Nonparametric predictive comparison of two diagnostic tests based on total numbers of correctly diagnosed individuals

Manal H. Alabdulhadi*
Department of Mathematics, Qassim University, Saudi Arabia
Frank P.A. Coolen†, Tahani Coolen-Maturi‡
Department of Mathematical Sciences, Durham University, UK

January 2, 2019

Abstract

In clinical applications, it is important to compare and study the ability of diagnostic tests to discriminate between individuals with and without the disease. In this paper, comparison of two diagnostic tests is presented and discussed using nonparametric predictive inference (NPI). We compare the two tests by considering the total numbers of correct diagnoses for specific numbers of future healthy individuals and future patients. This NPI approach for comparison of diagnostic tests is also generalized by the use of weighted sums for the healthy and patients groups, reflecting possibly different importance of correct diagnoses. Examples are provided to illustrate the new method.

Keywords: Comparison of diagnostic tests; lower and upper probabilities; nonparametric predictive inference.

1 Introduction

Developing and improving diagnostic tests to detect the presence or absence of a particular disease are important in medical applications. Often, researchers are asked to confirm the superiority of a new diagnostic test to an existing test. In practice, diagnostic tests are not

*Email: manalhamd@hotmail.com
†Email: frank.coolen@durham.ac.uk (corresponding author)
‡Email: tahani.maturi@durham.ac.uk

perfect. The tests can have two types of errors, namely false-negative (FN) and false-positive (FP) errors. This raises the question how one can compare the qualities of different diagnostic tests. Various methods to compare two diagnostic tests have been presented in the literature [26, 28]. The performance of a diagnostic test can be evaluated by indicators such as sensitivity, specificity, positive and negative likelihood ratio, or positive and negative predictive values. Using these indicators for comparison of two tests may not be straightforward, typically one test may have higher specificity while the other test may have higher sensitivity.

Measures such as the Youden index, have been suggested as global measures of diagnostic accuracy [28]. However, the Youden index can be misleading when comparing two diagnostic tests. The Youden index is not taking into account the differences in the specificity and sensitivity of the diagnostic test, and it treats the FN and FP errors as equally undesirable. The area under the receiver operating characteristic (ROC) curve (AUC) also provides a summary measure of the diagnostic test ability [28]. Although the AUC has been used to compare different diagnostic tests, it has some limitations. For example, the areas under the ROC curves of two diagnostic tests can be equal, yet the shapes of the two ROC curves can be different over the part of the ROC curves of main clinical relevance. According to Dodd and Pepe [21], the area under the ROC curve might summarize the performance of a diagnostic test over regions of the curve of no clinical and practical interest. Alternatively, the partial area under the ROC curve can provide more information for some diagnostic tests which require false-positive rates to be within a specific range of medical interest [21, 25]. Researchers have also presented the use of hypothesis testing to compare sensitivities, specificities or the areas under the ROC curves of two diagnostic tests [28].

As an alternative to the methods for comparison of two diagnostic tests mentioned above, we present a Nonparametric Predictive Inference (NPI) method for such comparisons [6, 7, 8]. NPI is a frequentist statistical method which is explicitly aimed at using few modelling assumptions, enabled through the use of lower and upper probabilities to quantify uncertainty [4, 5, 27]. NPI has been introduced for many application areas where the predictive nature of this method is attractive, including reliability, survival analysis, operations research and finance (see www.npi-statistics.com for more information). Restricting attention to one future observation, NPI has been developed for diagnostic test accuracy considering different types of data. Coolen-Maturi et al. [18] introduced NPI for diagnostic test accuracy with binary data, while Elkhafifi and Coolen [22] presented NPI for diagnostic tests with ordinal data. Coolen-Maturi et al. [17, 19]

proposed NPI for two- and three-group ROC analysis with continuous data. The results by Elkhafifi and Coolen [22] have been generalised by Coolen-Maturi [14] for three-group ROC analysis with ordinal data. Coolen-Maturi [15] considered NPI for scenarios where two or more diagnostic tests are combined in order to improve the overall accuracy. Recently, we have presented NPI methods for determining an optimal test threshold for diagnostic tests with real-valued outcomes, explicitly considering given numbers of future individuals from the healthy and disease groups [1, 16]. As an alternative to the approach in this paper, we have also presented comparison of two diagnostic tests with real-valued outcomes for diagnostic tests with two or three groups of individuals, where the comparisons consider the NPI lower and upper probabilities of the events that at least a specified proportion of future individuals, for each of the groups, will be correctly classified [2]. This differs from the method presented in this paper in two ways, namely the use of different events of interest as we consider the (possibly weighted) total number of correctly diagnosed individuals in this paper, and the NPI methods used in the two papers differ. In Alabdulhadi et al. [2], NPI for future order statistics [3, 12] was used, while in this paper the problem formulation requires the use NPI for Bernoulli quantities [6], for which also new results are presented.

Classical methods often focus on estimation rather than prediction. The end goal of studying the accuracy of diagnostic tests is to apply these tests to future individuals. Thus, it is of interest to consider the use of a frequentist predictive inference method for comparison of diagnostic tests as an alternative to the classical methods that have been presented in the literature. It will be useful to apply the NPI approach together with some other approaches, to see if they provide similar conclusions about the different tests. If the NPI approach comes to quite a different conclusion than classical methods, then it is likely due to the model assumptions underlying the other methods, as only few assumptions are made in the NPI method.

In this paper, we present NPI for comparing two diagnostic tests, assuming that the tests are applied on the same individuals from two groups, namely, healthy and diseased individuals. In Section 2 we provide a brief review of NPI for Bernoulli quantities [6]. Section 3 presents the main method for comparison of two diagnostic tests by considering the total sum of correctly classified individuals from both the healthy and disease groups. We also show how this method can be generalized to include weights for the two different groups, to express possibly different importance of getting the diagnosis right for either healthy or diseased individuals. This section contains some new results for NPI for Bernoulli quantities which can also be applied to different

problems than the comparison of diagnostic tests. Section 4 presents some examples to illustrate and discuss the new method. Finally, some concluding remarks are made in Section 5.

