NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Comparison of predictive values from two diagnostic tests in large samples

by

Clara-Cecilie Günther, Øyvind Bakke, Stian Lydersen and
Mette Langaas

PREPRINT STATISTICS NO. 9/2008
DEPARTMENT OF MATHEMATICAL SCIENCES

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
`http://www.math.ntnu.no/preprint/statistics/2008/S9-2008.pdf`

Mette Langaas has homepage: `http://www.math.ntnu.no/~mettela`
E-mail: `Mette.Langaas@math.ntnu.no`
Address: Department of Mathematical Sciences, Norwegian University of Science and
Technology, N-7491 Trondheim, Norway.

Comparison of predictive values from two diagnostic tests in large samples

Clara-Cecilie Günther*, Øyvind Bakke*, Stian Lydersen# and Mette Langaas*
* Department of Mathematical Sciences.
# Department of Cancer Research and Molecular Medicine.
The Norwegian University of Science and Technology,
NO-7491 Trondheim, Norway.

December 2008

Summary

Within the field of diagnostic tests, the positive predictive value is the probability of being diseased given that the diagnostic test is positive. Two diagnostic tests are applied to each subject in a study and in this report we look at statistical hypothesis tests for large samples to compare the positive predictive values of the two diagnostic tests. We propose a likelihood ratio test and a restricted difference test, and we perform simulation experiments to compare these tests with existing tests. For comparing the negative predictive values of the diagnostic tests, i.e. the probability of not being diseased given that the test is negative, we propose negative predictive versions of the same tests. The simulation experiments show that the restricted difference test performs well in terms of test size.

1 Introduction

Diagnostic tests are used in medicine to e.g. detect diseases and give prognoses. Diagnostic tests can for example be based on blood samples, X-ray scans, mammography, ultrasound or computed tomography (CT). Mammography is used for detecting breast cancer, a blood sample may show if an individual has an infection, fractures may be detected from X-ray images, gallstones in the gallbladder can be found using ultrasound, and CT scans are useful for identifying tumours in the liver. A diagnostic test can have several outcomes or the outcome may be continuous, but it can often be dichotomized in terms of presence or absence of a disease and we will only consider diagnostic tests for which the disease status is binary.

When evaluating the performance of diagnostic tests, the sensitivity, specificity and positive and negative predictive values are the common accuracy measures. The sensitivity and specificity are probabilities of the test outcomes given the disease status. The sensitivity is the probability of a positive test outcome given that the disease is present and the specificity is the probability of a negative outcome given no disease. These measures tell us the degree to which the test reflects the true disease status.

The predictive values are probabilities of disease given the test result. The positive predictive value (PPV) is the probability that a subject who has a positive test outcome has the disease and the negative predictive value (NPV) is the probability that a subject who has a negative test outcome does not have the disease. The predictive values give information about the prediction capabilities of the test. For a perfect test both the PPV and NPV are 1, the test result will then give the true disease status for each subject.

When there are several diagnostic tests available for the same disease, we are interested in knowing which test is the best to use, but depending on what we mean by best, there are different methods available. If we want to find the best test regarding the ability to give a correct test outcome given the disease status then e.g. McNemar's test, see Alan Agresti (2002), can be used for comparing the sensitivity or specificity of two tests evaluated on the same subjects.

A test that has a high sensitivity and specificity will be most likely to give the patient the correct test result. However, for the patient it is utterly important to be correctly diagnosed and thereby getting the right treatment. We need to take into account the prevalence of the disease. If the prevalence is low, the probability that the patient does have the disease when the test result is positive, will be small even if the sensitivity of the applied test is high. Therefore, comparing the positive or negative predictive values is often more relevant in clinical practise as discussed by Guggenmoos-Holzmann and van Houwelingen (2000).

In the remainder of this work, we wish to test if the positive or negative predictive values of two diagnostic tests are equal. In this report we apply existing tests by Leisenring, Alonzo and Pepe (2000) and Wang, Davis and Soong (2006), we propose a likelihood ratio test, and suggest improvements for some of the already existing tests in the large sample case.

In Section 2 we describe the model and the structure of the data and define the predictive values. The null hypothesis, along with our proposed methods and already existing methods are presented in Section 3. A simulation study is conducted to compare the methods in Section 4. In Section 5 the methods are applied to data from the literature. We also present an alternative model and test statistic for the likelihood ratio test in Section 6. The results are summarised in the conclusions in Section 7.

2 Model and data

Next we define the random variables and the model used to describe the situation when comparing the predictive values.

2.1 Definitions

Two tests, test A and test B, are evaluated on each subject in a study. Each test can have a positive or negative outcome, i.e. indicating whether the subject has the disease under study or not. The true disease status for each subject is assumed to be known. For each subject, we define three events:

- D: The subject has the disease.
- A: Test A is positive.
- B: Test B is positive.

Let D^*, A^* and B^* denote the complementary events. The situation can then be illustrated by a Venn diagram as in Figure 1. There are eight mutually exclusive events and we define the random variable

N_i, $i = 1, ..., 8$, to be the number of times event i occurs. In total there are $N = N_1 + \ldots + N_8$ subjects in the study. Table 1 gives an overview of the notation for the eight random variables in terms of the events A, B, D and their complements.

Notation	Alternative notation	Explanation
N_1	$N_{A \cap B \cap D^*}$	number of non-diseased subjects with positive tests A and B
N_2	$N_{A \cap B^* \cap D^*}$	number of non-diseased subjects with positive test A and negative test B
N_3	$N_{A^* \cap B \cap D^*}$	number of non-diseased subjects with negative test A and positive test B
N_4	$N_{A^* \cap B^* \cap D^*}$	number of non-diseased subjects with negative tests A and B
N_5	$N_{A \cap B \cap D}$	number of diseased subjects with positive tests A and B
N_6	$N_{A \cap B^* \cap D}$	number of diseased subjects with positive test A and negative test B
N_7	$N_{A^* \cap B \cap D}$	number of diseased subjects with negative test A and positive test B
N_8	$N_{A^* \cap B^* \cap D}$	number of diseased subjects with negative tests A and B

TABLE 1: Notation for the random variables defined by the events A, B and D and their complements.

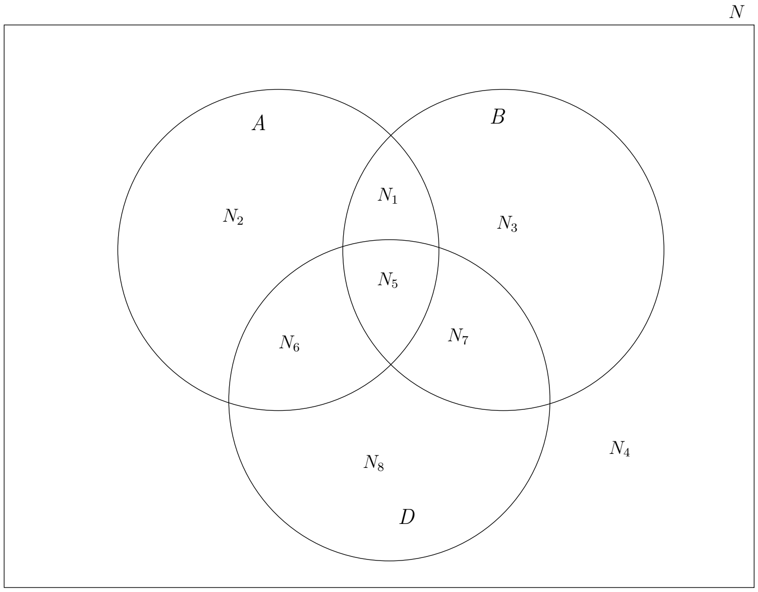

FIGURE 1: Venn diagram for the events D, A and B showing which events the random variables $N_1, ..., N_8$ correspond to.

To each of the eight mutually exclusive events there corresponds an unknown probability p_i, $i = 1, \ldots, 8$, where $\sum_{i=1}^{8} p_i = 1$, which is the probability that event i occurs for a randomly chosen subject. The positive predictive values of test A and test B can be expressed in terms of these probabilities and are given as

$$\text{PPV}_A = P(D|A) = \frac{P(D \cap A)}{P(A)} = \frac{p_5 + p_6}{p_1 + p_2 + p_5 + p_6}$$

and

$$\text{PPV}_B = P(D|B) = \frac{P(D \cap B)}{P(B)} = \frac{p_5 + p_7}{p_1 + p_3 + p_5 + p_7}.$$

Similarly, the negative predictive values of test A and B are

$$\text{NPV}_A = P(D^*|A^*) = \frac{P(D^* \cap A^*)}{P(A^*)} = \frac{p_3 + p_4}{p_3 + p_4 + p_7 + p_8}$$

and

$$\text{NPV}_B = P(D^*|B^*) = \frac{P(D^* \cap B^*)}{P(B^*)} = \frac{p_2 + p_4}{p_2 + p_4 + p_6 + p_8}.$$

The predictive values are dependent on the prevalence of the disease, $P(D)$, which is the probability that a randomly chosen subject has the disease. For the positive predictive value,

$$\text{PPV}_A = P(D|A) = \frac{P(D \cap A)}{P(A)} = \frac{P(A|D) \cdot P(D)}{P(A|D) \cdot P(D) + (1 - P(A^*|D^*)) \cdot (1 - P(D))},$$

where $P(A|D)$ is the sensitivity and $P(A^*|D^*)$ is the specificity of test A. When $P(A) = P(B)$ testing if $\text{PPV}_A = \text{PPV}_B$ is equivalent to testing if $P(A|D) = P(B|D)$, i.e. testing whether the sensitivities of the two tests are equal. We assume that the prevalence among the subjects in the study is the same as the prevalence in the population, and this can be achieved with a cohort study in which the subjects are randomly selected.

2.2 The multinomial model

Given the total number of subjects N in the study, the random variables $N_1, N_2, ..., N_8$ can be seen to be multinomially distributed with parameters $\boldsymbol{p} = (p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8)$ and N, where $\sum_{i=1}^{8} p_i = 1$. The joint probability distribution of $N_1, N_2, ..., N_8$ is

$$P\left(\bigcap_{i=1}^{8}(N_i = n_i)\right) = N! \prod_{i=1}^{8} \frac{{p_i}^{n_i}}{n_i!}.$$

The expectation of N_i is

$$\text{E}(N_i) = \mu_i = Np_i$$

for $i = 1, ..., 8$, and the variance is

$$\text{Var}(N_i) = Np_i(1 - p_i).$$

The covariance between N_i and N_j is

$$\text{Cov}(N_i, N_j) = -Np_ip_j$$

for $i \neq j$. This leads to the covariance matrix

$$\boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{N}) = N(\text{Diag}(\boldsymbol{p}) - \boldsymbol{p}^T\boldsymbol{p}),$$

for the multinomial distribution, Johnson, Kotz and Balakrishan (1997). The general unrestricted maximum likelihood estimator of p_i is

$$\hat{p}_i = n_i/N \tag{1}$$

for $i = 1, ..., 8$.

2.3 Data

For a number of subjects under study, we observe for each $i = 1, ..., 8$, the number of times event i occurs among the N subjects, n_i. Table 2 shows the observed data in a 2^3 contingency table. In the following, let $\boldsymbol{n} = (n_1, n_2, n_3, n_4, n_5, n_6, n_7, n_8)$ be the vector of the observed data. Using the unrestricted maximum likelihood estimators for $\boldsymbol{p}$, we can then estimate the positive and negative predictive values of test A and B as follows:

$$\widehat{\text{PPV}}_A = \frac{n_5 + n_6}{n_1 + n_2 + n_5 + n_6}, \quad \widehat{\text{PPV}}_B = \frac{n_5 + n_7}{n_1 + n_3 + n_5 + n_7}$$

$$\widehat{\text{NPV}}_A = \frac{n_3 + n_4}{n_3 + n_4 + n_7 + n_8}, \quad \widehat{\text{NPV}}_B = \frac{n_2 + n_4}{n_2 + n_4 + n_6 + n_8}.$$

		Subjects without disease		Subjects with disease	
		Test B		Test B	
		$+$	$-$	$+$	$-$
Test A	$+$	n_1	n_2	n_5	n_6
	$-$	n_3	n_4	n_7	n_8

TABLE 2: Observed data $n_1, ..., n_8$ presented in a 2^3 contingency table.

