Nonparametric predictive inference for diagnostic test thresholds

ALABDULHADI, MANAL,HAMAD,M

Nonparametric predictive inference for diagnostic test thresholds

Manal H. Alabdulhadi

A thesis presented for the degree of
Doctor of Philosophy

Department of Mathematical Sciences
Durham University
United Kingdom

March 2018

Nonparametric predictive inference for diagnostic test thresholds

Manal H. Alabdulhadi

Submitted for the degree of Doctor of Philosophy

March 2018

Abstract

Nonparametric Predictive Inference (NPI) is a frequentist statistical method that is explicitly aimed at using few modelling assumptions, with inferences in terms of one or more future observations. NPI has been introduced for diagnostic test accuracy, yet mostly restricting attention to one future observation. In this thesis, NPI for the accuracy of diagnostic tests will be developed for multiple future observations. The present thesis consists of three main contributions related to studying the accuracy of diagnostic tests. We introduce NPI for selecting the optimal diagnostic test thresholds for two-group and three-group classification, and we compare two diagnostic tests for multiple future individuals.

For the two- and three-group classification problems, we present new NPI approaches for selecting the optimal diagnostic test thresholds based on multiple future observations. We compare the proposed methods with some classical methods, including the two-group and three-group Youden index and the maximum area (volume) methods. The results of simulation studies are presented to investigate the predictive performance of the proposed methods along with the classical methods, and example applications using data from the literature are used to illustrate and discuss the methods.

NPI for comparison of two diagnostic tests is presented, assuming the tests are applied on the same individuals from two groups, namely healthy and diseased individuals. We also introduce weights to reflect the relative importance of the two groups.

Declaration

The work in this thesis is based on research carried out in the Department of Mathematical Sciences at Durham University. No part of this thesis has been submitted elsewhere for any degree or qualification.

Acknowledgements

Alhamdullilah, Allah my God, I am truly grateful for the countless blessings you have bestowed on me generally and in accomplishing this thesis especially.

My sincere gratitude and thanks go to my supervisors Prof. Coolen and Dr. Tahani Coolen-Maturi for their unlimited support, expert advice and guidance. I really appreciate their patience, kindness and invaluable suggestions with immense knowledge. I am very grateful that I have great supervisors.

My special thanks to my husband Abdullah for being very supportive and my lovely children for understanding that my work kept me very busy. I am internally grateful to my mother for all her unlimited love and prayers and all my brothers and sisters for encouraging and believing.

I would like to thank the sponsors and supporters of my study, Saudi Arabian Cultural Bureau in London, the Ministry of Higher Education in Saudi Arabia and Qassim University.

Many thanks to the Durham University for offering an enjoyable academic atmosphere and for the facilities that have enabled me to study smoothly.

Also, I would like to thank Prof. Sarafidis for providing the interleukin-6 (IL-6) data [61], and Prof. Nakas for providing the n-acetyl aspartate over creatine (NAA/Cr) data [54].

Contents

Chapter 1

Introduction

Diagnostic tests are often used to differentiate patients between two states, healthy and diseased. The results of the diagnostic test may take two values (binary tests), or real values (continuous tests), or a value in a finite number of ordered categories (ordinal tests) [59]. The focus in this thesis is on tests that yield real-valued results.

Assessing the accuracy of diagnostic tests is crucial in many application areas including medicine and health care. The receiver operating characteristic (ROC) curve is a useful tool to assess the diagnostic test accuracy, and the area under the ROC curve (AUC) is often used as a single global measure of the overall performance of the diagnostics test. For medical applications, it is important to select an appropriate threshold, or differentiation value, such that a person is assessed to be diseased or healthy, depending on ehether their corresponding diagnostic test result is greater than the threshold value or not. Therefore, threshold selection methods have been an active field of study [11, 35, 47, 72]. Several methods for selecting thresholds are based on the ROC curve, including the Youden index [33, 72], closest-to-(0,1) [11, 65], maximum area [47] methods and other methods as discussed by Greiner et al. [35].

In this thesis, we introduce a nonparametric predictive approach, called NPI, for selecting the optimal threshold of a diagnostic test, where the inferences focus on future observations. The NPI method uses a direct predictive method to select an optimal threshold, focusing on a limited number of future individuals. NPI is a frequentist statistical method that is explicitly aimed at using few modelling assumptions, enabled through the use of lower and upper probabilities to quantify uncertainty. NPI has been introduced for many application areas where the predictive nature of this method plays an important role, including reliability, survival analysis, operations research and finance. Restricting attention to one future observation, NPI has been developed for diagnostic test accuracy considering different types of data. For example, Coolen-Maturi et al. [25] introduced NPI for diagnostic test accuracy with binary data, while Elkhafifi and Coolen [32] presented NPI for diagnostic tests with ordinal data. Coolen-Maturi et al. [24, 26] proposed NPI for two- and three-group ROC analysis with continuous data. The results in [32] have been generalised by Coolen-Maturi [21] for three-group ROC analysis with ordinal data. Recently, Coolen-Maturi [22] considered NPI for scenarios where two or more diagnostic tests are combined in order to improve the overall accuracy.

This thesis develops a new NPI approach, based on multiple future individuals, for selecting the optimal diagnostic test threshold for the two-group scenario and also for selecting the two thresholds needed in a three-group scenario. We focus on the two- and three-group classification problems which are the most used in practice. However, the proposed NPI method is straightforward to generalise for a disease with k groups (stages), as will be briefly mentioned in Section 3.8, the concluding remarks.

Classical methods often focus on estimation rather than prediction. The end goal of studying the accuracy of diagnostic tests is to apply these tests on future patients. Thus, it is of interest to consider the use of a predictive inference method. Another issue would be the validity of the underlying assumptions required by some of these classical methods, which are often difficult to justify in practice.

The important difference of the NPI approach compared with the alternatives in the literature is that the inferences are explicitly in terms of a given number of future individuals. In this thesis, we will show that the number of future individuals considered might influence the choice of the optimal thresholds. If one should make a decision for a predetermined number of future patients, the direct prediction of NPI-based inferences in terms of m patients is clearly attractive. We compare our proposed methods with some empirical classical methods, including the empirical Youden index and maximum area methods, as these methods also take only few model assumptions.

This chapter is organised as follows. Section 1.1 presents an introduction to the concepts of the accuracy of diagnostic tests. Section 1.2 introduces methods in the literature for establishing the thresholds. In Section 1.3, we provide a brief introduction to NPI. A detailed outline of this thesis is given in Section 1.4.

1.1 Accuracy of diagnostic tests

In two-group classification, accuracy of a diagnostic test is determined by the ability of a test to distinguish between healthy and diseased individuals. Measuring the accuracy of diagnostic tests is an important goal in medical research. Parametric and nonparametric approaches have been introduced for accuracy of diagnostic tests [59, 73]. Test outcomes can be either binary, continuous or ordinal. The focus in this thesis is on continuous diagnostic tests. Let Y be a continuous random quantity representing the outcome of a diagnostic test. Studying a suitable choice of a value of c, called threshold, is the main objective for the accuracy of diagnostic tests. We assume through this thesis that for a specific value of a threshold $c \in \mathbb{R}$, the test result indicates disease if $Y > c$ ('positive' test results), and if $Y \leq c$ the test result indicates non-disease ('negative' test results) [59]. Sensitivity (Sn) of a diagnostic test is the probability of a positive test result for an individual with the disease, it is also known as True Positive Fraction (TPF). Specificity (Sp) is the probability of a negative test result for an individual without the disease [59]. A diagnostic test is considered ideal if it has both sensitivity and specificity equal to one [59]. The False Positive Fraction (FPF) is the probability of a positive test result for an individual without the disease, so FPF=$1 - Sp$.

Let X be used to refer to the test result for the healthy group and let Y be used to refer to the test result for the diseased group, and let n_x and n_y be the numbers of individuals in the healthy and the diseased groups, respectively. Let the FPF and TPF corresponding to the threshold c be FPF(c) and TPF(c), respectively, so

$$\begin{aligned} TPF(c) &= P[Y > c] \qquad (1.1) \\ FPF(c) &= P[X \geq c] \qquad (1.2) \end{aligned}$$

The Receiver Operating Characteristic (ROC) curve plots TPF(c) versus FPF(c) over all possible diagnostic thresholds $c \in \mathbb{R}$. The ROC curve has become a popular statistical tool for assessing the accuracy of a diagnostic test. The ROC curve can be defined as

$$ROC = \{(FPF(c), TPF(c)), c \in (-\infty, \infty)\} \qquad (1.3)$$

A perfect diagnostic test completely distinguishes between healthy and diseased individuals for a particular threshold $c^\star$, so FPF($c^\star$) $= 0$ and TPF($c^\star$) $= 1$. In contrast, the diagnostic test has no ability to separate individuals with and without disease if FPF(c)=TPF(c) for all $c \in \mathbb{R}$ [59].

The ROC curve depends on the distributions of X and Y, however these distributions are usually unknown. A nonparametric empirical approach has been introduced for estimating the ROC curve for a diagnostic test with continuous results [59]. This approach is commonly used due to its flexibility to adjust entirely to the available data [59, 73]. The corresponding ROC curve is called the empirical ROC curve.

To introduce the empirical ROC curve, we use the following notation. Suppose that we have test data on n_x individuals from a healthy group and n_y individuals from a disease group, denoted by $\{x_j, j = 1, ..., n_x\}$ and $\{y_i, i = 1, ..., n_y\}$, respectively. Assume that these two groups are fully independent, in the sense that any information about measurements on individuals in one group does not contain any information about measurements on individuals in the other group. For the empirical ROC approach, these observations for both groups are assumed to be realisations of random quantities that are identically distributed as X for the healthy group, and as Y for the disease group. The empirical ROC curve is defined by [59]

$$ROC_e = \{(FPF_e(c), TPF_e(c)), c \in (-\infty, \infty)\} \qquad (1.4)$$

with

$$TPF_e(c) = \frac{\sum_{i=1}^{n_y} 1[y_i > c]}{n_y} \quad (1.5)$$

$$FPF_e(c) = \frac{\sum_{j=1}^{n_x} 1[x_j \geq c]}{n_x} \quad (1.6)$$

where $1[A]$ is the indicator function which is equal to 1 if A is true and 0 otherwise.

The area under the ROC curve, AUC, is a global measure of the overall ability of the diagnostic test to distinguish among those individuals with and without disease, which has been widely studied in the literature [42, 59, 73]. It is equal to the probability that a randomly selected individual from the diseased group has a test result that is higher than that of a randomly chosen individual from the healthy group, so $P[Y > X]$ [59]. The maximum possible value of the AUC is 1, which indicates an ideal test, and AUC$= 0.5$ indicates an uninformative test [59, 73]. Pepe [59] and Zhou et al. [73] presented overviews of statistical methodology for diagnostic test accuracy and ROC curve, considering parametric and nonparametric methods of inference on the ROC curve. The ROC curve has been applied in a variety of areas such as medical imaging and radiology [48], credit scoring [9], psychiatry [40] and epidemiology [3].

1.2 Methods for selection of a threshold

To completely define a diagnostic test, selecting the optimal threshold is needed such that the test provides good differentiation of the individuals with and without the disease. Methods for the selection of the optimal threshold based on the ROC analyses have been discussed by Greiner et al. [35] and Schäfer [62]; one of these methods is to maximise the Youden index (YI) [33, 72]. Formally, the Youden index is defined as

$$\text{YI} = \max_c \{Sn(c) + Sp(c) - 1\} \quad (1.7)$$

Geometrically, YI represents the maximum vertical distance between the ROC curve and the diagonal line. The empirical estimate of the Youden index (EYI) is given by

$$\text{EYI}(c) = \frac{1}{n_x} \sum_{i=1}^{n_x} \mathbf{1}\{x_i \leq c\} + \frac{1}{n_y} \sum_{j=1}^{n_y} \mathbf{1}\{y_j > c\} - 1 \quad (1.8)$$

where perfect separation of the two groups results in EYI= 1 whereas complete overlap yields EYI= 0 [64].

In medical applications, the Youden index is presented as a useful measure for evaluating the diagnostic test procedures. For example, Aoki et al. [3] identified the optimal threshold level of serum pepsinogens for gastric cancer screening using the Youden index. They suggested that the Youden index is useful for identifying the optimal threshold level of serum pepsinogens for gastric cancer screening. Pekkanen and Pearce [58] examined the assessment between bronchial hyperresponsiveness (BHR) and symptom questionnaires of discriminating between asthma and nonasthma by computing the Youden index. The results showed that the symptom questionnaires have a higher Youden index, which could be considered more accurate than BHR. Demir et al. [30] applied the Youden index to measure and compare the assessment of eight discrimination indices in differentiating between thalassemia and iron deficiency anemia (IDA). First, they calculated eight discrimination indices in a number of patients with IDA and a number of patients with thalassemia, then they applied the Youden index for each index to determine which is the best for differentiating thalassemia from IDA. The Youden index was shown to be useful to obtain accurate indices in differentiating thalassemia from IDA. Jalali and Rezaie [41] compared the predicting pressure ulcer risk (PrUs) validity of 4 commonly used PrUs assessment tools using the Youden index as measure of validity between them.

There is a recognizable large body in literature of the Youden index, which addresses other issues such as the estimation of the Youden index and its optimal threshold [33, 43, 50, 63, 64]. This is not directly related to our work.

Another approach for establishing the optimal threshold is the closest-to-(0,1) method (MD). This method selects the optimal threshold that corresponds to the point on the curve closest to (0,1) (i.e. the point closest to perfection with $Se(c) = 1$ and $Sp(c) = 1$). The optimal threshold is the value that minimises the distance between a point on the curve and (0,1) point. This method can be found mathematically by

$$\text{MD} = \min_{c}\{\sqrt{(1 - Sp(c))^2 - (1 - Sn(c))^2}\} \tag{1.9}$$

Perkins and Schisterman [60] discussed a comparison of optimal thresholds selected by this method and the Youden index method. They recommend the use of the Youden index as it offers clear clinical meaning in terms of the probability of correct classification rate. In the literature, the closest-to-(0,1) method has received little attention compared to the Youden index [60].

Recently, Liu [47] proposed an alternative to these methods based on the concept of the area under the ROC curve (AUC), which is the maximum area method (MA). This method defines the optimal threshold as the point that maximising the product of specificity and sensitivity, given by

$$\mathrm{MA} = \max_{c}\{Sp(c) \times Sn(c)\} \tag{1.10}$$

Liu [47] also discussed a comparison of optimal thresholds selected by this method, the Youden index and the closest-to-(0,1) methods, via a simulation study. The maximum area criterion has a simple and more meaningful maximising function, which evaluates the classification accuracy of binary classification at threshold c. The empirical estimator for the maximum area method (EMA) is given by

$$\mathrm{EMA}(c) = \frac{1}{n_x}\sum_{i=1}^{n_x}\mathbf{1}\{x_i \leq c\} \times \frac{1}{n_y}\sum_{j=1}^{n_y}\mathbf{1}\{y_j > c\} \tag{1.11}$$

Several other approaches for selecting the optimal threshold based on the ROC curve are discussed by [35, 59, 62, 66]. For example, Unal [66] proposed an approach called Index of Union (IU). In this method the value of AUC is computed first, then we search for a threshold c from the coordinates of the ROC curve whose specificity and sensitivity values are simultaneously very close or equal to the value of AUC. Mathematically, the IU method can be defined by the following equation

$$IU = \min_{c}(|Se(c) - AUC| + |Sp(c) - AUC|) \tag{1.12}$$

such that the optimal threshold c can be found by minimising the IU(c) function [66]. A different method for the optimal threshold selection, which is not based on the ROC curve, employs the use of a maximally selected statistics that maximises a measure of difference among the two groups [10, 39, 49]. For example, the minimum P value method (min P) presented by Miller and Siegmund [49], defines the optimal threshold that maximises the standard chi-square statistic with one degree of freedom. In Section 2.5 we will compare our proposed NPI method with the EYI method, Equation (1.7), and EMA method, Equation (1.11), since both methods also take only few model assumptions. It is of interest to compare our NPI approach with, for example, IU and min P methods, but we leave that for further research.

1.3 Nonparametric Predictive Inference

1.3.1 A brief introduction

Nonparametric Predictive Inference (NPI) is a frequentist statistical framework based on Hill's assumption $A_{(n)}$ [37], which yields direct probabilities for one or more future observations, based on n observations for related random quantities. $A_{(n)}$ does not assume anything else and it can be considered as a post-data assumption related to exchangeability. Inferences based on $A_{(n)}$ are nonparametric and predictive, and can be considered appropriate if there is hardly any information or knowledge about the random quantities of interest, other than the n observations [38]. Such inferences based on limited knowledge have also been called 'low structure' predictive inferences [34].

The assumption $A_{(n)}$ partially specifies a predictive probability distribution for one future observation as follows. Suppose that $X_1, \ldots, X_n, X_{n+1}$ are continuous, real-valued and exchangeable random quantities. Suppose that the ordered observations of $X_1, \ldots, X_n$ are denoted by $x_1 < x_2 < ... < x_n$, and define $x_0 = -\infty$ and $x_{n+1} = \infty$ for ease of notation (or $x_0 = 0$ when dealing with non-negative random quantities). We assume that ties do not occur between the data observations; ties can be dealt with by assuming that tied observations differ by small amounts, a common approach to break ties in statistics [38]. These n observations partition the real-line into $n+1$ intervals $I_j = (x_{j-1}, x_j)$, for $j = 1, 2, \ldots, n+1$. The assumption $A_{(n)}$ is that the future observation X_{n+1} is equally likely to fall in any of these intervals with probability $\frac{1}{n+1}$ [14], for each $j = 1, \ldots, n+1$,

$$P(X_{n+1} \in I_j) = \frac{1}{n+1} \tag{1.13}$$

NPI has been introduced as predictive methodology, based only on the $A_{(n)}$ assumption. It is important to emphasize that no further assumptions are made on the distribution of probability $\frac{1}{n+1}$ within an interval I_j. In NPI uncertainty is quantified by lower and upper probabilities for events of interest. Augustin and Coolen [6] introduced predictive lower and upper probabilities based on $A_{(n)}$, which are in line with De Finetti's fundamental theorem of probability [29]. The lower probability $\underline{P}(.)$ and upper probability $\overline{P}(.)$ for the event $X_{n+1} \in B$ with $B \subset \mathbb{R}$, based on the intervals $I_j = (x_{j-1}, x_j)$ for $j = 1, 2, \ldots, n+1$, created by n real-valued non-tied observations, and the assumption $A_{(n)}$, are given by

$$\underline{P}(X_{n+1} \in B) = \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbf{1}\{I_j \subseteq B\} \tag{1.14}$$

$$\overline{P}(X_{n+1} \in B) = \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbf{1}\{I_j \cap B \neq \emptyset\} \tag{1.15}$$

The lower probability (1.14) is achieved by taking only probability mass into account that is necessarily within B, which is only the case for the probability mass $\frac{1}{n+1}$ per interval I_j if this interval is completely contained within B. The upper probability (1.15) is achieved by taking all the probability mass into account that could possibly be within B, which is the case for the probability mass $\frac{1}{n+1}$, per interval I_j, if the intersection of I_j and B is non-empty. NPI has strong consistency properties in the theory of interval probability [6, 69], and it never leads to results that are in conflict with inference based on empirical distributions.

NPI has been introduced for a variety of data types, NPI for multinomial data with an unknown number of unordered categories was presented by Coolen and Augustin [15] and Baker [7]. Elkhafifi and Coolen [32] presented NPI for ordinal data, based on a latent variable representation with the categories represented by intervals on the real line to reflect the known ordering of the categories. NPI for right-censored data was introduced by Coolen and Yan [19, 20]. In Chapters 2 and 3, we apply NPI for future order statistics as presented by Coolen et al. [16] and Alqifari [2], and in Chapter 4 we apply NPI for

Bernoulli data introduced by Coolen [13].

1.3.2 NPI for future order statistics

In Section 1.3 NPI was only introduced for one future observation, but it can also be generalized for multiple future observations, where we are interested in $m \geq 1$ future observations, X_{n+i} for $i = 1, \ldots, m$. It is important to emphasize that the future observations X_{n+i} are assumed to derive from the same data collection process as the n data observations. We link the data and future observations via Hill's assumption $A_{(n)}$ [37], or more precisely, via consecutive application of $A_{(n)}, A_{(n+1)}, \ldots, A_{(n+m-1)}$, we refer to these all together as $A_{(.)}$, which can be considered as a post-data version of a finite exchangeability assumption for $n+m$ random quantities. $A_{(.)}$ implies that all possible orderings of the n data observations and the m future observations are equally likely, where the n data observations are not distinguished among each other, and neither are the m future observations. Let $S_j = \#\{X_{n+i} \in I_j,\ i = 1, \ldots, m\}$, then assuming $A_{(.)}$ we have [16]

$$P(\bigcap_{j=1}^{n+1} \{S_j = s_j\}) = \binom{n+m}{n}^{-1} \tag{1.16}$$

for any non-negative integers s_j with $\sum_{j=1}^{n+1} s_j = m$. Equation (1.16) implies that all $\binom{n+m}{n}$ orderings of m future observations among the n observations are equally likely.

The probability distribution of a single order statistic of m future observations is important in this thesis which will be used in Chapters 2 and 3. Let $X_{(r)}$, for $r = 1, \ldots, m$, be the r-th ordered future observation, so $X_{(r)} = X_{n+i}$ for one $i = 1, \ldots, m$ and $X_{(1)} < X_{(2)} < \ldots < X_{(m)}$. The following probabilities are derived by counting the relevant orderings, and hold for $j = 1, \ldots, n+1$ and $r = 1, \ldots, m$ [16]

$$P(X_{(r)} \in I_j) = \binom{j+r-2}{j-1}\binom{n-j+1+m-r}{n-j+1}\binom{n+m}{n}^{-1} \tag{1.17}$$

For this event NPI provides a precise probability, as each of the $\binom{n+m}{n}$ equally likely orderings of n past and m future observations has the r-th ordered future observation in precisely one interval I_j [16]. Generally, consider the event $X_{(r)} \in B$, where $B \subset \mathbb{R}$. NPI provides bounds for the probability for such an event, where the maximum lower bound and minimum upper bound are the lower and upper probabilities, respectively [5, 6, 68, 69]. Following Equations (1.14) and (1.15) in Section 1.3, we can derive the lower and upper probabilities

$$\underline{P}(X_r \in B) = \sum_{j=1}^{n+1} \mathbf{1}\{I_j \subseteq B\} P(X_{(r)} \in I_j) \tag{1.18}$$

$$\overline{P}(X_r \in B) = \sum_{j=1}^{n+1} \mathbf{1}\{I_j \cap B \neq \emptyset\} P(X_{(r)} \in I_j) \tag{1.19}$$

The event that the number of future observations in an interval (x_a, x_b), with $1 \leq a < b \leq n+1$ and denoted by $S_{a,b}^m$, is greater than or equal to a particular value $v \in \mathbb{N}$, has the following precise probability [2],

$$P(S_{a,b}^m \geq v) = \sum_{i=v}^{m} \binom{n+m}{n}^{-1} \binom{b-a-1+i}{i} \binom{n-b+a+m-i}{m-i} \tag{1.20}$$

Equation 1.20 will be used in Chapter 3.

1.3.3 NPI for Bernoulli quantities

Coolen [13] presented NPI for Bernoulli quantities, which is based on the $A_{(.)}$ assumption, for m future observations given n observed values, and a latent variable representation of Bernoulli quantities represented as observations on the real line, with a threshold such that observations to one side are successes and to the other side failures. Suppose that there is a sequence of $n+m$ exchangeable Bernoulli trials, each with ‘success’ and ‘failure’ as possible outcomes, and data consisting of s successes in n trials. Let Y_1^n denote the random number of successes in trials 1 to n; then a sufficient representation of the data for NPI is $Y_1^n = s$, due to assumed exchangeablility of all trials. Let Y_{n+1}^{n+m} denote the random number of successes in trials $n+1$ to $n+m$. Coolen and Coolen-Schrijner [18] presented the lower and upper probabilities for events $Y_{n+1}^{n+m} \geq y$ and $Y_{n+1}^{n+m} < y$. The upper probabilities for these events are as follows. For $y \in \{0, 1, ..., m\}$ and $0 < s < n$,

$$\overline{P}(Y_{n+1}^{n+m} \geq y|Y_1^n = s) = \binom{n+m}{n}^{-1}\left[\binom{s+y}{s}\binom{n-s+m-y}{n-s} + \sum_{l=y+1}^{m}\binom{s+l-1}{s-1}\binom{n-s+m-l}{n-s}\right] \tag{1.21}$$

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

$$\overline{P}(Y_{n+1}^{n+m} < y|Y_1^n = s) = \binom{n+m}{n}^{-1}\left[\binom{n-s+m}{n-s} + \sum_{l=1}^{y-1}\binom{s+l-1}{s-1}\binom{n-s+m-l}{n-s}\right] \tag{1.22}$$

The corresponding lower probabilities can be derived via the conjugacy property [13],

$$\underline{P}(Y_{n+1}^{n+m} \geq y|Y_1^n = s) = 1 - \overline{P}(Y_{n+1}^{n+m} < y|Y_1^n = s)$$

$$\underline{P}(Y_{n+1}^{n+m} < y|Y_1^n = s) = 1 - \overline{P}(Y_{n+1}^{n+m} \geq y|Y_1^n = s)$$

For $m = 1$, the two non-trivial values of these upper probabilities are $\overline{P}(Y_{n+1}^{n+1} \geq 1|Y_1^n = s) = (s+1)/(n+1)$ and $\overline{P}(Y_{n+1}^{n+1} < 1|Y_1^n = s) = (n-s+1)/(n+1)$.

If the observed data are all successes, so $s = n$, or all failures, so $s = 0$, then these upper probabilities are, for all $y \in \{0, 1, ..., m\}$,

$$\overline{P}(Y_{n+1}^{n+m} \geq y|Y_1^n = n) = 1,$$

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

and for all $y \in \{0, 1, ..., m+1\}$,

$$\overline{P}(Y_{n+1}^{n+m} < y|Y_1^n = n) = \frac{\binom{n+y-1}{n}}{\binom{n+m}{n}},$$

$$\overline{P}(Y_{n+1}^{n+m} < y|Y_1^n = 0) = 1.$$

The results in this section will be used in Chapter 4.

1.4 Outline of thesis

This thesis is organized as follows. In Chapter 2, we introduce NPI for selecting the optimal diagnostic test threshold with two groups, healthy or diseased individuals, taking into account a fixed number of future individuals per group. We also introduce NPI method related to the two-group Youden index. Chapter 3 extends the NPI methods to three ordered groups of test outcomes. We further present NPI method related to the three-group Youden index. We investigate the performance of the two- and three-group NPI methods via simulation studies.

The results in Chapters 2 and 3 have been presented at the International Conference of the Royal Statistical Society (RSS) at Manchester University in September 2016, and at the 9th International Conference of the ERCIM WG on Computational and Methodological Statistics and 10th International Conference on Computational and Financial Econometrics (CFE-CMStatistics) at University of Seville, Spain in December 2016. The results of Chapters 2 and 3 are included in the paper “ Nonparametric predictive inference for diagnostic test thresholds”, which is in submission.

Chapter 4 presents a comparison of two diagnostic tests applied on the same individuals from two groups, healthy and diseased individuals, based on NPI for future order statistics and also based on NPI for Bernoulli quantities. Further, to reflect the relative importance of the groups, weights are added. This chapter has been presented at the Research Students' Conference in Probability and Statistics in Durham in April 2017. A journal paper representing the results in Chapter 4 is being prepared for submission. Chapter 5, provides some concluding remarks.

Chapter 2

NPI for two-group diagnostic test threshold

2.1 Introduction

The goal in a two-group classification study is to measure the ability of a diagnostic test to differentiate individuals with the disease of interest ('positive' test results) from those without the disease ('negative' test results). The critical point in measuring the accuracy of a diagnostic test is to select an optimal threshold to identify the positive and negative test results. There is a recognisable inverse relationship between the specificity and sensitivity, meaning that shifting the threshold leads to increasing one of these while decreasing the other. Selecting a classification threshold c usually leads to two different kinds of misclassification, as healthy individuals maybe classified as diseased, and diseased individuals maybe classified as healthy. Ideally, one would choose an optimal c, which effectively reflects one's belief of which group is more important to be correctly diagnosed.

Researchers in the literature use the utility concept, for example Hand [36] discussed the choice of c if one believes that misclassifying a healthy person as diseased is a more serious error than misclassifying a diseased person as healthy, or vice versa. In this chapter, we introduce NPI for selecting the optimal diagnostic test threshold for two-group classification settings, where the inference is based on multiple future individuals.

We present a direct criterion for introducing the relative importance of the two groups.

It is important to discuss a general feature in the NPI approach, which is for small number of future observations, there is relatively more variability in the values than for large m. This is close in nature to the classical situation covered by the central limit theorem, except in NPI where we do not assume an underling population, therefore we also do not use characteristics of population such as mean value. We will refer to this feature latter in this thesis as *randomness effect*.

Section 2.2 introduces NPI for selecting the optimal threshold for two-group diagnostic tests. In Section 2.3, we also introduce a NPI method related to the two-group Youden index. Section 2.4 discusses a property of searching for the optimal threshold. Section 2.5 presents some examples to illustrate and discuss the new approaches. We compare and investigate the performance of the two-group NPI methods and some classical methods via a simulation study in Section 2.6. Finally, some concluding remarks are made in Section 2.7.

2.2 NPI for two-group diagnostic test threshold

Assume that we have real-valued data from a diagnostic test on individuals from two groups, and there are n_x observations from the healthy group X and n_y observations from the disease group Y. Throughout this thesis it is assumed that these two groups are fully independent, in the sense that any information about the individuals in one group does not contain any information about the individuals in the other group. The ordered data of groups X and Y are denoted by $x_1 < x_2 < \ldots < x_{n_x}$ and $y_1 < y_2 < \ldots < y_{n_y}$, respectively. For ease of presentation, we define $x_0 = y_0 = -\infty$ and $x_{n_x+1} = y_{n_y+1} = \infty$. These n_x observations partition the real-line into $n_x + 1$ intervals $I_i^X = (x_{i-1}, x_i)$, for $i = 1, 2, \ldots, n_x + 1$, and the n_y observations partition the real-line into $n_y + 1$ intervals $I_j^Y = (y_{j-1}, y_j)$, for $j = 1, \ldots, n_y + 1$. In this section, we consider m_x future individuals from group X, with diagnostic test results X_{n_x+r}, $r = 1, \ldots, m_x$, and m_y future individuals from group Y, with diagnostic test results Y_{n_y+s}, $s = 1, \ldots, m_y$. Let the m_x and m_y

ordered future observations from groups X and Y be denoted by $X_{(1)} < X_{(2)} < \ldots < X_{(m_x)}$ and $Y_{(1)} < Y_{(2)} < \ldots < Y_{(m_y)}$, respectively.

Small values of the diagnostic test results are assumed to be associated with absence of the disease and large values of the test results with presence of the disease. To this end, a threshold $c \in \mathbb{R}$ can be used to classify individuals to either being healthy (absence of the disease) if their test result is below or equal to the threshold c, or having the disease if their test result is greater than the threshold c. Then the main question is how to find or select the optimal threshold c that maximizes the correct classification of patients and healthy people. As the NPI-based inferences are in terms of future observations, we will select the value c that gives the best classification based on the m_x and m_y future individuals. To this end, we will make use of NPI for future order statistics as summarized in Section 1.3.2, but first we need to introduce further notation.

