Nonparametric Predictive Inference for Diagnostic Accuracy

Tahani Coolen-Maturi, Pauline Coolen-Schrijner[1], Frank P.A. Coolen*

Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, UK

Abstract

Measuring the accuracy of diagnostic tests is crucial in many application areas including medicine and health care. 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) methods for accuracy of diagnostic tests with continuous test results are presented and discussed. For such tests, Receiver Operating Characteristic (ROC) curves have become popular tools for describing the performance of diagnostic tests. We present the NPI approach to ROC curves, and some important summaries of these curves. As NPI does not aim at inference for an entire population but instead explicitly considers a future observation, this provides an attractive alternative to standard methods. We show how NPI can be used to compare two continuous diagnostic tests.

Keywords: area under the ROC curve (AUC), diagnostic accuracy, lower and upper probability, nonparametric predictive inference (NPI), partial area under the ROC curve (pAUC), receiver operating characteristic (ROC)

*Corresponding author.

Email addresses: tahani.maturi@durham.ac.uk (Tahani Coolen-Maturi), frank.coolen@durham.ac.uk (Frank P.A. Coolen)

[1]Pauline died on 23.04.2008, when the research reported in this paper was at an advanced stage.

1. Introduction

The evaluation of the accuracy of diagnostic tests has particular importance in medicine and health care. Diagnostic test results may have only two values (binary test), or a value in a finite number of ordered categories (ordinal test), or real values (continuous test). There are several accuracy measures which vary depending on the type of diagnostic test results. For example, the Receiver Operating Characteristic (ROC) curve is a common statistical tool for describing the performance of certain medical tests. It is used to measure the accuracy of a diagnostic test that yields ordinal or continuous results. The ROC curve plays an important role in many areas such as signal detection, radiology, machine learning, data mining and credit scoring.

In this paper we introduce Nonparametric Predictive Inference (NPI) for diagnostic accuracy for continuous test results. The case of ordinal test results requires further development of NPI for ordinal data and is is left for future research. For accuracy of binary tests we are investigating an approach using NPI for Bernoulli data (Coolen, 1998), we hope to report on this in the near future. NPI is a statistical method based on Hill's assumption $A_{(n)}$ (Hill, 1968), which gives direct probabilities for a future observable random quantity, given observed values of related random quantities (Augustin and Coolen, 2004; Coolen, 2006). During the last decade, NPI has been developed for different applications in statistics, operational research, and reliability and risk analysis.

Section 2 gives a brief overview of NPI, and Section 3 gives an introduction to the concepts of continuous diagnostic tests used in this paper. NPI for such tests is introduced in Section 4. This includes NPI for ROC curves, the area under the ROC curve, and the partial area under the ROC curve, which are commonly used summaries of the ROC curve. Comparison of diagnostic tests is important with regard to guidance on most useful test methods, the NPI approach to such comparison of continuous diagnostic tests is discussed, and illustrated via an example, in Section 5. The paper is finished with some concluding remarks in Section 6.

2. Nonparametric Predictive Inference (NPI)

Nonparametric Predictive Inference (NPI) is a statistical method based on Hill's assumption $A_{(n)}$ (Hill, 1968), which gives direct probabilities for a

future observable random quantity, given observed values of related random quantities (Augustin and Coolen, 2004; Coolen, 2006). To introduce $A_{(n)}$ we use the following notation. Suppose that $X_1, \ldots, X_n$, X_{n+1} are real-valued absolutely continuous and exchangeable random quantities. Let the ordered observed values of $X_1, \ldots, X_n$ be denoted by $x_1 < x_2 < \ldots < x_n$, and let $x_0 = -\infty$ and $x_{n+1} = \infty$ for ease of notation. For ease of presentation, throughout this paper we assume that no ties occur, it is easy to generalize the method to allow ties by breaking the ties in all possible ways and deriving the overall NPI lower and upper probabilities as the minimum and maximum, respectively, of the lower and upper probabilities corresponding to each way of breaking the ties (Maturi, 2010). For X_{n+1}, representing a future observation, based on n observations, $A_{(n)}$ (Hill, 1968) is

$$P(X_{n+1} \in (x_j, x_{j+1})) = \frac{1}{n+1} \quad , \quad j = 0, 1, \ldots, n \tag{1}$$

$A_{(n)}$ does not assume anything else, and can be considered to be a post-data assumption related to finite exchangeability (De Finetti, 1974). Hill (1988) discusses $A_{(n)}$ in detail, including its justification from several foundational perspectives. Inferences based on $A_{(n)}$ are predictive and nonparametric, and can be considered suitable if there is hardly any knowledge about the random quantity of interest, other than the n observations, or if one does not want to use such information, e.g. to study effects of additional assumptions underlying other statistical methods. We consider NPI as a frequentist statistical framework, which has the important advantages of being exactly calibrated (Lawless and Fredette, 2005) and not requiring the use of counterfactuals nor relying on asymptotic justifications as is the use for many commonly used frequentist statistics procedures. $A_{(n)}$ is not sufficient to derive precise probabilities for many events of interest, but it provides bounds for probabilities via the 'fundamental theorem of probability' (De Finetti, 1974), which are lower and upper probabilities in interval probability theory (Walley, 1991; Weichselberger, 2001). For more details on NPI, including references to applications in statistics and related topics, see *www.npi-statistics.com*.

3. Continuous diagnostic tests

In this section, some common measures of diagnostic accuracy are briefly reviewed, 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 a continuous random quantity of a diagnostic test result, and that large values of Y are considered more indicative of disease. Using a threshold c, the test result is called positive if $Y > c$, so if it indicates the disease, and negative if $Y \leq c$, where $c \in (-\infty, \infty)$.

The Sensitivity (SN) of a test is the probability of a positive test result for an individual with the condition (disease), and is also known as True Positive Fraction (TPF). The Specificity (SP) is the probability of a negative test result for an individual without the condition (non-disease). Clearly, an accurate diagnostic test will have sensitivity and specificity both close to one. The False Positive Fraction (FPF) is the probability of a positive test result for an individual without the condition, hence SP= 1−FPF.

