Nonparametric predictive inference for binary diagnostic tests

Tahani Coolen-Maturi, *Kent Business School, University of Kent, UK.*
Email: T.Coolen-Maturi@kent.ac.uk
Pauline Coolen-Schrijner[1], Frank P.A. Coolen[2], *Department of Mathematical Sciences, Durham University, UK.*
Email: frank.coolen@durham.ac.uk

Received: February 2012 *Revised:* May 2012

Abstract

Measuring the accuracy of diagnostic tests is crucial in many application areas including medicine, health care and data mining. Good methods for determining diagnostic accuracy provide useful guidance on selection of patient treatment, and the ability to compare different diagnostic tests has a direct impact on quality of care. In this paper nonparametric predictive inference (NPI) for accuracy of diagnostic tests with binary test results is presented and discussed, together with methods for comparison of two such tests. NPI does not aim at inference for an entire population but instead explicitly considers future observations, which is particularly suitable for inference to support decisions on medical diagnosis for one future patient, or for a pre-determined number of future patients, so the NPI approach provides an attractive alternative to standard methods.

AMS Subject Classification: 60A99, 62G99, 62P10

Keywords: Binary data; diagnostic test accuracy; effect size; lower and upper probability; nonparametric predictive inference; pairwise comparison.

---

[1]Pauline died in 2008, when the research reported in this paper was at an advanced stage.
[2]Corresponding author

1 Introduction

The evaluation of the accuracy of a diagnostic test is crucial in many research and application areas such as medicine, radiology, machine learning and data mining. Traditional statistical methods tend to use concepts like 'sensitivity' and 'specificity' to express such accuracy. These are conditional probabilities which are properties of assumed populations and are estimated from available data, in line with the traditional frequentist approach to statistics. In recent years, nonparametric predictive inference (NPI) has been developed as an alternative frequentist statistical framework which is based on few modelling assumptions and considers one or more future observations instead of a population. This predictive nature of NPI can be attractive for diagnostic tests as one may wish to consider explicitly the quality of the test for one or more future individuals.

In this paper we show how NPI can be used to provide an alternative approach to inference on accuracy of binary diagnostic tests and for comparison of two such tests. In Section 2, some common measures of diagnostic accuracy are briefly reviewed, while NPI for Bernoulli quantities is reviewed in Section 3, where it is also generalized by introducing new results that are related to the concept of 'effect size'. In Section 4 we present NPI for binary diagnostic tests, both considering the test accuracy and comparison of two tests. Section 5 provides an example to illustrate and discuss the new method presented in this paper. Some concluding remarks are given in Section 6.

2 Binary diagnostic tests

In this section we briefly review some common measures of diagnostic accuracy, following Pepe (2003) and Dodd and Pepe (2003). Let D be a binary variable describing the disease status, i.e. $D = 1$ for disease and $D = 0$ for non-disease. Suppose that Y is the diagnostic test result which takes binary outcomes, with $Y = 1$ reflecting a test result which indicates disease ('positive') and $Y = 0$ indicating non-disease ('negative'). Table 1 shows the classification and notation of test results according to disease status. For example, 'true positive' occurs when an individual who has the disease is correctly classified by a positive test result, whereas 'false positive' occurs when an individual who does not have the disease has a positive test result.

	$D=0$	$D=1$
$Y=0$	True Negative	False Negative
$Y=1$	False Positive	True Positive

	$D=0$	$D=1$	
$Y=0$	n_0^-	n_1^-	n^-
$Y=1$	n_0^+	n_1^+	n^+
	n_0	n_1	

Table 1: The classification of test results by disease status

Table 1 also introduces the data notation corresponding to such a binary diagnosis of n individuals, where n_1 (n_0) is the number of individuals who (do not) have the disease and n^+ (n^-) the number of individuals who are classified positive (negative) to the disease by the diagnostic test used. This use of subscripts and superscripts is combined in the logical way, so n_1^+ (n_1^-) is the number of individuals who have the

disease and their test result is positive (negative). Similarly, n_0^+ (n_0^-) is the number of individuals who do not have the disease and their test result is positive (negative). Of course, these numbers sum up row- and columnwise in Table 1, and $n_0 + n_1 = n^- + n^+ = n$.

The Sensitivity (SN) of a test is the probability of a positive test result for an individual who has the disease, this is also known as True Positive Fraction (TPF $= P[Y = 1|D = 1]$). The Specificity (SP) is the probability of a negative test result for an individual without the disease. Clearly, an accurate diagnostic test will have both sensitivity and specificity close to one. The False Positive Fraction (FPF) is the probability of a positive test result for an individual without the disease (FPF $= P[Y = 1|D = 0]$), hence SP $= 1 -$ FPF. The empirical estimators for the TPF and FPF are $\widehat{\text{TPF}} = n_1^+/n_1$ and $\widehat{\text{FPF}} = n_0^+/n_0$. Two further important concepts are the Positive Predictive Value (PPV) and the Negative Predictive Value (NPV) (Weinstein et al., 2005). The PPV is the probability that an individual has the disease when the test result is positive (PPV $= P[D = 1|Y = 1]$), while the NPV is the probability that the individual does not have the disease when the test result is negative (NPV $= P[D = 0|Y = 0]$). The empirical estimators for these probabilities are $\widehat{\text{PPV}} = n_1^+/n^+$ and $\widehat{\text{NPV}} = n_0^-/n^-$. The sensitivity and specificity are normally used to evaluate the accuracy of a diagnostic test, and consequently to choose the best test if there are several possible tests. The PPV and NPV, on the other hand, are important for determining how to treat the individual once the test result is available. In Section 4 we introduce the same underlying concepts for these measures in NPI, i.e. FPF, TPF, PPV and NPV, however, the inferences will explicitly be in terms of a given number of future individuals and will not be properties of an assumed infinite-size population. Note that particularly the suggested use of PPV and NPV are predictive, that is setting a prognosis for a specific individual, and as such a predictive inferential approach like NPI provides an attractive alternative to the established methods.

To compare the accuracy of two tests, paired and unpaired designs can be used. In a paired design, each individual undergoes both tests while in an unpaired design each individual undergoes only a single test. According to Pepe (2003), the unpaired study design is more applicable in many circumstances than a paired design for several reasons, which include possible discomfort or risk for individuals undergoing two treatments and the possibility that the performance of one test may interfere with the performance of the other test. Throughout this paper the unpaired design is considered, which allows the assumption of independence of random quantities corresponding to different groups of individuals according to the test they undergo.

To compare two diagnostic tests, say A and B, one can use some function of FPF and TPF such as the absolute difference, the odds ratio or the relative probabilities which are defined by

$$r\text{TPF}(A, B) = \text{TPF}_A/\text{TPF}_B \qquad \text{and} \qquad r\text{FPF}(A, B) = \text{FPF}_A/\text{FPF}_B$$

and similarly rPPV and rNPV can be defined. Test A is more accurate than test B if $r\text{TPF}(A, B) > 1$ and $r\text{FPF}(A, B) < 1$ or, equivalently, if $r\text{PPV}(A, B) > 1$ and $r\text{NPV}(A, B) > 1$, in other cases it would be less straightforward to decide which test performs better (Pepe, 2003).