2 NPI for Bernoulli quantities

Coolen [6] presented NPI for Bernoulli quantities, which is based on Hill's assumption $A_{(n)}$ [23, 24], sequentially applied to derive at inference for $m \geq 1$ future observations given n observed values, together with a latent variable representation of Bernoulli quantities represented as observations on the real line, with a threshold such that observations to one side are successes and to the other side failures. Suppose that there is a sequence of $n+m$ exchangeable Bernoulli trials, each with success and failure as possible outcomes, and data consisting of s successes in n trials. Let Y_1^n denote the random number of successes in trials 1 to n; then a sufficient representation of the data for NPI is $Y_1^n = s$, due to assumed exchangeablility of all trials. Let Y_{n+1}^{n+m} denote the random number of successes in trials $n+1$ to $n+m$. Based on the basic method presented by Coolen [6], Coolen and Coolen-Schrijner [13] introduced the NPI lower and upper probabilities for events $Y_{n+1}^{n+m} \geq y$ and $Y_{n+1}^{n+m} < y$, these are the only lower and upper probabilities needed in this paper. The upper probabilities for these events are as follows. For $y \in \{0, 1, ..., m\}$ and $0 < s < n$,

$$\overline{P}(Y_{n+1}^{n+m} \geq y | Y_1^n = s) = \binom{n+m}{n}^{-1} \left[\binom{s+y}{s} \binom{n-s+m-y}{n-s} + \sum_{l=y+1}^{m} \binom{s+l-1}{s-1} \binom{n-s+m-l}{n-s} \right]$$

and for $y \in \{1, ..., m+1\}$ and $0 < s < n$,

$$\overline{P}(Y_{n+1}^{n+m} < y | Y_1^n = s) = \binom{n+m}{n}^{-1} \left[\binom{n-s+m}{n-s} + \sum_{l=1}^{y-1} \binom{s+l-1}{s-1} \binom{n-s+m-l}{n-s} \right]$$

The corresponding NPI lower probabilities can be derived via the conjugacy property $\underline{P}(A) = 1 - \overline{P}(A^c)$, for any event A and its complementary event A^c, which holds generally in imprecise probability theory and also in NPI for Bernoulli quantity [5, 6].

For $m = 1$, the two non-trivial values of these upper probabilities are $\overline{P}(Y_{n+1}^{n+1} \geq 1 | Y_1^n = s) = (s+1)/(n+1)$ and $\overline{P}(Y_{n+1}^{n+1} < 1 | Y_1^n = s) = (n-s+1)/(n+1)$. If the observed data are all successes, so $s = n$, or all failures, so $s = 0$, then these upper probabilities are, for all

$y \in \{0, 1, ..., m\}$, $\overline{P}(Y_{n+1}^{n+m} \geq y | Y_1^n = n) = 1$ and $\overline{P}(Y_{n+1}^{n+m} \geq y | Y_1^n = 0) = \frac{\binom{n+m-y}{n}}{\binom{n+m}{n}}$, and for all $y \in \{0, 1, ..., m+1\}$, $\overline{P}(Y_{n+1}^{n+m} < y | Y_1^n = n) = \frac{\binom{n+y-1}{n}}{\binom{n+m}{n}}$ and $\overline{P}(Y_{n+1}^{n+m} < y | Y_1^n = 0) = 1$.

3 Comparison of tests using NPI for Bernoulli quantities

In this section, we compare the accuracy of two diagnostic tests to classify individuals into one of two groups, which we indicate as 'healthy group' X and 'disease group' Y. Throughout this paper, we use either subscripts x and y or superscripts X and Y to refer to groups X and Y, respectively. We explicitly consider multiple future individuals from each group, with the inference based on observed data for individuals known to belong to either the healthy group or the disease group. Throughout this paper we assume that the two groups are fully independent in the sense that any information about one group does not provide any information about the other group.

We compare the two tests by considering the total number of correct diagnoses for m_x future healthy individuals and m_y future patients for one test with those for the other test, using NPI for Bernoulli quantities for each group separately. In this paper, it does not matter what kind of measurements are actually used in the diagnostic tests, the only relevant aspect is whether or not the diagnosis is correct. However, the model underlying NPI for Bernoulli quantities [6] assumes a latent variable representation for successes and failures using real-valued observations and a threshold, such that an observation to one side of the threshold is a success and to the other side is a failure. This provides a natural link to diagnostic tests which provide real-valued outcomes, with an optimal threshold determined on the basis of the data and some optimality criterion. We have recently presented NPI methods for determination of an optimal diagnostic threshold for such a scenario [1, 16], and this motivated us to develop the method presented in this paper. We also considered comparison of two diagnostic tests which are restricted to the real-valued case, and with criterion to maximize the NPI lower or upper probability of correctly classifying at least two specified proportions of the future individuals from the healthy and diseased group [2]. That work uses NPI for future order statistics [12], and cannot be used for the criterion on (possibly weighted) total number of correct future diagnoses considered in this paper. The method presented in the current paper can also be applied in different diagnostic scenarios as long as one can identify whether or not a diagnosis is correct.

The number of correct diagnoses by test t, for $t = 1, 2$, in n_x and n_y data observations from

groups X and Y, are denoted by s_x^t and s_y^t, respectively. Let $C_{m_x}^{X^t}$ denote the random number of successful diagnoses for m_x healthy future individuals according to test t, and let $C_{m_y}^{Y^t}$ denote the random number of successful diagnoses for m_y diseased future individuals for test t. We compare the two tests by considering the random total number of correct diagnoses for the $m_x + m_y$ future individuals, when each test would be applied to them. Hence we consider the event $C_{m_x}^{X^1} + C_{m_y}^{Y^1} > C_{m_x}^{X^2} + C_{m_y}^{Y^2}$ and develop the NPI lower and upper probabilities for this event. These results have not been presented before for such quantities, and of course these NPI lower and upper probabilities can also be useful for scenarios other than comparison of diagnostic tests. For $C_{m_x}^{X^1}, C_{m_x}^{X^2} \in \{0, ..., m_x\}$ and $C_{m_y}^{Y^1}, C_{m_y}^{Y^2} \in \{0, ..., m_y\}$, and based on data $(n_x, s_x^1), (n_y, s_y^1)$ and $(n_x, s_x^2), (n_y, s_y^2)$, the NPI upper probability for this event is derived as follows

