Determining sample size to evaluate and compare the accuracy of binary diagnostic tests in the presence of partial disease verification

J.A. Roldán Nofuentes, J.D. Luna del Castillo, M.A. Montero Alonso

Abstract

Calculating sample size to evaluate the accuracy of a binary diagnostic test and to compare the accuracy of two binary diagnostic tests is an important question in the study of diagnostic statistical methods. In the presence of partial disease verification, the disease status of some patients in the sample is unknown, so that the calculation of sample size can be complicated. A method to calculate sample size when evaluating the sensitivity and the specificity of a binary diagnostic test and when comparing the sensitivity and specificity of two binary tests in the presence of partial disease verification is proposed. The results obtained were applied to the diagnosis of coronary stenosis.

1. Introduction

The accuracy of a binary diagnostic test is normally measured through its sensitivity and specificity. The sensitivity is the probability of the diagnostic test giving a positive result when the patient is diseased, and the specificity is the probability of the test giving a negative result when the patient is non-diseased. In order to evaluate the sensitivity and the specificity of a binary test, it is necessary to know the true disease status (present or absent) of each patient through the application of a gold standard (e.g. a biopsy, a surgical operation). However, in clinical practice it is not common to apply the gold standard to all of the patients in the sample, giving rise to the problem known as partial disease verification (Begg and Greenes, 1983). Therefore, if the gold standard is a test which involves a risk for the patient or a very expensive test, then not all patients are verified with the gold standard. For example, the diagnosis of coronary stenosis the gold standard is made with a coronary angiography, whose application can cause different reactions in some patients (heart attack, infections, ...), and therefore not all patients are verified with the gold standard. The evaluation of a binary diagnostic test in the presence of partial disease verification has been the object of different studies (Zhou, 1993, 1994, 1998a; Kosinski and Barnhart, 2003a,b; Harel and Zhou, 2006; Roldán Nofuentes and Luna del Castillo, 2007a,b). The problem of partial disease verification can also arise when we compare the accuracy of two binary tests. If the probability of verifying a patient with the gold standard does not depend on disease status, Zhou (1998b) deduced hypothesis tests to compare the sensitivities (specificities) of two binary tests, by estimating the variance–covariance matrix applying the delta method. Roldán Nofuentes and Luna del Castillo (2005, 2006, 2007c,d,e) studied the comparison of different measurements of two binary tests in the presence of partial disease verification. Harel and Zhou (2007) used multiple imputation to compare the accuracy of two binary tests. In the absence of the gold standard, Hadgu et al. (2005) have studied the estimation of the accuracy of a binary diagnostic test and Dendukuri et al. (2004) have studied the sample size determination for estimating the disease prevalence.

In the evaluation and comparison of the accuracy of binary diagnostic tests, an important question is the determination of the sample size necessary to carry out the study. When trying to evaluate the accuracy of a binary test, researchers normally
have to consider calculating the sample size necessary to estimate the sensitivity (specificity) with a determined precision; and when we want to compare the accuracy of two binary tests, researchers have to consider calculating the sample size necessary to estimate the difference between sensitivities (specificities) with a determined precision or to compare the sensitivities (specificities) of the two diagnostic tests to an error $\alpha$ and a power $1 - \beta$. A review of the methods to calculate sample size to evaluate and compare diagnostic methods can be seen in Obuchowski (1998) and in Zhou et al. (2002). Alonzo et al. (2002) studied the calculation of sample size when we compare two diagnostic tests under different types of designs. Li and Fine (2004) studied the calculation of sample size in the estimation of the sensitivity and the specificity of a diagnostic test in prospective studies. Fosgate (2005) developed an algorithm to calculate the sample size to estimate the accuracy of a binary diagnostic test.

In the presence of partial disease verification, the calculation of sample size to evaluate the accuracy of a binary test and to compare the accuracy of two binary tests cannot be carried out applying traditional methods, since sensitivity and specificity cannot be estimated as binomial proportions. In this study, we propose a method to calculate sample size to estimate the sensitivity and the specificity of a binary diagnostic test in the presence of partial disease verification and when we compare the sensitivities and the specificities of two binary tests in the presence of partial disease verification.

2. A single binary test

A very important question in the evaluation of a binary diagnostic test is the calculation of the sample size necessary to estimate the accuracy of the test with a given precision at a confidence level of $100(1 - \alpha)\%$ (Obuchowski, 1998; Zhou et al., 2002). The equation to calculate the sample size to build a two-tailed confidence interval for the accuracy (sensitivity or specificity) of the diagnostic test is

$$m = \frac{z_{1-\alpha/2}^2 \nu}{L^2}, \quad (1)$$

where $z_{\gamma}$ is the 100$\gamma$th percentile of the standard normal distribution, $L$ is the precision of the estimation (the length of half of the confidence interval) and $V(\hat{\theta})$ is the variance function (McCullagh and Nelder, 1989) of $\hat{\theta}$, where $\theta$ is the sensitivity ($Se$) or the specificity ($Sp$). Eq. (1) presents the problem of estimating the variance function of $\hat{\theta}$. Arkin and Wachtel (1990) recommend estimating $V(\hat{\theta})$ using binomial distribution, so that $V(\hat{\theta}) = \theta(1 - \theta)$, where $\theta$ is the sensitivity or the specificity conjectured for the diagnostic test. In the presence of partial disease verification, for some patients the disease status is unknown, so that the assessment of the diagnostic test cannot be carried out only using those patients verified with the gold standard (Begg and Greenes, 1983; Zhou, 1993), and therefore it is not possible to calculate the sample size to estimate the accuracy of the diagnostic test from the patients verified using the traditional methods in $2 \times 2$ tables. We now propose a method to calculate the sample size to estimate the accuracy of a binary diagnostic test in the presence partial disease verification.

2.1. The method

Let us consider a binary diagnostic which is applied to all of the patients in a random sample sized $n$. Let $T$ be the random variable which models the result of the diagnostic test, in such a way that $T = 1$ when the test is positive and $T = 0$ when the test is negative. Let $V$ be the random variable which models the verification process, $V = 1$ when the patient is verified with the gold standard and $V = 0$ when the patient is not verified; and let $D$ be the random variable which models the result of the gold standard, $D = 1$ when the patient is diseased and $D = 0$ when the patient is non-disease. The application of the diagnostic test to all of the patients in the sample and the application of the gold standard to a part of the $n$ patients gives Table 1. Let $Se = P(T = 1|D = 1)$, $Sp = P(T = 0|D = 0)$ and $p = P(D = 1)$ be the sensitivity, the specificity and the disease prevalence, respectively. Let $\lambda_{ij} = P(V = 1|D = i, T = j)$ be the probability of verifying a patient with the gold standard when $D = i$ and $T = j$ (i, j = 0, 1). If $\lambda_{ij} = P(V = 1|D = i, T = j)$ = $P(V = 1)|T = j)$ = $\lambda_{j}$, the verification process only depends on the result of the diagnostic test, and the verification process is missing at random (MAR) (Rubin, 1976). Subject to this assumption, the probabilities of each cell in Table 1 (assessment of a binary test) are:

$$\xi_j = P(V = 1, D = 1, T = j) = p\lambda_{j}Se^j (1 - Se)^{1-j},$$
$$\psi_j = P(V = 1, D = 0, T = j) = (1 - p)\lambda_{j}Sp^{1-j} (1 - Sp)^j,$$

and

$$\xi_j = P(V = 0, T = j) = (1 - \lambda_{j}) \left[ pSe^j (1 - Se)^{1-j} + (1 - p)Sp^{1-j} (1 - Sp)^j \right],$$

with $j = 0, 1$.