For a specific value of c, C_c^X denotes the number of correctly classified future individuals from the healthy group X, that is those with test results $X_{n_x+r} \leq c$ (for $r = 1, \ldots, m_x$), and C_c^Y denotes the number of correctly classified future individuals from the disease group Y, that is those with test results $Y_{n_y+s} > c$ (for $s = 1, \ldots, m_y$). Let α and β be any two values in $(0, 1]$ that are selected to reflect the desired importance of one group over another. We consider the aim that the number of correctly classified future individuals of the healthy group X is at least αm_x, and that the number of correctly classified future individuals of the disease group Y is at least βm_y. To gain intuitive insight, varying the values of α and β will depend on one's believes of which group is more important to be correctly diagnosed, for example, if giving medication to diseased patients is crucial, yet does not have serious adverse effects for healthy people, one can take the value of β higher than the value of α. This would be expected to lead to a higher proportion of diseased persons being correctly diagnosed than healthy persons. Of course one can choose α and β to be equal if one prefers to give the same importance of correct classification of the future individuals to both groups. This criterion in terms of the proportions of successful diagnoses seems to be sensible from predictive perspective. Note that α and β are target proportions per group, hence their is no constraint on their values except being in $(0, 1]$.

As the two groups are assumed to be independent, the joint NPI lower and upper probabilities can be derived as the products of the corresponding lower and upper probabilities for the individual events that involve C_c^X and C_c^Y, thus

$$\underline{P}(C_c^X \geq \alpha m_x, C_c^Y \geq \beta m_y) = \underline{P}(C_c^X \geq \alpha m_x) \times \underline{P}(C_c^Y \geq \beta m_y) \tag{2.1}$$

$$\overline{P}(C_c^X \geq \alpha m_x, C_c^Y \geq \beta m_y) = \overline{P}(C_c^X \geq \alpha m_x) \times \overline{P}(C_c^Y \geq \beta m_y) \tag{2.2}$$

We will refer to Equations (2.1) and (2.2) as 2-NPI-L and 2-NPI-U, respectively, and to the method in general as 2-NPI.

Next we will use the NPI results for future order statistics in Section 1.3.2, in particular Equation (1.17), to derive the NPI lower and upper probabilities in Equations (2.1) and (2.2). We first present the results for group X in detail, followed by those for group Y, for which deriving the results follows similar steps. We note that the event $C_c^X \geq \alpha m_x$ is equivalent to $X_{(\lceil \alpha m_x \rceil)} \leq c$, where $\lceil \alpha m_x \rceil$ is the smallest integer greater than αm_x, and similarly that the event $C_c^Y \geq \beta m_y$ is equivalent to $Y_{(m_y - \lceil \beta m_y \rceil + 1)} > c$, where $\lceil \beta m_y \rceil$ is the smallest integer greater than βm_y.

For $I_i^X = (x_{i-1}, x_i)$, $i = 1, \ldots, n_x + 1$, and $c \in I_{i_c}^X = (x_{i_c - 1}, x_{i_c})$, $i_c = 2, 3, \ldots, n_x$, the NPI lower and upper probabilities for the event $C_c^X \geq \alpha m_x$ are given by

$$\underline{P}(C_c^X \geq \alpha m_x) = \underline{P}(X_{(\lceil \alpha m_x \rceil)} \leq c) = \sum_{i=1}^{i_c - 1} P(X_{(\lceil \alpha m_x \rceil)} \in I_i^X) \tag{2.3}$$

$$\overline{P}(C_c^X \geq \alpha m_x) = \overline{P}(X_{(\lceil \alpha m_x \rceil)} \leq c) = \sum_{i=1}^{i_c} P(X_{(\lceil \alpha m_x \rceil)} \in I_i^X) \tag{2.4}$$

where the precise probabilities on the right hand sides of Equations (2.3) and (2.4) can be obtained from Equation (1.17). For $i_c = 1$, Equations (2.3) and (2.4) become

$$\underline{P}(C_c^X \geq \alpha m_x) = 0 \quad \text{and} \quad \overline{P}(C_c^X \geq \alpha m_x) = P(X_{(\lceil \alpha m_x \rceil)} \in I_1^X)$$

and for $i_c = n_x + 1$,

$$\underline{P}(C_c^X \geq \alpha m_x) = 1 - P(X_{(\lceil \alpha m_x \rceil)} \in I_{n_x+1}^X) \quad \text{and} \quad \overline{P}(C_c^X \geq \alpha m_x) = 1$$

If c is equal to one of the observations x_i, say $c = x_{i_c}$ for the specific value $i_c \in \{2, ..., n_x\}$, then this event has the following precise probability,

$$P(C_c^X \geq \alpha m_x) = P(X_{(\lceil \alpha m_x \rceil)} \leq c) = \sum_{i=1}^{i_c} P(X_{(\lceil \alpha m_x \rceil)} \in I_i^X) \tag{2.5}$$

Of course, this means that for such a value of c we have $\underline{P}(C_c^X \geq \alpha m_x) = \overline{P}(C_c^X \geq \alpha m_x) = P(C_c^X \geq \alpha m_x)$.

The NPI lower and upper probabilities for the event $C_c^Y \geq \beta m_y$ are derived similarly. For $I_j^Y = (y_{j-1}, y_j)$, $j = 1, \ldots, n_y + 1$, and $c \in I_{j_c}^Y = (y_{j_c-1}, y_{j_c})$, $j_c = 2, 3, \ldots, n_y$, the NPI lower and upper probabilities for the event $C_c^Y \geq \beta m_y$ are

$$\underline{P}(C_c^Y \geq \beta m_y) = \underline{P}(Y_{(m_y - \lceil \beta m_y \rceil + 1)} > c) = \sum_{j=j_c+1}^{n_y+1} P(Y_{(m_y - \lceil \beta m_y \rceil + 1)} \in I_j^Y) \tag{2.6}$$

$$\overline{P}(C_c^Y \geq \beta m_y) = \overline{P}(Y_{(m_y - \lceil \beta m_y \rceil + 1)} > c) = \sum_{j=j_c}^{n_y+1} P(Y_{(m_y - \lceil \beta m_y \rceil + 1)} \in I_j^Y) \tag{2.7}$$

For $j_c = 1$, Equations (2.6) and (2.7) become

$$\underline{P}(C_c^Y \geq \beta m_y) = 1 - P(Y_{(m_y - \lceil \beta m_y \rceil + 1)} \in I_1^Y) \quad \text{and} \qquad \overline{P}(C_c^Y \geq \beta m_y) = 1 \tag{2.8}$$

and for $j_c = n_y + 1$,

$$\underline{P}(C_c^Y \geq \beta m_y) = 0 \quad \text{and} \qquad \overline{P}(C_c^Y \geq \beta m_y) = P(Y_{(m_y - \lceil \beta m_y \rceil + 1)} \in I_{n_y+1}^Y)$$

Furthermore, for $c = y_{j_c}$ we have

$$P(C_c^Y \geq \beta m_y) = P(Y_{(m_y - \lceil \beta m_y \rceil + 1)} > c) = \sum_{j=j_c+1}^{n_y+1} P(Y_{(m_y - \lceil \beta m_y \rceil + 1)} \in I_j^Y) \tag{2.9}$$

Of course, this means that for such a value of c we have $\underline{P}(Y_{(m_y - \lceil \beta m_y \rceil + 1)} > c) = \overline{P}(Y_{(m_y - \lceil \beta m_y \rceil + 1)} > c) = P(Y_{(m_y - \lceil \beta m_y \rceil + 1)} > c))$.

The optimal diagnostic threshold is selected by maximisation of Equation (2.1) for the lower probability or Equation (2.2) for the upper probability. It should be emphasised that the 2-NPI-L and 2-NPI-U are different criterion, hence they may lead to different optimal thresholds. This method will be illustrated in examples in Section 2.5.

2.3 NPI method related to the two-group Youden index

In this section we introduce the NPI method for the two groups classification problem related to the Youden index procedure. We apply the 2-NPI method presented in Section 2.2, specifically Equations (2.1) and (2.2), to the Youden index method, in the sense that the criterion of the Youden index maximises the sum of the probabilities of the correct classification for the two groups. Let the NPI lower and upper probabilities related to the Youden index be denoted by 2-NPI-Y-L and 2-NPI-Y-U, respectively, and the method in general as 2-NPI-Y, and they are given by

$$\text{2-NPI-Y-L} = \underline{P}(C_c^Y \geq \beta m_y) + \underline{P}(C_c^X \geq \alpha m_x) - 1 \tag{2.10}$$

$$\text{2-NPI-Y-U} = \overline{P}(C_c^Y \geq \beta m_y) + \overline{P}(C_c^X \geq \alpha m_x) - 1 \tag{2.11}$$

These probabilities are calculated as explained in Section 2.2. The 2-NPI-Y-L and 2-NPI-Y-U may lead to different optimal thresholds. This method will be illustrated in examples in Section 2.5.

2.4 Searching for the optimal threshold

Following the setting introduced in Section 2.2, to find the optimal threshold c, there is no need to go through each of the $n_x + n_y + 1$ intervals created by the data observations. As for any sensible method, if c is moved such that one more data observation is correctly classified for one group while not changing the number of correctly classified data observation for the other group, it is an improvement. In this reasoning, we call a method 'sensible' if such a move of the threshold leads to a greater value of the target function, so typically our NPI lower and upper probabilities. Our methods are indeed sensible in this way, which follows from the expressions of the NPI lower and upper probabilities involved. Thus, the optimal threshold c for the two groups classification setting can only be in intervals where the left end point of the interval is an observation from group X and the right end point is an observation from group Y, that is $c \in (x_i, y_j)$. We should also consider the first and the last interval for the optimal threshold c. In the simulation study which will be presented in Section 2.6, we use this property to speed up the derivation of the optimal threshold c.

2.5 Examples

In the following examples, we illustrate the 2-NPI and 2-NPI-Y methods as presented in Sections 2.2 and 2.3, and we compare them with the empirical estimate of maximum area (EMA) and Youden index (EYI) methods presented in Section 1.2. As it is irrelevant how c is chosen within the respective intervals, the reported values of c in these examples are set be a value in the interval that is between two consecutive observations of the X and Y data combined. In the tables for all the examples, we represent the interval for the optimal threshold c by its left end point.

Example 2.1. For a specific gene, the relative gene expression intensities for 23 non-disease ovarian tissues, and 30 disease ovarian tumor tissues, are displayed in Table 2.1 [59]. This data set has three pairs of tied observations between the two groups $(0.571, 0.628$ and $0.641)$, we avoid the ties by adding 0.0001 to the three relevant observations from the cancer tissues group [24], see Table 2.2.

Normal tissues	0.442	0.500	0.510	0.568	0.571	0.574	0.588	0.595	0.595	0.595	0.598	0.606	0.617
	0.628	0.641	0.641	0.680	0.699	0.746	0.793	0.884	1.149	1.785			
Cancer tissues	0.543	0.571	0.602	0.609	0.628	0.641	0.666	0.694	0.769	0.800	0.800	0.847	0.877
	0.892	0.925	0.943	1.041	1.075	1.086	1.123	1.136	1.90	1.234	1.315	1.428	1.562
	1.612	1.666	1.666	2.127									

Table 2.1: The relative gene expression intensities

0.442	0.500	0.510	0.543	0.568	0.571	0.572	0.574	0.588	0.595	0.5951
0.5952	0.598	0.602	0.606	0.609	0.617	0.628	0.629	0.641	0.6411	0.642
0.666	0.680	0.694	0.699	0.746	0.769	0.793	0.800	0.8001	0.847	0.877
0.884	0.892	0.925	0.943	1.041	1.075	1.086	1.123	1.136	1.149	1.190
1.234	1.315	1.428	1.562	1.612	1.666	1.6661	1.785	2.127		

Table 2.2: The relative gene expression intensities, the healthy group (black) and diseased group (red)

m	2-NPI method				2-NPI-Y method			
	Lower case		Upper case		Lower case		Upper case	
	c	2-NPI-L	c	2-NPI-U	c	2-NPI-Y-L	c	2-NPI-Y-U
$\alpha = \beta = 0.6$								
5	0.746	0.7651	0.746	0.8282	0.746	0.7514	0.746	0.8214
10	0.746	0.7783	0.746	0.8506	0.746	0.7671	0.746	0.8464
30	0.746	0.8243	0.746	0.8993	0.746	0.8184	0.746	0.8979
100	0.746	0.8635	0.746	0.9328	0.746	0.8605	0.746	0.9323
$\alpha = \beta = 0.8$								
5	0.746	0.3954	0.746	0.4893	0.793	0.2839	0.793	0.4183
10	0.746	0.2800	0.746	0.3886	0.793	0.1100	0.793	0.2828
30	0.746	0.1574	0.746	0.2743	0.510	- 0.0053	0.793	0.1267
100	0.746	0.0955	0.746	0.2077	0.510	- 0.0023	0.793	0.0407
$\alpha = \beta = 0.2$								
5	0.746	0.9948	0.746	0.9970	0.746	0.9948	0.746	0.9970
10	0.746	0.9992	0.746	0.9996	0.746	0.9992	0.746	0.9996
30	0.746	1.0000	0.746	1.0000	0.746	1.0000	0.746	1.0000
100	0.746	1.0000	0.746	1.0000	0.746	1.0000	0.746	1.0000
$\alpha = 0.4,\ \beta = 0.7$								
5	0.628	0.7064	0.628	0.7847	0.628	0.6826	0.628	0.7724
10	0.6411	0.8259	0.6411	0.8888	0.6411	0.8201	0.6411	0.8867
30	0.628	0.8715	0.628	0.9355	0.628	0.8671	0.628	0.9345
100	0.628	0.9127	0.628	0.9646	0.628	0.9107	0.628	0.9643
$\alpha = 0.1,\ \beta = 0.9$								
5	0.598	0.5813	0.598	0.6986	0.598	0.5574	0.598	0.6866
10	0.598	0.7389	0.571	0.8564	0.598	0.7371	0.598	0.8512
30	0.571	0.7277	0.571	0.8887	0.571	0.7072	0.571	0.8854
100	0.571	0.7422	0.571	0.9178	0.571	0.7256	0.571	0.9161

Table 2.3: Optimal threshold c and corresponding value of 2-NPI-L, 2-NPI-U, 2-NPI-Y-L, 2-NPI-Y-U, using the 2-NPI and 2-NPI-Y methods and $m_x = m_y = m$

Table 2.3 provides the optimal threshold value c obtained from the two NPI-based methods along with their corresponding lower and upper probabilities, for $m_x = m_y$. We have considered different scenarios of α and β. As we can see from the table, for $\alpha = \beta =$

0.6, both NPI-based methods give the same optimal threshold value, $c \in (0.746, 0.769)$, regardless of the value of m.

On increasing the values of α and β ($\alpha = \beta = 0.8$), the 2-NPI method gives the same optimal threshold value as $\alpha = \beta = 0.6$ scenario, whereas for the 2-NPI-Y the optimal threshold is $c \in (0.793, 0.800)$, regardless of the value of m; except for the 2-NPI-Y-L, the optimal threshold is $c \in (0.510, 0.543)$ for $m = 30, 100$. In this scenario the values of lower and upper probabilities for both the methods are very low as they struggle to meet the required criterion. It is noticed that the 2-NPI-Y-L can be less than zero, this is because the lower probability of the number of correctly classified future individuals from groups X and Y in Equation (2.10) are very low. When the required criteria are easy to achieve ($\alpha = \beta = 0.2$), both the methods perform well as these corresponding lower and upper probabilities are very high and both the 2-NPI and 2-NPI-Y methods provide the same optimal threshold, which is $c \in (0.746, 0.769)$, regardless of the value of m.

For $\alpha = 0.4$, $\beta = 0.7$, as this scenario requests to put more emphasis on the number of correctly classified future individuals from group Y than that of group X, it is clear that the optimal threshold c for both methods decreases in order to achieve the desired criteria in comparison to the $\alpha = \beta$ scenario. In addition, the optimal threshold changes with different values of m, for example, for $m = 10$ the optimal threshold for both the NPI-based lower and upper probabilities is $c \in (0.6411, 0.642)$, whereas for $m = 5, 30, 100$, the optimal threshold is $c \in (0.628, 629)$. For the extreme case with $\alpha = 0.1, \beta = 0.9$ where the desired criterion strongly emphasises the number of future observations from group Y, the optimal threshold value c decreases to achieve the required criterion in comparison to the $\alpha = \beta$ scenario, which is $c \in (0.598, 0.602)$ for $m = 5, 10$ for both the methods, except for $m = 10$, the optimal threshold for the 2-NPI-U is $c \in (0.571, 0.572)$, and for larger values of m, $m = 30, 100$, the optimal threshold for both the methods is $c \in (0.571, 0.572)$.

m_x	m_y	2-NPI method				2-NPI-Y method			
		Lower case		Upper case		Lower case		Upper case	
		c	2-NPI-L	c	2-NPI-U	c	2-NPI-Y-L	c	2-NPI-Y-U
$\alpha = \beta = 0.6$									
10	6	0.746	0.6971	0.746	0.7723	0.746	0.6799	0.746	0.7651
30	40	0.746	0.8315	0.746	0.9070	0.746	0.8259	0.746	0.9057
60	100	0.746	0.8592	0.746	0.9306	0.746	0.8557	0.746	0.9299
$\alpha = \beta = 0.8$									
10	6	0.746	0.2955	0.746	0.3962	0.793	0.1457	0.793	0.3066
30	40	0.746	0.1404	0.746	0.2550	0.510	-0.0042	0.793	0.0964
60	100	0.746	0.0981	0.746	0.2087	0.510	-0.0023	0.793	0.0371
$\alpha = 0.4, \beta = 0.7$									
10	6	0.628	0.6656	0.628	0.7585	0.628	0.6387	0.628	0.7458
30	40	0.628	0.8774	0.628	0.9398	0.628	0.8734	0.628	0.9389
60	100	0.628	0.9056	0.628	0.9597	0.628	0.9033	0.628	0.9593

Table 2.4: Optimal threshold c and corresponding value of 2-NPI-L, 2-NPI-U, 2-NPI-Y-L, 2-NPI-Y-U, using the 2-NPI and 2-NPI-Y methods and $m_x \neq m_y$

Table 2.4 provides the optimal threshold value c obtained from the two NPI-based methods along with their corresponding lower and upper probabilities for $m_x \neq m_y$. Comparing this table with Table 2.4, with respect to the optimal threshold, the optimal thresholds for $\alpha = \beta = 0.6$ and $\alpha = \beta = 0.8$ are found to be the same, whereas for $\alpha = 0.4, \beta = 0.7$, the optimal threshold is $c \in (0.628, 0.629)$, regardless of the values of m_x and m_y. Again, in this table, the 2-NPI-Y-L can be less than zero since the lower probability of the number of correctly classified future individuals from groups X and Y in Equation (2.10) are very low.

Over all, it is clear from the results in this example that the optimal threshold can change depending on the values of α and β and also on the value of m. The maximum values of the empirical Youden index (EYI) and maximum area (EMA) are equal to 0.5696 and 0.6087, respectively, and the optimal threshold for both methods is $c \in (0.793, 0.800)$

As Example 2.1 involved a data set with the data from the two groups quite a bit overlapping, we now consider a small example with more separate data for the two groups.

Example 2.2. Consider an artificial data set for groups X and Y with $n_x = n_y = 10$, consisting of the ranks, $X = \{1, 2, 3, 4, 6, 7, 8, 9, 10, 12\}$ and $Y = \{5, 11, 13, 14, 15, 16, 17, 18, 19, 20\}$.

m	2-NPI method				2-NPI-Y method			
	Lower case		Upper case		Lower case		Upper case	
	c	2-NPI-L	c	2-NPI-U	c	2-NPI-Y-L	c	2-NPI-Y-U
$\alpha = \beta = 0.6$								
5	10	0.8521	10	0.9565	10	0.8462	10	0.9560
10	10	0.8641	10	0.9678	10	0.8591	10	0.9675
25	10	0.8843	10	0.9787	10	0.8807	10	0.9786
100	10	0.9019	10	0.9856	10	0.8993	10	0.9855
$\alpha = \beta = 0.8$								
5	10	0.5749	10	0.8186	10	0.5165	10	0.8095
10	10	0.5027	10	0.8006	10	0.4180	10	0.7895
25	10	0.4424	10	0.7937	10	0.3303	10	0.7818
100	10	0.4092	10	0.7949	10	0.2715	10	0.7831

Table 2.5: Optimal threshold c and corresponding value of 2-NPI-L, 2-NPI-U, 2-NPI-Y-L, 2-NPI-Y-U, using the 2-NPI and 2-NPI-Y methods and $m_x = m_y = m$

Table 2.5 provides the optimal threshold values c obtained from the 2-NPI and 2-NPI-Y methods along with their corresponding lower and upper probabilities, for $m_x = m_y = m$. We have considered two different scenarios of α and β. For $\alpha = \beta = 0.6$, both NPI-based methods give the same optimal threshold, $c \in (10, 11)$, regardless of the value of m, with high values of the lower and upper probabilities since the data from each group are less overlapping. The same results hold for $\alpha = \beta = 0.8$, but with lower values of the lower and upper probabilities. The maximum values of the empirical Youden index (EYI) and maximum area (EMA) are equal to 0.8000 and 0.8100, respectively, and the optimal threshold value for both the methods is $c \in (10, 11)$. It is clear that both the methods provide high values of the probability.

2.6 Simulation

In order to study the performance of the methods presented in this chapter, a simulation study was conducted for the two-group scenarios. We have considered two main cases, in which the data are simulated from the following normal distributions:

Case A: $X \sim N(0, 2^2)$ and $Y \sim N(1, 2^2)$.

Case B: $X \sim N(0, 1^2)$ and $Y \sim N(1, 1^2)$.

Due to the larger variance in Case A, the groups in that case overlap more than in Case B, with the means in case A being one standard deviation apart while they are 0.5 standard deviation apart in case B. We simulate n_x and n_y from the two normal distributions. Then, the n_x and n_y simulated data observations will be used to find the optimal thresholds c according to these methods and for specific values of (α, β) when applicable, where the threshold values are set to the midpoint in the partition of $\mathbb{R}$ used by the data. After that, we simulate m_x and m_y future observations from the same underlying normal distributions as the n_x and n_y simulated data observations to see how the methods perform.

The m_x and m_y simulated future observations are compared with the optimal thresholds to obtain the number of correctly classified observations per group. We have studied the predictive performance of all methods in terms of the number of correctly classified future observations that are achieved using the desired criterion, that is when the number of correctly classified future observations from group X and Y exceed αm_x and βm_y, respectively. Let us denote by '+' when the desired criterion is achieved and '−' otherwise. Throughout this simulation we assume that $n_x = n_y$ and $m_x = m_y$, and $j_x, j_y \in \{0, 1, \ldots, m\}$.

We have run the simulation for $n = 10$ and $m = 5, 30$, and we have chosen different values of α and β. Obviously the empirical Youden index and the maximum area methods do not depend on the values of α and β in terms of selecting the optimal thresholds. However for the comparison of predictive performance we have considered the same desired criterion of the number of future observations that are correctly classified from groups X and Y being at least αm_x and βm_y, respectively. The results in this section are based on 10,000 simulations per case per method.

To search for the optimal threshold c, rather than searching for the value c that maximises the probability within each of the $n_x + n_y + 1$ intervals created by the data observations, which could be computationally demanding especially in the simulation, we just consider the intervals as discussed in Section 2.2, that is we only consider the threshold c to be in intervals between an observation from group X to the left and an observation from group Y to the right, and we also consider the first and the last intervals.

The predictive performance results for Case A are given in Tables 2.6 and 2.7 for $m = 5$ and $m = 30$, respectively, and in Tables 2.8 and 2.9 for Case B. We have studied the performance in two shapes for $\alpha = \beta$ with values $0.2, 0.6$ and 0.8, and for $\alpha = 0.4, \beta = 0.7$, for the NPI-based methods (2-NPI and 2-NPI-Y) and the empirical estimates of the Youden index and maximum area methods (2-EYI and 2-EMA).

Consider Table 2.7, for example, where '$+\ +$' indicates that the desired criteria are achieved for both groups while '$-\ -$' indicates that the desired criteria for both groups are not achieved. For example, for 2-NPI-Y-U and $\alpha = \beta = 0.2$, the desired criteria have been achieved for both groups in 9886 out of 10,000 simulations, that is, at least 6 future observations ($\alpha m = 0.2 \times 30$ and $\beta m = 0.2 \times 30$) are correctly classified from each of the disease and non-disease groups. On the other hand, in 62 out of 10,000 simulations, the desired criterion is achieved (6 or more out of 30 are correctly classified) for group X, but the desired criterion is not achieved for group Y.

From Tables 2.6-2.9, the 2-NPI method outperforms all the other methods and for all the settings that have been considered for achieving the desired criterion for both groups. While for small values of α and β, it appears that the 2-NPI and 2-NPI-Y perform similarly, the 2-NPI-Y method performs poorly for larger values of α and β. One possible explanation is that the 2-NPI-Y method is based on the sum of the probabilities of correct classification rather than the product, which does not seem ideal if one tries to achieve higher proportions of those who are correctly classified. Yet for small values of α and β, as we have mentioned earlier, the 2-NPI-Y method performs equally well as the 2-NPI method.

Interestingly, the maximum area method (MA) is the closest in terms of performance to the 2-NPI method over all settings, yet the NPI method can be better, considering its predictive nature. It is not surprising that the maximum area method performs better than the Youden index method, as we have already discussed that summing the probabilities of correct classification may not be ideal when considering the prediction performance.

In addition, we can see from these tables that for $\alpha = \beta = 0.6$ and $\alpha = \beta = 0.8$, all the methods perform better for small value of m than for larger m, while for $\alpha = \beta = 0.2$, all the methods perform better for large m than for small m; this is because of the randomness effect as discussed in Section 2.1. In general, we notice that all the methods, when they are not achieving the desired criterion on both groups X and Y, tend to reach the desired criterion for either group X or group Y. However, for larger values of α and β, the 2-NPI and 2-EMA methods mostly fail the desired criterion for each group. This result becomes clearer for larger m; for example, in Table 2.7, for $\alpha = \beta = 0.8$ the 2-NPI and 2-EMA methods mostly fail the desired criterion for each group, whereas, the 2-NPI-Y method prefers to reach the desired criterion for either group X or group Y. It is obvious that if the values of α and β vary ($\alpha = 0.4$, $\beta = 0.7$), the required criterion becomes either harder or easier to achieve, which depends on these values and the value of m. Clearly, all methods perform poorly with the increase of α and β as the criteria become harder to achieve, especially for $\alpha = \beta = 0.8$. Finally, and not surprisingly, all methods perform much better in Case B than in Case A, as the groups in Case B are more separated than in Case A.

We summarise the number of correctly classified future observations in all simulations from groups X and Y using bar-plots as follows. Let the number of successfully classified future observations from group X with regards to the event of interest, which include α, be denoted by $S^X_{j_x}$ and the number of successfully classified future observations from group Y with regards to the event of interest, which include β, be denoted by $S^Y_{j_y}$, where $j_x \in \{0, 1, \ldots, m_x\}$ and $j_y \in \{0, 1, \ldots, m_y\}$, respectively. Figures 2.1-2.4 show the distributions of the numbers of future observations out of m in all 10,000 simulations, that are correctly classified for each group. For Case A, we can see that for larger values of $\alpha = \beta$, all methods struggle to meet the required criterion. Obviously, the performance for all methods becomes better for Case B since the groups have less overlap.

The results of this simulation show that the number of future observations considered and the values of α and β have an impact with regard to achieving the required criterion of the number of future observations that are correctly classified from groups X and Y.