$$
\begin{aligned}
&\overline{P}(C_{m_x}^{X^1} + C_{m_y}^{Y^1} > C_{m_x}^{X^2} + C_{m_y}^{Y^2}) \\
&= \sum_{k=0}^{m_x+m_y} \overline{P}(C_{m_x}^{X^2} + C_{m_y}^{Y^2} < k) \times [\overline{P}(C_{m_x}^{X^1} + C_{m_y}^{Y^1} \geq k) - \overline{P}(C_{m_x}^{X^1} + C_{m_y}^{Y^1} \geq k+1)] \\
&= \sum_{k=0}^{m_x+m_y} \Big[\sum_{v=0}^{m_y} \overline{P}(C_{m_x}^{X^2} < k-v) \times [\overline{P}(C_{m_y}^{Y^2} \leq v) - \overline{P}(C_{m_y}^{Y^2} \leq v-1)]\Big] \\
&\times \Big[\sum_{v=0}^{m_y} \overline{P}(C_{m_x}^{X^1} \geq k-v) \times [\overline{P}(C_{m_y}^{Y^1} \geq v) - \overline{P}(C_{m_y}^{Y^1} \geq v+1)] \\
&- \sum_{v=0}^{m_y} \overline{P}(C_{m_x}^{X^1} \geq k+1-v) \times [\overline{P}(C_{m_y}^{Y^1} \geq v) - \overline{P}(C_{m_y}^{Y^1} \geq v+1)]\Big] \\
&= \sum_{k=0}^{m_x+m_y} \Big[\sum_{v=0}^{m_y} \overline{P}(C_{m_x}^{X^2} < k-v) \times [\overline{P}(C_{m_y}^{Y^2} \leq v) - \overline{P}(C_{m_y}^{Y^2} \leq v-1)]\Big] \\
&\times \Big[\sum_{v=0}^{m_y} [\overline{P}(C_{m_x}^{X^1} \geq k-v) - \overline{P}(C_{m_x}^{X^1} \geq k+1-v)] \times [\overline{P}(C_{m_y}^{Y^1} \geq v) - \overline{P}(C_{m_y}^{Y^1} \geq v+1)]\Big] \qquad (1)
\end{aligned}
$$

The first equation in this derivation follows from the fact that $\overline{P}(C_{m_x}^{X^2} + C_{m_y}^{Y^2} < k)$ is increasing in k. Hence, to derive the NPI upper probability for the event of interest, we put the maximum possible probability mass for $C_{m_x}^{X^1} + C_{m_y}^{Y^1}$ at the event $C_{m_x}^{X^1} + C_{m_y}^{Y^1} \geq m_x + m_y$, followed by assigning the maximum possible remaining probability mass for $C_{m_x}^{X^1} + C_{m_y}^{Y^1}$ at the event $C_{m_x}^{X^1} + C_{m_y}^{Y^1} \geq m_x + m_y - 1$, etc [13]. We can interpret Equation (1) as if we are optimistic for Test 1 by putting the maximum possible probability masses for this test at the larger values of $C_{m_x}^{X^1}$ and $C_{m_y}^{Y^1}$, while we are pessimistic for Test 2 so we put the maximum possible probability masses

for this test at the smaller values of $C_{m_x}^{X^2}$ and $C_{m_y}^{Y^2}$. The NPI lower and upper probabilities for the individual sums of Bernoulli quantities in the final formula above are as given in Section 2.

We also consider the event $C_{m_x}^{X^1} + C_{m_y}^{Y^1} \geq C_{m_x}^{X^2} + C_{m_y}^{Y^2}$, for which the NPI upper probability is derived as above, with just the first term $\overline{P}(C_{m_x}^{X^2} < k - v)$ on the right-hand side after the final equality in Equation (1) replaced by $\overline{P}(C_{m_x}^{X^2} \leq k - v)$. The corresponding NPI lower probabilities for these two events can be derived via the conjugacy property $\underline{P}(A) = 1 - \overline{P}(A^c)$, together with the obvious swapping of the Test 1 and Test 2 indicators in the respective formulae.

It is important to note that the NPI method presented in this paper, where the predictive inferences are done separately for the future individuals from group X and group Y, following which we consider the sum of numbers of correct diagnoses, differs from the possible simpler approach to only count the total number of successful diagnoses, both in the data and for future individuals, without taking the different groups into account. The latter approach would straightforwardly use the NPI for Bernoulli data method for comparison of different groups [13] and would lead to less imprecision, that is corresponding NPI lower and upper probabilities would differ less. Particularly in situations where the sample sizes for the two groups differ substantially, one could get quite different results if one neglects the fact that there are two groups. In addition, our approach can be generalized to reflect that correct diagnoses may be more important for one group than for the other group.

We can take different importance of correct diagnosis for the two groups into account by using weighted totals of correctly diagnosed individuals. As we will consider the same weighted total for both tests, the weights used can be scaled to any total. For ease of presentation, we will use positive integer valued weights w_x for group X and w_y for group Y. We now compare the two diagnostic tests by considering the event $w_x C_{m_x}^{X^1} + w_y C_{m_y}^{Y^1} > w_x C_{m_x}^{X^2} + w_y C_{m_y}^{Y^2}$. The NPI upper probability for this event, which also has not been presented elsewhere and may have applications to a wider range of statistical problems, is derived as follows

$$
\begin{aligned}
&\overline{P}(w_x C_{m_x}^{X^1} + w_y C_{m_y}^{Y^1} > w_x C_{m_x}^{X^2} + w_y C_{m_y}^{Y^2}) \\
&= \sum_{k=0}^{w_x m_x + w_y m_y} \overline{P}(w_x C_{m_x}^{X^2} + w_y C_{m_y}^{Y^2} < k) \times [\overline{P}(w_x C_{m_x}^{X^1} + w_y C_{m_y}^{Y^1} \geq k) - \overline{P}(w_x C_{m_x}^{X^1} + w_y C_{m_y}^{Y^1} \geq k+1)] \\
&= \sum_{k=0}^{w_x m_x + w_y m_y} \Big[\sum_{v=0}^{m_y} \overline{P}(C_{m_x}^{X^2} < \frac{k - w_y v}{w_x}) \times [\overline{P}(C_{m_y}^{Y^2} \leq v) - \overline{P}(C_{m_y}^{Y^2} \leq v-1)]\Big] \\
&\times \Big[\sum_{v=0}^{m_y} \overline{P}(C_{m_x}^{X^1} \geq \frac{k - w_y v}{w_x}) \times [\overline{P}(C_{m_y}^{Y^1} \geq v) - \overline{P}(C_{m_y}^{Y^1} \geq v+1)] \\
&- \sum_{v=0}^{m_y} \overline{P}(C_{m_x}^{X^1} \geq \frac{k+1 - w_y v}{w_x}) \times [\overline{P}(C_{m_y}^{Y^1} \geq v) - \overline{P}(C_{m_y}^{Y^1} \geq v+1)]\Big] \\
&= \sum_{K=0}^{w_x m_x + w_y m_y} \Big[\sum_{v=0}^{m_y} \overline{P}(C_{m_x}^{X^2} < \frac{k - w_y v}{w_x}) \times [\overline{P}(C_{m_y}^{Y^2} \leq v) - \overline{P}(C_{m_y}^{Y^2} \leq v-1)]\Big] \\
&\times \Big[\sum_{v=0}^{m_y} [\overline{P}(C_{m_x}^{X^1} \geq \frac{k - w_y v}{w_x}) - \overline{P}(C_{m_x}^{X^1} \geq \frac{k+1 - w_y v}{w_x})] \times [\overline{P}(C_{m_y}^{Y^1} \geq v) - \overline{P}(C_{m_y}^{Y^1} \geq v+1)]\Big] \qquad (2)
\end{aligned}
$$

The NPI upper probability for the event $w_x C_{m_x}^{X^1} + w_y C_{m_y}^{Y^1} \geq w_x C_{m_x}^{X^2} + w_y C_{m_y}^{Y^2}$ is again derived by replacing the first term after the final equality in Equation (2), $\overline{P}(C_{m_x}^{X^2} < (k - w_y v)/w_x)$, by $\overline{P}(C_{m_x}^{X^2} \leq (k - w_y v)/w_x)$, and the corresponding lower probabilities can again be derived via the conjugacy property.

