Nonparametric Predictive Inference for Multiple Future Ordinal Observations

ALHARBI, ABDULMAJEED,ABDULLAH,R

Nonparametric Predictive Inference for Multiple Future Ordinal Observations

Abdulmajeed Abdullah R. Alharbi

A Thesis presented for the degree of
Doctor of Philosophy

Department of Mathematical Sciences
Durham University
United Kingdom

October 2024

Dedication

For the sake of Allah, all praise is due to Him.

To my dear parents, for their unlimited love and prayers.

To the soul of my brother, may Allah have mercy on him.

To my siblings, for their encouragement and belief in me.

To my wife, for being supportive and standing by me.

To my wonderful children, you are the stars that light my path.

Nonparametric Predictive Inference for Multiple Future Ordinal Observations

Abdulmajeed Abdullah R. Alharbi

Submitted for the degree of Doctor of Philosophy
October 2024

Abstract

Nonparametric predictive inference (NPI) is a statistical methodology based on the assumption $A_{(n)}$ proposed by Hill for the prediction of a future observation [60]. NPI uses lower and upper probabilities to quantify uncertainty. NPI has been developed for various data types, and the explicitly predictive nature of NPI makes the method particularly attractive and well-suited for a wide variety of statistical applications. This thesis proposes novel contributions to statistical methods for ordinal data using the NPI method with multiple future observations. The method uses a latent variable representation of the data observations and ordered categories on the real-line.

NPI lower and upper probabilities for several events involving multiple future ordinal observations are presented. The NPI method is applied to selection problems involving multiple future ordinal observations. Pairwise comparison of future observations from two independent groups is presented. The accuracy of diagnostic tests with ordinal outcomes is considered, with NPI-based methods introduced for selecting the optimal thresholds of a diagnostic test, initially for two-group classification and then extended to three-group classification.

To illustrate the proposed NPI methods, examples using data from the literature are provided. Simulation studies are conducted to investigate the predictive performance of the proposed methods for selecting diagnostic test thresholds and to compare these methods with classical methods, such as the Youden index, Liu index and maximum volume methods. The results indicate that the NPI methods tend to outperform the classical approaches by correctly classifying more individuals in each group. Overall, the number of future observations considered influences the NPI lower and upper probabilities, affecting category selection, pairwise comparison, and diagnostic threshold selection.

Declaration

The work in this thesis is based on research carried out at the Department of Mathematical Sciences, Durham University, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

In the name of Allah, the Most Merciful, the Most Compassionate. First and foremost, I would like to say “Alhamdulillah”. All praise and gratitude are due to the Great Almighty, Allah, for the countless blessings He has bestowed upon me. I am truly grateful for His mercy and guidance, which have illuminated my life, and for giving me the strength and opportunity to carry out and complete this thesis.

Special appreciation and thanks go to my supervisors, Prof. Frank Coolen and Prof. Tahani Coolen-Maturi, for being sources of support throughout my PhD studies. Thank you for constantly providing expert advice, guidance, and support. It is difficult to find the right words to express my gratitude.

To my precious mother, your tireless efforts, unwavering faith, and endless love have shaped not only my achievements but also the person I have become. You are the heart of everything I hold dear, and I am forever grateful. To my dear father, words cannot capture the depth of my emotions. I am profoundly thankful for your endless support, not only in my educational pursuits but throughout my life’s journey. I would also like to extend my sincere gratitude to my siblings for their encouraging and unwavering belief in me.

My special thanks and warm appreciation go to my beloved wife Fatma, for her everlasting support. When everything seemed too hard, her unerring confidence in me motivated me to persevere. Thank you for your patience, love, kindness, prayers, and for standing by my side. I would also like to thank my wonderful children, please accept my heartfelt gratitude. Thank you for lighting up my life with your smiles.

I would like to thank my friends, who have been a constant source of support and motivation. I am also grateful to King Saud University in Riyadh and the Saudi Arabian Cultural Bureau in London for their financial support and for providing me with the scholarship that enabled me to complete my studies abroad. Finally, I would like to thank Durham University for offering an enjoyable academic atmosphere.

Contents

Chapter 1

Introduction

In many applications involving categorical data, the categories are ordered. For example, categories may represent different levels of disease severity [61]. In statistics, this type of data is referred to as ordinal data. Ordinal data occur in many fields, such as medicine, public health, marketing, education, and the social sciences [2, 3, 79]. For instance, in the social sciences, surveys often evaluate opinions and attitudes using ordered categories like *Strongly disagree*, *Disagree*, *Neither agree nor disagree*, *Agree*, and *Strongly agree*. Similarly, in medicine, patients can be categorized into ordered pain levels such as *No Pain*, *Mild*, *Moderate*, *Severe*, and *Very Severe Pain*.

This thesis proposes novel contributions to statistical methods for ordinal data using the nonparametric predictive inference (NPI) method [10, 26]. NPI is a statistical methodology based on Hill's assumption $A_{(n)}$ [60], quantifying uncertainty through lower and upper probabilities and explicitly focusing on future observations. The explicitly predictive nature of NPI makes it particularly attractive and well suited for a wide variety of statistical applications and for different data types [28].

