REVSTAT – Statistical Journal
Volume 16, Number 4, October 2018, 405–428

COMPARISON OF THE AVERAGE KAPPA COEFFICIENTS OF BINARY DIAGNOSTIC TESTS DONE ON THE SAME SUBJECTS

Authors: José Antonio Roldán-Nofuentes
– Statistics (Biostatistics), School of Medicine, University of Granada, Granada, Spain
jaroldan@ugr.es

Carmen Olvera-Porcel
– Statistics (Biostatistics), School of Medicine, University of Granada, Granada, Spain mcolvera@ugr.es

Received: October 2015 Revised: September 2016 Accepted: September 2016

Abstract:

- The average kappa coefficient of a binary diagnostic test is a chance corrected index between the binary diagnostic test and the gold standard, and it depends on the sensitivity and the specificity of the diagnostic test and on the disease prevalence. In this article, several hypothesis tests are studied to compare the average kappa coefficients of two (o more) binary diagnostic tests done on the same subjects. Simulation experiments were carried out to study the type I errors and the powers of the hypothesis tests studied. A program in R was written to solve the problem studied and it can be freely downloaded from the Internet. The results were applied to a real example on the diagnosis of coronary disease.

Key-Words:

- *average kappa coefficient; binary diagnostic test; weighted kappa coefficient.*

AMS Subject Classification:

- 62P10, 62H15.

1. INTRODUCTION

The fundamental parameters to assess and compare the performance of binary diagnostic tests are sensitivity and specificity. Sensitivity is the probability of the result of the binary diagnostic test (BDT) being positive when the individual has the disease, and specificity is the probability of the result of the BDT being negative when the individual does not have the disease. Both parameters depend only on the specific characteristics of the BDT, i.e. the intrinsic properties (physical, biological, chemical, etc.) of the BDT. When comparing two BDTs in paired designs, i.e. when the two BDTs and the gold standard (GS) are applied to all of the individuals in a random sample, the comparison of the two sensitivities (specificities) is made conditioning in the total of individuals with the disease (without the disease) and applying the exact test to compare two binomial proportions or its asymptotic version (McNemar's test).

When considering the losses associated with an erroneous classification with the BDT, the parameter that is used to assess the BDT is the weighted kappa coefficient [1,2,3]. The weighted kappa coefficient depends on the sensitivity and the specificity of the BDT, on the disease prevalence in the population studied and on the relative loss between the false positives and the false negatives (weighting index). The value of the weighting index is set by the clinical laboratory researcher based on his or her knowledge about the problem to be solved. Bloch [4] studied the comparison of the weighted kappa coefficients of two BDTs in relation to the same GS subject to a paired design.

The problem posed by the weighted kappa coefficient as a measure to assess and compare the performance of BDTs is the allocation of the value to the weighting index, since the clinical laboratory researcher does not have enough knowledge about the problem to be able to allocate that value, and two clinicians might even allocate different values to that index in the same problem. In order to solve this problem, Roldán-Nofuentes and Olvera-Porcel [5] defined a new parameter called the average kappa coefficient. The average kappa coefficient of the BDT depends on the sensitivity and the specificity of the BDT and on the disease prevalence, and does not depend on the weighting index. This new parameter has properties that make it valid to assess and compare BDTs. In this study, several hypothesis tests are studied to compare the average kappa coefficients of two BDTs in a paired design. In Section 2, the weighted kappa coefficient and the average kappa coefficient are explained. In Section 3, we present several asymptotic hypothesis tests to compare the average kappa coefficients of two BDTs subject to paired design. In Section 4, simulation experiments are carried out to study the type I errors and the powers of the hypothesis tests presented in Section 3. In Section 5, we study the situation in which more than two BDTs are compared. In Section 6, we present a program written in R which allows us to solve the problem posed. In Section 7, the results obtained are applied to a real example, and in Section 8 the results obtained are discussed.

2. WEIGHTED KAPPA COEFFICIENT AND AVERAGE KAPPA COEFFICIENT

Let L and L' be the losses associated with an erroneous classification with the BDT: L is the loss that occurs when for an individual the BDT is negative and the GS is positive, and L' is the loss that occurs when the BDT is positive and the GS is negative. Losses L and L' are zero when an individual (with or without the disease) is classified correctly with the BDT. The weighted kappa coefficient of a BDT is [1,2,3,4,6]

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

where Se is the sensitivity of the BDT, Sp the specificity, p the disease prevalence, $q = 1 - p$, $Q = pSe + q\,(1 - Sp)$ and $c = L/(L + L')$ is the weighting index. When loss L is equal to zero then $c = 0$, and the weighted kappa coefficient is

$$\kappa(0) = \frac{Sp - (1-Q)}{Q},$$

and when loss L' is equal to zero then $c = 1$, and the weighted kappa coefficient is

$$\kappa(1) = \frac{Se - Q}{1 - Q}.$$

The weighted kappa coefficient can also be written as

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

and therefore it is a weighted mean of $\kappa(0)$ and $\kappa(1)$. Weighting index c varies between 0 and 1 and represents the relative loss between the false positives and the false negatives. In practice the weighting index c is unknown, but its values can be assumed according to the objective for which the diagnostic test is going to be used. If the diagnostic test is going to be used as a previous step for a risky treatment (e.g. surgery), there is more concern about the false positives and the c index is lower than 0.5; if the diagnostic test is going to be used as a screening test, there is more concern about the false negatives and the c index is greater than 0.5; and the c index is 0.5 when the diagnostic test is used for a simple diagnosis. If $L = L'$, then $c = 0.5$ and $\kappa(0.5)$ is called the Cohen kappa coefficient; if $L > L'$, then $0.5 < c < 1$, and if $L' > L$ then $0 < c < 0.5$. The properties of the weighted kappa coefficient can be seen in the manuscript of Kraemer *et al.* [3] and in that of Roldán-Nofuentes *et al.* [6]. The problem posed by the weighted kappa coefficient is the allocation of a value to the weighting index. Allocating values 0 or 1 means that one of the losses is equal to zero, which is not realistic. In practice, the allocation is made based on the knowledge that the clinical laboratory researcher has about the problem that is being analyzed. This procedure can lead to some disagreement, since two different clinicians may allocate different values and their conclusions may not be the same.

In order to solve this problem of the allocation of values to the weighting index, Roldán-Nofuentes and Olvera-Porcel [5] defined a new parameter: the average kappa coefficient. The average kappa coefficient is a measure of the weighted kappa coefficients, and only depends on the sensitivity and the specificity of the BDT and the disease prevalence, and does not depend on the weighting index. If the clinical laboratory researcher considers that the loss associated with a false positive is greater than the loss associated with a false negative, $L' > L$ and $0 < c < 0.5$, the average kappa coefficient is

$$\kappa_1 = \frac{1}{0.5}\int_0^{0.5} \kappa\,(c)\,dc = \begin{cases} \frac{2\kappa(0)\kappa(1)}{\kappa(0)-\kappa(1)} \ln\left[\frac{\kappa(0)+\kappa(1)}{2\kappa(1)}\right], & p \neq Q \\ \quad Se + Sp - 1, & p = Q, \end{cases} \tag{2.1}$$

i.e. the average kappa coefficient κ_1 is the average value of $\kappa\,(c)$ when $0 < c < 0.5$. If the clinical laboratory researcher considers that the loss associated with a false negative is greater than the loss associated with a false positive, $L > L'$ and $0.5 < c < 1$, the average kappa coefficient is

$$\kappa_2 = \frac{1}{0.5}\int_{0.5}^{1} \kappa\,(c)\,dc = \begin{cases} \frac{2\kappa(0)\kappa(1)}{\kappa(0)-\kappa(1)} \ln\left[\frac{2\kappa(0)}{\kappa(0)+\kappa(1)}\right], & p \neq Q \\ \quad Se + Sp - 1, & p = Q. \end{cases} \tag{2.2}$$

As the weighted kappa coefficient is a measure of the beyond-chance agreement between the BDT and the GS, then κ_1 and κ_2 are measures of the average beyond-chance agreement between the BDT and the GS. The properties of κ_1 and κ_2 can be seen in the manuscript by Roldán-Nofuentes and Olvera-Porcel [5], and they are parameters that allow us to assess and compare the performance of BDTs. The comparison of the average kappa coefficients of two BDTs subject to paired design is now studied.