4 Examples

In this section we illustrate the NPI method for comparison of two diagnostic tests introduced in Section 3. A special feature of our method is that the number of future individuals from both the healthy and disease groups must be specified for the event of interest in the comparison. We therefore consider the application of the method for different values of m_x and m_y, which we mostly assume to be equal but we also consider what happens when they are not equal. The first two examples use made up data in order to illustrate the approach and discuss its important features. Example 3 uses data from the literature and is linked to an application of our recently presented NPI method to determine the optimal diagnostic threshold for real-valued data [1, 16].

Example 1 Assume that two diagnostic tests have been applied to the same $n_x = 10$ individuals from healthy group X and $n_y = 10$ individuals from disease group Y. The numbers of correctly diagnosed individuals when Test 1 is used are $s_x^1 = s_y^1 = 8$ from both groups, while

m	$[\underline{P}, \overline{P}](T^1 > T^2)$	$[\underline{P}, \overline{P}](T^1 \geq T^2)$	$[\underline{P}, \overline{P}](T^2 > T^1)$	$[\underline{P}, \overline{P}](T^2 \geq T^1)$
1	[0.3672, 0.5317]	[0.7748, 0.8853]	[0.1147, 0.2252]	[0.4683, 0.6328]
3	[0.5133, 0.7564]	[0.7286, 0.8996]	[0.1004, 0.2714]	[0.2436, 0.4867]
5	[0.5702, 0.8342]	[0.7232, 0.9179]	[0.0821, 0.2768]	[0.1658, 0.4298]
15	[0.6644, 0.9321]	[0.7298, 0.9535]	[0.0465, 0.2702]	[0.0679, 0.3356]
30	[0.7019, 0.9578]	[0.7374, 0.9664]	[0.0336, 0.2626]	[0.0422, 0.2981]
50	[0.7199, 0.9675]	[0.7421, 0.9721]	[0.0279, 0.2579]	[0.0325, 0.2801]
100	[0.7350, 0.9743]	[0.7464, 0.9764]	[0.0236, 0.2536]	[0.0257, 0.2650]

Table 1: NPI lower and upper probabilities for comparison of two tests with $m_x = m_y = m$

m_x	m_y	$[\underline{P}, \overline{P}](T^1 > T^2)$	$[\underline{P}, \overline{P}](T^1 \geq T^2)$	$[\underline{P}, \overline{P}](T^2 > T^1)$	$[\underline{P}, \overline{P}](T^2 \geq T^1)$
3	5	[0.5459, 0.8008]	[0.7230, 0.9078]	[0.0922, 0.2770]	[0.1992, 0.4541]
5	3	[0.5459, 0.8008]	[0.7230, 0.9078]	[0.0922, 0.2770]	[0.1992, 0.4541]
30	15	[0.6823, 0.9430]	[0.7272, 0.9564]	[0.0436, 0.2728]	[0.0570, 0.3177]
50	70	[0.7225, 0.9683]	[0.7410, 0.9720]	[0.0280, 0.2590]	[0.0317, 0.2775]
100	80	[0.7322, 0.9729]	[0.7447, 0.9752]	[0.0248, 0.2553]	[0.0271, 0.2678]

Table 2: NPI lower and upper probabilities for comparison of two tests with $m_x \neq m_y$

for Test 2 these numbers are $s_x^2 = s_y^2 = 6$ for both groups. To denote the events of interest concisely, we introduce notation $T^1 = C_{m_x}^{X^1} + C_{m_y}^{Y^1}$ and $T^2 = C_{m_x}^{X^2} + C_{m_y}^{Y^2}$, where the values m_x and m_y will be clear from the tables or the context.

Tables 1 and 2 present the NPI lower and upper probabilities for the events $T^1 > T^2$, $T^1 \geq T^2$, $T^2 > T^1$ and $T^2 \geq T^1$ for different values of m_x and m_y, which are set equal in Table 1 but differ in Table 2. It is obvious from the data that Test 1 has performed better for the observed individuals than Test 2, for both healthy and diseased groups. The aim of this example is to show how such a better performance is reflected by the predictive inferences to compare the two tests if they are applied to m_x and m_y future individuals from the groups X and Y.

The first thing to note from Table 1 is that the entries in the last two columns, that is the NPI lower and upper probabilities for the events $T^2 > T^1$ and $T^2 \geq T^1$, could have been deleted as they follow from the entries for the events $T^1 \geq T^2$ and $T^1 > T^2$, respectively, by use of the conjugacy property. However, we have included them because it simplifies comparison of the NPI lower and upper probabilities for all these events. The better performance of Test 1 than of Test 2 is reflected by larger values of the lower and upper probabilities for the event $T^1 > T^2$ than for the event $T^2 > T^1$, and larger values for $T^1 \geq T^2$ than for $T^2 \geq T^1$.

Comparing the lower and upper probabilities for the events $T^1 > T^2$ and $T^1 \geq T^2$, for the same value of $m = m_x = m_y$, shows that these differ a lot for small m yet the differences decrease for increasing m, to become very small for $m = 100$. This is of course due to the

m	$[\underline{P}, \overline{P}](T^1 > T^2)$	$[\underline{P}, \overline{P}](T^1 \geq T^2)$	$[\underline{P}, \overline{P}](T^2 > T^1)$	$[\underline{P}, \overline{P}](T^2 \geq T^1)$
1	[0.2331, 0.3920]	[0.6970, 0.8371]	[0.1629, 0.3030]	[0.6080, 0.7669]
5	[0.3294, 0.6610]	[0.5109, 0.8121]	[0.1879, 0.4891]	[0.3390, 0.6706]
6	[0.3344, 0.6851]	[0.4948, 0.8153]	[0.1847, 0.5052]	[0.3149, 0.6656]
10	[0.3442, 0.7432]	[0.4552, 0.8269]	[0.1731, 0.5448]	[0.2568, 0.6558]
50	[0.3534, 0.8420]	[0.3819, 0.8595]	[0.1405, 0.6181]	[0.1580, 0.6466]
100	[0.3540, 0.8577]	[0.3688, 0.8664]	[0.1336, 0.6312]	[0.1423, 0.6460]

Table 3: NPI lower and upper probabilities for comparison of two tests with $m_x = m_y = m$

fact that, for small m, it is quite likely that one gets $T^1 = T^2$, yet for larger m this becomes unlikely. Due to this effect, it is easiest to study the effect of different choices for the value m by looking at the event $T^1 \geq T^2$. We note that the lower and upper probabilities for this event vary with m, the upper probability increases while the lower probability first decreases and then increases slightly. This is not a pattern observed in all such examples, it varies from case to case. But overall the imprecision, that is the difference between corresponding upper and lower probabilities, tends to increase for larger values of m, unless a lower probability gets close to 1 (or an upper probability close to 0), which forces imprecision to become small as the corresponding upper probability cannot exceed 1 (and the lower probability cannot be less than 0). This example shows that, for the predictive criterion chosen in this paper to compare two diagnostic tests, the actual choice of the numbers of future individuals considered has some influence on the results.