Coolen et al. [32, 33] and Elkhafifi [47] developed NPI for ordinal data, focusing on a single future observation. They presented NPI for ordinal data using a latent variable representation of the observations, with the categories ordered and represented by intervals on the real-line. They also derived closed-form formulae for the NPI lower and upper probabilities for events involving the next future observation and briefly compared these inferences to NPI for multinomial data where the order of the categories is not taken into account. This thesis extends NPI for ordinal data to multiple future observations.

This chapter provides the necessary background for the later chapters. Section 1.1 provides an overview of the NPI method. Section 1.2 summarises the NPI results for ordinal data with a single future observation. Section 1.3 presents an overview of NPI-based selection methods. Section 1.4 introduces the concepts of diagnostic test accuracy and main methods used in the literature to determine diagnostic thresholds. Finally, Section 1.5 outlines the structure of this thesis.

1.1 Nonparametric Predictive Inference (NPI)

The concept of imprecise probabilities dates back to the mid-19th century when it was introduced by Boole in 1854 [19]. In classical probability theory, uncertainty about an event A is quantified using a single precise probability $P(A) \in [0, 1]$ that satisfies Kolmogorov's axioms [11]. However, if information about A is incomplete or vague, a unique probability may be too restrictive, so using an imprecise probability is an appropriate alternative approach. The imprecise probability concept uses an interval probability for uncertainty quantification instead of a single precise probability $P(A)$ [11]. The interval probability is bounded by a lower probability $\underline{P}(A)$ and an upper probability $\overline{P}(A)$, such that $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$. Therefore, lower and upper probabilities generalise the classical theory of precise probability. When $\underline{P}(A) = \overline{P}(A)$, classical probability occurs as a special case of imprecise probability. The case when $[\underline{P}(A), \overline{P}(A)] = [0, 1]$ reflects complete lack of knowledge or absence of information about the event A. The statistical method used in this thesis, known as Nonparametric Predictive Inference (NPI), uses lower and upper probabilities to quantify uncertainty [11, 28].

NPI is based on Hill's assumption $A_{(n)}$ [60] which gives a direct conditional probability for a future observable random quantity, conditional on observed values of related random quantities [10, 27]. This assumption is introduced to predict one or more future observations and is suitable for cases where no prior knowledge about the underlying distribution exists, or where one does not want to use any such knowledge or further assumptions. Introducing the assumption $A_{(n)}$ requires some notation. Let $X_1, \ldots, X_n, X_{n+1}$ be exchangeable and continuous real-valued random quantities. The corresponding ordered observed values of $X_1, \ldots, X_n$ are represented by $x_1 < x_2 < \ldots < x_n$, with $x_0 = -\infty$ and $x_{n+1} = \infty$ defined for ease of notation. It is assumed that no ties exist among the real-valued data for ease of presentation; therefore, the probability of having ties is assumed to be 0. However, if there are ties in the data, then these can be handled in NPI by assuming that tied observations differ by small amounts that tend to zero, but this is irrelevant in this work as we assume no ties occur in the assumed latent variable representation (as discussed in Section 1.2). Based on n observations that divide the real-line into $n+1$ intervals, $I_i = (x_{i-1}, x_i)$ for $i = 1, \ldots, n+1$, the assumption $A_{(n)}$ is that a future observation X_{n+1} will be in interval I_i with probability $\frac{1}{n+1}$ [28], that is, for each $i = 1, \ldots, n+1$,

$$P(X_{n+1} \in I_i) = \frac{1}{n+1} \tag{1.1}$$

The assumption $A_{(n)}$ is a post-data assumption related to finite exchangeability [45]. Inferences based on $A_{(n)}$ are predictive and nonparametric and they can be appropriate to be applied if there is hardly any knowledge about the random quantity of interest, or if one does not want to use such information. For many events of interest, the assumption $A_{(n)}$ is not sufficient to derive precise probabilities, however, it yields lower and upper bounds for probabilities [45]. Based on the assumption $A_{(n)}$, Augustin and Coolen [10] introduced NPI lower and upper probabilities.

The NPI method can be used for $m \geq 1$ future observations via Hill's assumptions $A_{(n)}, A_{(n+1)}, \ldots, A_{(n+m-1)}$. These jointly are referred to as 'the $A_{(\cdot)}$ assumptions'. The $A_{(\cdot)}$ assumptions imply that all possible orderings of n data observations and m future observations are equally likely [35], where the m future observations are not distinguished or ordered in any way. Let S_j denote the number of future observations that are in I_j, so $S_j = \#\{X_{n+l} \in I_j, l = 1, \ldots, m\}$. The $A_{(\cdot)}$ assumptions lead to

$$P(\bigcap_{j=1}^{n+1}\{S_j = s_j\}) = \binom{n+m}{m}^{-1} \tag{1.2}$$

where s_j, for $j = 1. \ldots, n+1$, are non-negative integers with $\sum_{j=1}^{n+1} s_j = m$.