Let us suppose that, subject to the MAR assumption, the MAR assumption, the disease prevalence and the verification probabilities are known. Let $\omega = (\xi_1, \xi_0, \psi_1, \psi_0, \zeta_1, \zeta_0)^T$. As $\omega$ is the vector of probabilities of a multinomial distribution, the variance–covariance matrix of $\omega$ is

$$\sum_\omega = \text{diag}(\omega) - \omega^T \omega.$$
samplesize is necessary to specify a value for the difference between \( \theta \) to an error of \( \alpha \). The variance functions of sensitivity and specificity, samplesize is easily calculated applying Eq. (7).

\[
\begin{align*}
\text{Frequencies observed in the evaluation of a binary test} & \\
\begin{array}{c|c|c|c|c}
T & T = 1 & T = 0 & \text{Total} \\
V & 1 & D = 1 & s_1 & s_0 & s \\
 & D = 0 & r_1 & r_0 & r \\
V & 0 & u_1 & u_0 & u \\
\hline
\text{Total} & n_1 & n_0 & n \\
\end{array}
\end{align*}
\]

\[
\begin{align*}
\text{Frequencies observed in the comparison of two binary tests} & \\
\begin{array}{c|c|c|c|c|c|c}
T & T_1 = 1 & T_2 = 0 & T_1 = 0 & T_2 = 0 & \text{Total} \\
V & 1 & D = 1 & s_{11} & s_{10} & s_{01} & s_{00} & s \\
 & D = 0 & r_{11} & r_{10} & r_{01} & r_{00} & r \\
V & 0 & u_{11} & u_{10} & u_{01} & u_{00} & u \\
\hline
\text{Total} & n_{11} & n_{10} & n_{01} & n_{00} & n \\
\end{array}
\end{align*}
\]

Let \( \phi_i = \xi_i + \psi_j + \zeta_j \). The sensitivity and specificity of the diagnostic test are written in terms of the probabilities \( \xi_i, \psi_j \) and \( \zeta_j \) as

\[
\begin{align*}
Se &= \frac{\xi_1 \phi_1}{\xi_1 + \psi_1} \left( \sum_{i = 0}^{1} \frac{\xi_i \phi_i}{\xi_i + \psi_i} \right) - (3)
\end{align*}
\]

and

\[
\begin{align*}
Sp &= \frac{\psi_0 \phi_0}{\xi_0 + \psi_0} \left( \sum_{i = 0}^{1} \frac{\psi_i \phi_i}{\xi_i + \psi_i} \right)
\end{align*}
\]

and applying the delta method (Agresti, 2002) it holds that

\[
\begin{align*}
V(Se) &= \left( \frac{\partial Se}{\partial \omega} \right) \sum_{\omega} \left( \frac{\partial Se}{\partial \omega} \right)^T
\end{align*}
\]

and

\[
\begin{align*}
V(Sp) &= \left( \frac{\partial Sp}{\partial \omega} \right) \sum_{\omega} \left( \frac{\partial Sp}{\partial \omega} \right)^T.
\end{align*}
\]

Performing the algebraic operations we obtain that

\[
\begin{align*}
V(Se) &= \frac{Se (1 - Se) [q (SeSp - Y) (pY - Sp) \lambda_0 - Se \{qSp (pY - Sp + 1) + p (1 - Se) \lambda_0\} \lambda_1]}{p (pY - Sp) (pY - Sp + 1) \lambda_0 \lambda_1}
\end{align*}
\]

and

\[
\begin{align*}
V(Sp) &= \frac{Sp (1 - Sp) [pSeSp (Sp - pY) \lambda_0 - (1 - Sp) \{p (Se - 1) (pY - Sp + 1) - qSp \lambda_0\} \lambda_1]}{p (pY - Sp) (pY - Sp + 1) \lambda_0 \lambda_1},
\end{align*}
\]

where \( q = 1 - p \) and \( Y = Se + Sp - 1 \) is the Youden’s index of the diagnostic test. As the probabilities given in Eq. (2) are the probabilities of a multinomial distribution, their maximum likelihood estimators (MLEs) are \( \hat{\xi_i} = s_i/n, \hat{\psi_j} = r_j/n \) and \( \hat{\zeta_j} = u_j/n \). Substituting in Eq. (3) the parameters with their corresponding MLEs, the MLEs of sensitivity and specificity which are obtained are the same as those obtained by Begg and Greenes (1983) and Zhou (1993). Once we have calculated the variance functions of sensitivity and specificity, sample size is easily calculated applying Eq. (1).

The method we propose can also be used to calculate the sample size necessary contrast \( H_0 : \theta = \theta_0 \) vs. \( H_1 : \theta \neq \theta_0 \) to an error of \( \alpha \) and a power of \( 1 - \beta \), when \( \theta \) is the sensitivity or specificity. In order to calculate the sample size, it is necessary to specify a value for the difference between \( \theta \) and \( \theta_0 \) subject to the alternative hypothesis. If we call \( \delta = \theta - \theta_0 \) the conjectured value of the difference between \( \theta \) and \( \theta_0 \) subject to the alternative hypothesis, the equation to calculate the sample size is

\[
\begin{align*}
m = \left( \frac{z_{\alpha/2} \sqrt{V_0(\hat{\theta})} + z_\beta \sqrt{V_1(\hat{\theta})}}{\delta} \right)^2,
\end{align*}
\]
where $V_0(\hat{\theta})$ is the variance function of $\hat{\theta}$ subject to the null hypothesis and $V_1(\hat{\theta})$ is the variance function of $\hat{\theta}$ subject to the alternative hypothesis. If we know the values of $\hat{\theta}$ subject to the null hypothesis and subject to the alternative hypothesis, with a disease prevalence and verification probabilities of $\lambda_1$ and $\lambda_0$, then the variance functions can be calculated through the same method used previously. Finally, the sample size is calculated using Eq. (4).

2.2. Simulation study

The method which we propose to calculate sample size requires knowledge of the sensitivity and specificity of the diagnostic test, disease prevalence and verification probabilities. Sensitivity, specificity and prevalence can be estimated from a pilot sample or another similar study, and the verification probabilities can be estimated from the pilot sample or can be previously established from verification costs. Once we have estimated the parameters, in order to calculate the sample size it is assumed that these estimators are the true values of the parameters. Therefore, it is necessary to study the robustness of the method to calculate sample size. For this purpose we have carried out simulation experiments which consisted of the generation of 5000 random samples with multinomial distributions whose probabilities have been calculated with Eq. (2). For sensitivity and specificity we took the values $(Se = 0.80, Sp = 0.70)$ and $(Se = 0.90, Sp = 0.85)$, which are values which appear with a certain frequency in clinical practice; as values of prevalence we took $10\%, 30\%, 50\%, 70\%$ and $90\%$. In clinical practice, the probability of verifying a patient with a positive test ($\lambda_1$) is greater than the probability of verifying a patient with a negative test ($\lambda_0$), and therefore as verification probabilities we have taken the following values: (a) $\lambda_1 = 0.80$ and $\lambda_0 = 0.10$; (b) $\lambda_1 = 0.80$ and $\lambda_0 = 0.30$; (c) $\lambda_1 = 0.95$ and $\lambda_0 = 0.10$; (d) $\lambda_1 = 0.95$ and $\lambda_0 = 0.30$; (e) $\lambda_1 = \lambda_0 = 1$. Case (a) can be considered as a scenario with low verification, cases (b) and (c) can be considered as scenarios with intermediate verifications, case (d) is a scenario with high verification and in case (e) all of the patients are verified with the gold standard and there is no verification bias. As precision (L) we took the value 0.05 and as error $\alpha = 5\%$. The sample size of 5000 random samples with the same multinomial distribution was calculated applying the method proposed in Section 2.1, except when $\lambda_i = 1$ in which case the sample size was calculated applying the method of Arkin and Wachtel (1990); and for each 5000 samples (in the presence of partial verification) we calculated the average sample size and the relative root mean squared error (RRMSE),

$$RRMSE = \sqrt{\frac{\sum_{i=1}^{5000} (\hat{m}_i - m)^2}{4999}} / m,$$

where $m$ is the sample size calculated from the values with which the multinomial samples were generated and $\hat{m}_i$ is the estimated value of the sample size calculated from the maximum likelihood estimators of sensitivity, specificity, prevalence and verification probabilities obtained from each random sample and whose expressions are (Begg and Greens, 1983; Zhou, 1993)