In Table 2 the NPI lower and upper probabilities for the comparison of these two diagnostic tests are given for some cases with $m_x \neq m_y$. Of course, due to the data for groups X and Y being the same for both tests, the first two reported cases lead to the same results. We furthermore see similar aspects as discussed above for the situation with equal numbers of future individuals for both groups.

Example 2 In this example, we consider two tests that have similar total numbers of correct diagnoses for the groups X and Y. As in the previous example, we set $n_x = n_y = 10$ and the observed numbers of correct diagnoses for Test 1 are $s_x^1 = 7$ for group X and $s_y^1 = 9$ for group Y, while for Test 2 the numbers are $s_x^2 = 9$ and $s_y^2 = 6$, respectively.

Tables 3 and 4 present the NPI lower and upper probabilities for the same four events for comparison of these two tests as in the previous example, with Table 3 presenting results for $m_x = m_y = m$ and Table 4 presenting some cases with $m_x \neq m_y$. The values of the lower and upper probabilities for the event $T^1 > T^2$ are a bit higher than for the event $T^2 > T^1$, for

m_x	m_y	$[\underline{P}, \overline{P}](T^1 > T^2)$	$[\underline{P}, \overline{P}](T^1 \geq T^2)$	$[\underline{P}, \overline{P}](T^2 > T^1)$	$[\underline{P}, \overline{P}](T^2 \geq T^1)$
15	30	[0.5510, 0.9101]	[0.6069, 0.9315]	[0.0685, 0.3931]	[0.0899, 0.4490]
30	15	[0.1850, 0.6315]	[0.2286, 0.6856]	[0.3144, 0.7714]	[0.3685, 0.8150]
50	70	[0.4670, 0.9034]	[0.4918, 0.9137]	[0.0863, 0.5082]	[0.0966, 0.5330]
70	50	[0.2515, 0.7658]	[0.2726, 0.7847]	[0.2153, 0.7274]	[0.2342, 0.7485]

Table 4: NPI lower and upper probabilities for comparison of two tests with $m_x \neq m_y$

the same value of m, and similar for the events including equality of T^1 and T^2, reflecting the slightly better performance of Test 1 on the 20 data observations than Test 2. Of course, the differences here are much smaller than in Example 1, as the tests have performed very similarly here but Test 1 had performed quite a bit better than Test 2 in Example 1. For larger values of m, where $T^1 = T^2$ becomes unlikely, all intervals created by the lower and upper probabilities in this example contain the value 0.5, which one could interpret as there not being a strong indication that either test is better than the other. Note that there is substantial imprecision in this example, in particular for the larger values of m. If we had larger data sets with similarly close performance, the imprecision would be less.

The results in Table 4 are quite different to those for the case with unequal values for m_x and m_y in Example 1. Since Test 1 is better for diagnoses for group Y while Test 2 is better for diagnoses for group X, this is reflected in the predictive inference for the future performance if one considers different numbers of individuals from these groups. For relatively small numbers, one of m_x and m_y equal to 15 and the other equal to 30, we see that for more future individuals from group Y Test 1 performs better than Test 2, while for more future individuals from group X Test 2 performs better. The differences between the entries in the first two rows of this table are large, which shows the influence that different choices of m_x and m_y can have, while there is also much imprecision, due to the small samples. However, once we consider larger numbers of future individuals, namely 50 and 70, Test 1 remains better than Test 2 if there are more future individuals from group Y, but even with more individuals from group X Test 1 is still marginally better than Test 2. This reflects that Test 1 was overall a little better for the observed data, while the values of m_x and m_y are relatively close. Note that there is again quite much imprecision, so with NPI lower and upper probabilities as presented here for the case $m_x = 70$ and $m_y = 50$ one would reach the conclusion that there is very little evidence that one test would be better than the other.

To illustrate the use of weights for the different groups, as also presented in Section 3, Table 5 presents the NPI lower and upper probabilities for comparison of the two diagnostic tests in this

m	$[\underline{P}, \overline{P}](T^1 > T^2)$	$[\underline{P}, \overline{P}](T^1 \geq T^2)$	$[\underline{P}, \overline{P}](T^2 > T^1)$	$[\underline{P}, \overline{P}](T^2 \geq T^1)$
	$w_x = 2, w_y = 1$			
1	[0.2398, 0.3986]	]0.5987, 0.7449]	[0.2551, 0.4013]	[0.6014, 0.7602]
5	[0.2418, 0.5521]	[0.3490, 0.6666]	[0.3334, 0.6510]	[0.4479, 0.7582]
15	[0.2063, 0.6207]	[0.2487, 0.6707]	[0.3293, 0.7513]	[0.3793, 0.7937]
50	[0.1785, 0.6629]	[0.1919, 0.6803]	[0.3197, 0.8081]	[0.3371, 0.8215]
100	[0.1702, 0.6747]	[0.1770, 0.6837]	[0.3163, 0.8230]	[0.3253, 0.8298]
	$w_x = 1, w_y = 2$			
1	[0.3315, 0.4843]	[0.6903, 0.8305]	[0.1695, 0.3097]	[0.5157, 0.6685]
5	[0.4954, 0.7880]	[0.6097, 0.8650]	[0.1350, 0.3903]	[0.2120, 0.5046]
15	[0.5464, 0.8896]	[0.5974, 0.9128]	[0.0872, 0.4026]	[0.1104, 0.4536]
50	[0.5764, 0.9346]	[0.5944, 0.9404]	[0.0596, 0.4056]	[0.0654, 0.4236]
100	[0.5847, 0.9446]	[0.5941, 0.9473]	[0.0527, 0.4059]	[0.0554, 0.4153]

Table 5: NPI lower and upper probabilities for comparison of two tests for $m_x = m_y = m$, using different weights

example, using weights to let successful diagnoses for one group be twice as important as for the other group. We restrict attention here to equal numbers of future individuals, $m_x = m_y = m$, to ensure that the effects illustrated are resulting from the use of the weights. Using weights $w_x = 2$ and $w_y = 1$, Test 2 is better than Test 1 for all considered values of m, albeit only marginally so for small m. This reflects that Test 2 had a better performance than Test 1 for individuals from group X in the data. For $w_x = 1$ and $w_y = 2$, Test 1 compares favourably to Test 2 for all considered values of m, also reflecting that Test 1 had performed better than Test 2 for group Y in the data. Note that, for the latter case, Test 1 is quite a bit stronger than Test 2, while the difference was not so large in the first case with the weights the other way around. This reflects that Test 1 had performed slightly better overall in the observed data. Also these lower and upper probabilities have quite some imprecision, which suggests that larger data samples may be needed before a final decision can be made on the choice of the diagnostic test for the future individuals.

