each DM_m matrix is the *DM* matrix in $X = x_m$. This decomposition of the *DM* matrix simplifies the calculations of the variance-covariance matrix.

An alternative method to the *SEM* algorithm for the estimation of the variance-covariance matrix consists of applying the Louis method [23]. The Louis method requires us to calculate the conditional expectation of the square of the complete-data score function and is a method which has been criticized by several authors [20, 24]. The advantage of the *SEM* algorithm is that it is a method which makes use of many of the calculations of the *EM* algorithm.

Once the model based on the *EM* and *SEM* algorithms was proposed, simulation experiments were carried out to study the size and the power of the hypothesis test when the covariate is binary. The choice of a binary covariate is justified by its practical usefulness, since in clinical studies it is frequent to have covariates of this type, such as

sex, family history, the presence or absence of a risk factor, etc. The results showed that the hypothesis test is a conservative test when the sample size is not excessively large, and the type I error fluctuates around the nominal error when the sample size is very large. The power of the test depends strongly on the covariances between the *BDTs*, on the verification probabilities and on the difference between the weighted kappa coefficients. In very general terms, when the covariances and the verification probabilities take low values, it is necessary to have a very large sample size, between 500 and 1000 depending on the difference between the two weighted kappa coefficients, so that the power is higher than 80%. When the covariances and the verification probabilities are high, with a sample size between 200 and 500, depending on the difference between the two weighted kappa coefficients, we obtain a power higher than 80%. Therefore, as the proposed hypothesis test is a conservative test with a sample size between 100 and 500, it may have a high power $(>80\%)$ with $200 \le n \le 500$, and a very high power ($>90\%$ or even close to 100%) with $n \ge 1000$, depending on the covariances and on the verification probabilities. Furthermore, for the sample sizes considered in the simulation experiments, the type I error does not exceed too much the nominal error and, therefore, the hypothesis test does not give too many false significances.

The problem of comparison of the weighted kappa coefficients in the situation posed here was solved from an unconditional point of view. That is to say, the *EM* algorithm was applied from the likelihood function based on the *n* individuals in the sample (equation (11)). Another way of solving the problem is conditioning in each one of the *M* 3×4 tables (i.e. conditioning in each value of the covariate), and applying again the *EM* and *SEM* algorithms In this situation, for the *m*-th 3×4 table $(X = x_m)$ the likelihood function based on n_m individuals is

$$l\left(\kappa_{1m}(0),\kappa_{1m}(1),\kappa_{2m}(0),\kappa_{2m}(1),p_m,\alpha_{1m},\alpha_{0m}\right) =$$
$$\sum_{i,j=0}^{1}\left(s_{ijm}+d_{ijm}\right)\log\left[P\left(T_1=i,T_2=i,D=1\middle|X=x_m\right)\right]+$$
$$\sum_{i,j=0}^{1}\left(r_{ijm}+u_{ijm}-d_{ijm}\right)\log\left[P\left(T_1=i,T_2=i,D=0\middle|X=x_m\right)\right].$$

For each one of the *M* 3×4 tables, the *EM* algorithm is applied from the previous function, and then the *SEM* algorithm is applied in each table, calculating the matrixes

$DM_1,\ldots,DM_M$. Then the variance-covariance matrixes are calculated in each one of the *M* tables. Finally, applying the delta method in a similar way to how it is applied at the end of Section 3.2, we calculate the variances-covariances of the estimators of the overall weighted kappa coefficients of the two *BDTs*. Both perspectives, unconditioned and conditioned, lead to the same solutions.

Finally, if the verification process depends on more than one discrete covariate, then we can consider a single covariate whose number of patterns would be the product of the number of patterns of each covariate [25]. For example, if in the study of Alzheimer's disease the probability of verifying with the *GS* the status of an individual conditionally depends on the results of both *BDTs* and also on sex and age (≥ 75 years or < 75), then we can consider a single covariate with four patterns (female ≥ 75 years, female < 75 years, male ≥ 75 years and male < 75 years).

Acknowledgements

This research was supported by the Spanish Ministry of Economy, Grant Number MTM2016-76938-P.

References

1. Bloch, D.A., Kraemer, H.C. (1989). 2×2 Kappa coefficients: measures of agreement or association. Biometrics, 45, 269-287.

2. Kraemer, H.C. (1992). *Evaluating Medical Tests. Objective and Quantitative Guidelines*. Newbury Park: Sage Publications.

3. Kraemer, H.C., Periyakoil, V.S., Noda, A. (2002). Kappa coefficients in medical research. Statistics in Medicine, 21, 2109-2129.

4. Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press.

5. Zhou, X.H., Obuchowski, N.A., McClish, D.K. (2011). *Statistical Methods in Diagnostic Medicine (Second Edition)*. New Jersey: John Wiley & Sons.