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

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

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

and

$$\hat{\lambda}_1 = \frac{s_1 + r_1}{n_1}.$$

The RRMSE is a measure of the quality of an estimator, and although there is no concrete value of the RRMSE below which the estimator is valid, we have considered, as happens in many practical situations, that a RRMSE value lower than 25% indicates that the estimator has good quality, whilst a value higher than 25% indicates that the estimator is not valid. In Table 2 we show the results obtained for sensitivity and for specificity, and which we reach the following conclusions. The sample size to estimate sensitivity decreases when there is an increase in disease prevalence, and the sample size to estimate specificity decreases when there is a decrease in prevalence. Regarding the effect of the verification probabilities, the sample size to estimate sensitivity decreases when there is an increase in verification probabilities, and the effect of the verification probability $\lambda_0$ is much more important than the effect of $\lambda_1$, for values of $\lambda_1$ which are sufficiently large (such as the ones we have used in the simulation experiments) and independent of disease prevalence. In a similar way, the sample size to estimate the specificity decreases when there is an increase in verification probabilities. Nevertheless, the effect of $\lambda_0$ is less important than in the case of sensitivity. Therefore, when the probability of verifying a patient with a positive test is relatively high (as usually occurs in clinical practice), in order for the estimated sample size to be lower it is necessary to increase the verification probability of the patients with a negative test, and it is not necessary to increase the verification probabilities with a positive test. Regarding total verification, partial disease verification means an important increase in sample size when estimating sensitivity (specificity), above all when the prevalence is not large (small). In relation to the relative root mean squared error, this is not usually higher than 25%, therefore, the results obtained confirm the quality of
Table 2
Sample size to estimate the sensitivity and specificity of a binary test ($L = 0.05$ and $\alpha = 5\%$)

<table>
<thead>
<tr>
<th>Sample size for sensitivity</th>
<th>$\lambda_1 = 0.80$ $\lambda_0 = 0.10$</th>
<th>$\lambda_1 = 0.95$ $\lambda_0 = 0.30$</th>
<th>$\lambda_1 = \lambda_0 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>p (%)</td>
<td>m</td>
<td>%AVP</td>
<td>ASS</td>
</tr>
<tr>
<td>10</td>
<td>19711</td>
<td>34.5</td>
<td>19527</td>
</tr>
<tr>
<td>30</td>
<td>6096</td>
<td>41.5</td>
<td>6040</td>
</tr>
<tr>
<td>50</td>
<td>3253</td>
<td>48.5</td>
<td>3218</td>
</tr>
<tr>
<td>70</td>
<td>1871</td>
<td>55.5</td>
<td>1850</td>
</tr>
<tr>
<td>90</td>
<td>825</td>
<td>62.5</td>
<td>809</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample size for specificity</th>
<th>$\lambda_1 = 0.80$ $\lambda_0 = 0.10$</th>
<th>$\lambda_1 = 0.95$ $\lambda_0 = 0.30$</th>
<th>$\lambda_1 = \lambda_0 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>p (%)</td>
<td>m</td>
<td>%AVP</td>
<td>ASS</td>
</tr>
<tr>
<td>10</td>
<td>403</td>
<td>34.5</td>
<td>443</td>
</tr>
<tr>
<td>30</td>
<td>640</td>
<td>41.5</td>
<td>658</td>
</tr>
<tr>
<td>50</td>
<td>1115</td>
<td>48.5</td>
<td>1141</td>
</tr>
<tr>
<td>70</td>
<td>2400</td>
<td>55.5</td>
<td>2463</td>
</tr>
<tr>
<td>90</td>
<td>10043</td>
<td>62.5</td>
<td>10257</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample size for specificity</th>
<th>$\lambda_1 = 0.80$ $\lambda_0 = 0.10$</th>
<th>$\lambda_1 = 0.95$ $\lambda_0 = 0.30$</th>
<th>$\lambda_1 = \lambda_0 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>p (%)</td>
<td>m</td>
<td>%AVP</td>
<td>ASS</td>
</tr>
<tr>
<td>10</td>
<td>16675</td>
<td>31.7</td>
<td>16409</td>
</tr>
<tr>
<td>30</td>
<td>5265</td>
<td>40.1</td>
<td>5198</td>
</tr>
<tr>
<td>50</td>
<td>2894</td>
<td>48.5</td>
<td>2851</td>
</tr>
<tr>
<td>70</td>
<td>1741</td>
<td>56.9</td>
<td>1706</td>
</tr>
<tr>
<td>90</td>
<td>813</td>
<td>65.3</td>
<td>789</td>
</tr>
</tbody>
</table>

$\text{p: prevalence; } m: \text{ sample size; } \%\text{AVP: } \% \text{ average of verified patients; } \text{ASS: } \text{average sample size; RRMSE: relative root mean squared error.}$

Fig. 1. The effect of verification probabilities on sample size to estimate the sensitivity.

the method which we propose in order to calculate sample size when we assess the accuracy of a binary diagnostic test. In some cases, the sample size obtained applying the method used in Section 2.1 may be very large, especially when the disease prevalence and the verification probabilities are small. In this situation, we recommend taking large verification probabilities, so that the estimated sample size applying the method used in Section 2.1 is lower than when the verification probabilities are low (in the Discussion of the article we analyse some criteria to establish the verification probabilities). Thus, for example in Table 2 when $Se = 0.80$, $Sp = 0.70$, $p = 10\%$, $\lambda_1 = 0.80$ and $\lambda_0 = 0.10$, the sample size is very large ($m = 19711$) which makes it difficult to carry out the study; increasing the verification probabilities ($\lambda_1 = 0.95$, $\lambda_0 = 0.30$) the sample size is 6927 patients.

In Figs. 1 and 2 we show the effect that probability $\lambda_0$ has on the sample size to estimate sensitivity and specificity, respectively, for the same value $\lambda_1 = 0.80$ (the figures taking $\lambda_1 = 0.95$ are very similar to these). In Fig. 1 it can be
observed how, for the same prevalence, the increase in $\lambda_0$ has an important effect on the sample size to estimate sensitivity (overall when the disease prevalence is not very high). The probability of verifying a patient with a negative test ($\lambda_0$) has a greater effect in the calculation of sample size than the probability of verifying a patient with a positive test ($\lambda_1$). An increase in $\lambda_0$, when $\lambda_1$ remains constant, implies an important reduction in the sample size necessary to estimate the sensitivity of the diagnostic test, especially when the disease prevalence is not very high. Nevertheless, as can be observed in Fig. 2, the effect that $\lambda_0$ has on the sample size to estimate the specificity is less important than in the case of sensitivity. Therefore, when trying to calculate the sample size to assess the accuracy (sensitivity or specificity) of a binary diagnostic test, the high values of $\lambda_0$ imply a lower sample size than with low values (for the same value of $\lambda_1$).