Throughout this paper, Y^1 and Y^0 are used to refer to test results for the disease and non-disease groups, respectively, and n_1 and n_0 are the corresponding numbers of individuals in the disease and the non-disease groups. With threshold $c \in (-\infty, \infty)$, $(\text{FPF}(c), \text{TPF}(c))$, FPF and TPF are the main probabilities considered in this paper for continuous diagnostic tests, with

$$\text{FPF}(c) = P(Y^0 > c|D = 0) = S_0(c) \quad (2)$$

$$\text{TPF}(c) = P(Y^1 > c|D = 1) = S_1(c) \quad (3)$$

where $S_0(c)$ and $S_1(c)$ are the survival functions for the random quantities Y^0 and Y^1 for the diagnostic test results for the non-disease and disease groups, respectively.

3.1. Receiver Operating Characteristic (ROC) curve

The Receiver Operating Characteristic (ROC) curve is a common statistical tool for describing the performance of diagnostic tests which yield ordinal or continuous results. For continuous test results, the ROC curve is defined as the combination of FPF and TPF over all values of the threshold c, i.e.

$$\text{ROC} = \{(\text{FPF}(c), \text{TPF}(c)), c \in (-\infty, \infty)\} \quad (4)$$

The ROC curve can also be written as

$$\text{ROC}(t) = S_1\left(S_0^{-1}(t)\right) \ , \ t \in (0, 1) \quad (5)$$

where $\text{FPF}(c) = S_0(c) = t$ (Pepe, 2003).

Geometrically, an ROC curve can be obtained by plotting FPF versus TPF, which gives a clear indication of the performance of the test. An

ideal test completely separates the patients with and without the disease for a threshold c, i.e. $\text{FPF}(c) = 0$ and $\text{TPF}(c) = 1$. As the other extreme situation, if $\text{FPF}(c) = \text{TPF}(c)$ for all thresholds c, then the test has no ability to distinguish between the patients with and without the disease.

The ROC curve depends on the distributions of Y^1 and Y^0, however these distributions are usually unknown. To estimate the ROC curve for diagnostic tests with continuous results, the nonparametric empirical method is popular due to its flexibility to adapt fully to the available data. This method yields the empirical ROC curve which we will consider later in this paper, in particular to compare with the NPI method introduced in this paper. Methods using assumed parametric distributions for both Y^1 and Y^0, together with methods for estimation of the parameters, are of course also used, but are less popular unless one actually has substantial information beyond the data to justify the model assumptions. More details on these methods can be found in Pepe (2003). In Section 4 we present NPI as an alternative approach for inference about the ROC curve. The NPI method is different from the nonparametric empirical method as it is explicitly predictive, considering a single next observations given the past observations instead of aiming at inference on an entire assumed underlying population.

Suppose that we have test data on n_1 individuals from a disease group and n_0 individuals from a non-disease group, denoted by $\{y_i^1, i = 1, \ldots, n_1\}$ and $\{y_j^0, j = 1, \ldots, n_0\}$, respectively. Throughtout this paper we assume that the two groups are fully independent, meaning that no information about any aspect related to one group contains information about any aspect of the other group. For the empirical method, these observations per group are assumed to be realisations of random quantities that are identically distributed as Y^1 and Y^0, for the disease and non-disease groups, with corresponding survival functions $S_1(y) = P[Y^1 > y]$ and $S_0(y) = P[Y^0 > y]$. The empirical estimator of the ROC is (Pepe, 2003)

$$\widehat{\text{ROC}} = \left\{ \left(\widehat{\text{FPF}}(c), \widehat{\text{TPF}}(c) \right), c \in (-\infty, \infty) \right\} \tag{6}$$

with

$$\widehat{\text{TPF}}(c) = \hat{S}_1(c) = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{1}\left\{ y_i^1 > c \right\} \tag{7}$$

$$\widehat{\text{FPF}}(c) = \hat{S}_0(c) = \frac{1}{n_0} \sum_{j=1}^{n_0} \mathbf{1}\left\{ y_j^0 > c \right\} \tag{8}$$

where $\mathbf{1}\{A\}$ is the indicator function which is equal to 1 if A is true and 0 else, and where $\hat{S}_1$ and $\hat{S}_0$ are the empirical survival functions for Y^1 and Y^0, respectively. The empirical estimator of the ROC can also be written as

$$\widehat{\mathrm{ROC}}(t) = \hat{S}_1\left(\hat{S}_0^{-1}(t)\right)$$

3.2. The area under the ROC curve (AUC)

In many cases, a single numerical value or summary may be useful to represent the accuracy of a diagnostic test or to compare two or more ROC curves (Pepe, 2003). A useful summary is the area under the ROC curve, AUC, which is

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(t)\, dt \tag{9}$$

The AUC measures the overall performance of the diagnostic test. Higher AUC values indicate more accurate tests, with AUC = 1 for perfect or ideal tests and AUC = 0.5 for uninformative tests. The AUC is equal to the probability that the test results from a randomly selected pair of diseased and non-diseased subjects are correctly ordered (Zhou et al., 2002), i.e.

$$\mathrm{AUC} = P\left[Y^1 > Y^0\right] \tag{10}$$

So the AUC represents the test's ability to correctly classify a randomly selected individual as being from either the disease group or the non-disease group. Equation (10) will be used in Section 4.2 to introduce NPI for the area under the ROC curve. The empirical estimator of the AUC is the well-known Mann-Whitney U statistic (Pepe, 2003), which is given by

$$\widehat{\mathrm{AUC}} = \frac{1}{n_1 n_0} \sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \psi(y_i^1, y_j^0) \tag{11}$$

where

$$\psi(y_i^1, y_j^0) = \begin{cases} 1 & \text{if} \quad y_i^1 > y_j^0 \\ \frac{1}{2} & \text{if} \quad y_i^1 = y_j^0 \\ 0 & \text{if} \quad y_i^1 < y_j^0 \end{cases} \tag{12}$$