6. Roldán-Nofuentes, J.A., Sidaty-Regad, S.B. (2019). Recommended methods to compare the accuracy of two binary diagnostic tests subject to a paired design. Journal of Statistical Computational and Simulation, 89, 2621-2644.

7. Bloch, D.A. (1997). Comparing two diagnostic tests against the same "gold standard" in the same sample. Biometrics, 53, 73-85.

8. Rubin, D. (1976). Inference and missing data. Biometrika, 4, 73-89.

9. Zhou, X.H. (1998). Comparing accuracies of two screening tests in a two-phase study for dementia. Journal of the Royal Statistical Society, Series C Applied Statistics, 47, 135-147.

10. Harel, O., Zhou, X.H. (2007). Multiple imputation for the comparison of two screening tests in two-phase Alzheimer studies. Statistics in Medicine, 26, 2370-2388

11. Roldán-Nofuentes, J.A., Luna del Castillo, J.D. (2008). EM algorithm for comparing two binary diagnostic tests when not all the patients are verified. Journal of Statistical Computation and Simulation, 78, 19-35.

12. Roldán-Nofuentes, J.A., Luna del Castillo, J.D., Femia-Marzo, P. (2009a). Computational methods for comparing two binary diagnostic tests in the presence of partial verification of the disease. Computational Statistics, 24, 695-718.

13. Roldán-Nofuentes, J.A., Luna del Castillo, J.D. (2006). Comparing two binary diagnostic tests in the presence of verification bias. Computational Statistics and Data Analysis, 50, 1551-1564.

14. 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. and Baiyewu, O. (1996). A cross-cultural community based study of dementias: methods and performance of survey instrument. International Journal of Methods in Psychiatric Research, 6, 129-142.

15. Youden, W.J. (1950). Index for rating diagnostic tests. Cancer, 3, 32-35.

16. Roldán-Nofuentes, J.A., Luna del Castillo, J.D., Montero-Alonso, M.A. (2009b). Confidence intervals of weighted kappa coefficient of a binary diagnostic test. Communications in Statistics - Simulation and Computation, 38, 1562-1578.

17. Roldán-Nofuentes, J.A., Amro, R. (2017). Approximate confidence intervals for the weighted kappa coefficient of a binary diagnostic test subject to a case-control design. Journal of Statistical Computation and Simulation, 87, 530-545.

18. Roldán-Nofuentes, J.A., Amro, R. (2018). Combination of the weighted kappa coefficients of two binary diagnostic tests. Journal of Biopharmaceutical Statistics, 28, 909-926.

19. Berry, G., Smith, C., Macaskill, P., Irwig, L. (2002). 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. Statistics in Medicine, 21, 853-862.

20. Meng, X., Rubin, D. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. Journal of the American Statistical Association, 86, 899-909.

21. Dempster, A., Laird, N., Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1-38.

22. Cicchetti, D.V. (2001). The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements, Journal of Clinical and Experimental Neuropsychology, 23, 695-700.

23. Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B, 44, 226-233.

24. Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. Journal of the Royal Statistical Society, Series B, 51, 127-138.

25. Hosmer, D.W., Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.

Supplementary material of the manuscript:

EM and *SEM* algorithms to compare the weighted kappa coefficients of two diagnostic tests in the presence of partial verification and discrete covariates

Appendix *A*

1. Partial verification: probabilities

Let us consider $X = x_m$ and the parameters $\left(Se_{1m}, Se_{2m}, Sp_{1m}, Sp_{2m}, \alpha_{1m}, \alpha_{0m}, \lambda_{ijm}, \delta_m, p_m\right)$ defined in Section 3, then the probabilities of the *m*-th 3×4 table are:

$$f_{ijm} = P\left(V=1, D=1, T_1=i, T_2=j, X=x_m\right) =$$
$$\delta_m p_m \lambda_{ijm}\left[Se_{1m}^{i}\left(1-Se_{1m}\right)^{1-i} Se_{2m}^{j}\left(1-Se_{2m}\right)^{1-j} + \Delta_{ij} Se_{1m} Se_{2m}\left(\alpha_{1m}-1\right)\right],$$

$$g_{ijm} = P\left(V=1, D=0, T_1=i, T_2=j, X=x_m\right) =$$
$$\delta_m \overline{p}_m \lambda_{ijm}\left[Sp_{1m}^{1-i}\left(1-Sp_{1m}\right)^{i} Sp_{2m}^{1-j}\left(1-Sp_{2m}\right)^{j} + \Delta_{ij}\left(1-Sp_{1m}\right)\left(1-Sp_{2m}\right)\left(\alpha_{0m}-1\right)\right]$$

and

$$h_{ijm} = P\left(V=0, T_1=i, T_2=j, X=x_m\right) = \frac{1-\lambda_{ijm}}{\lambda_{ijm}}\left(f_{ijm}+g_{ijm}\right),$$

with $i, j = 0,1$ and where $\Delta_{ij} = 1$ if $i = j$ and $\Delta_{ij} = -1$ if $i \neq j$. It is easy to check that $\sum_{i,j=0}^{1}\left(f_{ijm}+g_{ijm}+h_{ijm}\right) = \delta_m$, and therefore it is verified that $\sum_{i,j=0}^{1}\sum_{m=1}^{M}\left(f_{ijm}+g_{ijm}+h_{ijm}\right) = 1$.