X	Y	2-NPI-L	2-NPI-U	2-NPI-Y-L	2-NPI-Y-U	2-EYI	2-EMA
$\alpha = \beta = 0.2$							
-	-	0	0	0	0	0	0
-	+	301	293	301	294	890	424
+	-	259	249	259	249	620	356
+	+	9440	9458	9440	9457	8490	9220
$\alpha = \beta = 0.6$							
-	-	793	795	664	747	540	741
-	+	2869	2854	3372	3040	3844	3039
+	-	2795	2787	2937	2882	3034	2911
+	+	3543	3564	3027	3331	2582	3309
$\alpha = \beta = 0.8$							
-	-	3556	3575	1684	2447	2734	3455
-	+	2885	2874	4686	3902	3749	2999
+	-	2797	2779	3325	3149	2962	2815
+	+	762	772	305	502	555	731
$\alpha = 0.4, \beta = 0.7$							
-	-	863	864	727	816	575	607
-	+	2523	2727	2458	2887	1828	1031
+	-	3072	2860	3833	2939	5121	5663
+	+	3542	3549	2982	3358	2476	2699

Table 2.6: Simulation results $(10,000$ runs) for case A with $n = 10$ and $m = 5$

X	Y	2-NPI-L	2-NPI-U	2-NPI-Y-L	2-NPI-Y-U	2-EYI	2-EMA
$\alpha = \beta = 0.2$							
-	-	0	0	0	0	0	0
-	+	52	50	54	52	752	185
+	-	63	65	62	62	542	172
+	+	9885	9885	9884	9886	8706	9643
$\alpha = \beta = 0.6$							
-	-	867	890	586	797	488	751
-	+	3943	3922	4753	4162	4905	4203
+	-	3624	3595	3606	3617	3748	3696
+	+	1566	1593	1055	1424	859	1350
$\alpha = \beta = 0.8$							
-	-	7043	7186	1461	2701	5003	6746
-	+	1495	1447	3327	4450	2899	1753
+	-	1460	1365	5212	2848	2097	1499
+	+	2	2	0	1	1	2
$\alpha = 0.4, \beta = 0.7$							
-	-	274	277	210	266	154	181
-	+	3105	3148	2718	3249	2630	1437
+	-	3556	3450	4620	3539	5379	6236
+	+	3065	3125	2452	2946	1837	2146

Table 2.7: Simulation results $(10,000$ runs) for case A with $n = 10$ and $m = 30$

X	Y	2-NPI-L	2-NPI-U	2-NPI-Y-L	2-NPI-Y-U	2-EYI	2-EMA
$\alpha = \beta = 0.2$							
-	-	0	0	0	0	0	0
-	+	116	112	116	113	347	172
+	-	95	94	95	94	182	134
+	+	9789	9794	9789	9793	9471	9694
$\alpha = \beta = 0.6$							
-	-	226	236	212	226	175	209
-	+	2095	2084	2199	2119	2843	2208
+	-	1992	1970	2108	2019	2089	2086
+	+	5687	5710	5481	5636	4893	5497
$\alpha = \beta = 0.8$							
-	-	1956	1975	1360	1696	1669	1904
-	+	3090	3076	3899	3418	3766	3162
+	-	3052	3022	3374	3228	2931	3067
+	+	1902	1927	1367	1658	1634	1867
$\alpha = 0.4, \beta = 0.7$							
-	-	287	297	261	284	208	193
-	+	1842	1900	1775	1930	1111	677
+	-	2449	2369	2829	2413	4392	4778
+	+	5422	5434	5135	5373	4289	4352

Table 2.8: Simulation results ($10,000$ runs) for case B with $n = 10$ and $m = 5$

X	Y	2-NPI-L	2-NPI-U	2-NPI-Y-L	2-NPI-Y-U	2-EYI	2-EMA
$\alpha = \beta = 0.2$							
-	-	0	0	0	0	0	0
-	+	9	9	10	9	163	41
+	-	11	11	11	10	88	26
+	+	9980	9980	9979	9981	9749	9933
$\alpha = \beta = 0.6$							
-	-	31	33	26	34	20	21
-	+	2571	2518	2905	2629	3723	2860
+	-	2377	2345	2546	2348	2570	2574
+	+	5021	5104	4523	4989	3687	4545
$\alpha = \beta = 0.8$							
-	-	4513	4627	1580	2987	3470	4257
-	+	2747	2726	3646	3747	3835	2998
+	-	2673	2579	4748	3220	2640	2684
+	+	67	68	26	46	55	61
$\alpha = 0.4, \beta = 0.7$							
-	-	7	7	7	7	2	4
-	+	1525	1432	1525	1452	1232	576
+	-	2035	2059	2447	2105	4189	4615
+	+	6433	6502	6021	6436	4577	4805

Table 2.9: Simulation results ($10,000$ runs) for case B with $n = 10$ and $m = 30$

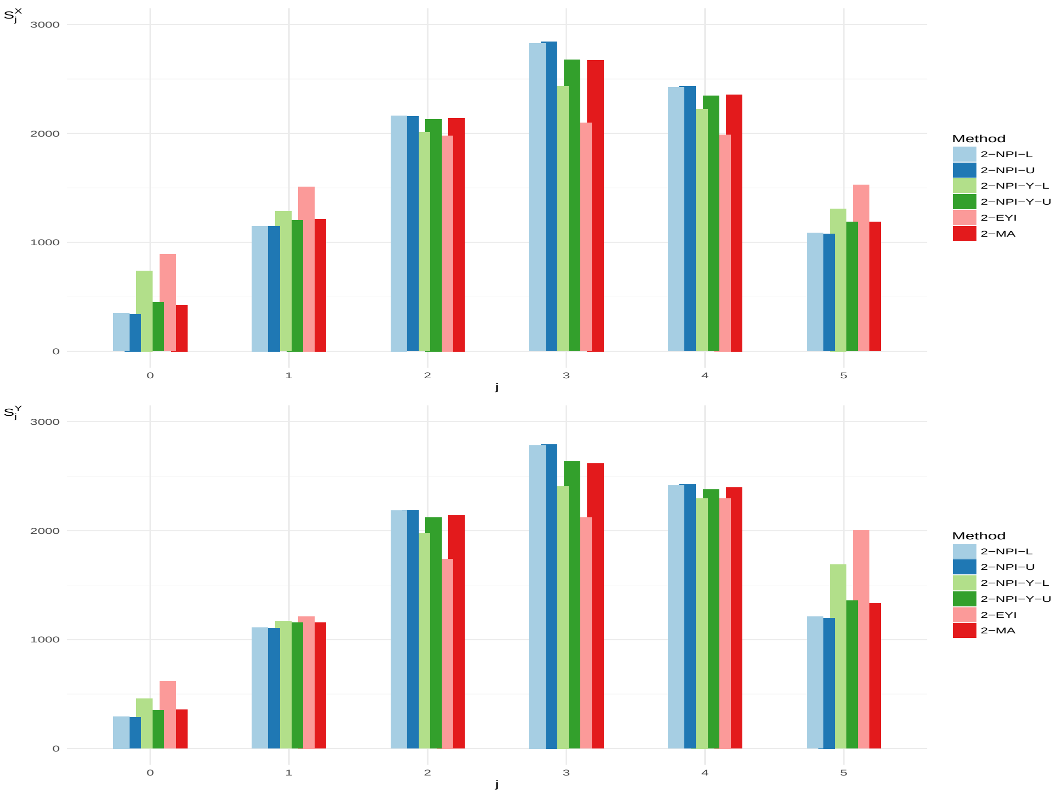

Figure 2.1: Simulation results (10,000 runs), when $\alpha = \beta = 0.6$ and $m = 5$ (case A)

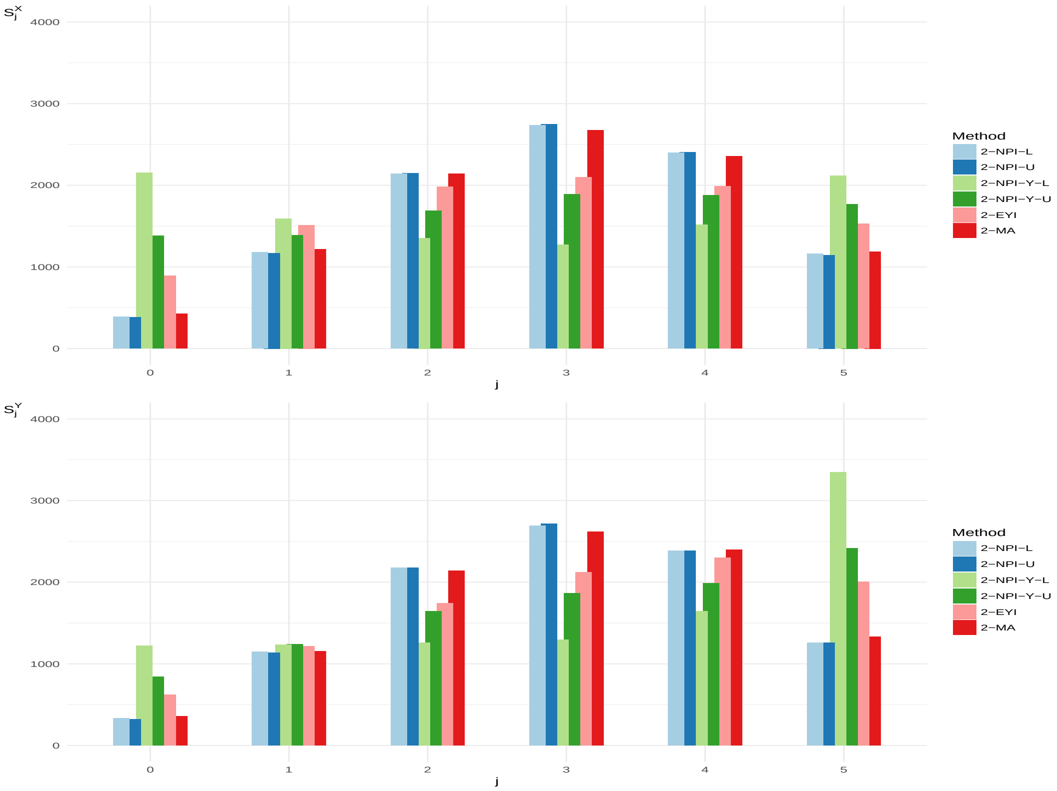

Figure 2.2: Simulation results (10,000 runs), when $\alpha = \beta = 0.8$ and $m = 5$ (case A)

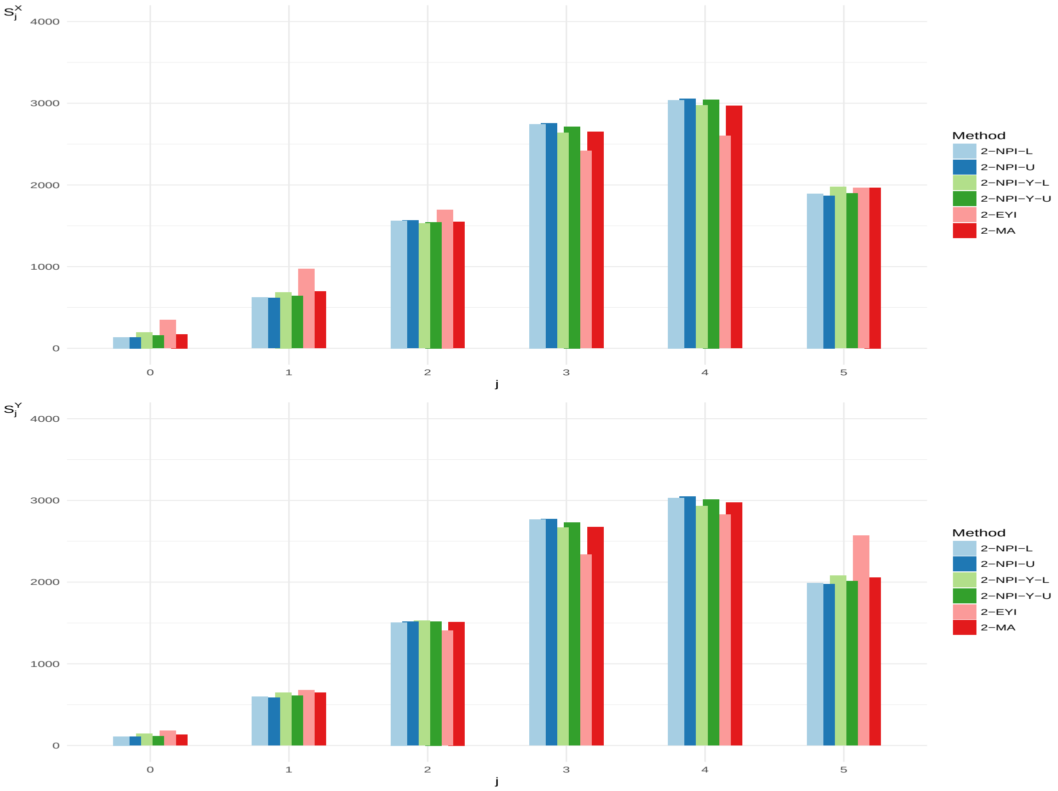

Figure 2.3: Simulation results (10, 000 runs), when $\alpha = \beta = 0.6$ and $m = 5$ (case B)

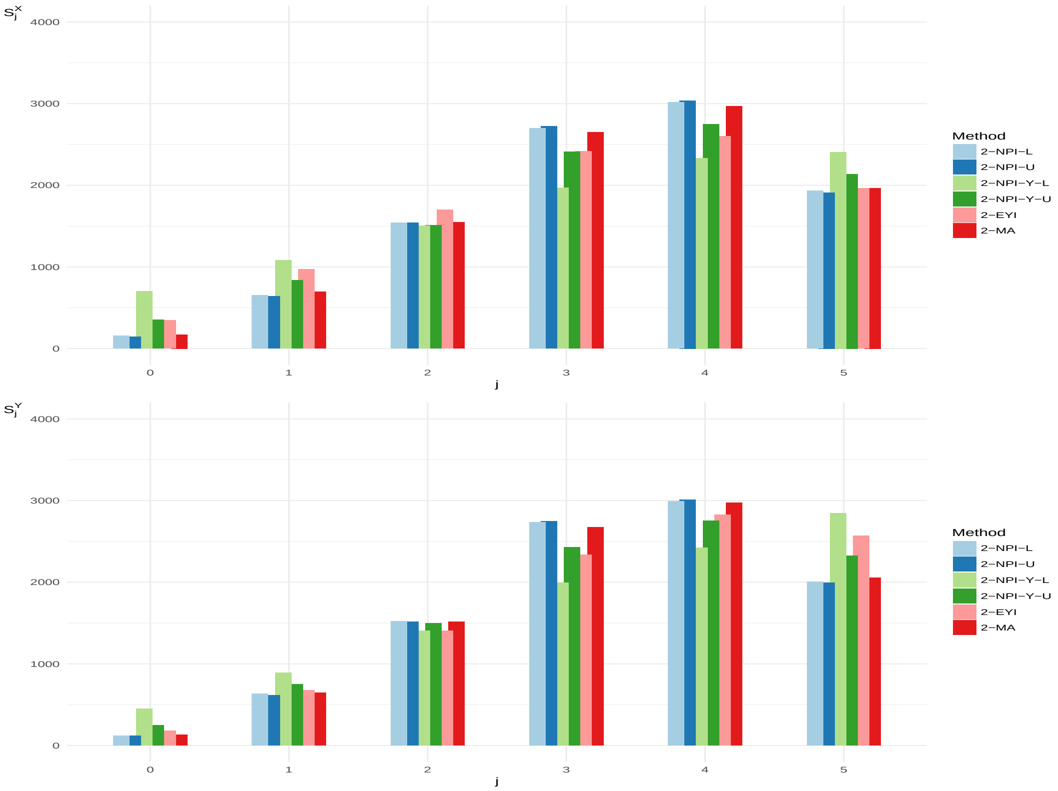

Figure 2.4: Simulation results (10, 000 runs), when $\alpha = \beta = 0.8$ and $m = 5$ (case B)

2.7 Concluding remarks

This chapter has presented methods for selecting the optimal diagnostic threshold with two groups, explicitly as a predictive problem instead of the classical approach based on estimation. We considered m future individuals in each group for who the threshold would be applied, and criteria in terms of the proportions of successful diagnoses. Nonparametric predictive inference was applied to derive the optimal thresholds, which were shown to depend on the target success proportions and also on the value of m. We have shown that the optimal threshold might change if the number of future individuals changes.

We have presented the use of the target proportions α and β in our method presented in Section 2.2. We consider this an attractive approach for the predictive method. It will be interesting to compare this approach to the use of utilities. We have restricted attention to introducing the NPI method for the two groups classification problem related to the Youden index procedure, and left further investigation of the use of other methods within the NPI framework, such as the minimum P value method (min P) as mentioned in Section 1.2. In the next chapter, we extend the two-group NPI approach for selecting the optimal threshold to three-group classification problems.

Chapter 3

NPI for three-group diagnostic test thresholds

3.1 Introduction

In this chapter, we extend the two groups NPI methods for selecting the optimal thresholds presented in Chapter 2 to three groups classification problems. Traditionally, measuring the diagnostic test accuracy dealt with binary outcomes where individuals can be in one of two states: healthy or diseased. Often, however, medicine studies involve discriminating between more than two stages. For example, in Alzheimer's disease (AD), there exists mild cognitive impairment (MCI) as an intermediate stage (transition stage) between normal aging and complete loss of memory [28, 51, 70]. The intermediate stage in the AD progress is crucial to detect as it is an indication of serious disease processing in the future. For the late stages of the disease, no medical treatments are efficient, whereas the intermediate stage can lead to early treatment with new drugs to slow the development of memory loss. The treatment for those in the intermediate stage can provide a more profound influence on the cognitive decline rate [70]. Therefore, it is important to improve diagnostic test accuracy for distinguishing among the three disease stages. In this setting, the ROC curve is generalized to the ROC surface by adding a third dimension [52, 53, 56] and considering two decision thresholds, $c_1 < c_2$, to classify individuals into one of these groups. Selecting appropriate threshold values c_1 and c_2 is the main aspect of analysis of a diagnostic test to distinguish between the three groups.

Section 3.2 provides a brief overview of existing methods for diagnostic test thresholds in the three groups setting. In Section 3.3, we discuss a pairwise approach for selecting the optimal thresholds in the three-group diagnostic test scenario. In Section 3.4, we introduce NPI for selecting the optimal thresholds for three-group diagnostic tests. Section 3.5 introduces a NPI method related to the three-group Youden index. Section 3.6 presents some examples to illustrate and discuss the new approaches. We compare and investigate the performance of the three groups NPI methods and some classical methods via a simulation study in Section 3.7. Finally, some concluding remarks are made in Section 3.8.

3.2 Thresholds selection in three-group classification

The ROC surface is a useful tool to assess the accuracy of a diagnostic test when three ordered groups are involved. To introduce the ROC surface let there be three separately ordered groups, denoted by X, Y and Z. Assume that we have real-valued data from diagnostic tests on individuals from the three groups; group X with n_x observations, group Y with n_y observations and group Z with n_z observations. Assume that a continuous diagnostic test is used to distinguish the individuals from the three groups. Suppose that the measurements from group X tend to be smaller than those from group Y, which in turn tend to be smaller than those from group Z. Let the cumulative distribution functions (CDFs) for the test outcomes of the three groups X, Y and Z be denoted by F_x, F_y and F_z, respectively.

For a decision rule, two thresholds $c_1 < c_2$ are required to classify individuals, based on their diagnostic test results, into one of the three groups, such that a test value which is less than or equal to c_1 is an indication that this individual belongs to group X, a test value which is greater than c_1 and less than or equal to c_2 is an indication that this individual belongs to group Y, and a test value which is greater than c_2 is an indication that this individual belongs to group Z. The probability of correct classification for the three groups with thresholds $c_1 < c_2$ are as follows; $p_1 = P(X \leq c_1) = F_x(c_1)$ is the probability of correct classification for individuals from group X, $p_2 = P(c_1 < Y \leq c_2) = F_y(c_2) - F_y(c_1)$ for individuals from group Y and $p_3 = P(Z \geq c_2) = 1 - F_z(c_2)$ for individuals from group Z. The three-class ROC surface is a plot of these probabilities of correct diagnosis for all possible values $c_1 < c_2$ [52, 53, 56]. For three-group classification problems, the volume under the ROC surface (VUS) has been extensively studied for assessment accuracy of a diagnostic test to differentiate among the three groups [1, 52, 70, 71]. The VUS is equal to the probability that three randomly selected measurements (one from each disease group) are ordered correctly. It takes the value 1 if the three groups are perfectly ordered and the value $\frac{1}{6}$ if the diagnostic test results for the three groups are identical.

Once the accuracy of a diagnostic test is determined over all the possible thresholds, the selection of optimal thresholds is required to discriminate between the three groups. The common approach is the generalization of the Youden index as introduced by Nakas et al. [54], which is an extension of the two-group Youden index, discussed in Section 1.2, to the three-group setting. The three-group Youden index (3-YI) is defined as

$$\text{3-YI} \quad = \quad \max_{(c_1 < c_2)} \{F_x(c_1) + F_y(c_2) - F_y(c_1) + 1 - F_z(c_2)\} \tag{3.1}$$

The optimal thresholds are the values of c_1 and c_2 which maximise the 3-YI, with the constraint $c_1 < c_2$, where 3-YI is equal to 1 when the three groups are identical, and equal to 3 where they are perfectly distinguished. In order to obtain the empirical estimator for the 3-YI, replace the CDFs by the corresponding empirical CDFs. The empirical estimate of the Youden index (3-EYI) is given by

$$\text{3-EYI}(c) = \frac{1}{n_x} \sum_{i=1}^{n_x} \mathbf{1}\{x_i \leq c_1\} + \frac{1}{n_y} \sum_{j=1}^{n_y} \mathbf{1}\{c_1 < y_j \leq c_2\} + \frac{1}{n_z} \sum_{l=1}^{n_z} \mathbf{1}\{z_l > c_2\}. \tag{3.2}$$

Other methods for three-group thresholds selection based on ROC analyses are the closest to perfection method (3-MD) and the maximum volume method (3-MV), as introduced by Attwood et al. [4]. Both approaches are generalisations of corresponding methods in two-group classification, namely the closest-to-(0,1) method [11, 65] and the maximum area method, respectively.

The 3-MD approach selects the optimal thresholds which generate the point on the ROC surface closest to the point of perfection (1,1,1) (i.e. the point closest to perfection with $p_1(c_1) = 1$, $p_2(c_1, c_2) = 1$ and $p_3(c_2) = 1$). The optimal thresholds are the values of c_1 and c_2, which minimise the distance, and this method is given by

$$\text{3-MD} = \min_{(c_1<c_2)} \{\sqrt{(1 - p_1(c_1))^2 + (1 - p_2(c_1, c_2))^2 + (1 - p_3(c_2))^2}\} \tag{3.3}$$

The 3-MV method can be defined as the maximum product of the correct classification probabilities for the three groups as follows

$$\text{3-MV} = \max_{(c_1<c_2)} \{p_1(c_1) \times p_2(c_1, c_2) \times p_3(c_2)\} \tag{3.4}$$

Attwood et al. [4] did not mention the empirical estimate of the maximum volume (3-EMV) method in their paper, which is defined by

$$\text{3-EMV}(c) = \frac{1}{n_x}\sum_{i=1}^{n_x} \mathbf{1}\{x_i \leq c_1\} \times \frac{1}{n_y}\sum_{j=1}^{n_y} \mathbf{1}\{c_1 < y_j \leq c_2\} \times \frac{1}{n_z}\sum_{l=1}^{n_z} \mathbf{1}\{z_l > c_2\}. \tag{3.5}$$

Attwood et al. [4] also discussed a comparison of optimal thresholds selected by their methods and the three-group Youden index method (3-YI). We review some results obtained by [4]. Although the 3-YI approach maximises the total number of correct classification rates for the three groups, it tends to have limitation in selecting the thresholds c_1 and c_2. The maximisation problem in Equation (3.1) can be written as the maximisation of two two-group problems, one between the healthy and intermediate groups and the other between the intermediate and diseased groups, given that $c_1 < c_2$. This can lead to imbalanced classification rates between the three groups, in favour of identifying healthy and diseased groups but poor identification of the intermediate group. This aspect is in line with the results of the simulation study, which will be presented it in Section 3.7.

Nakas et al. [55] applied the Youden index method for pairwise analysis in Montreal

Cognitive Assessment (MOCA) when screening cognitive impairment in Parkinson disease (PD). The study sample of patients was classified into three groups as patients with dementia (PD-D), patients with mild cognitive impairment (PD-MCI) and normal cognition (PD-N). The PD levels are anticipated to be lowest among the PD-D group and highest among the PD-N group, with the PD-MCI group being intermediate to the other two. The optimal thresholds are derived by using the pairwise Youden index, selecting the optimal threshold c_1 from groups PD-D and PD-MCI and selecting the optimal threshold c_2 from groups PD-MCI and PD-N. In addition, Nakas et al. [55] discussed the comparison of the optimal thresholds c_1 and c_2 selected by the pairwise Youden index and the three-group Youden index. The results showed that the optimal thresholds c_1 and c_2 derived by both approaches are the same because the maximisation problem in Equation (3.1) can be seen as two two-group maximisation problems. Moreover, the value of the Youden index for the three-group problem is equal tothe sum of the values of the Youden indexes for the two two-group problems. This result generally holds given that $(c_1 < c_2)$ for the two two-group problems.

In this chapter, we will consider the 3-EYI and 3-EMV methods to compare them with our proposed methods.

3.3 NPI pairwise analysis for three-group diagnostic test thresholds

One possible way to find the optimal thresholds c_1 and c_2 for the three groups setting, is by naively using the 2-NPI method, presented in Section 2.2, twice. Thus, in addition to the notation introduced in Section 2.2 for groups X and Y, we need to introduce further notation for group Z as follows. Suppose we have n_z observations from group Z, and the ordered data from this group are denoted by $z_1 < z_2 < \ldots < z_{n_z}$, and we define $z_0 = -\infty$ and $z_{n_z+1} = \infty$. These n_z observations partition the real-line into $n_z + 1$ intervals $I_l^Z = (z_{l-1}, z_l)$, for $l = 1, 2, \ldots, n_z + 1$. Let the diagnostic test results of m_z future individuals be denoted by Z_{n_z+t}, $t = 1, \ldots, m_z$, and let the corresponding ordered future observations be denoted by $Z_{(1)} < Z_{(2)} < \ldots < Z_{(m_z)}$. Similarly, we assume that the three groups are fully independent as explained in Section 2.2. Assume that the three groups are ordered in the sense that observations from group X tend to be smaller than those from group Y, which in turn tend to be smaller than those from group Z.

From Equations (2.1) and (2.2) we find the optimal threshold c_1 for groups X and Y. Similarly, for groups Y and Z we find the optimal c_2 by maximising either the lower or upper probabilities for the events $C^Y_{c_2} \geq \beta m_y$ and $C^Z_{c_2} \geq \gamma m_z$. These lower or upper probabilities are derived as follows.

$$\underline{P}(C^Y_{c_2} \geq \beta m_y, C^Z_{c_2} \geq \gamma m_z) = \underline{P}(C^Y_{c_2} \geq \beta m_y) \times \underline{P}(C^Z_{(c_2} \geq \gamma m_z) \tag{3.6}$$

$$\overline{P}(C^Y_{c_2} \geq \beta m_y, C^Z_{c_2} \geq \gamma m_z) = \overline{P}(C^Y_{c_2} \geq \beta m_y) \times \overline{P}(C^Z_{(c_2} \geq \gamma m_z) \tag{3.7}$$

where

$$\underline{P}(C^Y_{c_2} \geq \beta m_y) = \underline{P}(Y_{\lceil \beta m_y \rceil} \leq c_2) = \sum_{j=1}^{j_{c_2}-1} P(Y_{\lceil \beta m_y \rceil} \in I^Y_j) \tag{3.8}$$

$$\overline{P}(C^Y_{c_2} \geq \beta m_y) = \overline{P}(Y_{\lceil \beta m_y \rceil} \leq c_2) = \sum_{j=1}^{j_{c_2}} P(Y_{\lceil \beta m_y \rceil} \in I^Y_j) \tag{3.9}$$

$$\underline{P}(C^Z_{c_2} \geq \gamma m_z) = \underline{P}(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} > c_2) = \sum_{l=l_{c_2}+1}^{n_z+1} P(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} \in I^Z_l) \tag{3.10}$$

$$\overline{P}(C^Z_{c_2} \geq \gamma m_z) = \overline{P}(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} > c_2) = \sum_{l=l_{c_2}}^{n_z+1} P(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} \in I^Z_l) \tag{3.11}$$

The precise probabilities in Equations (3.8)-(3.11) can be calculated using Equation (1.17) in Section 1.3.2.

We will refer to this pairwise method as NPI-PW and the corresponding approach that utilises the lower (upper) probabilities in Equations (2.1) and (3.6) (in Equations (2.2) and (3.7)) to obtain the optimal thresholds (c_1, c_2) as NPI-PW-L (NPI-PW-U). It is important to emphasise that selecting the optimal thresholds based on this method may not satisfy the condition that $c_1 < c_2$. It might be that the groups are ordered in a different way, so one could investigate a change of the order of the three groups X, Y and Z. Generally, this method is not suggested to be applicable. However we introduce it for comparison in the examples and simulation study that will be presented later in this chapter. The problem consideration of the NPI-PW method motivates us to develop a better method for the three groups classification setting in the next section.

3.4 NPI for three-group diagnostic test thresholds

In this section, we extend the two-group NPI method for selecting the optimal threshold presented in Section 2.2, to a three-group setting. We follow the same notation as presented in Section 3.3, but we need to add the following. Let us assume that the three groups are ordered in the sense that observations from group X tend to be smaller than those from group Y, which in turn tend to be smaller than those from group Z. For a decision rule, two thresholds $c_1 < c_2$ are required to classify individuals, based on their diagnostic test results, into one of the three groups, such that a test value which is less than or equal to c_1 is an indication that this individual belongs to group X, a test value which is greater than c_1 and less than or equal to c_2 is an indication that this individual belongs to group Y, and a test value which is greater than c_2 is an indication that this individual belongs to group Z.

For specific values of c_1 and c_2, with $c_1 < c_2$, $C^X_{c_1}$ denotes the number of correctly classified future individuals from group X, that is those with test results $X_{n_x+r} \leq c_1$ (for $r = 1, \ldots, m_x$), $C^Y_{(c_1,c_2)}$ denotes the number of correctly classified future individuals from group Y, that is those with test results $c_1 < Y_{n_y+s} \leq c_2$ (for $s = 1, \ldots, m_y$), and $C^Z_{c_2}$ denotes the number of correctly classified future individuals from group Z, that is those with test results $Z_{n_z+t} > c_2$ (for $t = 1, \ldots, m_z$).

Let α, β and γ be any values in $(0, 1]$ that are selected to reflect the desired importance of the groups. Following the same events of interest for the two groups as presented in Section 2.2, the events of interest for the groups X, Y and Z that we focus on are $C^X_{c_1} \geq \alpha m_x, C^Y_{(c_1,c_2)} \geq \beta m_y$ and $C^Z_{c_2} \geq \gamma m_z$, respectively. Varying the values of α, β and γ will depend on one's beliefs of which group is more important to be correctly diagnosed. Of course one can choose α, β and γ to be equal if one prefers to give the same importance of correct classification to all future individuals.

Under the independence assumption of the three groups, the joint NPI lower and upper probabilities can be derived as the products of the corresponding lower and upper probabilities for the individual events involving $C^X_{c_1}$, $C^Y_{(c_1,c_2)}$, and $C^Z_{c_2}$, thus

$$\begin{aligned}
&\underline{P}(C^X_{c_1} \geq \alpha m_x, C^Y_{(c_1,c_2)} \geq \beta m_y, C^Z_{c_2} \geq \gamma m_z) = \\
&\underline{P}(C^X_{c_1} \geq \alpha m_x) \times \underline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y) \times \underline{P}(C^Z_{c_2} \geq \gamma m_z) \qquad (3.12)\\
&\overline{P}(C^X_{c_1} \geq \alpha m_x, C^Y_{(c_1,c_2)} \geq \beta m_y, C^Z_{c_2} \geq \gamma m_z) = \\
&\overline{P}(C^X_{c_1} \geq \alpha m_x) \times \overline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y) \times \overline{P}(C^Z_{c_2} \geq \gamma m_z) \qquad (3.13)
\end{aligned}$$

We refer to the use of Equations (3.12) and (3.13) as 3-NPI-L and 3-NPI-U, respectively, and the method in general as 3-NPI.

For $I^X_i = (x_{i-1}, x_i)$ with $i = 1, \ldots, n_x + 1$ and $c_1 \in I^X_{i_{c_1}} = (x_{i_{c_1}-1}, x_{i_{c_1}})$, $i_{c_1} \in \{2, 3, \ldots, n_x\}$, the NPI lower and upper probabilities for the event $C^X_{c_1} \geq \alpha m_x$ are given by

$$\underline{P}(C^X_{c_1} \geq \alpha m_x) = \underline{P}(X_{\lceil \alpha m_x \rceil} \leq c_1) = \sum_{i=1}^{i_{c_1}-1} P(X_{\lceil \alpha m_x \rceil} \in I^X_i) \qquad (3.14)$$

$$\overline{P}(C^X_{c_1} \geq \alpha m_x) = \overline{P}(X_{\lceil \alpha m_x \rceil} \leq c_1) = \sum_{i=1}^{i_{c_1}} P(X_{\lceil \alpha m_x \rceil} \in I^X_i) \qquad (3.15)$$

For $i_{c_1} = 1$, Equations (3.14) and (3.15) become

$$\underline{P}(C^X_{c_1} \geq \alpha m_x) = 0 \quad \text{and} \qquad \overline{P}(C^X_{c_1} \geq \alpha m_x) = P(X_{(\lceil \alpha m_x \rceil)} \in I^X_1)$$

and for $i_{c_1} = n_x + 1$,

$$\underline{P}(C^X_{c_1} \geq \alpha m_x) = 1 - P(X_{(\lceil \alpha m_x \rceil)} \in I^X_{n_x+1}) \quad \text{and} \qquad \overline{P}(C^X_{c_1} \geq \alpha m_x) = 1$$

If c_1 is equal to one of the observations x_i, say $c_1 = x_{i_{c_1}}$ for the specific value $i_{c_1} \in \{2, ..., n_x\}$, then this event has the following precise probability,

$$P(C^X_{c_1} \geq \alpha m_x) = P(X_{(\lceil \alpha m_x \rceil)} \leq c_1) = \sum_{i=1}^{i_{c_1}} P(X_{(\lceil \alpha m_x \rceil)} \in I^X_i)$$

For $I^Y_j = (y_{j-1}, y_j)$ with $j = 1, \ldots, n_y + 1$ and $c_1 \in I^Y_{j_{c_1}} = (y_{j_{c_1}-1}, y_{j_{c_1}})$ and $c_2 \in I^Y_{j_{c_2}} =$

$(y_{j_{c_2}-1}, y_{j_{c_2}})$, with $j_{c_1} \in \{1,\ldots,n_y{+}1\}$ and $j_{c_2} \in \{1,\ldots,n_y{+}1\}$, with $c_2 \geq c_1$, which implies that $j_{c_2} \geq j_{c_1}$, the NPI approach leads to the following lower and upper probabilities $\underline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y)$ and $\overline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y)$,

$$\underline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y) = P(C^Y_{(y_{j_{c_1}}, y_{j_{c_2}-1})} \geq \beta m_y) \tag{3.16}$$

$$\overline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y) = P(C^Y_{(y_{j_{c_1}-1}, y_{j_{c_2}})} \geq \beta m_y) \tag{3.17}$$

For $j_{c_1} = 1$ and $j_{c_2} = 2$, Equations (3.16) and (3.17) become

$$\underline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y) = 0 \quad \text{and} \qquad \overline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y) = P(C^Y_{(-\infty, y_{j_{c_2}})} \geq \beta m_y)$$

For $j_{c_1} = 1$ and $j_{c_2} = \{3, ..., n_y + 1\}$,

$$\underline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y) = P(C^Y_{(y_{j_{c_1}}, y_{j_{c_2}-1})} \geq \beta m_y) \quad \text{and}$$

$$\overline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y) = P(C^Y_{(-\infty, y_{j_{c_2}})} \geq \beta m_y)$$

For $j_{c_1} = n_y$ and $j_{c_2} = n_y + 1$,

$$\underline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y) = 0 \quad \text{and} \qquad \overline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y) = P(C^Y_{(y_{j_{c_1}-1}, \infty)} \geq \beta m_y)$$

.