3 Nonparametric predictive inference for Bernoulli quantities

Coolen (1998) introduced direct lower and upper probabilities for the number of successes in m future trials, given information about n past trials. These lower and upper probabilities follow from an assumed underlying latent variable representation together with Hill's assumption $A_{(n)}$ (Hill, 1968), which has an explicitly predictive nature, and fit in the framework of nonparametric predictive inference (NPI) (Augustin and Coolen, 2004; Coolen, 2006). Several inferential problems involving Bernoulli data have been addressed using this NPI approach, for example comparisons of groups of Bernoulli data (Coolen and Coolen-Schrijner, 2006, 2007), acceptance sampling (Coolen and Elsaeiti, 2009) and system reliability (Coolen-Schrijner et al., 2008). We briefly summarize this approach, for more details and justification we refer to Coolen (1998) and Coolen and Coolen-Schrijner (2006, 2007).

Suppose that, given data consisting of s successes in n trials, denoted by $X_1^n = s$, we are interested in the number of successes in m future trials, denoted by X_{n+1}^{n+m}, where the outcomes of all trials are considered to be exchangeable. Let $R_t = \{r_1, \ldots, r_t\}$, with $1 \leq t \leq m+1$ and $0 \leq r_1 < r_2 < \ldots < r_t \leq m$, and let $\binom{s+r_0}{s} = 0$. The NPI upper probability for the event $X_{n+1}^{n+m} \in R_t$, given data $X_1^n = s$ with $s \in \{0, \ldots, n\}$, is (Coolen, 1998)

$$\overline{P}(X_{n+1}^{n+m} \in R_t | X_1^n = s) = \binom{n+m}{n}^{-1} \sum_{j=1}^{t} \left[\binom{s+r_j}{s} - \binom{s+r_{j-1}}{s} \right] \binom{n-s+m-r_j}{n-s} \tag{1}$$

The corresponding lower probability can be derived via

$$\underline{P}(X_{n+1}^{n+m} \in R_t | X_1^n = s) = 1 - \overline{P}(X_{n+1}^{n+m} \in R_t^c | X_1^n = s) \tag{2}$$

where $R_t^c = \{0, 1, \ldots, m\} \setminus R_t$. If one is interested in the event that precisely r successes occur in m trials, then (1) and (2) can be used to calculate the NPI lower and upper probabilities for this event with $t = 1$ and $R_1 = \{r\}$. It also follows from (1) and (2) that, for $x \in \{0, 1, \ldots, m\}$ and $0 < s < n$,

$$\overline{P}(X_{n+1}^{n+m} \geq x | X_1^n = s) = \binom{n+m}{n}^{-1} \left[\binom{s+x}{s} \binom{n-s+m-x}{n-s} + \sum_{l=x+1}^{m} \binom{s+l-1}{s-1} \binom{n-s+m-l}{n-s} \right] \tag{3}$$

and for $x \in \{1, \ldots, m+1\}$ and $0 < s < n$,

$$\overline{P}(X_{n+1}^{n+m} < x | X_1^n = s) = \binom{n+m}{n}^{-1} \left[\binom{n-s+m}{n-s} + \sum_{l=1}^{x-1} \binom{s+l-1}{s-1} \binom{n-s+m-l}{n-s} \right] \tag{4}$$

It is again easiest to derive the corresponding NPI lower probabilities via the conjugacy property $\underline{P}(A) = 1 - \overline{P}(A^c)$.

Coolen and Coolen-Schrijner (2007) introduced both pairwise and multiple comparisons to compare numbers of successes in m future Bernoulli trials from different groups, in this paper we restrict attention to pairwise comparisons. The NPI upper probabilities for such predictive comparison of two groups, A and B, with related notation using index a and b, respectively, are (Coolen and Coolen-Schrijner, 2007)

$$\overline{P}\left(X_{a,n_a+1}^{n_a+m} \geq X_{b,n_b+1}^{n_b+m} | X_{a,1}^{n_a} = s_a, X_{b,1}^{n_b} = s_b\right) = \sum_{x=0}^{m} \overline{A}_x \, \overline{P}\left(X_{b,n_b+1}^{n_b+m} \leq x | X_{b,1}^{n_b} = s_b\right) \tag{5}$$

$$\overline{P}\left(X_{a,n_a+1}^{n_a+m} > X_{b,n_b+1}^{n_b+m} | X_{a,1}^{n_a} = s_a, X_{b,1}^{n_b} = s_b\right) = \sum_{x=0}^{m} \overline{A}_x \, \overline{P}\left(X_{b,n_b+1}^{n_b+m} < x | X_{b,1}^{n_b} = s_b\right) \tag{6}$$

with

$$\overline{A}_x = \overline{P}(X_{a,n_a+1}^{n_a+m} \geq x | X_{a,1}^{n_a} = s_a) - \overline{P}(X_{a,n_a+1}^{n_a+m} \geq x+1 | X_{a,1}^{n_a} = s_a) \tag{7}$$

The corresponding NPI lower probabilities are derived by replacing all upper probabilities in formulae (5), (6) and (7) by their corresponding lower probabilities.

These NPI lower and upper probabilities for pairwise comparison of two groups can be used for comparing two different diagnostic tests and also for the comparison of the performance of a single test on individuals from the disease and non-disease groups, this use of NPI for binary diagnostic tests is described in Section 4.

It is also of interest to consider in more detail the difference between groups A and B. We introduce an attractive and natural way of doing this in NPI, that is similar to the use of the so-called 'effect size' in hypothesis testing (Borenstein et al., 2009). Continuing with the notation and concepts introduced above, we consider the following generalizations of (5) and (6), for $d = 0, 1, \ldots m-1$,

$$\overline{P}\left(X_{a,n_a+1}^{n_a+m} \geq X_{b,n_b+1}^{n_b+m} + d | X_{a,1}^{n_a} = s_a, X_{b,1}^{n_b} = s_b\right) = \sum_{x=0}^{m} \overline{A}_x \, \overline{P}\left(X_{b,n_b+1}^{n_b+m} \leq x - d | X_{b,1}^{n_b} = s_b\right) \tag{8}$$

$$\overline{P}\left(X_{a,n_a+1}^{n_a+m} > X_{b,n_b+1}^{n_b+m} + d | X_{a,1}^{n_a} = s_a, X_{b,1}^{n_b} = s_b\right) = \sum_{x=0}^{m} \overline{A}_x \, \overline{P}\left(X_{b,n_b+1}^{n_b+m} < x - d | X_{b,1}^{n_b} = s_b\right) \tag{9}$$

with $\overline{A}_x$ as given by (7). The corresponding NPI lower probabilities are again derived by replacing all upper probabilities in these formulae by their corresponding lower probabilities. Considering these NPI lower and upper probabilities as functions of d provides valuable insight into the actual strength of the evidence in the data with regard to the differences for the future individuals of the two groups considered. We illustrate this here for some general values, we will further discuss these NPI lower and upper probabilities as functions of d in the example in Section 5. In the figures that follow we link up the values of these upper probabilities at different integer values of d by

straight lines, and similarly for the lower probabilities, to give clearer illustrations of these functions.