3. COMPARISON OF TWO AVERAGE KAPPA COEFFICIENTS

Let us consider two BDTs that are compared in relation to the same GS. The frequencies obtained applying the two BDTs and the GS to a sample of n individuals and theoretical probabilities are shown in Table 1, where the variable T_i models the result of the i-th BDT ($T_i = 1$ when the result is positive and $T_i = 0$ when it is negative) and the variable D models the result of the GS ($D = 1$ when the individual has the disease and $D = 0$ when this is not the case). If the clinical laboratory researcher assumes a value for the weighting index c, Bloch [4] has studied the comparison of the weighted kappa coefficients of two BDTs subject to a paired design. Using the notation in Table 1, the estimators of the weighted kappa coefficients deduced by Bloch [4] are

$$\hat{\kappa}_1\,(c) = \frac{(s_{11}+s_{10})\,(r_{01}+r_{00}) - (s_{01}+s_{00})\,(r_{10}+r_{11})}{sc\sum\limits_{k=0}^{1}(s_{0k}+r_{0k}) + r\,(1-c)\sum\limits_{k=0}^{1}(s_{1k}+r_{1k})}$$

and

$$\hat{\kappa}_2(c) = \frac{(s_{11}+s_{01})(r_{10}+r_{00})-(s_{10}+s_{00})(r_{01}+r_{11})}{sc\sum\limits_{h=0}^{1}(s_{h0}+r_{h0})+r(1-c)\sum\limits_{h=0}^{1}(s_{h1}+r_{h1})},$$

and the statistic for $H_0 : \kappa_1(c) = \kappa_2(c)$ vs $H_1 : \kappa_1(c) \neq \kappa_2(c)$ is

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

where the expressions of the variances and the covariance have been obtained by Bloch [4] applying the delta method.

Table 1: Observed frequencies and probabilities subject to paired design.

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

Probabilities					
	$T_1=1$		$T_1=0$		Total
	$T_2=1$	$T_2=0$	$T_2=1$	$T_2=0$	
$D=1$	p_{11}	p_{10}	p_{01}	p_{00}	p
$D=0$	q_{11}	q_{10}	q_{01}	q_{00}	q
Total	$p_{11}+q_{11}$	$p_{10}+q_{10}$	$p_{01}+q_{01}$	$p_{00}+q_{00}$	1

We then study the comparison of the average kappa coefficients of the two BDTs. Firstly, we study the comparison of the two average kappa coefficients when the clinical laboratory researcher considers that $L' > L$ $(0 < c < 0.5)$ and after when $L > L'$ $(0.5 < c < 1)$.

When $L' > L$ the hypothesis test to compare the two average kappa coefficients is $H_0 : \kappa_{11} = \kappa_{21}$ vs $H_1 : \kappa_{11} \neq \kappa_{21}$, where κ_{i1} is the average kappa coefficient of the i-th BDT when the clinical laboratory researcher considers that $L' > L$. In terms of the probabilities in Table 1, the sensitivity and the specificity of each BDT are written as $Se_1 = (p_{10}+p_{11})/p$, $Sp_1 = (q_{00}+q_{01})/q$, $Se_2 = (p_{01}+p_{11})/p$ and $Sp_2 = (q_{00}+q_{10})/q$, where $p = \sum\limits_{ij} p_{ij}$ is the disease prevalence and $q = 1-p = \sum\limits_{ij} q_{ij}$. Replacing in equation (2.1) each parameter with its expression, the average kappa coefficient κ_{11} is written as

$$\begin{aligned}
\kappa_{11} &= \frac{2\kappa_1(0)\,\kappa_1(1)}{\kappa_1(0)-\kappa_1(1)} \ln\left\{\frac{\kappa_1(0)+\kappa_1(1)}{2\kappa_1(1)}\right\} \\
&= 2\left(\frac{\sum\limits_{j=0}^{1}(p_{0j}+q_{0j})}{\frac{1}{p}\sum\limits_{j=0}^{1}p_{1j}-\sum\limits_{j=0}^{1}(p_{1j}+q_{1j})} - \frac{\sum\limits_{j=0}^{1}(p_{1j}+q_{1j})}{\frac{1}{q}\sum\limits_{j=0}^{1}q_{0j}-\sum\limits_{j=0}^{1}(p_{0j}+q_{0j})}\right)^{-1} \\
&\quad \times \ln\left[\frac{1}{2}\left(\frac{\left(\sum\limits_{j=0}^{1}(p_{0j}+q_{0j})\right)\left(\frac{1}{q}\sum\limits_{j=0}^{1}q_{0j}-\sum\limits_{j=0}^{1}(p_{0j}+q_{0j})\right)}{\left(\sum\limits_{j=0}^{1}(p_{1j}+q_{1j})\right)\left(\frac{1}{p}\sum\limits_{j=0}^{1}p_{1j}-\sum\limits_{j=0}^{1}(p_{1j}+q_{1j})\right)}+1\right)\right]
\end{aligned}$$

when $p \neq Q_1$ and $\kappa_{11} = \frac{1}{p}\sum\limits_{j=0}^{1}p_{1j}+\frac{1}{q}\sum\limits_{j=0}^{1}q_{0j}-1$ when $p = Q_1$. Regarding κ_{21}, its expression is

$$\begin{aligned}
\kappa_{21} &= \frac{2\kappa_2(0)\,\kappa_2(1)}{\kappa_2(0)-\kappa_2(1)} \ln\left\{\frac{\kappa_2(0)+\kappa_2(1)}{2\kappa_2(1)}\right\} \\
&= 2\left(\frac{\sum\limits_{i=0}^{1}(p_{i0}+q_{i0})}{\frac{1}{p}\sum\limits_{i=0}^{1}p_{i1}-\sum\limits_{i=0}^{1}(p_{i1}+q_{i1})} - \frac{\sum\limits_{i=0}^{1}(p_{i1}+q_{i1})}{\frac{1}{q}\sum\limits_{i=0}^{1}q_{i0}-\sum\limits_{i=0}^{1}(p_{i0}+q_{i0})}\right)^{-1} \\
&\quad \times \ln\left[\frac{1}{2}\left(\frac{\left(\sum\limits_{i=0}^{1}(p_{i0}+q_{i0})\right)\left(\frac{1}{q}\sum\limits_{i=0}^{1}q_{i0}-\sum\limits_{i=0}^{1}(p_{i0}+q_{i0})\right)}{\left(\sum\limits_{i=0}^{1}(p_{i1}+q_{i1})\right)\left(\frac{1}{p}\sum\limits_{i=0}^{1}p_{i1}-\sum\limits_{i=0}^{1}(p_{i1}+q_{i1})\right)}+1\right)\right]
\end{aligned}$$

when $p \neq Q_2$ and $\kappa_{21} = \frac{1}{p}\sum\limits_{j=0}^{1}p_{1j}+\frac{1}{q}\sum\limits_{j=0}^{1}q_{0j}-1$ when $p = Q_2$. As the probabilities p_{ij} and q_{ij} are probabilities of a multinomial distribution, their estimators are $\hat{p}_{ij} = s_{ij}/n$ and $\hat{q}_{ij} = r_{ij}/n$. Therefore, the estimator of κ_{11} is

$$\begin{aligned}
\hat{\kappa}_{11} &= \frac{2\left\{(s_{10}+s_{11})(r_{00}+r_{01})-(s_{00}+s_{01})(r_{10}+r_{11})\right\}}{n\left(\sum\limits_{j=0}^{1}(s_{0j}-r_{1j})\right)} \\
&\quad \times \ln\left[\frac{1}{2}\left(\frac{s\sum\limits_{j=0}^{1}(s_{0j}+r_{0j})}{r\sum\limits_{j=0}^{1}(s_{1j}+r_{1j})}+1\right)\right]
\end{aligned}$$

when $\hat{p} \neq \hat{Q}_1$, i.e. when $s_{01}+s_{00} \neq r_{10}+r_{11}$, and

$$\hat{\kappa}_{11} = \frac{(s_{10}+s_{11})(r_{00}+r_{01})-(s_{00}+s_{01})(r_{10}+r_{11})}{sr}$$

when $\hat{p} = \hat{Q}_1$, i.e. when $s_{01} + s_{00} = r_{10} + r_{11}$. Regarding the estimator of κ_{21}, its expression is

$$\hat{\kappa}_{21} = \frac{2\left\{\left(s_{01}+s_{11}\right)\left(r_{00}+r_{10}\right)-\left(s_{00}+s_{10}\right)\left(r_{01}+r_{11}\right)\right\}}{n\left(\sum\limits_{i=0}^{1}\left(s_{i0}-r_{i1}\right)\right)} \times \ln\left[\frac{1}{2}\left(\frac{s\sum\limits_{i=0}^{1}\left(s_{i0}+r_{i0}\right)}{r\sum\limits_{i=0}^{1}\left(s_{i1}+r_{i1}\right)}+1\right)\right]$$

when $\hat{p} \neq \hat{Q}_2$, i.e. $s_{10} + s_{00} \neq r_{01} + r_{11}$, and

$$\hat{\kappa}_{21} = \frac{\left(s_{01}+s_{11}\right)\left(r_{00}+r_{10}\right)-\left(s_{00}+s_{10}\right)\left(r_{01}+r_{11}\right)}{sr}$$