Equation (1.2) implies that all $\binom{n+m}{m}$ orderings of m future observations among the n observations are equally likely. The NPI lower probability for an event involving the m future observations is the proportion of the orderings for which the event must hold, and the corresponding NPI upper probability is the proportion of the orderings for which the event can hold [10, 27].

The number $\binom{n+m}{m}$ of possible orderings of m future observations among n data observations is very large in many situations. Hence, it can be computationally expensive, or even impossible, to determine the NPI lower and upper probabilities exactly. The sampling of orderings method, originally introduced for studying test reproducibility [34], offers a practical solution, as it allows for the estimation of the NPI lower and upper probabilities by sampling of orderings, reducing computation time. The sampling of orderings method is particularly attractive when closed-form expressions for the NPI lower and upper probabilities are unavailable for specific events of interest, but where for each ordering it is easily verified if the event of interest must hold, can hold, or cannot hold. In this thesis, the sampling of orderings method is utilized to estimate NPI lower and upper probabilities, as presented in Section 3.4.

The sampling of orderings procedure is based on simple random sampling (SRS). The process ensures that each possible ordering is equally likely to be chosen during each selection, and that each selection of an ordering is independent of other selections [34]. It is important to note that, as the total number of orderings is large, any possible differences between sampling with or without replacement can be neglected, so sampling with replacement is applied.

Implementation of the sampling of orderings method involves sampling a vector of integers $(r_1, \ldots, r_n)$ such that $r_1 \geq 1$, $r_l > r_{l-1}$ for each $l = 2, \ldots, n$, and $r_n \leq m + n$. This set of integers establishes the ranks of the n data observations in the combined ranking of the n data and m future observations. By defining $r_0 = 0$ and $r_{n+1} = m + n + 1$, and with a sampled vector $(r_1, \ldots, r_n)$, we compute $S_l = r_l - r_{l-1} - 1$ for $l = 1, \ldots, n + 1$. This creates each future observation ordering in the SRS process, where the ordering is with respect to the combined ranking of the n data observations and the m future observations. This process ensures that each possible ordering has an equal probability of being selected and is independent of the other selections, which satisfies the requirements for SRS [34].

NPI has been introduced for a wide range of applications such as survival analysis, reliability testing, topics in operational research and finance [24, 28, 59]. In addition, NPI has been introduced for different types of data, including Bernoulli data [26], multinomial data [29], real-valued data [36], right-censored data [37] and bivariate data [41]. For more details about NPI, we refer to `www.npi-statistics.com`.

For example, Coolen and Augustin [29] presented the NPI method for multinomial data in the absence of prior knowledge of the relationship between the categories. Inferences for such data about the next future observation are based on the latent variable representation of the data using a probability wheel, where each category on the wheel is assumed to only be represented by one segment of the wheel. Note that the probability wheel representation of the data is partitioned into n equally-sized slices that are fully or partially within the segment. To reflect the knowledge of the order of the categories, Coolen et al. [32, 33] and Elkhafifi [47] developed NPI for ordinal data, in which the categories are ordered and represented by intervals on the real-line. Section 1.2 briefly presents NPI results for ordinal data, considering a single future observation.

The NPI method has also been developed to assess diagnostic test accuracy with different data types. For example, Coolen-Maturi et al. [43] introduced NPI for diagnostic test accuracy with binary data, while Elkhafifi and Coolen [48] presented NPI for diagnostic tests with ordinal data, considering a single future observation. Coolen-Maturi [38] generalised the results presented by Elkhafifi and Coolen [48] by developing NPI for three-group receiver operating characteristic (ROC) analysis, and this was generalised for more than three groups by Coolen-Maturi and Coolen [39]. Recently, Alabdulhadi [4] and Coolen-Maturi et al. [40] introduced the NPI approach for selecting the optimal threshold(s) for two- and three-group classification problems based on tests that yield real-valued results for a given number of future observations from each group. This thesis presents new methods for selecting the optimal diagnostic test threshold, considering multiple future ordinal individuals, in a two-group scenario and for selecting two thresholds in a three-group scenario. While the primary focus is on these two- and three-group problems, which are most commonly used in practice, the proposed methods can be generalised for scenarios with more than three groups, as will be discussed briefly in Section 5.9.

1.2 NPI for one future ordinal observation

This section summarises NPI for ordinal data considering a single future observation, based on Coolen et al. [33] and Elkhafifi [47]. They presented NPI for ordinal data using a latent variable representation of the observations, with the categories ordered and represented by intervals on the real-line. They derived closed-form formulae for the NPI lower and upper probabilities for events involving the next future observation, and these inferences were briefly compared to NPI for multinomial data [29]. The notation and assumed latent variable representation introduced here are also used in Chapter 2 where the method is generalised to multiple future observations.