3.3. The partial area under the ROC curve (pAUC)

In applications a high FPF value may typically lead to high costs and not be realistic, in which case the area under the ROC curve corresponding

only to small values of FPF is relevant. Generally, it may be of interest to use as summary of the ROC curve the area under it between two values of FPF, say t_0 and t_1. This is known as the partial area under the ROC curve, pAUC, and is

$$\text{pAUC}(t_0, t_1) = \text{pAUC}(t_0 \leq \text{FPF} \leq t_1) = \int_{t_0}^{t_1} \text{ROC}(t)\, dt \tag{13}$$

For perfect tests the $\text{pAUC}(t_0, t_1)$ is equal to $(t_1 - t_0)$, for uninformative tests it is equal to $(t_1 + t_0)(t_1 - t_0)/2$. The pAUC can also be written as (Dodd and Pepe, 2003),

$$\text{pAUC}(t_0, t_1) = P\left[Y^1 > Y^0, Y^0 \in (S_0^{-1}(t_1), S_0^{-1}(t_0))\right] \tag{14}$$

The value $\text{pAUC}(t_0, t_1)/(t_1 - t_0)$ is the probability of correctly ordering a disease and a non-disease observation chosen at random, given that the non-disease observation is in the range between the $1-t_1$ and $1-t_0$ quantiles of the non-disease distribution (Pepe, 2003). In other words, $\text{pAUC}(t_0 \leq \text{FPF} \leq t_1)$ is the probability that a randomly chosen patient with the disease will be correctly distinguished from a randomly chosen patient without the disease, in case the latter patient tested positive in a diagnostic test with $\text{FPF} \in [t_0, t_1]$ (Zhou et al., 2002). Equation (14) will be used in Section 4.3 to introduce NPI for the partial area under the ROC curve.

The empirical estimator of pAUC (Dodd and Pepe, 2003) is given by

$$\widehat{\text{pAUC}}(t_0, t_1) = \frac{1}{n_1 n_0} \sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \psi^*(y_i^1, y_j^0) \tag{15}$$

where

$$\psi^*(y_i^1, y_j^0) = \begin{cases} \psi(y_i^1, y_j^0) & \text{if } y_j^0 \in (\hat{S}_0^{-1}(t_1), \hat{S}_0^{-1}(t_0)) \\ 0 & \text{else} \end{cases} \tag{16}$$

with $\psi(y_i^1, y_j^0)$ as given by (12).

Alternatively, one can also consider the partial area under the ROC curve corresponding to an interval of values of TPF, for example if one specifically wishes to have a test with high ability to detect patients with the disease. In such a case, one needs to calculate $\text{pAUC}(q_0 \leq \text{TPF} \leq q_1)$, for suitable values q_0 and q_1 between 0 and 1. We do not discuss this, the theory and development of NPI are very similar to the case with FPF values restricted. For more details we refer to Dodd and Pepe (2003).

4. NPI for continuous diagnostic tests

In this section we present NPI for ROC curves, and for AUC and pAUC. In NPI the uncertainty is quantified by lower and upper probabilities for events of interest. In effect, the optimal lower and upper bounds for the ROC, AUC and pAUC are derived, corresponding to the assumptions $A_{(n_1)}$ for the disease group and $A_{(n_0)}$ for the non-disease group, where the inferences are explicitly predictive, so involving one further patient from each group.

4.1. NPI lower and upper ROC curves

To present NPI for ROC curve and their summaries, the same notation will be used as in Section 3, with the following additions. Suppose that $\{Y_i^1,\ i = 1, \ldots, n_1, n_1 + 1\}$ are continuous and exchangeable random quantities from the disease group and $\{Y_j^0,\ j = 1, \ldots, n_0, n_0 + 1\}$ are continuous and exchangeable random quantities from the non-disease group, where $Y_{n_1+1}^1$ and $Y_{n_0+1}^0$ are the next observations from the disease and non-disease groups following n_1 and n_0 observations, respectively. As mentioned before, we assume that both groups are fully independent. Let $y_1^1 < \ldots < y_{n_1}^1$ be the ordered observed values for the first n_1 individuals from the disease group and $y_1^0 < \ldots < y_{n_0}^0$ the ordered observed values for the first n_0 individuals from the non-disease group. For ease of notation, let $y_0^1 = y_0^0 = -\infty$ and $y_{n_1+1}^1 = y_{n_0+1}^0 = \infty$. We assume that there are no ties in the data.

Coolen et al. (2002) introduced NPI for some reliability applications when the data represent failure times or general event times, which are non-negative. For example, they introduced the NPI lower and upper survival functions for the next observation based on n observations. This can be easily extended for data that may contain negative values. The NPI lower and upper survival functions for $Y_{n_1+1}^1$ and $Y_{n_0+1}^0$ are

$$\underline{S}_1(c) = \underline{P}(Y_{n_1+1}^1 > c) = \frac{\sum_{i=1}^{n_1} \mathbf{1}\left\{y_i^1 > c\right\}}{n_1 + 1} \tag{17}$$

$$\overline{S}_1(c) = \overline{P}(Y_{n_1+1}^1 > c) = \frac{\sum_{i=1}^{n_1} \mathbf{1}\left\{y_i^1 > c\right\} + 1}{n_1 + 1} \tag{18}$$

$$\underline{S}_0(c) = \underline{P}(Y_{n_0+1}^0 > c) = \frac{\sum_{j=1}^{n_0} \mathbf{1}\left\{y_j^0 > c\right\}}{n_0 + 1} \tag{19}$$

$$\overline{S}_0(c) = \overline{P}(Y_{n_0+1}^0 > c) = \frac{\sum_{j=1}^{n_0} \mathbf{1}\left\{y_j^0 > c\right\} + 1}{n_0 + 1} \tag{20}$$

where $\underline{P}$ and $\overline{P}$ are NPI lower and upper probabilities (Augustin and Coolen, 2004).