### 3. Two binary tests

Comparison of the accuracy of two binary diagnostic tests is another important problem in the study of diagnostic statistical methods and has been the subject of numerous studies. The equation used to calculate sample size when trying to construct a two-tailed confidence interval for the difference between the two sensitivities or specificities is

$$m = \frac{z^2_{1-\alpha/2} V(\hat{\theta}_1 - \hat{\theta}_2)}{L^2},$$

where $z_{\gamma}$ is the $100\gamma$-th percentile of the normal standard distribution, $L$ is the precision of the estimation and $V(\hat{\theta}_1 - \hat{\theta}_2)$ is the variance function (McCullagh and Nelder, 1989) of $\hat{\theta}_1 - \hat{\theta}_2$, when $\theta_i$ is the sensitivity ($Se_i$) or the specificity ($Sp_i$) of each diagnostic test. Therefore, in order to calculate the sample size it is necessary to previously determine $V(\hat{\theta}_1 - \hat{\theta}_2)$, for which we are going to apply a similar method to the one used in the previous section.

#### 3.1. The method

Let us consider two diagnostic binary tests which are applied independently to the same random sample of $n$ patients. Let $T_1$ and $T_2$ be random variables which model the results diagnostic tests 1 and 2, respectively, in such a way that $T_h = 1$ when the result of the $h$th test is positive and $T_h = 0$ when the result of the $h$th test is negative. Let $V$ and $D$ be the random variables defined in the previous section. Let $Se_h = P(T_h = 1 | D = 1)$ and $Sp_h = P(T_h = 0 | D = 0)$ be the sensitivity and the specificity of the $h$th diagnostic test ($h = 1, 2$), $p = P(D = 1)$ the disease prevalence, and $\lambda_{ijk} = P(V = 1 | T_1 = i, T_2 = j, D = k)$ the probability of verifying a patient with results $T_1 = i, T_2 = j$ and $D = k$, with $i, j, k = 0, 1$. The application of the two diagnostic tests to all of the patients in a random sample sized $n$ and the application of the gold standard to a part of the sample gives us Table 1. When the verification process only depends on the results of the two binary tests and not on the result of the gold standard, it is verified that $\lambda_{ijk} = \lambda_{ij}, i, j, k = 0, 1$. This assumption is equivalent to supposing that the verification process is missing at random (MAR) (Rubin, 1976). Subject to the MAR assumption, let the probabilities be

$$\xi_{ij} = P(V = 1, D = 1, T_1 = i, T_2 = j),$$

$$\psi_{ij} = P(V = 1, D = 0, T_1 = i, T_2 = j),$$

and

$$\xi_{ij} = P(V = 0, T_1 = i, T_2 = j),$$

with $i, j = 0, 1$, and $\sum_{i,j=0}^{1} \xi_{ij} + \sum_{i,j=0}^{1} \psi_{ij} + \sum_{i,j=0}^{1} \xi_{ij} = 1$. In general, and as happens in most practical situations, both diagnostic tests are conditionally dependent on the disease (Vacek, 1985; Torrance-Rynard and Walter, 1997), so that the probabilities (6) are expressed in terms of sensitivity, specificity, prevalence and dependence factors such as

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

$$\psi_{ij} = (1-p) \lambda_{ij} \left( Sp_1^{1-i} (1-Sp_1)^{1-i} Sp_2^{1-j} (1-Sp_2)^{1-j} + \delta_{ij} (1-Sp_1) (1-Sp_2) (\varepsilon_0 - 1) \right)$$

where $\lambda_{ij}$ is the conditional probability of verifying a patient with a negative test ($\lambda_0$) when the disease prevalence is $\varepsilon_1$ for tests 1 and $\varepsilon_0$ for tests 2.
and

\[ \xi_{ij} = \frac{1 - \lambda_{ij}}{\lambda_{ij}} (\xi_{ij} + \psi_{ij}), \]

where \( \delta_{ij} = 1 \) if \( i = j \) and \( \delta_{ij} = -1 \) if \( i \neq j \), and the parameters \( \xi_1 \) and \( \xi_0 \) are the dependence factors between the two diagnostic tests. The dependence factor \( \xi_1(\xi_0) \) is the covariance between both diagnostic tests when \( D = 1 \) \((D = 0) \) (Berry et al., 2002), and it is verified that \( 1 \leq \xi_1 \leq 1/\max(Se_h) \) and \( 1 \leq \xi_0 \leq 1/\max(1 - Sp_h) \). If \( \xi_1 = \xi_0 = 1 \), both diagnostic tests are conditionally independent on the disease. When \( D = 1 \), the correlation between the two diagnostic tests is

\[ \rho_1 = \text{Corr}(T = 1, T_2 = 1|D = 1) = (\xi_1 - 1) \sqrt{\frac{Se_1 Se_2}{(1 - Se_1)(1 - Se_2)}}, \]

and when \( D = 0 \) the correlation is

\[ \rho_0 = \text{Corr}(T_1 = 1, T_2 = 1|D = 0) = (\xi_0 - 1) \sqrt{\frac{(1 - Sp_1)(1 - Sp_2)}{Sp_1 Sp_2}}. \]

A positive correlation means that the proportion of concordant results (positive or negative) of both diagnostic tests is greater than when the two tests are conditionally independent. A negative correlation indicates a greater proportion of discordant results.

Let \( \omega = (\xi_{11}, \xi_{10}, \xi_{01}, \xi_{00}, \psi_{11}, \psi_{10}, \psi_{01}, \psi_{00}, \xi_{11}, \xi_{10}, \xi_{01}, \xi_{00})^T \). As \( \xi_{ij}, \psi_{ij} \) and \( \xi_{ij} \) are the probabilities of a multinomial distribution, the variance–covariance matrix of \( \omega \) is

\[ \sum_{\omega} = \text{diag}(\omega) - \omega^T \omega. \]

Let \( \phi_{ij} = \xi_{ij} + \psi_{ij} + \xi_{ij} \) with \( i, j = 0, 1 \). The sensitivity and the specificity of each diagnostic test can be written in terms of the previous probabilities as

\[ Se_1 = \sum_{j=0}^{1} \frac{\xi_{ij} \phi_{ij}}{\xi_{ij} + \psi_{ij}} / \sum_{j=0}^{1} \frac{\xi_{ij} \phi_{ij}}{\xi_{ij} + \psi_{ij}} \]

and

\[ Sp_1 = \sum_{j=0}^{1} \frac{\psi_{ij} \phi_{ij}}{\xi_{ij} + \psi_{ij}} / \sum_{j=0}^{1} \frac{\psi_{ij} \phi_{ij}}{\xi_{ij} + \psi_{ij}} \]

for diagnostic test 1, and

\[ Se_2 = \sum_{i=0}^{1} \frac{\xi_{ii} \phi_{10}}{\xi_{ii} + \psi_{10}} / \sum_{i=0}^{1} \frac{\xi_{ii} \phi_{10}}{\xi_{ii} + \psi_{10}} \]

and

\[ Sp_2 = \sum_{i=0}^{1} \frac{\psi_{i0} \phi_{10}}{\xi_{i0} + \psi_{i0}} / \sum_{i=0}^{1} \frac{\psi_{i0} \phi_{10}}{\xi_{i0} + \psi_{i0}} \]