Note that $\underline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y) = 0$ for all $j_{c_2} = j_{c_1} + 1$. Further a special case occurs when c_1 and c_2 occur in the same interval, that is $j_{c_1} = j_{c_2}$. Then the lower probability in Equation (3.16) is equal to zero.

The upper probability in Equation (3.17) can be calculated as follows. In order to assign the probability masses within the interval $(y_{j_{c_1}-1}, y_{j_{c_1}})$ to derive the NPI upper probability in Equation (3.17), let the number of observations from groups X and Z between $y_{j_{c_1}-1}$ and $y_{j_{c_1}}$ be denoted by $n_x^{j_{c_1}}$ and $n_z^{j_{c_1}}$, respectively. These observations create a partition of the interval $(y_{j_{c_1}-1}, y_{j_{c_1}})$ into $n_x^{j_{c_1}} + n_z^{j_{c_1}} + 1$ sub-intervals. If c_1 is in sub-interval (y_{j-1}, x_i), then we put the probability mass to the right end point x_i. Simultaneously, if c_2 is in sub-interval (z_l, y_j), then we put the probability mass to the left end point z_l, $l = 1, ..., n_z + 1$. If the observations are only from group X, so $n_z^{j_{c_1}} = 0$, then we put the probability mass to the right end point x_i, and if they are only from group Z, $n_x^{j_{c_1}} = 0$, then we put the probability mass to the left end point z_l. If there are no observations from groups X and Z in the interval $(y_{j_{c_1}-1}, y_{j_{c_1}})$, then we put all the probability masses in between c_1 and c_2, as long as c_1 is to the left of c_2.

For $I_l^Z = (z_{l-1}, z_l)$ with $l = 1, \ldots, n_z + 1$ and $c_2 \in I_{l_{c_2}}^Z = (z_{l_{c_2}-1}, z_{l_{c_2}})$, $l_{c_2} = 1, 2, 3, \ldots, n_z$, the NPI approach leads to the following lower and upper probabilities $\underline{P}(C_{c_2}^Z \geq \gamma m_z)$ and $\overline{P}(C_{c_2}^Z \geq \gamma m_z)$,

$$\underline{P}(C_{c_2}^Z \geq \gamma m_z) = \underline{P}(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} > c_2) = \sum_{l=l_{c_2}+1}^{n_z+1} P(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} \in I_l^Z) \tag{3.18}$$

$$\overline{P}(C_{c_2}^Z \geq \gamma m_z) = \overline{P}(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} > c_2) = \sum_{l=l_{c_2}}^{n_z+1} P(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} \in I_l^Z) \tag{3.19}$$

For $l_{c_2} = 1$, Equations (3.18) and (3.19) become

$$\underline{P}(C_{c_2}^Z \geq \gamma m_z) = 1 - P(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} \in I_1^Z) \quad \text{and} \qquad \overline{P}(C_{c_2}^Z \geq \gamma m_z) = 1$$

and for $l_{c_2} = n_z + 1$,

$$\underline{P}(C_{c_2}^Z \geq \gamma m_z) = 0 \quad \text{and} \qquad \overline{P}(C_{c_2}^Z \geq \gamma m_z) = P(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} \in I_{n_z+1}^Z)$$

Furthermore, for $c = z_{l_{c_2}}$ we have

$$P(C_{c_2}^Z \geq \gamma m_z) = P(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} > c_2) = \sum_{l=l_{c_2}+1}^{n_z+1} P(Z_{(m_z - \lceil \gamma m_z \rceil + 1)} \in I_j^Z)$$

The optimal thresholds c_1 and c_2 can be obtained by maximising Equations (3.12) and (3.13). To search for the optimal thresholds c_1 and c_2, we follow a similar process as presented in Section 2.4, with the addition of group Z. Thus, we need to search for the values c_1 and c_2 that maximise the lower or the upper probability for Equations (3.12) or (3.13), respectively, within each of the $n_x + n_y + n_z + 1$ intervals created by the data observations. However, the optimal threshold c_1 can only be in intervals where the left end point of the interval is an observation from group X and the right end point is an observation from group Y, that is $c_1 \in (x_i, y_j)$. Any observations from group Z are irrelevant here and must be ignored. On the other hand, the optimal threshold c_2 can only be in intervals where the left end point of the interval is an observation from group Y and the right end point is an observation from group Z, that is $c_2 \in (y_j, z_l)$. Any observations from group X are irrelevant here and must be ignored. We should also consider the first interval for the optimal threshold c_1 and the last interval for the optimal threshold c_2. In the simulation study which will be presented in Section 3.7, we use this property to speed up the derivation of the optimal thresholds c_1 and c_2.

3.5 NPI method related to the three-group Youden index

Similarly as Section 2.3 introduced the NPI method for the two groups classification problem related to the Youden index procedure, in this section, we introduce a NPI method for the three-group classification problem related to the Youden index procedure. We apply the 3-NPI method presented in Section 3.4, especially Equations (3.12) and (3.13), to the three-group Youden index method, in the sense that the criterion of the Youden index maximises the sum of the probabilities of the correct classification for the three groups. Let the NPI-based lower and upper probabilities for the three-group Youden index be denoted by 3-NPI-Y-L and 3-NPI-Y-U, respectively, and the method in general by 3-NPI-Y and they are given by

$$\text{3-NPI-Y-L} = \underline{P}(C^X_{c_1} \geq \alpha m_x) + \underline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y) + \underline{P}(C^Z_{c_2} \geq \gamma m_z) \tag{3.20}$$

$$\text{3-NPI-Y-U} = \overline{P}(C^X_{c_1} \geq \alpha m_x) + \overline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y) + \overline{P}(C^Z_{c_2} \geq \gamma m_z) \tag{3.21}$$

These probabilities are calculated as explained in Section 3.4.

3.6 Examples

In the following examples, we illustrate the three NPI-based methods presented in Sections 3.3, 3.4 and 3.5, namely NPI-PW, 3-NPI and 3-NPI-Y, and compare them with the three groups empirical Youden index (3-EYI) and maximum volume (3-EMV) methods presented in Section 3.2. As it is irrelevant how c_1 and c_2 are chosen within the respective intervals, the reported values of c_1 and c_2 in these examples are set to be a value in the interval that is between two consecutive observations of the X, Y and Z data combined. In the tables for all the examples, we represent the interval for the optimal thresholds $c_1$and c_2 by their left end point.

Example 3.1. The n-acetyl aspartate over creatine (NAA/Cr) is a neuronal metabolism marker in the brain used to distinguish between different levels of human immunodeficiency virus (HIV) in patients [47, 54]. Decreased levels of NAA/Cr have been observed in patients with mild to severe AIDS dementia complex (ADC). The NAA/Cr levels were available for 137 patients, of whom 61 were HIV-positive subjects with AIDS dementia complex (ADC), 39 were HIV-positive non-symptomatic subjects (NAS), and 37 were HIV-negative individuals (NEG). The NAA/Cr levels were anticipated to be lowest among the ADC group and highest among the NEG group, with the NAS group being intermediate to the other two. We refer to these groups as X, Y and Z, respectively. Nakas et al. [54] used this data set to illustrate the generalized Youden index for thresholds selection in three-group classification problems. The maximum empirical Youden index is 1.434 at the threshold values $c_1 = 1.83$ and $c_2 = 1.99$. This data set has tie observations between the three groups, we avoid the ties by adding 0.001 to group Y and 0.002 to group Z. We also applied our method without this specific breaking of the ties, and observed that the results were close.

Figure 3.1 shows the box-plots of the NAA/Cr levels for ADC, NAS and NEG, where a noticeable overlap between the three groups can be observed, in particular between the NAS and NEG groups. We may not be surprised if we find that the diagnostic test may struggle to distinguish between the latter two groups.

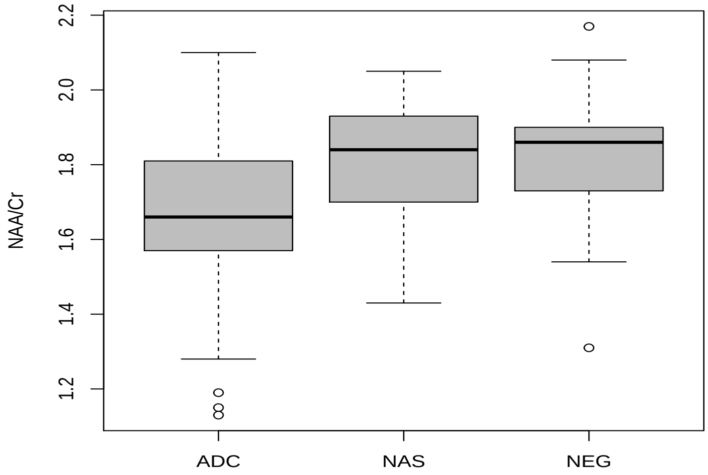

Figure 3.1: Box-plots of NAA/Cr levels for ADC, NAS and NEG

Method	Lower case			Upper case			Lower case			Upper case		
	c_1	c_2	value	c_1	c_2	value	c_1	c_2	value	c_1	c_2	value
$m = 5$	$\alpha = \beta = \gamma = 0.5$						$\alpha = \beta = \gamma = 0.7$					
3-NPI	1.66	1.861	0.0919	1.66	1.861	0.1318	1.66	1.861	0.0049	1.66	1.861	0.0089
3-NPI-Y	1.76	2.05	1.556	1.76	2.05	1.6566	1.66	1.661	1.0097	1.66	1.661	1.0770
NPI-PW (X, Y)	1.76	-	0.6188	1.76	-	0.6631	1.76	-	0.2267	1.76	-	0.2656
NPI-PW (Y, Z)	-	1.861	0.3373	-	1.861	0.3912	-	1.861	0.0708	-	1.861	0.0931
$m = 10$	$\alpha = \beta = \gamma = 0.5$						$\alpha = \beta = \gamma = 0.7$					
3-NPI	1.66	1.861	0.1629	1.66	1.861	0.2399	1.66	1.861	0.0038	1.66	1.861	0.0086
3-NPI-Y	1.76	2.05	1.7980	1.66	1.861	1.8705	1.83	1.83	1.1116	1.76	2.05	1.2473
NPI-PW (X, Y)	1.76	-	0.823	1.76	-	0.8582	1.76	-	0.3237	1.76	-	0.3826
NPI-PW (Y, Z)	-	1.861	0.4924	-	1.861	0.5646	-	1.861	0.0799	-	1.861	0.1126
$m = 25$	$\alpha = \beta = \gamma = 0.5$						$\alpha = \beta = \gamma = 0.7$					
3-NPI	1.66	1.861	0.0683	1.66	1.861	0.1361	1.66	1.861	0.00003	1.66	1.861	0.0003
3-NPI-Y	1.76	2.05	1.822	1.76	2.05	1.8913	2.1	2.1	0.9999	1.66	1.661	1.0164
NPI-PW (X, Y)	1.76	-	0.8532	1.76	-	0.8932	1.76	-	0.1687	1.76	-	0.2297
NPI-PW (Y, Z)	-	1.861	0.3951	-	1.861	0.4922	-	1.861	0.0124	-	1.861	0.0240

Table 3.1: Optimal thresholds (c_1, c_2) using NPI-based methods, where value represents the value of the ..NPI... corresponding to the specific cases

Tables 3.1 and 3.3 provide the optimal threshold values (c_1, c_2) obtained from the three NPI-based methods along with their corresponding lower and upper probabilities, for $m_x = m_y = m_z = m$. We have considered four different scenarios of α, β and γ. As we can see from Table 3.1, for $\alpha = \beta = \gamma = 0.5$, all the methods provide the same optimal thresholds c_1 and c_2 regardless of the m value; except the 3-NPI-Y-U method, the optimal thresholds are $c_1 \in (1.66, 1.661)$ and $c_2 \in (1.861, 1.862)$ for $m = 10$. It is noticed that the lower and upper NPI-PW method based on groups Y and Z are lower than that based on groups X and Y, which is due to the fact that groups Y and Z overlap more than groups X and Y.

For $\alpha = \beta = \gamma = 0.7$, the optimal thresholds (c_1, c_2) are the same as the $\alpha = \beta = \gamma = 0.5$ scenario for both the 3-NPI and NPI-PW methods. The optimal thresholds for the 3-NPI-Y vary with m, for example for $m = 10$ both the optimal thresholds for the NPI-Y-L are $c_1, c_2 \in (1.83, 1.831)$, and for $m = 25$ both the optimal thresholds are $c_1, c_2 \in (2.1, 2.17)$. We notice that the corresponding lower and upper probabilities for all the methods become lower than for scenario $\alpha = \beta = \gamma = 0.5$ as the required criteria become harder to achieve.

An interesting point is that the 3-NPI-Y method often tries to squeeze one of the groups in order to maximise the corresponding lower and upper probabilities (as it is based on summing up the individual probabilities rather than taking the product), while the 3-NPI method actually tries to balance between the groups in order to find the optimal thresholds c_1 and c_2. To illustrate this further, we have calculated the individual probabilities for the groups X, Y and Z, the optimal thresholds and the corresponding lower and upper probabilities of the 3-NPI and 3-NPI-Y methods, which are presented in Table 3.2, where (c_1^L, c_2^L) and (c_1^U, c_2^U) are the corresponding thresholds of the lower and upper probabilities, respectively. As we can see from this table, the 3-NPI-Y-L method squeezes group Y in order to obtain the optimal thresholds that maximise the lower probability in Equation (3.20) and focuses on maximising the number of correctly classified future observations from group X. Whereas, the 3-NPI-Y-U method squeezes group Z in order to obtain the optimal thresholds that maximise the upper probability in Equation (3.21), and focuses on maximising the number of correctly classified future observations from groups X and Y. On the other hand, the 3-NPI method tries to balance between the three groups in order to obtain the optimal thresholds that maximise both the lower and upper probabilities, but we also notice a slightly smaller value for the Y group in the lower and upper probabilities.

c_1^L	c_2^L	$\underline{P}(C^X_{c_1} \geq \alpha m_x)$	$\underline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y)$	$\underline{P}(C^Z_{c_2} \geq \gamma m_z)$	3-NPI-L	3-NPI-Y-L
1.66	1.861	0.1902	0.0997	0.2005	0.0038	–
1.83	1.83	0.8149	0.0000	0.2967	–	1.1116

c_1^U	c_2^U	$\overline{P}(C^X_{c_1} \geq \alpha m_x)$	$\overline{P}(C^Y_{(c_1,c_2)} \geq \beta m_y)$	$\overline{P}(C^Z_{(c_2)} \geq \gamma m_z)$	3-NPI-U	3-NPI-Y-U
1.66	1.861	0.2177	0.1611	0.2460	0.0086	–
1.76	2.05	0.5457	0.7011	0.0005	–	1.2473

Table 3.2: Comparison of 3-NPI and 3-NPI-Y methods, for $m = 10$ and $\alpha = \beta = \gamma = 0.7$.

Method	Lower case			Upper case			Lower case			Upper case		
	c_1	c_2	value	c_1	c_2	value	c_1	c_2	value	c_1	c_2	value
$m = 5$	$\alpha = \beta = 0.7$, $\gamma = 0.4$						$\alpha = \beta = \gamma = 0.2$					
3-NPI	1.76	1.941	0.0248	1.76	1.941	0.0423	1.66	1.861	0.8568	1.66	1.861	0.8929
3-NPI-Y	1.86	1.861	1.5442	1.86	1.861	1.6074	1.66	1.861	2.8498	1.66	1.861	2.8890
NPI-PW (X, Y)	1.76	-	0.2267	1.76	-	0.2656	1.76	-	0.9886	1.76	-	0.9913
NPI-PW (Y, Z)	-	1.861	0.2782	-	1.861	0.3248	-	1.861	0.9467	-	1.861	0.9587
$m = 10$	$\alpha = \beta = 0.7$, $\gamma = 0.4$						$\alpha = \beta = \gamma = 0.2$					
3-NPI	1.76	1.941	0.0164	1.76	1.941	0.0341	1.66	1.861	0.9167	1.66	1.861	0.9477
3-NPI-Y	1.83	1.83	1.6889	1.83	1.83	1.7494	1.66	1.861	2.9147	1.66	1.861	2.9468
NPI-PW (X, Y)	1.76	-	0.3237	1.76	-	0.3826	1.76	-	0.9981	1.76	-	0.9987
NPI-PW (Y, Z)	-	1.861	0.3187	-	1.861	0.3841	-	1.861	0.9761	-	1.861	0.9836
$m = 25$	$\alpha = \beta = 0.7$, $\gamma = 0.4$						$\alpha = \beta = \gamma = 0.2$					
3-NPI	1.76	1.941	0.0011	1.76	1.941	0.0044	1.66	1.861	0.9737	1.66	1.861	0.9889
3-NPI-Y	1.86	1.861	1.6940	1.86	1.861	1.7777	1.66	1.861	2.9735	1.66	1.861	2.9889
NPI-PW (X, Y)	1.76	-	0.1687	1.76	-	0.2297	1.76	-	1.0000	1.76	-	1.0000
NPI-PW (Y, Z)	-	1.861	0.1771	-	1.861	0.2438	-	1.84	0.9953	-	1.84	0.9975

Table 3.3: Optimal thresholds (c_1, c_2) using NPI-based methods, where value represents the value of the ..NPI... corresponding to the specific cases

From Table 3.3, for $\alpha = \beta = 0.7, \gamma = 0.4$, as this scenario request to put more emphasis on the number of correctly classified future observations from groups X and Y than that from groups Z, it is noticed that the optimal thresholds (c_1, c_2) for the 3-NPI method increase in order to achieve the desired criteria in comparison to $\alpha = \beta = \gamma$ scenario, as $c_1 \in (1.76, 1.761)$ and $c_2 \in (1.941, 951)$. Whereas, the NPI-Y method, also at this scenario, tries to squeeze the group Y in order to find the optimal thresholds (c_1, c_2), for example for $m = 10$, both the optimal thresholds for the NPI-Y-U are $c_1, c_2 \in (1.83, 1.831)$, while for $m = 5, 25$ the optimal thresholds are $c_1 \in (1.86, 1.861)$ and $c_1 \in (1.861, 1.862)$ which they are next to each other. The optimal thresholds for the NPI-PW method stay the same as in the $\alpha = \beta = \gamma$ scenario.

Finally, when the required criteria are easy to achieve ($\alpha = \beta = \gamma = 0.2$), all the methods perform well as the values of the lower and upper probabilities are very high, and the 3-NPI and 3-NPI-Y methods provide the same optimal threshold values c_1 and c_2 where $c_1 \in (1.66, 1.661)$ and $c_2 \in (1.861, 1.862)$. The optimal threshold values for the NPI-PW method stay the same in all the scenarios except that the NPI-PW (Y, Z) for $m = 25$ the optimal threshold c_2 changes to $c_2 \in (1.84, 1.841)$.

The maximum value of the empirical maximum volume (3-EMV) is equal to 0.1205 and at the optimal thresholds $c_1 \in (1.66, 1.661)$ and $c_2 \in (1.861, 1.862)$. The point to be highlighted is that the optimal threshold values (c_1, c_2) for the NPI method (with $\alpha = \beta = \gamma$) are the same as they are for the 3-EMV method since both criteria are based on the product of the probabilities which tends to provide more balanced classification between the three groups. However, the NPI-Y method based on the sum of the probabilities does not seem to be ideal to obtain the optimal thresholds that maximise the probability for every individual group.

m	2-NPI method				2-NPI-Y method			
	Lower case		Upper case		Lower case		Upper case	
	c	2-NPI-L	c	2-NPI-U	c	2-NPI-Y-L	c	2-NPI-Y-U
$\alpha = \beta = 0.5$								
5	1.76	0.6112	1.76	0.6446	1.76	1.5650	1.76	1.6067
10	1.76	0.8225	1.76	0.8504	1.76	1.8145	1.76	1.8448
25	1.76	0.8586	1.76	0.8910	1.76	1.8542	1.76	1.8542

Table 3.4: Selecting the optimal threshold c and corresponding value of 2-NPI-L, 2-NPI-U, 2-NPI-Y-L, 2-NPI-Y-U, using the 2-NPI and 2-NPI-Y methods, when NAS and NEG are combined

Now, we use this example with some change in the data to illustrate some further aspects of our approaches. Since this data set shows more overlapping between groups Y and Z and group X is more separated than these two groups, we combine groups Y and Z together and run the analysis again. Then the remaining NPI-based methods, 2-NPI and 2-NPI-Y as presented in Sections 2.2 and 2.3, are illustrated in Table 3.4. As we can see from this table, all NPI-based methods give the same optimal threshold $c \in (1.76, 1.761)$ regardless of the value of m. The maximum value of the empirical maximum area (2-EMA) is equal to 0.4573 at the same threshold value as the NPI-based methods ($c \in (1.76, 1.761)$), while the maximum value of the empirical Youden index (2-EYI) is equal to 0.3635, which gives a different threshold $c \in (1.66, 1.661)$.

3.48	7.38	7.93	8.57	9.73	10.95	12.43	13.03	13.60	14.38	15.42
15.84	17.19	17.84	18.42	18.71	28.76	39.16	41.87	43.24	50.23	60.31
65.27	66.69	82.00	87.29	97.55	101.10	104.50	109.00	115.10	135.80	139.00
219.10	226.70	301.80	311.80	313.30	322.30	325.70	326.80	330.70	332.50	335.40
336.60	337.50	337.60	339.90	340.80	341.10	355.00				

Table 3.5: IL-6 data set, where group X is black, group Y is blue and group Z is red

Example 3.2. The interleukin-6 (IL-6) is a common diagnostic test for detection of late onset sepsis (LOS) in neonates [57, 61, 67]. The cases in the study consisted of 52 neonates assessed as suspicious for LOS. They were classified into three groups, 22 confirmed sepsis (positive blood cultures for fungi and microbes), 9 possible sepsis (laboratory evidence of sepsis however negative blood cultures) and 21 non-infected neonates (no laboratory evidence of sepsis and negative blood cultures), one missing value is excluded from the confirmed sepsis group. We refer to these groups as X, Y and Z, respectively. Table 3.5 show the IL-6 data set for groups X, Y and Z, where a noticeable overlap between the three groups can be observed.

In this example the number of future individuals from groups X, Y and Z are considered to be equal to the number of individuals from groups X, Y and Z, respectively, so $m_x = 21, m_y = 9, m_z = 21$. Table 3.6 provides the optimal threshold values (c_1, c_2)

obtained from the three NPI-based methods along with their corresponding lower and upper probabilities for $m_x = 21, m_y = 9, m_z = 21$. We have considered three different scenarios of α, β and γ.

Method	Lower case			Upper case		
	c_1	c_2	value	c_1	c_2	value
$\alpha = \beta = \gamma = 0.6$						
3-NPI	82	322.3	0.0817	82	322.3	0.2565
3-NPI-Y	65.27	65.27	1.8191	82	322.3	2.0517
NPI-PW (X, Y)	82	-	0.6605	82	-	0.8391
NPI-PW (Y, Z)	-	226.7	0.2500	-	226.7	0.4308
$\alpha = \beta = \gamma = 0.8$						
3-NPI	82	322.3	0.0011	82	322.3	0.0115
3-NPI-Y	115.1	139	1.2920	65.27	65.27	1.3037
NPI-PW (X, Y)	82	-	0.1879	82	-	0.3975
NPI-PW (Y, Z)	-	226.7	0.0133	-	226.7	0.0441
$\alpha = \beta = 0.5, \gamma = 0.7$						
3-NPI	82	322.3	0.0515	82	226.7	0.1855
3-NPI-Y	41.87	41.87	1.8309	82	322.3	1.9998
NPI-PW (X, Y)	82	-	0.8241	82	-	0.9314
NPI-PW (Y, Z)	-	139	0.1861	-	139	0.3407

Table 3.6: Optimal thresholds (c_1, c_2) using NPI-based methods, where value represents the value of the ..NPI... corresponding to the specific cases, for $m_x = 21, m_y = 9, m_z = 21$

The 3-NPI and NPI-PW methods are noticed to provide the same optimal threshold (c_1, c_2) for both scenarios, $\alpha = \beta = \gamma = 0.6$ and $\alpha = \beta = \gamma = 0.8$. Whereas the 3-NPI-Y-L for $\alpha = \beta = \gamma = 0.6$ and the 3-NPI-Y-U for $\alpha = \beta = \gamma = 0.8$ squeeze group Y as both the optimal thresholds are $c_1, c_2 \in (65.27, 66.69)$. For $\alpha = \beta = 0.5$, $\gamma = 0.7$, as this scenario requests to put more emphasis on the number of correctly classified future observations from group Z than the number of correctly classified future observations from groups X and Y, the optimal threshold c_2 for the 2-NPI-U and NPI-PW methods decreases in order to achieve the desired criteria in comparison to the $\alpha = \beta = \gamma$ scenario, as $c_2 \in (226.7, 301.8)$ and $c_2 \in (139, 219.1)$, respectively. While, both the optimal thresholds for the 3-NPI-Y-L are $c_1, c_2 \in (41.87, 43.24)$, and for the 3-NPI-Y-U are $c_1 \in (82, 87.29)$ and $c_2 \in (322.3, 325.7)$. In addition, it is noticed in this table that the values of the lower and upper probabilities for the NPI-PW (X, Y) are higher than for the NPI-PW (Y, Z)

since this data set has more overlapping between groups Y and Z, and group X is a bit separated than these two groups.

Method	Lower case			Upper case		
	c_1	c_2	value	c_1	c_2	value
$\alpha = \beta = \gamma = 0.6$						
3-NPI	82	322.3	0.0742	82	322.3	0.2473
3-NPI-Y	65.27	65.27	1.8638	82	322.3	2.0876
NPI-PW (X, Y)	82	-	0.7528	82	-	0.9041
NPI-PW (Y, Z)	-	226.7	0.2698	-	226.7	0.4745
$\alpha = \beta = \gamma = 0.8$						
3-NPI	82	322.3	0.0003	82	226.7	0.0059
3-NPI-Y	115.1	341.1	1.067	82	341.1	1.4095
NPI-PW (X, Y)	82	-	0.2394	82	-	0.4925
NPI-PW (Y, Z)	-	226.7	0.0082	-	139	0.0336
$\alpha = \beta = 0.5, \gamma = 0.7$						
3-NPI	82	226.7	0.0268	82	226.7	0.1537
3-NPI-Y	41.87	41.87	1.8682	82	322.3	1.9611
NPI-PW (X, Y)	82	-	0.8526	82	-	0.9505
NPI-PW (Y, Z)	-	139	0.1539	-	139	0.3186

Table 3.7: Optimal thresholds (c_1, c_2) using NPI-based methods, where value represents the value of the ..NPI... corresponding to the specific cases, for $m_x = 33, m_y = 15, m_z = 52$

In Table 3.7, we increase the number of future individuals from groups X, Y and Z, with $m_x = 33, m_y = 15, m_z = 52$. Now comparing this table with Table 3.6, with respect of the optimal thresholds (c_1, c_2), the 3-NPI method provides the same optimal threshold (c_1, c_2) when we increase the values of m_x, m_y, m_z, except for $\alpha = \beta = \gamma = 0.8$ and $\alpha = \beta = 0.5, \gamma = 0.7$ the optimal threshold c_2 for the 3-NPI-U and 3-NPI-L change to $c_2 \in (226.7, 301.8)$. Also, the 3-NPI-Y method provides the same optimal threshold (c_1, c_2) when we increase the values of m_x, m_y, m_z, except that for $\alpha = \beta = \gamma = 0.8$ the optimal threshold c_2 for the 3-NPI-Y-L changes to $c_2 \in (341.1, 355)$ and for the 3-NPI-Y-U $c_1 \in (82, 87.29)$ and $c_2 \in (341.1, 355)$. The NPI-PW method also provides the same optimal thresholds (c_1, c_2) when we increase the values of m_x, m_y, m_z, except that for $\alpha = \beta = \gamma = 0.8$ the optimal threshold c_2 for the NPI-PW-L (Y, Z) change to $c_2 \in (139, 219.1)$.

It is clear from both the tables that the optimal thresholds (c_1, c_2) can change with changing the number of future individuals. The maximum values of the empirical maximum volume (3-EMV) are equal to 0.2993 at the thresholds $c_1 \in (82, 87.29)$ and $c_2 \in (322.3, 325.7)$, while the empirical Youden index (3-EYI) is equal to 2.0794 at the thresholds $c_1 \in (115.1, 135.8)$ and $c_2 \in (322.3, 325.7)$. Overall, in the criterion that consider the product of the number of correct classification between the three groups, i.e. 3-EMV and 3-NPI (with $\alpha = \beta = \gamma$), the corresponding optimal thresholds c_1 and c_2 seem to be widely apart, which yields more identification of the group Y than for the criterion based on the sum.

Example 3.3. Consider an artificial data set for groups X, Y and Z, with $n_x = 5$, $n_y = 7$ and $n_z = 8$, consisting of the ranks $X = \{5, 8, 11, 12, 15\}$, $Y = \{1, 2, 3, 4, 6, 10, 18\}$ and $Z = \{7, 9, 13, 14, 16, 17, 19, 20\}$. In this example, we show a special case where the optimal $c_1 > c_2$ for the NPI-PW method for $\alpha = \beta = \gamma = 0.6$, and we resolve this problem by investigating a different ordering of the three groups.