3 Method

Assume that we would like to test the null hypothesis that the positive predictive values are equal for test A and B, i.e. $\text{PPV}_A = \text{PPV}_B$. The null hypothesis can be written as

$$\text{H}_0^P : P(D|A) = P(D|B), \text{ i.e. } \text{H}_0^P : \frac{p_5 + p_6}{p_1 + p_2 + p_5 + p_6} = \frac{p_5 + p_7}{p_1 + p_3 + p_5 + p_7}. \tag{2}$$

Alternatively, if we would like to test whether the negative predictive values are equal for test A and B, i.e. if $\text{NPV}_A = \text{NPV}_B$, the null hypothesis is

$$\text{H}_0^N : P(D^*|A^*) = P(D^*|B^*), \text{ i.e. } \text{H}_0^N : \frac{p_3 + p_4}{p_3 + p_4 + p_7 + p_8} = \frac{p_2 + p_4}{p_2 + p_4 + p_6 + p_8}. \tag{3}$$

Our alternative hypotheses will be that the predictive values are not equal, i.e.
$H_1^P : P(D|A) \neq P(D|B)$ and $H_1^N : P(D^*|A^*) \neq P(D^*|B^*)$.

3.1 Likelihood ratio test

One possibility to test the null hypothesis in (2) is to use a likelihood ratio test. We first write down the test statistic and then describe how to find the maximum likelihood estimates of parameters.

3.1.1 Test statistic

In a general setting, if we want to test $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ versus $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0^c$ where $\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_0^c = \boldsymbol{\Theta}$ and $\boldsymbol{\Theta}$ denotes the entire parameter space, we may use a likelihood ratio test. This approach was also suggested by Leisenring et al. (2000), who faced numerical difficulties trying to implement it. The likelihood ratio test statistic is in general defined as

$$\lambda(\boldsymbol{n}) = \frac{\sup_{\boldsymbol{\Theta}_0} L(\boldsymbol{\theta}|\boldsymbol{n})}{\sup_{\boldsymbol{\Theta}} L(\boldsymbol{\theta}|\boldsymbol{n})}$$

where $\boldsymbol{n}$ is the observed data, Casella and Berger (2002). The denominator of $\lambda(\boldsymbol{n})$ is the maximum likelihood of the observed sample over the entire parameter space and the numerator is the maximum likelihood of the observed sample over the parameters satisfying the null hypothesis. Let $\boldsymbol{N} = (N_1, N_2, N_3, N_4, N_5, N_6, N_7, N_8)$ be the vector of the random variables. When the sample size is large,

$$-2 \cdot \log\lambda(\boldsymbol{N}) \approx \chi^2_k$$

i.e. $-2 \cdot \log\lambda(\boldsymbol{N})$ is χ^2 distributed with k degrees of freedom where k is the difference between the number of free parameters in the unrestricted case and under the null hypothesis.

Let $\boldsymbol{\theta} = \boldsymbol{p} = (p_1, \dots, p_8)$ be the parameters in the multinomial distribution and $\boldsymbol{n} = (n_1, \dots, n_8)$ the observed data. The log-likelihood to be maximized for the multinomial distribution is

$$l(\boldsymbol{p}) = \log L(\boldsymbol{p}|\boldsymbol{n}) = c + \sum_{i=1}^{8} n_i \cdot \log(p_i) \tag{4}$$

where c is a constant.

The sum of p_1, p_2, ..., p_8 must equal 1,

$$\sum_{i=1}^{8} p_i = 1. \tag{5}$$

Under the null hypothesis that the positive predictive values for the two tests are equal, their difference δ_P is zero, i.e.

$$\delta_P = \frac{p_5 + p_6}{p_1 + p_2 + p_5 + p_6} - \frac{p_5 + p_7}{p_1 + p_3 + p_5 + p_7} = 0. \tag{6}$$

In the unrestricted case (i.e. $H_0 \cup H_1$), the maximum likelihood estimates for $p_1, \dots, p_8$ are the estimates given by (1), which satisfy (5). Under the null hypothesis, the estimates cannot be given in closed form and we will need to use an optimization routine to estimate $p_1, \dots, p_8$ by maximizing the log-likelihood (4) under the constraints (5) and (6).

Let $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5, \hat{p}_6, \hat{p}_7, \hat{p}_8)$ be the unconstrained maximum likelihood estimates and $\tilde{\boldsymbol{p}} = (\tilde{p}_1, \tilde{p}_2, \tilde{p}_3, \tilde{p}_4, \tilde{p}_5, \tilde{p}_6, \tilde{p}_7, \tilde{p}_8)$ the maximum likelihood estimates under the null hypothesis. Then, in our model, asymptotically as N is large,

$$-2 \cdot \log(\lambda(\boldsymbol{n})) = -2 \left(\sum_{i=1}^{8} n_i \cdot (\log(\tilde{p}_i) - \log(\hat{p}_i)) \right) \approx \chi^2_1. \tag{7}$$

We have one less free parameter in the restricted case because of the constraint (6).

For testing whether the negative predictive values for the two tests are equal, the constraint δ_P (6) is replaced by δ_N, where

$$\delta_N = \frac{p_3 + p_4}{p_3 + p_4 + p_7 + p_8} - \frac{p_2 + p_4}{p_2 + p_4 + p_6 + p_8} = 0. \tag{8}$$

3.1.2 Finding maximum likelihood estimates under the null hypotheses

To find the maximum likelihood estimates under the null hypothesis, we can either maximize the likelihood function under the given constraints using a numerical optimization routine or find the estimates analytically by solving a system of equations. In both approaches we use Lagrange multipliers and in either case we have two constraints.

Numerical maximization of the log-likelihood If we want to find the maximum likelihood estimates using an optimization routine, the goal is to find the values $\tilde{\boldsymbol{p}}$ under the null hypothesis such that $\log L(\tilde{\boldsymbol{p}}) \geq \log L(\boldsymbol{p})$ for all $\boldsymbol{p}$ that satisfies the two constraints (5) and (6).

To maximize the log-likelihood (4) under the null hypotheses, we use the R interface version of TANGO (Trustable Algorithms for Nonlinear General Optimization), see Andreani, Birgin E. G., Martinez and Schuverdt (2007) and Andreani, Birgin, Martinez and Schuverdt (2008), which is a set of Fortran routines for optimization. In order to run the program, one must specify the objective function and the constraint and their corresponding first order derivatives. We reparametrize the problem by setting

$$\begin{aligned}
p_1 &= \frac{1}{1 + e^{y_1} + ... + e^{y_7}}, \\
p_2 &= \frac{e^{y_1}}{1 + e^{y_1} + e^{y_2} + ... + e^{y_7}}, \\
p_3 &= \frac{e^{y_2}}{1 + e^{y_1} + e^{y_2} + ... + e^{y_7}}, \\
&\vdots \\
p_8 &= \frac{e^{y_7}}{1 + e^{y_1} + e^{y_2} + ... + e^{y_7}}
\end{aligned}$$

where $-\infty < y_i < \infty$, $i = 1, \ldots, 7$. This reparametrization ensures that the constraint (5) is satisfied, in addition to restricting the estimated probabilities to be $0 \leq p_i \leq 1$, $i = 1, \ldots, 8$. Let $\boldsymbol{y} = (y_1, y_2, y_3, y_4, y_5, y_6, y_7)$. The constraint under the null hypothesis (2) is then

$$\delta_{P,\boldsymbol{y}} = \frac{e^{y_4} + e^{y_5}}{1 + e^{y_1} + e^{y_4} + e^{y_5}} - \frac{e^{y_4} + e^{y_6}}{1 + e^{y_2} + e^{y_4} + e^{y_6}} = 0 \tag{9}$$

and the constraint under the null hypothesis (3) is

$$\delta_{N,\boldsymbol{y}} = \frac{e^{y_2} + e^{y_3}}{e^{y_2} + e^{y_3} + e^{y_6} + e^{y_7}} - \frac{e^{y_1} + e^{y_3}}{e^{y_1} + e^{y_3} + e^{y_5} + e^{y_7}} = 0. \tag{10}$$

These constraints are both non-linear equality constraints. The TANGO program uses an augmented Lagrangian algorithm to find the minimum of the negative log-likelihood while ensuring that the H_0 constraints (9) and (10) are satisfied when testing the null hypotheses (2) and (3) respectively. The

Lagrangian multiplier is updated successively starting by an initial value that must be set. We also set the initial value of $\boldsymbol{y}$ and its lower and upper bounds. The value of $\boldsymbol{y}$ at the optimum is returned. Some computational remarks are given in Appendix D.

ANALYTICAL MAXIMIZATION OF THE LOG-LIKELIHOOD Another approach is to find the estimates analytically by solving a system of equations arising from the method of Lagrange multipliers, for an introduction see Edwards and Penney (1998). The constraint under the null hypothesis can be rewritten as

$$k(\boldsymbol{p}) = p_1p_7 + p_2p_7 + p_2p_5 - p_1p_6 - p_3p_5 - p_3p_6 = 0. \tag{11}$$

In addition, let $h(\boldsymbol{p})$ be the constraint that $p_1, ..., p_8$ must sum to one,

$$h(\boldsymbol{p}) = \sum_{i=1}^{8} p_i = 1, \tag{12}$$

and let $l(\boldsymbol{p})$ be the log-likelihood function given in (4).

The system of equations to be solved then consists of

$$\nabla l = \gamma \nabla h + \kappa \nabla k \tag{13}$$

where γ and κ are Lagrangian multipliers, together with the above constraints.

The partial derivatives of the log-likelihood l and the constraints h and k with respect to p_1, p_2, p_3, p_4, p_5, p_6, p_7 and p_8 are given by

$$\nabla l = \left(\frac{n_1}{p_1}, \frac{n_2}{p_2}, \frac{n_3}{p_3}, \frac{n_4}{p_4}, \frac{n_5}{p_5}, \frac{n_6}{p_6}, \frac{n_7}{p_7}, \frac{n_8}{p_8} \right), \tag{14}$$

$$\nabla k = (p_7 - p_6, p_5 + p_7, -p_5 - p_6, 0, p_2 - p_3, -p_1 - p_3, p_1 + p_2, 0), \tag{15}$$

and

$$\nabla h = (1, 1, 1, 1, 1, 1, 1, 1). \tag{16}$$

From Equations (11) – (16) we obtain the following system of equations, which consists of ten equations and ten unknown variables

$$\begin{aligned}
n_1 &= p_1(\gamma + \kappa(p_7 - p_6)) \\
n_2 &= p_2(\gamma + \kappa(p_5 + p_7)) \\
n_3 &= p_3(\gamma + \kappa(-p_5 - p_6)) \\
n_4 &= p_4\gamma \\
n_5 &= p_5(\gamma + \kappa(p_2 - p_3)) \\
n_6 &= p_6(\gamma + \kappa(-p_1 - p_3)) \\
n_7 &= p_7(\gamma + \kappa(p_1 + p_2)) \\
n_8 &= p_8\gamma \\
\sum_{i=1}^{8} p_i &= 1 \\
p_1p_7 + p_2p_7 + p_2p_5 - p_1p_6 - p_3p_5 - p_3p_6 &= 0.
\end{aligned} \tag{17}$$

The denominators of (14) have been multiplied over to the right hand side in order to allow for $p_i = 0$ as a possible solution for $n_i = 0$. Obviously, l cannot have a maximum value $p_i = 0$ if $n_i \neq 0$, as $l(\boldsymbol{p})$ would be $-\infty$ in this case. The solutions of this system of equations involve roots of third degree polynomials, and we have used the Maple 12 command solve to find solutions. Among its solutions, the one that maximizes $l(\boldsymbol{p})$ and where all $p_i \geq 0$ yields the likelihood estimates $\tilde{\boldsymbol{p}}_i$ under the null hypothesis. We can show that when $n_i = 0$, the corresponding likelihood estimate under the null hypothesis $\tilde{p}_i$ is 0 for $i = 1, 4, 5, 8$, but that this is not necessarily true for $i = 2, 3, 6, 7$. For $\tilde{p}_4$ and $\tilde{p}_8$ we have the more general result that $\tilde{p}_4 = n_4/N$ and $\tilde{p}_8 = n_8/N$, see Appendix A for the proofs.