Consider ordinal data with $K \geq 2$ categories, denoted by $C_1, \ldots, C_K$, where the ordering between them is indicated by the notation $C_1 < C_2 < \ldots < C_K$. Let the number of observations in category C_k denoted by n_k, for $k = 1, \ldots, K$, and let n be the total number of observations, so $\sum_{k=1}^{K} n_k = n$. The observations in each category are represented using an assumed underlying latent variable representation with the K categories represented by intervals on the real-line.

In the latent variable representation, the category C_k is assumed to be represented by the interval IC_k for $k = 1, \ldots, K$, where the K ordered intervals $IC_1, \ldots, IC_K$ form a partition of the real-line. Let the random quantity X_{n+1} represent a future ordinal observation, and let Y_{n+1} denote a latent observation on the real-line that corresponds to the future observation X_{n+1}, then $X_{n+1} \in C_k$ corresponds to the event $Y_{n+1} \in IC_k$. We further assume that the n observations are represented by $y_1 < \ldots < y_n$, of which n_k are in interval IC_k, these are also denoted by y_i^k, for $i = 1, \ldots, n_k$. The $A_{(n)}$ assumption can be applied to the latent variable Y_{n+1} and then transformed for inference on the random quantity X_{n+1}. The ordinal data structure is presented in Figure 1.1 [32, 47].

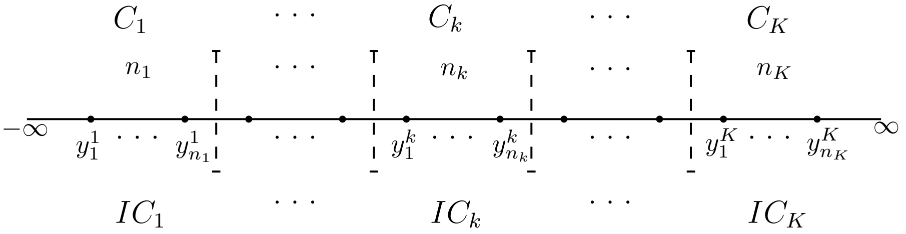


Figure 1.1: Ordinal data structure

The NPI lower and upper probabilities for general events of the form $X_{n+1} \in \mathcal{C}_T$ have been derived with $\mathcal{C}_T = \bigcup_{k \in T} C_k$ and T is assumed to be a strict subset of $\{1, \ldots, K\}$, so $T \subset \{1, \ldots, K\}$ [32, 47]. Using the latent variable representation, $\mathcal{C}_T$ is assumed to be represented by $\mathcal{IC}_T = \bigcup_{k \in T} IC_k$. Note that, if $T = \{1, \ldots, K\}$, then both NPI lower and upper probabilities for this event are equal to 1 and if $T = \emptyset$ both lower and upper probabilities are 0. Using the $A_{(n)}$ assumption for Y_{n+1} in the latent variable representation, each interval $I_j = (y_{j-1}, y_j)$, for $j = 1, \ldots, n+1$, has been assigned probability mass $\frac{1}{n+1}$. It should be emphasized that since the values of y_j only exist in the latent variable representation, their exact values are unknown, indeed they do not even exist, but this also conveniently enables us to assume that there are no ties between the y_j values. The NPI lower probability for the event $X_{n+1} \in \mathcal{C}_T$ is equal to the NPI lower probability for the corresponding latent observation event $Y_{n+1} \in \mathcal{IC}_T$. The NPI lower probability is

$$\underline{P}(X_{n+1} \in \mathcal{C}_T) = \underline{P}(Y_{n+1} \in \mathcal{IC}_T) = \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbf{1}\{I_j \subset \mathcal{IC}_T\} \tag{1.3}$$

where $\mathbf{1}\{E\}$ is equal to 1 if E is true and equal to 0 otherwise [32, 47]. The precise locations of the intervals IC_k are unknown, however, the fact that how many of y_i included within each interval IC_k is known leads to unique values for the NPI lower probability. The NPI upper probability for the event $X_{n+1} \in \mathcal{C}_T$ is similarly defined,

$$\overline{P}(X_{n+1} \in \mathcal{C}_T) = \overline{P}(Y_{n+1} \in \mathcal{IC}_T) = \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbf{1}\{I_j \cap \mathcal{IC}_T \neq \emptyset\} \tag{1.4}$$

While the NPI lower probability for the event of interest is derived by minimising the total probability mass assigned to $\mathcal{IC}_T$, the corresponding NPI upper probability can be derived by maximising the total probability mass that can be in $\mathcal{IC}_T$. Thus, the NPI upper probability is derived by summing all the probability masses that can be assigned to I_j which have a non-empty intersection with $\mathcal{IC}_T$.

In Chapter 2, NPI for ordinal data is generalised to events involving multiple future observations. Then, these events are applied to selection problems in Chapter 3, and NPI-based pairwise comparisons are developed. The following section provides an overview of the NPI-based selection methods.