Solving the system of equations $\kappa_{hm}(0) = \left(Sp_{hm} - \overline{Q}_{hm}\right)/Q_{hm}$ and $\kappa_{hm}(1) = \left(Se_{hm} - Q_{hm}\right)/\overline{Q}_{hm}$ it is obtained that the sensitivity and the specificity of each *BDT* in $X = x_m$ are given by equations (23) and (24), i.e.

$$Se_{hm} = \frac{p_m \kappa_{hm}(1) + \overline{p}_m \kappa_{hm}(0)\kappa_{hm}(1)}{\overline{p}_m \kappa_{hm}(0) + p_m \kappa_{hm}(1)} \text{ and } Sp_{hm} = \frac{\overline{p}_m \kappa_{hm}(0) + p_m \kappa_{hm}(0)\kappa_{hm}(1)}{\overline{p}_m \kappa_{hm}(0) + p_m \kappa_{hm}(1)},$$

and substituting the parameters Se_{hm} and Sp_{hm} in the expressions of the previous probabilities $\left(f_{ijm}, f_{ijm}, h_{ijm}\right)$ and performing algebraic operations we obtain the probabilities of the multinomial distributions in terms of the weighted kappa coefficients $\kappa_{hm}(0)$ and $\kappa_{hm}(1)$, which are final expressions that have been used to generate the random samples in the simulation experiments.

2. Overall weighted kappa coefficients

For $X = x_m$ the sensitivity and specificity of each *BDT* are defined as $Se_{hm} = P\left(T_h = 1 \middle| D = 1, X = x_m\right)$ and $Sp_{hm} = P\left(T_h = 0 \middle| D = 0, X = x_m\right)$, with $h = 1,2$. Let $\delta_m = P\left(X = x_m\right)$ and $p_m = P\left(D = 1 \middle| X = x_m\right)$ defined in Section 3. Then the overall sensitivity and the overall specificity of each *BDT* are $Se_h = \dfrac{\sum_{m=1}^{M} \delta_m p_m Se_{hm}}{p}$ and $Sp_h = \dfrac{\sum_{m=1}^{M} \delta_m \overline{p}_m Sp_{hm}}{\overline{p}}$, where $p = \sum_{m=1}^{M} \delta_m p_m$ is the overall prevalence and $\overline{p} = 1 - p = \sum_{m=1}^{M} \delta_m \overline{p}_m$. Substituting the overall sensitivity, the overall specificity and the overall prevalence in equations $\kappa_h(0) = \left(Sp_h - \overline{Q}_h\right)/Q_h$ and $\kappa_h(1) = \left(Se_h - Q_h\right)/\overline{Q}_h$, with $Q_h = pSe_h + \overline{p}\left(1 - Sp_h\right)$ and $\overline{Q}_h = 1 - Q_h$, and performing algebraic operations we obtain the expressions of $\kappa_h(0)$ and $\kappa_h(1)$, in terms of $\kappa_{hm}(0)$ and $\kappa_{hm}(1)$, given in (4) and (5).

3. Complete data: probabilities

Let us consider that the *GS* was applied to all of the individuals, then $X = x_m$ we obtain the 2×4 frequency table given in Table 1 (Complete data). The probability of each one of the cells in this table is $P\left(T_1 = i, T_2 = i, X = x_m, D = 1\right) = \delta_m \phi_{ijm}$ and $P\left(T_1 = i, T_2 = i, X = x_m, D = 0\right) = \delta_m \varphi_{ijm}$, where $\phi_{ijm} = P\left(T_1 = i, T_2 = i, D = 1 \middle| X = x_m\right)$ and

$\varphi_{ijm} = P\left(T_1 = i, T_2 = i, D = 0 \middle| X = x_m\right)$. Applying the conditional dependence model of Berry et al [19] it holds that

$$\begin{aligned}\phi_{ijm} = P\left(T_1 = i \middle| D = 1, X = x_m\right) \times P\left(T_2 = j \middle| D = 1, X = x_m\right) + \Delta_{ij}\varepsilon_{1m} = \\ Se_{1m}^{i}\left(1 - Se_{1m}\right)^{1-i} Se_{2m}^{j}\left(1 - Se_{2m}\right)^{1-j} + \Delta_{ij} Se_{1m} Se_{2m}\left(\alpha_{1m} - 1\right),\end{aligned} \tag{25}$$

and

$$\begin{aligned}\varphi_{ijm} = P\left(T_1 = i \middle| D = 0, X = x_m\right) \times P\left(T_2 = j \middle| D = 0, X = x_m\right) + \Delta_{ij}\varepsilon_{0m} = \\ Sp_{1m}^{1-i}\left(1 - Sp_{1m}\right)^{i} Sp_{2m}^{1-j}\left(1 - Sp_{2m}\right)^{j} + \Delta_{ij}\left(1 - Sp_{1m}\right)\left(1 - Sp_{2m}\right)\left(\alpha_{0m} - 1\right).\end{aligned} \tag{26}$$

Substituting in these probabilities each sensitivity and specificity with (23) and (24) respectively, and performing algebraic operations we obtain the probabilities of the cells of the table of the complete data in terms of the weighted kappa coefficients.