for test 2. From these expressions, it holds that

\[ Se_1 - Se_2 = \left\{ \frac{\xi_{10} \phi_{10}}{\xi_{10} + \psi_{10}} - \frac{\xi_{01} \phi_{01}}{\xi_{01} + \psi_{01}} \right\} / \sum_{i,j=0}^{1} \frac{\xi_{ij} \phi_{ij}}{\xi_{ij} + \psi_{ij}} \]

and

\[ Sp_1 - Sp_2 = \left\{ \frac{\psi_{01} \phi_{01}}{\xi_{01} + \psi_{01}} - \frac{\psi_{10} \phi_{10}}{\xi_{10} + \psi_{10}} \right\} / \sum_{i,j=0}^{1} \frac{\psi_{ij} \phi_{ij}}{\xi_{ij} + \psi_{ij}} \]

and applying the delta method (Agresti, 2002), it holds that

\[ V(\hat{Se}_1 - \hat{Se}_2) = \left( \frac{\partial (Se_1 - Se_2)}{\partial \omega} \right) \sum_{\omega} \left( \frac{\partial (Se_1 - Se_2)}{\partial \omega} \right)^T \]

and

\[ V(\hat{Sp}_1 - \hat{Sp}_2) = \left( \frac{\partial (Sp_1 - Sp_2)}{\partial \omega} \right) \sum_{\omega} \left( \frac{\partial (Sp_1 - Sp_2)}{\partial \omega} \right)^T, \]
where the partial derivatives $\partial (\hat{Se}_1 - \hat{Se}_2) / \partial \omega$ and $\partial (\hat{Sp}_1 - \hat{Sp}_2) / \partial \omega$, and therefore $V(\hat{Se}_1 - \hat{Se}_2)$ and $V(\hat{Sp}_1 - \hat{Sp}_2)$, are long and complicated expressions which need to be assessed with mathematical or statistical software. Therefore, if from a pilot sample or some other previous study we know the sensitivity and specificity of each diagnostic test, the disease prevalence, the covariances $\varepsilon_1$ and $\varepsilon_0$ (or the correlations $\rho_1$ and $\rho_0$) and the verification probabilities, applying the Eq. (7) it is easy to calculate the sample size necessary to compare the two sensitivities (specificities) with precision $L$ to confidence 100(1 - $\alpha$)%.

The method which we propose to calculate sample size requires us to assume that all of the parameters (sensitivities, specificities, verification probabilities, prevalence and covariances) are known, whether from a pilot sample or any other previous study. The sensitivities, specificities and the prevalence can be calculated applying the equations proposed by Zhou (1998a,b), and the verification probabilities can be estimated from the pilot sample $\left(\hat{\lambda}_{ij} = (s_{ij} + r_{ij}) / n_{ij}\right)$ or can be established from verification costs. Nevertheless, the covariances $\varepsilon_1$ and $\varepsilon_0$ are not easy to calculate, since not all patients are verified with the gold standard. In Appendix A we show an EM algorithm allows us to estimate the covariances between the two diagnostic tests in the presence of partial disease verification; this EM algorithm also allows us to estimate the rest of the parameters.

The method which we have proposed in this section can also be applied when calculating the sample size to contrast $H_0 : \theta_1 = \theta_2$ with $H_1 : \theta_1 \neq \theta_2$, where $\theta$ is the sensitivity or specificity. In this situation, in order to calculate the sample size it is necessary to specify a value for the difference between $\theta_1$ and $\theta_2$, subject to the alternative hypothesis. If we denote $\delta$ as the difference between $\theta_1$ and $\theta_2$, i.e., $\delta = \theta_1 - \theta_2$, the formula to determine the sample size in paired designs is

$$m = \left(\frac{z_{1-\alpha/2} + z_\beta}{\delta}\right)^2 V(\hat{\theta}_1 - \hat{\theta}_2),$$

(8)

where $z_\gamma$ is the 100$\gamma$th percentile of the normal standard distribution, $\alpha$ is the type I error, $\beta$ is the type II error, and where the variance of $\theta_1 - \theta_2$ is calculated applying the method which we have described in this Section.

3.2. Simulation study

As in Section 2, we conducted simulation experiments to study the robustness of the method proposed to calculate sample size when estimating the difference between the two sensitivities (specificities). Therefore, we generated 5000 random samples with multinomial distributions and probabilities given as (7). As sensitivities and specificities we took the values

$$\begin{align*}
(\hat{Se}_1 = 0.90, \hat{Sp}_1 = 0.85, \hat{Se}_2 = 0.85, \hat{Sp}_2 = 0.80), \\
(\hat{Se}_1 = 0.90, \hat{Sp}_1 = 0.85, \hat{Se}_2 = 0.80, \hat{Sp}_2 = 0.70)
\end{align*}$$

and

$$\begin{align*}
(\hat{Se}_1 = 0.95, \hat{Sp}_1 = 0.95, \hat{Se}_2 = 0.90, \hat{Sp}_2 = 0.90),
\end{align*}$$

which are values which appear with a certain frequency in clinical practice; as prevalence we considered the same values as in Section 2.2, and as verification probabilities we took

$$\begin{align*}
(\lambda_{11} = 0.80, \lambda_{10} = \lambda_{01} = 0.40, \lambda_{11} = 0.10), \\
(\lambda_{11} = 0.95, \lambda_{10} = \lambda_{01} = 0.60, \lambda_{11} = 0.30)
\end{align*}$$

and

$$\begin{align*}
(\lambda_{11} = \lambda_{10} = \lambda_{01} = \lambda_{11} = 1).
\end{align*}$$

For 5000 samples with the same multinomial distribution we calculated the average sample size and the relative root mean squared error (RRMSE) in a similar way to in Section 2.2, and we consider that method of calculation of sample size is valid when the RRMSE of the estimator of sample size is lower than 25%. The sample size was calculated through the method proposed in Section 3.1, and the maximum likelihood estimators for sensitivities, specificities, prevalence and correlations were obtained applying the EM algorithm described in Appendix A; the maximum likelihood estimators for the verification probabilities were obtained through the equations $\hat{\lambda}_{ij} = (s_{ij} + r_{ij}) / n_{ij}$, with $i, j = 0, 1$. When $\lambda_{ij} = 1$, the estimators were obtained applying the classic equations in Tables 2 × 4.

In Tables 3 and 4 we show some of the results obtained for $(\hat{Se}_1 = 0.90, \hat{Sp}_1 = 0.85, \hat{Se}_2 = 0.85, \hat{Sp}_2 = 0.80)$ and different values of $\rho_1$ and $\rho_0$. From the results obtained with these experiments we can deduce the following conclusions. For sensitivities, in general terms, when the correlation $\rho_1$ between the two binary tests is low or intermediate (independent of the value of the correlation $\rho_0$), the estimator of size to compare the two sensitivities has a relative root mean square error (RRMSE) lower than 25%; while if the correlation $\rho_1$ is high, its RRMSE is higher than 25%. Regarding specificities, in general terms, when the correlation $\rho_0$ between the two binary tests is low or intermediate, the estimator of sample size to compare the two specificities has an RRMSE lower than 25%; whilst if the correlation $\rho_0$ is high its RRMSE is higher than 25%. Therefore, the correlations $(\rho_1$ and $\rho_0$) between the two diagnostic tests has an important effect on the
method which we propose to calculate sample size, and we think that this is due to the minor effect that the correlations have on the proportion of verified patients (for the same prevalence and verification probabilities), since the increase
Table 5
Application conditions for the method of calculation of sample size to compare the accuracy of two binary tests