1.3 Overview of NPI selection methods

The NPI methods developed for selection problems differ fundamentally from the classical methods in the literature. These classical methods include the indifference zone method introduced by Bechhofer [17] and Gupta's subset selection method [56]. Both methods are non-predictive, relying solely on hypothesis testing rather than incorporating predictive inferences for selection problems. Chapter 3 presents NPI-based selection methods considering multiple future ordinal observations where these methods use predictive inferences based on past observations and make use of Hill's assumptions $A_{(\cdot)}$ [33].

Selection methods based on NPI have been applied to real-valued data, including right-censored data, to select the group, or groups most likely to yield the largest next observation [36, 42, 44]. Only the next observation was taken into account when making inferences for such data. The NPI-based selection methods have also been applied to Bernoulli data [30, 31] and to multinomial data [12] with multiple future observations. This section presents an overview of these predictive selection methods.

The NPI selection method for real-valued data from different groups was developed by Coolen and van der Laan [36]. By making inferences about a single future observation from each group, the group which is most likely to provide the largest next observation was identified. The NPI lower and upper probabilities were derived for the event that one group's next observation will exceed the next observation of each other group. Additionally, a subset of groups was selected in two ways. The first way involves determining the NPI lower and upper probabilities for a subset containing the group providing the largest next observation. The second way involves determining the NPI lower and upper probabilities for the event that the next observation from every group in a subset, exceeds the next observation from each non-selected group.

The NPI selection method by Coolen and van der Laan [36] was generalised by Coolen-Maturi et al. [44] for right-censored observations. The NPI lower and upper probabilities for the event that a specific group will yield the largest next lifetime were obtained by applying the comparison of multiple groups using the rc-$A_{(n)}$ assumption, which is a generalisation of $A_{(n)}$ for right-censored data [37]. Using their method, experiments may be terminated early to save time and costs. This means that, when the experiment ends, all units in all groups that have not yet failed are right-censored [42].

The NPI selection method for Bernoulli data was developed by Coolen and Coolen-Schrijner [30, 31] to select, from different groups, the group with the highest number of future successes. Instead of focusing solely on a single future observation, they made inferences based on multiple future observations and conducted a pairwise comparison of the groups to obtain the NPI lower and upper probabilities for the event that one group would be more likely than another to have more future successes. Furthermore, a multiple comparison analysis was performed to determine the NPI lower and upper probabilities for the event that one group would have more future successes than all other groups [31]. Additionally, subsets of the groups were studied, providing NPI lower and upper probabilities for two scenarios: one in which a specific subset contains the group with the most future successes, and another in which all groups within a chosen subset will have more future successes than every group which is not in the subset [30]. For example, their method can be applied to screening experiments in which one starts with all treatments available and wishes to continue with only a subset which is likely to contain the best treatments. In clinical trials, the method can also be used to select a subset of treatments that are most likely to be effective.

For multinomial data, Baker and Coolen [12, 13] developed NPI-based methods with multiple future observations for selecting either a single category with the largest lower or upper probability of occurrence or the smallest subset of categories that meets a specified probability requirement. Chapter 3 of this thesis presents NPI selection methods based on multiple future ordinal observations.

1.4 Diagnostic test accuracy

This section introduces the concepts of diagnostic test accuracy and main methods used in the literature to determine a diagnostic threshold; these concepts and methods will be used in Chapters 4 and 5. Diagnostic tests are evaluated according to their ability to discriminate between healthy and diseased individuals in two-group classifications. Assessing the accuracy of diagnostic tests is essential in many application areas, particularly in medicine and healthcare [77]. The diagnostic accuracy of a test refers to its ability to distinguish between different conditions. The diagnostic test may produce a binary outcome, a continuous outcome, or an ordinal outcome. This thesis focuses on diagnostic tests for ordinal data with K categories.

Assume that there is a threshold $k \in \{1, \ldots, K\}$ such that a test result in categories $\{C_{k+1}, \ldots, C_K\}$ indicates the presence of the disease, called a positive test result, while a test result in categories $\{C_1, \ldots, C_k\}$ indicates the absence of the disease, called a negative test result [91, 96]. A main goal for statistical inference in this scenario is the study of appropriate choice for the value k, referred to as the 'optimal threshold' k'. Let T_i^j with $j = 0, 1$ and $i = 1, \ldots, n^j$, denote the test result for individuals in the healthy group and the disease group, respectively. Let n_k^0 and n_k^1 be the number of observations in the healthy and disease groups, respectively, in category C_k, for $k = 1, \ldots, K$. Let n^0 and n^1 be the total number of observations in the healthy and disease groups, respectively, so $\sum_{k=1}^{K} n_k^0 = n^0$ and $\sum_{k=1}^{K} n_k^1 = n^1$. In a diagnostic test, sensitivity (*sens*) is defined as the probability of a positive test result for an individual who has the disease. This is also referred to as the true positive fraction (*TPF*). Specificity (*spec*) is the probability of the test result being negative given the absence of disease. The term false positive fraction (*FPF*) is defined as the probability of a positive test result for an individual who does not have the disease, so *FPF* = 1 - *spec*. Let random quantity T^1 denote the test result for an individual of the disease group and T^0 the test result for an individual of the non-disease group, then for threshold k, $TPF(k) = P(T^1 \in \{C_{k+1}, \ldots, C_K\})$ and $FPF(k) = P(T^0 \in \{C_{k+1}, \ldots, C_K\})$.