We will use these NPI lower and upper survival functions to predict the FPF and TPF for the next future individual per group, for different threshold values c, and combine these to derive the corresponding NPI lower and upper ROC curves. The NPI lower and upper survival functions are optimal bounds for all survival functions corresponding to $A_{(n)}$ (Coolen et al., 2002), so they immediately lead to optimal bounds for the TPF and FPF, using (2) and (3). We introduce

$$\underline{\mathrm{TPF}}(c) = \underline{S}_1(c) \quad (21)$$
$$\overline{\mathrm{TPF}}(c) = \overline{S}_1(c) \quad (22)$$
$$\underline{\mathrm{FPF}}(c) = \underline{S}_0(c) \quad (23)$$
$$\overline{\mathrm{FPF}}(c) = \overline{S}_0(c) \quad (24)$$

where the NPI lower and upper survival functions follow from (17)-(20). As the ROC combines the survival functions for the two groups, the NPI lower and upper ROC curves are again defined to be the optimal bounds for all such curves corresponding to any pair of survival functions $S_1(t)$ and $S_0(t)$ for $Y^1_{n_1+1}$ and $Y^0_{n_0+1}$ in between their respective NPI lower and upper survival functions as given by (17)-(20). As the ROC curve clearly depends monotonously on the survival functions, it is easily seen that the optimal bounds, which we define to be the NPI lower and upper ROC curves, are

$$\underline{\mathrm{ROC}} = \left\{ \left(\overline{\mathrm{FPF}}(c), \underline{\mathrm{TPF}}(c) \right), c \in (-\infty, \infty) \right\} \quad (25)$$
$$\overline{\mathrm{ROC}} = \left\{ \left(\underline{\mathrm{FPF}}(c), \overline{\mathrm{TPF}}(c) \right), c \in (-\infty, \infty) \right\} \quad (26)$$

In line with (5), these NPI lower and upper ROC curves are equal to

$$\underline{\mathrm{ROC}}(t) = \underline{S}_1(\overline{S}_0^{-1}(t)) \quad (27)$$
$$\overline{\mathrm{ROC}}(t) = \overline{S}_1(\underline{S}_0^{-1}(t)) \quad (28)$$

Here some care is required as the NPI lower and upper survival functions for the non-disease group, $\underline{S}_0$ and $\overline{S}_0$, are step-functions taking on values $w/(n+1)$ for $w \in \{0, 1, \ldots, n+1\}$. To avoid problems due to their inverse functions not being uniquely defined at these values, we define $\underline{S}_0^{-1}(w/(n+1))$ to be the infimum of all values t for which $\underline{S}_0(t) = w/(n+1)$. This is only of little relevance, for example the exact definition of the inverse function in

such points does not affect the area under the ROC curve, but it should be kept in mind also when we discuss NPI for the partial area under the ROC curve.

It is easily seen that $\underline{\mathrm{FPF}}(c) \leq \widehat{\mathrm{FPF}}(c) \leq \overline{\mathrm{FPF}}(c)$ and $\underline{\mathrm{TPF}}(c) \leq \widehat{\mathrm{TPF}}(c) \leq \overline{\mathrm{TPF}}(c)$ for all c, which implies that the empirical ROC curve is bounded by the NPI lower and upper ROC curves.

4.2. NPI lower and upper AUC

We are interested in the NPI lower and upper probabilities for the event that the test result for the next individual from the disease group is greater than the test result for the next individual from the non-disease group. Following (10) these NPI lower and upper probabilities can directly be defined as the NPI lower and upper AUC, and using the results by Coolen (1996) who introduced NPI for comparison of two fully independent groups, we have

$$\begin{aligned} \underline{\mathrm{AUC}} &= \underline{P}\left(Y_{n_1+1}^1 > Y_{n_0+1}^0\right) \\ &= \frac{1}{(n_1+1)(n_0+1)} \sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \mathbf{1}\left\{y_i^1 > y_j^0\right\} \qquad (29) \\ \overline{\mathrm{AUC}} &= \overline{P}\left(Y_{n_1+1}^1 > Y_{n_0+1}^0\right) \\ &= \frac{1}{(n_1+1)(n_0+1)} \left[\sum_{j=1}^{n_0} \sum_{i=1}^{n_1} \mathbf{1}\left\{y_i^1 > y_j^0\right\} + n_1 + n_0 + 1\right] \qquad (30) \end{aligned}$$

It is interesting to notice that the imprecision

$$\overline{\mathrm{AUC}} - \underline{\mathrm{AUC}} = \frac{n_1 + n_0 + 1}{(n_1+1)(n_0+1)}$$

depends only on the two sample sizes n_0 and n_1.

The NPI lower AUC (29) and upper AUC (30) are logically linked to the NPI lower ROC (25) and upper ROC (26), as $\underline{\mathrm{AUC}}$ is the area under the $\underline{\mathrm{ROC}}$ and $\overline{\mathrm{AUC}}$ is the area under the $\overline{\mathrm{ROC}}$. The proofs of these results are similar to the proof linking the empirical ROC and AUC, so that $\widehat{\mathrm{AUC}}$ is the area under $\widehat{\mathrm{ROC}}$, as given by Pepe (2003, p. 103). To adapt that proof, note that the horizontal step in $\underline{\mathrm{ROC}}$ ($\overline{\mathrm{ROC}}$) corresponding to the point y_j^0 adds a rectangular area of size $\frac{1}{n_0+1}\underline{\mathrm{TPF}}(y_j^0)\left(\frac{1}{n_0+1}\overline{\mathrm{TPF}}(y_j^0)\right)$. The result then follows because $\underline{\mathrm{AUC}}$ ($\overline{\mathrm{AUC}}$) is the sum (over j) of these $n_0 + 1$ rectangular

areas corresponding to the partition created by the n_0 observations from the non-disease group. We therefore see that the introduced NPI approach to ROC and AUC logically combines the generalizations of these concepts of the standard theory into the framework of lower and upper probabilities. Next we will introduce NPI for the partial area under the ROC curve, which again will be consistent with the concepts already introduced.