<table>
<thead>
<tr>
<th>Comparison of the two sensitivities</th>
<th>( \rho_0 )</th>
<th>Low or intermediate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho_1 )</td>
<td>Low or intermediate</td>
<td>Method valid</td>
<td>Method not valid</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>Method not valid</td>
<td>Method not valid</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparison of the two specificities</th>
<th>( \rho_0 )</th>
<th>Low or intermediate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho_1 )</td>
<td>Low or intermediate</td>
<td>Method valid</td>
<td>Method valid</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>Method not valid</td>
<td>Method not valid</td>
</tr>
</tbody>
</table>

Table 6
Data from the study of coronary stenosis

<table>
<thead>
<tr>
<th>Evaluation of the dobutamine echocardiography</th>
<th>( T_1 = 1 )</th>
<th>( T_1 = 0 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V = 1 )</td>
<td>109</td>
<td>7</td>
<td>116</td>
</tr>
<tr>
<td>( D = 1 )</td>
<td>90</td>
<td>126</td>
<td>216</td>
</tr>
<tr>
<td>( D = 0 )</td>
<td>40</td>
<td>283</td>
<td>323</td>
</tr>
<tr>
<td>( V = 0 )</td>
<td>195</td>
<td>44</td>
<td>655</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparison of the dobutamine and stress echocardiographies</th>
<th>( T_1 = 1 )</th>
<th>( T_1 = 0 )</th>
<th>( T_2 = 1 )</th>
<th>( T_2 = 0 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V = 1 )</td>
<td>95</td>
<td>14</td>
<td>5</td>
<td>2</td>
<td>116</td>
</tr>
<tr>
<td>( D = 1 )</td>
<td>70</td>
<td>20</td>
<td>47</td>
<td>79</td>
<td>216</td>
</tr>
<tr>
<td>( D = 0 )</td>
<td>30</td>
<td>10</td>
<td>24</td>
<td>259</td>
<td>323</td>
</tr>
<tr>
<td>( V = 0 )</td>
<td>195</td>
<td>44</td>
<td>76</td>
<td>340</td>
<td>655</td>
</tr>
</tbody>
</table>

of \( \rho_1 \) and/or \( \rho_0 \) does not have an important effect on the increase in the percentage of verified patients. Regarding disease prevalence, in general terms, the increase (or decrease) in prevalence has little effect on the robustness of the method which we have proposed in the Section 3.1, above all when calculating the sample size to estimate the differences between the sensitivities. Similar conclusions are obtained for \((S_{e1} = 0.90, S_{p1} = 0.85, S_{e2} = 0.80, S_{p2} = 0.70)\) and for \((S_{e1} = 0.95, S_{p1} = 0.95, S_{e2} = 0.90, S_{p2} = 0.90)\).

In Table 5 we summarise the general conditions subject to which it is possible to apply the method which we propose to calculate the sample size to compare the sensitivities (specificities) in the presence of partial disease verification. In this table, the valid method indicates that the RRMSE of the estimator of sample size is lower than 25% and, therefore, the method that we propose to estimate the sample size offers good results, whilst a non-valid method indicates that the RRMSE of the estimator of sample size is higher than 25% and, therefore, the method that we propose to estimate sample size does not offer satisfactory results.

4. Applications

We applied the methods proposed in Sections 2 and 3 to the diagnosis of coronary stenosis. Coronary stenosis is a disease which consists of the obstruction of the coronary artery and its diagnosis can be made through a dobutamine echocardiography or a dobutamine stress echocardiography, and as the gold standard a coronary angiography is used. As the angiography can cause reactions in some patients (heart attacks, infections, thrombosis) not all of the patients are verified. In Table 6 we show the results obtained when assessing the dobutamine echocardiogram, and when comparing the dobutamine echocardiography and stress echocardiography with a sample of 655 patients. Random variable \( T_1 \) models the result of the dobutamine echocardiography, random variable \( T_2 \) models the result of the stress echocardiography and random variable \( D \) models the result of the coronary angiography. In all of the study, the selection of patients to verify disease status with the angiography is carried out only depending on the results of the dobutamine echocardiography and the stress echocardiography. The selection to verify the disease status of each patient was carried out based on the results of each diagnostic test. The study corresponds to a design in two phases: in the first phase, the two diagnostic tests are applied to all of the patients, and in the second phase, the gold standard is only applied to a subsample of patients based on the results of the diagnostic tests. Therefore, we can state that the verification process is MAR. We are going to use the results
obtained by analysing this data to illustrate how to calculate the sample size when assessing a diagnostic test and when comparing two diagnostic tests in the presence of partial disease verification.

4.1. Calculation of sample size for a binary test

We applied the method in Section 2 to the estimation of the sensitivity of the dobutamine echocardiography (see Table 6: evaluation of the dobutamine echocardiography). As the verification process is MAR, applying the results of Begg and Greener (1983) or Zhou et al. (2002) the estimated sensitivity and specificity of the dobutamine echocardiography are 0.86 and 0.78, and the corresponding 95% confidence intervals are $\hat{Se} \in (0.77, 0.95)$ and $\hat{Sp} \in (0.74, 0.82)$, and their respective precisions are 9% and 4%. The estimated disease prevalence is 23% and the estimation of the verification probabilities are $\lambda_1 = 0.83$ and $\lambda_0 = 0.32$. If we consider that the precision of the estimation of sensitivity is not large enough, we have to deal with the problem of calculating the minimum size necessary to estimate the sensitivity with a precision of 5% ($L = 0.05$) to a confidence of 95%. Using the results obtained from the sample of 665 patients, applying the method proposed in Section 2 it holds that $V \left( \hat{Se} \right) = 1.4384$ and applying Eq. (1) it holds that $m = 2211$. Therefore, we need a sample of at least 2211 patients to estimate the sensitivity with a precision of 5% to a confidence of 95%, and therefore it is necessary to add to the initial sample of 665 patients a minimum of 1546 patients in order to estimate the sensitivity with a precision of 5% to a confidence of 95%.

4.2. Calculation of sample size to compare two binary tests

We applied the results from Section 3 to the comparison of the dobutamine echocardiography and the stress echocardiography. As the verification process is MAR, applying the method proposed by Zhou (1998a,b) the estimated sensitivities and specificities of each diagnostic test and the 95% confidence intervals for the difference in sensitivity and specificity are $\hat{Se}_1 = 0.89$, $\hat{Se}_2 = 0.82$, $\hat{Sp}_1 = 0.79$, $\hat{Sp}_2 = 0.70$, $\hat{Se}_1 - \hat{Se}_2 \in (-0.0007, 0.1487)$ and $\hat{Sp}_1 - \hat{Sp}_2 \in (0.0461, 0.1221)$. Moreover, the estimated prevalence and verification probabilities are $\hat{p} = 0.22$, $\hat{\lambda}_{11} = 0.85$, $\hat{\lambda}_{10} = 0.77$, $\hat{\lambda}_{01} = 0.68$ and $\hat{\lambda}_{00} = 0.24$. Applying the EM algorithm (see Appendix A), the estimated covariances (correlations) are $\hat{\sigma}_1 = 1.052$ ($\hat{\rho}_1 = 0.32$) and $\hat{\sigma}_0 = 2.560$ ($\hat{\rho}_0 = 0.53$). If we consider that the estimation of the difference in sensitivities is not large enough, using the previous sample as a pilot sample, we have to deal with the problem of calculating the minimum sample size which is necessary in order to estimate the difference in sensitivities with a precision of 5% to a confidence of 95%. Applying the method in Section 3 it holds that $V \left( \hat{Se}_1 - \hat{Se}_2 \right) = 0.976$ and applying Eq. (5) it holds that $m = 1500$. Therefore, we need a sample of at least 1500 patients in order to estimate the difference between the two sensitivities with a precision of 5% to a confidence of 95% (it is necessary to add to the initial sample of 665 patients a minimum of 835 patients).