Figure 1 shows, for the case with $n_a = n_b = n$ and $m_a = m_b = m$, the effect of different values of n and m, assuming that $s_a = 0.7n$ and $s_b = 0.4n$. We present the NPI lower and upper probabilities for the event $X_{a,n+1}^{n+m} > X_{b,n+1}^{n+m} + d$ given data $X_{a,1}^{n} = s_a, X_{b,1}^{n} = s_b$ as function of d, for different values for m and n ($m = 5, 25, 100$ and $n = 10, 100, 1000$). As usual in NPI, imprecision tends to decrease as function of n and to increase as function of m. Consider for example the case with $m = 100$ and $n = 100$, so with $s_a = 70$ and $s_b = 40$. We see that the NPI lower and upper probabilities are quite large for d up to about 20, while they are small for d larger than about 40. So, for a further 100 individuals from each of the two groups, the event that there will be at least about 20 more successes from group A than from group B would be very likely to happen, but this difference would be unlikely to exceed 40. This clearly reflects, in a predictive manner, the data from the 100 observations from each group, which had 70 successes for group A and 40 for group B.

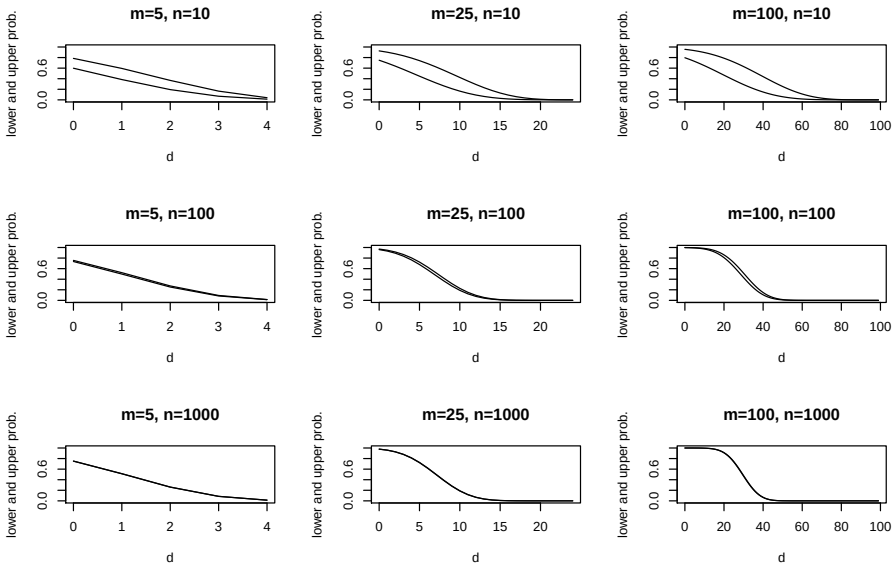

Figure 1: NPI lower and upper probabilities for $X_{a,n+1}^{n+m} > X_{b,n+1}^{n+m} + d | X_{a,1}^{n} = 0.7n, X_{b,1}^{n} = 0.4n$

Figure 2 illustrates the effect of different s_a and s_b for the case with $n = 100, m = 25$. The first figure corresponds to the case with 90 successes observed in 100 trials for group A with only 20 successes out of 100 trials for group B. For a further 25 observations from each group, the number of successes for group A is very likely to be at least 10 larger than for group B, but this difference is unlikely to exceed 20. The second figure relates to the case with 60 successes observed in 100 trials for group A and 50 for group B. In this case it is already pretty unlikely that group A will have 5 or

more successes more than group B for 25 further trials. Finally, the third figure shows the results for a case where group A had fewer successes in the data than group B, namely 20 and 30 respectively. This shows clearly in the predictive inferences through very small lower and upper probabilities for the event considered, for all values of d.

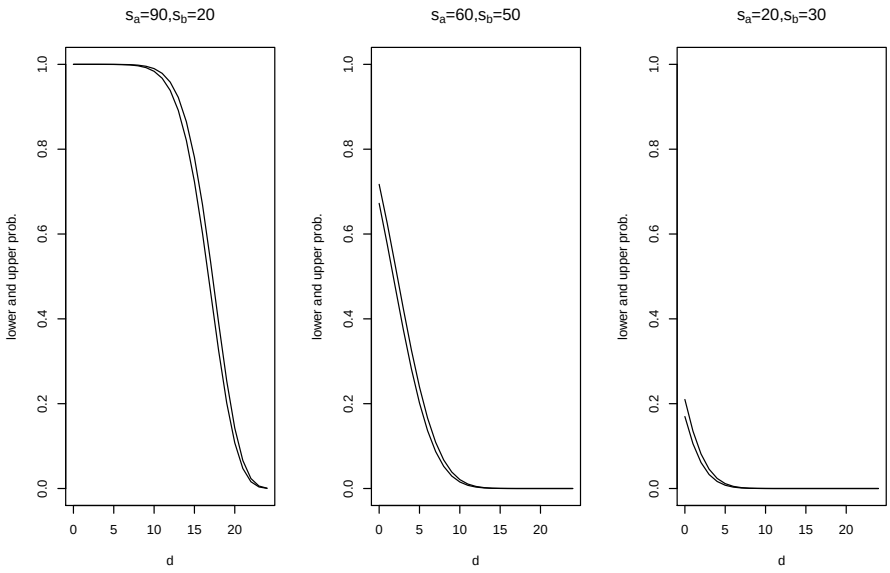

Figure 2: NPI lower and upper probabilities for $X_{a,101}^{125} > X_{b,101}^{125} + d | X_{a,1}^{100} = s_a, X_{b,1}^{100} = s_b$

In Figure 3 we briefly consider the effect of different numbers of data observations for the two groups, for the case with $m = 25$ and, as in Figure 1, assuming $s_a = 0.7n_a$ and $s_b = 0.4n_b$. We present the NPI lower and upper probabilities for the event $X_{a,n_a+1}^{n_a+25} > X_{b,n_b+1}^{n_b+25} + d$ given data $X_{a,1}^{n_a} = 0.7n_a, X_{b,1}^{n_b} = 0.4n_b$, as function of d, for the three different cases $(n_a = 10, n_b = 15)$, $(n_a = 100, n_b = 75)$ and $(n_a = 1000, n_b = 400)$, in the first, second and third plot, respectively. The main difference between these three plots is the decreasing imprecision, which is the difference between corresponding upper and lower probabilities, for increasing total number of data observations.