A gradient based optimization routine searches for the global minimum across the negative log-likelihood surface and it can get stuck in a local minimum. In our experience this especially happens when some of the cell counts in the contingency table are small. The analytical maximization might yield more accurate estimates in these cases, see Appendix D.

3.2 Difference based tests

Other possible test statistics start out by looking at the difference of the PPVs, and then these test statistics can be standardized by using Taylor series expansion. We also suggest some improvement to these tests.

3.2.1 Test statistics

Based on the difference δ_P given in Equation (6), which equals zero under the null hypothesis, we may suggest a variety of possible test statistics.

Wang et al. (2006) suggested the test statistics

$$g_1(\boldsymbol{N}) = \frac{N_5 + N_6}{N_1 + N_2 + N_5 + N_6} - \frac{N_5 + N_7}{N_1 + N_3 + N_5 + N_7} \tag{18}$$

and

$$g_2(\boldsymbol{N}) = \log \frac{(N_5+N_6)(N_1+N_3+N_5+N_7)}{(N_5+N_7)(N_1+N_2+N_5+N_6)}. \tag{19}$$

For a more detailed description of their work, see Appendix B.1. Moskowitz and Pepe (2006) also suggest a similar test statistic to $g_2(\boldsymbol{N})$, see Appendix B.2.

Since the null hypothesis can be written

$$H_0^P : \frac{p_1+p_3}{p_5+p_7} = \frac{p_1+p_2}{p_5+p_6}$$

another test statistic to be used may be

$$g_3(\boldsymbol{N}) = \frac{N_1+N_3}{N_5+N_7} - \frac{N_1+N_2}{N_5+N_6}.$$

Another possibility is to use the log ratio of the terms, instead of their difference,

$$g_4(\boldsymbol{N}) = \log \frac{(N_1+N_3)(N_5+N_6)}{(N_1+N_2)(N_5+N_7)}$$

or we may rewrite the null hypothesis in order to obtain

$$g_5(\boldsymbol{N}) = \frac{N_5+N_6}{N_1+N_2} - \frac{N_5+N_7}{N_1+N_3}.$$

3.2.2 Standardization by Taylor series expansion

For a general test statistic, $g(\boldsymbol{N})$, we may construct a standardized test statistic by subtracting the expectation of the test statistic, $\mathrm{E}(g(\boldsymbol{N}))$, and dividing by its standard deviation, $\sqrt{\mathrm{Var}(g(\boldsymbol{N}))}$. In the large sample case the square of the standardized test statistics may then be assumed to be approximately χ_1^2-distributed,

$$T(\boldsymbol{N}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left\{ \frac{g(\boldsymbol{N}) - \mathrm{E}(g(\boldsymbol{N}))}{\sqrt{\mathrm{Var}(g(\boldsymbol{N}))}} \right\}^2 \approx \chi_1^2. \tag{20}$$

The expectation and variance of the test statistic can be approximated with the aid of Taylor series expansion as suggested by Wang et al. (2006). Let $\mathrm{E}(\boldsymbol{N}) = \boldsymbol{\mu}$ be the point around which the expansion is centered. As before, $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{N})$. A second order Taylor expansion in matrix notation is given as

$$g(\boldsymbol{N}) \approx g(\boldsymbol{\mu}) + \boldsymbol{G}^T(\boldsymbol{\mu})(\boldsymbol{N}-\boldsymbol{\mu}) + \frac{1}{2}(\boldsymbol{N}-\boldsymbol{\mu})^T \boldsymbol{H}(\boldsymbol{\mu})(\boldsymbol{N}-\boldsymbol{\mu}) \tag{21}$$

where $\boldsymbol{G}$ is a vector containing the first order partial derivatives of $g(\boldsymbol{N})$ with respect to the components of $\boldsymbol{N}$ and $\boldsymbol{G}^T$ is the transpose of $\boldsymbol{G}$. Further $\boldsymbol{H}$ is a matrix with second order partial derivatives of $g(\boldsymbol{N})$ with respect to the components of $\boldsymbol{N}$, i.e. the Hessian matrix.

The expectation of $g(\boldsymbol{N})$ can then be approximated as

$$\mathrm{E}(g(\boldsymbol{N})) \approx g(\boldsymbol{\mu})$$

for the first order Taylor expansion and as

$$\mathrm{E}(g(\boldsymbol{N})) \approx g(\boldsymbol{\mu}) + \frac{1}{2}\mathrm{tr}(\boldsymbol{H}(\boldsymbol{\mu})\boldsymbol{\Sigma}) \tag{22}$$

for the second order Taylor expansion, since

$$\begin{aligned} & \mathrm{E}\left(\tfrac{1}{2}(\boldsymbol{N}-\boldsymbol{\mu})^T\boldsymbol{H}(\boldsymbol{\mu})(\boldsymbol{N}-\boldsymbol{\mu})\right) \\ = \; & \mathrm{E}\left(\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{H}(\boldsymbol{\mu})(\boldsymbol{N}-\boldsymbol{\mu})^T(\boldsymbol{N}-\boldsymbol{\mu})\right)\right) \\ = \; & \tfrac{1}{2}\mathrm{tr}\left(\boldsymbol{H}(\boldsymbol{\mu})\mathrm{E}((\boldsymbol{N}-\boldsymbol{\mu})^T(\boldsymbol{N}-\boldsymbol{\mu}))\right) \end{aligned}$$

where we have used the result $\boldsymbol{x}^TA\boldsymbol{x} = \mathrm{tr}(\boldsymbol{x}^TA\boldsymbol{x}) = \mathrm{tr}(A\boldsymbol{x}\boldsymbol{x}^T)$ where $\boldsymbol{x} = \boldsymbol{N} - \boldsymbol{\mu}$ and A is the Hessian matrix $\boldsymbol{H}$. $\mathrm{E}((\boldsymbol{N}-\boldsymbol{\mu})(\boldsymbol{N}-\boldsymbol{\mu})^T)$ is the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{N}$.

The variance of $g(\boldsymbol{N})$ can be approximated as

$$\mathrm{Var}(g(\boldsymbol{N})) \approx \boldsymbol{G}^T(\boldsymbol{\mu})\boldsymbol{\Sigma}\,\boldsymbol{G}(\boldsymbol{\mu})$$

for the first order Taylor expansion. Using a second order Taylor expansion for the variance requires finding the third and fourth order moments of $\boldsymbol{N}$.

Using the first order Taylor approximation of $\mathrm{E}(g(\boldsymbol{N}))$ and $\mathrm{Var}(g(\boldsymbol{N}))$ in the standardized test statistic of (20) yields

$$T(\boldsymbol{N}) = \frac{(g(\boldsymbol{N}) - g(\boldsymbol{\mu}))^2}{\boldsymbol{G}^T(\boldsymbol{\mu})\boldsymbol{\Sigma}\boldsymbol{G}(\boldsymbol{\mu})} \approx \chi^2_1. \tag{23}$$

Under the null hypothesis, $g(\boldsymbol{\mu}) = 0$. $\boldsymbol{G}(\boldsymbol{\mu})$ and $\boldsymbol{\Sigma}$ are functions of the unknown parameters $\boldsymbol{p}$ and needs to be estimated. We can either insert the general maximum likelihood estimates $\hat{p}_i = n_i/N$ or the maximum likelihood estimates $\tilde{p}_i$ under H_0^P, as found in Section 3.1.2. When we use the standardized test statistic (23) with $g_1(\boldsymbol{N})$ and estimate the variance using the maximum likelihood estimates under H_0 we refer to it as the *restricted difference test*. If we instead use the unrestricted maximum likelihood estimates to estimate the variance, we refer to it as the *unrestricted difference test*.

We have investigated two possible improvements of the standardized test statistics. In addition to using the restricted maximum likelihood estimates to estimate the variance of (23), we have looked at the effect of using a second order Taylor series approximation to $\mathrm{E}(g(\boldsymbol{N}))$ as an attempt to arrive at a more accurate approximation to a χ^2_1 distributed test statistic. The expectation and variance in the standardized test statistic given in (20) is found using a first order Taylor series expansion and the difference between using the first order and the second order Taylor series approximation to $\mathrm{E}(g(\boldsymbol{N}))$ is the term $1/2 \cdot \mathrm{tr}(\boldsymbol{H}(\boldsymbol{\mu})\boldsymbol{\Sigma})$. For the simulation experiment in Section 4 this turned out to be very small as compared to the denominator of (23).

3.3 TEST BY LEISENRING, ALONZO AND PEPE (LAP)

Leisenring et al. (2000) present a test for the null hypothesis given in (2). We will denote this the LAP test. They define three binary random variables; D_{ij} that denotes disease status, Z_{ij} that indicates which test was used and X_{ij} that describes the outcome of the diagnostic test for test j, $j = 1, 2$, for subject i, $i = 1, \ldots, N$.

$$D_{ij} = \begin{cases} 0, & \text{non-diseased} \\ 1, & \text{diseased} \end{cases}$$

$$Z_{ij} = \begin{cases} 0, & \text{test A} \\ 1, & \text{test B} \end{cases}$$

$$X_{ij} = \begin{cases} 0, & \text{negative} \\ 1, & \text{positive} \end{cases}$$

The PPV of test A can be written as PPV$_A = P(D_{ij} = 1 \mid Z_{ij} = 0, X_{ij} = 1)$ and the PPV of test B as PPV$_B = P(D_{ij} = 1 \mid Z_{ij} = 1, X_{ij} = 1)$. Based on generalized estimation equations Leisenring et al. (2000) fit the generalized linear model

$$\text{logit}(P(D_{ij} = 1 \mid Z_{ij}, X_{ij} = 1)) = \alpha_P + \beta_P Z_{ij}.$$

Testing the null hypothesis H_0: PPV$_A$ = PPV$_B$ is equivalent to testing the null hypothesis $H_0 : \ \beta_P = 0$. To derive the generalized score statistic, an independent working correlation structure is assumed for the score function and the corresponding variance function is $v_{ij} = \mu_{ij}(1 - \mu_{ij})$ where $\mu_{ij} = E(D_{ij})$. The score function is then $S_P = \sum_{i=1}^{N_p} \sum_{j=1}^{m_i} Z_{ij}(D_{ij} - \bar{D})$ which also can be written as $S_P = \sum_{i=1}^{N_p} \sum_{j=1}^{m_i} D_{ij}(Z_{ij} - \bar{Z})$. Here N_p is the number of subjects with at least one positive test outcome and m_i is the number of positive test results for subject i.

$$\bar{Z} = \frac{\sum_{i=1}^{N_p} m_i Z_i D_i}{\sum_{i=1}^{N_p} m_i}$$

is the proportion of positive B tests for the diseased subjects among all the positive tests and

$$\bar{D} = \frac{\sum_{i=1}^{N_P} m_i D_i}{\sum_{i=1}^{N_P} m_i}$$

is the proportion of positive tests for the diseased subjects among all the positive tests.