Method	Lower case			Upper case		
	c_1	c_2	value	c_1	c_2	value
3-NPI	5	10, 12	0.0028	5	10, 12	0.0717
3-NPI-Y	12	12	1.4887	0	6	1.9318
NPI-PW (X, Y)	15, 17	-	0.0417	8, 9	-	0.1553
NPI-PW (Y, Z)	-	6	0.6653	-	6	0.8485

Table 3.8: Optimal thresholds (c_1, c_2) using NPI-based methods, where value represents the value of the ..NPI... corresponding to the specific cases, for $m = 5$

Table 3.8 shows that for $m = 5$ and $\alpha = \beta = \gamma = 0.6$, the NPI-PW method $c_1 > c_2$, where the corresponding lower and upper probabilities for NPI-PW (X, Y) are very low, and these lower and upper probabilities for the NPI-PW (Y, Z) are high. Moreover, the corresponding lower and upper probabilities for the 3-NPI method are very low. While both the optimal thresholds for the NPI-Y-L are $c_1, c_2 \in (12, 13)$ (squeezing group Y), and the optimal threshold c_1 for the NPI-Y-U occurs in the first interval (squeezing group X). These results can be an indication for considering a different ordering of these three groups. For example, let $Y = \{5, 8, 11, 12, 15\}$, $X = \{1, 2, 3, 4, 6, 10, 18\}$ and

$Z = \{7, 9, 13, 14, 16, 17, 19, 20\}$. Table 3.9 shows that these lower and upper probabilities for the NPI-PW (X, Y), 3-NPI and 3-NPI-Y increase, while the NPI-PW (Y, Z) decrease a bit.

Method	Lower case			Upper case		
	c_1	c_2	value	c_1	c_2	value
3-NPI	4	12	0.1876	4	12	0.5568
3-NPI-Y	4	12	1.7506	4	12	2.4872
NPI-PW (X, Y)	6, 7	-	0.5088	6, 7	-	0.7778
NPI-PW (Y, Z)	-	12	0.5540	-	12	0.8077

Table 3.9: Optimal thresholds (c_1, c_2) using NPI-based methods , where value represents the value of the ..NPI... corresponding to the specific cases, for $m = 5$

Example 3.4. In this example, we show a special case where the optimal c_1 occurs in the first interval for the 3-NPI-U, 3-NPI-Y-U and NPI-PW-U methods. Consider an artificial data set for groups X, Y and Z with $n_x = 3$, $n_y = 7$ and $n_z = 4$, consisting of the ranks, $X = \{4, 6, 8\}$, $Y = \{1, 2, 3, 5, 7, 9, 12\}$ and $Z = \{10, 11, 13, 14\}$.

Method	Lower case			Upper case		
	c_1	c_2	value	c_1	c_2	value
3-NPI	4	12	0.0631	0	12	0.5952
3-NPI-Y	8, 9	8, 9	1.9742	0	12	2.5774
NPI-PW (X, Y)	4	-	0.1547	0	-	0.6250
NPI-PW (Y, Z)	-	12	0.7071	-	12	0.9524

Table 3.10: Optimal thresholds (c_1, c_2) using NPI-based methods, where value represents the value of the ..NPI... corresponding to the specific cases, for $m = 5$

Table 3.10 shows that, for $m = 5$ and $\alpha = \gamma = 0.2, \beta = 0.8$, the optimal threshold value c_1 occurs in the first interval for the corresponding upper probabilities for all methods. Whereas, the optimal threshold value c_1 will never be at the first interval for the corresponding lower probability for the 3-NPI and NPI-PW methods, as the value of the lower probability for these methods in this case would be equal to zero because it would imply $\underline{P}(C^X_{c1} \leq \alpha m_x) = 0$

Example 3.5. In this example, we show a special case where the optimal threshold values c_1 and c_2 occur in the same interval for the 3-NPI-U, 3-NPI-Y-L and 3-NPI-Y-U methods. Consider an artificial data set for groups X, Y and Z, with $n_x = 9$, $n_y = 2$ and $n_z = 2$, consisting of the ranks $X = \{1, 2, 4, 6, 7, 8, 9, 10, 11\}$, $Y = \{3, 5\}$ and $Z = \{12, 13\}$.

Method	Lower case			Upper case		
	c_1	c_2	value	c_1	c_2	value
3-NPI	2	5, 11	0.0156	11	11	0.7143
3-NPI-Y	11	11	1.8425	11	11	2.7143
NPI-PW (X, Y)	4	-	0.045	11, 13	-	0.7143
NPI-PW (Y, Z)	-	5, 11	0.9070	-	5, 11	1.0000

Table 3.11: Optimal thresholds (c_1, c_2) using NPI-based methods, where value represents the value of the ..NPI... corresponding to the specific cases, for $m = 5$

Table 3.11 shows that for $m = 5$, when $\alpha = \gamma = 0.2, \beta = 0.8$, the optimal thresholds c_1 and c_2 occur in the same interval for the 3-NPI-U method, whereas the optimal thresholds c_1 and c_2 would never be in the same interval for the 3-NPI-L, as the value of the lower probability for this method in this case would be equal to zero because $\underline{P}(C^Y_{c_1,c_2} \leq \beta m_y) = 0$. In comparison, the optimal thresholds c_1 and c_2 occur in the same interval for the corresponding lower and upper probability for the 3-NPI-Y.

3.7 Simulation

This section extends the simulation study for two groups presented in Section 2.6, to the three-group scenario. Following the same simulation process for groups X and Y as presented in Section 2.6, we group Z as follows. The two main cases in which the data are simulated are:

Case A: $X \sim N(0, 2^2)$, $Y \sim N(1, 2^2)$, and $Z \sim N(3, 2^2)$.

Case B: $X \sim N(0, 1^2)$, $Y \sim N(1, 1^2)$, and $Z \sim N(3, 1^2)$.

Due to the larger variance in Case A, the groups in that case overlap more than in Case B. We simulate n_x, n_y and n_z from the two normal distributions. Then the n_x, n_y

and n_z simulated data observations will be used to find the optimal thresholds c_1 and c_2 according to these methods and for specific values of (α, β, γ) when applicable, where the threshold values are set to the midpoint in the partition of $\mathbb{R}$ used by the data. After that, we simulate m_x, m_y and m_z future observations from the same underlying normal distributions as the n_x, n_y and n_z simulated data observations to see how the methods perform.

The m_x, m_y and m_z simulated future observations are compared with the optimal thresholds to obtain the number of correctly classified observations per group. We have studied the predictive performance of all methods in terms of the number of correctly classified future observations that are achieved using the desired criteria, that is when the number of correctly classified future observations from group X, Y, and Z exceed αm_x, βm_y and γm_z, respectively. Let us denote by '$+$' when the desired criteria are achieved and '$-$' otherwise. Throughout this section we assume that $n_x = n_y = n_z = n$ and $m_x = m_y = m_z = m$, and $j_x, j_y, j_z \in \{0, 1, \dots, m\}$.

We have run the simulation for $n = 20$ and $m = 10, 30$ and we have chosen different values of α, β and γ. Obviously the empirical Youden index and the maximum volume methods do not depend on the values of α, β and γ in terms of selecting the optimal thresholds, however, for the comparison of predictive performance we have considered the same desired criterion of the number of future observations that are correctly classified from groups X, Y and Z being at least αm_x, βm_y and γm_z, respectively. The results in this section are based on 10,000 simulations per case per method.

To search for the optimal threshold c, rather than searching for the value c that maximises the probability within each of the $n_x + n_y + n_z + 1$ intervals created by the data observations, which could be computationally demanding especially in the simulation, we just consider the intervals as discussed in Section 3.4. We excluded the possibility for the optimal thresholds c_1 and c_2 to occur in the same interval. It should be mentioned that for the NPI-PW method, it may occur that $c_1 > c_2$, due to the fact that the optimal thresholds c_1 and c_2 are obtained separately. In this case we set the optimal threshold $\tilde{c}_2 := c_1$, as threshold between the X and Y groups.

The predictive performance results for Case A are given in Tables 3.12 and 3.13 for $m = 10$ and $m = 30$, respectively, and in Tables 3.14 and 3.15 for Case B. We have studied the performance in two shapes for $\alpha = \beta = \gamma$ with values $0.2, 0.6$ and 0.8, and for $\alpha = \beta = 0.5, \gamma = 0.7$, for the NPI-based methods (3-NPI, 3-NPI-Y and NPI-PW) and the empirical estimates of Youden index and maximum volume methods.

Consider Table 3.12, for example, where '$+\ +\ +$' indicates that the desired criteria have been achieved for all groups while '$-\ -\ -$' indicates that the desired criterion for all groups have not been achieved. For example, for 3-NPI-L and $\alpha = \beta = \gamma = 0.2$ the desired criterion have been achieved for all groups is 9303 out of $10,000$ simulations, that is at least 2 future observations ($\alpha m = 0.2 \times 10$, $\beta m = 0.2 \times 10$ and $\gamma m = 0.2 \times 10$) have been correctly classified from each of the three groups. On the other hand, in 160 out of $10,000$ simulations the desired criterion is achieved (2 or more out of 10 are correctly classified) for groups X and Y but the desired criterion has not been achieved for group Z.

From Tables 3.12-3.15, we observe a similar behaviour in the two-groups scenario. Generally, the 2-NPI method performs better than the other methods, while for small values of α and β all methods preform equally well. So for $\alpha = \beta = \gamma = 0.2$, all the methods perform similarly since the desired criteria are easily achieved, while for $\alpha = \beta = \gamma = 0.6$, the 3-NPI method can achieve the desired criteria better than the other methods. The results in these tables suggest that in general the 3-EMV method is the closest to the 3-NPI method with regards to the performance, yet the NPI method can be better considering its predictive nature. Interestingly, the NPI-PW method has better performance than the empirical Youden index (3-EYI) for $\alpha = \beta = \gamma = 0.6$ for Case A.

We also notice that for $\alpha = \beta = \gamma = 0.6$, in Tables 3.12 and 3.13 (Case A), the 3-NPI-Y tends to squeeze the middle group Y substantially, the reason is that the 3-NPI-Y method is based on maximising the sum of the probabilities of correct classification rather than the product, which does not seem ideal if one tries to achieve higher proportions of those who are correctly classified, and that is clearly shows in the three groups setting as the 3-NPI-Y method does not tend to achieve higher proportions of those who are correctly classified from the three groups simultaneously. While the empirical Youden index tends to squeeze group Y in some occasions or squeeze groups X and Y and achieve the desired criterion for just group Z in other occasions. The 3-NPI, NPI-PW and 3-EMV methods tend to squeeze both groups X and Y and achieve the desired criterion for just group Z in some occasions. Also, the 3-NPI and 3-EMV methods fail the desired criterion for each group in other occasions.

For $\alpha = \beta = \gamma = 0.8$, all methods struggle to meet the required criteria, especially in Case A where the groups have more overlap. For example, the 3-NPI, NPI-PW and 3-EMV methods mostly fail the desired criterion for each group. The 3-NPI-U tends to squeeze group Y substantially, while the 3-NPI-Y-L tend to squeeze both the groups X and Y and achieve the desired criterion for just group Z in some occasions or squeeze both the groups Y and Z and achieve the desired criterion for just group X in other occasions. For $\alpha = \beta = 0.5$ and $\gamma = 0.7$ in the tables, all methods achieve the desired criteria more than for $\alpha = \beta = \gamma = 0.6$ in both the cases, due to the fact that the group Z is more separated in comparison to the other two groups and also the value of γ is higher.

In addition, we observe similar behaviour as discussed in Section 2.6 for the two-group scenario, that is for $\alpha = \beta = \gamma = 0.6$ and $\alpha = \beta = \gamma = 0.8$, all the methods perform better for small value of m than for larger m while for $\alpha = \beta = \gamma = 0.2$ all the methods perform better for larger m than for smaller m. That is because of the randomness effect as discussed in Section 2.1. Obviously, all methods perform much better in Case B than in Case A, as the groups in Case B are more separated.

We summarise the number of correctly classified future observations in all simulations from groups X, Y and Z using bar-plots as follows. Let the number of successfully classified future observations from group X with regard to the event of interest, which include α denoted by $S^X_{j_x}$, the number of successfully classified future observations from group Y with regard to the event of interest, which include β denoted by $S^Y_{j_y}$, and the number of successfully classified future observations from group Z with regard to the event of interest, which include γ denoted by $S^Z_{j_z}$, where $j_x \in \{0, 1, \ldots, m_x\}$, $j_y \in \{0, 1, \ldots, m_y\}$

and $j_z \in \{0, 1, \ldots, m_z\}$, respectively. Figures 3.2 - 3.5 show the distributions of the number of future observations out of m in all 10,000 simulations, that are correctly classified for each group.

For $\alpha = \beta = \gamma = 0.6$, Figure 3.2 clearly shows the squeezing behaviour of the 3-NPI-Y method for group Y, leading to correctly classifying more future observations from groups X and Z. This can be an indication that for most of the simulation runs the optimal c_1 and c_2 are next to each other and it is more likely that there is no future observation of group Y between them. This also supports the results explained above in Tables 3.12 and 3.13, that for maximisation of the sum of the probabilities of groups X, Y and Z does not seem ideal to achieve higher proportions of those who are correctly classified from the three groups, which can cause us to correctly classify more future individuals from groups X and Z and leading to squeezing of the group Y. The figure shows that the 3-NPI-Y squeezes group Y more than 6000 out of 10,000 times, whereas it correctly classifies groups X and Z more than 4000 times out of $10,000$. Also, the figure shows the squeezing behaviour for the 3-EYI and NPI-PW methods but not as much as for 3-NPI-Y. The 3-NPI and 3-MV methods try to balance classification between the three groups.

For $\alpha = \beta = \gamma = 0.8$, Figure 3.3 shows that the behaviour of the 3-NPI-Y method in squeezing group Y and correctly classifying more future observations from groups X and Z, becomes much clearer than for $\alpha = \beta = \gamma = 0.6$. The 3-NPI and 3-MV methods, similar to the $\alpha = \beta = \gamma = 0.6$ scenario, try to balance classification between the three groups. Figure 3.4 (Case B) shows that for $\alpha = \beta = \gamma = 0.6$, the performance becomes better than Case A for all methods, as the groups in this case are more separated. In addition, the number of correctly classified future observations from group Z is much larger than that from groups X and Y, as group Z is more separated from the other two groups. For $\alpha = \beta = \gamma = 0.8$, Figure 3.5 shows the performance becomes poor for all methods, while again the 3-NPI-Y method shows the squeezing of group Y and correctly classify more future individuals from the other two groups.

Over all, the predictive performance for the NPI-based methods depend on the number of future observations considered and the values of α, β and γ. More attention should be paid to maximising the sum of the probabilities of the correct classification for the three groups, which may lead to squeezing the intermediate group.

X	Y	Z	3-NPI-L	3-NPI-U	3-NPI-Y-L	3-NPI-Y-U	NPI-PW-L	NPI-PW-U	3-EYI	3-EMV
						$\alpha = \beta = \gamma = 0.2$				
-	-	-	0	0	0	0	0	0	0	0
-	-	+	0	1	0	1	6	7	43	15
-	+	-	1	0	1	0	0	0	2	0
-	+	+	251	183	257	187	51	51	645	337
+	-	-	5	6	5	6	3	3	9	9
+	-	+	280	382	281	380	2421	2419	3124	903
+	+	-	160	128	158	126	6	6	58	110
+	+	+	9303	9300	9298	9300	7513	7514	6119	8626
						$\alpha = \beta = \gamma = 0.6$				
-	-	-	1323	1245	217	387	579	575	530	1012
-	-	+	2360	2440	1608	889	2981	2985	3061	2856
-	+	-	969	772	206	245	144	138	505	585
-	+	+	1154	1007	598	345	329	318	889	940
+	-	-	1631	1754	984	492	1267	1241	1103	1765
+	-	+	1574	1860	6251	7374	4380	4425	3569	2135
+	+	-	556	485	67	136	107	99	171	361
+	+	+	433	437	69	132	213	219	172	346
						$\alpha = \beta = \gamma = 0.8$				
-	-	-	6375	6225	1380	12	4596	4602	3968	5835
-	-	+	1915	2021	3780	349	3104	3113	3252	2305
-	+	-	411	307	150	2	87	78	318	294
-	+	+	71	62	33	1	23	23	65	52
+	-	-	1094	1214	3054	252	1496	1479	1290	1317
+	-	+	124	157	1603	9384	692	703	1105	187
+	+	-	10	13	0	0	2	2	2	10
+	+	+	0	1	0	0	0	0	0	0
						$\alpha = \beta = 0.5\ \gamma = 0.7$				
-	-	-	775	741	356	417	313	321	442	656
-	-	+	1628	1758	1320	978	1594	1584	1512	1165
-	+	-	995	749	535	431	99	96	827	971
-	+	+	1228	1075	728	588	209	203	897	857
+	-	-	1675	1738	1431	970	1665	1663	1815	2487
+	-	+	1689	2114	4482	5598	5563	5573	3422	1728
+	+	-	1229	1040	845	584	224	218	689	1485
+	+	+	781	785	303	434	333	342	396	651

Table 3.12: Simulation results (10, 000 runs) for case A with $m = 10$ and $n = 20$

X	Y	Z	3-NPI-L	3-NPI-U	3-NPI-Y-L	3-NPI-Y-U	NPI-PW-L	NPI-PW-U	3-EYI	3-EMV
$\alpha = \beta = \gamma = 0.2$										
-	-	-	0	0	0	0	0	0	0	0
-	-	+	0	0	0	0	0	0	14	7
-	+	-	0	0	0	0	0	0	1	0
-	+	+	73	44	75	44	4	4	583	178
+	-	-	0	0	0	0	0	0	2	5
+	-	+	64	120	64	121	2359	2358	3210	664
+	+	-	35	27	35	27	1	1	26	29
+	+	+	9828	9809	9826	9808	7636	7637	6164	9117
$\alpha = \beta = \gamma = 0.6$										
-	-	-	2284	2160	149	447	644	633	664	1462
-	-	+	3026	3158	1311	770	3790	3825	3856	3778
-	+	-	1078	815	142	148	68	67	487	549
-	+	+	619	481	524	118	87	85	506	434
+	-	-	1809	1985	691	386	1197	1166	1206	2085
+	-	+	951	1191	7168	8091	4191	4201	3232	1569
+	+	-	193	173	14	32	14	14	41	100
+	+	+	40	37	1	8	9	9	8	23
$\alpha = \beta = \gamma = 0.8$										
-	-	-	8618	8551	1386	8	6890	6959	5507	8041
-	-	+	935	992	4675	249	2154	2135	2808	1349
-	+	-	65	39	75	1	3	2	104	41
-	+	+	2	0	2	0	0	0	2	1
+	-	-	378	414	3473	170	811	774	927	563
+	-	+	2	4	389	9572	142	130	652	5
+	+	-	0	0	0	0	0	0	0	0
+	+	+	0	0	0	0	0	0	0	0
$\alpha = \beta = 0.5$ $\gamma = 0.7$										
-	-	-	1136	1124	216	516	253	277	454	864
-	-	+	2115	2299	1118	989	1665	1646	1833	1384
-	+	-	1306	909	440	423	63	59	1040	1209
-	+	+	822	684	520	323	53	46	619	537
+	-	-	2235	2355	1338	1038	1853	1830	2344	3490
+	-	+	1370	1748	5644	6308	6002	6031	3207	1340
+	+	-	811	683	685	313	69	67	436	1038
+	+	+	205	198	39	90	42	44	67	138

Table 3.13: Simulation results (10,000 runs) for case A with $m = 30$ and $n = 20$

X	Y	Z	3-NPI-L	3-NPI-U	3-NPI-Y-L	3-NPI-Y-U	NPI-PW-L	NPI-PW-U	3-EYI	3-EMV
$\alpha = \beta = \gamma = 0.2$										
-	-	-	0	0	0	0	0	0	0	0
-	-	+	0	0	0	0	0	0	0	0
-	+	-	0	0	0	0	0	0	0	0
-	+	+	26	17	26	17	13	12	135	54
+	-	-	0	0	0	0	0	0	0	0
+	-	+	26	46	28	46	208	208	539	158
+	+	-	5	5	5	4	0	0	0	0
+	+	+	9943	9932	9941	9933	9779	9780	9326	9788
$\alpha = \beta = \gamma = 0.6$										
-	-	-	54	55	37	47	25	23	18	34
-	-	+	651	657	629	684	913	925	1072	869
-	+	-	239	192	169	171	47	41	73	110
-	+	+	2006	1720	1803	1796	1195	1162	2024	1775
+	-	-	287	297	210	282	125	118	130	246
+	-	+	2510	2827	3663	3035	4569	4576	4173	3555
+	+	-	551	534	373	453	120	118	151	297
+	+	+	3702	3718	3116	3532	3006	3037	2359	3114
$\alpha = \beta = \gamma = 0.8$										
-	-	-	1799	1777	82	271	1097	1102	961	1456
-	-	+	3425	3405	1107	878	4165	4209	4271	3775
-	+	-	758	636	32	107	275	276	425	513
-	+	+	980	863	100	174	605	590	910	843
+	-	-	1185	1273	783	338	925	920	816	1220
+	-	+	1472	1673	7884	8176	2701	2668	2450	1903
+	+	-	204	198	6	26	82	81	58	146
+	+	+	177	175	6	30	150	154	109	144
$\alpha = \beta = 0.5$ $\gamma = 0.7$										
-	-	-	13	11	9	11	5	6	9	11
-	-	+	262	253	269	265	309	292	297	209
-	+	-	131	107	114	105	23	22	127	150
-	+	+	1377	1158	1369	1181	639	642	1453	1097
+	-	-	193	189	194	188	125	134	223	362
+	-	+	1857	2165	2213	2239	4468	4381	3260	2313
+	+	-	842	773	759	749	216	225	576	1045
+	+	+	5325	5344	5073	5262	4215	4298	4055	4813

Table 3.14: Simulation results ($10,000$ runs) for case B with $m = 10$ and $n = 20$

X	Y	Z	3-NPI-L	3-NPI-U	3-NPI-Y-L	3-NPI-Y-U	NPI-PW-L	NPI-PW-U	3-EYI	3-EMV
$\alpha = \beta = \gamma = 0.2$										
-	-	-	0	0	0	0	0	0	0	0
-	-	+	0	0	0	0	0	0	0	0
-	+	-	0	0	0	0	0	0	0	0
-	+	+	0	0	1	0	0	0	75	19
+	-	-	0	0	0	0	0	0	0	0
+	-	+	0	2	0	2	55	54	390	61
+	+	-	0	0	0	0	0	0	0	0
+	+	+	10000	9998	9999	9998	9945	9946	9535	9920
$\alpha = \beta = \gamma = 0.6$										
-	-	-	21	20	16	15	5	5	5	9
-	-	+	595	591	486	632	943	948	1202	929
-	+	-	211	158	119	125	13	12	57	66
-	+	+	2445	2008	2025	2047	1128	1092	2324	2040
+	-	-	266	281	166	249	46	43	84	214
+	-	+	2767	3282	4678	3597	5724	5740	4816	4303
+	+	-	517	483	254	387	54	56	79	198
+	+	+	3178	3177	2256	2948	2087	2104	1433	2241
$\alpha = \beta = \gamma = 0.8$										
-	-	-	3533	3496	32	238	1825	1851	1558	2602
-	-	+	4092	4090	984	517	5600	5657	5493	4770
-	+	-	497	386	27	37	135	123	321	309
-	+	+	290	231	66	66	124	122	301	225
+	-	-	957	1074	629	163	659	639	640	1042
+	-	+	616	706	8262	8979	1652	1604	1684	1041
+	+	-	13	15	0	0	5	4	3	9
+	+	+	2	2	0	0	0	0	0	2
$\alpha = \beta = 0.5$ $\gamma = 0.7$										
-	-	-	2	2	1	2	0	0	0	1
-	-	+	140	148	150	158	185	174	218	121
-	+	-	89	62	87	59	5	3	156	125
-	+	+	1386	1088	1384	1154	409	418	1590	1075
+	-	-	131	137	113	132	54	57	215	378
+	-	+	1745	2149	2345	2208	5569	5438	3685	2530
+	+	-	817	731	702	704	127	134	475	1079
+	+	+	5690	5683	5218	5583	3651	3776	3661	4691

Table 3.15: Simulation results (10, 000 runs) for case B with $m = 30$ and $n = 20$

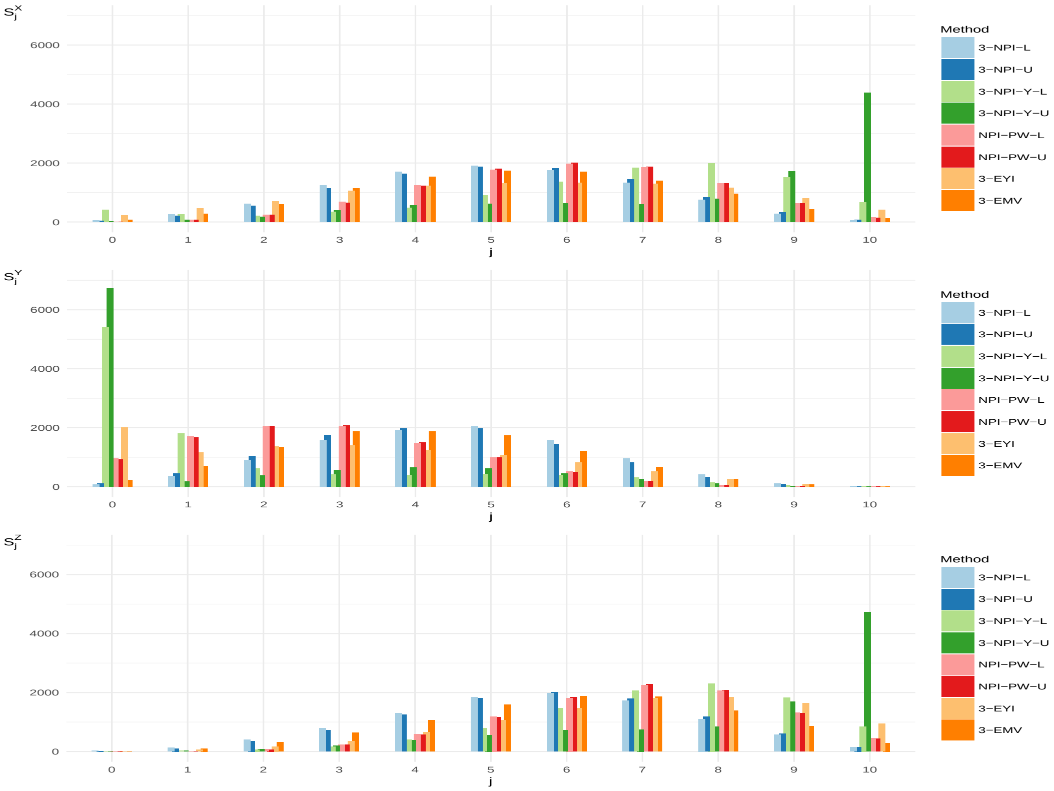

Figure 3.2: Simulation results (10, 000 runs), when $\alpha = \beta = \gamma = 0.6$ and $m = 10$ (case A)

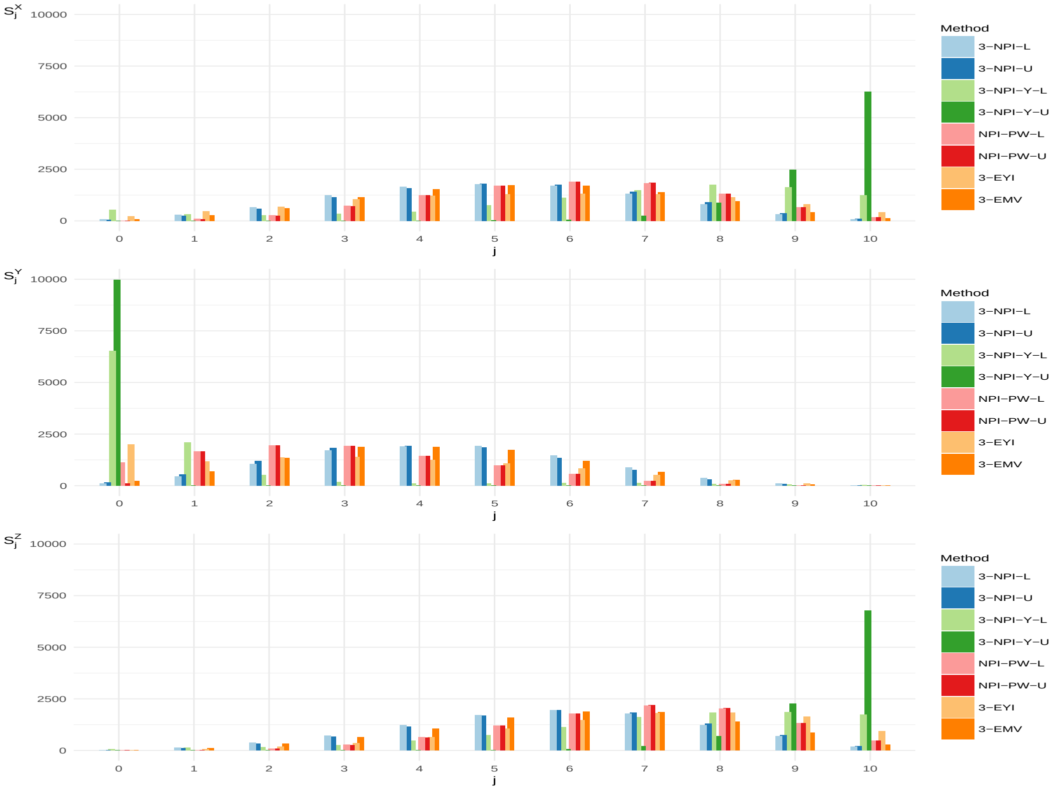

Figure 3.3: Simulation results (10, 000 runs), when $\alpha = \beta = \gamma = 0.8$ and $m = 10$ (case A)

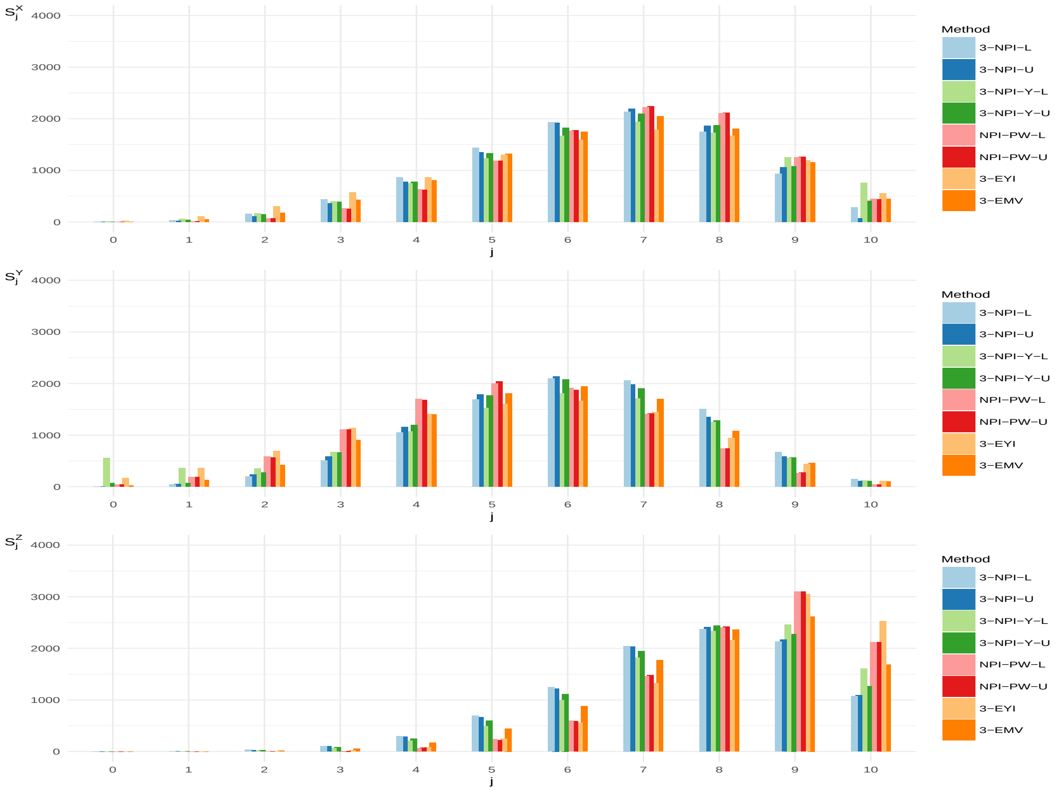

Figure 3.4: Simulation results (10, 000 runs), when $\alpha = \beta = \gamma = 0.6$ and $m = 10$ (case B)

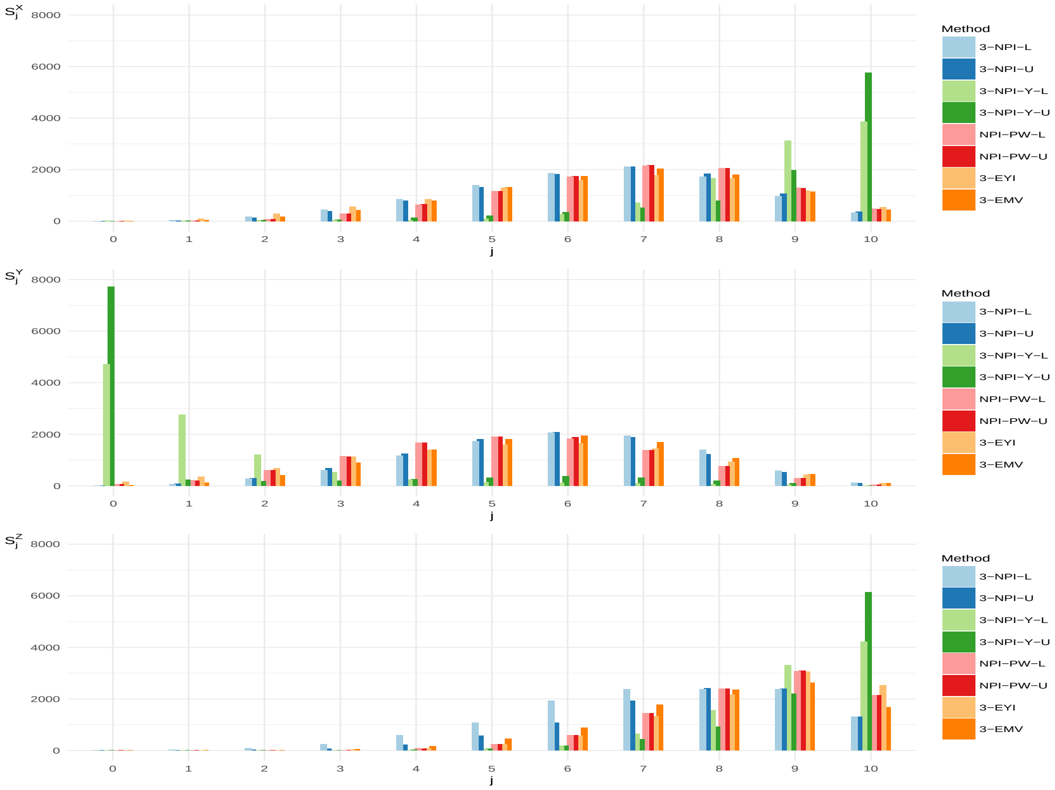

Figure 3.5: Simulation results (10, 000 runs), when $\alpha = \beta = \gamma = 0.8$ and $m = 10$ (case B)

3.8 Concluding remarks

This chapter extended the NPI methods for the selection of optimal threshold for two-group classification problems, presented in Chapter 2, to the three groups scenario. We have considered a specific number of future individuals in each group. We have shown in the examples that the optimal thresholds c_1 and c_2 can change with changing the number of future individuals. The performance for the three groups NPI methods was evaluated through simulation studies. These revealed that, in the case of the three groups scenario for which the 3-NPI-Y and classical Youden index approach have been used, the intermediate group may have very poor predictive performance. The 3-NPI method overcomes such problem since the optimal thresholds c_1 and c_2 yield more reasonable identification of the intermediate group.