when $\hat{p} = \hat{Q}_2$, i.e. $s_{10} + s_{00} = r_{01} + r_{11}$. Applying the delta method, the asymptotic variance-covariance matrix of $\hat{\kappa}_{11}$ and $\hat{\kappa}_{21}$ is

$$\sum\nolimits_{\hat{\boldsymbol{\kappa}}_1} = \left(\frac{\partial \boldsymbol{\kappa}_1}{\partial \boldsymbol{\pi}}\right) \sum\nolimits_{\hat{\boldsymbol{\pi}}} \left(\frac{\partial \boldsymbol{\kappa}_1}{\partial \boldsymbol{\pi}}\right)^T,$$

where $\boldsymbol{\kappa}_1 = \left(\kappa_{11}, \kappa_{21}\right)^T$, $\boldsymbol{\pi} = \left(p_{11}, p_{10}, p_{01}, p_{00}, q_{11}, q_{10}, q_{01}, q_{00}\right)^T$ and

$$\sum\nolimits_{\hat{\boldsymbol{\pi}}} = \frac{\text{Diag}\left(\boldsymbol{\pi}\right) - \boldsymbol{\pi}\boldsymbol{\pi}^T}{n}$$

is the variance-covariance matrix of the probabilities in Table 1. Replacing in the expression of $\sum_{\hat{\boldsymbol{\kappa}}_1}$ each parameter with its estimator, we obtain the expressions of the estimated asymptotic variances-covariances of $\hat{\boldsymbol{\kappa}}_1$. These expressions are not presented here as they are very long and complicated (they were calculated using the R programming approach created to solve this hypothesis test). Finally, the statistic to contrast the equality of the average kappa coefficients when $L' > L$ is

$$z = \frac{\hat{\kappa}_{11} - \hat{\kappa}_{21}}{\sqrt{\hat{V}ar\left(\hat{\kappa}_{11}\right) + \hat{V}ar\left(\hat{\kappa}_{21}\right) - 2\hat{C}ov\left(\hat{\kappa}_{11}, \hat{\kappa}_{21}\right)}} \xrightarrow[n\to\infty]{} N\left(0,1\right).$$

Furthermore, an asymptotic confidence interval for the difference of the average kappa coefficients is

$$\kappa_{11} - \kappa_{21} \in \hat{\kappa}_{11} - \hat{\kappa}_{21} \pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\hat{\kappa}_{11}\right) + \hat{V}ar\left(\hat{\kappa}_{21}\right) - 2\hat{C}ov\left(\hat{\kappa}_{11}, \hat{\kappa}_{21}\right)},$$

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

If the clinical laboratory researcher considers that $L > L'$, and therefore that $0.5 < c < 1$, the hypothesis test to compare the two average kappa coefficients is $H_0 : \kappa_{12} = \kappa_{22}$ vs $H_1 : \kappa_{12} \neq \kappa_{22}$, where κ_{i2} is the average kappa coefficient of the i-th BDT when $L > L'$. The process to solve this hypothesis test is similar to the previous case, and the statistic is

$$z = \frac{\hat{\kappa}_{12} - \hat{\kappa}_{22}}{\sqrt{\hat{V}ar\left(\hat{\kappa}_{12}\right) + \hat{V}ar\left(\hat{\kappa}_{22}\right) - 2\hat{C}ov\left(\hat{\kappa}_{12}, \hat{\kappa}_{22}\right)}} \xrightarrow[n\to\infty]{} N\left(0,1\right).$$

Replacing in equation (2.2) each parameter with its expression, the estimators of κ_{12} is

$$\begin{aligned}\hat{\kappa}_{12} &= \frac{2\left\{\left(s_{10}+s_{11}\right)\left(r_{00}+r_{01}\right)-\left(s_{00}+s_{01}\right)\left(r_{10}+r_{11}\right)\right\}}{n\left(\sum\limits_{j=0}^{1}\left(s_{0j}-r_{1j}\right)\right)} \\ &\times \ln\left[2\frac{s\sum\limits_{j=0}^{1}\left(s_{0j}+r_{0j}\right)}{s\sum\limits_{j=0}^{1}\left(s_{0j}+r_{0j}\right)+r\sum\limits_{j=0}^{1}\left(s_{1j}+r_{1j}\right)}\right]\end{aligned}$$

when $\hat{p} \neq \hat{Q}_1$, i.e. $s_{01}+s_{00} \neq r_{10}+r_{11}$, and

$$\hat{\kappa}_{12} = \frac{\left(s_{10}+s_{11}\right)\left(r_{00}+r_{01}\right)-\left(s_{00}+s_{01}\right)\left(r_{10}+r_{11}\right)}{sr}$$

when $\hat{p} = \hat{Q}_1$, i.e. $s_{01}+s_{00} = r_{10}+r_{11}$. Regarding κ_{22}, it holds that

$$\begin{aligned}\hat{\kappa}_{22} &= \frac{2\left\{\left(s_{01}+s_{11}\right)\left(r_{00}+r_{10}\right)-\left(s_{00}+s_{10}\right)\left(r_{01}+r_{11}\right)\right\}}{n\left(\sum\limits_{i=0}^{1}\left(s_{i0}-r_{i1}\right)\right)} \\ &\times \ln\left[2\frac{s\sum\limits_{i=0}^{1}\left(s_{i0}+r_{i0}\right)}{s\sum\limits_{i=0}^{1}\left(s_{i0}+r_{i0}\right)+r\sum\limits_{i=0}^{1}\left(s_{i1}+r_{i1}\right)}\right]\end{aligned}$$

when $\hat{p} \neq \hat{Q}_2$, i.e. $s_{10}+s_{00} \neq r_{01}+r_{11}$, and

$$\hat{\kappa}_{22} = \frac{\left(s_{01}+s_{11}\right)\left(r_{00}+r_{10}\right)-\left(s_{00}+s_{10}\right)\left(r_{01}+r_{11}\right)}{sr}$$

when $\hat{p} = \hat{Q}_2$, i.e. $s_{10}+s_{00} = r_{01}+r_{11}$. The asymptotic variance-covariance matrix is estimated in a similar way to the previous case. Moreover, an asymptotic confidence interval for the difference of the average kappa coefficients is

$$\kappa_{12}-\kappa_{22} \in \hat{\kappa}_{12}-\hat{\kappa}_{22} \pm z_{1-\alpha/2}\sqrt{\hat{V}ar\left(\hat{\kappa}_{12}\right)+\hat{V}ar\left(\hat{\kappa}_{22}\right)-2\hat{C}ov\left(\hat{\kappa}_{12},\hat{\kappa}_{22}\right)}.$$

The comparison of the average kappa coefficients can also be made using transformations, such as the logarithm and the logit transformations. In this case, the hypothesis test is $H_0 : F\left(\kappa_{1k}\right) = F\left(\kappa_{2k}\right)$ vs $H_1 : F\left(\kappa_{1k}\right) \neq F\left(\kappa_{2k}\right)$, where F is the logarithm or the logit respectively. The problem is solved in a similar way to in the previous case. These transformations aim to improve the convergence of the distribution of the estimators to the normal distribution.

4. SIMULATION EXPERIMENTS

Simulation experiments were carried out to study the type I errors and the powers of the hypothesis tests $H_0 : \kappa_{1k} = \kappa_{2k}$ and $H_0 : F(\kappa_{1k}) = F(\kappa_{2k})$. Therefore, 5000 random samples of multinomial distributions were generated with sizes of 100, 200, 300, 400, 500, 1000 and 2000, which are sizes in a wide range to show the behaviour of the hypothesis tests. The probabilities of the multinomial distributions were calculated using the conditional dependence model proposed by Vacek [7], i.e.

$$p_{ij} = P(T_1 = i, T_2 = j \,|D = 1) = P(T_1 = i \,|D = 1) \times P(T_2 = j \,|D = 1) + \delta_{ij}\varepsilon_1$$

and

$$q_{ij} = P(T_1 = i, T_2 = j \,|D = 0) = P(T_1 = i \,|D = 0) \times P(T_2 = j \,|D = 0) + \delta_{ij}\varepsilon_0,$$

where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = -1$ if $i \neq j$, and ε_i is the covariance between the two BDTs when $D = i$. Vacek [7] demonstrated that

$$0 \leq \varepsilon_1 \leq Min\{Se_1(1 - Se_2), (1 - Se_1)Se_2\}$$

and that

$$0 \leq \varepsilon_0 \leq Min\{Sp_1(1 - Sp_2), (1 - Sp_1)Sp_2\}.$$

If $\varepsilon_1 = \varepsilon_0 = 0$ then the two BDTs are conditionally independent on the disease status. In practice, the assumption of conditional independence is not very realistic and therefore $\varepsilon_1 > 0$ and/or $\varepsilon_0 > 0$.