4.3. NPI lower and upper pAUC

As mentioned before, one may be interested in the partial area under the ROC curve (pAUC), with restriction of either the values of FPF or of TPF to belong to an interval of specific interest. The NPI method can also be used if we want to focus on particular values of FPF or of TPF. As before, we will focus on the area under the ROC curve between two values of FPF, the method is similar for the partial area under the ROC curve between two TPF values.

We first consider the NPI lower pAUC. We know from (25) that the NPI lower ROC combines the upper FPF and the lower TPF, or alternatively by (27) the NPI lower survival function for the disease group, $\underline{S}_1$, with the NPI upper survival function for the non-disease group, $\overline{S}_0$. Hence, we restrict the upper FPF to belong to an interval between two values t_0 and t_1, using (24), which implies that the restricted area is defined via the NPI upper survival function $\overline{S}_0$ for $Y^0_{n_0+1}$.

Let L_0 and U_0 be such that $L_0 = \overline{S}_0^{-1}(t_1)$ and $U_0 = \overline{S}_0^{-1}(t_0)$, where the comment near the end of Subsection 4.1 with regard to the inverse of such a step-function should be taken into account. Suppose that r_0 of the n_0 observations from the non-disease group are between L_0 and U_0, and let these be denoted by $y^0_{(j)}$, with $L_0 \leq y^0_{(1)} < \ldots < y^0_{(r_0)} \leq U_0$. In line with (14), the NPI lower and upper pAUC are defined as the NPI lower and upper probabilities for the combined event that $Y^1_{n_1+1} > Y^0_{n_0+1}$ and $Y^0_{n_0+1} \in (L_0, U_0)$.

These NPI lower and upper pAUC are

$$\underline{\text{pAUC}}(t_0,t_1) = \frac{1}{(n_1+1)(n_0+1)} \times \sum_{i=1}^{n_1}\left\{1\{U_0 < y_i^1\} + \sum_{j=1}^{r_0} 1\{y_{(j)}^0 < y_i^1\}\right\} \quad (31)$$

$$\overline{\text{pAUC}}(t_0,t_1) = \frac{1}{(n_1+1)(n_0+1)} \times \left[(r_0+1) + \sum_{i=1}^{n_1}\left\{1\{L_0 < y_i^1\} + \sum_{j=1}^{r_0} 1\{y_{(j)}^0 < y_i^1\}\right\}\right] \quad (32)$$

These equalities follow from the same arguments as Coolen (1996) used for comparison of two groups of real-valued data, but now only including the probabilities based on $A_{(n_0)}$ for $Y_{n_0+1}^0$ assigned to intervals that are within (L_0, U_0). Actually, due to the fact that $\overline{S}_0$ is a step function, with steps only at observations y_j^0 (and the way in which we defined the inverse for such a function, if the function value is equal to $w/(n_0+1)$ for some integer w as explained before), L_0 and U_0 are equal to some of these observed values of the non-disease group (or 0 or ∞). This implies that the NPI probability for the event $Y_{n_0+1}^0 \in (L_0, U_0)$ is equal to $(r_0+1)/(n_0+1)$, so this is a precise probability. Hence, we can also derive at these expressions (31) and (32) by a conditioning argument, which then involves deriving the NPI lower and upper probabilities for the event that $Y_{n_1+1}^1 > Y_{n_0+1}^0$ given that $Y_{n_0+1}^0 \in (L_0, U_0)$. In this case, we would only use the data $y_{(1)}^0 < \ldots < y_{(r_0)}^0$ between L_0 and U_0 (so not the data points which are equal to these) together with $A_{(r_0)}$ to calculate the NPI lower and upper probabilities for this conditional event, which would again be done following the method presented by Coolen (1996).

These definitions (31) and (32) of the NPI lower and upper pAUC are again fully consistent with the concepts introduced earlier in this paper. If we set $t_0 = 0$ and $t_1 = 1$, so we do not actually restrict to a partial area, we get $\underline{\text{pAUC}}(0,1) = \underline{\text{AUC}}$ and $\overline{\text{pAUC}}(0,1) = \overline{\text{AUC}}$. More importantly, these NPI lower and upper pAUC are actually the partial areas under the NPI lower and upper ROC with FPF restricted to the interval between t_0 and t_1, so $\underline{\text{pAUC}}(t_0,t_1)$ is the area under $\underline{\text{ROC}}(t)$ with $t \in (t_0,t_1)$ and similarly $\underline{\text{pAUC}}(t_0,t_1)$ is the area under $\underline{\text{ROC}}(t)$ with $t \in (t_0,t_1)$. This suitable geometric interpretation of these quantities is proven similarly to the corresponding proof for the NPI lower and upper AUC

in Subsection 4.2, and results from the fact that the horizontal step in $\underline{\text{ROC}}$ ($\overline{\text{ROC}}$) corresponding to the point $y^0_{(j)}$ adds a rectangular area of size $\frac{1}{n_0+1}\underline{\text{TPF}}(y^0_{(j)})\left(\frac{1}{n_0+1}\overline{\text{TPF}}(y^0_{(j)})\right)$. The result then follows because $\underline{\text{pAUC}}$ ($\overline{\text{pAUC}}$) is the sum (over j) of these r_0+1 rectangular areas corresponding to the partition created by $L_0, y^0_{(1)}, \ldots, y^0_{(r_0)}, U_0$. The NPI concepts introduced in this section are illustrated in Example 4.1.

Example 4.1. For a particular gene, the relative gene expression intensities for 23 non-diseased ('normal') ovarian tissues, Y^0, and 30 ovarian tumor ('cancer') tissues, Y^1, are given in Table 1 (Pepe, 2003). As we use this example to illustrate the method presented in this paper, we avoid the three pairs of tied observations between the two groups in the original data by adding 0.001 to the three relevant observations from the cancer tissues group (i.e. original values 0.571, 0.628 and 0.641). Further comments on breaking ties between groups in NPI were made in Section 2, we could use the original data including ties but it would require going through all 8 possible ways of breaking these 3 ties, which is straightforward to do but does not add to understanding of the main concepts presented here. Note that ties within a group do not really affect the NPI methods presented in this paper, but to avoid mathematical complications we also add a very small value to one of such a pair of tied observations when performing the calculations.