We have also discussed the NPI-PW method such that the optimal thresholds c_1 and c_2 are selected independently, which may not satisfy the condition that $c_1 < c_2$. Whereas, the 3-NPI method selects the optimal thresholds c_1 and c_2 jointly which tends to produce a balanced classification of the three groups.

In the simulation study we only considered the normal distributions to investigate the general performance of the proposed methods, it is also interesting to simulate from other distributions. For example, for skewed distributions, it is interesting to study the effect of the values of α, β and γ for setting the optimal threshold values (c_1, c_2).

This line of work provides many questions and opportunities for future research. For example, setting meaningful target proportions for the predictive inferences should be discussed. Further research might to be developed similar approaches for different kind of data, e.g. ordinal data [32]. If one measures multiple markers per patient, their optimal combination together with optimal selection of thresholds is of interest, while also taking dependence of such multivariate data [23] into account provides interesting challenges. A further challenge is to develop such methods for data containing right-censored observations [19, 20]. Some of these topics require further development of NPI, including methods for multivariate data and for multiple future observations based on right-censored data. Generally, considering such problems from a predictive perspective, in particular how the number of future individuals considered might influence the optimal thresholds, provides interesting new insights which may also have substantial practical relevance.

For disease with k groups ($k > 3$), the 3-NPI method (Equations (3.12) and (3.13)) can be easily generalised by considering

$$\underline{P}(C_{c_1}^{X} \geq \alpha m_x, C_{(c_1,c_2)}^{Y} \geq \beta m_y, C_{(c_2,c_3)}^{Z} \geq \gamma m_z, C_{(c_3,c_4)}^{V} \geq \zeta m_v, \ldots, C_{c_k}^{W} \geq \xi m_w).$$
$$\overline{P}(C_{c_1}^{X} \geq \alpha m_x, C_{(c_1,c_2)}^{Y} \geq \beta m_y, C_{(c_2,c_3)}^{Z} \geq \gamma m_z, C_{(c_3,c_4)}^{V} \geq \zeta m_v, \ldots, C_{c_k}^{W} \geq \xi m_w).$$

Nakas et al. [55] also introduced the Youden index method for $k > 3$ groups by maximiseing the total number of correct classification rates for the k groups. This method is more likely to face the squeezing problem as it separates out into multiple optimisation problems, whereas the generalisation of the NPI method might perform better in term of reducing such squeezing.

Chapter 4

NPI for comparison of two diagnostic tests

4.1 Introduction

Developing and improving diagnostic tests to detect a particular disease are important in medical applications. Often, researchers are asked to confirm the superiority of a new diagnostic test to the existing test. In practice, most diagnostic tests do not always provide the correct classification. The tests can have two types of possible errors, false-negative errors (FN) and false-positive errors (FP). This raises the question how one can compare the qualities of two or more diagnostic tests. Various methods to compare two diagnostic tests have been presented in the literature [59, 73]. The performance of diagnostic tests can be evaluated by a single indicator such as sensitivity, specificity, positive and negative likelihood ratio or positive and negative predictive values. Such comparisons of two tests are rarely straightforward as one test may have higher specificity while the other test has higher sensitivity.

Measures such as the Youden index, have been suggested as global measures of diagnostic accuracy [73]. However, the Youden index can be misleading when comparing two diagnostic tests. The Youden index is not taking into account the differences in the specificity and sensitivity of the diagnostic test, and treats the FN and FP errors as equally undesirable. For example, assume that test A has a specificity of 0.9 and sensitivity of 0.4 and test B has specificity of 0.6 and sensitivity of 0.7. The Youden index of for each of these tests is 0.3. It is obvious that these tests have different discriminative properties.

The area under the ROC curve (AUC) also provides a summary measure of the diagnostic test ability [73]. Although the AUC has been used to compare different diagnostic tests, it has some limitations. For example, the areas under the ROC curves of two diagnostic tests can be equal, yet the shapes of the two ROC curves can be different over a certain part of the ROC curves of clinical relevance. According to Dodd and Pepe [31], the area under the ROC curve might summarize the performance of a diagnostic test over regions of the curve of no clinical and practical interest. Alternatively, the partial area under the ROC curve can provide more information for some diagnostic tests which require false-positive rates to be within the medical interest range [31, 45]. In addition, researchers use hypothesis testing to compare sensitivity, specificity or the area under the curve of two diagnostic tests [73].

In this chapter, we present NPI for comparing two diagnostic tests. The predictive nature of the NPI approach can be attractive for diagnostic tests as one tends to assess the quality of the diagnostic tests for a given number of future individuals. In Section 4.2, we introduce NPI of two diagnostic tests based on order statistics. NPI for comparison of two diagnostic tests based on Bernoulli quantities is presented in Section 4.3. Section 4.4 introduces weights to reflect the relative importance of two groups. Section 4.5 presents some examples to illustrate and discuss the new approaches. Finally, some concluding remarks are made in Section 4.6.

4.2 NPI of two diagnostic tests based on order statistics

In this chapter, we compare the accuracy of two diagnostic tests explicitly considering multiple future individuals. We assume that both diagnostic tests are applied to the same people. Assume that we have real-valued data from two different diagnostic tests on individuals from two independent groups in each test, and there are n_x observations from the healthy group X and n_y observations from the disease group Y. We refer to the two tests with superscript t; $t = 1, 2$, so we assume that we have data $(x_i^1, x_i^2), i = 1, ...n_x$ and $(y_j^1, y_j^2), j = 1, ...n_y$, where superscript 1 indicates test results of diagnostic test one and 2 indicates test results of diagnostic test two. We assume that the outcomes of the two tests are independent given the disease state of the individuals. The intention of this section is to compare between two diagnostic tests for m_x and m_y future individuals. The natural question is whether one test is better than the other for the m_x and m_y future individuals from groups X and Y, respectively, and we investigate the possible influence of the choice of m. We use the 2-NPI lower and upper method introduced in Section 2.2 for each diagnostic test. Of course other methods to determine the diagnostic test threshold can be used instead. The same notations and definitions will be used as in Section 2.2, with the superscript t to differentiate between the two tests. For a specific value of threshold c^t and for fixed α and β, the 2-NPI lower and upper probabilities for the event $C_{c^t}^{X^t} \geq \alpha m_x$, $C_{c^t}^{Y^t} \geq \beta m_y$ are given by

$$\underline{P}(C_{c^t}^{X^t} \geq \alpha m_x, C_{c^t}^{Y^t} \geq \beta m_y) = \underline{P}(C_{c^t}^{X^t} \geq \alpha m_x) \times \underline{P}(C_{c^t}^{Y^t} \geq \beta m_y) \tag{4.1}$$

$$\overline{P}(C_{c^t}^{X^t} \geq \alpha m_x, C_{c^t}^{Y^t} \geq \beta m_y) = \overline{P}(C_{c^t}^{X^t} \geq \alpha m_x) \times \overline{P}(C_{c^t}^{Y^t} \geq \beta m_y) \tag{4.2}$$

As we introduced the 2-NPI method in Section 2.2, we are going to use the NPI results for future order statistics in Section 1.3.2, in particular Equation (1.17), to derive the NPI lower and upper probabilities in Equations (4.1) and (4.2). The NPI lower and upper probabilities for the event $C_{c^t}^{X^t} \geq \alpha m_x$ are given by

$$\underline{P}(C_{c^t}^{X^t} \geq \alpha m_x) = \underline{P}(X_{(\lceil \alpha m_x \rceil)} \leq c) = \sum_{i=1}^{i_c - 1} P(X_{(\lceil \alpha m_x \rceil)} \in I_i^X) \tag{4.3}$$

$$\overline{P}(C_{c^t}^{X^t} \geq \alpha m_x) = \overline{P}(X_{(\lceil \alpha m_x \rceil)} \leq c) = \sum_{i=1}^{i_c} P(X_{(\lceil \alpha m_x \rceil)} \in I_i^X) \tag{4.4}$$

where the precise probabilities on the right hand sides of Equations (4.3) and (4.4) can be obtained from Equation (1.17). The NPI lower and upper probabilities for the event

$C^{Yt}_{c^t} \geq \beta m_y$ are derived similarly,

$$\underline{P}(C^{Yt}_{c^t} \geq \beta m_y) = \underline{P}(Y_{(m_y - \lceil \beta m_y \rceil + 1)} > c) = \sum_{j=j_c+1}^{n_y+1} P(Y_{(m_y - \lceil \beta m_y \rceil + 1)} \in I^Y_j) \tag{4.5}$$

$$\overline{P}(C^{Yt}_{c^t} \geq \beta m_y) = \overline{P}(Y_{(m_y - \lceil \beta m_y \rceil + 1)} > c) = \sum_{j=j_c}^{n_y+1} P(Y_{(m_y - \lceil \beta m_y \rceil + 1)} \in I^Y_j) \tag{4.6}$$

To define the NPI lower and upper probabilities for such predictive comparison, we consider the following. If the corresponding lower probability in Equation (4.1) for Test 1 is greater than the corresponding upper probability Equation (4.2) for Test 2, that is

$$\underline{P}(C^{X^1}_{c^1} \geq \alpha m_x) \times \underline{P}(C^{Y^1}_{c^1} \geq \beta m_y) > \overline{P}(C^{X^2}_{c^2} \geq \alpha m_x) \times \overline{P}(C^{Y^2}_{c^2} \geq \beta m_y) \tag{4.7}$$

we can regard this as a strong indication that Test 1 is better than Test 2. Whereas, if the corresponding lower probability for Test 1 is greater than the corresponding lower probability for Test 2, that is

$$\underline{P}(C^{X^1}_{c^1} \geq \alpha m_x) \times \underline{P}(C^{Y^1}_{c^1} \geq \beta m_y) > \underline{P}(C^{X^2}_{c^2} \geq \alpha m_x) \times \underline{P}(C^{Y^2}_{c^2} \geq \beta m_y) \tag{4.8}$$

then we can regard this as a weak indication that Test 1 is better than Test 2. Of course, the roles of Test 1 and Test 2 can be exchanged to get an indication of Test 2 being better than Test 1. The method will be illustrated in the following examples.

Example 4.1. Consider an artificial data set from two different diagnostic tests applied to the same individuals from two groups, X^1 and Y^1 for Test 1, and X^2 and Y^2 for Test 2, with $n_x = n_y = 10$, consisting of the following ranks $X^1 = \{1, 2, 3, 4, 5, 7, 9, 10, 11, 12\}$ and $Y^1 = \{6, 8, 13, 14, 15, 16, 17, 18, 19, 20\}$ for Test 1, and $X^2 = \{1, 2, 6, 7, 10, 11, 12, 13, 16, 18\}$ and $Y^2 = \{3, 4, 5, 8, 9, 14, 15, 17, 19, 20\}$ for Test 2. Based on these data, Test 1 seems to differentiate between groups X and Y more than Test 2.

The 2-NPI lower and upper probabilities as given in Equations (4.1) and (4.2) for Test 1 and Test 2, for $m = 1, \ldots, 30$ are displayed in Figure 4.1. We have considered two different scenarios of α and β. It is obviously that every test has a different location of the optimal threshold, thus has a different number of correctly classified individuals from groups X and Y. To find the optimal thresholds for the two tests, we use the 2-

NPI-L method, the optimal thresholds are $c^1 \in (12, 13)$ and $c^2 \in (13, 14)$ regardless of the value of m. Then if we apply the optimal threshold c^1 to the empirical data for Test 1, the number of correctly classified individuals from group X is 10 out of 10 and that from group Y is 8 out of 10. If we apply the optimal threshold c^2 to the empirical data for Test 2, then the number of correctly classified individuals from group X is 8 out of 10 and 5 out of 10 from group Y. Therefore, under the scenario $\alpha = \beta = 0.6$, the empirical data for Test 1 exceeds the proportions of correctly classified observations from both groups, whereas the empirical data for Test 2 does not achieve the proportions of correctly classified observations from group Y. Hence, it is likely that the number of correctly classified future individuals from groups X and Y for every test reflect their empirical data proportions. The results are shown in the first plot in Figure 4.1, there is a strong indication that Test 1 is better than Test 2.

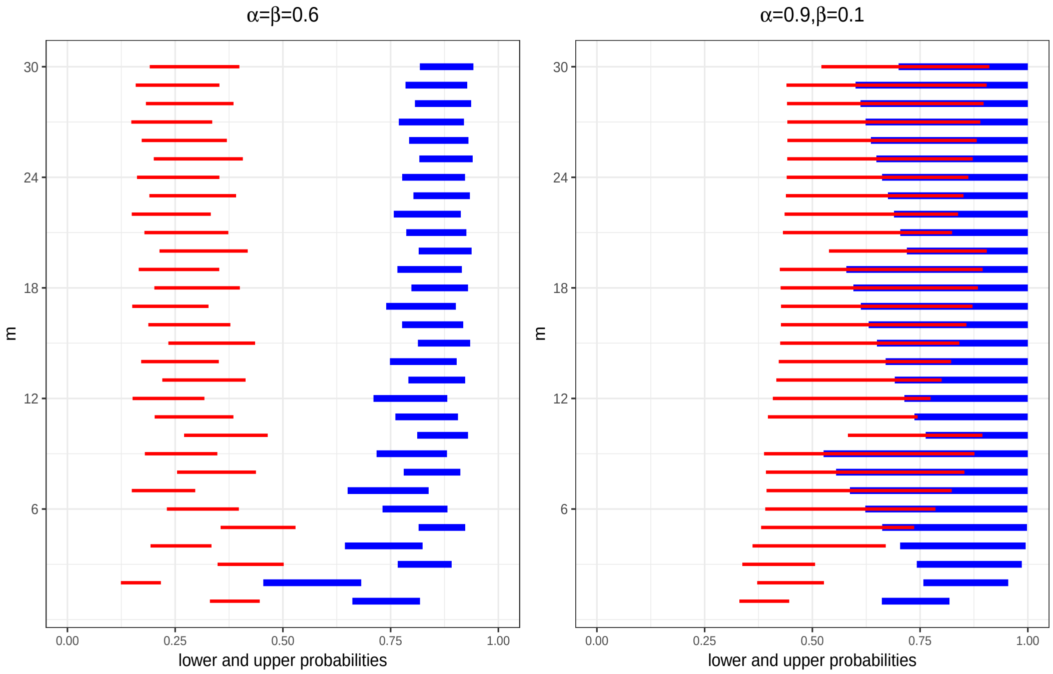

Figure 4.1: Comparison of Test 1 (blue) and Test 2 (red)

We notice that the values of these lower and upper probabilities vary. With a change in the value of m, the required number of correctly classified future individuals in Equations (4.1) and (4.2) of course changes, and then, the values of the corresponding lower and upper probabilities in these equations vary. For example, for $m = 2$, the required number of correctly classified future individuals from both groups is 2 out of 2, which means that both the future individuals must be correctly classified for each group, and that is hard to achieve. Consequently, the values of these lower and upper probabilities are small. For $m = 3$, the required number of correctly classified future individuals from both groups is 2 out of 3, which is easier to achieve than 2 out of 2. Thus, the values of these lower and upper probabilities are higher for $m = 3$ than for $m = 2$. For $m = 4$, the required number of correctly classified future individuals from both groups is 3 out of 4, which is harder to achieve than 2 out of 3. Therefore, the values of these lower and upper probabilities are smaller for $m = 4$ than for $m = 3$.

When the desired criterion strongly emphasizes the number of correctly classified future observations from group X, using the values $\alpha = 0.9, \beta = 0.1$, the optimal threshold for Test 1 stays the same as in the previous scenario, but for Test 2 the optimal threshold is $c^2 \in (18, 19)$ for $m = 4, \ldots, 30$ and $c^2 \in (13, 14)$ for $m = 1, 2, 3$. On applying the optimal threshold $c^2 \in (18, 19)$ to the empirical data for Test 2, the number of correctly classified individuals from group X is 10 out of 10 and from group Y is 2 out of 10. According to the locations of these optimal thresholds from both groups, the required numbers of correctly classified future individuals from both tests are easy to achieve. However, the empirical data for Test 1 exceeds the proportion of correctly classified observations from group Y more than Test 2. Thus, the corresponding lower probabilities in Equation (4.1) for Test 1 are greater than for Test 2, and also the corresponding upper probabilities in Equation (4.2) for Test 1 are greater than for Test 2, so we can say that there is a weak indication that Test 1 is better than Test 2. However, there is not a strong indication, as was the case in the previous scenario ($\alpha = \beta = 0.6$), because the required numbers of correctly classified future individuals from both tests are easier to achieve. However, due to the randomness effect as discussed in Section 2.1, for small values of m there is a strong indication that Test 1 is better than Test 2. The results are shown in the second plot of Figure 4.1.

For both tests, the differences between the upper and lower probabilities, called the imprecision, are observed to increase from $m = 1$ to $m = 10$. This occurs because all the future individuals must be correctly classified from group X, therefore the corresponding lower and upper probabilities for the events $C^{X^1}_{c^1} \geq \alpha m_x$ and $C^{X^2}_{c^2} \geq \alpha m_x$ decrease gradually with increasing the value m. Whereas, the required number of correctly classified future individuals from group Y is just one future individual, therefore the corresponding lower and upper probabilities for the events $C^{Y^1}_{c^1} \geq \beta m_y$ and $C^{Y^2}_{c^2} \geq \beta m_y$ are close to one. Thus, the imprecision entirely occurs by the effect of group X.

Figure 4.1 shows some step-like pattern, in particular for the lower probabilities for both tests in the case $\alpha = 0.9, \beta = 0.1$. This pattern is explained as follows. For example, considering $m = 9$ and $m = 10$ the lower probability in Equation (4.1) for both tests is greater for $m = 10$ than for $m = 9$ because for $m = 9$ the required numbers of correctly classified future individuals from group Y is 1 out of 9 and from group X is 9 out of 9. Whereas for $m = 10$, the required numbers of correctly classified future individuals from group Y is 1 out of 10 and from group X is 9 out 10. Thus, for $m = 10$, the required numbers of correctly classified future individuals from both groups are easier to achieve than the required numbers for $m = 9$. This increase does not occur for the corresponding upper probabilities in Equation (4.2) for both tests since the values of the upper probabilities for group X are equal to 1 for $m = 4, 5, \ldots 30$ for Test 2, and equal to 1 for Test 1 for all the values of m, since the thresholds c^1 and c^2 are greater than all X data observations. After that, for $m = 11, \ldots, 19$, the lower probabilities for both tests start to decrease gradually because the required numbers for such values of m are harder to achieve than the required numbers for $m = 10$.

Generally, for large values of m, the imprecision is higher for scenario $\alpha = 0.9, \beta = 0.1$ than that for the scenario $\alpha = \beta = 0.6$. It is because the required numbers of correctly classified future individuals are easier to achieve under this scenario, so the corresponding upper probabilities in Equations (4.1) and (4.2) are high for both groups and tests, especially for group X where the upper probabilities are equal to one as the thresholds c^1 and c^2 are greater than all X data observations.

Example 4.2. Example 4.1 consisted of two tests that have different performance in the sense that Test 1 seems to separate between groups X and Y more than Test 2.

In this example, we consider two tests that have a similar level of overlap between groups X and Y. Let $n_x = n_y = 10$ observations have the following ranks; $X^1 = \{1, 2, 5, 6, 8, 10, 11, 12, 13, 17\}$ and $Y^1 = \{3, 4, 7, 9, 14, 15, 16, 18, 19, 20\}$ for Test 1, and $X^2 = \{1, 2, 3, 7, 8, 9, 10, 15, 16, 18\}$, $Y^2 = \{4, 5, 6, 11, 12, 13, 14, 17, 19, 20\}$ for Test 2.

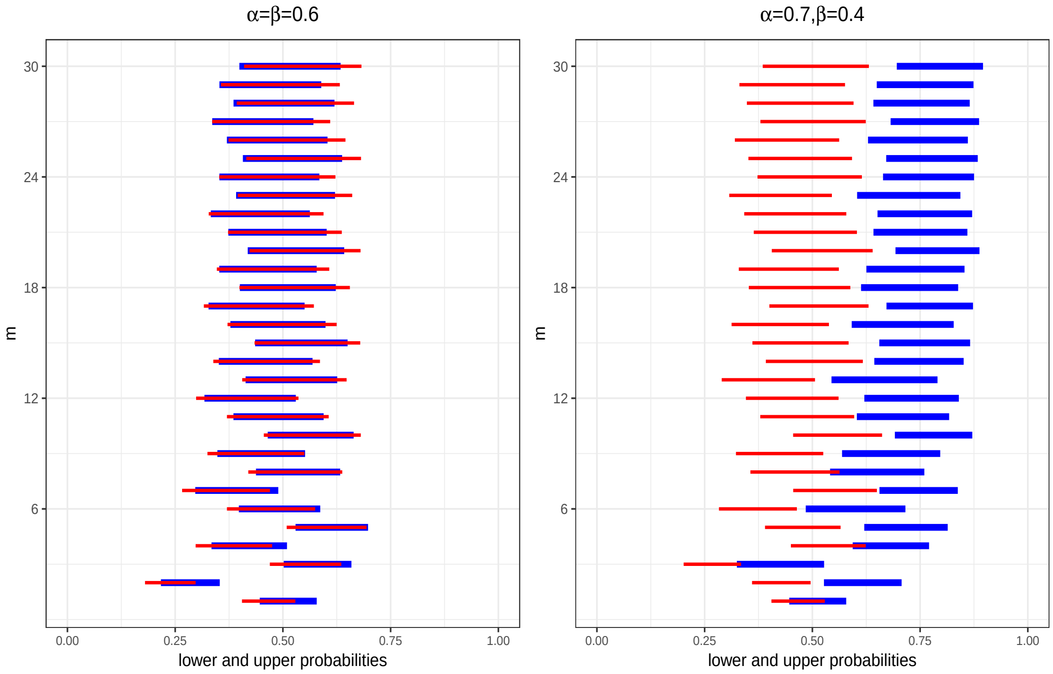

Figure 4.2: Comparison of Test 1 (blue) and Test 2 (red)

The 2-NPI lower and upper probabilities as given by Equations (4.1) and (4.2) for Test 1 and Test 2, for $m = 1, \ldots, 30$ are displayed in Figure 4.2. We have considered two different scenarios of α and β. For $\alpha = \beta = 0.6$, the optimal thresholds are $c^1 \in (13, 14)$ and $c^2 \in (10, 11)$ for all the values of m. If we apply the optimal threshold c^1 to the empirical data for Test 1, the numbers of correctly classified individuals from group X is 9 out of 10 and from group Y it is 6 out of 10. If we apply the optimal threshold c^2 to the empirical data for Test 2, the numbers of correctly classified individuals from each of the groups, X and Y, is 7 out of 10. Therefore, under the scenario $\alpha = \beta = 0.6$, the empirical data for both tests are quite similar regarding the number of correctly classified individuals from both groups and the empirical data for both tests achieve the required numbers of correctly classified individuals, but it is not clear which test is better. The results are shown in the first plot in Figure 4.2, there is a weak indication that Test 1 is better than Test 2 for small values of m, whereas for larger values of m, these lower and upper probabilities for Test 1 are nested within those for Test 2.

The optimal thresholds c^1 and c^2 for $\alpha = 0.7, \beta = 0.4$ are the same as those for the scenario $\alpha = \beta = 0.6$. As the results are shown in the second plot in Figure 4.2, there is a strong indication that Test 1 is better than Test 2 for almost all the values of m, since the empirical data for Test 1 exceeds the proportion of correctly classified individuals from both groups. However, for some small values of m, because of the randomness effect, there is only a weak indication that Test 1 is better than Test 2.

Example 4.3. In this example, we use the data set from a study to develop screening methods to detect carriers of a rare genetic disorder. The data were first discussed by Cox et al. [27], and are available from Carnegie Mellon University Statlib Datasets Archive at http://lib.stat.cmu.edu/datasets/. Four measurements M1, M2, M3 and M4 were made on blood samples. For some patients, there are several samples of which the average is considered, and five missing values are excluded from the analysis. The remaining sample, which is used in this example, consists of 120 observations, 38 for carriers of the rare genetic disorder and 82 for non-carriers. Coolen-Maturi [22] used this data set to combine two or more of these diagnostic tests in order to improve the overall accuracy using the area under the ROC curve, based on the NPI setting for one future individual. In this example, we use this data set for pairwise comparisons of these four diagnostic tests, using the NPI method presented in Section 4.2. To compare two of these four diagnostic tests, for test Mt, for $t = 1, 2, 3, 4$, we define $T^{Mt} = C^{X^t}_{c^t} \geq \alpha m_x \times C^{Y^t}_{c^t} \geq \beta m_y$. Comparison any of two of these four tests is derived by Equations (4.1) and (4.2).

To compare two of these tests, the 2-NPI lower and upper probabilities as given in Equations (4.1) and (4.2), for the four diagnostic tests, for $m = 1, \ldots, 30$, are displayed in Figures 4.3 and 4.4, for the scenarios $\alpha = \beta = 0.5$ and $\alpha = 0.5, \beta = 0.7$, respectively. The heading in each plot states the two diagnostic tests, the first named test is presented in blue and the second named test in red.