5. Discussion

The calculation of sample size is an important topic in the study of diagnostic statistical methods, and has been the subject of numerous studies in statistical literature. In the presence of partial disease verification, the calculation of sample size cannot be carried out using the traditional methods, since the accuracy of the diagnostic test cannot be estimated as a binomial proportion, since a sub sample of patients does not have its disease status verified. In this study, we propose a method to calculate sample size when we evaluate the accuracy of a single binary test and when we compare the accuracy of two binary tests when not all of the patients are verified with the gold standard. In each case, the method which we propose requires the verification to not depend on the result of the gold standard. Therefore, the method which we propose to calculate sample size (both when trying to assess the accuracy of a single binary test and when trying to compare the accuracy of two binary tests) requires that the verification process be MAR. Nevertheless, in some practical situations the verification process does not only depend on the result of the diagnostic test (or on the results of the diagnostic tests) but also depends on disease status. This, if the verification process depends on non-observed variables which are related to the disease status, then the verification process is not MAR and, therefore, the method which we propose cannot be applied. It is also necessary to know the values of sensitivity and specificity of each test, disease prevalence, covariances between the two diagnostic tests (when two binary tests are compared) and the verification probabilities (values which can be conjectured through an initial study or a pilot sample), as well as the specifications to calculate sample size (confidence and precision). We carried out simulation experiments to study the robustness of the method which we have proposed to calculate the sample size, both when assessing a single diagnostic test and when comparing two diagnostic tests. In the first case (a single diagnostic test), we have studied the effect that verification probabilities and prevalence have on sample size, concluding that the method which we have proposed offers very satisfactory results, since the simulation experiments have shown that, in all of the cases studied, the relative root mean square errors are always lower than 25%, which indicates the validity of our method to calculate the sample size when assessing the accuracy of a binary test in the presence of partial disease verification. In the second case (two binary tests) we have studied the effect that prevalence, verification probabilities and the correlations ($\rho_1$ and $\rho_0$) have on sample size to estimate the difference in sensitivities (specificities) to a precision $L$ with a confidence $100 (1 - \alpha) \%$. The results of the simulation experiments have shown that the correlations $\rho_1$ and $\rho_0$ are the
parameters which have the greatest effect on sample size and it holds that the method which we propose to calculate sample size to estimate the difference in sensitivities (specificities) is valid when the correlation $\rho_1$ ($\rho_0$) is low or intermediate, and is not a good method when the correlation $\rho_1$ ($\rho_0$) is high (see Table 5). These correlations can be estimated from a pilot sample, or from previous studies, applying the EM algorithm which we propose in Appendix A, and their corresponding standard errors can be estimated applying the SEM algorithm (Meng and Rubin, 1991). Another alternative method to estimate the correlations (or covariances) involves applying the bootstrap method in a similar way to that used in the study conducted by Royston et al. (2003).

The problem which we study in this article could be posed as a study of power when the sample size is known. When comparing two binary tests, with an established sample size, the power can be determined from the Eq. (8). Nevertheless, when assessing a single binary test, the problem is rather more complicated. In this situation, the sample size to contrast $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, where $\theta$ is the sensitivity or specificity, is

$$m = \left( \frac{z_{\alpha/2} \sqrt{V_0(\hat{\theta})} + z_{\beta} \sqrt{V_1(\hat{\theta})}}{\delta} \right)^2,$$

where $\delta = \theta - \theta_0$, $V_0(\hat{\theta})$ is the variance function of $\hat{\theta}$ subject to the null hypothesis and $V_1(\hat{\theta})$ is the variance function of $\hat{\theta}$ subject to the alternative hypothesis, and where both variance functions are calculated applying the method used in Section 2.

Furthermore, an important question when assessing the accuracy of a diagnostic test in the presence of partial verification is how to determine the strategy to establish the verification probabilities. Although this question requires a more complete study, from the simulation experiments carried out in Sections 2.2 and 3.2, in general terms, we can obtain some conclusions. When we seek to calculate the sample size to estimate sensitivity to a precision $L$ to a confidence $100(1 - \alpha)\%$, it is preferable to take as verification probabilities high values rather than small ones, since the cost of verifying patients with the gold standard is lower. This is due to the fact that the sample size is smaller when the verification probabilities are higher than when these probabilities are low, and the average number of verified patients is lower when the verification probabilities are high than when these probabilities are low. Thus, for example, for $Se = 0.80$, $Sp = 0.70$ and $p = 50\%$, if $\lambda_1 = 0.80$ and $\lambda_0 = 0.10$, the sample size necessary to estimate the sensitivity with a precision $L = 5\%$ and a confidence level of 95\% is $m = 3253$, verifying on average 48.5\% of patients (1578 patients); if $\lambda_1 = 0.95$ and $\lambda_0 = 0.30$, the sample size is $m = 1208$, verifying 65.7\% of patients (794 patients on average). Therefore, in the second case the medical costs of verification are reduced, due to the fact that the sample size is much smaller. Nevertheless, when we seek to calculate the sample size to estimate the specificity to a precision $L$ to a confidence $100(1 - \alpha)\%$, the disease prevalence has an important effect in terms of establishing the verification probabilities. In general, when $p \leq 30\%$ it is recommendable to establish low verification probabilities so that the verification costs are lower, since the average number of verified patients is lower when the verification probabilities are low than when these probabilities are high; and when $p > 30\%$ it is recommendable to establish high verification probabilities, since the average number of verified patients is lower when the verification probabilities are high than when they are low. This same question arises when comparing the accuracy of two binary tests in the presence of partial verification. When calculating the sample size to estimate the difference between the two sensitivities, for the same value of prevalence and the correlations $\rho_1$ and $\rho_0$, it is recommendable to establish low verification probabilities, since the average number of verified patients is lower when the verification probabilities $\lambda_{ij}$ are low than when they are high. Nevertheless, when calculating the sample size to estimate the difference between the specificities, disease prevalence once again has an important effect on the choice of verification probabilities. In general, when $p \leq 50\%$ it is recommendable to establish low verification probabilities (for the same reasons as before), and when $p > 50\%$ it is recommendable to establish high verification probabilities.

On the other hand, the method which we propose to calculate the sample size is based on the normal approach to estimators. When this approach is not valid, the method cannot be applied. Another limitation of our method lies in the fact that the equations of sample size are based on Wald confidence intervals, so that when the values of the accuracy of the diagnostic test are very near to 0 or to 1 (Agresti and Coull, 1998; Brown et al., 2001), the method used to calculate the sample size is affected by the bad performance of the confidence interval. We relieve that a possible solution to these problems could be found by applying exact inference methods. Therefore, it is necessary to carry out further research into exact inference methods which allows us to calculate the sample size when trying to assess or compare the accuracy of binary diagnostic tests in the presence of partial disease verification. Another solution to this problem could be the application of multiple imputation and the calculation of sample size applying the methods in $2 \times 2$ tables.

Finally, the method which we propose in Section 3 can be applied when we try to calculate the sample size necessary to compare the accuracy of more than two binary tests in the presence of partial verification. In this new situation, we take as the value of $\delta$ in Eq. (8) the greatest difference between the sensitivities (specificities), and $V(\hat{\theta}_i - \hat{\theta}_j)$ is the variance function of the difference between the sensitivities which give the value $\delta$. Thus, for three binary tests it is necessary to know beforehand (through a pilot sample or a previous study) the values of the sensitivities, specificities, prevalence, verification probabilities and covariances between the three binary tests. The problem is more complex than in the case of two binary
tests, since the covariances between the three binary tests are difficult to estimate, and therefore an adaptation of the method used by Royston et al. (2003) could be the ideal method to solve this problem.