Example 3 In this example, we use a data set from a study to develop screening methods to detect carriers of a rare genetic disorder. The data were discussed by Cox et al. [20] (available from http://lib.stat.cmu.edu/datasets/). Four tests are applied on the same blood samples, each taking a real-valued measurement. The tests are indicated by $M1$, $M2$, $M3$ and $M4$. For some patients, there were several measurements for the same test, in such cases the average is taken, and five patients with some missing values are excluded from the analysis. The remaining sample, which is used in this example, consists of 120 individuals, 38 carriers of the rare genetic disorder, which we call group X, and 82 non-carriers, group Y.

m :	1	5	10	30	100
s_x^{M1}, s_y^{M1}	70, 32	70, 32	70, 32	70, 32	70, 32
s_x^{M2}, s_y^{M2}	56, 28	58, 27	58, 27	58, 27	58, 27
s_x^{M3}, s_y^{M3}	74, 24	70, 25	70, 25	57, 27	57, 27
s_x^{M4}, s_y^{M4}	67, 31	67, 31	67, 31	67, 31	67, 31

Table 6: The number of correct diagnoses in the data from groups X and Y for Tests $M1$, $M2$, $M3$ and $M4$

To illustrate our method for comparison of two diagnostic tests, we first decided on the optimal diagnostic threshold for each test. To stay within the NPI framework, we applied the recently presented method [1, 16] where we choose the threshold which maximizes the NPI lower probability that at least half of the m_x future individuals from group X will be correctly diagnoses, and also at least half of the m_y future individuals from group Y. Throughout this example, we set $m_x = m_y = m$. How the specific thresholds are chosen is in itself not important for the illustration of our method for comparison of the tests, but by choosing this NPI method we will see an important feature of such comparisons that may otherwise have gone unnoticed.

First, we applied the above mentioned method to find the optimal diagnostic thresholds for the four tests and for different values of m. It is important here to note that the threshold, using the NPI method to determine it, can vary for different values of m. We only need the numbers of correctly diagnosed individuals from both groups X and Y for our comparison method, we denote the numbers by s_x^{Mt} and s_y^{Mt}, respectively, for $t = 1, 2, 3, 4$. We further denote the random number of correctly diagnosed future individuals for Test Mt for group X by $C_m^{X^t}$, and for group Y by $C_m^{Y^t}$. We base our predictive comparison of the tests on the random total numbers $T^{Mt} = C_m^{X^t} + C_m^{Y^t}$ for the four tests.

Table 6 shows the number of successful diagnoses in the data from healthy and diseased groups for each test, for different values of m. Test $M1$ performs best overall for the data observations, if we consider the total observed correct diagnoses. Test $M4$ is second best, and both these tests had the same optimal threshold for all the considered values of m. For Tests $M2$ and $M3$ the situation is less clear, and the optimal threshold is not the same for all m. For Test $M2$, the optimal threshold is slightly different for $m = 1$ than for the larger values of m considered, but for Test $M3$ the optimal threshold differs much more, leading to substantially different numbers of correctly diagnosed individuals from both groups for small m compared to larger values of m. It should be noted here that this is due to the multi-modal shape of our criterion function for specific values of m, as function of the threshold, while also our criterion

changes with m. This multi-modality also happens for other methods to determine the optimal threshold, so it is not a peculiarity of the NPI approach, although of course other methods presented in the literature are not predictive, hence do not depend on m, and hence they tend not to show this feature. The criterion functions have very similar values at several modes, but picking the threshold by overall optimisation of the functions, for different m, can lead to quite different thresholds and hence quite different numbers of correctly diagnosed individuals from the two groups in the data set. We will see that this feature can substantially impact on the comparison of the diagnostic tests.

We present the pairwise comparisons for all pairs of these tests, by considering the NPI lower and upper probabilities, as presented in Section 3, for different values of m in Table 7. Test $M1$ was the best for the data, and this shows in the comparisons of this test with each of the other tests. For small m, there is again a considerable possibility that any two tests considered lead to the same total number of correct future diagnoses, as can be seen from the differences in the first and second, and third and fourth, columns with lower and upper probabilities in this table. This effect decreases for larger m, and Test $M1$ has high lower and upper probabilities to be better than Tests $M2$ and $M3$ for $m = 100$, while it is also quite likely to be better in this case than Test $M4$. Test $M4$ is also likely to perform better than $M2$ and $M3$, so this is all in line with the conclusions drawn from the observed data, although these predictive inferences provide far more detailed information and they provide much insight into the role of m for the predictions. Note further here that imprecision is far smaller than in Examples 1 and 2, reflecting that there is considerably more information from the data in this example.

The most interesting pairwise comparison here is between Tests $M2$ and $M3$, mainly due to the changes of optimal thresholds as discussed above, and the corresponding changes in numbers of correctly diagnosed individuals from groups X and Y. For smaller values of m, here $m = 1, 5, 10$, the future performance of Test $M3$ is likely to be slightly better than that of Test $M2$, but for larger values of m, here $m = 30, 100$, it is the other way around, with only a very small difference. The latter reflects that, for these larger m, Test $M2$ has one more correct diagnosis for group X in the data than Test $M3$, with the same number of correct diagnoses for group Y. For the smaller values of m, it is quite different as Test $M3$ then performed considerably better on the data for group X but worse for group Y. It turns out, however, that using these data for predictive inference, with $m_x = m_y = m$, indicates a better performance to be likely for Test $M3$ than for Test $M2$ for these smaller values of m, something which would

have been quite impossible to foresee without this formal predictive inference method being used.

