Nonparametric predictive inference for three-group ROC analysis

Tahani Coolen-Maturi[a], Faiza F. Elkhafifi[b], Frank P.A. Coolen[c,*]

[a]*Durham University Business School, Durham University, Durham, DH1 3LB, UK*
[b]*Department of Statistics, Benghazi University, Benghazi, LIBYA*
[c]*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, health care and credit scoring. Good methods for determining diagnostic accuracy in medicine provide useful guidance on selection of patient treatment according to the severity of their health status. The receiver operating characteristic (ROC) surface is a useful tool to assess the ability of a diagnostic test to discriminate among three ordered classes or groups. In this paper, nonparametric predictive inference (NPI) for three-group ROC analysis for continuous data is presented. NPI is a frequentist statistical method that is explicitly aimed at using few modelling assumptions in addition to data, enabled through the use of lower and upper probabilities to quantify uncertainty. Furthermore, it focuses exclusively on future observations, which provides an alternative perspective to the usual approaches which typically aim at estimation of characteristics of assumed underlying populations. This focus on prediction may be particularly relevant if one considers decisions about a diagnostic test that must be applied to a future patient. This paper presents the NPI approach to three-group ROC analysis, including results on the volumes under the ROC surfaces and consideration of the choice of decision thresholds for the diagnosis.

Keywords: Diagnostic accuracy, Imprecise probability, Nonparametric predictive inference, Receiver operating characteristic (ROC) surface, Youden's index.

*Corresponding author

Email addresses: tahani.maturi@durham.ac.uk (Tahani Coolen-Maturi), f_elkhafifi@yahoo.com (Faiza F. Elkhafifi), frank.coolen@durham.ac.uk (Frank P.A. Coolen)

1. Introduction

Measuring the accuracy of diagnostic tests is crucial in many application areas including medicine, health care (Wians et al., 2001; Xiong et al., 2007; Lopez-de Ullibarri et al., 2008; Tian et al., 2011; Rodriguez-Alvarez et al., 2011a,b), including new developments involving genomic classifiers (Chen et al., 2012), and credit scoring (Xanthopoulos and Nakas, 2007). Good methods for determining diagnostic accuracy provide useful guidance on selection of patient treatment according to the severity of their health status. The receiver operating characteristic (ROC) surface has proven to be a useful tool to assess the ability of a diagnostic test to discriminate among three ordered classes or groups. The construction of the ROC surface based on the probabilities of correct classification for three classes has been introduced by Mossman (1999), Nakas and Yiannoutsos (2004) and Nakas and Alonzo (2007). They also considered the volume under the ROC surface (VUS), and its relation to the probability of correctly ordered observations from the three groups. The three-group ROC surface generalizes the popular two-group ROC curve, which in recent years has attracted much theoretical attention and has been widely applied for analysis of accuracy of diagnostic tests.

Statistical inference for such accuracy using ROC curves or surfaces has mostly focused on estimating the relevant probabilities of correct classification for the different groups, with these probabilities being considered as properties of assumed underlying populations. While this is a well-established approach, both from frequentist and Bayesian perspective, the practical importance of diagnostic tests may well be in their use for future patients. As such, it is of interest to study a predictive statistical approach to such inferences on accuracy of diagnostic tests. The importance of prediction is well understood, e.g. Airola et al. (2011); van Calster et al. (2012) explicitly mention 'predictive models' and 'prediction models', but thus far the statistical approaches used in this field have mostly been based on estimation, with their predictive performance investigated via numerical studies.

In recent years, nonparametric predictive inference (NPI) has been presented as a frequentist statistical method which uses few modelling assumptions, and hence is strongly data-driven, which is enabled by the use of lower and upper probabilities to quantify uncertainty (Augustin and Coolen, 2004; Coolen, 2006, 2011). Lower and upper probabilities generalize the classical theory of (precise) probability (Coolen et al., 2011), with the difference between the upper and lower probabilities for an event typically reflecting the amount of information available, with little information leading to large imprecision while in case of much information the corresponding lower and up-

per probabilities are nearly identical. In NPI, the lower and upper probabilities always provide bounds for empirical probabilities, hence the NPI-based statistical conclusions are never contradictory to those based on empirical probabilities (Coolen, 2006). Due to the importance of prediction of the accuracy of diagnostic tests for a future patient, NPI provides an attractive alternative approach to the established methods in this field. NPI has recently been introduced for assessing the accuracy of a classifier's ability to discriminate between two outcomes (or two groups) for binary data (Coolen-Maturi et al., 2012a) and for diagnostic tests with ordinal observations (Elkhafifi and Coolen, 2012) and with real-valued observations (Coolen-Maturi et al., 2012b). In this paper we introduce NPI for the three-group ROC analysis for continuous data, where we also consider the volume under the ROC surfaces and selection of optimal decision thresholds for the diagnosis.

The outline of this paper is as follows. Section 2 provides a brief introduction to NPI, followed in Section 3 by an introduction to three-group ROC analysis. In Section 4 NPI is presented for three-group ROC analysis, which includes the introduction of NPI lower and upper ROC surfaces. We will first derive the lower and upper envelopes for the set of all ROC surfaces that follow from applying NPI to the observed data per group. Then, by considering explicitly the link of the volume under an ROC surface to the probability of correctly ordered observations from the three groups, we define the NPI lower and upper ROC surfaces. We also consider the selection of optimal decision thresholds for the diagnosis based on these ROC surfaces. Three examples are provided for illustration in Section 5. The paper ends with some concluding remarks in Section 6, and an appendix which presents several proofs of results presented in Section 4.

2. Nonparametric predictive inference

Nonparametric predictive inference (NPI) (Augustin and Coolen, 2004; Coolen, 2006, 2011) is based on the assumption $A_{(n)}$ proposed by Hill (1968). Let $X_1, \ldots, X_n, X_{n+1}$ be real-valued absolutely continuous and exchangeable random quantities. Let the ordered observed values of $X_1, X_2, \ldots, X_n$ be denoted by $x_1 < x_2 < \ldots < x_n$ and let $x_0 = -\infty$ and $x_{n+1} = \infty$ for ease of notation. We assume that no ties occur; ties can be dealt with in NPI by assuming that tied observations differ by small amounts which tend to zero (Coolen, 2006). For X_{n+1}, representing a future observation, $A_{(n)}$ partially specifies a probability distribution by $P(X_{n+1} \in (x_{i-1}, x_i)) = \frac{1}{n+1}$ for $i = 1, \ldots, n+1$. $A_{(n)}$ does not assume anything else, and can be considered to be a post-data assumption related to exchangeability (De Finetti, 1974). It is convenient to introduce the set of precise probability distributions which

correspond to the partial specification by $A_{(n)}$, so which have probability $\frac{1}{n+1}$ in each of the $n+1$ intervals (x_{i-1}, x_i). This set is called a 'structure' by Weichselberger (2000, 2001), we will denote it by $\mathcal{P}_x$.

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 any such further information in order to derive at inferences that are strongly based on the data. The assumption $A_{(n)}$ 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, 2000, 2001; Coolen et al., 2011).

In NPI, uncertainty about the future observation X_{n+1} is quantified by lower and upper probabilities for events of interest. Lower and upper probabilities generalize classical ('precise') probabilities, and a lower (upper) probability for event A, denoted by $\underline{P}(A)$ ($\overline{P}(A)$), can be interpreted as supremum buying (infimum selling) price for a gamble on the event A (Walley, 1991), or just as the maximum lower (minimum upper) bound for the probability of A that follows from the assumptions made (Coolen, 2006). This latter interpretation is used in NPI (Coolen, 2011). We wish to explore application of $A_{(n)}$ for inference without making further assumptions. So, NPI lower and upper probabilities are the sharpest bounds on a probability for an event of interest when only $A_{(n)}$ is assumed. Using the $A_{(n)}$-based structure, the NPI lower probability for event A is

$$\underline{P}(A) = \inf_{P \in \mathcal{P}_x} P(A)$$

and the corresponding NPI upper probability for event A is

$$\overline{P}(A) = \sup_{P \in \mathcal{P}_x} P(A)$$

Informally, $\underline{P}(A)$ ($\overline{P}(A)$) can be considered to reflect the evidence in favour of (against) event A (Coolen et al., 2011). Augustin and Coolen (2004) proved that NPI has strong consistency properties in the theory of interval probability (Walley, 1991; Weichselberger, 2000, 2001; Coolen et al., 2011), it is also exactly calibrated from frequentist statistics perspective (Lawless and Fredette, 2005). We should emphasize that, in this paper, inferences are restricted to a single future observation X_{n+1} (per group). NPI also provides interesting opportunities and challenges when one considers multiple future observations (Arts et al., 2004), where it is important that the interdependence of the multiple future observations is taken into account. This

also provides possibilities to generalize the results presented in this paper to diagnostic tests for multiple future patients, which is left as an interesting challenge for future research.

3. Three-group ROC analysis

In this section we present an introduction to three-group ROC analysis (Mossman, 1999; Nakas and Yiannoutsos, 2004; Nakas and Alonzo, 2007), including concepts and notation which are important for the remainder of the paper. Let there be three groups or classes, denoted by X, Y and Z. Throughout this paper, we assume that these groups are fully independent, in the sense that any information about one of the groups does not hold any information about another group. Let $x_1, x_2, ..., x_{n_x}$ denote the observed test results for n_x subjects from group X, $y_1, y_2, ..., y_{n_y}$ the observed test results for n_y subjects from group Y and $z_1, z_2, ..., z_{n_z}$ the observed test results for n_z subjects from group Z. All these test results are assumed to be real-valued observations. Suppose that a continuous diagnostic test is used to discriminate the subjects from these groups. We assume that the three groups are ordered in the sense that observations from group X tend to be lower than those from group Y, which in turn tend to be lower than those from group Z. There will typically be overlap of observations from different groups, but the practical diagnostic setting is assumed to be such that observations from the three groups tend to be ordered in this way. If there is no such practical knowledge, the groups can be re-labelled and the methods in this paper can be applied to several (or all) of the six possible orderings of the three groups. The cumulative distribution functions (CDF) for the test outcomes of groups X, Y and Z are denoted by F_x, F_y and F_z, respectively.

For continuous test results, as considered in this paper, two ordered decision threshold points, say $c_1 < c_2$, are required in order to classify a subject into one of the three groups. A diagnostic decision for each subject is based on the following rule, where T_j represents the test results for subject j: If $T_j \leq c_1$ then subject j is classified into group X, if $c_1 < T_j \leq c_2$ then subject j is classified into group Y, and if $T_j > c_2$ then subject j is classified into group Z. It is important to emphasize that the test data are assumed to consist of measurements for individuals known to belong to specific groups, while the goal of the inferences is to develop a diagnostic classification method for patients for who the group is unknown. We assume throughout the paper that the test data does not contain mistakes, generalization of the NPI approach to deal with such possible mistakes provides an interesting topic for future research. There are interesting further topics of practical interest, for example Shiu and Gatsonis (2012) consider continuous disease states which

also need to be classified into groups, leading to an additional decision problem for diagnostic methods. Development of NPI for such scenarios is also left as an interesting topic for future research.

For a pair of thresholds (c_1, c_2), the probability of correct classification of a subject from group X is $p_1 = P(X \leq c_1) = F_x(c_1)$, the probability of correct classification of a subject from group Y is $p_2 = P(c_1 < Y \leq c_2) = F_y(c_2) - F_y(c_1)$, and the probability of correct classification of a subject from group Z is $p_3 = P(Z > c_2) = 1 - F_z(c_2)$. The ROC surface, denoted by ROC_s, can be constructed by plotting these probabilities of correct classification for all possible real-valued $c_1 < c_2$, so by plotting all triples (p_1, p_2, p_3) resulting from varying c_1 and c_2 over all real values with the constraint $c_1 < c_2$. A mathematically convenient way to define this ROC surface is as follows, for $p_1, p_3 \in [0, 1]$ (Inacio et al., 2011; Nakas and Yiannoutsos, 2004; Tian et al., 2011)

$$ROC_s(p_1, p_3) = \begin{cases} F_y(F_z^{-1}(1-p_3)) - F_y(F_x^{-1}(p_1)) & \text{if } F_x^{-1}(p_1) \leq F_z^{-1}(1-p_3) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $F_x^{-1}(p)$ is the inverse function of the CDF F_x, and similar for F_z.