Acknowledgements

This research was supported by Consejería de Innovación, Ciencia y Empresa, Junta de Andalucía, Spain, grant number FQM-01459. We thank the two referees, Professor Norbert Victor (editor associate) and editor of Computational Statistics and Data Analysis for their helpful comments that improved the quality of the paper.

Appendix

Roldán Nofuentes and Luna del Castillo (2007e) deduced an EM algorithm to estimate the accuracy of two binary tests in the presence of partial disease verification and which allows us to estimate the covariances between the two binary tests. In the situation of partial verification, the missing information is the result of the gold standard of each non-verified patient. This information is reconstructed in step E of the algorithm, and in step M we impute the maximum likelihood estimators from the reconstructed data in the previous step. Let us suppose that from each frequency $u_{ij}$ of non-verified patients $x_i$ patients are diseased and $u_{ij} - x_i$ patients are non-diseased. Therefore, the data observed can be expressed in the form of a $2 \times 4$ table with $s_j + x_{ij}$ frequencies for $D = 1$ and $r_j + u_{ij} - x_{ij}$ for $D = 0$. Subject to the MAR assumption, the logarithm of the likelihood function of the data in this $2 \times 4$ table is

$$
\log l \propto \sum_{i,j=0}^{1} (s_{ij} + x_{ij}) \log \{P(T_1 = i, T_2 = j, D = 1)\} + \sum_{i,j=0}^{1} (r_j + u_{ij} - x_{ij}) \log \{P(T_1 = i, T_2 = j, D = 0)\}
$$

Let $Se_1$ and $Sp_1$ be the sensitivity and specificity of each diagnostic test, $p$ be the disease prevalence and $\alpha_k$ the covariance between the two diagnostic tests when $D = k$. Subject to the MAR assumption, let $x^{(k)}_i$ be the values of $x_i$ in the $k$th iteration of the EM algorithm. The values of the MLEs in the $k$th iteration are calculated through the expressions:

$$
\hat{Se}_1^{(k)} = \frac{\sum_{i=0}^{1} (s_{ij} + x^{(k)}_{ij})}{s + x^{(k)}}, \\
\hat{Sp}_1^{(k)} = \frac{\sum_{i=0}^{1} (r_j + u_{ij} - x^{(k)}_i)}{r + u - x^{(k)}}, \\
\hat{Se}_2^{(k)} = \frac{\sum_{i=0}^{1} (s_{ij} + x^{(k)}_{ij})}{s + x^{(k)}}, \\
\hat{Sp}_2^{(k)} = \frac{\sum_{i=0}^{1} (r_j + u_{ij} - x^{(k)}_i)}{r + u - x^{(k)}},
$$

and

$$
\hat{p}^{(k)} = \frac{s + x^{(k)}}{n}, \\
\hat{\varepsilon}^{(k)}_1 = \frac{(s + x^{(k)})(s_{11} + x^{(k)}_{11})}{\left\{\sum_{i=0}^{1} (s_{ij} + x^{(k)}_{ij})\right\} \left\{\sum_{j=0}^{1} (s_{ij} + x^{(k)}_{ij})\right\}},
$$

and

$$
\hat{\varepsilon}^{(k)}_0 = \frac{(u + r - x^{(k)})(r_{11} + u_{11} - x^{(k)}_{11})}{\left\{\sum_{i=0}^{1} (r_{ij} + u_{ij} - x^{(k)}_{ij})\right\} \left\{\sum_{j=0}^{1} (r_{ij} + u_{ij} - x^{(k)}_{ij})\right\}},
$$

when $s = \sum_{i,j=0}^{1} s_{ij}$, $r = \sum_{i,j=0}^{1} r_{ij}$, $u = \sum_{i,j=0}^{1} u_{ij}$, $x^{(k)} = \sum_{i,j=0}^{1} x^{(k)}_{ij}$ and $n = s + r + u$. The estimators in the following iteration are

$$
\hat{Se}_1^{(k+1)} = \frac{\sum_{i=0}^{1} (s_{ij} + x^{(k+1)}_{ij})}{s + x^{(k+1)}}, \\
\hat{Sp}_1^{(k+1)} = \frac{\sum_{i=0}^{1} (r_j + u_{ij} - x^{(k+1)}_i)}{r + u - x^{(k+1)}}, \\
\hat{Se}_2^{(k+1)} = \frac{\sum_{i=0}^{1} (s_{ij} + x^{(k+1)}_{ij})}{s + x^{(k+1)}}, \\
\hat{Sp}_2^{(k+1)} = \frac{\sum_{i=0}^{1} (r_j + u_{ij} - x^{(k+1)}_i)}{r + u - x^{(k+1)}},
$$

and

$$
\hat{p}^{(k+1)} = \frac{s + x^{(k+1)}}{n}, \\
\hat{\varepsilon}^{(k+1)}_1 = \frac{(s + x^{(k+1)})(s_{11} + x^{(k+1)}_{11})}{\left\{\sum_{i=0}^{1} (s_{ij} + x^{(k+1)}_{ij})\right\} \left\{\sum_{j=0}^{1} (s_{ij} + x^{(k+1)}_{ij})\right\}},
$$

and

$$
\hat{\varepsilon}^{(k+1)}_0 = \frac{(u + r - x^{(k+1)})(r_{11} + u_{11} - x^{(k+1)}_{11})}{\left\{\sum_{i=0}^{1} (r_{ij} + u_{ij} - x^{(k+1)}_{ij})\right\} \left\{\sum_{j=0}^{1} (r_{ij} + u_{ij} - x^{(k+1)}_{ij})\right\}}.
and

\[ z_0^{(k+1)} = \left( u + r - x_0^{(k+1)} \right) \left( r_{11} + u_{11} - x_{11}^{(k+1)} \right) \left\{ \sum_{i=0}^{1} \left( r_{i1} + u_{i1} - x_{i1}^{(k+1)} \right) \right\}^{-1} \left\{ \frac{1}{2} \sum_{j=0}^{1} \left( r_{ij} + u_{ij} - x_{ij}^{(k+1)} \right) \right\}, \]

where

\[ x_{ij}^{(k+1)} = u_{ij} P^{(k)} (T_1 = i, T_2 = j, D = 1) \]

and \( P^{(k)} (T_1 = i, T_2 = j, D) = P (T_1 = i, T_2 = j, D = 0) \) in the kth iteration of the EM algorithm. This process is repeated until the estimators converge. As initial values \( x_{ij}^{(0)} \) we can take any value between 0 and \( u_{ij} \). The EM algorithm stops when the difference in the values of the logarithms of the likelihood function of two consecutive iterations is less than a sufficiently small value, e.g. \( 10^{-10} \) or \( 10^{-12} \). The EM algorithm proposed to estimate the covariates between the two binary tests can also be used to obtain the values of the estimators of sensitivity and specificity for each diagnostic test. In this situation, the estimations of the variances–covariances of the estimators can be obtained applying the SEM algorithm (Meng and Rubin, 1991).

References

Berry, G., Smith, C.L., Macaskill, P., Irwig, L., 2002. Analytic methods for comparing two dichomotous screening or diagnostic tests applied to two populations of differing disease prevalence when individuals negative on both tests are unverfied. Statistics in Medicine 21, 853–862.