4 NPI for binary diagnostic tests

With the overview of the results of Coolen (1998) and Coolen and Coolen-Schrijner (2006, 2007) and the generalization involving a difference of at least d successes between the two groups, as presented in Section 3, we can now introduce NPI for binary diagnostic tests. We consider first NPI for accuracy of a single diagnostic test applied to both a disease and a non-disease group, followed by comparison of two different diagnostic tests.

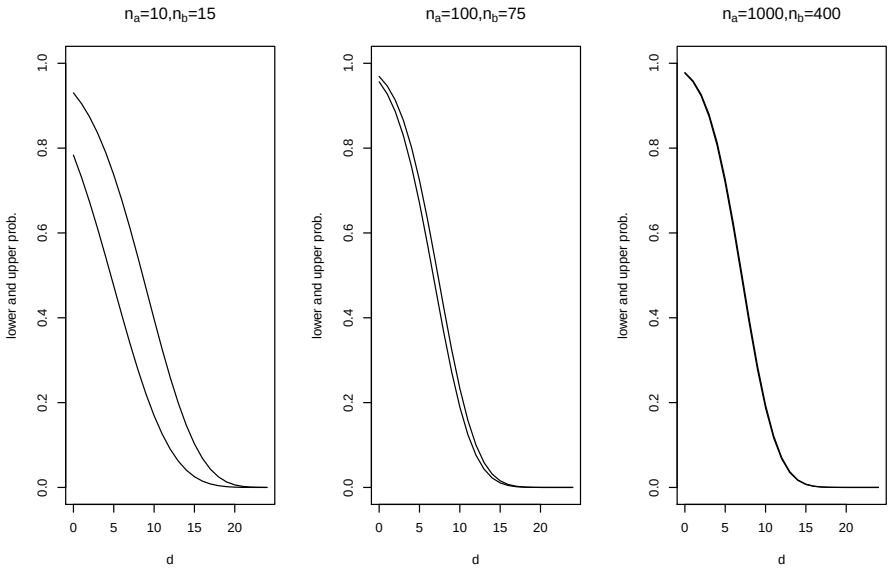

Figure 3: NPI lower and upper probabilities for $X_{a,n_a+1}^{n_a+25} > X_{b,n_b+1}^{n_b+25} + d | X_{a,1}^{n_a} = 0.7n_a, X_{b,1}^{n_b} = 0.4n_b$

Suppose that a diagnostic test has been applied to n_1 individuals in the disease group ($D = 1$) and to n_0 individuals in the non-disease group ($D = 0$). Let $\{Y_i^1; i = 1, \ldots, n_1\}$ and $\{Y_j^0; j = 1, \ldots, n_0\}$ represent the results of the test applied to the individuals from the disease and non-disease groups, respectively. The event that individual i (j) from the disease (non-disease) group has a positive test result is denoted by $Y_i^1 = 1$ ($Y_j^0 = 1$) and a negative test result is denoted by $Y_i^1 = 0$ ($Y_j^0 = 0$). Let $X_{1,1}^{n_1}$ ($X_{0,1}^{n_0}$) be the number of individuals from the disease (non-disease) group with positive test results, so $X_{1,1}^{n_1} = \sum_{i=1}^{n_1} \mathbf{1}\{Y_i^1 = 1\} = n_1^+$ ($X_{0,1}^{n_0} = \sum_{j=1}^{n_0} \mathbf{1}\{Y_j^0 = 1\} = n_0^+$), with indicator function $\mathbf{1}\{A\} = 1$ if A is true and 0 else. Suppose that there are m_1 and m_0 further ('future') individuals from the disease and non-disease groups, respectively, in NPI the inferences are explicitly on these future individuals. Let $X_{1,n_1+1}^{n_1+m_1}$ ($X_{0,n_0+1}^{n_0+m_0}$) be the number of individuals out of these m_1 (m_0) future individuals from the disease (non-disease) group who will have positive test results, so $X_{1,n_1+1}^{n_1+m_1} = \sum_{i=n_1+1}^{n_1+m_1} \mathbf{1}\{Y_i^1 = 1\}$ ($X_{0,n_0+1}^{n_0+m_0} = \sum_{j=n_0+1}^{n_0+m_0} \mathbf{1}\{Y_j^0 = 1\}$). These two random quantities $X_{1,n_1+1}^{n_1+m_1}$ and $X_{0,n_0+1}^{n_0+m_0}$ are logically related to the concepts of TPF and FPF as introduced in Section 2.

Of course, one would prefer the number of individuals from the disease group who have positive test results to be as large as possible and the number of individuals from the non-disease group who have positive test results to be as small as possible. In predictive inference, there are several options for events of interest involving $X_{1,n_1+1}^{n_1+m_1}$ and $X_{0,n_0+1}^{n_0+m_0}$ that can be used to express this general aim. For example, one may want

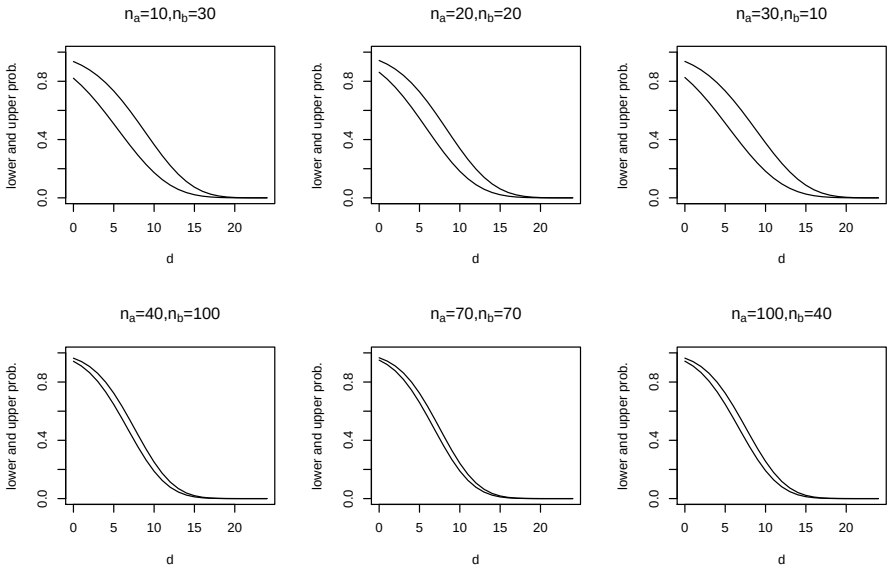

Figure 4: NPI lower and upper probabilities for $X_{a,n_a+1}^{n_a+25} > X_{b,n_b+1}^{n_b+25} + d | X_{a,1}^{n_a} = 0.7n_a, X_{b,1}^{n_b} = 0.4n_b$

the number of future individuals from the disease group who will test positive to exceed a number x to ensure the accuracy of this test, so $X_{1,n_1+1}^{n_1+m_1} \geq x$. Given the data on the first n_1 individuals from the disease group, the NPI lower and upper probabilities for this event can be calculated from (3), (4) and the conjugacy property (2), with $n = n_1$, $m = m_1$, $s = n_1^+$. One may also want the number of future individuals from the non-disease group who will test positive not to exceed a number y, so $X_{0,n_0+1}^{n_0+m_0} < y$. The NPI lower and upper probabilities for this event can also be calculated from (3), (4) and the conjugacy property (2), with $n = n_0$, $m = m_0$, $s = n_0^+$. These NPI lower and upper probabilities will be illustrated via an example in Section 5.