The simulation experiments were designed based on the equations of the average kappa coefficients of the two BDTs, i.e.

$$\kappa_{i1} = \frac{2\kappa_i(0)\kappa_i(1)}{\kappa_i(0) - \kappa_i(1)} \ln\left[\frac{\kappa_i(0) + \kappa_i(1)}{2\kappa_i(1)}\right] \tag{4.1}$$

and

$$\kappa_{i2} = \frac{2\kappa_i(0)\kappa_i(1)}{\kappa_i(0) - \kappa_i(1)} \ln\left[\frac{2\kappa_i(0)}{\kappa_i(0) + \kappa_i(1)}\right]. \tag{4.2}$$

As the disease prevalence, we took the values 5%, 10%, 30% and 50%. The first two values correspond to a scenario with low prevalence and the last two with a high disease prevalence, and they are a range of values that allow us to study the effect of the prevalence on the behaviour of each hypothesis test. Regarding the average kappa coefficients we took the values 0.2, 0.4, 0.6 and 0.8. Therefore, following the idea of Cicchetti [8] we took values of average kappa coefficients with different levels of significance: poor (< 0.40), fair ($0.40 - 0.59$), good ($0.60 - 0.74$) and excellent ($0.75 - 1$). Once the values for the prevalence and the average kappa coefficient were set, using the Newton–Raphson method, the system made up of equations (4.1) and (4.2) was solved to thus obtain the values of $\kappa_i(0)$ and $\kappa_i(1)$, only considering those values whose solutions are between 0 and 1. Finally, in order to obtain the values of the sensitivity and the specificity of each BDT (Se_i and Sp_i) the system made up of the equations $\kappa_i(0) = \{Sp_i - (1 - Q_i)\}/Q_i$ and $\kappa_i(1) = (Se_i - Q_i)/(1 - Q_i)$ was solved. Once the values for Se_i and Sp_i were obtained, the maximum values for the covariances ε_1 and ε_0 were calculated. Finally, the probabilities of the multinomial distributions were calculated based on the model proposed by Vacek [7]. Furthermore, the samples were generated in such a way that in all cases it was possible to estimate all of the parameters and their variances-covariances. In all of the study, we took as the nominal error $\alpha = 5\%$. In the tables with results, Test 1 refers to the hypothesis test without logarithmic transformation and Test 2 refers to the hypothesis test with logarithmic transformation. The results with the logit transformation are not shown as they are very similar to those obtained with the logarithmic transformation.

4.1. Type I errors

In Table 2, we can see some of the results obtained for the type I errors of the hypothesis tests $H_0 : \kappa_{11} = \kappa_{21}$ (Test 1) and $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$ (Test 2), i.e. when comparing the average kappa coefficients considering that $L' > L$. In Table 3, we can see some results for the type I errors of the hypothesis test $H_0 : \kappa_{12} = \kappa_{22}$ (Test 1) and $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ (Test 2), i.e. when comparing the average kappa coefficients considering that $L > L'$. In these tables we can see the values of the sensitivities, specificities, prevalence and covariances with which the multinomial samples were generated.

When $L' > L$ (Table 2), the disease prevalence and the covariances between the two BDTs have an important effect upon the type I error of the test $H_0 : \kappa_{11} = \kappa_{21}$. The increase in the prevalence implies an increase in the type I error, especially in samples of 100 and 200, although without overwhelming the nominal error (a situation which has been considered when the type I error is greater than 6.5%). The increase in the values of the covariances implies a decrease in the type I error, especially for $n \leq 500$. In general terms, when the values of the covariances are high, the hypothesis test $H_0 : \kappa_{11} = \kappa_{21}$ is conservative (its type I error is lower than the nominal error) for a sample size $n \leq 500$ (depending on the disease prevalence). The prevalence and the covariances have practically no effect upon the type I error when the samples are very large ($n \geq 1000$). Therefore, in general terms, the type I error of the test $H_0 : \kappa_{11} = \kappa_{21}$ is lower than the nominal error and starting from a certain sample size it fluctuates around the nominal error without overwhelming it. Regarding the type I error of the test $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$, its behavior is, in general terms, very similar to that of the test $H_0 : \kappa_{11} = \kappa_{21}$, although for sample sizes of 100 and 200 its type I error is somewhat lower than that of the hypothesis test without transformation.

Table 2: Type I errors of the hypothesis tests when $L' > L$.

$\kappa_{11} = \kappa_{21} = 0.2$
$Se_1 = 0.7773,\ Sp_1 = 0.7308,\ Se_2 = 0.7773,\ Sp_2 = 0.7308$
$p = 10\%,\ \varepsilon_1 \leq 0.1731,\ \varepsilon_0 \leq 0.1967$

n	$\varepsilon_1 = 0,\ \varepsilon_0 = 0$		$\varepsilon_1 = 0.08,\ \varepsilon_0 = 0.09$		$\varepsilon_1 = 0.16,\ \varepsilon_0 = 0.18$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.022	0.009	0.012	0.008	0	0
200	0.044	0.026	0.031	0.022	0.001	0
300	0.047	0.040	0.035	0.029	0.004	0.004
400	0.045	0.040	0.050	0.042	0.004	0.004
500	0.050	0.048	0.044	0.042	0.010	0.008
1000	0.048	0.046	0.047	0.046	0.020	0.020
2000	0.055	0.056	0.056	0.055	0.044	0.043

$\kappa_{11} = \kappa_{21} = 0.4$
$Se_1 = 0.8864,\ Sp_1 = 0.6746,\ Se_2 = 0.8864,\ Sp_2 = 0.6746$
$p = 30\%,\ \varepsilon_1 \leq 0.1007,\ \varepsilon_0 \leq 0.2195$

n	$\varepsilon_1 = 0,\ \varepsilon_0 = 0$		$\varepsilon_1 = 0.04,\ \varepsilon_0 = 0.10$		$\varepsilon_1 = 0.08,\ \varepsilon_0 = 0.20$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.058	0.049	0.045	0.058	0.049	0.045
200	0.050	0.046	0.049	0.050	0.046	0.049
300	0.047	0.046	0.052	0.047	0.046	0.052
400	0.052	0.051	0.048	0.052	0.051	0.048
500	0.048	0.047	0.040	0.048	0.047	0.040
1000	0.049	0.048	0.050	0.049	0.048	0.050
2000	0.046	0.046	0.048	0.046	0.046	0.048

$\kappa_{11} = \kappa_{21} = 0.6$
$Se_1 = 0.43,\ Sp_1 = 0.97,\ Se_2 = 0.43,\ Sp_2 = 0.97$
$p = 5\%,\ \varepsilon_1 \leq 0.2425,\ \varepsilon_0 \leq 0.0291$

n	$\varepsilon_1 = 0,\ \varepsilon_0 = 0$		$\varepsilon_1 = 0.10,\ \varepsilon_0 = 0.01$		$\varepsilon_1 = 0.20,\ \varepsilon_0 = 0.02$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.011	0.002	0.003	0	0	0
200	0.035	0.017	0.019	0.008	0.001	0
300	0.049	0.026	0.024	0.015	0.001	0
400	0.054	0.033	0.040	0.027	0.007	0.006
500	0.054	0.028	0.033	0.023	0.011	0.009
1000	0.047	0.041	0.049	0.044	0.027	0.023
2000	0.055	0.050	0.049	0.046	0.042	0.040

$\kappa_{11} = \kappa_{21} = 0.8$
$Se_1 = 0.8063,\ Sp_1 = 0.9392,\ Se_2 = 0.8063,\ Sp_2 = 0.9392$
$p = 50\%,\ \varepsilon_1 \leq 0.1562,\ \varepsilon_0 \leq 0.0571$

n	$\varepsilon_1 = 0,\ \varepsilon_0 = 0$		$\varepsilon_1 = 0.07,\ \varepsilon_0 = 0.02$		$\varepsilon_1 = 0.14,\ \varepsilon_0 = 0.04$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.033	0.025	0.023	0.019	0.002	0.001
200	0.048	0.045	0.043	0.039	0.011	0.008
300	0.045	0.044	0.036	0.034	0.027	0.024
400	0.053	0.049	0.049	0.047	0.040	0.037
500	0.056	0.055	0.056	0.055	0.037	0.036
1000	0.048	0.048	0.053	0.052	0.043	0.043
2000	0.045	0.045	0.056	0.055	0.051	0.050

Table 3: Type I errors of the hypothesis tests when $L > L'$.