The resulting test statistic for testing the null hypothesis $H_0 : \beta_P = 0$ is obtained by taking the square of the score function divided by its variance:

$$T_{PPV} = \frac{\left\{\sum_{i=1}^{N_p} \sum_{j=1}^{m_i} D_{ij}(Z_{ij} - \bar{Z})\right\}^2}{\sum_{i=1}^{N_p} \left\{\sum_{j=1}^{m_i} (D_{ij} - \bar{D})(Z_{ij} - \bar{Z})\right\}^2}. \tag{24}$$

Under the null hypothesis, this test statistic is asymptotically χ_1^2-distributed. It is worth noting that only the subjects with at least one positive test outcome contribute to the test statistic (24).

The test statistic in (24) is general and can be used even if the disease status is not constant within a subject. Usually the disease status will be constant within the subject and the test statistic can be then simplified. By defining $T_i = \sum_{j=1}^{m_i} Z_{ij}$, the number of positive B tests subject i contributes to the analysis, the statistic can be written

$$T_{\text{PPV}} = \frac{\left\{\sum_{i=1}^{N_p} D_i(T_i - m_i\bar{Z})\right\}^2}{\sum_{i=1}^{N_p} (D_i - \bar{D})^2 (T_i - m_i\bar{Z})^2}.$$

We derived the test statistic by using our notation of the eight mutually exclusive events in Figure 1. The numerator can be separated into six terms, in three of which the disease status $D = 0$ and three where $D = 1$, by noting that $T_i = 0$ and $m_i = 1$ when only test A is positive, $T_i = 1$ and $m_i = 1$ when only test B is positive and $T_i = 1$ and $m_i = 2$ when both tests are positive. Then

$$T_{\text{PPV}} = \frac{((N_1 + N_2 + N_5 + N_6)(N_5 + N_7) - (N_1 + N_3 + N_5 + N_7)(N_5 + N_6))^2}{f(N_1, N_2, N_3, N_5, N_6, N_7)} \qquad (25)$$

where

$$\begin{aligned} &f(N_1, N_2, N_3, N_5, N_6, N_7) \\ &\qquad = N_1(N_2 - N_3 + N_6 - N_7)^2 \left(\frac{2N_5 + N_6 + N_7}{2N_1 + N_2 + N_3 + 2N_5 + N_6 + N_7}\right)^2 \\ &\qquad + N_2(N_1 + N_3 + N_5 + N_7)^2 \left(\frac{2N_5 + N_6 + N_7}{2N_1 + N_2 + N_3 + 2N_5 + N_6 + N_7}\right)^2 \\ &\qquad + N_3(N_1 + N_2 + N_5 + N_6)^2 \left(\frac{2N_5 + N_6 + N_7}{2N_1 + N_2 + N_3 + 2N_5 + N_6 + N_7}\right)^2 \\ &\qquad + N_5(N_2 - N_3 + N_6 - N_7)^2 \left(1 - \frac{2N_5 + N_6 + N_7}{2N_1 + N_2 + N_3 + 2N_5 + N_6 + N_7}\right)^2 \\ &\qquad + N_6(N_1 + N_3 + N_5 + N_7)^2 \left(1 - \frac{2N_5 + N_6 + N_7}{2N_1 + N_2 + N_3 + 2N_5 + N_6 + N_7}\right)^2 \\ &\qquad + N_7(N_1 + N_2 + N_5 + N_6)^2 \left(1 - \frac{2N_5 + N_6 + N_7}{2N_1 + N_2 + N_3 + 2N_5 + N_6 + N_7}\right)^2 . \end{aligned}$$

To compare the NPVs for test A and test B, Leisenring et al. (2000) fit the generalized linear model

$$\text{logit}(P(D_{ij} = 1 | Z_{ij}, X_{ij} = 0)) = \alpha_N + \beta_N Z_{ij}.$$

by using the generalized estimating equations method. The null hypothesis in this case is $H_0 : \beta_N = 0$. Under the assumption that disease status is constant within a subject, this leads to the test statistic

$$T_{\text{NPV}} = \frac{\left\{\sum_{i=1}^{N_n} D_i(T_i - k_i\bar{Z})\right\}^2}{\sum_{i=1}^{N_n}(D_i - \bar{D})^2(T_i - k_i\bar{Z})^2}$$

where N_n is the number of subjects with at least one negative test outcome and k_i is the number of negative test results for subject i. Only the subjects with at least one negative test outcome contribute to this test statistic.

Leisenring et al. (2000) also propose a Wald test based on the estimates of the regression coefficients, but their simulation studies show that the score test performs better.

4 Simulation study

In order to compare the test size under the null hypothesis for the tests presented in Section 3 and to assess the power of the tests under the alternative hypothesis, we perform a simulation experiment.

All the tests are asymptotic tests, but it is not clear how large the sample size has to be for the tests to preserve their test size. Therefore we will consider different sample sizes. Two different simulation strategies for generating datasets will be presented. The maximum likelihood estimates under the null hypotheses needed for the likelihood ratio test and the restricted difference test are found using TANGO as described in Section 3.1.2. All analyses are performed using the R language, R Development Core Team (2008).

4.1 Simulation experiment from Leisenring, Alonzo and Pepe

The first simulation experiment is based on the simulation experiment of Leisenring et al. (2000) and we use their algorithm to generate the data. Therefore we denote this simulation experiment the LAP simulation experiment.

4.1.1 Algorithm

We generate datasets by using the algorithm presented in Appendix B in Leisenring et al. (2000). Let I_D denote the disease status,

$$I_D = \begin{cases} 1, & \text{diseased} \\ 0, & \text{non-diseased} \end{cases}$$

and I_A and I_B the test results of test A and B,

$$I_A = \begin{cases} 1, & \text{test A positive} \\ 0, & \text{test A negative} \end{cases}$$

$$I_B = \begin{cases} 1, & \text{test B positive} \\ 0, & \text{test B negative} \end{cases}$$

In order to generate the datasets, the number of subjects tested, N, the positive and negative predictive values for both tests, the prevalence of the disease $P(D)$ and the variance σ^2 for the random effect for each subject must be set. The random effect introduces correlation between the test outcomes for each subject. Our interpretation of the simulation algorithm is as follows:

1. Set N, $P(D)$, PPV_A, NPV_A, PPV_B, NPV_B and σ.
2. Calculate the true positive rate TP and the false positive rate FP for test A and test B defined by the equations

$$\text{TP} = \frac{(1 - P(D) - \text{NPV}) \cdot \text{PPV}}{(1 - \text{PPV} - \text{NPV}) \cdot P(D)}$$

   and

$$\text{FP} = \frac{1 - P(D) - \text{NPV}}{(1 - P(D))(1 - \text{NPV} - \frac{\text{PPV}\cdot\text{NPV}}{1-\text{PPV}})}.$$

3. Given TP and FP, the parameters α_i and β_i, $i = 1, 2$, for each test are calculated from the following equations,

$$\alpha_i = \Phi^{-1}(\text{FP})\sqrt{1+\sigma^2}$$
$$\beta_i = \Phi^{-1}(\text{TP})\sqrt{1+\sigma^2},$$

   where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

Case no.	N	$P(D)$	σ	PPV$_A$	PPV$_B$	NPV$_A$	NPV$_B$
1	100	0.25	0.1	0.75	0.75	0.85	0.85
2	500	0.25	0.1	0.75	0.75	0.85	0.85
3	100	0.50	0.1	0.75	0.75	0.85	0.85
4	500	0.50	0.1	0.75	0.75	0.85	0.85
5	100	0.25	1.0	0.75	0.75	0.85	0.85
6	500	0.25	1.0	0.75	0.75	0.85	0.85
7	100	0.50	1.0	0.75	0.75	0.85	0.85
8	500	0.50	1.0	0.75	0.75	0.85	0.85

TABLE 3: Specifications of the cases under the null hypotheses in the LAP simulation experiment.

4. For each subject the disease status I_D is drawn independently with probability $P(D)$.
5. A random effect $r \sim N(0, \sigma^2)$ is generated for each subject.
6. Given the disease status and the random effect r, the probability of a positive test outcome for each subject is given by

$$P(I_A = 1|I_D, r) = \Phi(\alpha_1(1 - I_D) + \beta_1 I_D + r)$$

   for test A and by

$$P(I_B = 1|I_D, r) = \Phi(\alpha_2(1 - I_D) + \beta_2 I_D + r)$$

   for test B. The test outcomes are drawn with these probabilities for each subject.

7. Find n_1, ..., n_8 by counting the number of subjects that belongs to each of the eight events described in Section 2, e.g. n_1 is the number of subjects for which $I_D = 0$, $I_A = 1$ and I_B=1, the number of subjects that are not diseased and have positive tests A and B.

The algorithm is repeated M times, providing M datasets of $n_1, \ldots, n_8$.

4.1.2 CASES UNDER STUDY

In the simulation experiment, we suggest eight cases by varying the input parameters N, $P(D)$ and σ in the LAP simulation algorithm. The setup of the experiment is a 2^3 factorial experiment, i.e. we have three factors, N, $P(D)$ and σ, and each factor has two levels. The low level for N is 100 and the high level is 500, while the low level for $P(D)$ is 0.25 and the high level is 0.50. For σ the low level is 0.1 and the high level is 1.0. The response in this experiment is the estimated test size for each test. The cases that are under the null hypotheses H_0^P and H_0^N in equations (2) and (3) are given in Table 3. For cases not under the null hypotheses, the parameters N, $P(D)$ and σ are the same, but the remaining parameters are changed and will be described below. For each of these eight cases we simulate $M = 5000$ datasets.

We generate data under the null hypotheses (2) and (3), by setting PPV$_A$ = PPV$_B$ $= 0.75$ and NPV$_1$ = NPV$_2$ $= 0.85$. These datasets are used to estimate the test size under H_0 for both the PPV and NPV tests. To estimate the power of the tests, we need datasets under H_1, and for PPV we generate datasets where $\text{PPV}_A = 0.85$ and $\text{PPV}_B = 0.75$ and $\text{NPV}_1 = \text{NPV}_2 = 0.85$. To compare the power for the NPV tests, we generate datasets where $\text{NPV}_1 = 0.85$ and $\text{NPV}_2 = 0.80$ and $\text{PPV}_A = \text{PPV}_B = 0.75$.

To compare the positive predictive values of test A and B, we calculate the test statistics for the LAP test, the likelihood ratio test and the unrestricted and restricted difference tests. To compare the negative predictive values for test A and B we use the negative predictive value versions of these test statistics. We calculate p-values based on the χ_1^2 distribution. We also assess the performance of the four other difference based tests as proposed in Section 3.2.1.

4.1.3 RESULTS

A summary of the results of the simulation experiment will follow. For each case and selected value of the nominal significance level α, let W be a random variable counting the number of p-values smaller than or equal to α. Then W is binomially distributed with size M, the number of p-values generated, and probability α. An estimate of the true significance level of the test, $\hat{\alpha}$ is then

$$\hat{\alpha} = \frac{W}{M}. \tag{26}$$

Let

$$\begin{aligned} \widetilde{W} &= W + 2 \\ \widetilde{M} &= M + 4 \\ \tilde{\alpha} &= \frac{\widetilde{W}}{\widetilde{M}}. \end{aligned}$$

A $100 \cdot (1 - \gamma)\%$ confidence interval for $\hat{\alpha}$ with limits $\hat{\alpha}_L$ and $\hat{\alpha}_U$, according to Agresti and Coull (1998) is given as

$$\hat{\alpha}_L = \tilde{\alpha} - z_{\frac{\gamma}{2}} \sqrt{\frac{\tilde{\alpha} \cdot (1 - \tilde{\alpha})}{\widetilde{M}}} \tag{27}$$

and

$$\hat{\alpha}_U = \tilde{\alpha} + z_{\frac{\gamma}{2}} \sqrt{\frac{\tilde{\alpha} \cdot (1 - \tilde{\alpha})}{\widetilde{M}}} \tag{28}$$

where $z_{\gamma/2}$ is the $\gamma/2$-quantile in the standard normal distribution. When the samples are drawn under H_0, $\hat{\alpha}$ will be an estimate of the test size, i.e. the probability of making a type I error, $P(\text{reject } H_0|H_0)$. A p-value is valid, as defined by Lloyd and Moldovan (2008), if the actual probability of rejecting the null hypothesis never exceeds the nominal significance level. We choose the nominal significance level to be 0.05 and we say that the test preserves its test size if the lower confidence limit is less than or equal to 0.05, i.e. if $\hat{\alpha}_L \leq 0.05$. If $\hat{\alpha}_U < 0.05$, the test is said to be conservative, while if $\hat{\alpha}_L > 0.05$ it does not keep its test size and it is then optimistic. If the samples are drawn under the alternative H_1, $\hat{\alpha}$ is an estimate of the power of the test, i.e. $P(\text{reject } H_0 \mid H_1)$, the probability to correctly reject the null hypothesis when it is not true.