One can also consider a direct comparison of $X_{1,n_1+1}^{n_1+m_1}$ and $X_{0,n_0+1}^{n_0+m_0}$. For example, if one considers equal numbers of future individuals for the disease and non-disease groups, so $m_1 = m_0 = m$, one may be interested in the event that $X_{n_1+1}^{n_1+m} > X_{n_0+1}^{n_0+m}$, so the event that the number of future individuals who will test positive to the disease is larger from the disease group than from the non-disease group given that m individuals are considered from each group. Assuming that the disease and non-disease groups are independent, the NPI upper probability for this event $X_{1,n_1+1}^{n_1+m} > X_{0,n_0+1}^{n_0+m}$, can be calculated from (6) with $n_a = n_1$, $n_b = n_0$, $s_a = n_1^+$ and $s_b = n_0^+$, the corresponding NPI lower probability can be calculated with the appropriate changes to this formula as indicated in Section 3.

We can also consider NPI applied to groups according to test results. Suppose that all individuals who test positive to the disease form the group D^+ and all individuals who test negative to the disease form the group D^-. If individual i (j) in D^+ (D^-)

actually has the disease we denote this by $D_i^+ = 1$ ($D_j^- = 1$) and if this individual does not have the disease it is denoted by $D_i^+ = 0$ ($D_j^- = 0$), where $i = 1, \ldots, n^+$ ($j = 1, \ldots, n^-$). Of course, the ideal situation would be that all individuals who test positive to the disease actually have the disease and all those who test negative actually do not have the disease. Let $Z_1^{n^+}$ be the number of individuals who actually have the disease within the group D^+ of individuals who tested positive, so $Z_1^{n^+} = \sum_{i=1}^{n^+} \mathbf{1}\{D_i^+ = 1\} = n_1^+$. Let $V_1^{n^-}$ be the number of individuals who actually do not have the disease within the group D^- of individuals who tested negative, so $V_1^{n^-} = \sum_{j=1}^{n^-} \mathbf{1}\{D_j^- = 0\} = n_0^-$. Now suppose that there are m^+ and m^- further (future) individuals who test positive and negative to the disease, respectively. Let $Z_{n^++1}^{n^++m^+}$ ($V_{n^-+1}^{n^-+m^-}$) denote the number of future individuals, out of these m^+ (m^-) who test positive (negative), who actually (do not) have the disease, so $Z_{n^++1}^{n^++m^+} = \sum_{i=n^++1}^{n^++m^+} \mathbf{1}\{D_i^+ = 1\}$ ($V_{n^-+1}^{n^-+m^-} = \sum_{j=n^-+1}^{n^-+m^-} \mathbf{1}\{D_j^- = 0\}$). These two random quantities $Z_{n^++1}^{n^++m^+}$ and $V_{n^-+1}^{n^-+m^-}$ are logically related to the concepts of PPV and NPV as introduced in Section 2.

There are also several options in NPI for expressing accuracy of a single diagnostic test in terms of events involving $Z_{n^++1}^{n^++m^+}$ and $V_{n^-+1}^{n^-+m^-}$. For example, one may want both of these quantities to be greater than a particular value, e.g. $Z_{n^++1}^{n^++m^+} \geq x$ and $V_{n^-+1}^{n^-+m^-} \geq y$. The NPI lower and upper probabilities for these events, and for alternative events that one may wish to focus on, can be calculated using the results of Section 3.

To compare two diagnostic tests, one can use the NPI upper probabilities (5) and (6) and the corresponding lower probabilities for such events involving any of the random quantities related to the diagnostic tests as introduced above. This will also be illustrated and discussed in the example in Section 5. For all these comparisons, the NPI upper probabilities (8) and (9) and the corresponding lower probabilities can be studied, as functions of d, to get useful insights into the size of the differences between the random quantities involved.

5 Example

Table 2 provides data introduced by Pepe (2003, p.38) considering a study of two diagnostic tests for fetal abnormality, chorionic villus sampling (CVS) and early amniocentesis (EA). We use these data to illustrate NPI for accuracy of binary diagnostic tests and for the comparison of two such tests, as introduced in this paper.

	$D=0$			$D=1$		
	$Y=0$	$Y=1$		$Y=0$	$Y=1$	
EA	4844	34	4878	6	116	122
CVS	4765	111	4876	13	111	124

Table 2: CVS and EA data

Let us first consider the two tests, EA and CVS, separately for different events of interest as discussed in Section 4. The NPI lower and upper probabilities for eight

		$m = 10$		$m = 100$	
	x	EA	CVS	EA	CVS
E_1	$0.1m$	0.9306, 0.9325	0.7929, 0.7945	1.0000, 1.0000	0.9999, 0.9999
E_2	$0.9m$	0.8867, 0.9114	0.6897, 0.7198	0.9251, 0.9571	0.4655, 0.5433
E_3	$0.9m$	0.2964, 0.3133	0.0117, 0.0126	0.0049, 0.0069	0.0000, 0.0000
E_4	$0.9m$	0.9999, 0.9999	0.9996, 0.9996	1.0000, 1.0000	1.0000, 1.0000
F_1	$0.1m$	0.9977, 0.9978	0.9789, 0.9792	1.0000, 1.0000	1.0000, 1.0000
F_2	$0.9m$	0.5684, 0.6174	0.3183, 0.3470	0.8839, 0.9296	0.3694, 0.4450
F_3	$0.9m$	0.0780, 0.0847	0.0011, 0.0012	0.0023, 0.0034	0.0000, 0.0000
F_4	$0.9m$	0.9857, 0.9877	0.9711, 0.9732	1.0000, 1.0000	1.0000, 1.0000