Normal tissues					Cancer tissues					
0.442	0.500	0.510	0.568	0.571	0.543	0.572	0.602	0.609	0.629	0.642
0.574	0.588	0.595	0.595	0.595	0.666	0.694	0.769	0.800	0.800	0.847
0.598	0.606	0.617	0.628	0.641	0.877	0.892	0.925	0.943	1.041	1.075
0.641	0.680	0.699	0.746	0.793	1.086	1.123	1.136	1.190	1.234	1.315
0.884	1.149	1.785			1.428	1.562	1.612	1.666	1.666	2.127

Table 1: The relative gene expression intensities

The NPI lower and upper ROC curves for the next future individuals from both groups, from equations (25) and (26), are plotted in Figure 1, together with the corresponding empirical ROC curve, illustrating that the NPI lower and upper ROC curves indeed bound the empirical ROC curve.

The empirical estimator for the area under the ROC curve is $\widehat{\text{AUC}} = 0.8116$. To illustrate the partial area under the ROC curve, suppose that we are interested in the values $0 \leq \text{FPF} \leq 0.25$. The corresponding empirical estimator is $\widehat{\text{pAUC}}(0, 0.25) = 0.1333$, to calculate this we have used linear

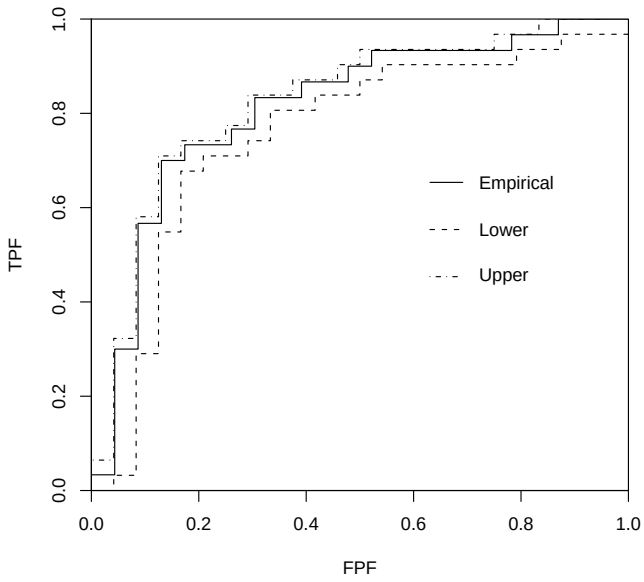

Figure 1: Empirical and NPI lower and upper ROC curves)

interpolation of the empirical distribution for the normal tissue group as suggested by Dodd and Pepe (2003).

The NPI lower and upper area under the ROC curve, for the next individuals from both the cancer tissue and normal tissue groups, calculated from (29) and (30), are $\underline{\text{AUC}} = \underline{P}(Y_{31}^1 > Y_{24}^0) = 0.7527$ and $\overline{\text{AUC}} = \overline{P}(Y_{31}^1 > Y_{24}^0) = 0.8253$. The NPI lower and upper partial area under the ROC curve for $0 \leq \text{FPF} \leq 0.25$ (i.e. $Y_{n_0+1}^0 \in (0.685, \infty)$ and $r_0 = 6$), calculated from (31) and (32), are $\underline{\text{pAUC}}(0, 0.25) = 0.1237$ and $\overline{\text{pAUC}}(0, 0.25) = 0.1640$. These NPI lower and upper AUC and pAUC indeed provide bounds for the corresponding empirical estimates.

5. Comparing continuous diagnostic tests

A common problem is to determine which of two diagnostic tests is better, this can be done by comparing the two corresponding ROC curves. Generally, such comparisons can be based on paired or unpaired designs. In a paired design each individual undergoes both tests, while in an unpaired de-

sign each individual is diagnosed using only one test. We restrict attention to unpaired designs, as paired designs result in two correlated ROC curves, and NPI has not yet been developed to take such correlations into account, this is left as an important topic for future research. Nevertheless, the comparison of continuous diagnostic tests in unpaired studies is of great practical importance, as often individuals cannot undergo both diagnostic tests.

There are several ways in which two diagnostic tests can be compared based on the corresponding ROC curves. For example, one may want to test the null-hypothesis that both full ROC curves correspond to the same underlying population of diagnostic test results, or alternatively test the null-hypothesis that this only applies at a particular FPF. Venkatraman and Begg (1996) and Venkatraman (2000) proposed permutation tests to compare the entire curves for the paired and unpaired cases in order to test whether the two ROC curves corrspond to the same population of diagnostic test values. Two diagnostic tests can also be compared by considering summaries of the ROC curves such as the area and the partial area under each ROC curve (AUC and pAUC), again this is possible via hypothesis testing. There are many other methods to compare diagnostic tests based on the corresponding ROC curves, including parametric, nonparametric and Bayesian techniques, for more details we refer to (DeLong et al., 1988; Krzanowski and Hand, 2009; Pepe, 2003; Zhou et al., 2002; Zhang et al., 2002). The important difference of the NPI method presented in this paper, compared to the alternatives in the literature, is the explicit predictive nature of the inferences. To compare two diagnostic tests in NPI, one does not start from a null-hypothesis but instead one considers one further future individual from each of the disease and non-disease groups, and compares the random performances of these tests based on the appropriate $A_{(n)}$ assumptions per group. Such a comparison can for example be based on the NPI lower and upper AUC or pAUC for the two diagnostic tests, which will be illustrated and discussed further in Example 5.1.