Table 4 shows the estimated test size with 95% confidence limits for the LAP test, the likelihood ratio test and the restricted and unrestricted difference tests in Case 1–8 where the data is generated under the null hypothesis that $\text{PPV}_A = \text{PPV}_B$.

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 1 LAP test	0.058	0.052	0.065
Case 1 Likelihood ratio test	0.065	0.059	0.072
Case 1 Restricted difference test	0.051	0.046	0.058
Case 1 Unrestricted difference test	0.067	0.060	0.074
Case 2 LAP test	0.056	0.050	0.063
Case 2 Likelihood ratio test	0.056	0.050	0.063
Case 2 Restricted difference test	0.055	0.049	0.062
Case 2 Unrestricted difference test	0.058	0.051	0.064
Case 3 LAP test	0.051	0.046	0.058
Case 3 Likelihood ratio test	0.050	0.044	0.056
Case 3 Restricted difference test	0.048	0.043	0.055
Case 3 Unrestricted difference test	0.051	0.045	0.058
Case 4 LAP test	0.057	0.051	0.064
Case 4 Likelihood ratio test	0.057	0.051	0.064
Case 4 Restricted difference test	0.057	0.051	0.064
Case 4 Unrestricted difference test	0.057	0.051	0.064
Case 5 LAP test	0.058	0.052	0.065
Case 5 Likelihood ratio test	0.070	0.063	0.077
Case 5 Restricted difference test	0.053	0.047	0.059
Case 5 Unrestricted difference test	0.065	0.058	0.072
Case 6 LAP test	0.054	0.048	0.060
Case 6 Likelihood ratio test	0.053	0.048	0.060
Case 6 Restricted difference test	0.052	0.046	0.058
Case 6 Unrestricted difference test	0.055	0.049	0.061
Case 7 LAP test	0.053	0.047	0.060
Case 7 Likelihood ratio test	0.055	0.049	0.061
Case 7 Restricted difference test	0.049	0.044	0.056
Case 7 Unrestricted difference test	0.054	0.048	0.060
Case 8 LAP test	0.055	0.049	0.062
Case 8 Likelihood ratio test	0.055	0.049	0.062
Case 8 Restricted difference test	0.054	0.048	0.061
Case 8 Unrestricted difference test	0.055	0.049	0.062

TABLE 4: Estimated test size with 95% confidence limits when testing $\text{PPV}_A = \text{PPV}_B$ for data generated under the null hypothesis using the LAP-simulation algorithm.

In Case 3, 6, 7 and 8 all four test preserve the test size. In Case 2 the unrestricted difference test is too optimistic, while the other tests preserve the test size.

In Case 1 and 5 the restricted difference test is the only test preserving the test size. The other tests are too optimistic. These cases have small cell counts, and it might be that the restricted difference test is more robust towards small cell counts than the other tests. Table 5 shows the mean observed cell counts in Case 1–8 for the data generated under the null hypothesis that PPV$_A$ = PPV$_B$. We see that in Case 1 and 5, $\bar{n}_1$ is 0.2 and 1.1 respectively, and thereby $n_1 = 0$ in many of the datasets, and also some of the other cell counts are small.

In Case 4 none of the four tests preserve the test size, i.e. all the tests are slightly optimistic. As all the cell counts are high in this case it is not surprising that the estimated test size is the same for all the tests, however we see no apparent reason why the test size is not preserved, and thus this may perhaps be a purely random event.

For the likelihood ratio test we analysed the 2^3 factorial experiment using $\hat{\alpha}$ as the response and found that the interaction between the factors N and $P(D)$ is the most significant effect on the the test size with a p-value of 0.012. When N is at its high level, $N = 500$, the test size is less affected by $P(D)$ which makes sense, since the high value of N ensures that all the cells will have large expected values unless the corresponding cell probabilities are very small. There is also a significant interaction between N and σ, when $N = 100$, the estimated test size is higher for $\sigma = 1.0$ than for $\sigma = 0.01$ and when $N = 500$, the estimated test size is lower for $\sigma = 1.0$ than for $\sigma = 0.01$.

Table 12 (see Appendix C) shows the estimated test size with 95% confidence limits for the NPV versions of the LAP test, the likelihood ratio test and the restricted and unrestricted difference tests in Case 1–8 where the data is generated under the null hypothesis that NPV$_1$ = NPV$_2$. In Case 1, 2, 4 and 8, all the cases preserve the test size. In Case 5 all the cases except the likelihood ratio test preserve the test size too. In Case 3 and 7 however, only the restricted difference test preserves the test size, none of the other tests do. It may be because it is more robust to the small cell counts in the eight cell, $\bar{n}_8$, which is 0.8 and 2.2 respectively in these two cases.

Table 13 and 14 (see Appendix C) show the estimated power with 95% confidence intervals for the PPV and NPV versions respectively of the LAP test, the likelihood ratio test and the restricted and unrestricted difference tests in Case 1–8 for the data generated under the two alternative hypotheses. The power of the restricted difference test is generally lower than the power of the other tests, which is not surprising since it preserves its test size when the other tests do not. The power increases with the number of subjects N as we would expect. For the PPV tests, it also increases when the prevalence $P(D)$ increases. When the prevalence increases it is more likely that a random subject has the disease, therefore more subjects will have the disease and there will be more positive tests. $P(D) = 0.50$ in Case 4 and 8 where the tests have higher power than in Case 2 and 6 where $P(D) = 0.25$. We also note that in general the test power is higher when $\sigma = 1$ compared to when $\sigma = 0.1$. For the NPV tests, the power increases when N increases and when $P(D)$ decreases. When $P(D) = 0.25$, $P(D^*) = 1 - P(D) = 0.75$, and the higher this probability is the more likely it is that a random subject does not have the disease. The number of subjects that do not have the disease are then expected to be higher than when $P(D) = 0.50$ and $P(D^*) = 0.50$. As for the PPV tests, the power increases when σ increases.

Table 6 shows the estimated test size with 95% confidence intervals in Case 1–8 for the four other difference based test statistics from Section 3.2.1. When calculating the observed value of the standardized test statistics the unrestricted maximum likelihood estimates in the variance are inserted

Case no.	$\bar{n}_1$	$\bar{n}_2$	$\bar{n}_3$	$\bar{n}_4$	$\bar{n}_5$	$\bar{n}_6$	$\bar{n}_7$	$\bar{n}_8$
1	0.2	3.9	3.9	67.0	6.3	6.2	6.2	6.3
2	1.2	19.7	19.6	334.6	31.4	31.1	31.0	31.5
3	4.3	10.3	10.2	25.1	38.4	5.4	5.5	0.8
4	21.4	51.5	51.3	125.7	191.8	27.1	27.1	4.0
5	1.1	3.1	3.1	67.8	8.3	4.2	4.2	8.4
6	5.3	15.6	15.5	338.6	41.7	20.9	20.8	41.6
7	7.5	7.0	7.0	28.3	39.8	4.1	4.0	2.2
8	37.6	35.3	35.2	141.9	198.7	20.1	20.0	11.2

TABLE 5: Mean cell counts for the cases in the LAP simulation study under H_0.

since in the LAP simulation experiment, the test size for the restricted difference test was lower than the test size for the unrestricted difference test. If we compare these results with the results for the unrestricted difference test, we see that the estimated test size depend highly on the choice of test statistic. The test based on $g_2(\boldsymbol{N})$ preserves its test size in all the cases except Case 4. It is however conservative in Case 1 and 5. The test based on $g_3(\boldsymbol{N})$ preserves its test size in all the cases, but it is conservative in all except Case 4 and 8. In Case 1 and 5 it is very conservative with an estimated test size of just 0.008 and 0.007 respectively. For the fourth difference based test statistic, $g_4(\boldsymbol{N})$, the test size is preserved in all the cases except Case 4. It is conservative in Case 1, 3 and 5. The test based on $g_5(\boldsymbol{N})$ is conservative in all the cases, and more conservative than the other tests. In Case 1 and 5 the estimated test size is 0 and 0.001 which shows that this test statistic almost never rejects the null hypothesis. The tests based on $g_2(\boldsymbol{N})$ and $g_4(\boldsymbol{N})$ can be used as their estimated test size is reasonable, although conservative in some of the cases. We do not recommend using the tests based on $g_3(\boldsymbol{N})$ and $g_5(\boldsymbol{N})$ as these are even more conservative than the other tests.

4.2 Multinomial simulation experiment

In the LAP-simulation algorithm, $n_1, \dots, n_8$ are not drawn from a particular probability distribution, but obtained from the disease status and test results which are drawn with the specified probabilities in Section 4.1.1. However, in our model for the likelihood ratio test we assume that $N_1, ..., N_8$ are multinomially distributed. This can be used in the sampling strategy and we simulate data by sampling $n_1, ..., n_8$ from the multinomial distribution given the total number of subjects N and the parameters $p_1, ..., p_8$. This is less challenging to implement than the LAP-simulation algorithm and when using the likelihood ratio test it is natural to sample data from the distribution assumed when deriving the test statistic.

4.2.1 Algorithm

Given $\boldsymbol{p} = (p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8)$ and the total number of subjects N, we can generate datasets by drawing $n_1, n_2, ..., n_8$ from a multinomial distribution with parameters $\boldsymbol{p}$ and N. We first need to set $\boldsymbol{p}$ and if we want to sample under the null hypotheses, we need to ensure that $\boldsymbol{p}$ satisfy the constraints δ_P in Equation (6) and/or δ_N in Equation (8). In addition $p_1, ..., p_8$ must sum to 1.

Our simulation algorithm is as follows:

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 1 $g_2(\boldsymbol{N})$	0.042	0.037	0.048
Case 1 $g_3(\boldsymbol{N})$	0.008	0.006	0.011
Case 1 $g_4(\boldsymbol{N})$	0.021	0.017	0.025
Case 1 $g_5(\boldsymbol{N})$	0.000	0.000	0.001
Case 2 $g_2(\boldsymbol{N})$	0.056	0.050	0.063
Case 2 $g_3(\boldsymbol{N})$	0.043	0.038	0.049
Case 2 $g_4(\boldsymbol{N})$	0.051	0.045	0.058
Case 2 $g_5(\boldsymbol{N})$	0.008	0.006	0.011
Case 3 $g_2(\boldsymbol{N})$	0.045	0.040	0.051
Case 3 $g_3(\boldsymbol{N})$	0.037	0.032	0.042
Case 3 $g_4(\boldsymbol{N})$	0.043	0.038	0.049
Case 3 $g_5(\boldsymbol{N})$	0.001	0.000	0.002
Case 4 $g_2(\boldsymbol{N})$	0.057	0.051	0.063
Case 4 $g_3(\boldsymbol{N})$	0.056	0.050	0.063
Case 4 $g_4(\boldsymbol{N})$	0.057	0.051	0.064
Case 4 $g_5(\boldsymbol{N})$	0.042	0.036	0.048
Case 5 $g_2(\boldsymbol{N})$	0.039	0.034	0.044
Case 5 $g_3(\boldsymbol{N})$	0.007	0.005	0.010
Case 5 $g_4(\boldsymbol{N})$	0.027	0.023	0.032
Case 5 $g_5(\boldsymbol{N})$	0.000	0.000	0.001
Case 6 $g_2(\boldsymbol{N})$	0.049	0.043	0.055
Case 6 $g_3(\boldsymbol{N})$	0.040	0.035	0.046
Case 6 $g_4(\boldsymbol{N})$	0.047	0.042	0.054
Case 6 $g_5(\boldsymbol{N})$	0.007	0.005	0.010
Case 7 $g_2(\boldsymbol{N})$	0.047	0.042	0.053
Case 7 $g_3(\boldsymbol{N})$	0.035	0.031	0.041
Case 7 $g_4(\boldsymbol{N})$	0.047	0.042	0.054
Case 7 $g_5(\boldsymbol{N})$	0.001	0.000	0.003
Case 8 $g_2(\boldsymbol{N})$	0.054	0.048	0.061
Case 8 $g_3(\boldsymbol{N})$	0.052	0.046	0.058
Case 8 $g_4(\boldsymbol{N})$	0.054	0.048	0.061
Case 8 $g_5(\boldsymbol{N})$	0.042	0.036	0.048

TABLE 6: Estimated test size with 95% confidence limits when testing $\text{PPV}_A = \text{PPV}_B$ using the difference based tests.