4. Algorithm EM: estimators

Let us consider $X = x_m$, the estimators of the sensitivities and specificities in the *t*-th iteration of the *EM* algorithm are

$$\hat{Se}_{1m}^{(t)} = \frac{s_{11m} + d_{11m}^{(t)} + s_{10m} + d_{10m}^{(t)}}{s_m + d_m^{(t)}},\ \hat{Se}_{2m}^{(t)} = \frac{s_{11m} + d_{11m}^{(t)} + s_{01m} + d_{01m}^{(t)}}{s_m + d_m^{(t)}},$$

$$\hat{Sp}_{1m}^{(t)} = \frac{r_{01m} + u_{01m} - d_{01m}^{(t)} + r_{00m} + u_{00m} - d_{00m}^{(t)}}{r_m + u_m - d_m^{(t)}}$$

and

$$\hat{Sp}_{2m}^{(t)} = \frac{r_{10m} + u_{10m} - d_{10m}^{(t)} + r_{00m} + u_{00m} - d_{00m}^{(t)}}{r_m + u_m - d_m^{(t)}},$$

and the estimator of the prevalence is $\hat{p}_m^{(h)} = \left(s_m + d_m^{(h)}\right) \Big/ n_m$. Substituting the previous expressions in the equations $\kappa_{hm}(0) = \left(Sp_{hm} - \overline{Q}_{hm}\right) \Big/ Q_{hm}$ and $\kappa_{hm}(1) = \left(Se_{hm} - Q_{hm}\right) \Big/ \overline{Q}_{hm}$ and performing algebraic operations, we obtain the expressions of estimators $\hat{\kappa}_{hm}^{(t)}(0)$ and $\hat{\kappa}_{hm}^{(t)}(1)$ in the *t*-th iteration of the *EM* algorithm.

Moreover, from equations (25) and (26) it holds that $\phi_{11m} = \alpha_{1m} Se_{1m} Se_{2m}$ and that $\varphi_{00m} = \alpha_{0m}\left(1 - Sp_{1m}\right)\left(1 - Sp_{2m}\right) + Sp_{1m} + Sp_{2m} - 1$. In the *t*-th iteration of the *EM* algorithm, the estimators of these probabilities (which are estimators of multinomial proportions) are

$$\hat{\phi}_{11m}^{(t)} = \hat{\alpha}_{1m}^{(t)} \hat{Se}_{1m}^{(t)} \hat{Se}_{2m}^{(t)} = \left(s_{11m} + d_{11m}^{(t)}\right) \Big/ n_m$$

and

$$\hat{\varphi}_{00m}^{(t)} = \hat{\alpha}_{0m}^{(t)} \left(1 - \hat{Sp}_{1m}^{(t)}\right)\left(1 - \hat{Sp}_{2m}^{(t)}\right) + \hat{Sp}_{1m}^{(t)} + \hat{Sp}_{2m}^{(t)} - 1 = \left(r_{00m} + u_{00m} - d_{00m}^{(t)}\right) \Big/ n_m \,.$$

From these two equations, the expressions of $\hat{\alpha}_{1m}^{(t)}$ and $\hat{\alpha}_{0m}^{(t)}$ are obtained.

Appendix *B*

To simplify the demonstration, let us consider that the covariate *X* is binary and that it takes the values $X = 1$ and $X = 2$. The extension of the demonstration to a covariate with $M \geq 3$ patterns is analogous considering $X = x_k$ and $X = x_l$. Let us suppose that the initial values of the *EM* algorithm in each one of these two patterns of the covariate are

$$\hat{\boldsymbol{\theta}}^{(0)}(1) = \left(\hat{\kappa}_{11}^{(0)}(0), \hat{\kappa}_{11}^{(0)}(1), \hat{\kappa}_{21}^{(0)}(0), \hat{\kappa}_{21}^{(0)}(1), \hat{p}_1^{(0)}, \hat{\alpha}_{11}^{(0)}, \hat{\alpha}_{01}^{(0)}\right)$$

and

$$\hat{\boldsymbol{\theta}}^{(0)}(2) = \left(\hat{\kappa}_{12}^{(0)}(0), \hat{\kappa}_{12}^{(0)}(1), \hat{\kappa}_{22}^{(0)}(0), \hat{\kappa}_{22}^{(0)}(1), \hat{p}_2^{(0)}, \hat{\alpha}_{12}^{(0)}, \hat{\alpha}_{02}^{(0)}\right),$$