Example 5.1. We consider a sample of 60 patients reporting to a hospital with severe head trauma. For each patient, the CK-BB isoenzyme is measured within 24 hours of the injury to predict whether the patient will have poor outcome (death, vegetative state, or severe disability) or good outcome (reasonable to full recovery) after suffering a severe head trauma. Later, 19 out of 60 patients had reasonable to full recovery (Good outcome) and 41 patients had poor or no recovery (Poor outcome) (Zhou et al., 2002, p. 138).

These patients are categorized according to their age into the younger group A (age < 20) or older group B (age ≥ 20). The data are presented in Table 2. To avoid complicating the presentation in order to deal fully with ties, we will assume that when ties occur between different outcome groups, the observation from the Poor outcome group is slightly larger than the corresponding observation from the Good outcome group.

	Poor Outcome							Good Outcome				
Group A	16	90	126	140	153	183	193	17	23	27	60	96
(Age < 20)	230	253	283	303	700	740	800	100	100	126	136	146
	1087	1256						200	220	281	286	
Group B	60	76	76	80	120	156	206	6	40	46	70	253
(Age ≥ 20)	216	230	303	323	350	353	356					
	443	463	490	509	523	543	576					
	671	913	1370	1560								

Table 2: Severe head trauma data

In this example two main issues are considered, namely the ability of CK-BB to distinguish between patients with a poor outcome and patients with a good outcome, and whether this ability varies according to patient's age, so to examine whether the CK-BB is a better predictor for group B than for group A in discriminating between patients with regard to the outcome. The empirical ROC curves for the combined data (no age considered) and for groups A and B are shown in Figure 2.

For the combined data we put groups A and B together, giving 41 observations in the poor outcome group and 19 in the good outcome group. We calculate the NPI lower and upper ROC curves for the next individuals from the poor outcome group and the good outcome group, say Y_{42}^0 (using the assumption $A_{(41)}$) and Y_{20}^1 (using the assumption $A_{(19)}$), respectively. These NPI lower and upper ROC curves, and the corresponding empirical ROC curve, are plotted in Figure 3. The empirical estimates for the area and the partial area, with $0 \leq \text{FPF} \leq 0.2$, under this ROC curve (so with groups A and B combined) are $\widehat{\text{AUC}} = 0.8306$ and $\widehat{\text{pAUC}}(0, 0.2) = 0.1220$. The corresponding NPI lower and upper AUC and pAUC are $\underline{\text{AUC}} = 0.7702$, $\overline{\text{AUC}} = 0.8429$, $\underline{\text{pAUC}}(0, 0.2) = 0.1131$ and $\overline{\text{pAUC}}(0, 0.2) = 0.1524$.

Figure 2 shows that the empirical ROC curve for groups A and B combined is almost everywhere in between the separate empirical ROC curves for groups A and B, except for two areas with FPF between 0.3 and 0.6. The empirical ROC curve for the group B is everywhere above the empiri-

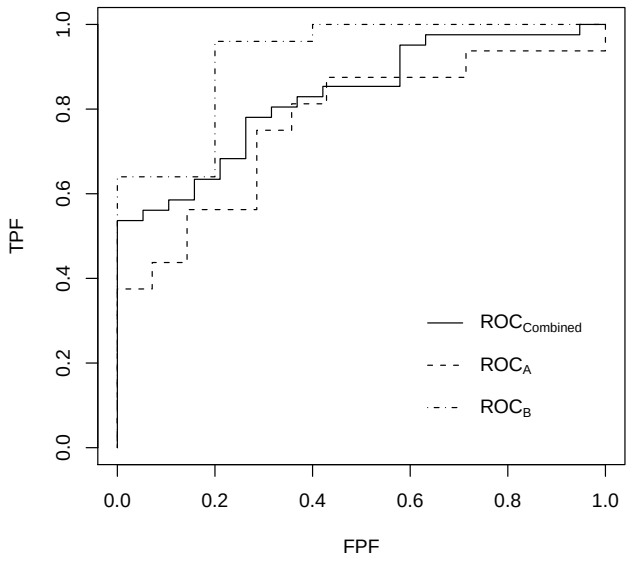

Figure 2: Empirical ROC curves for the combined groups, for group A and group B

cal ROC curve for group A. The empirical estimates for the areas and the partial areas (again with $0 \leq \text{FPF} \leq 0.2$) under the ROC curves for groups A and B are $\widehat{\text{AUC}}_A = 0.7679$, $\widehat{\text{AUC}}_B = 0.9200$, $\widehat{\text{pAUC}}_A(0, 0.2) = 0.0982$ and $\widehat{\text{pAUC}}_B(0, 0.2) = 0.1280$.

Zhou et al. (2002) discuss several methods to compare two diagnostic tests through (summaries of) their ROC curves. For example, they present (Zhou et al., 2002, Sect. 5.2.4) a test for the null-hypothesis that the AUC corresponding to each ROC curve is the same, so meaning that they could result from a single underlying diagnostic test with differences just due to expected random fluctuations. Applying this test leads to test statistic $z = 1.31$, which approximately the standard normal null-distribution, hence at 5% significance one would not reject that the AUC for group A and the AUC for group B might result from the same diagnostic test in this example.