The empirical estimator of the ROC surface can be obtained by replacing the CDFs in (1) with their empirical counterparts (Beck, 2005; Inacio et al., 2011), that is, for $p_1, p_3 \in [0, 1]$,

$$\widehat{ROC}_s(p_1, p_3) = \begin{cases} \hat{F}_y(\hat{F}_z^{-1}(1-p_3)) - \hat{F}_y(\hat{F}_x^{-1}(p_1)) & \text{if } \hat{F}_x^{-1}(p_1) \leq \hat{F}_z^{-1}(1-p_3) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\hat{F}_x^{-1}(p) = x_i$ if $p \in (\frac{i-1}{n_x}, \frac{i}{n_x}]$, $i = 1, \ldots, n_x$, and $\hat{F}_x^{-1}(p) = -\infty$ if $p = 0$, with $\hat{F}_z^{-1}(p)$ defined similarly.

For three-group diagnostic tests the volume under the ROC surface (VUS) can be considered as a global summary measure of the test's ability to discriminate between the three groups. It should be mentioned that there are alternative measures that can provide useful insights (van Calster et al., 2012), investigation of such measures within the NPI framework is left as a topic for future research. The VUS is equal to the probability that three randomly selected measurements, one from each group, are correctly ordered, so that the observation from group X is less than the observation from group Y and the latter is less than the observation from group Z (Mossman, 1999; Nakas and Yiannoutsos, 2010). The VUS can be considered as a generalization of the area under the ROC curve (AUC) for two-group diagnostic tests, where the corresponding equality between the AUC and the probability of correctly ordered observations from the two groups also holds (Airola et al.,

2011). An unbiased nonparametric estimator of the VUS is given by (Nakas and Yiannoutsos, 2004, 2010)

$$\widehat{VUS} = \frac{1}{n_x n_y n_z} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \sum_{l=1}^{n_z} I(x_i < y_j < z_l) \tag{3}$$

with indicator function $I(A)$ equal to 1 if A is true and 0 else. Equation (3) gives the proportion of all possible triple combinations from the data that are correctly ordered, and it is therefore the empirical probability for this event based on the information from the data. $\widehat{VUS}$ can take values from 0 to 1. It is (about) equal to 1/6 if the diagnostic test outcomes for the three groups completely overlap, in which case the data suggest that the test is not useful for the diagnosis. If there is a perfect separation of the test results for the three groups, that is $x_i < y_j < z_l$ for all i,j and l, then $\widehat{VUS} = 1$.

In practice, ties between measurements may occur, in this case a modified version of (3) should be used (Nakas and Yiannoutsos, 2004, 2010). In this paper, for ease of presentation we assume that no ties occur in the data. In NPI, ties can be dealt with by breaking them, that is by assuming that tied observations differ by very small amounts (Hill, 1988). The use of lower and upper probabilities makes this conceptually easy, as one can break ties in all possible ways (that is creating all corresponding orderings of the data), and defining the overall NPI lower and upper probabilities as the minimum and maximum, respectively, of the NPI lower and upper probabilities corresponding to all manners in which the ties could be broken.

The selection of the optimal cut-off points c_1 and c_2, also called the threshold values, is an important aspect of defining the diagnostic test and analyzing its quality. Several approaches for choosing these cut-off points have been proposed in the literature and their statistical properties have been investigated (Greiner et al., 2000; Schafer, 1989; Yousef et al., 2009; Lai et al., 2012). In this paper we consider Youden's index (Youden, 1950), which for three-group diagnostic tests was introduced by Nakas et al. (2010) and is given by

$$\begin{aligned} J(c_1, c_2) &= P(X \leq c_1) + P(c_1 < Y \leq c_2) - P(Z \leq c_2) + 1 \\ &= F_x(c_1) + F_y(c_2) - F_y(c_1) - F_z(c_2) + 1 \end{aligned} \tag{4}$$

Using this index, the optimal cut-off points are the values of c_1 and c_2 which maximise $J(c_1, c_2)$. This index $J(c_1, c_2)$ is equal to 1 if the three CDFs F_x, F_y and F_z are identical, while $J(c_1, c_2) = 3$ if the three groups are perfectly separated, that is if $P(X < Y < Z) = 1$. The empirical estimator for $J(c_1, c_2)$ is obtained by replacing the CDFs by the corresponding empirical CDFs.

4. NPI for three-group ROC analysis

In this section the main results of this paper are presented. The nonparametric predictive inference (NPI) approach for three-group ROC analysis is introduced and corresponding results for the volumes under the ROC surfaces are derived. First the notation required is introduced in Section 4.1, which includes the introduction of the NPI-based structures for the next observation from each of the three groups. The set of all ROC surfaces corresponding to probability distributions in these NPI-based structures is of main interest, in Section 4.2 the lower and upper envelopes of this set of ROC surfaces is derived. These envelopes are derived by pointwise optimisations and represent this set well, but they are too wide in the sense that the volumes under their surfaces are not generally the infimum and supremum of the volumes under the ROC surfaces in this set. We wish to define NPI lower and upper ROC surfaces for which the volumes under them are equal to this infimum and supremum, respectively. To achieve this, we consider the relation between the volume under an ROC surface and the probability of correctly ordered observations from the three groups. The NPI lower and upper probabilities for this event are presented in Section 4.3, followed by the introduction of the corresponding NPI lower and upper ROC surfaces in Section 4.4. In Section 4.5 the choice of decision threshold for the diagnosis is considered within the NPI approach. Computation of the NPI lower and upper ROC surfaces is not straightforward, and it may be attractive to quickly derive bounds for them. The envelopes presented in Section 4.2 provide a lower bound for the NPI lower ROC surface and an upper bound for the NPI upper ROC surface. In addition, in Section 4.6 we present an upper bound for the NPI lower ROC surface and a lower bound for the NPI upper ROC surface, both of which are also straightforward to derive. Together with the bounds provided by the envelopes, these bounds may well provide sufficient information about the actual NPI lower and upper ROC surfaces such that their detailed calculation may not be required. A further argument for including these bounds lies in the possibility to generalize this NPI approach to ROC-based comparison of more than three groups, leading to ROC hypersurfaces. Such a generalization would follow the concepts as presented in this paper, but would lead to increasingly complex computations to derive the exact NPI ROC hypersurfaces, so the idea of approximations based on lower and upper bounds which are straightforward to compute would be even more attractive. We leave this generalization as an interesting topic for future research.

4.1. Notation

To develop the NPI approach for three-group ROC analysis, let X_{n_x+1}, Y_{n_y+1} and Z_{n_z+1} be the next (future) observations from groups X, Y and Z, respectively. We apply $A_{(n)}$ for each group as follows. Let the n_x ordered observations from group X be denoted by $x_1 < x_2 < \ldots < x_{n_x}$ and let $x_0 = -\infty$ and $x_{n_x+1} = \infty$ for ease of notation. For X_{n_x+1}, representing a future observation from group X, $A_{(n_x)}$ partially specifies a probability distribution by $P(X_{n_x+1} \in (x_{i-1}, x_i)) = \frac{1}{n_x+1}$ for $i = 1, \ldots, n_x + 1$. Similarly, let the n_y ordered observations from group Y be denoted by $y_1 < y_2 < \ldots < y_{n_y}$ and let $y_0 = -\infty$ and $y_{n_y+1} = \infty$. For Y_{n_y+1}, representing a future observation from group Y, $A_{(n_y)}$ partially specifies a probability distribution by $P(Y_{n_y+1} \in (y_{j-1}, y_j)) = \frac{1}{n_y+1}$ for $j = 1, \ldots, n_y + 1$. Finally, let the n_z ordered observations from group Z be denoted by $z_1 < z_2 < \ldots < z_{n_z}$ and let $z_0 = -\infty$ and $z_{n_z+1} = \infty$. For Z_{n_z+1}, representing a future observation from group Z, $A_{(n_z)}$ partially specifies a probability distribution by $P(Z_{n_z+1} \in (z_{l-1}, z_l)) = \frac{1}{n_z+1}$ for $l = 1, \ldots, n_z+1$. The sets of all probability distributions that correspond to these partial specifications, for X_{n_x+1}, Y_{n_y+1} and Z_{n_z+1}, are the NPI-based structures and are denoted by $\mathcal{P}_x$, $\mathcal{P}_y$ and $\mathcal{P}_z$, respectively.

For $x \in [x_{i-1}, x_i)$ the NPI lower CDF for X_{n_x+1} is $\underline{F}_x(x) = \frac{i-1}{n_x+1}$, $i = 1, \ldots, n_x + 1$, and for $x \in (x_{i-1}, x_i]$ the NPI upper CDF for X_{n_x+1} is $\overline{F}_x(x) = \frac{i}{n_x+1}$, $i = 1, \ldots, n_x + 1$. Note that there is no imprecision at the x_i, as $\underline{F}_x(x_i) = \overline{F}_x(x_i) = \frac{i}{n_x+1}$ for $i = 0, 1, \ldots, n_x + 1$. These lower and upper CDFs are derived as the pointwise infima and suprema over all corresponding CDFs in the structure $\mathcal{P}_x$. The NPI lower and upper CDFs for Y_{n_y+1} and Z_{n_z+1} are similarly defined.

4.2. Lower and upper envelopes of the set of NPI-based ROC surfaces

For each combination of probability distributions for X_{n_x+1}, Y_{n_y+1} and Z_{n_z+1} in their respective NPI-based structures, $\mathcal{P}_x$, $\mathcal{P}_y$ and $\mathcal{P}_z$, the corresponding ROC surface as presented in Equation (1) can, in principle, be created. This will lead to a set of such NPI-based ROC surfaces, which we denote by $\mathcal{S}_{roc}$. The lower and upper envelopes of this set are of interest, they consist of the pointwise infima and suprema for this set. These envelopes are presented in Theorem 4.1, but first their construction is explained with the aid of Figure 1.

To derive the lower and upper envelopes of the set $\mathcal{S}_{roc}$, we need to derive the infima and suprema of the values $ROC_s(p_1, p_3)$ for ROC surfaces in the set $\mathcal{S}_{roc}$. Consider a value for $p_1 \in (0, 1)$ that is not equal to a value $i/(n_x+1)$ for any $i = 1, \ldots, n_x$. There is a unique $i \in \{1, \ldots, n_x + 1\}$ such that $x_{i-1} < F_x^{-1}(p_1) < x_i$ for every CDF F_x corresponding to all probability distributions