both calculated applying the *EM* algorithm taking as initial values $0 \leq d_{ij1}^{(0)} \leq u_{ij1}$ and $0 \leq d_{ij2}^{(0)} \leq u_{ij2}$, and when the final estimations obtained in *T* iterations

$$\hat{\boldsymbol{\theta}}(1) = \left(\hat{\kappa}_{11}(0), \hat{\kappa}_{11}(1), \hat{\kappa}_{21}(0), \hat{\kappa}_{21}(1), \hat{p}_1, \hat{\alpha}_{11}, \hat{\alpha}_{01}\right)$$

and

$$\hat{\boldsymbol{\theta}}(2) = \left(\hat{\kappa}_{12}(0), \hat{\kappa}_{12}(1), \hat{\kappa}_{22}(0), \hat{\kappa}_{22}(1), \hat{p}_2, \hat{\alpha}_{12}, \hat{\alpha}_{02}\right).$$

From $\hat{\boldsymbol{\theta}}(1)$ and $\hat{\boldsymbol{\theta}}(2)$ it is possible to calculate the probabilities of the last 2×4 table obtained by applying the *EM* algorithm, both for $X=1$ and for $X=2$. Let $\hat{\phi}_{ijm}^{(T)}$ and $\hat{\varphi}_{ijm}^{(T)}$, with $m=1,2$, be these probabilities which are calculated substituting in the expressions of ϕ_{ijm} and of φ_{ijm} each parameter with its final estimator. In these last two tables it is verified that

$$d_{ijm}^{(T)} = u_{ij2}\frac{\hat{\phi}_{ijm}^{(T)}}{\hat{\phi}_{ijm}^{(T)}+\hat{\varphi}_{ijm}^{(T)}} =,\ \ i,j=0,1,\ \ m=1,2\,.$$

For $X=1$ let the vectors be

$$\hat{\boldsymbol{\tau}}_1^{(0)}(1)=\left(\hat{\kappa}_{11}^{(0)}(0),\hat{\kappa}_{11}(1),\hat{\kappa}_{21}(0),\hat{\kappa}_{21}(1),\hat{p}_1,\hat{\alpha}_{11},\hat{\alpha}_{0k}\right)$$

.

.

.

$$\hat{\boldsymbol{\tau}}_7^{(0)}(1)=\left(\hat{\kappa}_{11}(0),\hat{\kappa}_{11}(1),\hat{\kappa}_{21}(0),\hat{\kappa}_{21}(1),\hat{p}_1,\hat{\alpha}_{11},\hat{\alpha}_{01}^{(0)}\right),$$

and for $X=2$

$$\hat{\boldsymbol{\tau}}_1^{(0)}(2)=\left(\hat{\kappa}_{12}^{(0)}(0),\hat{\kappa}_{12}(1),\hat{\kappa}_{22}(0),\hat{\kappa}_{22}(1),\hat{p}_2,\hat{\alpha}_{12},\hat{\alpha}_{02}\right)$$

.

.

.

$$\hat{\boldsymbol{\tau}}_7^{(0)}(2)=\left(\hat{\kappa}_{12}(0),\hat{\kappa}_{12}(1),\hat{\kappa}_{22}(0),\hat{\kappa}_{22}(1),\hat{p}_2,\hat{\alpha}_{12},\hat{\alpha}_{02}^{(0)}\right).$$

Let the vectors also be

$$\hat{\mathbf{v}}_1^{(0)}(1)=\left(\hat{\boldsymbol{\tau}}_1^{(0)}(1),\hat{\boldsymbol{\theta}}(2)\right),\ \hat{\mathbf{v}}_2^{(0)}(1)=\left(\hat{\boldsymbol{\tau}}_2^{(0)}(1),\hat{\boldsymbol{\theta}}(2)\right),...,\ \hat{\mathbf{v}}_7^{(0)}(1)=\left(\hat{\boldsymbol{\tau}}_7^{(0)}(1),\hat{\boldsymbol{\theta}}(2)\right)$$

.

.