Table 3: NPI lower and upper probabilities for individual diagnostic tests

different events, each for two values of m, are presented in Table 3. The first row, E_1, considers the event that the number of individuals from m future individuals from the non-disease group who will have positive test results is less than a particular value x that is related to m as indicated in the second column. So the reported values for E_1 are

$$\underline{P}, \overline{P}(X^{n_0+m}_{\text{EA},n_0+1} < x | X^{n_0}_{\text{EA},1} = n^+_{\text{EA},0}) \text{ and } \underline{P}, \overline{P}(X^{n_0+m}_{\text{CVS},n_0+1} < x | X^{n_0}_{\text{CVS},1} = n^+_{\text{CVS},0})$$

where we have introduced the notation $\underline{P}, \overline{P}(\cdot)$ to indicate that the corresponding NPI lower and upper probabilities are reported in Table 3 as a pair of values separated by a comma. The second row, E_2, considers the event that the number of individuals from m future individuals from the disease group who will have positive test results is greater than or equal to x, so the reported values for E_2 are

$$\underline{P}, \overline{P}(X^{n_1+m}_{\text{EA},n_1+1} \geq x | X^{n_1}_{\text{EA},1} = n^+_{\text{EA},1}) \text{ and } \underline{P}, \overline{P}(X^{n_1+m}_{\text{CVS},n_1+1} \geq x | X^{n_1}_{\text{CVS},1} = n^+_{\text{CVS},1})$$

The third row, E_3, considers the event that the number of individuals from m future individuals who will have the disease given that they get positive test results is greater than or equal to x, so the reported values for E_3 are

$$\underline{P}, \overline{P}(Z^{n^+ + m}_{\text{EA},n^+ + 1} \geq x | Z^{n^+}_{\text{EA},1} = n^+_{\text{EA},1}) \text{ and } \underline{P}, \overline{P}(Z^{n^+ + m}_{\text{CVS},n^+ + 1} \geq x | Z^{n^+}_{\text{CVS},1} = n^+_{\text{CVS},1})$$

The fourth row, E_4, considers the event that the number of individuals from m future individuals who do not have the disease given that they get negative test results is greater than or equal to x, so the reported values for E_4 are

$$\underline{P}, \overline{P}(V^{n^- + m}_{\text{EA},n^- + 1} \geq x | V^{n^-}_{\text{EA},1} = n^-_{\text{EA},0}) \text{ and } \underline{P}, \overline{P}(V^{n^- + m}_{\text{CVS},n^- + 1} \geq x | V^{n^-}_{\text{CVS},1} = n^-_{\text{CVS},0})$$

Table 3 shows that these NPI lower and upper probabilities for EA are all larger than the corresponding ones for CVS, but for $m = 100$ there is no substantial difference between the values for EA and CVS except for event E_2.

Rows 5-8 of Table 3 give the NPI lower and upper probabilities for events that correspond to the events in $E_1 - E_4$ in rows 1-4. In row 5, F_1 corresponds to E_1 but with the $<$ replaced by $\leq$ in the event of interest, so the reported values for F_1 are

$$\underline{P}, \overline{P}(X^{n_0+m}_{\text{EA},n_0+1} \leq x | X^{n_0}_{\text{EA},1} = n^+_{\text{EA},0}) \text{ and } \underline{P}, \overline{P}(X^{n_0+m}_{\text{CVS},n_0+1} \leq x | X^{n_0}_{\text{CVS},1} = n^+_{\text{CVS},0})$$

The sixth row, F_2, corresponds to E_2 with $\geq$ replaced by $>$, so the reported values for F_2 are

$$\underline{P}, \overline{P}(X_{\text{EA},n_1+1}^{n_1+m} > x | X_{\text{EA},1}^{n_1} = n_{\text{EA},1}^{+}) \text{ and } \underline{P}, \overline{P}(X_{\text{CVS},n_1+1}^{n_1+m} > x | X_{\text{CVS},1}^{n_1} = n_{\text{CVS},1}^{+})$$

The seventh row, F_3, similarly corresponds to E_3 and the reported values are

$$\underline{P}, \overline{P}(Z_{\text{EA},n^+ +1}^{n^+ +m} > x | Z_{\text{EA},1}^{n^+} = n_{\text{EA},1}^{+}) \text{ and } \underline{P}, \overline{P}(Z_{\text{CVS},n^+ +1}^{n^+ +m} > x | Z_{\text{CVS},1}^{n^+} = n_{\text{CVS},1}^{+})$$

The final row, F_4, similarly corresponds to E_4 and the reported values are

$$\underline{P}, \overline{P}(V_{\text{EA},n^- +1}^{n^- +m} > x | V_{\text{EA},1}^{n^-} = n_{\text{EA},0}^{-}) \text{ and } \underline{P}, \overline{P}(V_{\text{CVS},n^- +1}^{n^- +m} > x | V_{\text{CVS},1}^{n^-} = n_{\text{CVS},0}^{-})$$

Again, these values are larger for EA than for CVS, and the values in the F-rows can vary substantially from those in the corresponding E-rows due to the discrete nature of the random quantities considered, which clearly has a stronger effect for $m = 10$ than for $m = 100$. This emphasizes that care must be taken on defining precisely the event of interest in such studies. Clearly, both these tests are quite good, with only relatively few errors, which also shows in the values in Table 3. Particularly for $m = 100$, several of the entries make clear that at least 90 or more of 100 tests will give the correct results with very high lower and upper probabilities (the values 1.0000 in the table are rounded upwards, none of them are precisely 1), in these cases the corresponding NPI lower and upper probabilities for $m = 10$ are of interest, as they are not always very close to 1. Also the difference between the NPI lower and upper probabilities for corresponding events such as E_1 and F_1 is substantially larger for $m = 10$ than for $m = 100$, this is due to the fact that outcomes being equal to a single value (the '$=$' which makes the difference between such events) are of course more likely for $m = 10$ than for $m = 100$.

For event E_2, the NPI lower and upper probabilities for CVS in Table 3 are perhaps somewhat surprising, as they are smaller for $m = 100$ than for $m = 10$. For this event, CVS gave positive test result to 111 out of 124 individuals in the disease group, which is a proportion of 0.895. For $m = 10$, the corresponding lower and upper probabilities for the event that out of $m = 10$ future individuals with the disease at least 9 test positive are quite high (0.6897 and 0.7198), but for $m = 100$ the corresponding lower and upper probabilities for the event that at least 90 test positive are smaller (0.4655 and 0.5433). For the corresponding event F_2, these lower and upper probabilities are quite a bit smaller, reflecting that the actual outcomes of 9 positive CVS test results for 10 individuals with the disease, and 90 out of 100, are actually quite likely to occur. This is a consequence of the observed proportion being so close to the future proportion 0.9 which is of interest in this event.

All events E_1 to F_4 are formulated such that, for an ideal situation, the NPI lower and upper probabilities presented in Table 3 would be large. Unfortunately, this is not the case for the events E_3 and F_3. Of course, this reflects the data; for EA there were 150 individuals with a positive test result, of whom 116 indeed had the disease, and for CVS there were 222 individuals with a positive test result, of whom 111 indeed had the disease. Both these proportions are well below 0.9, leading to very small lower and upper probabilities for the events that at least 90% of future individuals with a positive test result do have the disease.