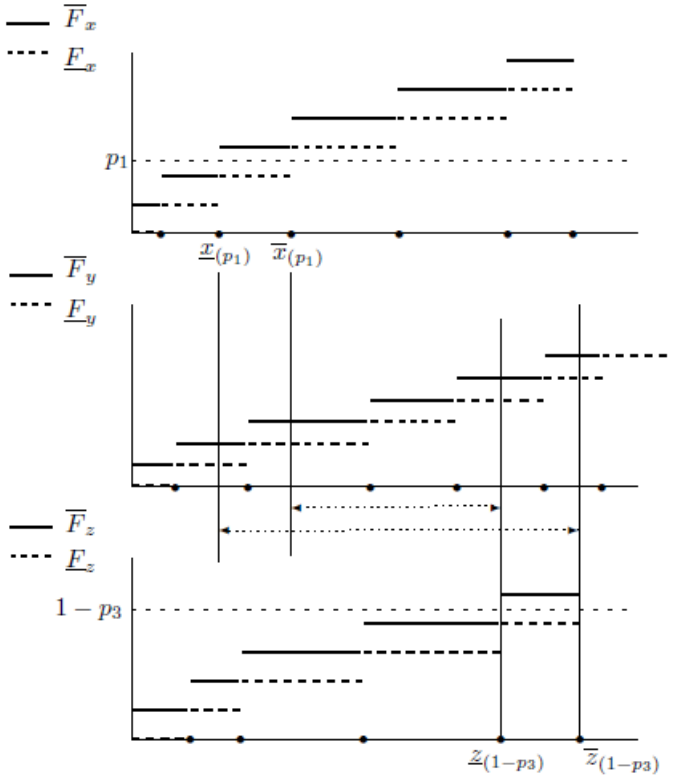


Figure 1: Construction of NPI lower and upper ROC surfaces

in $\mathcal{P}_x$. As indicated in Figure 1, we denote these x_{i-1} and x_i by $\underline{x}_{(p_1)}$ and $\overline{x}_{(p_1)}$, respectively, so $\underline{F}_x(\underline{x}_{(p_1)}) < p_1 < \overline{F}_x(\overline{x}_{(p_1)})$ for the CDFs corresponding to all probability distributions in $\mathcal{P}_x$. For $p_1 = \frac{i}{n_x+1}$, $i = 1, \ldots, n_x$, we would have $x_{i-1} < F_x^{-1}(p_1) < x_{i+1}$, but for ease of presentation we neglect this further as it only describes what happens to the envelopes we create at the finite number of points corresponding to the observations. Also for the volumes under these lower and upper envelopes of all the ROC surfaces in $\mathcal{S}_{roc}$, which are important later in this paper, it is irrelevant what happens exactly at this finite number of points. Similarly, consider a value $p_3 \in (0, 1)$ which is not equal to a value $l/(n_z + 1)$ for any $l = 1, \ldots, n_z$. We now consider all the inverse CDFs F_z^{-1}, corresponding to all probability distributions in $\mathcal{P}_z$, and we are interested in their value at $1 - p_3$. It is clear that there are two consecutive observations which we can denote by $\underline{z}_{(1-p_3)}$ and $\overline{z}_{(1-p_3)}$, with $\underline{z}_{(1-p_3)} < F_z^{-1}(1 - p_3) < \overline{z}_{(1-p_3)}$ and therefore $\underline{F}_z(\underline{z}_{(1-p_3)}) < 1 - p_3 < \overline{F}_z(\overline{z}_{(1-p_3)})$. As before, we neglect the values of p_3 such that $1 - p_3 = \frac{l}{n_z+1}$ for $l \in \{1, \ldots, n_z\}$, for which $z_{l-1} < F_z^{-1}(1 - p_3) < z_{l+1}$ but these do not

affect what follows.

For any (p_1, p_3) as described above, the infimum of the values $ROC_s(p_1, p_3)$, as given by Equation (1), for all ROC surfaces in the set $\mathcal{S}_{roc}$, can be derived as follows (see also Figure 1). We must find the infimum for the NPI-based probability for the event $Y_{n_y+1} \in (\overline{x}_{(p_1)}, \underline{z}_{(1-p_3)})$. This restricts the interval corresponding to the inverse CDFs to be as small as possible. The probability for the event that Y_{n_y+1} is in this interval must be minimised. As usual for NPI lower probabilities, this is achieved by counting the number of intervals (y_{j-1}, y_j) that are totally included in $(\overline{x}_{(p_1)}, \underline{z}_{(1-p_3)})$. We denote the resulting lower envelope at the point (p_1, p_3) by $\underline{ROC}_s^L(p_1, p_3)$, it is presented in Theorem 4.1. The derivation of the upper envelope follows the same reasoning, with first the interval corresponding to the inverse CDFs made as large as possible, that is $(\underline{x}_{(p_1)}, \overline{z}_{(1-p_3)})$, and secondly taking the NPI upper probability for the event that Y_{n_y+1} will be in this interval. This second step is achieved by counting the number of intervals (y_{j-1}, y_j) that have non-empty intersection with $(\underline{x}_{(p_1)}, \overline{z}_{(1-p_3)})$. We denote the resulting upper envelope at the point (p_1, p_3) by $\overline{ROC}_s^U(p_1, p_3)$, it is also presented in Theorem 4.1. No formal proof of this theorem is included, the steps follow the explanation just given, so the theorem applies formally to the values of (p_1, p_3) as described above.

Theorem 4.1. The lower envelope of all NPI-based ROC surfaces in $\mathcal{S}_{roc}$ is

$$\underline{ROC}_s^L(p_1, p_3) = \begin{cases} \underline{F}_y(\underline{z}_{(1-p_3)}) - \overline{F}_y(\overline{x}_{(p_1)}) & \text{if } \underline{F}_y(\underline{z}_{(1-p_3)}) \geq \overline{F}_y(\overline{x}_{(p_1)}) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The upper envelope of all NPI-based ROC surfaces in $\mathcal{S}_{roc}$ is

$$\overline{ROC}_s^U(p_1, p_3) = \begin{cases} \overline{F}_y(\overline{z}_{(1-p_3)}) - \underline{F}_y(\underline{x}_{(p_1)}) & \text{if } \underline{x}_{(p_1)} \leq \overline{z}_{(1-p_3)} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

It is interesting to consider the volumes under these lower and upper envelopes, which we denote by $\underline{VUS}^L$ and $\overline{VUS}^U$, respectively. These are given in Theorem 4.2, the proofs are presented in the appendix.

Theorem 4.2. The volumes under the lower and upper envelopes of all NPI-

based ROC surfaces in $\mathcal{S}_{roc}$ are

$$\underline{VUS}^L = A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I(x_i < y_{j-1} \wedge y_j < z_{l-1}) \tag{7}$$

$$\overline{VUS}^U = A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I(x_{i-1} < y_j \wedge x_{i-1} < z_l \wedge y_{j-1} < z_l) \tag{8}$$

where $A = \frac{1}{(n_x+1)(n_y+1)(n_z+1)}$.

These lower and upper envelopes of all NPI-based ROC surfaces in $\mathcal{S}_{roc}$ are themselves not elements of $\mathcal{S}_{roc}$. This follows easily from the above described construction of these envelopes. For the lower envelope, the minimisation performed to find its value at a specific point (p_1, p_3) involves putting the minimum possible NPI-based probability mass for Y_{n_y+1} in the interval $(\overline{x}_{(p_1)}, \underline{z}_{(1-p_3)})$. This pointwise optimisation gives, for all such points (p_1, p_3), solutions that cannot be obtained simultaneously, particularly because it always minimizes probability mass for Y_{n_y+1} and hence, when all the solutions are taken together, not a total probability of 1 is used for Y_{n_y+1}. Note that with regard to X_{n_x+1} and Z_{n_z+1} this problem does not occur, as all optimisations with regard to the probability distributions for these random quantities have solutions that can be obtained simultaneously (quite obviously, they are achieved by either putting all probability masses as far left or all as far right within their respective intervals). These envelopes are useful, as they adequately describe the whole set of all NPI-based ROC surfaces in $\mathcal{S}_{roc}$, but they are too wide which shows from the fact that the volumes under them, $\underline{VUS}^L$ and $\overline{VUS}^U$, are not generally equal to the infimum and supremum of the volumes under all the NPI-based ROC surfaces in $\mathcal{S}_{roc}$, which will become clear from Section 4.4 and the examples in Section 5.

We wish to identify ROC surfaces corresponding to $\mathcal{S}_{roc}$ such that their VUS values are indeed generally equal to the infimum and supremum of the VUS values for all the ROC surfaces in $\mathcal{S}_{roc}$. We achieve this, with the main concepts and ideas presented in Section 4.4, by focusing on the volumes under the ROC surfaces and their relations to NPI lower and upper probabilities for correctly ordered observations, so for the event $X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1}$. However, as the NPI lower and upper probabilities for such correctly ordered observations have not yet been presented in the literature, they are first derived in Section 4.3.

4.3. NPI lower and upper probabilities for the event $X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1}$

In this section we present the NPI lower and upper probabilities for the event $X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1}$, with notation as introduced in Section 4.1. These NPI lower and upper probabilities for a specific ordering of three such future observations have not yet been presented in the literature, and have applicability to a variety of problems beyond their use in Section 4.4. They are not expressable in a closed analytical form, but are derived as specified in the following theorem.

Theorem 4.3. The NPI lower and upper probabilities for the event $X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1}$ are

$$\underline{P}(X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1}) = A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I(x_i < t^j_{\min} < z_{l-1}) \qquad (9)$$

$$\overline{P}(X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1}) = A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I(x_{i-1} < t^j_{\max} < z_l) \qquad (10)$$

where $A = \frac{1}{(n_x+1)(n_y+1)(n_z+1)}$ and $t^j_{\min}$ ($t^j_{\max}$) is any value belonging to a sub-interval of (y_{j-1}, y_j), for $j = 1, \ldots, n_y+1$, where the sub-intervals are created by the observations from groups X and Z within this interval (y_{j-1}, y_j), such that the probability for the event $X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1}$ is minimal (maximal).

These NPI lower and upper probabilities are, of course, formally derived as the infimum and supremum, respectively, over all precise probabilities for this event corresponding to precise probability distributions for X_{n_x+1} in $\mathcal{P}_x$, for Y_{n_y+1} in $\mathcal{P}_y$, and for Z_{n_z+1} in $\mathcal{P}_z$. The proof of this theorem, given below, contains further explanation of the remaining optimisations that are required to exactly determine these NPI lower and upper probabilities, which is of course related to finding the values $t^j_{\min}$ and $t^j_{\max}$ in this theorem.

Proof: If the probability distributions of the random quantities X_{n_x+1}, Y_{n_y+1} and Z_{n_z+1} were fully known, the probability for the event $X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1}$ could be derived by

$$P(X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1})$$
$$= \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} P\left\{X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1} | X_{n_x+1} \in (x_{i-1}, x_i), Y_{n_y+1} \in (y_{j-1}, y_j)\right.,$$
$$\left. Z_{n_z+1} \in (z_{l-1}, z_l)\right\} \times P(X_{n_x+1} \in (x_{i-1}, x_i)) P(Y_{n_y+1} \in (y_{j-1}, y_j)) P(Z_{n_z+1} \in (z_{l-1}, z_l))$$

So, this holds for all combinations of probability distributions for X_{n_x+1} in $\mathcal{P}_x$, for Y_{n_y+1} in $\mathcal{P}_y$, and for Z_{n_z+1} in $\mathcal{P}_z$, and we need to find the infimum and supremum for this probability over all these combinations.

To derive the NPI lower probability for this event, the probability $1/(n_x+1)$ for X_{n_x+1}, as assigned to each interval in the partition of the real-line created by the observations from group X, is put at the right-end point of each interval. Simultaneously, the probability $1/(n_z+1)$ for Z_{n_z+1}, as assigned to each interval in the partition of the real-line created by the observations from group Z, is put at the left-end point of each interval. It is straightforward to see that these assignments lead to the infimum of this probability over the respective structures $\mathcal{P}_x$ and $\mathcal{P}_z$, for any probability distribution for Y_{n_y+1}; they can be considered to be as 'pessimistic' as possible for the event of interest. This leads to

$$\inf_{\mathcal{P}_x,\mathcal{P}_y,\mathcal{P}_z} P(X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1}) = \frac{1}{(n_x+1)(n_z+1)} \times$$
$$\inf_{\mathcal{P}_y} \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} P(x_i < Y_{n_y+1} < z_{l-1} | Y_{n_y+1} \in (y_{j-1}, y_j)) P(Y_{n_y+1} \in (y_{j-1}, y_j)) \tag{11}$$

Here the infima are with regard to all probability distributions in the respective structures. In the remaining terms in the sum on the right-hand side, the infima with regard to the probability distribution for Y_{n_y+1} in the structure $\mathcal{P}_y$ must still be considered, we return to this shortly but first present the similar steps for the derivation of the NPI upper probability.

To derive the corresponding NPI upper probability the reasoning with regard to groups X and Z is similar, but now of course the probability masses are placed as 'optimistic' as possible for the event of interest. Hence, the probability masses $1/(n_x+1)$ for X_{n_x+1}, as assigned to each interval in the partition of the real-line created by the observations from group X, are assigned to the left-end points of each interval. Simultaneously, the probability masses $1/(n_z+1)$ for Z_{n_z+1}, as assigned to each interval in the partition of the real-line created by the observations from group Z, are assigned to the right-end points of each interval. It is again straightforward to see that these assignments lead to the supremum of this probability over the respective structures $\mathcal{P}_x$ and $\mathcal{P}_z$, for any probability distribution for Y_{n_y+1}. This leads

to

$$\sup_{\mathcal{P}_x,\mathcal{P}_y,\mathcal{P}_z} P(X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1}) = \frac{1}{(n_x+1)(n_z+1)} \times$$
$$\sup_{\mathcal{P}_y} \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} P(x_{i-1} < Y_{n_y+1} < z_l | Y_{n_y+1} \in (y_{j-1}, y_j)) P(Y_{n_y+1} \in (y_{j-1}, y_j)) \tag{12}$$

Of course, the suprema are with regard over all probability distributions in the respective structures, and determining the suprema with regard to the probability distribution for Y_{n_y+1} in the structure $\mathcal{P}_y$ must still be considered.