.

$$\hat{\mathbf{v}}_1^{(0)}(2)=\left(\hat{\boldsymbol{\theta}}(1),\hat{\boldsymbol{\tau}}_1^{(0)}(2)\right),\ \hat{\mathbf{v}}_2^{(0)}(2)=\left(\hat{\boldsymbol{\theta}}(1),\hat{\boldsymbol{\tau}}_2^{(0)}(2)\right),...,\ \hat{\mathbf{v}}_7^{(0)}(2)=\left(\hat{\boldsymbol{\theta}}(1),\hat{\boldsymbol{\tau}}_7^{(0)}(2)\right).$$

The second step of the *SEM* algorithm consists of applying the first iteration of the *EM* algorithm with each one of the previous vectors. For example, using the vector $\hat{\mathbf{v}}_1^{(0)}(1)$ it holds that for $X=1$

$$d_{ij1}^{(1)} = u_{ij1} \frac{\hat{\phi}_{ij1}^{(1)}}{\hat{\phi}_{ij1}^{(1)} + \hat{\varphi}_{ij1}^{(1)}}, \quad i, j = 0,1,$$

where

$$\hat{\phi}_{111}^{(1)} = \hat{p}_1 \hat{\alpha}_{11} \hat{\xi}_{11} \hat{\kappa}_{11}(1) \hat{\xi}_{21} \hat{\kappa}_{21}(1), \; \hat{\phi}_{101}^{(1)} = \hat{p}_1 \hat{\xi}_{11} \hat{\kappa}_{11}(1) \left[1 - \hat{\alpha}_{11} \hat{\xi}_{21} \hat{\kappa}_{21}(1) \right],$$

$$\hat{\phi}_{011}^{(1)} = \hat{p}_1 \hat{\xi}_{21} \hat{\kappa}_{21}(1) \left[1 - \hat{\alpha}_{11} \hat{\xi}_{11} \hat{\kappa}_{11}(1) \right],$$

$$\hat{\phi}_{001}^{(1)} = \hat{p}_1 \left[1 - \hat{\xi}_{11} \hat{\kappa}_{11}(1) - \hat{\xi}_{21} \hat{\kappa}_{21}(1) + \hat{\alpha}_{11} \hat{\xi}_{11} \hat{\kappa}_{11}(1) \hat{\xi}_{21} \hat{\kappa}_{21}(1) \right]$$

$$\hat{\varphi}_{111}^{(1)} = \hat{\bar{p}}_1 \hat{\alpha}_{01} \left[1 - \hat{\psi}_{11} \hat{\kappa}_{11}^{(0)}(0) \right] \left[1 - \hat{\psi}_{21} \hat{\kappa}_{21}(0) \right],$$

$$\hat{\varphi}_{101}^{(1)} = \hat{\bar{p}}_1 \left[1 - \hat{\psi}_{11} \hat{\kappa}_{11}^{(0)}(0) \right] \left[1 - \hat{\alpha}_{01} + \hat{\alpha}_{01} \hat{\psi}_{21} \hat{\kappa}_{21}(0) \right],$$

$$\hat{\varphi}_{011}^{(1)} = \hat{\bar{p}}_1 \left[1 - \hat{\psi}_{21} \hat{\kappa}_{21}(0) \right] \left[1 - \hat{\alpha}_{01} + \hat{\alpha}_{01} \hat{\psi}_{11} \hat{\kappa}_{11}(0) \right]$$

and

$$\hat{\varphi}_{001}^{(1)} = \hat{\bar{p}}_1 \left\{ \hat{\psi}_{11} \hat{\kappa}_{11}^{(0)}(0) + \hat{\psi}_{21} \hat{\kappa}_{21}(0) - 1 + \hat{\alpha}_{01} \left[1 - \hat{\psi}_{11} \hat{\kappa}_{11}^{(0)} \right] \left[1 - \hat{\psi}_{21} \hat{\kappa}_{21}(0) \right] \right\},$$

with

$$\hat{\xi}_{11} = \frac{\hat{\bar{p}}_m \hat{\kappa}_{11}^{(0)}(0) + \hat{p}_1}{\hat{\bar{p}}_1 \hat{\kappa}_{11}^{(0)}(0) + \hat{p}_1 \hat{\kappa}_{11}(1)}, \quad \hat{\xi}_{21} = \frac{\hat{\bar{p}}_1 \hat{\kappa}_{21}(0) + \hat{p}_1}{\hat{\bar{p}}_1 \hat{\kappa}_{21}(0) + \hat{p}_1 \hat{\kappa}_{21}(1)}, \quad \hat{\psi}_{11} = \frac{\hat{p}_1 \hat{\kappa}_{11}(1) + \hat{\bar{p}}_1}{\hat{\bar{p}}_1 \hat{\kappa}_{11}^{(0)}(0) + \hat{p}_1 \hat{\kappa}_{11}(1)},$$

and

$$\hat{\psi}_{21} = \frac{\hat{p}_1 \hat{\kappa}_{21}(1) + \hat{\bar{p}}_1}{\hat{\bar{p}}_1 \hat{\kappa}_{21}(1) + \hat{p}_1 \hat{\kappa}_{21}(1)}.$$