Case	N	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
Case 3MN	100	0.05	0.10	0.10	0.25	0.39	0.05	0.05	0.01
Case 5MN	100	0.01	0.03	0.03	0.68	0.08	0.04	0.04	0.09

TABLE 7: Specification of the parameters in the multinomial simulation experiment.

1. Set $\boldsymbol{p} = (p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8)$ and N.
2. Draw $n_1, n_2, ..., n_8 \sim$ multinom$(\boldsymbol{p}, N)$. Repeat M times.

4.2.2 CASES UNDER STUDY

We performed a small simulation study by drawing data from a multinomial distribution. Under the null hypothesis (2) we defined two cases called Case 3MN and Case 5MN. The parameters for these cases are given in Table 7.

The parameters $p_1, ..., p_8$ for each of the cases sum to one and the δ_P-constraint (6) and δ_N-constraint (8) are both satisfied. The parameters were set in order to represent Case 3 and 5 from the LAP-simulation experiment. In both of these cases $N = 100$, while $P(D)$ is 0.5 in Case 3MN and 0.25 in Case 5MN as in Case 3 and 5 in the LAP-simulation experiment. For both Case 3MN and 5MN the PPVs are equal and approximately 0.75, the NPVs are equal and approximately 0.85. However, since the datasets in the LAP simulation experiment were not drawn from a multinomial distribution, the mean and the variance of $\boldsymbol{n}$ will not be exactly the same in Case 3MN and 5MN as in Case 3 and 5.

The parameters $p_1, ..., p_8$ were found by setting the value of $P(D)$, the values of PPV$_1$ $=$ PPV$_2$ and NPV$_1$ $=$ NPV$_2$ and by considering the mean observed values for Case 3 and 5 in Table 5. These two cases were chosen because we would like to test the multinomial sampling strategy for one case where the likelihood ratio test did not preserve its test size (Case 5) as well as one case where the test size was preserved (Case 3) in the LAP simulation experiment when testing if the positive predictive values are equal.

For each of the cases we draw $M = 5000$ samples from the multinomial distribution with parameters as given in Table 7.

4.2.3 RESULTS

The estimated test size and 95% confidence limits for the LAP test, the likelihood ratio test and the restricted and unrestricted difference tests for the two cases in the simulation study using the multinomial simulation algorithm are given in Table 8.

In Case 3MN all the tests preserve the test size. We note that the estimated test size is lower for the restricted difference test than for the other tests. In Case 5MN only the restricted difference test and the LAP test preserve their test size.

If we compare the results to Case 3 in the LAP simulation experiment we see that $\hat{\alpha}$ is higher in Case 3MN than in Case 3 for all the tests. In Case 5MN, $\hat{\alpha}$ is higher for the likelihood ratio test and lower for the other tests compared to Case 5 in the LAP simulation experiment. The datasets in the two simulation experiments are not identical, but since they are generated with approximately the same

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 3MN LAP	0.054	0.048	0.060
Case 3MN Likelihood ratio test	0.054	0.048	0.060
Case 3MN Restricted difference test	0.052	0.046	0.058
Case 3MN Unrestricted difference test	0.054	0.048	0.061
Case 5MN LAP	0.056	0.050	0.063
Case 5MN Likelihood ratio test	0.072	0.065	0.079
Case 5MN Restricted difference test	0.050	0.044	0.057
Case 5MN Unrestricted difference test	0.064	0.058	0.071

TABLE 8: Estimated test size with 95% confidence limits for testing $\text{PPV}_1 = \text{PPV}_2$ under the null hypothesis using the multinomial simulation algorithm.

values for PPV_1, PPV_2, NPV_1, NPV_2 and $P(D)$ we find it surprising that the estimated test size for the likelihood ratio test is higher in the multinomial simulation experiment than in the LAP simulation experiment. We would expect the likelihood ratio test to perform better, i.e. have a lower test size, on datasets that are drawn from the model on which the test statistic is based, namely the multinomial model.

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 3MN LAP	0.059	0.053	0.066
Case 3MN Likelihood ratio test	0.061	0.054	0.068
Case 3MN Restricted difference test	0.052	0.046	0.058
Case 3MN Unrestricted difference test	0.060	0.054	0.067
Case 5MN LAP	0.049	0.044	0.056
Case 5MN Likelihood ratio test	0.062	0.056	0.069
Case 5MN Restricted difference test	0.051	0.045	0.057
Case 5MN Unrestricted difference test	0.049	0.044	0.056

TABLE 9: Estimated test size with 95% confidence limits for testing $\text{NPV}_1 = \text{NPV}_2$ under the null hypothesis using the multinomial simulation algorithm.

The estimated test size with 95% confidence limits for testing if the NPVs are equal in the same cases are shown in Table 9. In Case 3MN only the restricted difference test preserves its test size, while in Case 5MN the LAP test and the unrestricted difference test also preserve their test size. The likelihood ratio test does not preserve its test size in any of these cases.

5 DATA FROM LITERATURE

We will use the dataset from Weiner, Ryan, McCabe, Kennedy, Schloss, Tristani and Fisher (1979) which is the same dataset as used in Leisenring et al. (2000) and Wang et al. (2006). There were 871 subjects of which 608 subjects had coronary artery disease (CAD) and 263 subjects did not have CAD. For all the subjects the results of clinical history (test A) and exercise stress testing (EST) (test

B) were registered. The dataset is shown in Table 10.

		Coronary artery disease -		Coronary artery disease +	
		Result of EST		Result of EST	
		+	-	+	-
Result of clinical history	+	22	44	473	81
	-	46	151	29	25

TABLE 10: Data from the coronary artery disease study.

Table 11 shows the resulting p-values for comparing the positive and negative predictive values using the LAP-test, the likelihood ratio test and the restricted and unrestricted difference test.

Test	PPV	NPV
LAP	0.3706	<0.0001
Likelihood ratio test	0.3710	<0.0001
Restricted difference test	0.3696	<0.0001
Unrestricted difference test	0.3705	<0.0001

TABLE 11: Comparison of p-values for the tests using data from the coronary artery disease study.

We see that all the tests yield the same results. We do not reject the null hypothesis that the PPVs are equal, but we reject the null hypothesis that the NPVs are equal. The estimated NPVs are 0.78 for the clinical history and 0.65 for EST. Therefore the clinical history is more likely to reflect the true disease status for subjects without CAD than without EST. Since all the cell counts in Table 10 are large, it is to be expected that the p-values are equal for all the tests, as seen in our simulation experiments.

6 ALTERNATIVE MODEL

When deriving the test statistic for comparing the positive predictive values for two tests, Leisenring et al. (2000) only consider the subjects that have at least one positive test result. The subjects that do not have any positive tests do not contribute to the test statistic, i.e. there is no information in how many subjects have two negative test results. Our multinomial setting with eight probabilities is useful because the null hypothesis for both the PPV and NPV can easily be expressed using the same model. However, for testing the equivalence of the PPVs, it is interesting to consider only using the subjects with at least one positive test result also for our likelihood ratio test as this will reduce the number of parameters and thereby reducing the dimension of the optimization problem. Similarly, for testing the equivalence of the NPVs for test A and test B, we only need to look at the subjects with at least one negative test.

This situation is illustrated in Figure 2. We still have the three main events A, B and D, but we only consider the data contained in A and/or B. The sample space is divided into six mutually exclusive events, to each of which a random variable N_i^*, $i = 1, ..., 6$, corresponds. We define N_i^* to be the number of subjects for which event i occurs and n_i^* to be the observed value of N_i^*. There are N^* subjects in total, i.e. $\sum_{i=1}^{6} N_i^* = N^*$. Let q_i be the probability that event i, $i = 1, ..., 6$, occurs. q_1 is then the probability that a subject has a positive test result for both test A and B and has the disease. $N_1^*, N_2^*, ..., N_6^*$ are multinomially distributed with parameters N^* and $\boldsymbol{q} = (q_1, q_2, q_3, q_4, q_5, q_6)$ where $\sum_{i=1}^{6} q_i = 1$.

The null hypothesis that the positive predictive value for test A is equal to the positive predictive value for test B can be written

$$H_0^{P,6}: \ \frac{q_4 + q_5}{q_1 + q_2 + q_4 + q_5} - \frac{q_4 + q_6}{q_1 + q_3 + q_4 + q_6} = 0 \tag{29}$$

The likelihood ratio test statistic in this case is then

$$-2 \cdot \log\lambda(\boldsymbol{n}^*) = -2 \left(\sum_{i=1}^{6} n_i^* \cdot (\log(\tilde{q}_i) - \log(\hat{q}_i) \right) \tag{30}$$

where $\boldsymbol{n}^* = (n_1^*, n_2^*, n_3^*, n_4^*, n_5^*, n_6^*)$, $\tilde{q}_i$ is the maximum likelihood estimate for q_i under the null hypothesis (29) and $\hat{q}_i = n_i^*/N^*$ is the general maximum likelihood estimate.

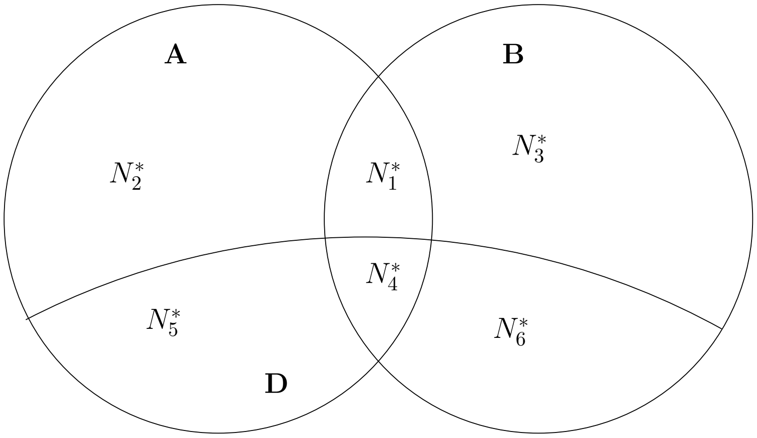

FIGURE 2: Venn diagram for the events D, A and B showing which events the random variables $N_1^*, ..., N_6^*$ correspond to.

If there is no information in the number of subjects not having at least one positive rest result, then n_4 and n_8 should not affect the value of the likelihood ratio test statistic. From the Lagrangian system of equations in Section 3.1.2, we can show that the estimates of p_4 and p_8 under H_0 are $\tilde{p}_4 = \frac{n_4}{N}$ and $\tilde{p}_8 = \frac{n_8}{N}$, see Appendix A.