The remaining steps in this proof concern the optimisation with regard to the probability distribution for Y_{n_y+1} over the NPI-based structure $\mathcal{P}_y$, so how to assign the probability masses $1/(n_y+1)$ within each interval (y_{j-1}, y_j), $j = 1, \ldots, n_y + 1$, for the NPI lower probability and for the NPI upper probability. Let the number of observations from groups X and Z between y_{j-1} and y_j be denoted by n_x^j and n_z^j, respectively. These observations create a partition of the interval (y_{j-1}, y_j) into $n_x^j + n_z^j + 1$ sub-intervals, where the assumption that the data contain no ties again simplifies notation but could be relaxed without affecting the approach substantially. Note that, if there are no observations from groups X and Z in the interval (y_{j-1}, y_j), then the following reasoning still applies with this whole interval being the only 'sub-interval'.

It is easy to see that this optimisation with regard to the probability distribution for Y_{n_y+1} can be achieved by putting the probability mass $1/(n_y+1)$ within an interval (y_{j-1}, y_j) in a single point, say t_{mi}^j related to the infimum and t_{ma}^j related to the supremum. Doing this for all $j = 1, \ldots, n_y + 1$, and using the NPI lower and upper CDFs for X_{n_x+1} and Z_{n_z+1}, the optimisation problem (11) is equivalent to

$$\inf \frac{1}{n_y+1} \sum_{j=1}^{n_y+1} \underline{F}_x(t_{mi}^j)(1 - \overline{F}_z(t_{mi}^j))$$

and the optimisation problem (12) is equivalent to

$$\sup \frac{1}{n_y+1} \sum_{j=1}^{n_y+1} \overline{F}_x(t_{ma}^j)(1 - \underline{F}_z(t_{ma}^j))$$

where the infimum and supremum are with regard to the values t_{mi}^j and t_{ma}^j over all possible sub-intervals of (y_{j-1}, y_j) for each $j \in \{1, \ldots, n_y+1\}$. These

optimisations can be solved by minimising and maximising, respectively, the products within the sums on the right-hand sides. As these lower and upper CDFs are step-functions, these optimisations can be quite easily performed. However, these products are not monotone over the intervals (y_{j-1}, y_j), so careful searches are required. This can be simplified by using the knowledge that the CDFs are non-decreasing step-functions, and the fact that it is irrelevant which specific point within a sub-interval (as created by the x and z observations) is chosen. It is quite straightforward to implement an algorithm for these optimisations, e.g. one can take the mid-point of each sub-interval as candidate point to be t^j_{mi} or t^j_{ma}.

Once these optimisations have been performed, we denote the points to which the probability masses for Y_{n_y+1} in the intervals (y_{j-1}, y_j) are assigned by $t^j_{\min}$ and $t^j_{\max}$, $j = 1, \ldots, n_y + 1$, which completes the proof.

□

We remind the reader that this section was included for the relation between the volume under and ROC surface and the probability for the event considered here. In the following section we define NPI lower and upper ROC surfaces, for which we introduce some further notation. Let F_y^* and F_y^{**} denote the CDFs of the probability distributions created in the optimisation procedure in the proof above, so with probability $1/(n_y + 1)$ assigned to the values $t^j_{\min}$ and $t^j_{\max}$, respectively, for $j = 1, \ldots, n_y + 1$. So, F_y^* is the step-function increasing from 0 to 1 with step $1/(n_y + 1)$ at each $t^j_{\min}$, and F_y^{**} is the step-function increasing from 0 to 1 with step $1/(n_y + 1)$ at each $t^j_{\max}$.

4.4. NPI lower and upper ROC surfaces

In Section 4.2 we presented the lower and upper envelopes of all ROC surfaces in $\mathcal{S}_{roc}$, the set of all ROC surfaces created by combining probability distributions for X_{n_x+1}, Y_{n_y+1} and Z_{n_z+1} in the respective NPI-based structures $\mathcal{P}_x$, $\mathcal{P}_y$ and $\mathcal{P}_z$. However, as these lower and upper envelopes result from pointwise optimisations they are too wide with regard to the set $\mathcal{S}_{roc}$ when the VUS values are considered. It should be emphasized that these envelopes are of interest as they characterize the set $\mathcal{S}_{roc}$ and can e.g. be used to graphically represent this set, as will be done in the examples in Section 5. But it is also interesting to identify surfaces that provide thight bounds to all ROC surfaces in the set $\mathcal{S}_{roc}$ when the VUS values are considered, as these values play an important role for summarizing the quality of the diagnostic tests and for interpreting the ROC surfaces. So, we wish to define ROC surfaces with VUS values equal to the infimum and supremum of the

VUS values for all ROC surfaces in $\mathcal{S}_{roc}$. The equality of the VUS and the probability of correctly ordered observations enables us to define lower and upper ROC surfaces in line with the optimisation procedures in Section 4.3, we will call these the NPI lower and upper ROC surfaces. This is presented in the following definition, using notation introduced in Sections 4.1, 4.2 and 4.3.

Definition 4.1. The NPI lower ROC surface is defined by, for $p_1, p_3 \in [0,1]$,

$$\underline{ROC}_s(p_1,p_3) = \begin{cases} F_y^*(\underline{z}_{(1-p_3)}) - F_y^*(\overline{x}_{(p_1)}) & \text{if } F_y^*(\underline{z}_{(1-p_3)}) \geq F_y^*(\overline{x}_{(p_1)}) \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

The NPI upper ROC surface is defined by, for $p_1, p_3 \in [0,1]$,

$$\overline{ROC}_s(p_1,p_3) = \begin{cases} F_y^{**}(\overline{z}_{(1-p_3)}) - F_y^{**}(\underline{x}_{(p_1)}) & \text{if } \underline{x}_{(p_1)} \leq \overline{z}_{(1-p_3)} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

The volumes under these lower and upper NPI ROC surfaces are given by Theorem 4.4.

Theorem 4.4. Let the volume under the NPI lower ROC surface $\underline{ROC}_s(p_1,p_3)$ be denoted by $\underline{VUS}$, then

$$\underline{VUS} = \underline{P}(X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1})$$

Similarly, let the volume under the NPI upper ROC surface $\overline{ROC}_s(p_1,p_3)$ be denoted by $\overline{VUS}$, then

$$\overline{VUS} = \overline{P}(X_{n_x+1} < Y_{n_y+1} < Z_{n_z+1})$$

The NPI lower and upper probabilities for correctly ordered observations, on the right-hand sides of the equalities in Theorem 4.4, are as presented in Theorem 4.3. Due to the fact that the NPI lower and upper ROC surfaces follow precisely the construction of the NPI lower and upper probabilities in Section 4.3, these results in Theorem 4.4 are logical. For completeness we present a proof directly from Definition 4.1 in the appendix.

From the construction of these NPI lower and upper ROC surfaces, it is easy to see that, for all $0 \leq p_1, p_3 \leq 1$,

$$\underline{ROC}_s^L(p_1,p_3) \leq \underline{ROC}_s(p_1,p_3) \leq \widehat{ROC}_s(p_1,p_3) \leq \overline{ROC}_s(p_1,p_3) \leq \overline{ROC}_s^U(p_1,p_3) \tag{15}$$

This can also be proven directly from the definitions of these surfaces, which is left as an exercise for the reader. Of course, this implies the same ordering for the corresponding volumes under these surfaces,

$$\underline{VUS}^L \leq \underline{VUS} \leq \widehat{VUS} \leq \overline{VUS} \leq \overline{VUS}^U \tag{16}$$

A special case of interest occurs when there are no observations from groups X or Z within an interval (y_{j-1}, y_j), for $j = 1, \ldots, n_y$, or when there are only observations from either group X or group Z within (y_{j-1}, y_j). If there are only observations from group X within (y_{j-1}, y_j), then the probability mass $1/(n_y + 1)$ for Y_{n_y+1} which is assigned to this interval can be put at $t^j_{\min} = y_{j-1}$ ($t^j_{\max} = y_j$) to get the optimal solution to the corresponding optimisation process in Section 4.3, so leading to the NPI lower (upper) probability for the event of correctly ordered next observations and hence also to the NPI lower (upper) ROC surface. Similarly, if we only have observations from group Z within (y_{j-1}, y_j), then the probability mass $1/(n_y + 1)$ for Y_{n_y+1} for this interval can be assigned to $t^j_{\min} = y_j$ ($t^j_{\max} = y_{j-1}$) in order to derive the NPI lower (upper) probability for the event of correctly ordered next observations and also to derive the NPI lower (upper) ROC surface. From these facts it follows that, if the data of the X and Z groups are fully separated, with $x_{n_x} < z_1$, and there is at least one $y_j \in (x_{n_x}, z_1)$, then the NPI lower and upper ROC surfaces introduced in Definition 4.1 are equal to the lower and upper envelopes of $\mathcal{S}_{roc}$ in Theorem 4.1, respectively, and of course also the corresponding volumes under these surfaces are equal. This will be illustrated in Example 5.3 in Section 5.

4.5. The NPI-based optimal decision thresholds for diagnosis

The choice of the decision thresholds c_1 and c_2 is an important aspect of designing the diagnostic method for the three groups case discussed in this paper. One method for this is by maximisation of the Youden's index as given in Equation (4). In the NPI approach, the NPI lower and upper CDFs can be used to get the NPI lower and upper probabilities of correct classifications, which can be combined into NPI lower and upper bounds for Youden's index which are the sharpest possible bounds for all Youden's indices corresponding to probability distributions for X_{n_x+1}, Y_{n_y+1} and Z_{n_z+1} in their respective NPI-based structures $\mathcal{P}_x$, $\mathcal{P}_y$ and $\mathcal{P}_z$. The NPI lower bound for Youden's index is

$$\begin{aligned}\underline{J}(c_1, c_2) &= \underline{P}(X_{n_x+1} \leq c_1) + \underline{P}(c_1 < Y_{n_y+1} \leq c_2) + \underline{P}(Z_{n_z+1} > c_2) \\ &= \underline{F}_x(c_1) + \left\{\underline{F}_y(c_2) - \overline{F}_y(c_1)\right\}^+ + 1 - \overline{F}_z(c_2)\end{aligned}$$

where $\{A\}^+ = \max\{A, 0\}$, and the corresponding NPI upper bound for Youden's index is

$$\begin{aligned}\overline{J}(c_1, c_2) &= \overline{P}(X_{n_x+1} \leq c_1) + \overline{P}(c_1 < Y_{n_y+1} \leq c_2) + \overline{P}(Z_{n_z+1} > c_2) \\ &= \overline{F}_x(c_1) + \overline{F}_y(c_2) - \underline{F}_y(c_1) + 1 - \underline{F}_z(c_2)\end{aligned}$$

If c_1 and c_2 do not coincide with any data observations, then it is straightforward to show that

$$\overline{J}(c_1, c_2) = \underline{J}(c_1, c_2) + \frac{1}{n_x + 1} + \frac{2}{n_y + 1} + \frac{1}{n_z + 1} \tag{17}$$

If either or both of c_1 and c_2 are equal to some data observations, then a similar relation but with fewer terms on the right-hand side is easily derived, but this is of little practical relevance. This constant difference between the NPI upper and lower Youden's indices implies that both will be maximised at the same values of c_1 and c_2. It is further easy to show that, for all c_1 and c_2,

$$\underline{J}(c_1, c_2) \leq \hat{J}(c_1, c_2) \leq \overline{J}(c_1, c_2)$$

where $\hat{J}(c_1, c_2)$ is the empirical estimate of Youden's index, obtained by using the empirical CDFs in Equation (4). Of course, these inequalities do not imply that the empirical estimate of Youden's index is maximal for the same values of c_1 and c_2 as the NPI lower and upper Youden's indices, but in most situations one would expect the maxima to be indeed at the same values and in case of any differences these are likely to be small.