For $X = 2$ we obtain

$$d_{ij2}^{(1)} = u_{ij2} \frac{\hat{\phi}_{ij2}^{(1)}}{\hat{\phi}_{ij2}^{(1)} + \hat{\varphi}_{ij2}^{(1)}}, \quad i, j = 0,1$$

where

$$\hat{\phi}_{112}^{(1)} = \hat{p}_2 \hat{\alpha}_{12} \hat{\xi}_{12} \hat{\kappa}_{12}(1) \hat{\xi}_{22} \hat{\kappa}_{22}(1),$$

$$\hat{\phi}_{102}^{(1)} = \hat{p}_1 \hat{\xi}_{12} \hat{\kappa}_{12}(1) \left[1 - \hat{\alpha}_{12} \hat{\xi}_{22} \hat{\kappa}_{22}(1) \right],$$

$$\hat{\phi}_{012}^{(1)} = \hat{p}_2 \hat{\xi}_{22} \hat{\kappa}_{22}(1)\left[1 - \hat{\alpha}_{12} \hat{\xi}_{12} \hat{\kappa}_{12}(1)\right],$$

$$\hat{\phi}_{002}^{(1)} = \hat{p}_2 \left[1 - \hat{\xi}_{12} \hat{\kappa}_{12}(1) - \hat{\xi}_{22} \hat{\kappa}_{22}(1) + \hat{\alpha}_{12} \hat{\xi}_{12} \hat{\kappa}_{12}(1) \hat{\xi}_{22} \hat{\kappa}_{22}(1)\right],$$

$$\hat{\varphi}_{112}^{(1)} = \hat{\bar{p}}_2 \hat{\alpha}_{02} \left[1 - \hat{\psi}_{12} \hat{\kappa}_{12}(0)\right]\left[1 - \hat{\psi}_{22} \hat{\kappa}_{22}(0)\right],$$

$$\hat{\varphi}_{102}^{(1)} = \hat{\bar{p}}_2 \left[1 - \hat{\psi}_{12} \hat{\kappa}_{12}(0)\right]\left[1 - \hat{\alpha}_{02} + \hat{\alpha}_{02} \hat{\psi}_{22} \hat{\kappa}_{22}(0)\right],$$

$$\hat{\varphi}_{012}^{(1)} = \hat{\bar{p}}_2 \left[1 - \hat{\psi}_{22} \hat{\kappa}_{22}(0)\right]\left[1 - \hat{\alpha}_{02} + \hat{\alpha}_{02} \hat{\psi}_{12} \hat{\kappa}_{12}(0)\right]$$

and

$$\hat{\varphi}_{002}^{(1)} = \hat{\bar{p}}_2 \left\{\hat{\psi}_{12} \hat{\kappa}_{12}(0) + \hat{\psi}_{22} \hat{\kappa}_{22}(0) - 1 + \hat{\alpha}_{02}\left[1 - \hat{\psi}_{12} \hat{\kappa}_{12}(0)\right]\left[1 - \hat{\psi}_{22} \hat{\kappa}_{22}(0)\right]\right\},$$

with

$$\hat{\xi}_{i2} = \frac{\hat{\bar{p}}_2 \hat{\kappa}_{i2}(0) + \hat{p}_2}{\hat{\bar{p}}_2 \hat{\kappa}_{i2}(0) + \hat{p}_2 \hat{\kappa}_{i2}(1)} \quad \text{and} \quad \hat{\psi}_{i2} = \frac{\hat{p}_2 \hat{\kappa}_{i2}(1) + \hat{\bar{p}}_2}{\hat{\bar{p}}_2 \hat{\kappa}_{i2}(0) + \hat{p}_2 \hat{\kappa}_{i2}(1)}, \quad i = 0,1.$$

For $X = 1$ it is verified that $\hat{\phi}_{ij1}^{(1)} \neq \hat{\phi}_{ij1}^{(T)}$ and that $\hat{\varphi}_{ij1}^{(1)} \neq \hat{\varphi}_{ij1}^{(T)}$, since in $\hat{\phi}_{ij1}^{(1)}$ and $\hat{\varphi}_{ij1}^{(1)}$ $\hat{\kappa}_{11}^{(0)}(0)$ intervenes instead of $\hat{\kappa}_{11}(0)$ (which is the value that intervenes in $\hat{\phi}_{ij1}^{(T)}$ and in $\hat{\varphi}_{ij1}^{(T)}$). It is evident that

$$d_{ij1}^{(1)} = u_{ij1} \frac{\hat{\phi}_{ij1}^{(1)}}{\hat{\phi}_{ij1}^{(1)} + \hat{\varphi}_{ij1}^{(1)}} \neq d_{ij1}^{(T)}, \quad i, j = 0,1.$$