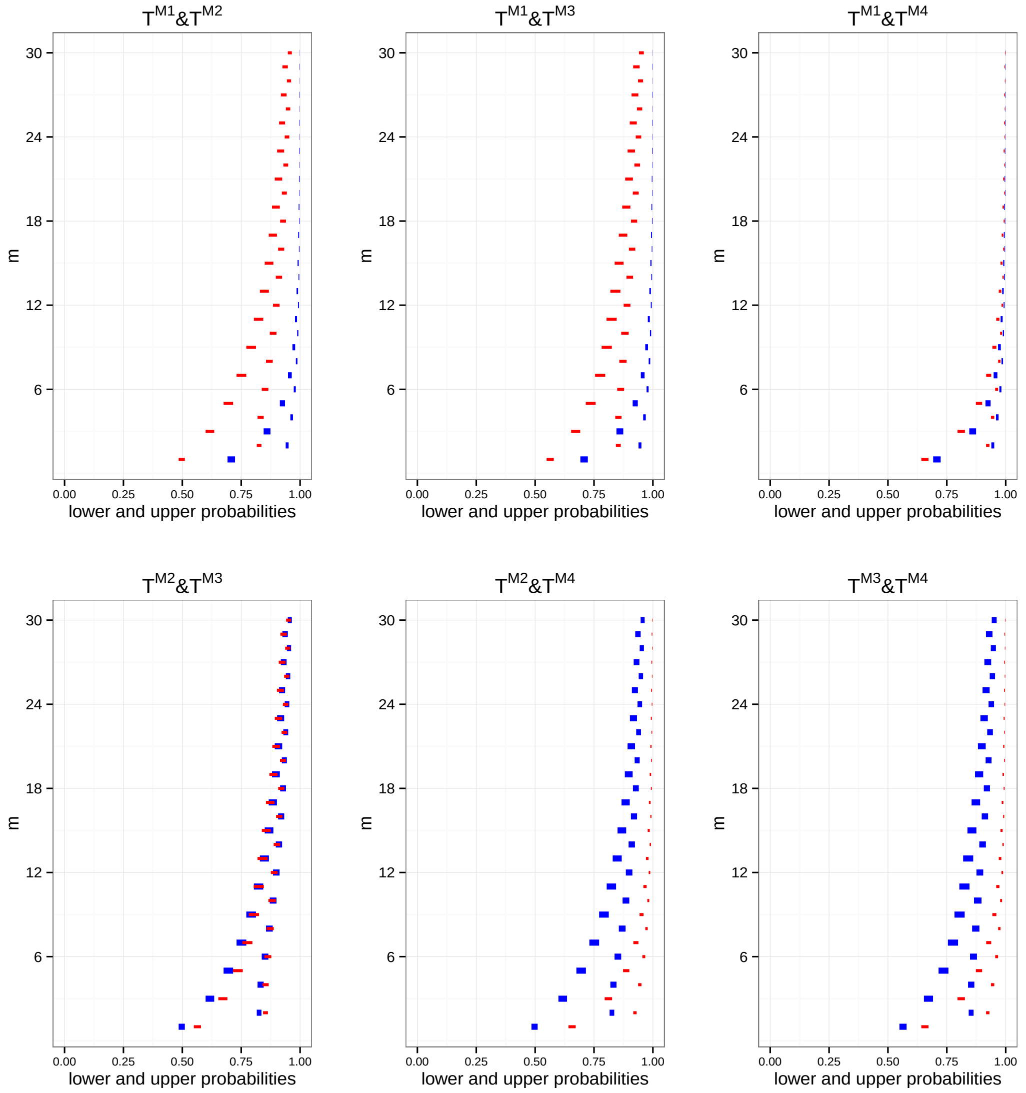

Figure 4.3: Pairwise comparisons of T^{M1}, T^{M2}, T^{M3} and T^{M4}, with $\alpha = \beta = 0.5$

First, we find the optimal threshold for each test, then we apply the optimal threshold in their empirical data to find the numbers of correctly classified individuals from groups X and Y. After doing so, we find the following. For $\alpha = \beta = 0.5$, the numbers of correctly classified individuals from groups X and Y for T^{M1} are 70 out of 82 and 32 out of 38, respectively, regardless of the value of m. The numbers of correctly classified individuals from groups X and Y for T^{M2} are 56 out of 82 and 28 out of 38, respectively, for $m = 1, 2$. and 58 out of 82 and 27 out of 38, respectively, for $m = 3, \ldots, 30$. The numbers of correctly classified individuals from groups X and Y for T^{M3} are 74 out of 82 and 24 out of 38, respectively, for $m = 1$, and 70 out of 82 and 25 out of 38, respectively, for $m = 2, \ldots, 11$, and 57 out of 82 and 27 out of 38, respectively, for $m = 12, \ldots, 30$. The numbers of correctly classified individuals from groups X and Y for T^{M4} are 67 out of 82 and 31 out of 38, respectively, regardless of the value m.

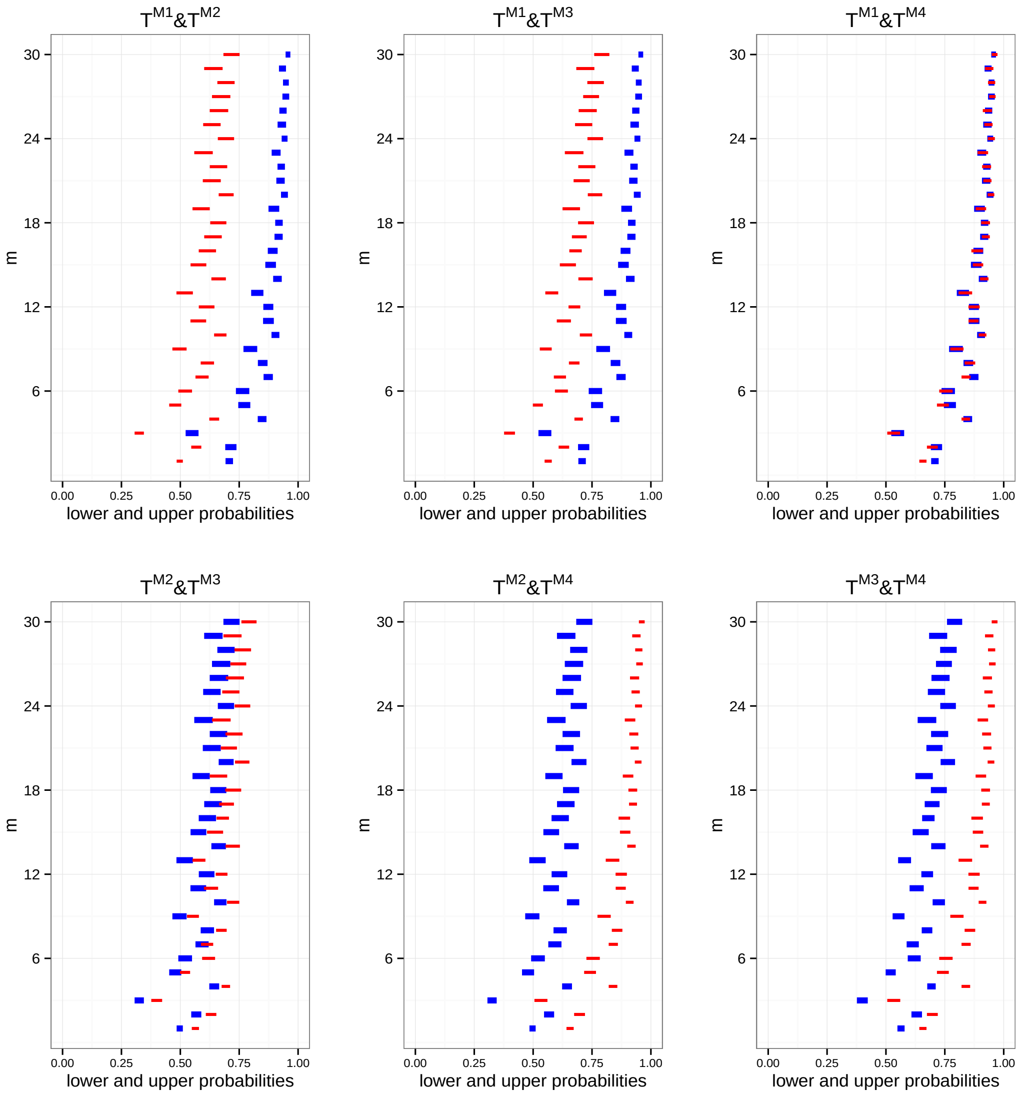

Figure 4.4: Pairwise comparisons of T^{M1}, T^{M2}, T^{M3} and T^{M4}, with $\alpha = 0.5, \beta = 0.7$

Based on these numbers, the numbers of correctly classified individuals from both groups for T^{M1} are the greatest, followed by the corresponding number for T^{M4}. While, the numbers of correctly classified individuals from both groups for T^{M3} are greater than the corresponding number for T^{M2}, for $m = 1, \ldots, 11$, whereas for $m = 12, \ldots, 30$ the number of correctly classified individuals from group X for T^{M2} is greater than the corresponding number for T^{M3} and the corresponding number from group Y are equal for the two tests.

For scenario where $\alpha = \beta = 0.5$ in Figure 4.3, in the first row we can say there is a strong indication that T^{M1} is better than T^{M2} and T^{M3} for all values of m. Whereas, there is only a weak indication that T^{M1} is better than T^{M4} for large values of m, since both have high values of the lower and upper probabilities, but for small values of m there is a strong indication that T^{M1} is better than T^{M4}, which is because of the randomness effect. In the second row, there is a strong indication that T^{M3} is better than T^{M2} for small values of m whereas for large values of m the corresponding lower and upper probabilities in Equations (4.1) and (4.2) for T^{M2} are nested within those for T^{M3}. T^{M4} is better than T^{M2} and T^{M3} with a strong indication for all values of m.

For $\alpha = 0.5, \beta = 0.7$, the numbers of correctly classified individuals from groups X and Y for T^{M1} are 70 out of 82 and 32 out of 38, respectively, for $m = 1, 2, 4, 5, 7, 9, 11$, and 56 out of 82 and 34 out of 38, respectively, for $m = 6, 8, 12, 13, 16, 18, 22, 23, 24, 26, 28, 29, 30$, and 60 out of 82 and 33 out of 38, respectively, for $m = 3, 10, 14, 15, 17, 19, 20, 21, 25, 27$. The numbers of correctly classified individuals from groups X and Y for T^{M2} are 56 out of 82 and 28 out of 38, respectively, for $m = 1, 5, 7$ and 36 out of 82 and 35 out of 38, respectively, for $m = 2, 3, 4, 6, 8, \ldots, 30$. The numbers of correctly classified individuals from groups X and Y for T^{M3} are 74 out of 82 and 24 out of 38, respectively, for $m = 1$, and 42 out of 82 and 35 out of 38, respectively, for $m = 2, 3, 4, 5, 6, 8, 9, 12, 13, 16$, and 52 out of 82 and 30 out of 38, respectively, for $m = 7, 10, 11, 15, 17, \ldots, 30$. The numbers of correctly classified individuals from groups X and Y for T^{M4} are 67 out of 82 and 31 out of 38, respectively, for $m = 1$, and 55 out of 82 and 34 out of 38, respectively, for $m = 2, 6$, and 61 out of 82 and 33 out of 38, respectively, for $m = 3, 4, 5, 7, \ldots, 30$.

Thus, under the scenario $\alpha = 0.5, \beta = 0.7$, the empirical data for T^{M1} exceed the proportion of correctly classified individuals from both groups more than the empirical data for T^{M2} and T^{M3} do. While the empirical data for both T^{M1} and T^{M4} are quite similar regarding to the number of correctly classified individuals from groups. The empirical data for T^{M3} exceed the proportion of correctly classified individuals from both groups more than the empirical data for T^{M2} does, for most values of m.

In Figure 4.4, for $\alpha = 0.5, \beta = 0.7$, similar results hold as for scenario $\alpha = \beta = 0.5$. However, the values of the corresponding lower and upper probabilities for these tests are lower, because the required numbers of correctly classified future individuals from group Y is harder to achieve than with $\beta = 0.5$. However, the third and fourth plots show different results than for scenario $\alpha = \beta = 0.5$. The third plot shows that T^{M1} is better than T^{M4} with only a weak indication for small values of m, but for large values of m, the corresponding lower and upper probabilities in Equations (4.1) and (4.2) for T^{M1} are nested within those for T^{M4}. The fourth plot shows that T^{M3} is better than T^{M2} with a strong indication for some values of m, while with a weak indication for others.

4.3 Comparison of two diagnostic tests using NPI for Bernoulli quantities

The method presented in Section 4.2 can be also set up using NPI for Bernoulli quantities as presented in Section 1.3.3. In this Section we compare the two tests by considering the total number of correct diagnoses for m_x future healthy individuals and m_y future patients for one test with those for the other test, using NPI for Bernoulli quantities. The same notations will be used as introduced in Section 4.2, and again the 2-NPI-L method presented in Section 2.2 is used to select the optimal threshold c^t. The number of successes in n_x and n_y data observations are denoted by s_x^t and s_y^t, respectively, for test t. Let $C_{m_x}^{X^t}$ denote the random number of successful diagnoses for the healthy future individuals out of m_x for test t, and $C_{m_y}^{Y^t}$ denote the random number of successful diagnoses for the diseased future individuals out of m_y for test t. The total number of correct diagnoses for m_x future healthy individuals and m_y future patients in Test 1 is $C_{m_x}^{X^1} + C_{m_y}^{Y^1}$, and the total number of correct diagnoses for m_x and m_y in Test 2 is $C_{m_x}^{X^2} + C_{m_y}^{Y^2}$. We now consider the event $C_{m_x}^{X^1} + C_{m_y}^{Y^1} > C_{m_x}^{X^2} + C_{m_y}^{Y^2}$. The NPI upper probability for this event, for $C_{m_x}^{X^1}, C_{m_x}^{X^2} \in \{0, ..., m_x\}$, and $C_{m_y}^{Y^1}, C_{m_y}^{Y^2} \in \{0, ..., m_y\}$, based on data $(n_x, s_x^1), (n_y, s_y^1)$ and $(n_x, s_x^2), (n_y, s_y^2)$, is

$$\begin{aligned}
&\overline{P}(C^{X^1}_{m_x} + C^{Y^1}_{m_y} > C^{X^2}_{m_x} + C^{Y^2}_{m_y}) \\
&= \sum_{k=0}^{m_x+m_y} \overline{P}(C^{X^2}_{m_x} + C^{Y^2}_{m_y} < k) \times [\overline{P}(C^{X^1}_{m_x} + C^{Y^1}_{m_y} \geq k) - \overline{P}(C^{X^1}_{m_x} + C^{Y^1}_{m_y} \geq k+1)]
\end{aligned} \tag{4.9}$$

This equation follows from the fact that $\overline{P}(C^{X^2}_{m_x}+C^{Y^2}_{m_y} < k)$ is increasing in k, thus, we put the maximum possible probability mass for $C^{X^1}_{m_x}+C^{Y^1}_{m_y}$ at the event $C^{X^1}_{m_x}+C^{Y^1}_{m_y} \geq m_x+m_y$, followed by assigning the maximum possible remaining probability mass for $C^{X^1}_{m_x}+C^{Y^1}_{m_y}$ at the event $C^{X^1}_{m_x}+C^{Y^1}_{m_y} \geq m_x+m_y-1$, etc [18]. Then, we can write Equation (4.9) as follows,

$$\begin{aligned}
&\sum_{k=0}^{m_x+m_y} \overline{P}(C^{X^2}_{m_x} + C^{Y^2}_{m_y} < k) \times [\overline{P}(C^{X^1}_{m_x} + C^{Y^1}_{m_y} \geq k) - \overline{P}(C^{X^1}_{m_x} + C^{Y^1}_{m_y} \geq k+1)] \\
&= \sum_{k=0}^{m_x+m_y} \left[\sum_{v=0}^{m_y} \overline{P}(C^{X^2}_{m_x} < k-v) \times [\overline{P}(C^{Y^2}_{m_y} \leq v) - \overline{P}(C^{Y^2}_{m_y} \leq v-1)]\right] \\
&\times \left[\sum_{v=0}^{m_y} \overline{P}(C^{X^1}_{m_x} \geq k-v) \times [\overline{P}(C^{Y^1}_{m_y} \geq v) - \overline{P}(C^{Y^1}_{m_y} \geq v+1)]\right. \\
&\left.- \sum_{v=0}^{m_y} \overline{P}(C^{X^1}_{m_x} \geq k+1-v) \times [\overline{P}(C^{Y^1}_{m_y} \geq v) - \overline{P}(C^{Y^1}_{m_y} \geq v+1)]\right] \\
&= \sum_{k=0}^{m_x+m_y} \left[\sum_{v=0}^{m_y} \overline{P}(C^{X^2}_{m_x} < k-v) \times [\overline{P}(C^{Y^2}_{m_y} \leq v) - \overline{P}(C^{Y^2}_{m_y} \leq v-1)]\right] \\
&\times \left[\sum_{v=0}^{m_y} [\overline{P}(C^{X^1}_{m_x} \geq k-v) - \overline{P}(C^{X^1}_{m_x} \geq k+1-v)] \times [\overline{P}(C^{Y^1}_{m_y} \geq v) - \overline{P}(C^{Y^1}_{m_y} \geq v+1)]\right]
\end{aligned} \tag{4.10}$$

In equation (4.10), we are optimistic for Test 1 by putting the maximum possible probability mass for this test at the larger value of $C^{X^1}_{m_x}$ and $C^{Y^1}_{m_y}$, and pessimistic for Test 2 by putting the maximum possible probability mass for this test at the smaller value of $C^{X^2}_{m_x}$ and $C^{Y^2}_{m_y}$. We may also be interested in the event $C^{X^1}_{m_x} + C^{Y^1}_{m_y} \geq C^{X^2}_{m_x} + C^{Y^2}_{m_y}$, for which the NPI upper probability follow similarly and is equal to

$$\begin{aligned}
&\overline{P}(C^{X^1}_{m_x} + C^{Y^1}_{m_y} \geq C^{X^2}_{m_x} + C^{Y^2}_{m_y}) \\
&= \sum_{k=0}^{m_x+m_y} \overline{P}(C^{X^2}_{m_x} + C^{Y^2}_{m_y} \leq k) \times [\overline{P}(C^{X^1}_{m_x} + C^{Y^1}_{m_y} \geq k) - \overline{P}(C^{X^1}_{m_x} + C^{Y^1}_{m_y} \geq k+1)]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=0}^{m_x+m_y} \Big[\sum_{v=0}^{m_y} \overline{P}(C^{X^2}_{m_x} \leq k-v) \times [\overline{P}(C^{Y^2}_{m_y} \leq v) - \overline{P}(C^{Y^2}_{m_y} \leq v-1)]\Big] \\
&\times \Big[\sum_{v=0}^{m_y}[\overline{P}(C^{X^1}_{m_x} \geq k-v) - \overline{P}(C^{X^1}_{m_x} \geq k+1-v)] \times [\overline{P}(C^{Y^1}_{m_y} \geq v) - \overline{P}(C^{Y^1}_{m_y} \geq v+1)]\Big]
\end{aligned}
\tag{4.11}
$$

The corresponding lower probability can again be derived via the conjugacy property $\underline{P}(A) = 1 - \overline{P}(A^c)$. This method will be illustrated in examples in Section 4.5, but before that we show how to include weights in the next section.

4.4 Comparison of tests using weighted numbers of successful diagnoses

In Section 4.3, we present a method for comparison of two diagnostic tests applied to the same individuals from two groups, healthy and diseased individuals, where both groups are treated equally in the event of interest. When unequal weights are requested to reflect the relative importance of the two groups, weights can be added to the method presented in Section 4.3. Let w_x, $w_y \in \mathbb{N}^+$ be the weights for group X and Y, respectively. We are interested in the event $w_x C^{X^1}_{m_x} + w_y C^{Y^1}_{m_y} > w_x C^{X^2}_{m_x} + w_y C^{Y^2}_{m_y}$, the NPI upper probability for this event is

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

$$= \sum_{k=0}^{w_x m_x + w_y m_y} \Big[\sum_{v=0}^{m_y} \overline{P}(C_{m_x}^{X^2} < \frac{k-(w_y v)}{w_x}) \times [\overline{P}(C_{m_y}^{Y^2} \leq v) - \overline{P}(C_{m_y}^{Y^2} \leq v-1)] \Big]$$
$$\times \Big[\sum_{v=0}^{m_y} [\overline{P}(C_{m_x}^{X^1} \geq \frac{k-(w_y v)}{w_x}) - \overline{P}(C_{m_x}^{X^1} \geq \frac{k+1-(w_y v)}{w_x})] \times [\overline{P}(C_{m_y}^{Y^1} \geq v) - \overline{P}(C_{m_y}^{Y^1} \geq v+1)] \Big] \tag{4.12}$$

The corresponding lower probability can again be derived via the conjugacy property $\underline{P}(A) = 1 - \overline{P}(A^c)$.

We choose the weights w_x and w_y as positive integers because this simplifies notation in the derivation of Equation 4.12, as this ensures that $w_x m_x + w_y m_y$ is integer. Of course, the overall inference for the event $w_x C_{m_x}^{X^1} + w_y C_{m_y}^{Y^1} > w_x C_{m_x}^{X^2} + w_y C_{m_y}^{Y^2}$ is not affected by multiplication of both w_x and w_y by the same positive constant, hence one could scale them, e.g. to be in (0.1] or even to sum up to 1.

In the case of $m_x = m_y = 1$ in Equation (4.12), we notice that there are three possible events depending on $w_x < w_y$, $w_x > w_y$ and $w_x = w_y$. Therefore, for $m_x = m_y = 1$, Equation (4.12) can be expressed in a simple expression as follows.

For $w_x < w_y$

$$\begin{aligned}
&\overline{P}(w_x C_1^{X^1} + w_y C_1^{Y^1} > w_x C_1^{X^2} + w_y C_1^{Y^2}) \\
&= \{\overline{P}(C_1^{Y^2} = 0)[\overline{P}(C_1^{X^2} = 0)\overline{P}(C_1^{X^1} = 1)\underline{P}(C_1^{Y^1} = 0) \\
&+ \overline{P}(C_1^{Y^1} = 1)(1 - \overline{P}(C_1^{X^1} = 1)\overline{P}(C_1^{X^2} = 0))]\} \\
&+ \overline{P}(C_1^{X^1} = 1)\overline{P}(C_1^{Y^1} = 1)\overline{P}(C_1^{X^2} = 0)
\end{aligned} \tag{4.13}$$

For $w_x > w_y$

$$\begin{aligned}
&\overline{P}(w_x C_1^{X^1} + w_y C_1^{Y^1} > w_x C_1^{X^2} + w_y C_1^{Y^2}) \\
&= \{\overline{P}(C_1^{X^2} = 0)[\overline{P}(C_1^{Y^2} = 0)\overline{P}(C_1^{Y^1} = 1)\underline{P}(C_1^{X^1} = 0) \\
&+ \overline{P}(C_1^{X^1} = 1)(1 - \overline{P}(C_1^{Y^1} = 1)\overline{P}(C_1^{Y^2} = 0))]\} \\
&+ \overline{P}(C_1^{Y^1} = 1)\overline{P}(C_1^{X^1} = 1)\overline{P}(C_1^{Y^2} = 0)
\end{aligned} \tag{4.14}$$

For $w_y = w_x$

$$
\begin{aligned}
&\overline{P}(w_xC_1^{X^1} + w_yC_1^{Y^1} > w_xC_1^{X^2} + w_yC_1^{Y^2})\\
&= \overline{P}(C_1^{X^2} = 0)\overline{P}(C_1^{Y^2} = 0)[\overline{P}(C_1^{X^1} = 1) + \overline{P}(C_1^{Y^1} = 1)]\\
&+ \overline{P}(C_1^{X^1} = 1)\overline{P}(C_1^{Y^1} = 1)[\overline{P}(C_1^{Y^2} = 0) + \overline{P}(C_1^{X^2} = 0)]\\
&- 3\overline{P}(C_1^{X^2} = 0)\overline{P}(C_1^{Y^2} = 0)\overline{P}(C_1^{X^1} = 1)\overline{P}(C_1^{Y^1} = 1) \qquad (4.15)
\end{aligned}
$$

The weights, w_x and w_y, are introduced to reflect the relative importance of one group over the other. Varying the values of w_x and w_y will depend on which group is more important to be successfully diagnosed. The proof for Equations 4.13-4.15 is given in Appendix A. This method will be illustrated in Example 4.5 in the next section.

4.5 Examples

In this section, three examples are given to illustrate the NPI comparison of two diagnostic tests, as presented in Section 4.3. The data for Examples 4.4 and 4.5 are small artificial data sets. Example 4.6 uses data from the literature. Also, Example 4.5 illustrates the comparison of tests using weighted numbers of successful diagnoses presented Section 4.4.

Example 4.4. Suppose that we are interested in comparing between two diagnostic tests for a particular disease. Each test is applied to the same individuals from two groups X and Y, with $n_x = n_y = 10$. Assume that the threshold values c^1 and c^2 are set before the comparison, and they provide the following numbers of successful diagnoses from groups X and Y, $s_x^1 = s_y^1 = 8$ for Test 1, and $s_x^2 = s_y^2 = 6$ for Test 2. In this example, T^1 refers to $C_{m_x}^{X^1} + C_{m_y}^{Y^1}$ and T^2 refers to $C_{m_x}^{X^2} + C_{m_y}^{Y^2}$.

Tables 4.1 and 4.2 present the NPI upper probabilities given by Equation (4.10) and the corresponding NPI lower probabilities, for comparison of T^1 and T^2 for different values of m_x and m_y. In Table 4.1, equal number of future individuals for diseased and healthy groups are considered, so $m_x = m_y = m$. Table 4.2 some cases with $m_x \neq m_y$ are presented. In this example, the total number of correct diagnoses for T^1 is greater than that for T^2. Hence, it is likely that the total number of correct diagnoses for m_x future healthy individuals and m_y future patients for T^1 is greater than that for T^2, if m is not too small. Indeed, Table 4.1 shows that the values of the lower and upper probabilities for $T^1 > (\geq)\ T^2$ are greater than for $T^2 > (\geq)\ T^1$. Moreover, these values show that the differences between the two tests become clearer for larger values of m, as equal outcomes become less likely. This is shown by the large differences of the lower and upper probabilities for the events $T^1 > T^2$ and $T^1 \geq T^2$ for small m.

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

Table 4.1: NPI lower and upper probabilities for comparison of two tests with $m_x = m_y = m$

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

Table 4.2: NPI lower and upper probabilities for comparison of two tests with $m_x \neq m_y$

Table 4.2 shows that when the number of future individuals from one group is increased more than the other, T^1 stays better than T^2 as in the results for $m_x = m_y$. However the values of the lower and upper probabilities for $T^1 > T^2$ are higher than for $m_x = m_y$. For example, for $m_x = m_y = 15$ the lower and upper probabilities for the event $T^1 > T^2$ are equal to 0.6644 and 0.9321 respectively, while for $m_x = 30$ and $m_y = 15$ these values become 0.6823 and 0.9430 respectively. It is also noticed that in this table, when

m_x and m_y are interchanged, these lower and upper probabilities are the same since the total number of correct diagnoses from X and Y are equal for both the tests ($s_x^1 = s_y^1 = 8$, $s_x^2 = s_y^2 = 6$).

Example 4.5. Example 4.4 consisted of two tests that have different performance in terms of the total number of successfully diagnosed individuals from groups X and Y. In this example, we consider two tests that have more similar total numbers of correct diagnoses from X and Y, with $n_x = n_y = 10$, the data are $s_x^1 = 7$ and $s_y^1 = 9$ for Test 1, and $s_x^2 = 9$ and $s_y^2 = 6$ for Test 2.

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

Table 4.3: NPI lower and upper probabilities for comparison of two tests with $m_x = m_y = m$

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

Table 4.4: NPI lower and upper probabilities for comparison of two tests for $m_x \neq m_y$

Tables 4.3 and 4.4 present the NPI upper probabilities for comparison of T^1 and T^2, for different values of m_x and m_y. Table 4.3 presents results for $m_x = m_y = m$, while Table 4.4 presents some cases with $m_x \neq m_y$. The values of the lower and upper probabilities for the events $T^1 > (\geq)\ T^2$ are a bit higher than for the events $T^2 > (\geq)\ T^1$. For all the values of m, these values for the event $T^1 > T^2$ are slightly increasing with m as equal outcomes become less likely. This is shown by the fact that the lower and upper probabilities for $T^1 > T^2$ become close to those for $T^1 \geq T^2$ for larger m.

There is a tendency for the imprecision to increase with m, which is intuitively attractive when considering more future observations. However, this does not always happen, for example if an event is quite unlikely to happen then its upper probability will be close to zero, hence the imprecision will be quite small also for larger values of m. The decision which test is the best can be supported by the use of relevant values of these lower and upper probabilities. Therefore, one can prefer the better test for the values of m that have the lower probability that exceed 0.5, so that can be a strong indication of this test being better than the other. So, in this example, one can conclude that Test 1 is at least as good as Test 2 for the next 5 patients and 5 non-patients as the lower probability for $T^1 \geq T^2$ is equal to 0.5109 for $m = 5$. However for m_x and m_y are equal to 6 or more, we could conclude that neither test is really better than the other.

Table 4.4 shows different behavior than Table 4.3, since the numbers of future individuals from diseased and healthy groups differ. From these tables, the decision of which test is the best clearly depends on the values of the number of successful diagnoses from diseased and healthy groups and also the number of future individuals from both the groups. For example, for $m_x = 15$ and $m_y = 30$, T^1 is better than T^2, whereas for $m_x = 30$ and $m_y = 15$, T^2 is better than T^1.

m	$[\underline{P}, \overline{P}](T^1 > T^2)$	$[\underline{P}, \overline{P}](T^1 \geq T^2)$	$[\underline{P}, \overline{P}](T^2 > T^1)$	$[\underline{P}, \overline{P}](T^2 \geq T^1)$
		$w_x = 4, w_y = 2$		
1	[0.2398, 0.3986]	]0.5987, 0.7449]	[0.2551, 0.4013]	[0.6014, 0.7602]
5	[0.2418, 0.5521]	[0.3490, 0.6666]	[0.3334, 0.6510]	[0.4479, 0.7582]
15	[0.2063, 0.6207]	[0.2487, 0.6707]	[0.3293, 0.7513]	[0.3793, 0.7937]
50	[0.1785, 0.6629]	[0.1919, 0.6803]	[0.3197, 0.8081]	[0.3371, 0.8215]
100	[0.1702, 0.6747]	[0.1770, 0.6837]	[0.3163, 0.8230]	[0.3253, 0.8298]
		$w_x = 2, w_y = 4$		
1	[0.3315, 0.4843]	[0.6903, 0.8305]	[0.1695, 0.3097]	[0.5157, 0.6685]
5	[0.4954, 0.7880]	[0.6097, 0.8650]	[0.1350, 0.3903]	[0.2120, 0.5046]
15	[0.5464, 0.8896]	[0.5974, 0.9128]	[0.0872, 0.4026]	[0.1104, 0.4536]
50	[0.5764, 0.9346]	[0.5944, 0.9404]	[0.0596, 0.4056]	[0.0654, 0.4236]
100	[0.5847, 0.9446]	[0.5941, 0.9473]	[0.0527, 0.4059]	[0.0554, 0.4153]

Table 4.5: NPI lower and upper probabilities for comparison of two tests for $m_x = m_y = m$, using different weights

Table 4.5 presents the NPI upper probabilities given by Equation (4.12) and the corresponding NPI lower probabilities, for comparison of the two tests using different weights. For $w_x = 4, w_y = 2$, T^2 is better than T^1 for all the different values of m since the number of successful diagnoses in the data from group X for T^2 is greater than the corresponding number for T^1. For $w_x = 2, w_y = 4$, T^1 is better than T^2 for all the different values of m since the number of successful diagnoses in the data from group Y for T^1 is greater than the corresponding number for T^2.

Example 4.6. In this example, the data set presented in Example 4.3 involving four tests: M1, M2, M3 and M4, is used. We define the number of successful diagnoses for all four tests by identifying the optimal thresholds c^t using the 2-NPI-L method, for different values of m and $\alpha = \beta = 0.5$, and then we count the number of successfully diagnosed individuals in the data for both the groups. To compare the two of these four diagnostic tests, for test Mt, for $t = 1, 2, 3, 4$, we define $T^{Mt} = C^{Xt}_{m_x} + C^{Yt}_{m_y}$, and s^{Mt}_x, s^{Mt}_y are the numbers of successful diagnoses from healthy and diseased groups for Mt. Comparison any of two of these four tests is derived by the upper probability in Equation (4.10) and the corresponding lower probability.