4.6. Upper (lower) bound for the NPI lower (upper) ROC surface

Obtaining the NPI lower and upper ROC surfaces, as introduced in Section 4.4, is not too problematic for small to medium data sets, but the optimisations involved in deriving the values $t^j_{\min}$ and $t^j_{\max}$ for each interval (y_{j-1}, y_j) can lead to much computational effort for large data sets, in particular if there is much overlap between the observations from the three groups. As discussed at the end of Section 4.4, the lower and upper envelopes of all NPI-based ROC surfaces in $\mathcal{S}_{roc}$, as presented in Section 4.2, are identical to the NPI lower and upper ROC surfaces if the data from groups X and Z do not overlap and there is at least one observation from group Y between them, while they are always lower and upper bounds for them as presented in the inequalities in (15). This raises the possibility to avoid the numerical optimisations required to derive the NPI lower and upper ROC surfaces by using the envelopes as approximations, benefiting from the fact that they are available in simple analytical expressions as given in Theorem 4.1. As

the lower envelope provides a lower bound for the NPI lower ROC surface, it will be useful to be able to derive, also without numerical optimisations, an upper bound for this NPI lower ROC surface; together these two bounds will give some further information about the quality of the approximation. Similarly, it is of interest to derive a lower bound for the NPI upper ROC surface. Of course, taking any combination of probability distributions for X_{n_x+1}, Y_{n_y+1} and Z_{n_z+1} in their respective NPI-based structures $\mathcal{P}_x$, $\mathcal{P}_y$ and $\mathcal{P}_z$ will provide such bounds, but it is attractive to find simple forms, so we propose bounds that are constructed in a manner that is closely related to the construction of the lower and upper envelopes in Section 4.2.

To construct a suitable upper bound for the NPI lower ROC surface $\underline{ROC}_s(p_1, p_3)$, which we will denote by $\underline{ROC}_s^U(p_1, p_3)$, we put the probability masses for X_{n_x+1} and Z_{n_z+1} at the same locations as for the lower envelope $\underline{ROC}_s^L(p_1, p_3)$, while for Y_{n_y+1} we take the probability distribution corresponding to the NPI lower CDF $\underline{F}_y$. Similarly, we construct a suitable lower bound for the NPI upper ROC surface $\overline{ROC}_s(p_1, p_3)$, which we will denote by $\overline{ROC}_s^L(p_1, p_3)$, by putting the probability masses for X_{n_x+1} and Z_{n_z+1} at the same locations as for the upper envelope $\overline{ROC}_s^U(p_1, p_3)$, while for Y_{n_y+1} we again take the probability distribution corresponding to the NPI lower CDF $\underline{F}_y$.

Definition 4.2. An upper bound for the NPI lower ROC surface can be defined by

$$\underline{ROC}_s^U(p_1, p_3) = \begin{cases} \underline{F}_y(\underline{z}_{(1-p_3)}) - \underline{F}_y(\overline{x}_{(p_1)}) & \text{if } \underline{F}_y(\underline{z}_{(1-p_3)}) \geq \underline{F}_y(\overline{x}_{(p_1)}) \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

A lower bound for the NPI upper ROC surface can be defined by

$$\overline{ROC}_s^L(p_1, p_3) = \begin{cases} \underline{F}_y(\overline{z}_{(1-p_3)}) - \underline{F}_y(\underline{x}_{(p_1)}) & \text{if } \underline{x}_{(p_1)} \leq \overline{z}_{(1-p_3)} \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

It should be remarked that, in these definitions, one can replace the NPI lower CDF $\underline{F}_y$ for Y_{n_y+1} by the corresponding NPI upper CDF $\overline{F}_y$, this would not make any difference as these upper and lower CDFs only differ by a term $1/(n_y + 1)$ at all of the observations for groups X and Z. Note that, if ties between data from different groups can occur, than the use of $\overline{F}_y$ instead of $\underline{F}_y$ could lead to a very small difference at the values of ties, but this is of very little practical relevance as, of course, using either CDF would still give valid bounds, and also the volume under these surfaces would not be

affected. The volumes under these bounding surfaces are given in Theorem 4.5, the proof of which is given in the appendix.

Theorem 4.5. The volume under the bounding surface $\underline{ROC}_s^U(p_1, p_3)$ is

$$\underline{VUS}^U = A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I(x_i < y_j < z_{l-1}) \tag{20}$$

and the volume under the bounding surface $\overline{ROC}_s^L(p_1, p_3)$ is

$$\overline{VUS}^L = A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I(x_{i-1} < y_j < z_l) \tag{21}$$

where $A = \frac{1}{(n_x+1)(n_y+1)(n_z+1)}$.

Of course, the relationship between the NPI lower ROC surface and its presented bounds is, for all $p_1, p_3 \in [0, 1]$,

$$\underline{ROC}_s^L(p_1, p_3) \leq \underline{ROC}_s(p_1, p_3) \leq \underline{ROC}_s^U(p_1, p_3)$$

and similarly for the NPI upper ROC surface and its presented bounds,

$$\overline{ROC}_s^L(p_1, p_3) \leq \overline{ROC}_s(p_1, p_3) \leq \overline{ROC}_s^U(p_1, p_3)$$

This trivially implies

$$\underline{VUS}^L \leq \underline{VUS} \leq \underline{VUS}^U$$

and

$$\overline{VUS}^L \leq \overline{VUS} \leq \overline{VUS}^U$$

However, while the lower and upper envelopes resulted from pointwise optimisations, the bounding surfaces $\underline{ROC}_s^U(p_1, p_3)$ and $\overline{ROC}_s^L(p_1, p_3)$ are not optimal in a specific sense; they are particularly proposed for ease of computation. It may be possible to propose other general bounds that perform better than these ones and which are also straigthforward to calculate, we have not considered this in great detail. While $\underline{ROC}_s^U(p_1, p_3) \leq \overline{ROC}_s^L(p_1, p_3)$ for all $p_1, p_3 \in [0, 1]$, they do not necessarily bound the empirical ROC surface as will be illustrated in Example 5.3.

5. Examples

In this section, the NPI approach for three-group diagnosis with ROC surfaces, as presented in this paper, is illustrated in three examples. With regard to the graphical illustrations, we restrict the presentation to the lower and upper envelopes $\underline{ROC}_s^L(p_1, p_3)$ and $\overline{ROC}_s^U(p_1, p_3)$ for $\mathcal{S}_{roc}$, together with the empirical ROC surfaces. The NPI lower and upper ROC surfaces $\underline{ROC}_s(p_1, p_3)$ and $\overline{ROC}_s(p_1, p_3)$ are close to the lower and upper envelopes, respectively (any differences are very hard to spot in such figures). This particularly shows through the values of the corresponding volumes under the surfaces, which are reported in each example. In the figures in this section, we denote the surfaces by the notation $\underline{S}_f$ for $\underline{ROC}_s^L(p_1, p_3)$, S_f for $\widehat{ROC}_s(p_1, p_3)$ and $\overline{S}_f$ for $\overline{ROC}_s^U(p_1, p_3)$. Furthermore, in all plots, p_1 and p_3 are both equal to 0 at the right-back corner, while at the left-front corner they are equal to 1, with arrows indicating the direction of increase of their values along the respective axes.

Example 5.1. The data in this example are simulated from 3 Normal distributions as follows: For group X, $n_x = 20$ observations from $N(0, 1)$; for group Y, $n_y = 24$ observations from $N(1, 1.1)$; from group Z, $n_z = 22$ observations from $N(1.3, 1.4)$. The boxplots of these simulated data are presented in Figure 2.

The lower and upper envelopes of the set $\mathcal{S}_{roc}$ of all NPI-based ROC surfaces, $\underline{ROC}_s^L(p_1, p_3)$ and $\overline{ROC}_s^U(p_1, p_3)$ as defined in Section 4.2, and the empirical ROC surface $\widehat{ROC}_s(p_1, p_3)$ are presented in Figure 3.

Although differences are not large and therefore not easy to see in Figure 3, the plots illustrate that the empirical ROC surface is entirely between the lower and upper envelopes. As the volume under an ROC surface (VUS) plays an important role in assessing the quality of the diagnosis, and enables interpretation through its relation with the probability of correctly ordered observations from the different groups as discussed before in this paper, Table 1 presents the values under the surfaces for the seven surfaces that we explicitly discussed: the empirical ROC surface ($\widehat{VUS}$), the lower and upper envelopes of all NPI-based ROC surfaces ($\underline{VUS}^L$ and $\overline{VUS}^U$), the NPI lower and upper ROC surfaces ($\underline{VUS}$ and $\overline{VUS}$), and the upper bound for the NPI lower ROC surface and lower bound for the NPI upper ROC surface ($\underline{VUS}^U$ and $\overline{VUS}^L$).

To interpret these values, it is important to remember that a VUS of about 1/6 would occur if the observations from the three groups would fully overlap, in such a way that the diagnostic method would perform no better than a

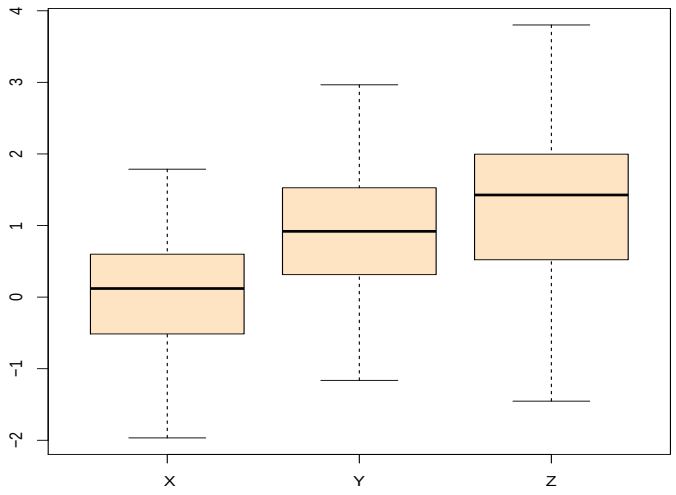
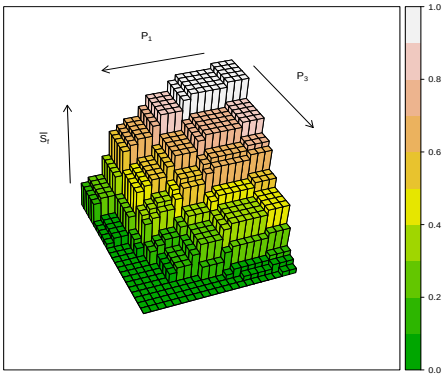


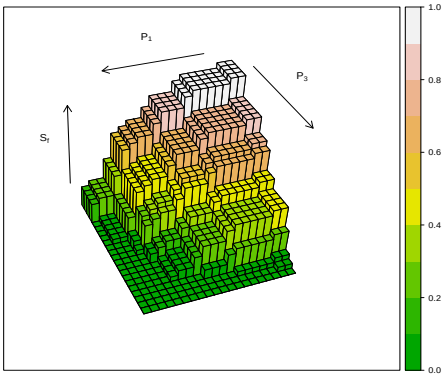
Figure 2: Boxplots for the data in Example 5.1

random allocation of patients to the three groups. As all VUS values are clearly greater than 1/6, this indicates that the diagnostic method is better than a random allocation. However, the VUS values are far away from 1, which would indicate a perfect diagnostic performance. It is clear from Figure 2 that particularly the simulated data from groups Y and Z substantially overlap, which therefore leads to substantial risk of misclassification of new patients from both groups. These VUS values also imply that the NPI lower and upper ROC surfaces are close to the corresponding envelopes, and that the empirical surface is relatively closer to the NPI upper ROC surface than to the NPI lower ROC surface. The upper bound for the NPI lower ROC surface and the lower bound for the NPI upper ROC surface are actually quite a bit further from them than the corresponding envelopes, but they nevertheless could be useful if one would not have gone through the efforts of calculating the NPI lower and upper ROC surfaces exactly, in which case they would, together with the envelopes, provide clear ranges within which the exact surfaces are.