More aspects of this example are considered in the PhD thesis of the first-named author [1], including a comparison of Tests $M2$ and $M3$ under the assumption that the thresholds used do not vary with m, but are the ones used for $m = 100$ in the analysis above. As that led to Test $M2$ diagnosing one more individual correctly, this test is then of course slightly better than Test $M3$ for all choices of m in our comparison. Finally, it is worth to mention that the empirical areas under the ROC curves (AUC) for these four tests are equal to $\widehat{AUC}_{M1} = 0.9034$, $\widehat{AUC}_{M2} = 0.7526$, $\widehat{AUC}_{M3} = 0.8232$ and $\widehat{AUC}_{M4} = 0.8798$. This is often considered to be a useful measure to distinguish between diagnostic accuracy of tests an NPI method for diagnostic accuracy leading to lower and upper AUCs, which always bound the empirical AUC, has also been presented [17]. While these results also indicate that Tests $M1$ and $M4$ are the two best tests, they do not show any possible further aspects of comparison for Tests $M2$ and $M3$, and it is also unclear what these quantities actually mean for future application of the tests. We should emphasize here that we are not advocating the use of our proposed method on its own, as there is certainly value in measures such as the empirical AUC, but considering several methods, including ours, and studying the results carefully can provide interesting insights for important applications.

5 Concluding remarks

This paper introduces a new method for comparison of two diagnostic tests to distinguish between two groups, based on the numbers of correctly diagnosed individuals from both groups in a data set. The method uses NPI for Bernoulli quantities, and leads to lower and upper probabilities for the event that the total number of correctly diagnosed future individuals from both groups is greater for one test than for the other, if we consider m_x future individuals from group X and m_y from group Y. We believe that such predictive inferences provide valuable insights and can be used together with more traditional ways for comparison of tests. The explicitly predictive nature can be natural when one considers that any decisions with regard to choice of test will be relevant for future individuals.

We have not discussed how to choose m_x and m_y, this is not a trivial issue and we mainly wish to emphasize in this paper that the actual values of these quantities can make a difference

to the overall conclusion on which test is best. If the results clearly indicate that one test is better than another one for some values of m_x and m_y, and the test is applied sequentially but one needs to select a single test to be used for multiple future individuals, then one could for example safely choose the better test for a number of future diagnoses that is equal to the minimum of these two numbers. This because one would of course not know whether the future individuals are from group X or group Y. Similar reasoning was used by Coolen [9] to determine the maximum group size for simultaneous testing in high potential risk scenarios. It is also possible that a practitioner may have a fair idea about the proportion of future individuals from either group, this could be used in our analysis by considering the m_x and m_y in similar proportion.

Another possible choice for these numbers of future individuals would be $m_x = n_x$ and $m_y = n_y$. This could be of particular interest for studying reproducibility characteristics of the tests, a topic that recently has received increasing interest as there is much confusion about it, and for which NPI methods have proven to be attractive due to their explicitly predictive nature [10, 11].

Acknowledgements

The authors gratefully acknowledge supportive statements of two reviewers and the journal's editor on the original manuscript.

Conflict of interest statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] Alabdulhadi, M.H. (2018). *Nonparametric Predictive Inference for Diagnostic Test Thresholds*. PhD thesis, Durham University (available from *www.npi-statistics.com*.)

[2] Alabdulhadi, M.H., Coolen-Maturi, T. and Coolen, F.P.A. (2018). Nonparametric predictive inference for comparison of two diagnostic tests. *In submission*.

[3] Alqifari, H.N. (2017). *Nonparametric Predictive Inference for Future Order Statistics*. PhD thesis, Durham University (available from *www.npi-statistics.com*.)

[4] Augustin, T. and Coolen, F.P.A. (2004). Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, **124**, 251-272.

[5] Augustin, T., Coolen, F.P.A., de Cooman, G. and Troffaes, M.C.M. (2014). *Introduction to Imprecise Probabilities*. Wiley, Chichester.

[6] Coolen, F.P.A. (1998). Low structure imprecise predictive inference for Bayes' problem. *Statistics and Probability Letters*, **36**, 349-357.

[7] Coolen, F.P.A. (2006). On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, **15**, 21-47.

[8] Coolen, F.P.A. (2011). Nonparametric predictive inference, In: *International Encyclopedia of Statistical Science*, Lovric, M. (ed.). Springer, Berlin, pp. 968-970.

[9] Coolen, F.P.A. (2013). Maximum group sizes for simultaneous testing in high potential risk scenarios. *Journal of Risk and Reliability*, **227**, 569-575.

[10] Coolen, F.P.A. and Alqifari, H.N. (2018). Nonparametric predictive inference for reproducibility of two basic tests based on order statistics. *Revstat - Statistical Journal*, **16**, 167-185.

[11] Coolen, F.P.A. and Bin Himd, S. (2014). Nonparametric predictive inference for reproducibility of basic nonparametric tests. *Journal of Statistical Theory and Practice*, **8**, 591-618.

[12] Coolen, F.P.A., Coolen-Maturi, T. and Alqifari, H.N. (2018). Nonparametric predictive inference for future order statistics. *Communications in Statistics - Theory and Methods*, **47**, 2527-2548.

[13] Coolen, F.P.A. and Coolen-Schrijner, P. (2007). Nonparametric predictive comparison of proportions. *Journal of Statistical Planning and Inference*, **137**, 23-33.

[14] Coolen-Maturi, T. (2017). Three-group ROC predictive analysis for ordinal outcomes. *Communications in Statistics: Theory and Methods*, **46**, 9476-9493.

[15] Coolen-Maturi, T. (2017). Predictive inference for best linear combination of biomarkers subject to limits of detection. *Statistics in Medicine*, 36, 2844-2874.

[16] Coolen-Maturi, T., Coolen, F.P.A. and Alabdulhadi, M.H. (2018). Nonparametric predictive inference for diagnostic test thresholds. *Communications in Statistics: Theory and Methods*, to appear.

[17] Coolen-Maturi, T., Coolen-Schrijner, P. and Coolen, F.P.A. (2012). Nonparametric predictive inference for diagnostic accuracy. *Journal of Statistical Planning and Inference*, **142**, 1141-1150.

[18] Coolen-Maturi, T., Coolen-Schrijner, P. and Coolen, F.P.A. (2012). Nonparametric predictive inference for binary diagnostic tests. *Journal of Statistical Theory and Practice*, **6**, 665-680.

[19] Coolen-Maturi, T., Elkhafifi, F.F. and Coolen, F.P.A. (2014). Three-group ROC analysis: A nonparametric predictive approach, *Computational Statistics and Data Analysis*, **78**, 69-81.

[20] Cox, L.H., Johnson, M.M. and Kafadar, K. (1982). Exposition of statistical graphics technology. *ASA Proceedings of the Statistical Computation Section*, 55-56.

[21] Dodd, L.E. and Pepe, M.S. (2003). Partial AUC estimation and regression. *Biometrics*, **59**, 614-623.

[22] Elkhafifi, F.F. and Coolen, F.P.A. (2012). Nonparametric predictive inference for accuracy of ordinal diagnostic tests. *Journal of Statistical Theory and Practice*, **6**, 681-697.