Diagnostic test accuracy is described and compared using a popular tool called the receiver operating characteristic (ROC) curve. Diagnostic tests can be considered perfect or ideal when they are able to completely distinguish between healthy and diseased individuals, such that $FPF(k_*) = 0$ and $TPF(k_*) = 1$, at a specific threshold k_* [77]. However, if $FPF(k)$ equals $TPF(k)$, for all $k \in \{1, \ldots, K\}$, the diagnostic test cannot distinguish between healthy and diseased individuals. The accuracy of a diagnostic test may be represented in many cases by a single numerical value or summary [77]. A useful summary is the area under the ROC curve, AUC, which can be used to compare two or more ROC curves. The AUC is a measure of the diagnostic test's overall performance and it has been widely studied in the literature [21, 77, 96].

To define a diagnostic test completely and analyze its quality, it is essential to determine an appropriate threshold. This ensures that the test can effectively discriminate between individuals with and without the disease. Although the AUC is a popular measure of the overall performance of a diagnostic test, it cannot be used to determine the optimal threshold. Methods for selecting the optimal threshold based on ROC analyses have been introduced in the literature. These methods include maximizing the Youden index [51, 95] and the Liu index [64]. The Youden index (YI) is defined as the sum of sensitivity and specificity minus one and is one of the most widely used measures of diagnostic accuracy [51, 82, 87]. The YI is given by

$$\mathrm{YI}(k) = \mathit{sens}(k) + \mathit{spec}(k) - 1 \tag{1.5}$$

The optimal threshold k', based on Youden's index, is defined as the value of k which maximises YI(k) [54, 82]. The empirical estimate of the Youden index (EYI) for ordinal data is

$$\mathrm{EYI}(k) = \left(\frac{1}{n^0}\sum_{i=1}^{k} n_i^0\right) + \left(\frac{1}{n^1}\sum_{i=k+1}^{K} n_i^1\right) - 1 \tag{1.6}$$

where n_i^0 and n_i^1 are the number of observations in the healthy and disease groups, respectively, in category C_i, for $i = 1, \ldots, K$. Additionally, n^0 and n^1 represent the total number of observations in the healthy and disease groups, respectively.

This index has a number of desirable features, as indicated by Youden [95]. For instance, the YI value ranges from 0 to 1. A value of 0 indicates a completely ineffective test, meaning no discriminatory ability between individuals with and without the condition, while a value of 1 indicates a perfect or ideal diagnostic test. This index can not only be used to evaluate the accuracy of a single diagnostic test, but also to compare one diagnostic test with another.

There are many applications of the YI in the medical sciences. For example, Demir et al. [46] applied the YI to determine the most reliable discrimination index for distinguishing between thalassemia trait and iron deficiency anemia. Schisterman et al. [83] conducted an analysis of the coronary calcium score, a marker of atherosclerosis, using the YI. Based on the YI, Aoki et al. [8] identified the optimal threshold level of serum pepsinogens for the detection of gastric cancer. Pekkanen and Pearce [76] computed the YI to distinguish asthma from non-asthma using bronchial hyperresponsiveness and symptom questionnaires. Moreover, YI can be used to make comparisons between different diagnostic testing procedures with regard to their accuracy. For example, Yerli et al. [94] utilized YI to compare two methods for diagnosing common parotid tumors. Similarly, Hawass [58] conducted a comparison of diagnostic tests applying YI, assessing the sensitivities and specificities of two diagnostic procedures within the same patient group. Additionally, YI has been applied to ordinal data to assess disease severity. For instance, a study by Mintoff et al. [68] evaluated the relationship between serum immunoglobulin G (IgG) levels and the severity of Hidradenitis Suppurativa, using YI to determine optimal threshold values for IgG levels, effectively distinguishing between mild, moderate, and severe stages of the disease.

While the Youden index is popular in applications and useful for identifying the optimal diagnostic threshold, it can sometimes yield a threshold with high sensitivity or specificity, leading to unbalanced classification rates [62]. When an illness is highly infectious or a condition is severe, sensitivity may be emphasized over specificity. For example, when assessing the clinical utility of prognostic biomarkers in cancer, the focus will be on developing diagnostic tests with high sensitivity [18]. The specificity of a test may, on the other hand, be emphasized in the event that a subsequent diagnostic test is risky or expensive [64]. The YI aims to maximise the sum of sensitivity and specificity, but this may not always result in a suitable threshold selection when well-balanced sensitivity and specificity is preferred.