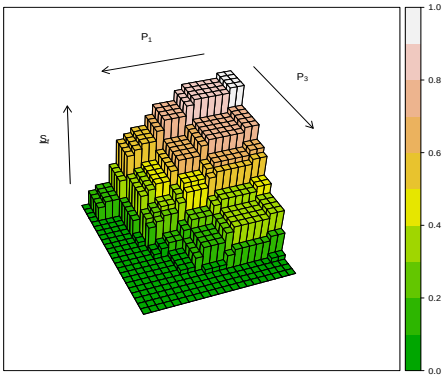
To decide on good values for the decision thresholds c_1 and c_2, which define the specific diagnostic method for the next patient in the NPI approach, we consider the maximum value of Youden's index, as discussed in Section 4.5. The Youden's index corresponding to the empirical ROC surface, $\hat{J}(c_1, c_2)$, has maximum value 1.6258, which occurs at two local maxima, namely at $(c_1, c_2) = (0.1293, 1.7715)$ and at $(c_1, c_2) = (0.7014, 1.7715)$.



(a) Upper envelope



(b) Empirical ROC surface



(c) Lower envelope

Figure 3: Upper and lower envelopes and empirical ROC surface for Example 5.1

Table 1: Volumes under ROC surfaces, Example 5.1

$\widehat{VUS}$	0.3854
$(\underline{VUS}^L, \overline{VUS}^U)$	(0.3091,0.4267)
$(\underline{VUS}, \overline{VUS})$	(0.3101,0.4237)
$(\underline{VUS}^U, \overline{VUS}^L)$	(0.3371,0.3942)

Of course, due to the fact that we are dealing with step-functions, there are ranges of values for (c_1, c_2) close to these reported values for which the Youden's index has the same value, but we just report single points resulting from a straightforward numerical routine which we implemented to find the optima. For the NPI lower and upper surfaces, the maxima of the corresponding Youden's indices $\underline{J}(c_1, c_2)$ and $\overline{J}(c_1, c_2)$ are 1.5151 and 1.6862, respectively, which both occur at $(c_1, c_2) = (0.1293, 1.7715)$, which was also one of the points which led to the maximum Youden's index corresponding to the empirical ROC surface. So these values for c_1 and c_2 would be good values to choose for the decision threshold for the diagnosis of the next patient. These maximum values for $\underline{J}(c_1, c_2)$ and $\overline{J}(c_1, c_2)$ are also in line with the relation (17) that holds generally between them. As values for the Youden's index for three group diagnostic methods that are better than random allocation are between 1 and 3, these values also indicate that the resulting diagnostic method for the next patient will be better than random classification but its performance is not very good. Using these values for c_1 and c_2, a test result less than or equal to 0.1293 means that the patient will be classified as belonging to group X, while a test result greater than 0.1293 and less than or equal to 1.7715 will lead to classification into group Y, and a test result greater than 1.7715 leads to classification into group Z.

Example 5.2. This example uses data from the literature concerning the diagnostic test NAA/Cr which is used to discriminate between different levels of HIV among patients (Chang et al., 2004; Yiannoutsos et al., 2008; Nakas et al., 2010). The data consist of observations for 135 patients, of whom 59 were HIV-positive with AIDS dementia complex (ADC), 39 were HIV-positive non-symptomatic subjects (NAS), and 37 were HIV-negative individuals (NEG) (Nakas et al., 2010; Inacio et al., 2011). The NAA/Cr levels are expected to be lowest among the ADC group and highest among the NEG group, with the NAS group being the intermediate group (Chang et al., 2004) (so in relation to the presentation in this paper, these would be

groups X, Z and Y, respectively). Figure 4 shows the boxplots of these data, from which it is clear that the data of the groups overlap considerably, particular for groups Y and Z. These data are quite similar, from the perspective of overlap among the different groups, to those in Example 5.1, but there are in total about twice as many observations in this example, which will lead to less imprecision as illustrated below. This is an important feature of NPI in general, with the difference between corresponding NPI upper and lower probabilities reflecting the amount of information on which the inferences are based.

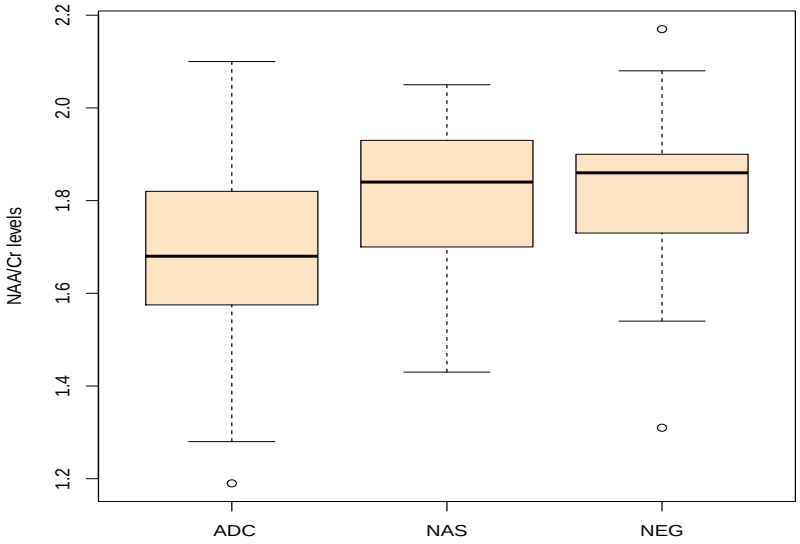


Figure 4: Boxplots of NAA/Cr levels for the ADC, NAS and NEG groups (Example 5.2)

Figure 5 presents the lower and upper envelopes for the set of all NPI-based ROC surfaces, together with the empirical ROC surface. Comparing these plots to Figure 3, we see that the surfaces in this example consists are build up by many more rectangular areas where the surface is constant, which directly reflects that there are more observations in this example than in Example 5.1. This also leads to less imprecision, with the differences between the three plots quite hard to spot. Of course, the empirical ROC surface is again between the two envelopes. The VUS values of the seven surfaces presented in this paper are given in Table 2. They show that the performance of the diagnostic test is worse than for Example 5.1, for which the VUS values were given in Table 1, but at least it is still better than

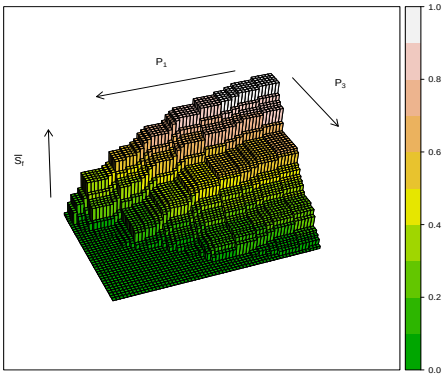
Table 2: Volumes under ROC surfaces, Example 5.2

$\widehat{VUS}$	0.2879
$(\underline{VUS}^L, \overline{VUS}^U)$	(0.2524,0.3131)
$(\underline{VUS}, \overline{VUS})$	(0.2548,0.3087)
$(\underline{VUS}^U, \overline{VUS}^L)$	(0.2688,0.2951)

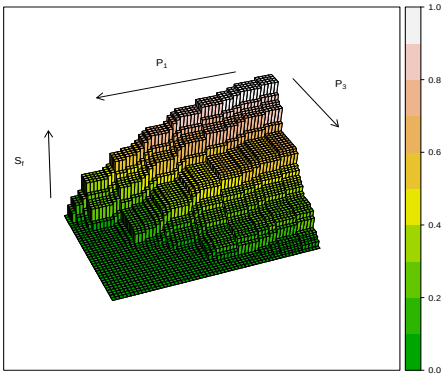
a random classification of the patients would be. The differences between corresponding upper and lower VUS values in this example are smaller than in Example 5.1, reflecting that there is less imprecision for corresponding upper and lower bounds or ROC surfaces as a result of the larger number of observations in this example.

The maximum value of Youden's index corresponding to the empirical ROC surface in this example is equal to 1.4362, which occurs for $(c_1, c_2) = (1.76, 2.05)$. The maximum values for the Youden's indices corresponding to the NPI lower and upper ROC surfaces are $\underline{J}(c_1, c_2) = 1.3803$ and $\overline{J}(c_1, c_2) = 1.4732$, which both occur for the same values of c_1 and c_2 as for the empirical ROC surface. As in the previous example, these maximum values for the Youden's indices indicate that the diagnostic performance of this test for the next patient is likely to be better than random classification, but it is not very good. These values also again illustrate relation (17). The difference between these maximum values for the respective Youden's indices is smaller than in Example 5.1, which of course also reflects that there is less imprecision in this example due to a larger number of observations. Using the maximasation of the Youden's index as criterion for selecting the decision thresholds for diagnosis of a future patient, a test result less than or equal to 1.76 leads to classification into the ADC group, a test result greater than 2.05 leads to classification into the NEG group, and a rest result in between these two values leads to classification into the NAS group. In this example there is small imprecision due to the large data sets.

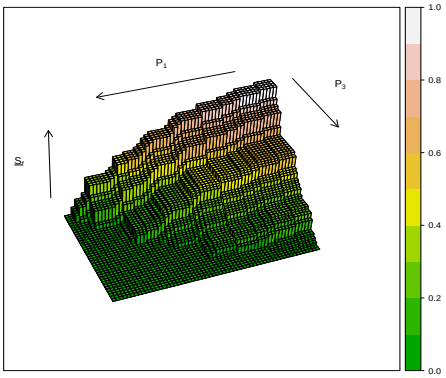
Example 5.3. Examples 5.1 and 5.2 both involved data sets with much overlap between the observations from the three groups, resulting in relatively poor performance of the diagnostic method as reflected through low values for the volumes under the considered surfaces. We present a further example to illustrate the approach for data with most observations for the three groups well separated. There are only 10 observations for each group, so $n_x = n_y = n_z = 10$, they are given in Table 3. It should be emphasized that the data from groups X and Z do not overlap.



(a) Upper envelope



(b) Empirical ROC surface



(c) Lower envelope

Figure 5: Upper and lower envelopes and empirical ROC surface for Example 5.2

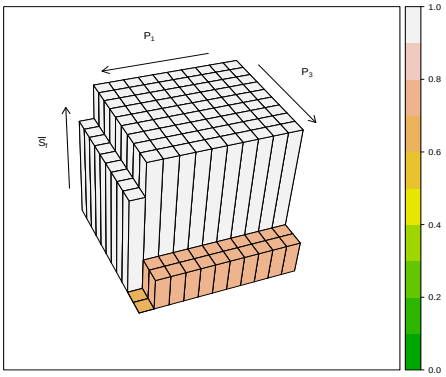
Table 3: Data for Example 5.3

	NAA/Cr levels									
X	1.28	1.43	1.52	1.53	1.55	1.57	1.60	1.63	1.64	1.66
Y	1.65	1.68	1.71	1.76	1.78	1.79	1.80	1.85	1.86	1.87
Z	1.83	1.84	1.88	1.89	1.90	1.93	1.96	1.99	2.06	2.08

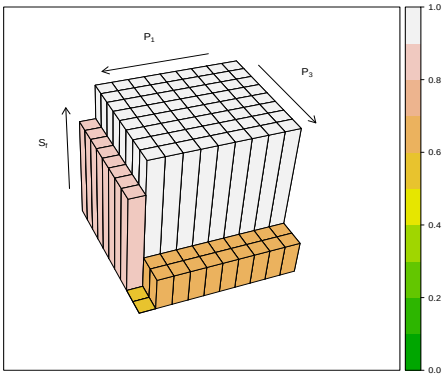
Figure 6 presents the lower and upper envelopes of the set of NPI-ROC surfaces, together with the empirical ROC surface. These surfaces appear to be quite different from the corresponding ones in the earlier examples: The surfaces are made up of far fewer rectangular areas where the surface is constant, which is due to there being far fewer observations. Furthermore, from these plots it can be easier seen that the empirical ROC surface is indeed in between the two envelopes, where it is clear that it is much closer to the upper envelope than to the lower envelope.