For $X = 2$ it is verified that $\hat{\phi}_{ij2}^{(1)} = \hat{\phi}_{ij2}^{(T)}$ and that $\hat{\varphi}_{ij2}^{(1)} = \hat{\varphi}_{ij2}^{(T)}$, since in the probabilities $\hat{\phi}_{ij2}^{(1)}$ and $\hat{\varphi}_{ij2}^{(1)}$ the only estimators that intervene are the final ones obtained by applying the *EM* algorithm. Consequently, for $X = 2$ it is verified that

$$d_{ij2}^{(1)} = u_{ij2} \frac{\hat{\phi}_{ij2}^{(1)}}{\hat{\phi}_{ij2}^{(1)} + \hat{\varphi}_{ij2}^{(1)}} = u_{ij2} \frac{\hat{\phi}_{ij2}^{(T)}}{\hat{\phi}_{ij2}^{(T)} + \hat{\varphi}_{ij2}^{(T)}} = d_{ij2}^{(T)}, \quad i, j = 0,1,$$

and therefore for $X = 2$ the estimators obtained by applying the first iteration of the *EM* algorithm with $\hat{\mathbf{v}}_1^{(0)}(1)$ (step 2 of the *SEM* algorithm) are equal to the final estimators obtained by applying the *EM* algorithm, i.e. $\hat{\kappa}_{12}^{(1)}(0) = \hat{\kappa}_{12}(0)$, $\hat{\kappa}_{12}^{(1)}(1) = \hat{\kappa}_{12}(1)$, $\hat{\kappa}_{22}^{(1)}(0) = \hat{\kappa}_{22}(0)$, $\hat{\kappa}_{22}^{(1)}(1) = \hat{\kappa}_{22}(1)$, $\hat{p}_2^{(1)} = \hat{p}_2$, $\hat{\alpha}_{12}^{(1)} = \hat{\alpha}_{12}$ and $\hat{\alpha}_{02}^{(1)} = \hat{\alpha}_{02}$. Therefore, by

calculating the elements β_{ij} between estimators of the different patterns of the covariate, it holds that these elements are equal to 0. In the previous situation for $\hat{\mathbf{v}}_1^{(0)}(1)$ it holds:

$$\beta_{18}^{(0)}(1,2)=\frac{\hat{\kappa}_{12}^{(1)}(0)-\hat{\kappa}_{12}(0)}{\hat{\kappa}_{11}^{(0)}(0)-\hat{\kappa}_{11}(0)}=0,\ \beta_{19}^{(0)}(1,2)=\frac{\hat{\kappa}_{12}^{(1)}(1)-\hat{\kappa}_{12}(1)}{\hat{\kappa}_{11}^{(0)}(0)-\hat{\kappa}_{11}(0)}=0,...,$$

$$\beta_{1,14}^{(0)}(1,2)=\frac{\hat{\alpha}_{02}^{(1)}-\hat{\alpha}_{02}}{\hat{\kappa}_{11}^{(0)}(0)-\hat{\kappa}_{11}(0)}=0,$$

since all the numerators are equal to 0. Repeating the process again, the elements are calculated $\beta_{1j}^{(1)}(1,2)$, with $j=8,...,14$, and again it is obtained that $\beta_{1j}^{(1)}(1,2)=0$, and the process stops because the difference between the two consecutive iterations is 0 $(<\gamma')$. The demonstration is identical to the rest of the vectors $\hat{\mathbf{v}}_i^{(0)}(1)$. For the same covariate pattern, the elements β_{ij} are calculated by applying equation (17).

For a binary covariate, two *DM* matrixes are obtained: DM_1 and DM_2. DM_1 is the matrix between the estimators in $X=1$ and DM_2 is the matrix between the estimators in $X=2$. Finally, the DM matrix is obtained as

$$DM=Diag\{DM_1,DM_2\},$$

since $\beta_{ij}(1,2)=0$ with $i=1,...,7$ and $j=8,...,14$, and $\beta_{ij}(2,1)=0$ with $i=8,...,14$ and $j=1,...,7$.

Appendix *C*

When all of the individuals are verified with the *GS* and in all of them we observe a discrete covariate, the comparison of the two weighted kappa coefficients is solved extending the Bloch method [7]. In this situation, for $X=x_m$ Table 1 is obtained with $u_{ijm}=u=0$. The method to solve the hypothesis test is:

1). In each pattern of the covariate the Bloch method is applied to estimate the weighted kappa coefficients $\kappa_{hm}(0)$ and $\kappa_{hm}(1)$. The parameters δ_m and p_m are estimated as $\hat{\delta}_m=n_m/n$ and $\hat{p}_m=s_m/n_m$, and $\hat{p}=\sum_{m=1}^{M}\hat{\delta}_m\hat{p}_m$.

2). Calculate $\hat{\kappa}_h(0)$ and $\hat{\kappa}_h(1)$ substituting in equations (4) and (5) the parameters with their estimators, and then calculate $\hat{\kappa}_1(c)$ and $\hat{\kappa}_2(c)$ applying equation (6).

3). Estimate the variances-covariances inverting the Fisher information matrix of the likelihood function of the complete data (equation (13) with $d_{ijm} = 0$).

4). Estimate the variances-covariances of $\hat{\kappa}_1(c)$ and $\hat{\kappa}_2(c)$ applying the delta method (equations (21) and (22)), and then solve the hypothesis test calculating the test statistic (equation (7)).