Maximizing the multinomial likelihood with six parameters yields the same test statistic and thereby the same p-value as when maximizing the multinomial likelihood with eight parameters as both the restricted and unrestricted maximum likelihood estimates of q_1, q_2, q_3, q_4, q_5, q_6 are obtained from the restricted and unrestricted estimates of p_1, p_2, p_3, p_5, p_6, p_7, p_8 respectively by scaling the estimates so they sum to one (see Appendix A).

7 CONCLUSIONS

In this report we have studied large sample tests for comparing the positive and negative predictive value of two diagnostic tests in a paired design.

Based on the simulation experiments in Section 4, we have found that our restricted difference test outperforms the existing methods (Leisenring et al. (2000) and Wang et al. (2006)) as well as our likelihood ratio test with respect to test size.

A very important prerequisite of our methods is the estimation of the maximum likelihood estimates for the parameters in the multinomial distribution under the null hypothesis, and this has shown to be a challenging task as is also mentioned in Leisenring et al. (2000). We have found these estimates in two different ways, by using numerical optimization and solving a system of equations.

We have seen that when the sample size decreases, the LAP test, likelihood ratio test and unrestricted difference test do not preserve their test size. In our future work we will abandon the large sample assumption and work with small sample versions of our test statistics.

REFERENCES

Agresti, A. and Coull, B. (1998). Approximate is better than ”exact” for interval estimation of binomial proportions, *The American Statistician* 52(2): 119–126.

Alan Agresti (2002). *Categorical data analysis*, second edn, John Wiley & Sons, Inc., Hoboken, NJ, chapter 10.1.1.

Andreani, R., Birgin E. G., Martinez, J. M. and Schuverdt, M. L. (2007). On Augmented Lagrangian methods with general lower-level constraints, *SIAM Journal on Optimization* 18: 1296–1309.

Andreani, R., Birgin, E. G., Martinez, J. M. and Schuverdt, M. L. (2008). Augmented Lagrangian methods under the constant positive linear dependence constraint qualification, *Mathematical Programming* 111: 5–32.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, second edn, Duxbury, chapter 8.

Edwards, C. H. and Penney, D. E. (1998). *Calculus with analytic geometry*, fifth edn, Prentice-Hall International, Inc., Upper Saddle River, New Jersey, chapter 13.9.

Guggenmoos-Holzmann, I. and van Houwelingen, H. C. (2000). The (in)validity of sensitivity and specificity, *Statistics in Medicine* 19: 1783–1792.

Johnson, N. L., Kotz, S. and Balakrishan, N. (1997). *Discrete multivariate distributions*, Wiley series in probability and statistics, chapter 35.

Leisenring, W., Alonzo, T. and Pepe, M. S. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs, *Biometrics* 56: 345–351.

Lloyd, C. J. and Moldovan, M. V. (2008). A more powerful exact test of noninferiority from binary matched-pairs data, *Statistics in Medicine* 27(18): 3540–3549.

Moskowitz, C. S. and Pepe, M. S. (2006). Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs, *Clinical Trials* 3: 272–279.

R Development Core Team (2008). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.
`http://www.R-project.org`

Wang, W., Davis, C. S. and Soong, S.-J. (2006). Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares, *Statistics in Medicine* 25: 2215–2229.

Weiner, D. A., Ryan, T., McCabe, C. H., Kennedy, J. W., Schloss, M., Tristani, F.and Chaitman, B. R. and Fisher, L. (1979). 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* 301: 230–235.

A Proofs of properties of the maximum likelihood estimators under the null hypothesis

We show some properties of the maximum likelihood estimators under the positive predictive value constraint (6). Similar properties can be shown for maximum likelihood estimators satisfying the negative predictive value constraint (8).

We start by showing that if $n_1 = 0$, then the maximum likelihood estimate of p_1 under the null hypothesis, $\tilde{p}_1$, is 0. In the following, let $p_1, \ldots, p_8$ denote estimates, not true multinomial probabilities.

First we rewrite the constraint (11) under the null hypothesis as

$$\frac{p_1 + p_2}{p_5 + p_6} = \frac{p_1 + p_3}{p_5 + p_7} \tag{31}$$

Assume that $n_1 = 0$, $\sum_{i=1}^{8} p_i = 1$, the H_0 constraint (31) is satisfied and that $p_1 > 0$. We will prove that when $n_1 = 0$, the maximum likelihood estimate of p_1 is zero, i.e. $\tilde{p}_1 = 0$.

Let $p'_1 = 0$, $p'_2 = k(p_1 + p_2)$, $p'_3 = k(p_1 + p_3)$, $p'_4 = p_4$, $p'_5 = p_5$, $p'_6 = p_6$, $p'_7 = p_7$ and $p'_8 = p_8$ where $k = \frac{p_1+p_2+p_3}{2p_1+p_2+p_3}$.

Then $\sum_{i=1}^{8} p'_i = 1$ and $\boldsymbol{p}'$ also satisfy H_0, since

$$\frac{0 + p'_2}{p_5 + p_6} = \frac{0 + p'_3}{p_5 + p_7}.$$

We will show that $p'_2 > p_2$ and $p'_3 > p_3$, implying $\log L(p'_1, ..., p'_8) > \log L(p_1, ..., p_8)$. We start by writing down the expression for p'_2 and check if it is greater than p_2.

$$\begin{aligned} k(p_1 + p_2) &> p_2 \\ \frac{p_1 + p_2 + p_3}{2p_1 + p_2 + p_3}(p_1 + p_2) &> p_2 \\ (p_1 + p_2 + p_3) \cdot (p_1 + p_2) &> (p_1 + p_1 + p_2 + p_3)p_2 \\ (p_1 + p_2 + p_3)p_1 &> p_1 \cdot p_2 \\ p_1 + p_2 + p_3 &> p_2 \end{aligned}$$

The inequality is satisfied and therefore $p'_2 > p_2$. The same argument can be used to show that $p'_3 > p_3$.

The non-constant part of the log likelihood function is $\sum_{i=1}^{8} n_i \cdot \text{log}p_i$, and when $n_1 = 0$, the first term in the sum is 0, regardless of the value of p_1. When $p_2' > p_2$ and $p_3' > p_3$ we see that

$$\text{log}L(0, p_2', p_3', p_4, p_5, p_6, p_7, p_8) > \text{log}L(p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8).$$

Therefore $\tilde{p}_1 = 0$. The same argument is valid for $\tilde{p}_5$, i.e. $\tilde{p}_5 = 0$ when $n_5 = 0$. When n_4 and/or n_8 is 0, then $\tilde{p}_4$ and/or $\tilde{p}_8$ are also 0, see below.

However, the argument does not hold for $\tilde{p}_2$, $\tilde{p}_3$, $\tilde{p}_6$ and $\tilde{p}_7$ when n_2, n_3, n_6 or n_7 is 0. Even though e.g. $\tilde{p}_2$ may sometimes be 0 when $n_2 = 0$, this is not always true. If e.g. $\tilde{p}_1 = 0$ and $\tilde{p}_3 > 0$, then $\tilde{p}_2$ cannot be equal to 0 even if $n_2 = 0$ because then the null hypothesis constraint (31) will not be satisfied. One example of this situation is the table $\boldsymbol{n} = (0, 0, 6, 0, 2, 6, 0, 0)$. The analytic solution of the Lagrangian system of equations is $\tilde{\boldsymbol{p}} = (0, 2/7, 1/7, 0, 2/7, 2/7, 0, 0)$ and we see that $\tilde{p}_2 \neq 0$ even though $n_2 = 0$.

We proceed to show that $\tilde{p}_4 = n_4/N$ and $\tilde{p}_8 = n_8/N$. If we add the first eight Lagrangian equations in (18), we get

$$N = \gamma h(\boldsymbol{p}) + 2\kappa k(\boldsymbol{p}) = \gamma,$$

where $h(\boldsymbol{p}) = 1$ and $k(\boldsymbol{p}) = 0$ are the two constraints. Thus $\gamma = N$, and $\tilde{p}_4 = n_4/N$ and $\tilde{p}_8 = n_8/N$ follow from (18).

So the maximum likelihood estimate $\tilde{\boldsymbol{p}}$ under the null hypothesis is among the $\boldsymbol{p} = (p_1, \ldots, p_8)$ for which $p_4 = n_4/N$ and $p_8 = n_8/N$. For such a $\boldsymbol{p}$, let $s(\boldsymbol{p}) = r \cdot (p_1, p_2, p_3, p_5, p_6, p_7)$, where $r = N/(N - n_4 - n_8)$ so that the sum of the components of $s(\boldsymbol{p})$ is 1. Let log$L$ and logL' denote the log-likelihood of the original multinomial model with eight parameters and the alternative multinomial model with six parameters in Section 6. Then log$L(\boldsymbol{p})$ − log$L'(s(\boldsymbol{p}))$ is constant, showing that log$L(\boldsymbol{p})$ is maximal if and only if log$L'(s(\boldsymbol{p}))$ is. Furthermore, $\boldsymbol{p}$ satisfies the null hypothesis for the multinomial model with eight parameters if and only if $s(\boldsymbol{p})$ does for the multinomial model with six parameters, showing that the maximum likelihood estimates under the null hypothesis for the multinomial model with eight and six parameters are obtained from the other model by up- and downscaling, respectively.

There are also other relationships between the restricted parameter estimates that can easily be shown and used in the estimation of the parameters:

$$p_1 + p_2 + p_3 = \frac{n_1 + n_2 + n_3}{N}$$

$$p_5 + p_6 + p_7 = \frac{n_5 + n_6 + n_7}{N}$$

$$\frac{n_1}{p_1} + N = \frac{n_2}{p_2} + \frac{n_3}{p_3}$$

$$\frac{n_5}{p_5} + N = \frac{n_6}{p_6} + \frac{n_7}{p_7}$$

B Existing difference based methods

The already published difference based tests for comparing predictive values will here be described briefly.

B.1 Test by Wang

Recently Wang et al. (2006) presented two tests, one based on the difference of the PPVs and one based on the log ratio of the PPVs for testing the null hypothesis in (2). The data are assumed to be multinomially distributed.

They fit the model $\text{PPV}_A - \text{PPV}_B = \beta_1^P$ using the weighted least squares approach.[1] Testing if the positive predictive values for test A and B are equal is equivalent to testing $H_0 : \beta_1^P = 0$.

The test statistic is

$$W_1^P = \left(\sqrt{\frac{N}{\hat{\Sigma}_1^P}} (\widehat{\text{PPV}}_A - \widehat{\text{PPV}}_B) \right)^2 , \tag{32}$$

which is asymptotically χ_1^2-distributed. $\hat{\Sigma}_1^P$ is the estimated variance of $\hat{\beta}_1^P = \widehat{\text{PPV}}_A - \widehat{\text{PPV}}_B$. To compare the negative predictive values the same approach is followed by looking at the difference of the two negative predictive values. They fit the model $\text{NPV}_A - \text{NPV}_B = \beta_1^N$ and test the null hypothesis $H_0 : \beta_1^N = 0$ using the following test statistic

$$W_1^N = \left(\sqrt{\frac{N}{\hat{\Sigma}_1^N}} (\widehat{\text{NPV}}_A - \widehat{\text{NPV}}_B) \right)^2 \tag{33}$$

where $\hat{\Sigma}_1^N$ is the estimated variance of $\hat{\beta}_1^N = \widehat{\text{NPV}}_A - \widehat{\text{NPV}_B}$. W_1^N is asymptotically χ_1^2-distributed.