Liu [64] proposed the so-called maximum area method, based on the AUC, to determine the optimal threshold that maximises the Liu index. Throughout this thesis, this method will be referred to as the Liu index method (LI). This approach selects the optimal threshold by considering the product of sensitivity and specificity as the objective function for threshold selection, which may result in a more balanced classification. Formally, the LI is defined as

$$\mathrm{LI}(k) = spec(k) \times sens(k) \tag{1.7}$$

The optimal threshold k', based on Liu's index, is defined as the value of k which maximises $\mathrm{LI}(k)$. The empirical estimator for the Liu index (ELI) method for ordinal data is given by

$$\mathrm{ELI}(k) = \left(\frac{1}{n^0} \sum_{i=1}^{k} n_i^0 \right) \times \left(\frac{1}{n^1} \sum_{i=k+1}^{K} n_i^1 \right) \tag{1.8}$$

where n_i^0 and n_i^1 are the number of observations in the healthy and disease groups, respectively, in category C_i, for $i = 1, \ldots, K$. Additionally, n^0 and n^1 represent the total number of observations in the healthy and disease groups, respectively.

According to Liu [64], the proposed approach demonstrated its relevance based on real-world data from a study of arsenic-induced skin lesions. In particular, the method identified individuals at risk for arsenic-induced skin lesions by determining a threshold for blood arsenic levels. His study examined a sample of individuals with and without arsenic-induced skin lesions. By applying the proposed method to the data, a threshold for blood arsenic levels that maximised the LI was selected. To identify individuals at risk of arsenic-induced skin lesions, the resulting threshold was used as a warning threshold. His study showed the value of the proposed method in determining an optimal threshold for blood arsenic levels, thus providing a practical approach for identifying individuals at risk of arsenic-induced skin lesions. In Chapter 4, the proposed NPI-based methods for two-group classification will be compared with the EYI and ELI methods, as given by Equations (1.6) and (1.8).

There are alternative methods for selecting the optimal threshold based on the ROC available in the literature. For instance, Unal [87] proposed a method known as the index of union. This method begins by computing the AUC value and then uses the coordinates of the ROC curve to find a threshold that has specificity and sensitivity values equal to or very close to the AUC value. There is also a method called the closest-to-(0,1) for determining the optimal threshold [78]. It is also called in the literature the northwest corner or the closest-to-perfection method. The objective of this method is to determine the point on the ROC curve that has the shortest distance to the point (0,1) on the graph. A comparison of optimal thresholds selected by the closest-to-(0,1) method and the YI method has been presented by Perkins and Schisterman [78]. In terms of the probability of a correct classification rate, they recommend using the YI since it provides clear clinical meaning. However, there has been little discussion of the closest-to-(0,1) method in the statistical literature compared to the YI method. In this thesis, we focus on the YI and LI, as they have attracted a lot of attention from researchers over the past decade. These methods are well-validated in the literature for threshold selection.

Threshold selection methods have been extended to three-group settings where two thresholds are required to classify individuals. The Youden index was extended by Nakas et al. [72] to a generalised Youden index for three groups. As an extension of the LI method, Attwood et al. [9] proposed the maximum volume method. In Section 5.3, these extended methods are discussed and in Section 5.8 they are compared to the NPI methods for three-group classification.

Classical methods for selecting the optimal threshold of a diagnostic test typically focus on the estimation of the optimal threshold rather than on prediction [4, 40]. Applying diagnostic tests on future patients is the end goal of studying their accuracy. Therefore, a predictive inference method is of interest. Recently, Alabdulhadi [4] and Coolen-Maturi et al. [40] introduced the NPI approach for selecting the optimal threshold(s) for two- and three-group classification problems based on tests that yield real-valued results for a given number of future observations from each group. Test results may take values within a finite number of categories which can be ordered or not; if ordered the test outcome is ordinal. The optimal diagnostic threshold will be selected such that the categories on one side of the threshold indicate disease, and the categories on the other side indicate non-disease. In Chapters 4 and 5, the NPI method for selecting the optional threshold(s) for two- and three-group classification problems is introduced for diagnostic tests that yield ordinal results.

1.5 Outline of the thesis

This thesis is organised as follows. In Chapter 2, the NPI lower and upper probabilities are derived for several events of interest involving multiple future ordinal observations. The results in Chapter 2 have been presented at the 15th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics), King's College London, 17-19 December 2022.

The use of the derived NPI lower and upper probabilities for some inferential problems is presented in Chapter 3 with a focus on both category selection and the selection of subsets of categories. Pairwise comparison for the future observations from two independent groups is also presented.

Chapter 4 presents NPI methods considering multiple future individuals to select the optimal diagnostic test threshold for two-group classification with ordinal outcomes. The NPI method related to the two-group Youden index is introduced. The results in Chapter 4 have been presented at the International Conference of the Royal Statistical Society (RSS) in Harrogate, United Kingdom, held in September 2023. This chapter has also been presented at the Durham Maths Postgraduate Research Day in Durham, United Kingdom, held in May 2024.