$\kappa_{11} = \kappa_{21} = 0.2$
$Se_1 = 0.4237,\ Sp_1 = 0.8131,\ Se_2 = 0.4237,\ Sp_2 = 0.8131$
$p = 50\%,\ \varepsilon_1 \leq 0.2442,\ \varepsilon_0 \leq 0.1520$

n	$\varepsilon_1 = 0,\ \varepsilon_0 = 0$		$\varepsilon_1 = 0.02,\ \varepsilon_0 = 0.10$		$\varepsilon_1 = 0.04,\ \varepsilon_0 = 0.20$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.047	0.006	0.045	0.013	0.003	0
200	0.048	0.021	0.049	0.027	0.020	0.006
300	0.056	0.037	0.042	0.030	0.030	0.021
400	0.055	0.044	0.052	0.043	0.045	0.034
500	0.058	0.051	0.042	0.037	0.040	0.034
1000	0.046	0.043	0.055	0.052	0.041	0.039
2000	0.046	0.044	0.048	0.048	0.058	0.057

$\kappa_{11} = \kappa_{21} = 0.4$
$Se_1 = 0.7773,\ Sp_1 = 0.7308,\ Se_2 = 0.7773,\ Sp_2 = 0.7308$
$p = 10\%,\ \varepsilon_1 \leq 0.1731,\ \varepsilon_0 \leq 0.1967$

n	$\varepsilon_1 = 0,\ \varepsilon_0 = 0$		$\varepsilon_1 = 0.08,\ \varepsilon_0 = 0.09$		$\varepsilon_1 = 0.16,\ \varepsilon_0 = 0.18$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.022	0.002	0.006	0	0	0
200	0.043	0.026	0.022	0.008	0	0
300	0.055	0.040	0.030	0.018	0.001	0
400	0.049	0.037	0.047	0.038	0.002	0.001
500	0.039	0.032	0.047	0.042	0.002	0.002
1000	0.049	0.047	0.053	0.050	0.014	0.011
2000	0.056	0.054	0.051	0.050	0.030	0.030

$\kappa_{11} = \kappa_{21} = 0.6$
$Se_1 = 0.8864,\ Sp_1 = 0.6746,\ Se_2 = 0.8864,\ Sp_2 = 0.6746$
$p = 30\%,\ \varepsilon_1 \leq 0.1007,\ \varepsilon_0 \leq 0.2195$

n	$\varepsilon_1 = 0,\ \varepsilon_0 = 0$		$\varepsilon_1 = 0.04,\ \varepsilon_0 = 0.10$		$\varepsilon_1 = 0.08,\ \varepsilon_0 = 0.20$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.050	0.034	0.024	0.014	0.001	0
200	0.053	0.048	0.044	0.036	0.009	0.006
300	0.044	0.040	0.058	0.051	0.017	0.015
400	0.053	0.050	0.052	0.048	0.030	0.028
500	0.052	0.050	0.054	0.050	0.033	0.032
1000	0.054	0.054	0.049	0.047	0.051	0.051
2000	0.055	0.054	0.063	0.062	0.056	0.055

$\kappa_{11} = \kappa_{21} = 0.8$
$Se_1 = 0.81,\ Sp_1 = 0.99,\ Se_2 = 0.81,\ Sp_2 = 0.99$
$p = 5\%,\ \varepsilon_1 \leq 0.1539,\ \varepsilon_0 \leq 0.0099$

n	$\varepsilon_1 = 0,\ \varepsilon_0 = 0$		$\varepsilon_1 = 0.07,\ \varepsilon_0 = 0.004$		$\varepsilon_1 = 0.14,\ \varepsilon_0 = 0.008$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.001	0	0	0	0	0
200	0.009	0.008	0.002	0.001	0	0
300	0.016	0.014	0.005	0.002	0	0
400	0.025	0.019	0.010	0.006	0	0
500	0.027	0.024	0.011	0.007	0	0
1000	0.044	0.040	0.037	0.033	0.006	0.003
2000	0.055	0.053	0.043	0.042	0.022	0.019

When $L > L'$ (Table 3), the prevalence and the covariances also have an important effect (and a similar one to the previous situation) upon the type I error of the test $H_0 : \kappa_{12} = \kappa_{22}$. As in the previous situation, the increase in the prevalence implies an increase in the type I error, especially in samples of 100 and 200, although it does not overwhelm the nominal error. The increase in the covariances implies a decrease in the type I error, especially for $n \leq 500$. Therefore, in general terms, when the values of the covariances are high, for a sample size $n \leq 500$ (depending on the disease prevalence) the hypothesis test $H_0 : \kappa_{12} = \kappa_{22}$ is conservative. The prevalence and the covariances have practically no effect upon the type I error when the sample size is very large ($n = 1000 - 2000$). Therefore, in general terms, the type I error of the test $H_0 : \kappa_{12} = \kappa_{22}$ shows very similar behavior to that of the hypothesis test of the comparison of the two average kappa coefficients when $L' > L$ ($H_0 : \kappa_{11} = \kappa_{21}$); i.e. it is a conservative test and starting from a determined sample size its type I error fluctuates around the nominal error without overwhelming it. Regarding the type I error of the test $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$, its behaviour is, in general terms, very similar to that of the test $H_0 : \kappa_{12} = \kappa_{22}$, although for $n = 100 - 200$ its type I error is, as in the case of $L' > L$, somewhat lower than that of the hypothesis test without transformation.

4.2. Powers

In Table 4, we can see some of the results for the power of the hypothesis tests $H_0 : \kappa_{11} = \kappa_{21}$ and $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$, and in Table 5, we can see some of the results for the power of the hypothesis tests $H_0 : \kappa_{12} = \kappa_{22}$ and $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$. In these tables we also indicate the values of the sensitivities, specificities, prevalence and covariances with which the multinomial samples were generated.

When $L' > L$ (Table 4), the disease prevalence has an important effect on the powers of the tests $H_0 : \kappa_{11} = \kappa_{21}$ and $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$. For the same sample size, the power of each hypothesis test rises with an increase in the prevalence. Regarding the covariances between the two BDTs, the power also rises with increase in the covariances, although its effect is, in general terms, less important than in the case of prevalence. Consequently, based on the prevalence we can reach the following general conclusions:

1. For a prevalence equal to 5% it is necessary to have very large sample size ($n \geq 1000$) so that the power is high (above 80%). If the prevalence is equal to 10%, with a sample size $n \geq 200$ high power is obtained (above 80%, depending on the covariances).
2. If the prevalence is high, p equal to 30% or 50%, with a sample size $n \geq 200$ the powers of both hypothesis tests are very high (higher than 80% or 90%, depending on the covariances).

Table 4: Powers of the hypothesis tests when $L' > L$.

$\kappa_{11} = 0.4$, $\kappa_{21} = 0.2$
$Se_1 = 0.8209$, $Sp_1 = 0.8670$, $Se_2 = 0.7773$, $Sp_2 = 0.7308$
$p = 10\%$, $\varepsilon_1 \leq 0.1392$, $\varepsilon_0 \leq 0.0972$

n	$\varepsilon_1 = 0$, $\varepsilon_0 = 0$		$\varepsilon_1 = 0.06$, $\varepsilon_0 = 0.04$		$\varepsilon_1 = 0.12$, $\varepsilon_0 = 0.08$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.498	0.452	0.613	0.593	0.767	0.755
200	0.831	0.837	0.927	0.935	0.995	0.996
300	0.937	0.941	0.987	0.988	1	1
400	0.986	0.987	1	1	1	1
500	0.990	0.991	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1

$\kappa_{11} = 0.6$, $\kappa_{21} = 0.4$
$Se_1 = 0.8495$, $Sp_1 = 0.8375$, $Se_2 = 0.8864$, $Sp_2 = 0.6746$
$p = 30\%$, $\varepsilon_1 \leq 0.0965$, $\varepsilon_0 \leq 0.1096$

n	$\varepsilon_1 = 0$, $\varepsilon_0 = 0$		$\varepsilon_1 = 0.04$, $\varepsilon_0 = 0.04$		$\varepsilon_1 = 0.08$, $\varepsilon_0 = 0.08$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.570	0.554	0.692	0.677	0.820	0.812
200	0.845	0.844	0.917	0.916	0.985	0.986
300	0.939	0.939	0.982	0.983	0.998	0.998
400	0.984	0.984	0.997	0.997	1	1
500	0.992	0.992	0.998	0.998	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1