The VUS values of the seven surfaces considered in this paper are given in Table 4. These values are considerably larger than for the previous examples, which shows that, not surprisingly, the diagnostic test is likely to perform quite well for a future patient in this case. Of course, this is directly due to the fact that the data hardly overlap and that the NPI approach is strongly based on the data with few further model assumptions. There is substantial imprecision in corresponding upper and lower VUS values, this is due to the relatively small numbers of observations. The VUS corresponding to the empirical ROC surface is close to the VUS corresponding to the NPI upper ROC surface, which reflects the earlier noted fact that the empirical ROC surface is close to the upper envelope. The lower and upper envelopes have the same VUS values as those corresponding to the NPI lower and upper ROC surfaces, and hence these surfaces are identical. This illustrates the general property, in case of data with the observations from groups X and Z not overlapping, that was mentioned at the end of Section 4.4. Note further that $\widehat{VUS} > \overline{VUS}^L$, which shows that the lower bound for the NPI upper ROC surface is not guaranteed to be always above the empirical ROC surface. Of course, in such cases one could just consider the empirical ROC surface as such a lower bound, or indeed it may be provide further motivation for finding another general lower bound for this NPI upper ROC surface that is closer to it and also easy to calculate.

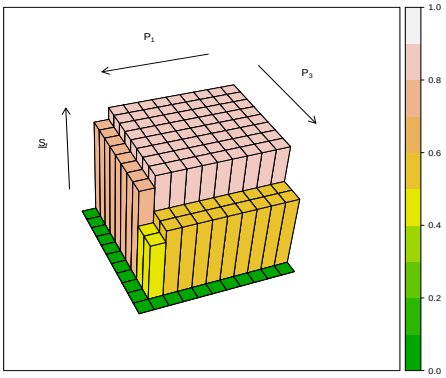
The maximum value of the Youden's index corresponding to the empirical ROC surface is equal to 2.7, which occurs for $(c_1, c_2) = (1.645, 1.875)$. The maximum values of the Youden's indices corresponding to the NPI lower and upper ROC surfaces, $\underline{J}(c_1, c_2)$ and $\overline{J}(c_1, c_2)$, are equal to 2.3636 and 2.7273, respectively, and both occur for the same values of c_1 and c_2 that



(a) Upper envelope



(b) Empirical ROC surface



(c) Lower envelope

Figure 6: Upper and lower envelopes and empirical ROC surface for Example 5.3

Table 4: Volumes under ROC surfaces, Example 5.3

$\widehat{VUS}$	0.9300
$(\underline{VUS}^L, \overline{VUS}^U)$	(0.6236,0.9421)
$(\underline{VUS}, \overline{VUS})$	(0.6236,0.9421)
$(\underline{VUS}^U, \overline{VUS}^L)$	(0.6987,0.8512)

maximize the Youden's index corresponding to the empirical ROC surface. These values are more imprecise than in the previous examples, which reflects the smaller number of observations. These values are also in line with relation (17). Compared to the two previous examples, these values are substantially closer to 3, which is the theoretically maximum possible value that would occur for the Youden's index corresponding to the upper ROC surface in case all observations from the three groups would be perfectly separated, so there would be no overlap at all between the data for the different groups. Even in such a case, the Youden's index corresponding to the lower ROC surface would be less than 3, with the difference depending on the numbers of observations according to relation (17). So, the diagnostic performance of this test for the next patient is likely to be quite good, which is of course no surprise given that the observations in the groups are well separated. With these values for the decision thresholds, a test result of less than or equal to 1.645 leads to the decision to classify the next patient into group X, while a test result greater than 1.645 and less than or equal to 1.875 leads to classification into group Y, and a test result greater than 1.875 to classification into group Z.

6. Concluding remarks

In this paper we introduced the NPI approach for three-group diagnostic tests using the ROC surface. This can be used to asses the accuracy of a diagnostic test, with the NPI setting ensuring, due to its predictive nature, specific focus on the next patient. NPI lower probabilities reflect the evidence in favour of the event of interest, while NPU upper probabilities reflect the evidence against the event of interest. The difference between corresponding upper and lower probabilities reflects that information from finite data is limited. When making decisions about diagnosis for a specific future patient, it seems useful to have the amount of information and the evidence it provides clearly reflected in this way.

Attention has been restricted to real-valued data, developing the related NPI theory for ROC surfaces in case of ordinal data is an interesting topic

for future research. The concepts and ideas presented can be generalized to classification into more than three categories (Waegeman et al., 2008), but the computation of NPI lower and upper ROC hypersurfaces, in line with Section 4.4, will require numerical optimisations that will quickly become complicated for larger data sets with substantial overlap between the observations from different groups. Generalization of the lower and upper envelopes of the set of all NPI-based ROC hypersurfaces is likely to remain feasible with more categories, but it has not yet been studied in detail. In any case, it is likely that heuristic methods to provide approximations to the NPI lower and upper ROC hypersurfaces would be required, where the quality of such approximations, in relation to the computational complexity for their implementation, requires detailed study. It is also important to develop NPI methods for ROC analysis including covariates (Lopez-de Ullibarri et al., 2008; Rodriguez-Alvarez et al., 2011a,b). Research of a general NPI approach for regression-type models is currently in progress, once fully developed we intend to also apply them to ROC scenarios. It is also possible to assume semi-parametric models in ROC analysis (Zhang, 2006; Wan and Zhang, 2008). It looks likely that the NPI approach can be combined with partial parametric model assumptions, which would also enable application to ROC problems, but this has not yet been developed and is left as an important topic for future research. Increasingly, statistical data are high-dimensional, which sets new challenges for analysis of diagnostic accuracy including ROC methods (Adler and Lausen, 2009). NPI has not yet been developed for multi-dimensional data, it is an important research challenge that is likely to require some investigation into the use of additional structural model assumptions due to the curse of dimensionality that generally affects nonparametric methods.

As the NPI approach does not aim at estimating characteristics for an assumed underlying population, but instead explicitly focuses on a future observation, it is quite different in nature to the established statistical approaches, but in practice a predictive formulation may often be natural. NPI for real-valued observation is also available for multiple future observations (Coolen, 2011), where the inter-dependence of these future observations is explicitly taken into account. Development of NPI-based methods for diagnostic accuracy with explicit focus on $m \geq 2$ future observations is an interesting topic for future research, where particularly the strength of the inferences as function of m should be studied carefully, see Coolen and Coolen-Schrijner (2007) for a similar study with focus on the role of m for comparison of groups of Bernoulli data. Typically, for increasing m the imprecision in inferences increases, which is likely to imply that, on the basis of the limited information in available data, a specific choice of diagnostic method including the

important decision thresholds can be inferred to be good for a number of future patients up to a specific value of m, but for larger values of m the evidence in the data would be too weak to make decisions that are strongly supported by the data without further modelling assumptions. This may be important for practical applications, as it might for example guide sequential experiments with new treatments.

We should emphasize that we do not advocate the NPI approach presented here as a replacement of more established methods, but as an interesting alternative approach to important problems which we recommend to be used alongside other methods. If the core results of different methods are quite close, that would provide a strong argument in favour of them, while substantial differences might suggest that further investigation would be beneficial. In particular, as most established statistical methods make stronger modelling assumptions, it would be logical in such cases to consider whether or not such assumptions are supported by the data.

Acknowledgement
We are grateful to Dr. Christos Nakas for stimulating discussions about this topic area and for providing the data used in Example 5.2.

Appendix

We provide the detailed proofs for six volumes under surfaces presented in three theorems in this paper. In these proofs, we use the notation $\{A\}^+ = \max\{A, 0\}$ and $\sum_{p_1}\sum_{p_3}$ to indicate the sums over pairs of values for p_1 and p_3 such that one value for p_1 is taken from each interval $(\frac{i-1}{n_x+1}, \frac{i}{n_x+1})$ for $i = 1, \ldots, n_x + 1$, and one value for p_3 from each interval $(\frac{l-1}{n_z+1}, \frac{l}{n_z+1})$ for $l = 1, \ldots, n_z + 1$. As the function respective ROC surfaces are constant for all values $p_1 \in (\frac{i-1}{n_x+1}, \frac{i}{n_x+1})$ and $p_3 \in (\frac{l-1}{n_z+1}, \frac{l}{n_z+1})$, it does not matter which specific values for p_1 and p_3 within these intervals are actually used in the calculations (e.g. mid-points of the intervals).

Proof of Theorem 4.2

$$
\begin{aligned}
\underline{VUS}^L &= \frac{1}{(n_x+1)(n_z+1)} \sum_{p_1}\sum_{p_3} \underline{ROC}_s^L(p_1, p_3) \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{p_1}\sum_{p_3} \left\{\underline{F}_y(\underline{z}_{(1-p_3)}) - \overline{F}_y(\overline{x}_{(p_1)})\right\}^+ \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{i=1}^{n_x+1}\sum_{l=1}^{n_z+1} \left\{\underline{F}_y(z_{l-1}) - \overline{F}_y(x_i)\right\}^+ \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{i=1}^{n_x+1}\sum_{l=1}^{n_z+1} \left\{\frac{\sum_{j=1}^{n_y} I(y_j \le z_{l-1})}{n_y+1} - \frac{\sum_{j=1}^{n_y} I(y_j \le x_i) + 1}{n_y+1}\right\}^+ \\
&= A \sum_{i=1}^{n_x+1}\sum_{l=1}^{n_z+1} \left\{\sum_{j=1}^{n_y+1} I(y_j \le z_{l-1}) - \sum_{j=1}^{n_y+1} I(y_{j-1} \le x_i)\right\}^+ \\
&= A \sum_{i=1}^{n_x+1}\sum_{j=1}^{n_y+1}\sum_{l=1}^{n_z+1} \left\{I(y_j \le z_{l-1}) - I(y_{j-1} \le x_i)\right\}^+ \\
&= A \sum_{i=1}^{n_x+1}\sum_{j=1}^{n_y+1}\sum_{l=1}^{n_z+1} I\left(y_j \le z_{l-1} \wedge y_{j-1} > x_i\right)
\end{aligned}
$$

$$
\begin{aligned}
\overline{VUS}^U &= \frac{1}{(n_x+1)(n_z+1)} \sum_{p_1} \sum_{p_3} \overline{ROC}_s^U(p_1, p_3) \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{p_1} \sum_{p_3} I(\underline{x}_{(p_1)} \leq \overline{z}_{(1-p_3)}) \left\{ \overline{F}_y(\overline{z}_{(1-p_3)}) - \underline{F}_y(\underline{x}_{(p_1)}) \right\} \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{i=1}^{n_x+1} \sum_{l=1}^{n_z+1} I(x_{i-1} \leq z_l) \left\{ \overline{F}_y(z_l) - \underline{F}_y(x_{i-1}) \right\} \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{i=1}^{n_x+1} \sum_{l=1}^{n_z+1} I(x_{i-1} \leq z_l) \left\{ \frac{\sum_{j=1}^{n_y} I(y_j \leq z_l) + 1}{n_y+1} - \frac{\sum_{j=1}^{n_y} I(y_j \leq x_{i-1})}{n_y+1} \right\} \\
&= A \sum_{i=1}^{n_x+1} \sum_{l=1}^{n_z+1} I(x_{i-1} \leq z_l) \left\{ \sum_{j=1}^{n_y+1} I(y_{j-1} \leq z_l) - \sum_{j=1}^{n_y+1} I(y_j \leq x_{i-1}) \right\} \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I(x_{i-1} \leq z_l) \left\{ I(y_{j-1} \leq z_l) - I(y_j \leq x_{i-1}) \right\} \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I(x_{i-1} \leq z_l) \left\{ I(y_{j-1} \leq z_l \wedge y_j > x_{i-1}) \right\} \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I\left(x_{i-1} < y_j \wedge x_{i-1} \leq z_l \wedge y_{j-1} \leq z_l \right)
\end{aligned}
$$