It is worth to emphasize that these NPI lower and upper probabilities always bound the empirical probability for the event considered based on the data. For example, the empirical probability corresponding to the event in row F_2 for EA with $m = 10$ is equal to $(116/122)^{10} = 0.6093$, which is within the interval $[0.5684, 0.6174]$. Table 3 only provides a small insight into the many events that can be studied using the NPI methods presented in this paper, as one can, and arguably should, vary the values of x and m to get a detailed perspective on the uncertainties involved. The choice of m is further discussed in Section 6.

A classical method to compare the two diagnostic tests EA and CVS is by using the relative probabilities as described in Section 2 (Pepe, 2003), the values of which are

$$r\text{TPF(EA,CVS)} = \frac{116/122}{111/124} = 1.062\ , \quad r\text{FPF(EA,CVS)} = \frac{34/4878}{111/4876} = 0.306$$

$$r\text{PPV(EA,CVS)} = \frac{116/150}{111/222} = 1.547\ , \quad r\text{NPV(EA,CVS)} = \frac{4844/4850}{4765/4778} = 1.002$$

These all indicate that EA is a better diagnostic test than CVS, but they do not clearly reflect the difference of the performances of these tests on further patients. We now illustrate the use of the NPI method presented in this paper for comparison of these two diagnostic tests. We do this by comparing four events of interest based on both tests, as described below. The NPI lower and upper probabilities for these events are presented in Table 4, these are based on the results presented in Sections 3 and 4.

The first row in Table 4, G_1, compares the proportions of individuals from m future individuals who will have positive test results if they are from the non-disease group, the reported values are

$$\underline{P}, \overline{P}(X_{\text{EA},n_0+1}^{n_0+m} > (\geq) X_{\text{CVS},n_0+1}^{n_0+m} | X_{\text{EA},1}^{n_0} = n_{\text{EA},0}^{+}, X_{\text{CVS},1}^{n_0} = n_{\text{CVS},0}^{+})$$

The second row, G_2, compares the proportions of individuals from m future individuals who will have positive test results if they are from the disease group, the reported values are

$$\underline{P}, \overline{P}(X_{\text{EA},n_1+1}^{n_1+m} > (\geq) X_{\text{CVS},n_1+1}^{n_1+m} | X_{\text{EA},1}^{n_1} = n_{\text{EA},1}^{+}, X_{\text{CVS},1}^{n_1} = n_{\text{CVS},1}^{+})$$

The third row, G_3, compares the proportions of individuals from m future individuals who will have the disease given that they get positive test results, the reported values are

$$\underline{P}, \overline{P}(Z_{\text{EA},n^+ +1}^{n^+ +m} > (\geq) Z_{\text{CVS},n^+ +1}^{n^+ +m} | Z_{\text{EA},1}^{n^+} = n_{\text{EA},1}^{+}, Z_{\text{CVS},1}^{n^+} = n_{\text{CVS},1}^{+})$$

The final row, G_4, compares the proportions of individuals from m future individuals who do not have the disease given that they get negative test results, the reported values are

$$\underline{P}, \overline{P}(V_{\text{EA},n^- +1}^{n^- +m} > (\geq) V_{\text{CVS},n^- +1}^{n^- +m} | V_{\text{EA},1}^{n^-} = n_{\text{EA},0}^{-}, V_{\text{CVS},1}^{n^-} = n_{\text{CVS},0}^{-})$$

The values in Table 4 also illustrate the conjugacy property for these NPI lower and upper probabilities, that is $\underline{P}(A) = 1 - \overline{P}(A^c)$. Of course, we would like all entries for G_2, G_3 and G_4 to be larger, and those for G_1 to be smaller, for EA$>$ $(\geq)$CVS than the corresponding values for CVS$>$ $(\geq)$EA, to indicate that EA is the best test from all perspectives considered, as was also suggested by the relative probabilities.

	$\underline{P}, \overline{P}$(EA > CVS)			$\underline{P}, \overline{P}$(EA $\geq$ CVS)		
	$m = 1$	$m = 10$	$m = 100$	$m = 1$	$m = 10$	$m = 100$
G_1	0.0068, 0.0070	0.0539, 0.0556	0.0966, 0.1016	0.9772, 0.9774	0.8055, 0.8074	0.2648, 0.2731
G_2	0.0981, 0.1065	0.4685, 0.5198	0.7984, 0.8785	0.9490, 0.9567	0.8037, 0.8421	0.8504, 0.9149
G_3	0.3824, 0.3892	0.8464, 0.8591	0.9992, 0.9995	0.8836, 0.8879	0.9303, 0.9375	0.9994, 0.9997
G_4	0.0027, 0.0029	0.0265, 0.0285	0.2088, 0.2269	0.9986, 0.9988	0.9861, 0.9881	0.8961, 0.9120
	$\underline{P}, \overline{P}$(CVS > EA)			$\underline{P}, \overline{P}$(CVS $\geq$ EA)		
	$m = 1$	$m = 10$	$m = 100$	$m = 1$	$m = 10$	$m = 100$
G_1	0.0226, 0.0228	0.1926, 0.1945	0.7269, 0.7352	0.9930, 0.9932	0.9444, 0.9461	0.8984, 0.9034
G_2	0.0433, 0.0510	0.1579, 0.1963	0.0851, 0.1496	0.8935, 0.9019	0.4802, 0.5315	0.1215, 0.2016
G_3	0.1121, 0.1164	0.0625, 0.0697	0.0003, 0.0006	0.6108, 0.6176	0.1409, 0.1536	0.0005, 0.0008
G_4	0.0012, 0.0014	0.0119, 0.0139	0.0880, 0.1039	0.9971, 0.9973	0.9715, 0.9735	0.7731, 0.7912

Table 4: NPI lower and upper probabilities for comparison of EA and CVS

This is indeed the case, as is easily seen in Table 4. These values show again that the differences between the two tests become clearer for larger values of m, which is both because equal outcomes become less likely and because the effect of the randomness, which can cause a worse test to perform better on a few patients, disappears when larger numbers of further patients are considered. It should be remarked that this study had quite substantial numbers of data, if these had been smaller there would have been substantially more imprecision in these predictive inferences.