$\kappa_{11} = 0.6$, $\kappa_{21} = 0.2$
$Se_1 = 0.8991$, $Sp_1 = 0.7458$, $Se_2 = 0.8131$, $Sp_2 = 0.4237$
$p = 50\%$, $\varepsilon_1 \leq 0.0820$, $\varepsilon_0 \leq 0.1076$

n	$\varepsilon_1 = 0$, $\varepsilon_0 = 0$		$\varepsilon_1 = 0.06$, $\varepsilon_0 = 0.01$		$\varepsilon_1 = 0.12$, $\varepsilon_0 = 0.02$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.954	0.907	0.973	0.949	0.982	0.961
200	0.998	0.998	0.998	0.998	0.999	0.999
300	1	1	1	1	1	1
400	1	1	1	1	1	1
500	1	1	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1

$\kappa_{11} = 0.8$, $\kappa_{21} = 0.6$
$Se_1 = 0.81$, $Sp_1 = 0.99$, $Se_2 = 0.62$, $Sp_2 = 0.98$
$p = 5\%$, $\varepsilon_1 \leq 0.1178$, $\varepsilon_0 \leq 0.0098$

n	$\varepsilon_1 = 0$, $\varepsilon_0 = 0$		$\varepsilon_1 = 0.05$, $\varepsilon_0 = 0.004$		$\varepsilon_1 = 0.10$, $\varepsilon_0 = 0.008$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.010	0.002	0.005	0	0.001	0
200	0.097	0.062	0.100	0.055	0.085	0.049
300	0.252	0.207	0.260	0.208	0.272	0.210
400	0.365	0.323	0.396	0.364	0.466	0.404
500	0.483	0.442	0.520	0.483	0.615	0.586
1000	0.735	0.721	0.801	0.797	0.842	0.842
2000	0.890	0.890	0.890	0.888	0.895	0.895

Table 5: Powers of the hypothesis tests when $L > L'$.

$\kappa_{11} = 0.4$, $\kappa_{21} = 0.2$ $Se_1 = 0.7021$, $Sp_1 = 0.6817$, $Se_2 = 0.3019$, $Sp_2 = 0.9030$ $p = 30\%$, $\varepsilon_1 \leq 0.0900$, $\varepsilon_0 \leq 0.0661$						
n	$\varepsilon_1 = 0$, $\varepsilon_0 = 0$		$\varepsilon_1 = 0.04$, $\varepsilon_0 = 0.03$		$\varepsilon_1 = 0.08$, $\varepsilon_0 = 0.06$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.410	0.238	0.416	0.252	0.433	0.278
200	0.630	0.587	0.733	0.693	0.784	0.749
300	0.790	0.773	0.862	0.851	0.931	0.927
400	0.878	0.876	0.938	0.936	0.978	0.977
500	0.941	0.940	0.970	0.969	0.991	0.991
1000	0.998	0.998	1	1	1	1
2000	1	1	1	1	1	1

$\kappa_{11} = 0.6$, $\kappa_{21} = 0.4$ $Se_1 = 0.8624$, $Sp_1 = 0.6816$, $Se_2 = 0.8112$, $Sp_2 = 0.5293$ $p = 50\%$, $\varepsilon_1 \leq 0.1116$, $\varepsilon_0 \leq 0.1686$						
n	$\varepsilon_1 = 0$, $\varepsilon_0 = 0$		$\varepsilon_1 = 0.05$, $\varepsilon_0 = 0.07$		$\varepsilon_1 = 0.10$, $\varepsilon_0 = 0.14$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.434	0.347	0.567	0.473	0.731	0.623
200	0.680	0.650	0.840	0.820	0.987	0.984
300	0.825	0.813	0.948	0.945	0.999	0.999
400	0.901	0.899	0.981	0.980	1	1
500	0.956	0.952	0.996	0.995	1	1
1000	1	0.999	1	1	1	1
2000	1	1	1	1	1	1

$\kappa_{11} = 0.6$, $\kappa_{21} = 0.2$ $Se_1 = 0.8209$, $Sp_1 = 0.8670$, $Se_2 = 0.2091$, $Sp_2 = 0.9715$ $p = 10\%$, $\varepsilon_1 \leq 0.0374$, $\varepsilon_0 \leq 0.0247$						
n	$\varepsilon_1 = 0$, $\varepsilon_0 = 0$		$\varepsilon_1 = 0.015$, $\varepsilon_0 = 0.01$		$\varepsilon_1 = 0.03$, $\varepsilon_0 = 0.02$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.376	0.135	0.359	0.146	0.411	0.164
200	0.805	0.683	0.814	0.693	0.838	0.720
300	0.945	0.914	0.965	0.928	0.874	0.947
400	1	0.978	0.993	0.972	0.996	0.990
500	1	1	1	1	1	1
1000	1	1	1	1	1	1
2000	1	1	1	1	1	1

$\kappa_{11} = 0.8$, $\kappa_{21} = 0.6$ $Se_1 = 0.9528$, $Sp_1 = 0.9598$, $Se_2 = 0.62$, $Sp_2 = 0.98$ $p = 5\%$, $\varepsilon_1 \leq 0.0292$, $\varepsilon_0 \leq 0.0191$						
n	$\varepsilon_1 = 0$, $\varepsilon_0 = 0$		$\varepsilon_1 = 0.01$, $\varepsilon_0 = 0.07$		$\varepsilon_1 = 0.02$, $\varepsilon_0 = 0.14$	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
100	0.017	0.005	0.024	0.004	0.019	0.007
200	0.112	0.067	0.109	0.057	0.123	0.067
300	0.233	0.189	0.229	0.164	0.243	0.191
400	0.391	0.331	0.401	0.325	0.368	0.308
500	0.483	0.440	0.480	0.428	0.510	0.468
1000	0.796	0.777	0.835	0.822	0.839	0.826
2000	0.953	0.953	0.944	0.944	0.951	0.951

Finally, in general terms, the test $H_0 : \kappa_{11} = \kappa_{21}$ is more powerful than the test $H_0 : \ln(\kappa_{11}) = \ln(\kappa_{21})$, especially when $n \leq 200$, since its type I error is slightly greater (without overwhelming the nominal error).

When $L > L'$ (Table 5), the powers of the hypothesis tests $H_0 : \kappa_{12} = \kappa_{22}$ and $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$ show very similar behaviour to that of the previous case ($L' > L$). The disease prevalence and the covariances have a very similar effect, and the conclusions about the powers are also very similar, although when the prevalence is 10% it is necessary to have a slightly larger sample size ($n \geq 200 - 300$) so that the power is high (above 80%). Finally, and as in the previous case, the test $H_0 : \kappa_{12} = \kappa_{22}$ is more powerful than the test $H_0 : \ln(\kappa_{12}) = \ln(\kappa_{22})$, especially when $n \leq 200$, since its type I error is also slightly greater (without overwhelming the nominal error).

5. EXTENSION TO MORE THAN TWO BDTS

Let us consider J BDTs ($J \geq 3$) and a GS that are applied to all of the n individuals in a random sample. When $L' > L$, the expression of the weighted kappa coefficient for the j-th BDT is

$$\kappa_{j1} = \begin{cases} \frac{2\kappa_j(0)\kappa_j(1)}{\kappa_j(0)-\kappa_j(1)} \ln\left[\frac{\kappa_j(0)+\kappa_j(1)}{2\kappa_j(1)}\right], & p \neq Q_j \\ Se_j + Sp_j - 1, & p = Q_j \end{cases}$$

and when $L > L'$ its expression is

$$\kappa_{j2} = \begin{cases} \frac{2\kappa_j(0)\kappa_j(1)}{\kappa_j(0)-\kappa_j(1)} \ln\left[\frac{2\kappa_j(0)}{\kappa_j(0)+\kappa_j(1)}\right], & p \neq Q_j \\ Se_j + Sp_j - 1, & p = Q_j, \end{cases}$$

with $\kappa_j(0) = \frac{Sp_j-(1-Q_j)}{Q_j}$, $\kappa_j(1) = \frac{Se_j-Q_j}{1-Q_j}$ and $Q_j = pSe_j + q(1 - Sp_j)$, and where $p = \sum_{i_1,\ldots,i_J=0}^{1} p_{i_1,\ldots,i_J}$ is the disease prevalence and $q = 1 - p = \sum_{i_1,\ldots,i_J=0}^{1} q_{i_1,\ldots,i_J}$. The sensitivity and the specificity of the j-th BDT are written as

$$Se_j = \frac{\sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\ldots,i_J}}{\sum\limits_{i_1,\ldots,i_J=0}^{1} p_{i_1,\ldots,i_J}}$$

and

$$Sp_j = \frac{\sum\limits_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\ldots,i_J}}{\sum\limits_{i_1,\ldots,i_J=0}^{1} q_{i_1,\ldots,i_J}},$$

respectively. Replacing these expressions with those of each average kappa coefficient, then

$$\kappa_{j1} = \begin{cases} \frac{2}{a_1-a_2} \times \ln\left[\frac{b_1+1}{2}\right], & p \neq Q_j \\ Se_j + Sp_j - 1, & p = Q_j \end{cases}$$