In the second test they consider the log ratio of the PPVs as their test statistic and fit the model $\log \frac{\text{PPV}_A}{\text{PPV}_B} = \beta_2^P$. Testing if the positive predictive values are equal is in this case equivalent to testing the null hypothesis $H_0 : \beta_2^P = 0$. The test statistic is

$$W_2^P = \left(\frac{\sqrt{N}}{\hat{\Sigma}_2^P} \log \frac{\widehat{\text{PPV}_A}}{\widehat{\text{PPV}_B}} \right)^2 \tag{34}$$

which is asymptotically χ_1^2 distributed. $\hat{\Sigma}_2^P$ is the estimated variance of $\hat{\beta}_2^P = \log \frac{\widehat{\text{PPV}_A}}{\widehat{\text{PPV}_B}}$. The same approach is followed to derive a second test for the negative predictive values by looking at the log ratio of the negative predictive values for test A and test B, the model fitted is $\log \left(\frac{\text{NPV}_A}{\text{NPV}_B} \right) = \beta_2^N$. To test if the negative predictive values are equal, the null hypothesis is $H_0 : \beta_2^N = 0$ and they use the following test statistic

$$W_2^N = \left(\frac{\sqrt{N}}{\hat{\Sigma}_2^N} \log \frac{\widehat{\text{NPV}_A}}{\widehat{\text{NPV}_B}} \right)^2 \tag{35}$$

where $\hat{\Sigma}_2^N$ is the estimated variance of $\hat{\beta}_2^N = \log \frac{\widehat{\text{NPV}_A}}{\widehat{\text{NPV}_B}}$. The test statistic in (35) is χ_1^2-distributed. They recommend using the tests based on the difference of the predictive values because it performs better than the tests based on the log ratio of the predictive values in terms of test size and power.

B.2 Test by Moskowitz and Pepe

Moskowitz and Pepe (2006) look at the relative predictive values, $\text{rPPV} = \frac{\text{PPV}_A}{\text{PPV}_B}$ and $\text{rNPV} = \frac{\text{NPV}_A}{\text{NPV}_B}$. By using the multivariate central limit theorem and the Delta method (which uses Taylor series ex-

[1] The notation in Appendix B.1 differs from the notation used in Wang et al. (2006).

pansions to derive the asymptotic variance), the following $100 \cdot (1-\alpha)\%$ confidence intervals can be estimated for log rPPV and log rNPV,

$$\text{log rPPV} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_P^2}{N}}$$

$$\text{log rNPV} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_N^2}{N}}$$

where $\widehat{\sigma}_P^2$ and $\widehat{\sigma}_N^2$ are the estimated variances of $\frac{1}{\sqrt{N}}\log \widehat{\text{rPPV}}$ and $\frac{1}{\sqrt{N}}\log \widehat{\text{rNPV}}$ respectively and N is the number of subjects under study. Moskowitz and Pepe (2006) do not present a hypothesis test, but based on the confidence intervals we have the asymptotically χ_1^2 distributed test statistic

$$Z_P = \left(\log\left(\frac{\sqrt{N}}{\widehat{\sigma}_P}\text{rPPV}\right)\right)^2 \tag{36}$$

for testing the null hypothesis (2) for the positive predictive values. When testing the null hypothesis (3) whether the negative predictive values are equal the test statistic

$$Z_N = \left(\log\left(\frac{\sqrt{N}}{\widehat{\sigma}_N}\text{rNPV}\right)\right)^2, \tag{37}$$

which has an asymptotic χ_1^2 distribution can be used. The test statistic in (36) only differs from the test statistic in (32) in the estimated variance. Moskowitz and Pepe (2006) use the multinomial Poisson transformation to simplify the variances.

C RESULTS FROM THE LAP SIMULATION EXPERIMENT

Table 12 shows the estimated test size with 95% confidence limits when comparing the negative predictive values for data generated the null hypothesis. Table 13 and 14 show the estimated test power when comparing the positive and negative predictive values respectively for data generated under the alternative hypothesis.

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 1 LAP test	0.052	0.046	0.059
Case 1 Likelihood ratio test	0.056	0.050	0.063
Case 1 Restricted difference test	0.048	0.043	0.055
Case 1 Unrestricted difference test	0.052	0.046	0.059
Case 2 LAP test	0.050	0.045	0.057
Case 2 Likelihood ratio test	0.050	0.044	0.057
Case 2 Restricted difference test	0.050	0.044	0.056
Case 2 Unrestricted difference test	0.050	0.045	0.057
Case 3 LAP test	0.058	0.052	0.065
Case 3 Likelihood ratio test	0.059	0.053	0.066
Case 3 Restricted difference test	0.052	0.047	0.059
Case 3 Unrestricted difference test	0.060	0.053	0.067
Case 4 LAP test	0.046	0.041	0.052
Case 4 Likelihood ratio test	0.046	0.040	0.052
Case 4 Restricted difference test 4	0.045	0.039	0.051
Case 4 Unrestricted difference test	0.047	0.041	0.053
Case 5 LAP test	0.050	0.044	0.056
Case 5 Likelihood ratio test	0.061	0.055	0.068
Case 5 Restricted difference test	0.049	0.044	0.056
Case 5 Unrestricted difference test	0.050	0.044	0.056
Case 6 LAP test	0.049	0.044	0.056
Case 6 Likelihood ratio test	0.049	0.044	0.056
Case 6 Restricted difference test	0.049	0.043	0.055
Case 6 Unrestricted difference test	0.049	0.044	0.056
Case 7 LAP test	0.060	0.053	0.067
Case 7 Likelihood ratio test	0.067	0.061	0.074
Case 7 Restricted difference test	0.056	0.050	0.063
Case 7 Unrestricted difference test	0.060	0.054	0.067
Case 8 LAP test	0.046	0.040	0.052
Case 8 Likelihood ratio test	0.047	0.041	0.053
Case 8 Restricted difference test	0.044	0.039	0.050
Case 8 Unrestricted difference test	0.046	0.040	0.052

TABLE 12: Estimated test size with 95% confidence limits when testing $\text{NPV}_A = \text{NPV}_B$ for data generated under the null hypothesis using the LAP simulation algorithm.

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 1 LAP test	0.125	0.116	0.135
Case 1 Likelihood ratio test	0.143	0.134	0.153
Case 1 Restricted difference test	0.111	0.102	0.120
Case 1 Unrestricted difference test	0.141	0.131	0.151
Case 2 LAP test	0.396	0.383	0.410
Case 2 Likelihood ratio test	0.390	0.376	0.403
Case 2 Restricted difference test	0.380	0.367	0.394
Case 2 Unrestricted difference test	0.400	0.386	0.413
Case 3 LAP test	0.369	0.356	0.383
Case 3 Likelihood ratio test	0.361	0.348	0.375
Case 3 Restricted difference test	0.349	0.336	0.363
Case 3 Unrestricted difference test	0.369	0.356	0.383
Case 4 LAP test	0.945	0.938	0.951
Case 4 Likelihood ratio test	0.944	0.937	0.950
Case 4 Restricted difference test	0.943	0.936	0.949
Case 4 Unrestricted difference test	0.944	0.938	0.950
Case 5 LAP test	0.146	0.137	0.156
Case 5 Likelihood ratio test	0.174	0.163	0.184
Case 5 Restricted difference test	0.124	0.116	0.134
Case 5 Unrestricted difference test	0.158	0.149	0.169
Case 6 LAP test	0.463	0.450	0.477
Case 6 Likelihood ratio test	0.458	0.444	0.472
Case 6 Restricted difference test	0.449	0.435	0.463
Case 6 Unrestricted difference test	0.466	0.452	0.479
Case 7 LAP test	0.485	0.471	0.498
Case 7 Likelihood ratio test	0.484	0.470	0.498
Case 7 Restricted difference test	0.468	0.454	0.482
Case 7 Unrestricted difference test	0.485	0.471	0.498
Case 8 LAP test	0.987	0.984	0.990
Case 8 Likelihood ratio test	0.987	0.984	0.990
Case 8 Restricted difference test	0.987	0.983	0.990
Case 8 Unrestricted difference test	0.987	0.984	0.990

TABLE 13: Estimated power with 95% confidence limits when testing $\text{PPV}_A = \text{PPV}_B$ for data generated under the alternative hypothesis.

Case/test	$\hat{\alpha}$	$\hat{\alpha}_L$	$\hat{\alpha}_U$
Case 1 LAP test	0.324	0.312	0.338
Case 1 Likelihood ratio test	0.336	0.323	0.349
Case 1 Restricted difference test	0.048	0.043	0.055
Case 1 Unrestricted difference test	0.052	0.046	0.059
Case 2 LAP test	0.929	0.921	0.935
Case 2 Likelihood ratio test	0.929	0.921	0.935
Case 2 Restricted difference test	0.050	0.044	0.056
Case 2 Unrestricted difference test	0.050	0.045	0.057
Case 3 LAP test	0.113	0.105	0.122
Case 3 Likelihood ratio test	0.114	0.105	0.123
Case 3 Restricted difference test	0.052	0.047	0.059
Case 3 Unrestricted difference test	0.060	0.053	0.067
Case 4 LAP test	0.350	0.337	0.364
Case 4 Likelihood ratio test	0.349	0.336	0.362
Case 4 Restricted difference test	0.045	0.039	0.051
Case 4 Unrestricted difference test	0.047	0.041	0.053
Case 5 LAP test	0.427	0.413	0.441
Case 5 Likelihood ratio test	0.458	0.444	0.471
Case 5 Restricted difference test	0.049	0.044	0.056
Case 5 Unrestricted difference test	0.050	0.044	0.056
Case 6 LAP test	0.986	0.982	0.989
Case 6 Likelihood ratio test	0.986	0.982	0.989
Case 6 Restricted difference test	0.049	0.043	0.055
Case 6 Unrestricted difference test	0.049	0.044	0.056
Case 7 LAP test	0.136	0.127	0.146
Case 7 Likelihood ratio test	0.146	0.136	0.156
Case 7 Restricted difference test	0.056	0.050	0.063
Case 7 Unrestricted difference test	0.060	0.054	0.067
Case 8 LAP test	0.465	0.451	0.478
Case 8 Likelihood ratio test	0.463	0.450	0.477
Case 8 Restricted difference test	0.044	0.039	0.050
Case 8 Unrestricted difference test	0.046	0.040	0.052

TABLE 14: Estimated power with 95% confidence limits when testing $\text{NPV}_A = \text{NPV}_B$ for data generated under the alternative hypothesis using the LAP simulation algorithm.

D Computational remarks

When using the TANGO program Andreani et al. (2007), Andreani et al. (2008), there are several parameters that can be set or modified by the user. Along with the specification of the objective function and the constraints, the initial estimates of the Lagrange multipliers, the initial values of the variables and their lower and upper bounds must be set. Other parameters have a default value, but these can be altered by the user. These parameters include tolerance limits and the maximum number of iterations.

In our simulation studies we have chosen the initial value 0.0 for all the variables with upper and lower bounds ± 200000. The initial value for the Lagrangian multiplier was set to 0.0 as advised in the program when one does not believe it should have a specific value. The feasibility and optimality tolerances are 10^{-4} by default. We found that with these tolerances, the resulting variable values depend on both the initial value of the Lagrange multiplier and the initial values of the variables. However, different initial values for the variables give more similar results than different initial values of the Lagrange multipliers. The smaller the tolerance is, the more similar the results will be, so in order to get results that do not depend on any of the initial values one should use smaller values for the tolerances and in our problems, smaller than 10^{-4}. The problem is then that it takes longer for the algorithm to converge. When performing the likelihood ratio test for one or a few datasets this is not an issue, but when performing simulation experiments with several thousand datasets this will slow down the experiment considerably.

Another problem is that of the algorithm converging to a local maximum. For example, the analytical restricted likelihood estimates for the table $\boldsymbol{n} = (0, 7, 0, 69, 5, 3, 11, 5)$ is -115.73 while it is -166.38 using the numerical estimates from TANGO. The difference is caused by the fact that $\tilde{p}_3 = 0$ using the numerical optimization routine, while it is 0.04 using the analytical optimization. For most of the large sample datasets the difference is less, with e.g. 15 of 5000 estimates in Case 1 in the LAP simulation experiment differ by more than 1.0 between the analytical and numerical estimates. We recommend using the analytical estimates.