[23] Hill, B.M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, **63**, 677-691.

[24] Hill, B.M. (1988). De Finetti's theorem, induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). In: *Bayesian Statistics 3*, Bernardo, J.M. et al. (eds.). Oxford University Press, pp. 211-241.

[25] Li, C.R., Liao, C.T. and Liu, J.P. (2008). A non-inferiority test for diagnostic accuracy based on the paired partial areas under ROC curves. *Statistics in Medicine*, **27**, 1762-1776.

[26] Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

[27] Weichselberger, K. (2000). The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, **24**, 149-170.

[28] Zhou, X.H., Obuchowski, N.A. and McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine*. Wiley, New York.

m	$[\underline{P}, \overline{P}](T^{M1} > T^{M2})$	$[\underline{P}, \overline{P}](T^{M1} \geq T^{M2})$	$[\underline{P}, \overline{P}](T^{M2} > T^{M1})$	$[\underline{P}, \overline{P}](T^{M2} \geq T^{M1})$
1	[0.3611, 0.3965]	[0.8312, 0.8568]	[0.1432, 0.1688]	[0.6035, 0.6389]
5	[0.6392, 0.7119]	[0.8115, 0.8621]	[0.1379, 0.1885]	[0.2881, 0.3608]
10	[0.7448, 0.8245]	[0.8445, 0.9014]	[0.0986, 0.1555]	[0.1755, 0.2552]
30	[0.8783, 0.9428]	[0.9104, 0.9605]	[0.0395, 0.0896]	[0.0572, 0.1217]
100	[0.9502, 0.9860]	[0.9571, 0.9883]	[0.0117, 0.0429]	[0.0140, 0.0498]
	$[\underline{P}, \overline{P}](T^{M1} > T^{M3})$	$[\underline{P}, \overline{P}](T^{M1} \geq T^{M3})$	$[\underline{P}, \overline{P}](T^{M3} > T^{M1})$	$[\underline{P}, \overline{P}](T^{M3} \geq T^{M1})$
1	[0.3008, 0.3373]	[0.8107, 0.8394]	[0.1606, 0.1893]	[0.6627, 0.6992]
5	[0.5473, 0.6290]	[0.7490, 0.8120]	[0.1880, 0.2510]	[0.3710, 0.4527]
10	[0.6345, 0.7350]	[0.7625, 0.8415]	[0.1585, 0.2375]	[0.2650, 0.3655]
30	[0.8900, 0.9492]	[0.9197, 0.9652]	[0.0348, 0.0803]	[0.0508, 0.1100]
100	[0.9580, 0.9886]	[0.9639, 0.9905]	[0.0095, 0.0361]	[0.0114, 0.0420]
	$[\underline{P}, \overline{P}](T^{M1} > T^{M4})$	$[\underline{P}, \overline{P}](T^{M1} \geq T^{M4})$	$[\underline{P}, \overline{P}](T^{M4} > T^{M1})$	$[\underline{P}, \overline{P}](T^{M4} \geq T^{M1})$
1	[0.2359, 0.2694]	[0.7847, 0.8157]	[0.1843, 0.2153]	[0.7306, 0.7641]
5	[0.4114, 0.4975]	[0.6418, 0.7199]	[0.2801, 0.3582]	[0.5025, 0.5886]
10	[0.4592, 0.5760]	[0.6142, 0.7209]	[0.2791, 0.3858]	[0.4240, 0.5408]
30	[0.5173, 0.6865]	[0.5946, 0.7525]	[0.2475, 0.4054]	[0.3135, 0.4827]
100	[0.5598, 0.7719]	[0.5907, 0.7950]	[0.2050, 0.4093]	[0.2281, 0.4402]
	$[\underline{P}, \overline{P}](T^{M2} > T^{M3})$	$[\underline{P}, \overline{P}](T^{M2} \geq T^{M3})$	$[\underline{P}, \overline{P}](T^{M3} > T^{M2})$	$[\underline{P}, \overline{P}](T^{M3} \geq T^{M2})$
1	[0.2185, 0.2475]	[0.6659, 0.6986]	[0.3014, 0.3341]	[0.7525, 0.7815]
5	[0.2846, 0.3515]	[0.4700, 0.5448]	[0.4552, 0.5300]	[0.6485, 0.7154]
10	[0.2733, 0.3633]	[0.3939, 0.4935]	[0.5065, 0.6061]	[0.6367, 0.7267]
30	[0.4188, 0.5638]	[0.4825, 0.6262]	[0.3738, 0.5175]	[0.4362, 0.5812]
100	[0.4208, 0.6148]	[0.4464, 0.6394]	[0.3606, 0.5536]	[0.3852, 0.5792]
	$[\underline{P}, \overline{P}](T^{M2} > T^{M4})$	$[\underline{P}, \overline{P}](T^{M2} \geq T^{M4})$	$[\underline{P}, \overline{P}](T^{M4} > T^{M2})$	$[\underline{P}, \overline{P}](T^{M4} \geq T^{M2})$
1	[0.1725, 0.1992]	[0.6263, 0.6604]	[0.3396, 0.3737]	[0.8008, 0.8275]
5	[0.1848, 0.2420]	[0.3505, 0.4251]	[0.5749, 0.6495]	[0.7580, 0.8152]
10	[0.1499, 0.2201]	[0.2448, 0.3347]	[0.6653, 0.7552]	[0.7799, 0.8501]
30	[0.0831, 0.1618]	[0.1129, 0.2077]	[0.7923, 0.8871]	[0.8382, 0.9169]
100	[0.0379, 0.1067]	[0.0440, 0.1201]	[0.8799, 0.9560]	[0.8933, 0.9621]
	$[\underline{P}, \overline{P}](T^{M3} > T^{M4})$	$[\underline{P}, \overline{P}](T^{M3} \geq T^{M4})$	$[\underline{P}, \overline{P}](T^{M4} > T^{M3})$	$[\underline{P}, \overline{P}](T^{M4} \geq T^{M3})$
1	[0.1931, 0.2228]	[0.6843, 0.7190]	[0.2810, 0.3157]	[0.7772, 0.8069]
5	[0.2451, 0.3137]	[0.4389, 0.5196]	[0.4804, 0.5611]	[0.6863, 0.7549]
10	[0.2279, 0.3188]	[0.3500, 0.4556]	[0.5444, 0.6500]	[0.6812, 0.7721]
30	[0.0746, 0.1477]	[0.1020, 0.1910]	[0.8090, 0.8980]	[0.8523, 0.9254]
100	[0.0318, 0.0927]	[0.0371, 0.1048]	[0.8952, 0.9629]	[0.9073, 0.9682]

Table 7: Lower and upper probabilities for pairwise comparisons of Tests $M1$, $M2$, $M3$ and $M4$