Proof of Theorem 4.4

$$
\begin{aligned}
\underline{VUS} &= \frac{1}{(n_x+1)(n_z+1)} \sum_{p_1} \sum_{p_3} \underline{ROC}_s(p_1, p_3) \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{p_1} \sum_{p_3} \left\{ F_y^*(\underline{z}_{(1-p_3)}) - F_y^*(\overline{x}_{(p_1)}) \right\}^+ \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{i=1}^{n_x+1} \sum_{l=1}^{n_z+1} \left\{ F_y^*(z_{l-1}) - F_y^*(x_i) \right\}^+ \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{i=1}^{n_x+1} \sum_{l=1}^{n_z+1} \left\{ \frac{\sum_{j=1}^{n_y+1} I(t_{\min}^j \le z_{l-1})}{n_y+1} - \frac{\sum_{j=1}^{n_y+1} I(t_{\min}^j \le x_i)}{n_y+1} \right\}^+ \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} \left\{ I(t_{\min}^j \le z_{l-1}) - I(t_{\min}^j \le x_i) \right\}^+ \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I\left(t_{\min}^j \le z_{l-1} \wedge t_{\min}^j > x_i \right) \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I\left(x_i < t_{\min}^j < z_{l-1} \right)
\end{aligned}
$$

$$
\begin{aligned}
\overline{VUS} &= \frac{1}{(n_x+1)(n_z+1)} \sum_{p_1} \sum_{p_3} \overline{ROC}_s(p_1, p_3) \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{p_1} \sum_{p_3} I(\underline{x}_{(p_1)} \leq \overline{z}_{(1-p_3)}) \left\{ F_y^{**}(\overline{z}_{(1-p_3)}) - F_y^{**}(\underline{x}_{(p_1)}) \right\} \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{i=1}^{n_x+1} \sum_{l=1}^{n_z+1} I(x_{i-1} \leq z_l) \left\{ F_y^{**}(z_l) - F_y^{**}(x_{i-1}) \right\} \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{i=1}^{n_x+1} \sum_{l=1}^{n_z+1} I(x_{i-1} \leq z_l) \left\{ \frac{\sum_{j=1}^{n_y+1} I(t^j_{\max} \leq z_l)}{n_y+1} - \frac{\sum_{j=1}^{n_y+1} I(t^j_{\max} \leq x_{i-1})}{n_y+1} \right\} \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I(x_{i-1} \leq z_l) \left\{ I(t^j_{\max} \leq z_l) - I(t^j_{\max} \leq x_{i-1}) \right\} \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I(x_{i-1} \leq z_l) \left\{ I(t^j_{\max} \leq z_l \wedge t^j_{\max} > x_{i-1}) \right\} \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I\left(x_{i-1} < t^j_{\max} \wedge x_{i-1} \leq z_l \wedge t^j_{\max} \leq z_l \right) \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I\left(x_{i-1} < t^j_{\max} < z_l \right)
\end{aligned}
$$

Proof of Theorem 4.5

$$
\begin{aligned}
\underline{VUS}^U &= \frac{1}{(n_x+1)(n_z+1)} \sum_{p_1} \sum_{p_3} \underline{ROC}_s^U(p_1, p_3) \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{p_1} \sum_{p_3} \left\{ \underline{F}_y(\underline{z}_{(1-p_3)}) - \underline{F}_y(\overline{x}_{(p_1)}) \right\}^+ \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{i=1}^{n_x+1} \sum_{l=1}^{n_z+1} \left\{ \underline{F}_y(z_{l-1}) - \underline{F}_y(x_i) \right\}^+ \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{i=1}^{n_x+1} \sum_{l=1}^{n_z+1} \left\{ \frac{\sum_{j=1}^{n_y} I(y_j \le z_{l-1})}{n_y+1} - \frac{\sum_{j=1}^{n_y} I(y_j \le x_i)}{n_y+1} \right\}^+ \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} \left\{ I(y_j \le z_{l-1}) - I(y_j \le x_i) \right\}^+ \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I\left(y_j \le z_{l-1} \wedge y_j > x_i \right) \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I\left(x_i < y_j < z_{l-1} \right)
\end{aligned}
$$

$$
\begin{aligned}
\overline{VUS}^L &= \frac{1}{(n_x+1)(n_z+1)} \sum_{p_1} \sum_{p_3} \overline{ROC}_s^L(p_1, p_3) \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{p_1} \sum_{p_3} I(\underline{x}_{(p_1)} \leq \overline{z}_{(1-p_3)}) \left\{ \underline{F}_y(\overline{z}_{(1-p_3)}) - \underline{F}_y(\underline{x}_{(p_1)}) \right\} \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{i=1}^{n_x+1} \sum_{l=1}^{n_z+1} I(x_{i-1} \leq z_l) \left\{ \underline{F}_y(z_l) - \underline{F}_y(x_{i-1}) \right\} \\
&= \frac{1}{(n_x+1)(n_z+1)} \sum_{i=1}^{n_x+1} \sum_{l=1}^{n_z+1} I(x_{i-1} \leq z_l) \left\{ \frac{\sum_{j=1}^{n_y} I(y_j \leq z_l)}{n_y+1} - \frac{\sum_{j=1}^{n_y} I(y_j \leq x_{i-1})}{n_y+1} \right\} \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I(x_{i-1} \leq z_l) \left\{ I(y_j \leq z_l) - I(y_j \leq x_{i-1}) \right\} \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I(x_{i-1} \leq z_l) \left\{ I(y_j \leq z_l \wedge y_j > x_{i-1}) \right\} \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I\left(x_{i-1} < y_j \wedge x_{i-1} \leq z_l \wedge y_j \leq z_l \right) \\
&= A \sum_{i=1}^{n_x+1} \sum_{j=1}^{n_y+1} \sum_{l=1}^{n_z+1} I\left(x_{i-1} < y_j < z_l \right)
\end{aligned}
$$

References

Adler, W., Lausen, B., 2009. Bootstrap estimated true and false positive rates and ROC curve. Computational Statistics & Data Analysis 53, 718–729.

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., Salakoski, T., 2011. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. Computational Statistics & Data Analysis 55, 1828–1844.

Arts, G.R.J., Coolen, F.P.A., van der Laan, P., 2004. Nonparametric predictive inference in statistical process control. Quality Technology and Quantitative Management 1, 201–216.

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

Beck, A.C., 2005. Receiver Operating Characteristic surfaces: Inference and Applications. Ph.D. thesis. University of Rochester. Rochester, New York.

van Calster, B., van Belle, V., Vergouwe, Y., Steyerberg, E.W., 2012. Discrimination ability of prediction models for ordinal outcomes: Relationships between existing measures and a new measure. Biometrical Journal 54, 674–685.

Chang, L., Lee, P.L., Yiannoutsos, C.T., Ernst, T., Marra, C.M., Richards, T., Kolson, D., Schifitto, G., Jarvik, J.G., Miller, E.N., Lenkinski, R., Gonzalez, G., Navia, B.A., 2004. A multicenter in vivo proton-mrs study of hiv-associated dementia and its relationship to age. NeuroImage 23, 1336–1347.

Chen, W., Yousef, W., Gallas, B., Hsu, E., Lababidi, S., Tang, R., Pennello, G., Symmans, W., Pusztai, L., 2012. Uncertainty estimation with a finite dataset in the assessment of classification models. Computational Statistics & Data Analysis 56, 1016–1027.

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: Lovric, M. (Ed.), International Encyclopedia of Statistical Science. Springer, pp. 968–970.

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., Troffaes, M.C., Augustin, T., 2011. Imprecise probability, in: Lovric, M. (Ed.), International Encyclopedia of Statistical Science. Springer, pp. 645–648.

Coolen-Maturi, T., Coolen-Schrijner, P., Coolen, F.P.A., 2012a. Nonparametric predictive inference for binary diagnostic tests. Journal of Statistical Theory and Practice 6, 665–680.

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

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

Elkhafifi, F.F., Coolen, F.P.A., 2012. Nonparametric predictive inference for accuracy of ordinal diagnostic tests. Journal of Statistical Theory and Practice 6, 681–697.

Greiner, M., Pfeiffer, D., Smith, R.D., 2000. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Preventive Veterinary Medicine 45, 23–41.

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: Bernando, J.M., DeGroot, M.H., Lindley, D.V., Smith, A. (Eds.), Bayesian Statistics 3. Oxford University Press, pp. 211–241.

Inacio, V., Turkman, A.A., Nakas, C.T., Alonzo, T.A., 2011. Nonparametric bayesian estimation of the the three-way receiver operating characteristic surface. Biometrical Journal 53, 1011–1024.

Lai, C., Tian, L., Schisterman, E., 2012. Exact confidence interval estimation for the youden index and its corresponding optimal cut-point. Computational Statistics & Data Analysis 56, 1103–1114.

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

Mossman, D., 1999. Three-way rocs. Medical Decision Making 19, 78–89.

Nakas, C.T., Alonzo, T.A., 2007. ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering. Biometrics 63, 603–609.

Nakas, C.T., Alonzo, T.A., Yiannoutsos, C.T., 2010. Accuracy and cut-off point selection in three-class classification problems using a generalization of the youden index. Statistics in Medicine 29, 2946–2955.

Nakas, C.T., Yiannoutsos, C.T., 2004. Ordered multiple-class ROC analysis with continuous measurements. Statistics in Medicine 23, 3437–3449.

Nakas, C.T., Yiannoutsos, C.T., 2010. Ordered multiple class receiver operating characteristic (ROC) analysis, in: Chow, S.C. (Ed.), Encyclopedia of Biopharmaceutical Statistics. Informa Healthcare, pp. 929–932.

Rodriguez-Alvarez, M., Roca-Pardinas, J., Cadarso-Suarez, C., 2011a. A new flexible direct ROC regression model: Application to the detection of cardiovascular risk factors by anthropometric measures. Computational Statistics & Data Analysis 55, 3257–3270.

Rodriguez-Alvarez, M., Tahoces, P., Cadarso-Suarez, C., Lado, M., 2011b. Comparative study of ROC regression techniques - applications for the computer-aided diagnostic system in breast cancer detection. Computational Statistics & Data Analysis 55, 888–902.

Schafer, H., 1989. Constructing a cut-off point for a quantitative diagnostic test. Statistics in Medicine 8, 1381–1391.

Shiu, S.Y., Gatsonis, C., 2012. On ROC analysis with nonbinary reference standard. Biometrical Journal 54, 457–480.

Tian, L., Xiong, C., Lai, Y., Vexler, A., 2011. Exact confidence interval estimation for the difference in diagnostic accuracy with three ordinal diagnostic groups. Journal of Statistical Planning and Inference 141, 549–558.

Lopez-de Ullibarri, I., Cao, R., Cadarso-Suarez, C., Lado, M., 2008. Nonparametric estimation of conditional ROC curves: Application to discrimination tasks in computerized detection of early breast cancer. Computational Statistics & Data Analysis 52, 2623–2631.

Waegeman, W., De Baets, B., Boullart, L., 2008. On the scalability of ordered multi-class ROC analysis. Computational Statistics & Data Analysis 52, 3371–3388.

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

Wan, S., Zhang, B., 2008. Comparing correlated ROC curves for continuous diagnostic tests under density ratio models. Computational Statistics & Data Analysis 52, 233–245.

Weichselberger, K., 2000. The theory of interval-probability as a unifying concept for uncertainty. International Journal of Approximate Reasoning 24, 149–170.

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

Wians, F.H.J., Urban, J.E., Keffer, J.H., Kroft, S.H., 2001. Discriminating between iron deficiency anemia and anemia of chronic disease using traditional indices of iron status vs transferrin receptor concentration. American Journal of Clinical Pathology 115, 112–118.

Xanthopoulos, S.Z., Nakas, C.T., 2007. A generalized ROC approach for the validation of credit rating systems and scorecards. The Journal of Risk Finance 8, 481 – 488.

Xiong, C., van Belle, G., Miller, J.P., Yan, Y., Gao, F., Yu, K., Morris, J.C., 2007. A parametric comparison of diagnostic accuracy with three ordinal diagnostic groups. Biometrical Journal 49, 682–693.

Yiannoutsos, C.T., Nakas, C.T., Navia, B.A., 2008. Assessing multiple-group diagnostic problems with multi-dimensional receiver operating characteristic surfaces: Application to proton mr spectroscopy (mrs) in hiv-related neurological injury. Neuroimage 40, 248–255.

Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3, 32–35.

Yousef, W., Kundu, S., Wagner, R., 2009. Nonparametric estimation of the threshold at an operating point on the ROC curve. Computational Statistics & Data Analysis 53, 4370–4383.

Zhang, B., 2006. A semiparametric hypothesis testing procedure for the ROC curve area under a density ratio model. Computational Statistics & Data Analysis 50, 1855–1876.