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

Table 4.6: The number of successful diagnoses in the data from groups X and Y for T^{M1}, T^{M2}, T^{M3} and T^{M4}

Table 4.6 shows the number of successful diagnoses in the data from healthy and diseased groups for every test, for different values of m. Based on these numbers, the total number of successfully diagnosed individuals in the data for T^{M1} is the greatest one, followed by the corresponding number for T^{M4}. While the total number of successfully diagnosed individuals in the data for T^{M3} is greater than the corresponding number for

T^{M2}, for $m = 1, 5, 10$, whereas for $m = 30, 100$ the number of successfully diagnosed from group X for T^{M2} is greater than the corresponding number for T^{M3} and the corresponding number from group Y are equal for the two tests.

m	$[\underline{P},\overline{P}](T^{M1} > T^{M2})$	$[\underline{P},\overline{P}](T^{M1} \geq T^{M2})$	$[\underline{P},\overline{P}](T^{M2} > T^{M1})$	$[\underline{P},\overline{P}](T^{M2} \geq T^{M1})$
1	[0.3611, 0.3955]	[0.8312, 0.8568]	[0.1432, 0.1688]	[0.6035, 0.6389]
5	[0.6392, 0.7119]	[0.8115, 0.8621]	[0.1379, 0.1885]	[0.2881, 0.3608]
10	[0.7448, 0.8245]	[0.8445, 0.9014]	[0.0986, 0.1555]	[0.1755, 0.2552]
30	[0.8783, 0.9428]	[0.9104, 0.9605]	[0.0395, 0.0896]	[0.0572, 0.1217]
100	[0.9500, 0.9860]	[0.9569, 0.9883]	[0.0117, 0.0431]	[0.0140, 0.0500]
	$[\underline{P},\overline{P}](T^{M1} > T^{M3})$	$[\underline{P},\overline{P}](T^{M1} \geq T^{M3})$	$[\underline{P},\overline{P}](T^{M3} > T^{M1})$	$[\underline{P},\overline{P}](T^{M3} \geq T^{M1})$
1	[0.3008, 0.3373]	[0.8107, 0.8394]	[0.1606, 0.1893]	[0.6627, 0.6992]
5	[0.5473, 0.6290]	[0.7490, 0.8120]	[0.3710, 0.4527]	[0.3710, 0.4527]
10	[0.6345, 0.7350]	[0.7625, 0.8415]	[0.2650, 0.3655]	[0.2650, 0.3655]
30	[0.8900, 0.9492]	[0.9197, 0.9652]	[0.0348, 0.0803]	[0.0508, 0.1100]
100	[0.9580, 0.9886]	[0.9639, 0.9905]	[0.0095, 0.0361]	[0.0114, 0.0420]
	$[\underline{P},\overline{P}](T^{M1} > T^{M4})$	$[\underline{P},\overline{P}](T^{M1} \geq T^{M4})$	$[\underline{P},\overline{P}](T^{M4} > T^{M1})$	$[\underline{P},\overline{P}](T^{M4} \geq T^{M1})$
1	[0.2359, 0.2694]	[0.7847, 0.8157]	[0.1843, 0.2153]	[0.7306, 0.7641]
5	[0.4114, 0.4975]	[0.6418, 0.7199]	[0.2801, 0.3582]	[0.5025, 0.5886]
10	[0.4592, 0.5760]	[0.6142, 0.7209]	[0.2791, 0.3858]	[0.4240, 0.5408]
30	[0.5173, 0.6865]	[0.5946, 0.7525]	[0.2475, 0.4054]	[0.3135, 0.4827]
100	[0.5598, 0.7719]	[0.5907, 0.7950]	[0.2050, 0.4093]	[0.2281, 0.4402]
	$[\underline{P},\overline{P}](T^{M2} > T^{M3})$	$[\underline{P},\overline{P}](T^{M2} \geq T^{M3})$	$[\underline{P},\overline{P}](T^{M3} > T^{M2})$	$[\underline{P},\overline{P}](T^{M3} \geq T^{M2})$
1	[0.2185, 0.2475]	[0.6659, 0.6986]	[0.3014, 0.3341]	[0.7525, 0.7815]
5	[0.2846, 0.3515]	[0.4700, 0.5448]	[0.4552, 0.5300]	[0.6485, 0.7154]
10	[0.2733, 0.3633]	[0.3939, 0.4935]	[0.5065, 0.6061]	[0.6367, 0.7267]
30	[0.4188, 0.5638]	[0.4825, 0.6262]	[0.3738, 0.5175]	[0.4362, 0.5812]
100	[0.4246, 0.6192]	[0.4503, 0.4638]	[0.3562, 0.5497]	[0.3808, 0.5754]
	$[\underline{P},\overline{P}](T^{M2} > T^{M4})$	$[\underline{P},\overline{P}](T^{M2} \geq T^{M4})$	$[\underline{P},\overline{P}](T^{M4} > T^{M2})$	$[\underline{P},\overline{P}](T^{M4} \geq T^{M2})$
1	[0.1725, 0.1992]	[0.6263, 0.6604]	[0.3396, 0.3737]	[0.8008, 0.8275]
5	[0.1848, 0.2420]	[0.3505, 0.4251]	[0.5749, 0.6495]	[0.7580, 0.8152]
10	[0.1499, 0.2201]	[0.2448, 0.3347]	[0.6653, 0.7552]	[0.7799, 0.8501]
30	[0.0831, 0.1618]	[0.1129, 0.2077]	[0.7923, 0.8871]	[0.8382, 0.9169]
100	[0.0381, 0.1075]	[0.0443, 0.1210]	[0.8790, 0.9557]	[0.8925, 0.9619]
	$[\underline{P},\overline{P}](T^{M3} > T^{M4})$	$[\underline{P},\overline{P}](T^{M3} \geq T^{M4})$	$[\underline{P},\overline{P}](T^{M4} > T^{M3})$	$[\underline{P},\overline{P}](T^{M4} \geq T^{M3})$
1	[0.1931, 0.2228]	[0.6843, 0.7190]	[0.2810, 0.3157]	[0.7772, 0.8069]
5	[0.2451, 0.3137]	[0.4389, 0.5196]	[0.4804, 0.5611]	[0.6863, 0.7549]
10	[0.2279, 0.3188]	[0.3500, 0.4556]	[0.5444, 0.6500]	[0.6812, 0.7721]
30	[0.0746, 0.1477]	[0.1020, 0.1910]	[0.8090, 0.8980]	[0.8523, 0.9254]
100	[0.0318, 0.0927]	[0.0371, 0.1048]	[0.8952, 0.9629]	[0.9073, 0.9682]

Table 4.7: NPI lower and upper probabilities for pairwise comparison for T^{M1}, T^{M2}, T^{M3} and T^{M4}

To compare two of these tests, the NPI upper probabilities as given in Equation (4.10) and the corresponding NPI lower probabilities are presented in Table 4.7, for $m_x = m_y = m$. It is noticed that T^{M1} is better than both T^{M2} and T^{M3}, and the differences between the two tests become greater for large values of m as equal outcomes become less likely. When we look at the total number of successfully diagnosed in the data in Table 4.6, the total number of successfully diagnosed for T^{M1} is greater than the corresponding number for both T^{M2} and T^{M3}. It is also noticed that the imprecision is very low since the lower and upper probabilities are both close to 1 in the cases of $T^{M1} > T^{M2}$ and $T^{M1} > T^{M3}$. T^{M1} is also better than T^{M4} but the values of these lower and upper probabilities for the events $T^{M1} > (\geq)\ T^{M4}$ are not very high, although they increase for large values of m. Further, the imprecision tends to increase for large values of m. The total number of successfully diagnosed in the data for T^{M1} is greater than the corresponding numbers for T^{M4}, but still the differences between these numbers for both groups are small.

To compare T^{M2} and T^{M3}, we notice that T^{M3} is better than T^{M2} for $m = 1, 5, 10$, while T^{M2} is better than T^{M3} for $m = 30, 100$. That is because the total number of successful diagnoses in the data for T^{M3} is greater than the corresponding number for T^{M2} for $m = 1, 5, 10$, whereas for $m = 30, 100$ the number of successful diagnoses from group X for T^{M2} is greater than the corresponding number for T^{M3} and the corresponding number from group Y are equal for the two tests. Finally, the last two tables from Table 4.7 show that T^{M4} is better than both T^{M2} and T^{M3} where the differences between the two tests become greater for large values of m. It is clear because the total number of successful diagnoses in the data for T^{M4} is greater than the corresponding number for T^{M2} and T^{M3}.

Table 4.8 presents the NPI lower and upper probabilities for comparison of T^{M2} and T^{M3}. Here, we use the same value of c for all m, in order to consider the impression for different m. Actually, we use c resulting from the 2-NPI-L method with $m = 30$ and $\alpha = \beta = 0.5$. So, the numbers of successful diagnoses from groups X and Y are $s_x^2 = 58, s_y^2 = 27$ for T^{M2} and $s_x^3 = 57, s_y^3 = 27$ for T^{M3}. This table shows that the values of the lower and upper probabilities for $T^{M2} > (\geq)\ T^{M3}$ are a bit higher than for $T^{M3} > (\geq)\ T^{M2}$, but the values between the two tests are close. The imprecision tends to increase with m. The decision which test is the best can be supported using the relevant values of these lower and upper probabilities. Therefore, one can prefer the better test for the values of m that have the lower probability that exceed 0.5, so that can be a strong indication of this test being better than the other. Thus, in this example, one can conclude that T^{M2} is at least as good as T^{M3} for the next 19 patients and 19 non-patients as the lower probability for $T^{M2}(\geq)\ T^{M3}$ is equal to 0.5009 for $m = 19$. However when m_x and m_y equal to 20 or more, we could conclude that neither test is really better than the other.

m	$[\underline{P}, \overline{P}](T^{M2} > T^{M3})$	$[\underline{P}, \overline{P}](T^{M2} \geq T^{M3})$	$[\underline{P}, \overline{P}](T^{M3} > T^{M2})$	$[\underline{P}, \overline{P}](T^{M3} \geq T^{M2})$
1	[0.2771, 0.3070]	[0.7027, 0.7323]	[0.2677, 0.2973]	[0.6930, 0.7229]
5	[0.3829, 0.4532]	[0.5695, 0.6389]	[0.3611, 0.4305]	[0.5468, 0.6171]
10	[0.4040, 0.5001]	[0.5309, 0.6258]	[0.3742, 0.4691]	[0.4999, 0.5960]
19	[0.4148, 0.5389]	[0.5009, 0.6236]	[0.3764, 0.4991]	[0.4611, 0.5852]
20	[0.4154, 0.5418]	[0.4987, 0.6238]	[0.3762, 0.5013]	[0.4582, 0.5846]
50	[0.4207, 0.5884]	[0.4649, 0.6314]	[0.3686, 0.5351]	[0.4116, 0.5793]
100	[0.4208, 0.6148]	[0.4464, 0.6394]	[0.3606, 0.5536]	[0.3852, 0.5792]

Table 4.8: NPI lower and upper probabilities for comparison of T^{M2} and T^{M3}

As mentioned in section 4.1, the area under the ROC curve (AUC) has been used in the literature for comparison of two diagnostic tests. We compare the results in this example with the empirical AUCs, which are equal to $\widehat{AUC}_{M1} = 0.9034$, $\widehat{AUC}_{M2} = 0.7526$, $\widehat{AUC}_{M3} = 0.8232$ and $\widehat{AUC}_{M4} = 0.8798$. These results are in line with our results, however T^{M2} can be better than T^{M3} for large values of m, which does not show in the comparison of the $\widehat{AUC}$ where m does not play a role.

4.6 Concluding remarks

This chapter introduced comparison of two diagnostic tests, assuming the tests are applied on the same individuals from two groups, healthy and diseased individuals, explicitly as a predictive problem where the inference is based on future individuals. We considered comparison of the total number of correct diagnoses for m_x future healthy individuals and m_y future patients in one test with those in the other test. We discussed the influence of the choice of the number of future individuals considered via examples.

If the tests perform similarly, it is possible that there is no strong, or even week, indication of one test being better than the other, due to the imprecision in our method. It may happen that there is a strong indication of one test being better than another. In such cases, one would recommend the better test only for such small numbers of future patients and ideally reconsider the decision once more information is available. Real world implementation of such recommendations will required further research. Similar reasoning is used in [12] to determine maximum group size for simultaneous testing in high potential risk scenarios. We also introduced weights to reflect the relative importance of the two groups.

The NPI approach can be attractive for inference to promote decisions on medical diagnoses for a predetermined number of future patients. We have restricted attention to comparison between two diagnostic tests on individuals from two groups. This can be generalized to such comparison on individuals from more than two groups. We leave that for future research. Comparison of more than two diagnostic tests is also an interesting topic for further research.

In some medical applications, the false-positive rates should be restricted within the medical interest. For example, to accept cancer screening tests, the false-positive rates have to be very small [46]. Many researchers have suggested the use of the partial area under the ROC carve for such problems, for example, Baker and Pinsky [8] designed a study using this in order to compare the performance of the digital and analog mammography for breast cancer screening over false-positive rates not exceeding 0.01. We have not linked our methods to such a setting, but this provides an interesting topic for future research.

Chapter 5

Concluding Remarks

In this thesis, we have presented new NPI methods to determine optimal diagnostic thresholds by considering specific numbers of future individuals in each of the groups. We have seen that the optimal thresholds might change if the numbers of future individuals change. This raises the question how to choose those numbers in practical applications, where it should be noted that we would not actually know the group to which a future individual, to whom the test is applied, belongs. Guidance on the choice of those numbers in practical situations is left as a topic for future research. One would expect that it is good to choose those numbers reflecting expected numbers of patients and healthy people over a specific period of time.

We have presented NPI methods for selecting optimal thresholds for two- and three-group classification problems. We have considered m future individuals in each group for whom the threshold would be applied, and criteria in terms of the proportions of successful diagnoses. These methods were shown to depend on the target success proportions (α, β and γ) and also on the value of m. How α, β and γ can be chosen in real applications would require further research in future. We have restricted attention to compare our method with the empirical Youden index and the maximum area (volume) methods, as these methods also take only few model assumptions, it is of interest to compare our NPI approach with other methods presented in Section 1.2.

In addition, we have presented the NPI for comparison of two diagnostic tests for a particular number of future individuals from two groups. If the tests preform similarly, it is possible that there is no strong, or even weak, indication that one test is better than the other, due to the imprecision in our method. It may be that there is a strong indication that one test is better than another. Then, the better test is recommended only for such small numbers of future patients and ideally we should reconsider the decision once more information is available. Weights have also been introduced to reflect the relative importance of the two groups.

Further research will be needed to consider aspects of practical implementation of our methods. One issue is that, if we wish to choose one test for implementation, based on the results presented in Chapter 4, then we may e.g. have a strong indication that Test 1 is better than Test 2 for a range of values of m_x and m_y. But in practice we may only be able to decide on the use of a test for the next total number of people, not knowing whether they are patients or healthy. One careful way to resolve this is to only recommend a Test 1 for the next m_{min} people, who can be either patients or healthy, with m_{min} the largest value such that Test 1 is strongly indicated to be better than Test 2 for $m_{min} = \min\{m_x, m_y\}$.

NPI is a statistics method with strong frequentist properties, in line with the notion of exact calibration as introduced by Lawless and Fredette [44]. Contrary to most classical frequentist statistics methods, NPI does not consider data as resulting from an assumed sampling method related to an assumed population. Instead, by focusing on future observations, the variation is in the possible orderings of the data observations and future observations, so the randomness is explicitly in the prediction. In absence of knowledge about the an underling population distribution, this is an alternative approach. If one had such additional knowledge, then one could attempt to combine NPI with aspects of sample variation; this is an interesting topic for future research.

Bibliography

[1] Alonzo, T.A., Nakas, C.T., Yiannoutsos, C.T. and Bucher, S. (2009). A comparison of tests for restricted orderings in the three-class case. *Statistics in Medicine*, 28, 1144-1158.

[2] Alqifari, H. (2017). *Nonparametric Predictive Inference for Future Order Statistics.* PhD thesis, Durham University.

[3] Aoki, K., Misumi, J., Kimura, T., Zhao, W. and Xie, T. (1997). Evaluation of cutoff levels for screening of gastric cancer using serum pepsinogens and distributions of levels of serum pepsinogen I, II and of PG I/PG II ratios in a gastric cancer case-control study. *Journal of Epidemiology*, 7, 143-151.

[4] Attwood, K., Tian, L. and Xiong, C. (2014). Diagnostic thresholds with three ordinal groups. *Journal of Biopharmaceutical Statistics*, 24, 608-633.

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

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

[7] Baker, R. (2010). *Multinomial Nonparametric Predictive Inference: Selection, Classification and Subcategory data.* PhD thesis, Durham University.

[8] Baker, S.G. and Pinsky, P.F. (2001). A proposed design and analysis for comparing digital and analog mammography: special receiver operating characteristic methods for cancer screening. *Journal of the American Statistical Association*, 96, 421-428.

[9] Bellotti, T. and Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, 36, 3302-3308.

[10] Betensky, R.A. and Rabinowitz, D. (1999). Maximally selected χ^2 statistics for $k \times 2$ tables. *Biometrics*, 55, 317-320.

[11] Coffin, M. and Sukhatme, S. (1997). Receiver operating characteristic studies and measurement errors. *Biometrics*, 53, 823-837.

[12] Coolen, F.P.A. (2013). Maximum group sizes for simultaneous testing in high potential risk scenarios. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability,* 227, 569-575.

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

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

[15] Coolen, F.P.A. and Augustin, T. (2005). Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model. In *Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, 5, 125-134.

[16] Coolen, F.P.A., Coolen-Maturi, T. and Alqifari, H.N. (2017). Nonparametric predictive inference for future order statistics. *Communications in Statistics-Theory and Methods*, to appear.

[17] Coolen, F.P.A and van der Laan, P. (2001). Imprecise predictive selection based on low structure assumptions. *Journal of Statistical Planning and Inference*, 98, 259-277.

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

[19] Coolen, F.P.A. and Yan, K.J. (2003). Nonparametric Predictive Comparison of Two Groups of Lifetime Data. In *Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications,* 3, 148-161.

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

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

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

[23] Coolen-Maturi, T., Coolen, F.P.A. and Muhammad, N. (2016). Predictive inference for bivariate data: Combining nonparametric predictive inference for marginals with an estimated copula. *Journal of Statistical Theory and Practice*, 10, 515-538.

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

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

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

[27] Cox, L.H, Johnson, M.M and Kafadar, K. (1982). Exposition of statistical graphics technology. *ASA Statistical Computing Section*, 55–56.

[28] Crystal, H., Dickson, D., Fuld, P., Masur, D., Scott, R., Mehler, M., Masdeu, J., Kawas, C., Aronson, M. and Wolfson, L. (1988). Clinico-pathologic studies in de-

mentia nondemented subjects with pathologically confirmed Alzheimer's disease. *Neurology*, 38, 1682-1682.

[29] De Finetti, B. (1974) *Theory of Probability.* Wiley, London.

[30] Demir, A., Yarali, N., Fisgin, T., Duru, F. and Kara, A. (2002). Most reliable indices in differentiation between thalassemia trait and iron deficiency anemia. *Pediatrics International*, 44, 612-616.

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

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

[33] Fluss, R., Faraggi, D. and Reiser, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biometrical Journal*, 47, 458-472.

[34] Geisser, S. (1993). *Predictive Inference: an Introduction.* Chapman and Hall, London.

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

[36] Hand, D.J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77, 103-123.

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

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

[39] Hothorn, T. and Zeileis, A. (2008). Generalized maximally selected statistics. *Biometrics*, 64, 1263-1269.

[40] Hsiao, J.K., Bartko, J.J. and Potter, W.Z. (1989). Diagnosing diagnoses: Receiver operating characteristic methods and psychiatry. *Archives of General Psychiatry*, 46, 664-667.

[41] Jalali, R. and Rezaie, M. (2005). Predicting pressure ulcer risk: comparing the predictive validity of 4 scales. *Advances in Skin & Wound Care*, 18, 92-97.

[42] Krzanowski, W.J. and Hand, D.J. (2009). *ROC Curves for Continuous Data.* CRC Press, Boca Raton.

[43] Lai, C.Y., Tian, L. and Schisterman, E.F. (2012). Exact confidence interval estimation for the Youden index and its corresponding optimal cut-point. *Computational Statistics & Data Analysis*, 56, 1103-1114.

[44] Lawless, J.F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, 92, 529-542.

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

[46] Lilienfeld, A.M. (1974). Some limitations and problems of screening for cancer. *Cancer*, 33, 1720-1724.

[47] Liu, X. (2012) Classification accuracy and cut point selection. *Statistics in Medicine*, 31, 2676-2686.

[48] Metz, C.E. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigative Radiology*, 24, 234-245.

[49] Miller, R. and Siegmund, D. (1982). Maximally selected chi square statistics. *Biometrics*, 38, 1011-1016.

[50] Molanes-López, E.M. and Letón, E. (2011). Inference of the Youden index and associated threshold using empirical likelihood for quantiles. *Statistics in Medicine*, 30, 2467-2480.

[51] Morris, J.C., McKeel, D.W., Storandt, M., Rubin, E.H., Price, J.L., Grant, E.A., Ball, M.J. and Berg, L. (1991). Very mild Alzheimer's disease Informant-based clinical, psychometric, and pathologic distinction from normal aging. *Neurology*, 41, 469-469.

[52] Mossman, D. (1999). Three-way ROCs. *Medical Decision Making*, 19, 78-89.

[53] Nakas, C.T. (2014). Developments in ROC surface analysis and assessment of diagnostic markers in three-class classification problems. *REVSTAT-Statistical Journal*, 12, 43-65.

[54] Nakas, C.T., Alonzo, T.A. and 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.

[55] Nakas, C.T., Dalrymple-Alford, J.C., Anderson, T.J. and Alonzo, T.A. (2013). Generalization of Youden index for multiple-class classification problems applied to the assessment of externally validated cognition in Parkinson disease screening. *Statistics in Medicine*, 32, 995-1003.

[56] Nakas, C.T. and Yiannoutsos, C.T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine*, 23, 3437-3449.

[57] Ng, P.C., Cheng, S.H., Chui, K.M., Fok, T.F., Wong, M.Y., Wong, W., Wong, R.P.O and Cheung, K.L. (1997). Diagnosis of late onset neonatal sepsis with cytokines, adhesion molecule, and C-reactive protein in preterm very low birthweight infants. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 77,F221-F227.

[58] Pekkanen, J. and Pearce, N. (1999). Defining asthma in epidemiological studies. *European Respiratory Journal*, 14, 951-957.

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

[60] Perkins, N. J. and Schisterman, E. F. (2006). The inconsistency of 'ptimal' cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, 163, 670-675.

[61] Sarafidis, K., Soubasi-Griva, V., Piretzi, K., Thomaidou, A., Agakidou, E., Taparkou, A., Diamanti, E. and Drossou-Agakidou, V. (2010). Diagnostic utility of elevated serum soluble triggering receptor expressed on myeloid cells (sTREM)-1 in infected neonates. *Intensive Care Medicine*, 36, 864-868.

[62] Schäfer, H. (1989). Constructing a cut-off point for a quantitative diagnostic test. *Statistics in Medicine*, 8, 1381-1391.

[63] Schisterman, E.F. and Perkins, N. (2007). Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistic: Simulation and Computation*, 36, 549-563.

[64] Schisterman, E.F., Perkins, N.J., Liu, A. and Bondell, H. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology*, 16, 73-81.

[65] Sharir, T., Berman, D.S., Waechter, P.B., Areeda, J., Kavanagh, P.B., Gerlach, J., Kang, X and Germano, G. (2001). Quantitative analysis of regional motion and thickening by gated myocardial perfusion SPECT: Normal heterogeneity and criteria for abnormality. *Journal of Nuclear Medicine*, 42, 1630-1638.

[66] Unal, I. (2017). Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach. *Computational and Mathematical Methods in Medicine.*

[67] Verboon-Maciolek, M.A., Thijsen, S.F., Hemels, M.A., Menses, M., van Loon, A.M., Krediet, T.G., Fleer, A., Voorbij, H.A. and Rijkers, G.T. (2006). Inflammatory mediators for the diagnosis and treatment of sepsis in early infancy. *Pediatric Research*, 59, 457-461.

[68] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities.* Chapman & Hall, London.

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

[70] Xiong, C., van Belle, G., Miller, J.P. and Morris, J.C. (2006). Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Statistics in Medicine*, 25, 1251-1273.

[71] Xiong, C., van Belle, G., Miller, J.P., Yan, Y., Gao, F., Yu, K. and Morris, J.C. (2007). A parametric comparison of diagnostic accuracy with three ordinal diagnostic groups. *Biometrical Journal*, 49, 682-693.

[72] Youden, W.J. (1950). Index for rating diagnostic tests. *Cancer*, 3, 32-35.

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

Appendix A

Proof for the case $m_x = m_y = 1$

$$
\begin{aligned}
&\overline{P}(w_x C_{m_x}^{X^1} + w_y C_{m_y}^{Y^1} > w_x C_{m_x}^{X^2} + w_y C_{m_y}^{Y^2}) \\
&= \sum_{k=0}^{w_x m_x + w_y m_y} \left[\sum_{v=0}^{m_y} \overline{P}(C_{m_x}^{X^2} < \frac{k-(w_y v)}{w_x}) \times [\overline{P}(C_{m_y}^{Y^2} \le v) - \overline{P}(C_{m_y}^{Y^2} \le v-1)] \right] \\
&\times \left[\sum_{v=0}^{m_y} [\overline{P}(C_{m_x}^{X^1} \ge \frac{k-(w_y v)}{w_x}) - \overline{P}(C_{m_x}^{X^1} \ge \frac{k+1-(w_y v)}{w_x})] \times [\overline{P}(C_{m_y}^{Y^1} \ge v) - \overline{P}(C_{m_y}^{Y^1} \ge v+1)] \right]
\end{aligned} \tag{A.1}
$$

For $m_x = m_y = 1$, the equation (A.1) becomes

$$
\begin{aligned}
&\overline{P}(w_x C_{m_x}^{X^1} + w_y C_{m_y}^{Y^1} > w_x C_{m_x}^{X^2} + w_y C_{m_y}^{Y^2}) \\
&= \sum_{k=0}^{w_x + w_y} \{ \overline{P}(C_1^{X^2} < \frac{k}{w_x}) \times \overline{P}(C_1^{Y^2} \le 0) \\
&+ \overline{P}(C_1^{X^2} < \frac{k-w_y}{w_x}) \times [\overline{P}(C_1^{Y^2} \le 1) - \overline{P}(C_1^{Y^2} \le 0)] \} \\
&\times \{ [\overline{P}(C_1^{X^1} \ge \frac{k}{w_x}) - \overline{P}(C_1^{X^1} \ge \frac{k+1}{w_x})] \times [\overline{P}(C_1^{Y^1} \ge 0) - \overline{P}(C_1^{Y^1} \ge 1)] \\
&+ [\overline{P}(C_1^{X^1} \ge \frac{k-w_y}{w_x}) - \overline{P}(C_1^{X^1} \ge \frac{k+1-w_y}{w_x})] \times \overline{P}(C_1^{Y^1} \ge 1) \}
\end{aligned} \tag{A.2}
$$

We consider all the events in the Equation (A.2).

1. $\overline{P}(C_1^{X^2} < \frac{k}{w_x}) = \begin{cases} \overline{P}(C_1^{X^2} < 1) = \overline{P}(C_1^{X^2} = 0) & \text{if} \quad k = 1, \ldots, w_x \\ \overline{P}(C_1^{X^2} < 2) = 1 & \text{if} \quad k = w_x + 1, \ldots, w_x + w_y \end{cases}$

2. $\overline{P}(C_1^{X^2} < \frac{k-w_y}{w_x}) = \begin{cases} \overline{P}(C_1^{X^2} < 0) = 0 & \text{if} \quad k = 1, \ldots, w_y \\ \overline{P}(C_1^{X^2} < 1) = \overline{P}(C_1^{X^2} = 0) & \text{if} \quad k = w_y + 1, \ldots, w_x + w_y \end{cases}$

3. $\overline{P}(C_1^{X^1} \geq \frac{k}{w_x}) = \begin{cases} \overline{P}(C_1^{X^1} \geq 1) = \overline{P}(C_1^{X^1} = 1) & \text{if} \quad k = 1, \ldots, w_x \\ \overline{P}(C_1^{X^1} \geq 2) = 0 & \text{if} \quad k = w_x + 1, \ldots, w_x + w_y \end{cases}$

4. $\overline{P}(C_1^{X^1} \geq \frac{k+1}{w_x}) = \begin{cases} \overline{P}(C_1^{X^1} \geq 1) = \overline{P}(C_1^{X^1} = 1) & \text{if} \quad k = 1, \ldots, w_x - 1 \\ \overline{P}(C_1^{X^1} \geq 2) = 0 & \text{if} \quad k = w_x, \ldots, w_x + w_y \end{cases}$

5. $\overline{P}(C_1^{X^1} \geq \frac{k-w_y}{w_x}) = \begin{cases} \overline{P}(C_1^{X^1} \geq 0) = 1 & \text{if} \quad k = 1, \ldots, w_y \\ \overline{P}(C_1^{X^1} \geq 1) = \overline{P}(C_1^{X^1} = 1) & \text{if} \quad k = w_y + 1, \ldots, w_x + w_y \end{cases}$

6. $\overline{P}(C_1^{X^1} \geq \frac{k+1-w_y}{w_x}) = \begin{cases} \overline{P}(C_1^{X^1} \geq 0) = 1 & \text{if} \quad k = 1, \ldots, w_y - 1 \\ \overline{P}(C_1^{X^1} \geq 1) = \overline{P}(C_1^{X^1} = 1) & \text{if} \quad k = w_y, \ldots, w_x + w_y - 1 \\ \overline{P}(C_1^{X^1} \geq 2) = 0 & \text{if} \quad k = w_x + w_y \end{cases}$

These simple form of the upper probabilities lead to the following results.

For $w_x < w_y$

$$\begin{aligned} &\overline{P}(w_x C_1^{X^1} + w_y C_1^{Y^1} > w_x C_1^{X^2} + w_y C_1^{Y^2}) \\ &= \{\overline{P}(C_1^{Y^2} = 0)[\overline{P}(C_1^{X^2} = 0)\overline{P}(C_1^{X^1} = 1)\underline{P}(C_1^{Y^1} = 0) \\ &+ \overline{P}(C_1^{Y^1} = 1)(1 - \overline{P}(C_1^{X^1} = 1)\overline{P}(C_1^{X^2} = 0))]\} \\ &+ \overline{P}(C_1^{X^1} = 1)\overline{P}(C_1^{Y^1} = 1)\overline{P}(C_1^{X^2} = 0) \end{aligned} \qquad \text{(A.3)}$$

For $w_x > w_y$, the expression is effectively the same, but with X and Y interchanged.

For $w_y = w_x$

$$\begin{aligned}
&\overline{P}(w_x C_1^{X^1} + w_y C_1^{Y^1} > w_x C_1^{X^2} + w_y C_1^{Y^2}) \\
&= \overline{P}(C_1^{X^2} = 0)\overline{P}(C_1^{Y^2} = 0)[\overline{P}(C_1^{X^1} = 1) + \overline{P}(C_1^{Y^1} = 1)] \\
&+ \overline{P}(C_1^{X^1} = 1)\overline{P}(C_1^{Y^1} = 1)[\overline{P}(C_1^{Y^2} = 0) + \overline{P}(C_1^{X^2} = 0)] \\
&- 3\overline{P}(C_1^{X^2} = 0)\overline{P}(C_1^{Y^2} = 0)\overline{P}(C_1^{X^1} = 1)\overline{P}(C_1^{Y^1} = 1) \qquad \text{(A.4)}
\end{aligned}$$