and

$$\kappa_{j2} = \begin{cases} \frac{2}{a_1-a_2} \times \ln\left[\frac{2}{b_2+1}\right], & p \neq Q_j \\ Se_j + Sp_j - 1, & p = Q_j, \end{cases}$$

where

$$a_1 = \frac{p - \sum\limits_{\substack{i_1,\dots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\dots,i_J} + \sum\limits_{\substack{i_1,\dots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\dots,i_J}}{\frac{\sum\limits_{\substack{i_1,\dots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\dots,i_J}}{\sum\limits_{i_1,\dots,i_J=0}^{1} p_{i_1,\dots,i_J}} - q - \sum\limits_{\substack{i_1,\dots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\dots,i_J} + \sum\limits_{\substack{i_1,\dots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\dots,i_J}},$$

$$a_2 = \frac{q + \sum\limits_{\substack{i_1,\dots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\dots,i_J} - \sum\limits_{\substack{i_1,\dots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\dots,i_J}}{\frac{\sum\limits_{\substack{i_1,\dots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\dots,i_J}}{\sum\limits_{i_1,\dots,i_J=0}^{1} q_{i_1,\dots,i_J}} - p + \sum\limits_{\substack{i_1,\dots,i_J=0 \\ i_j=1}}^{1} p_{i_1,\dots,i_J} - \sum\limits_{\substack{i_1,\dots,i_J=0 \\ i_j=0}}^{1} q_{i_1,\dots,i_J}},$$

$b_1 = \frac{a_1}{a_2}$ and $b_2 = \frac{1}{b_1}$. As the maximum likelihood estimators of the probabilities $p_{i_1,\dots,i_J}$ and $q_{i_1,\dots,i_J}$ are $\hat{p}_{i_1,\dots,i_J} = s_{i_1,\dots,i_J}/n$ and $\hat{q}_{i_1,\dots,i_J} = r_{i_1,\dots,i_J}/n$, with $i_1, \dots, i_J = 0, 1$, the estimator of each average kappa coefficient is obtained replacing in the expressions of κ_{j1} and κ_{j2} each parameter $p_{i_1,\dots,i_J}$ and $q_{i_1,\dots,i_J}$ with its corresponding estimator. Let $\boldsymbol{\kappa}_i = (\kappa_{1i}, \kappa_{2i}, \dots, \kappa_{Ji})^T$ be the vector of average kappa coefficients and $\hat{\boldsymbol{\kappa}}_i = (\hat{\kappa}_{1i}, \hat{\kappa}_{2i}, \dots, \hat{\kappa}_{Ji})^T$ its estimator, where $i = 1$ when $L' > L$ and $i = 2$ when $L > L'$. Applying the delta method, the asymptotic variances-covariances matrix of the vector $\hat{\boldsymbol{\kappa}}_i$ is $\sum_{\hat{\boldsymbol{\kappa}}_i} = \left(\frac{\partial \boldsymbol{\kappa}_i}{\partial \boldsymbol{\pi}}\right) \sum_{\hat{\boldsymbol{\pi}}} \left(\frac{\partial \boldsymbol{\kappa}_i}{\partial \boldsymbol{\pi}}\right)^T$, where $\boldsymbol{\pi}$ is the vector of probabilities. Performing algebraic operations and replacing in this expression each parameter with its estimator, the estimated asymptotic variances-covariances matrix $\hat{\sum}_{\hat{\boldsymbol{\kappa}}_i}$ is obtained. The global hypothesis test to contrast the equality of the J average kappa coefficients is $H_0 : \kappa_{1i} = \kappa_{2i} = \dots = \kappa_{Ji}$ vs H_1 : at least one equality is not true. This hypothesis test is equivalent to $H_0 : \boldsymbol{\varphi}\boldsymbol{\kappa}_i = \mathbf{0}$ vs $H_1 : \boldsymbol{\varphi}\boldsymbol{\kappa}_i \neq \mathbf{0}$, where $\boldsymbol{\varphi}$ is a complete range matrix whose dimension is $(J-1) \times J$. For example, for three BDTs the matrix $\boldsymbol{\varphi}$ is

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

Applying the multivariate central limit theorem it is verified that

$$\sqrt{n}\left(\hat{\boldsymbol{\kappa}}_i - \boldsymbol{\kappa}_i\right) \xrightarrow[n\to\infty]{} N_{J-1}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\kappa}_i}\right),$$

so that the statistic $Q^2 = \hat{\boldsymbol{\kappa}}_i{}^T \boldsymbol{\varphi}^T \left(\boldsymbol{\varphi} \hat{\sum}_{\hat{\boldsymbol{\kappa}}_i} \boldsymbol{\varphi}^T\right)^{-1} \boldsymbol{\varphi}\hat{\boldsymbol{\kappa}}_i$ is distributed according to a distribution T^2 of Hotelling sized $J-1$ and n degrees of freedom, where $J-1$ is the dimension of vector $\boldsymbol{\varphi}\hat{\boldsymbol{\kappa}}_i$. For a large n, the statistic Q^2 is distributed according to a chi-squared central distribution with $J-1$ degrees of freedom when the null hypothesis is true, i.e. $Q^2 = \hat{\boldsymbol{\kappa}}_i{}^T \boldsymbol{\varphi}^T \left(\boldsymbol{\varphi} \hat{\sum}_{\hat{\boldsymbol{\kappa}}_i} \boldsymbol{\varphi}^T\right)^{-1} \boldsymbol{\varphi}\hat{\boldsymbol{\kappa}}_i \xrightarrow[n\to\infty]{} \chi^2_{J-1}$.

The procedure to solve the hypothesis test would be very similar to that used by Roldán-Nofuentes *et al.* [9] to simultaneously compare the weighted kappa coefficients of multiple BDTs: 1) solve the global test to an error of α; 2) if the global test is not significant at that error rate, then the homogeneity of the J average kappa coefficients is not rejected, and if the test is significant then the investigation into the causes of the significance is carried out comparing the pairs of average kappa coefficients using the results in Section 3 and penalizing the level of significance through some method of multiple comparisons, for example Bonferroni [10], Holm [11] or Hochberg [12].

Finally, as in the case of two BDTs, the comparison of multiple average kappa coefficients can be made using logarithmic transformation, and the procedure is similar to that used in the case without transformation.

6. THE "CAKCTBT" PROGRAM

The "cakctbt" program (Comparison of Average Kappa Coefficients of Two Binary Tests) is a program written in R that solves the hypothesis tests to contrast the equality of the average kappa coefficients of two BDTs, i.e. $H_0 : \kappa_{11} = \kappa_{21}$ and $H_0 : \kappa_{12} = \kappa_{22}$. This program runs with the command

$$\text{cakctbt}\,(s_{11}, s_{10}, s_{01}, s_{00}, r_{11}, r_{10}, r_{01}, r_{00})$$

when $\alpha = 5\%$, and with the command

$$\text{cakctbt}\,(s_{11}, s_{10}, s_{01}, s_{00}, r_{11}, r_{10}, r_{01}, r_{00}, \alpha)$$

when $\alpha \neq 5\%$. The program provides the estimation of each average kappa coefficient and its respective standard error, the value of the contrast statistic and the p-value of each hypothesis test. It also provides the confidence intervals for the difference of the average kappa coefficients in each situation ($L' > L$ and $L > L'$). The results obtained when running the program are kept in a file called "Results_cakctbt.txt" in the same folder from where the program is run. The program is available for free at URL:

"`http://www.ugr.es/~bioest/software/cmd.php?seccion=mdb`".

7. EXAMPLE

The results in Section 3 were applied to the study of Weiner *et al.* [13] about the diagnosis of coronary disease, which is a classic example when comparing the parameters of two BDTs subject to a paired design. In Table 6 (Observed frequencies), we can see the results when applying two BDTs, a cardiac stress test and the individual's clinical history in relation to coronary disease, and the GS (coronary arteriography) to a sample of 871 individuals, and where the variable T_1 models the result of the stress test, T_2 models the result of the individual's clinical history and the variable D models the result of the coronary angiography.

Table 6: Data of the study of Weiner *et al.* and results.