Using the NPI approach, such a comparison would be made on the basis of the NPI lower and upper AUC or pAUC. The NPI lower and upper AUC for the next individuals from the poor and good outcome groups are $\underline{\text{AUC}}_A = 0.6745$ and $\overline{\text{AUC}}_A = 0.7961$ for group A and $\underline{\text{AUC}}_B = 0.7372$ and

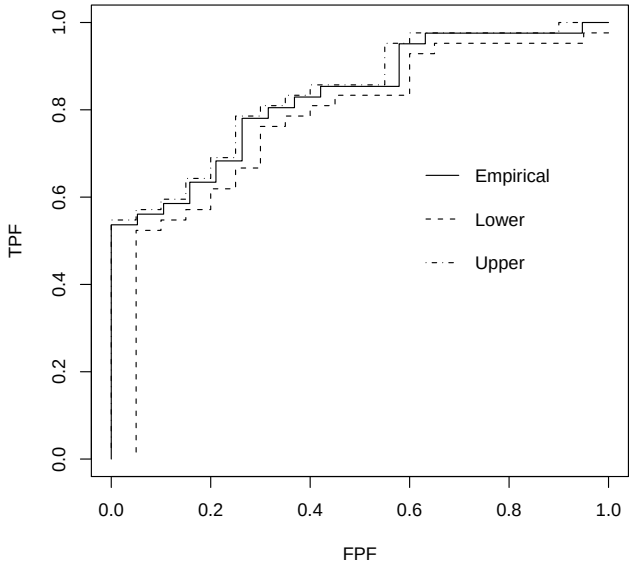

Figure 3: Empirical and NPI lower and upper ROC curves for the combined groups

$\overline{\text{AUC}}_B = 0.9359$ for group B. As $\underline{\text{AUC}}_B > \underline{\text{AUC}}_A$ and $\overline{\text{AUC}}_B > \overline{\text{AUC}}_A$, one could interpret this as providing a weak indication that the area under the ROC curve is larger for group B than for group A, and hence that the CK-BB isoenzyme method is more accurate for individuals from the older group than from the younger group. Such an indication would have been stronger if $\underline{\text{AUC}}_B > \overline{\text{AUC}}_A$, which is not the case in this example, we refer to Maturi (2010) for further discussion of comparisons based on NPI lower and upper probabilities with examples for groups of lifetime data. If we compared instead the pAUC with $0 \leq \text{FPF} \leq 0.2$, then we would have a similar conclusion as $\underline{\text{pAUC}}_A(0, 0.2) = 0.0863$, $\overline{\text{pAUC}}_A(0, 0.2) = 0.1373$, $\underline{\text{pAUC}}_B(0, 0.2) = 0.1026$ and $\overline{\text{pAUC}}_B(0, 0.2) = 0.2692$.

6. Concluding remarks

In this paper we introduced NPI for diagnostic tests with continuous test results. In particular, we introduced NPI for ROC curves and two popular summaries, AUC and pAUC. We assumed that the available data did not

contain censored observations, but the concepts and methods introduced can easily be generalized to deal with right-censored observations in the data, using NPI for right-censored data (Coolen and Yan, 2003, 2004), see Maturi (2010) for more results on comparison of different groups of lifetime data including right-censored observations and also for simplified formulae for the corresponding NPI lower and upper survival functions. For ordinal data it is also possible to base inference on accuracy of diagnostic tests on the corresponding ROC curves, the development of NPI for this case is left as a topic for future research, which will be achievable following general development of NPI for ordinal data which is currently in progress. NPI for accuracy of binary diagnostic tests is in development, we hope to report on this in the near future.

References

Augustin, T., Coolen, F. P. A., 2004. Nonparametric predictive inference and interval probability. Journal of Statistical Planning and Inference 124 (2), 251–272.

Coolen, F. P. A., 1996. Comparing two populations based on low stochastic structure assumptions. Statistics & Probability Letters 29 (4), 297–305.

Coolen, F. P. A., 1998. Low structure imprecise predictive inference for bayes' problem. Statistics & Probability Letters 36 (4), 349–357.

Coolen, F. P. A., 2006. On nonparametric predictive inference and objective bayesianism. Journal of Logic, Language and Information 15 (1-2), 21–47.

Coolen, F. P. A., Coolen-Schrijner, P., Yan, K. J., 2002. Nonparametric predictive inference in reliability. Reliability Engineering & System Safety 78 (2), 185–193.

Coolen, F. P. A., Yan, K. J., 2003. Comparing two groups of lifetime data. In: Bernard, J. M., Seidenfeld, T., Zaffalon, M. (Eds.), ISIPTA'03: Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications. pp. 148–161.

Coolen, F. P. A., Yan, K. J., 2004. Nonparametric predictive inference with right-censored data. Journal of Statistical Planning and Inference 126 (1), 25–54.

De Finetti, B., 1974. Theory of Probability: A Critical Introductory Treatment. Wiley, London.

DeLong, E. R., DeLong, D. M., Clarke-Pearson, D. L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 44 (3), 837–845.

Dodd, L. E., Pepe, M. S., 2003. Partial auc estimation and regression. Biometrics 59 (3), 614–623.

Hill, B. M., 1968. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. Journal of the American Statistical Association 63 (322), 677–691.

Hill, B. M., 1988. De finetti's theorem, induction, and a_n, or bayesian nonparametric predictive inference (with discussion). In: Bernando, J. M., DeGroot, M. H., Lindley, D. V., Smith, A. (Eds.), Bayesian Statistics 3. Oxford University Press, pp. 211–241.

Krzanowski, W. J., Hand, D. J., 2009. ROC Curves for Continuous Data. Chapman & Hall, Boca Raton.

Lawless, J. F., Fredette, M., 2005. Frequentist prediction intervals and predictive distributions. Biometrika 92, 529–542.

Maturi, T. A., 2010. Nonparametric predictive inference for multiple comparisons. Ph.D. thesis, Durham University, Durham, UK, available from www.npi-statistics.com.

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

Venkatraman, E. S., 2000. A permutation test to compare receiver operating characteristic curves. Biometrics 56 (4), 1134–1138.

Venkatraman, E. S., Begg, C. B., 1996. A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. Biometrika 83 (4), 835–848.

Walley, P., 1991. Statistical Reasoning with Imprecise Probabilities. Chapman & Hall, London.

Weichselberger, K., 2001. Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept. Physika, Heidelberg.

Zhang, D. D., Zhou, X.-H., Freeman, D. H., Freeman, J. L., 2002. A non-parametric method for the comparison of partial areas under roc curves and its application to large health care data sets. Statistics in Medicine 21 (5), 701–715.

Zhou, X.-H., McClish, D. K., Obuchowski, N. A., 2002. Statistical Methods in Diagnostic Medicine. Wiley-Interscience, New York.