As introduced in Section 4, one possibility to get a better feeling for the actual difference between two random quantities of the type considered in this paper is by considering the NPI lower and upper probabilities for the event that they differ by at least d. This is illustrated in Figure 5 for the events G_2 and G_3 (for $m = 10$ and 100 and considering only the events with $>$), so these events are now generalized to

$$\underline{P}, \overline{P}(X_{\text{EA},n_1+1}^{n_1+m} > X_{\text{CVS},n_1+1}^{n_1+m} + d | X_{\text{EA},1}^{n_1} = n_{\text{EA},1}^{+}, X_{\text{CVS},1}^{n_1} = n_{\text{CVS},1}^{+})$$

and

$$\underline{P}, \overline{P}(Z_{\text{EA},n^+ +1}^{n^+ +m} > Z_{\text{CVS},n^+ +1}^{n^+ +m} + d | Z_{\text{EA},1}^{n^+} = n_{\text{EA},1}^{+}, Z_{\text{CVS},1}^{n^+} = n_{\text{CVS},1}^{+})$$

These plots provide a more detailed insight into the size of the difference between the performances of these two diagnostic tests on further patients. For example, the bottom-right plot for the case G_3 with $m = 100$ shows that, for 100 further patients who test positive using test EA and for 100 other further patients who test positive using test CVS, it is likely (NPI lower probability close to 0.8) that at least 20 more of these patients actually have the disease for test EA than for test CVS, while this number is unlikely to exceed 40 (NPI upper probability well below 0.1). This actually provides a detailed insight into the strength of the evidence in the data, and can provide important input into decisions about choice of diagnostic test.

6 Concluding remarks

In this paper, NPI for accuracy of binary diagnostic tests has been presented. It provides an attractive alternative to classical methods for such problems, in particular when used together these can provide valuable insights into the actual quality of a single test,

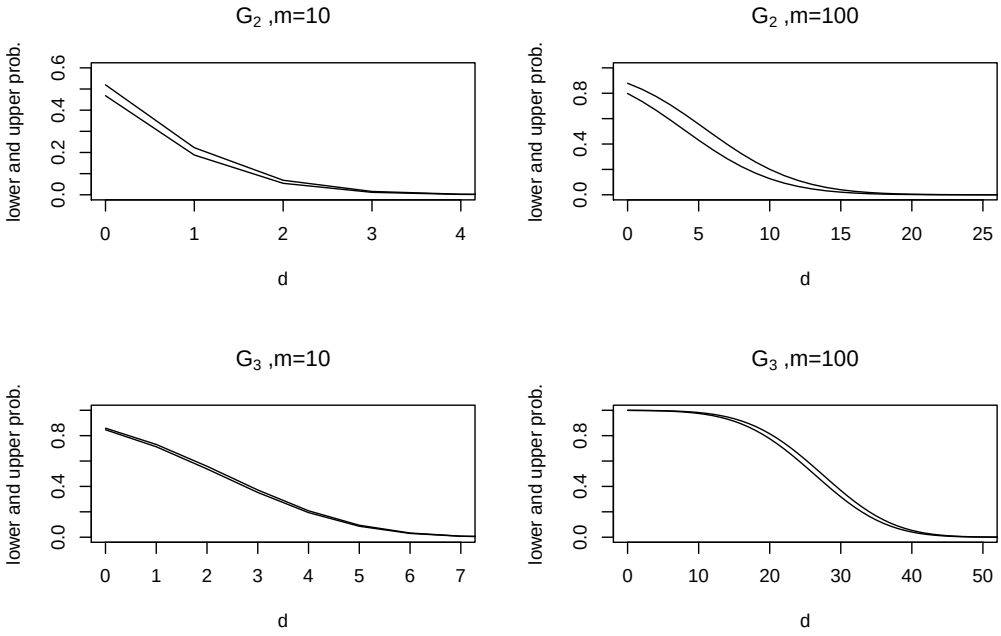

Figure 5: NPI lower and upper probabilities for the events G_2 and G_3 generalized to EA $>$ CVS $+ d$

and the difference in quality of two tests, with regard to application to future patients. NPI does not aim at inference for an entire population but instead explicitly considers future observations, which is particularly suitable for inference to support decisions on medical diagnosis for one future patient, or for a pre-determined number of future patients, so the NPI approach provides an attractive alternative to standard methods for such inferences. It has been shown how different criteria can be considered, which are in line with the well-known criteria used in the classical methods. If one must make a decision for a known number m of future patients, the predictive inference directly in terms of m patients is clearly attractive. More generally, by studying the NPI results for varying values of m one gets deeper insight into the variability and the strength of evidence in the data, where also the use of the difference d can provide useful information. We do not recommend specific values for m and d to be used, it is particularly the opportunity to study these inferences over a range of these values that will give insights which can contribute to well informed decision making. We have restricted attention to comparison of two diagnostic tests. This can easily be generalized to comparison of more such tests, using the NPI approach for multiple comparisons of groups of Bernoulli data (Coolen and Coolen-Schrijner, 2006, 2007).

The NPI approach has also been developed for other data types (Coolen, 2011), including real-valued and ordinal data. Diagnostic accuracy has recently also been presented within the NPI framework for real-valued data (Coolen-Maturi, et al., 2012a)

and for ordinal data (Elkhafifi and Coolen, 2012), but for both those scenarios only inferences involving a single future patient from each of the disease and non-disease groups were considered. Extending those approaches to multiple future patients is an interesting research challenge.

Acknowledgements

We gratefully acknowledge comments and suggestions by two reviewers of this paper, which led to improved presentation.

References

Augustin T., Coolen F.P.A., 2004. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124, 251–272.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., Rothstein, H.R., 2009. *Introduction to Meta-Analysis*. Wiley, Chichester.

Coolen F.P.A., 1998. Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters*, 36, 349–357.

Coolen F.P.A., 2006. On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, 15, 21–47.

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

Coolen, F.P.A., Coolen-Schrijner, P., 2006. Nonparametric predictive subset selection for proportions. *Statistics & Probability Letters*, 76, 1675–1684.

Coolen, F.P.A., Coolen-Schrijner, P., 2007. Nonparametric predictive comparison of proportions. *Journal of Statistical Planning and Inference*, 137, 23–33.

Coolen, F.P.A., Elsaeiti, M.A., 2009. Nonparametric predictive methods for acceptance sampling. *Journal of Statistical Theory and Practice*, 3, 907–921.

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

Coolen-Schrijner, P., Coolen, F.P.A., MacPhee, I.M., 2008. Nonparametric predictive inference for system reliability with redundancy allocation. *Journal of Risk and Reliability*, 222, 463–476.

Coolen-Schrijner P., Maturi T.A., Coolen F.P.A., 2009. Nonparametric predictive precedence testing for two groups. *Journal of Statistical Theory and Practice*, 3, 273–287.

Dodd, L.E., Pepe, M.S., 2003. Partial AUC estimation and regression. *Biometrics*, 59, 614–623.

Elkhafifi, F.F.G.A., Coolen, F.P.A., 2012. Nonparametric predictive inference for accuracy of ordinal diagnostic tests. *Journal of Statistical Theory and Practice*, to appear.

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

Hill B.M., 1988. De Finetti's theorem, induction, and A_n, or Bayesian nonparametric predictive inference (with discussion). In: D.V. Lindley, J.M. Bernardo, M.H. DeGroot, A.F.M. Smith (eds.), *Bayesian Statistics 3*. Oxford University Press, pp. 211–241.

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

Weinstein, S., Obuchowski, N.A., Lieber, M.L., 2005. Clinical evaluation of diagnostic tests. *American Journal of Roentgenology*, 184, 14–19.