Observed frequencies					
	$T_1 = 1$		$T_1 = 0$		Total
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	
$D = 1$	473	29	81	25	608
$D = 0$	22	46	44	151	263
Total	495	75	125	176	871

Results
$L' > L$
$\hat{\kappa}_{11} = 0.574$, $\hat{\kappa}_{21} = 0.658$ $\hat{Var}\,(\hat{\kappa}_{11}) = 0.031820$, $\hat{Var}\,(\hat{\kappa}_{21}) = 0.029746$ $\hat{Cov}\,(\hat{\kappa}_{11}, \hat{\kappa}_{21}) = 0.000112$
$H_0 : \kappa_{11} = \kappa_{21}$ vs $H_1 : \kappa_{11} \neq \kappa_{21}$ $z = 2.06$, p-value $= 0.039$ 95% *CI* for $\kappa_{21} - \kappa_{11}$: (0.0041 ; 0.1644)
$L > L'$
$\hat{\kappa}_{12} = 0.519$, $\hat{\kappa}_{22} = 0.680$ $\hat{Var}\,(\hat{\kappa}_{12}) = 0.031303$, $\hat{Var}\,(\hat{\kappa}_{22}) = 0.029260$ $\hat{Cov}\,(\hat{\kappa}_{11}, \hat{\kappa}_{21}) = 0.000229$
$H_0 : \kappa_{12} = \kappa_{22}$ vs $H_1 : \kappa_{12} \neq \kappa_{22}$ $z = 4.33$, p-value $= 1.46 \times 10^{-5}$ 95% *CI* for $\kappa_{22} - \kappa_{12}$: (0.0881 ; 0.2336)

In Table 6 (Results), we can see the estimations of the parameters, the results of the hypothesis tests ($\alpha = 5\%$) and the confidence intervals to 95%. Based on these results, if the clinical laboratory researcher is more concerned about the false positives than the false negatives ($L' > L$), then the equality of the average kappa coefficients is rejected, and it holds that the average kappa coefficient of the clinical history (which has a "good" value in terms of point estimation) is significantly larger than that of the stress test (which has a "moderate" value in terms of point estimation). Therefore, the average beyond-chance agreement between the clinical history and the angiography is, with a confidence of 95%, a value between 0.0041 and 0.1644 greater than the average beyond-chance agreement between the stress tests and the angiography. Similar conclusions are obtained if the clinical laboratory researcher is more concerned about the false negatives than the false positives ($L > L'$). In this situation, the average beyond-chance agreement between the clinical history and the angiography, with a confidence of 95%, is a value between 0.0881 and 0.2336 higher than the average beyond-chance agreement between the stress tests and the angiography.

8. DISCUSSION

The comparison of the performance of two BDTs in relation to a GS can be made through a paired design or an unpaired one. Paired design consists of applying the two BDTs to all of the individuals in a simple, whereas in unpaired design each individual is only tested with one of the two BDTs. Paired design is used more in practice and has more advantages than unpaired design [14]. Paired design was chosen to develop the method proposed in this article.

In clinical practice, when we consider the losses in an erroneous classification with two BDTs, the appropriate parameters to compare the two BDTs are weighted kappa coefficients. In this situation, it is necessary to assume a value for the weighting index c and solve the test $H_0 : \kappa_1(c) = \kappa_2(c)$ vs $H_1 : \kappa_1(c) \neq \kappa_2(c)$ applying the Bloch method [4]. The value of the weighting index c is set by the clinical laboratory researcher based on his or her knowledge about the problem in question. If the clinical laboratory researcher does not have enough knowledge to allow them to allocate a value to the weighting index c, the comparison of the performance of the two (or more) BDTs can be made through the average kappa coefficients, which are measures of the beyond-chance agreement between each BDT and the GS and do not depend on the weighting index c. Therefore, if the clinical laboratory researcher can assume a value of the weighting index c, then compare the weighted kappa coefficients of the two BDTs applying the Bloch method [4]. In the opposite case, compare the weighted kappa coefficients κ_{i1} if there is a greater concern about the false positives than about the false negatives, or compare the weighted kappa coefficients κ_{i2} if there is a greater concern about the false negatives than the false positives.

In this article, we have studied the comparison of the average kappa coefficients of two (and more) BDTs when the clinical laboratory researcher considers that loss associated with the false positives is greater than that associated with the false negatives ($L' > L$), and when the clinical laboratory researcher considers the opposite ($L > L'$). The hypothesis tests studied are asymptotic and the simulation experiments carried out have demonstrated that the type I errors do not overwhelm the nominal error of 5%. Regarding the power of each hypothesis test, this increases with the prevalence, and so when the prevalence is small (e.g. 5%) it is necessary to have a very large sample size ($n \geq 1000$) so that the power is high (above 80%); whereas with a large prevalence (e.g. 30% or 50%), with a sample size $n \geq 200$ a high power is obtained.

In the expressions of the statistics deduced to solve the hypothesis tests, the variances-covariances have been estimated applying the delta method. An alternative method is to estimate these variances-covariances through bootstrap. Simulation experiments (similar to those in Section 4) have shown that there is no important difference in terms of the type I error and the power between both methods of estimation of the variances-covariances.

The results were extended to the case of more than two BDTs, finding that the solution to the hypothesis test is also asymptotic and a method based on multiple comparisons is proposed to solve the problem. This method is very similar to that used in the analysis of the variance. Firstly, the global test is solved to an error of α and if the test is significant then the causes of the significance are investigated making paired comparisons and applying a multiple comparison method. For our problem, we have chosen the Bonferroni, Holm or Hochberg methods, which are very easy to apply and have been used in the field of BDTs [15,16].

The method that we have proposed requires knowledge of the disease status of all of the individuals in a sample through the application of the GS. If the disease status of any individual is unknown, leading to the problem known as partial disease verification, the method proposed cannot be applied. If the verification process with the GS only depends on the results of the BDTs, a solution to this problem could be obtained following a method similar to that used by Roldán-Nofuentes and Luna del Castillo [17] and Roldán-Nofuentes *et al.* [18].

If case-control sampling is being used, the method that we have proposed cannot be used either as it is necessary to know the disease prevalence. An extension of the study of Roldán-Nofuentes and Amro [19] to the situation of two BDTs may be a solution to this problem.

ACKNOWLEDGMENTS

This research was supported by the Spanish Ministry of Economy, Grant Number MTM2016-76938-P. We thank the two Referees, Editor Associate and Editor of REVSTAT – Statistical Journal for their helpful comments that improved the quality of the paper.

REFERENCES

[1] Bloch, D.A. and 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*, Sage Publications, Newbury Park.

[3] Kraemer, H.C.; Periyakoil, V.S. and Noda, A. (1992). Kappa coefficients in medical research, *Statistics in Medicine*, **21**, 2109–2129.

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

[5] Roldán-Nofuentes, J.A. and Olvera-Porcel, C. (2015). Average kappa coefficient: a new measure to assess a binary test considering the losses associated with an erroneous classification, *Journal of Statistical Computation and Simulation*, **85**, 1601–1620.

[6] Roldán-Nofuentes, J.A.; Luna del Castillo, J.D. and 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.

[7] Vacek, P. (1985). The effect of conditional dependence on the evaluation of diagnostic tests, *Biometrics*, **41**, 959–968.

[8] Cicchetti, D.V. (2001). The precision of reliability and validity estimates revisited: distinguishing between clinical and statistical significance of sample size requirements, *Journal of Clinical and Experimental Neuropsychology*, **23**, 695–700.

[9] Roldán-Nofuentes, J.A. and Luna del Castillo, J.D. (2010). Comparison of weighted kappa coefficients of multiple binary diagnostic tests done on the same subjects, *Statistics in Medicine*, **29**, 2149–2165.

[10] Bonferroni, C.E. (1936). Teoria statistica delle classi e calcolo delle probabilitá, *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

[11] Holm, S. (1979). A simple sequential rejective multiple testing procedure, *Scandinavian Journal of Statistics*, **6**, 65–70.

[12] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika*, **75**, 800–802.

[13] Weiner, D.A.; Ryan, T.J.; McCabe, C.H.; Kennedy, J.W.; Schloss, M.; Tristani, F.; Chaitman, B.R. and Fisher, L.D. (1979). Exercise stress testing. Correlations among history of angina, ST-segment response and prevalence of coronary-artery disease in the coronary artery surgery study (CASS), *The New England Journal of Medicine*, **31**, 230–235.

[14] Pepe, M.S. (2003). *The statistical evaluation of medical tests for classification and prediction*, Oxford University Press, New York.

[15] 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, Special issue "Computational Statistics for Clinical Research"*, **56**, 1161–1173.

[16] Marín-Jiménez, E. and Roldán-Nofuentes, J.A. (2015). Comparison of the predictive values of multiple binary diagnostic tests in the presence of ignorable missing data, *Revstat*, **15**, 45–64.

[17] Roldán-Nofuentes, J.A. and 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.

[18] Roldán-Nofuentes, J.A.; Marín-Jiménez, E. and Luna del Castillo, J.D. (2014). Asymptotic hypothesis test to simultaneously compare the weighted kappa coefficients of multiple binary diagnostic tests in the presence of ignorable missing data, *Journal of Statistical Computation and Simulation*, **84**, 273–289.

[19] Roldán-Nofuentes, J.A. and 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.