Chapter 5 extends the NPI-based methods to three-group classification problems. Chapter 6 summarises the conclusions of this thesis and discusses some related topics for future research. The analyses and calculations in this thesis have been done using `R` [80], with the `ggplot2` package [92] for data visualization.

Chapter 2

NPI for ordinal data

2.1 Introduction

In many applications, categories have a natural order, such as levels of disease severity. This type of data is referred to as ordinal data. This chapter presents NPI for ordinal data with multiple future observations, using an assumed underlying latent variable representation. The categories are represented by intervals on the real-line, reflecting their natural order and allowing the application of Hill's assumption [60].

NPI lower and upper probabilities are presented for several events involving multiple future ordinal observations. The chapter begins by introducing the method for two future observations, providing a foundation for extending the approach to multiple future observations. The generalisation to multiple future observations is then presented. The events considered in this chapter have been chosen to ensure that the methodology can be applied to a wide range of practical scenarios in ordinal data analysis, enabling a variety of statistical inferences, several of which are presented in Chapters 3, 4, and 5.

The overview of NPI for ordinal data, provided in Section 1.2, was restricted to a single future observation. This chapter extends NPI for ordinal data to $m \geq 1$ future observations, denoted by X_{n+l} for $l = 1, \ldots, m$. It is important to note that the use of indices $n + l$ does not imply that the X_{n+l} are ordered in any particular way. The data and future observations are linked via the assumed latent variable representation, where observations on the real-line fall into intervals representing the categories, and via Hill's assumptions $A_{(n)}, A_{(n+1)}, \ldots, A_{(n+m-1)}$ [60]. These are jointly referred to as 'the $A_{(\cdot)}$ assumptions', as outlined in Section 1.1, which imply that all possible orderings of n data observations and m future observations are equally likely [35], where the m future observations are not distinguished or ordered in any way. In NPI, the lower probability for an event of interest is the proportion of the orderings for which the event must hold, and the corresponding upper probability is the proportion of the orderings for which the event can hold [10, 27].

This chapter is structured as follows. Section 2.2 presents the NPI lower and upper probabilities for events involving two future ordinal observations. Section 2.3 generalises the method presented in Section 2.2 to $m \geq 2$ future observations. Section 2.4 extends the focus to more events of interest involving multiple future ordinal observations using a path counting technique. Finally, Section 2.5 provides some concluding remarks.

2.2 NPI for two future ordinal observations

This section presents NPI lower and upper probabilities for two different events of interest, involving two future observations. Two future observations are considered first to gain a clear understanding of the methodology of counting the orderings of the future observations before it is extended to more complex scenarios with multiple observations. Furthermore, the method presented in this section will be generalised to $m \geq 2$ future observations in Sections 2.3 and 2.4, applications of these results are presented in later chapters of this thesis.

The first event considered is that one of the two future observations is in a specific category, while the other observation is in any of the remaining categories. Consider ordinal data with $K \geq 2$ categories, denoted by $C_1, \ldots, C_K$, where the ordering between them is indicated by the notation $C_1 < C_2 < \ldots < C_K$. Let the number of observations in category C_k be denoted by n_k for $k = 1, \ldots, K$, and let n be the total number of the observations, so $\sum_{k=1}^{K} n_k = n$.

Using the latent variable representation, explained in Section 1.2, the category C_k is represented by the interval IC_k, for $k = 1, \ldots, K$, where the K ordered intervals

$IC_1, \ldots, IC_K$ form a partition of the real-line. Each interval IC has neighbouring intervals IC_{k-1} to its left and IC_{k+1} to its right on the real-line (or only one of these neighbours if $k = 1$ or $k = K$) [32, 48]. In the latent variable representation, the n observations are assumed to be represented by $y_1 < \ldots < y_n$, of which n_k are in IC_k, these are also denoted by y_i^k for $i = 1, \ldots, n_k$. As presented in Section 1.2, there are $I_j = (y_{j-1}, y_j)$, for $j = 1, \ldots, n+1$. Although the values y_j are unknown, since they only exist in the latent variable representation, the number of y_j in each interval IC_k is known, therefore, the number of I_j that must be or can be in each IC_k is also known. Let the random quantity M_k represent the number of future observations in category C_k, for $k = 1, \ldots, K$, so the event considered is $M_k = 1$, with $m = 2$.

Generally in NPI, the lower probability for an event of interest involving m future observations is derived by counting all the orderings of the m future observations among the past observations for which the event of interest must hold, and the corresponding NPI upper probability is derived by counting all the orderings for which the event can hold [4]. As a first step, it is important to note that the NPI lower and upper probabilities for the event that one of the two future observations is in a specific category and the other observation is in any of the remaining categories require presenting two different cases related to the value of k, one involves the first or last categories ($k = 1$ or $k = K$), and the other involves the middle categories ($2 \leq k \leq K - 1$). For ordinal data with $n_k \geq 1$ for all k, each IC_k is represented by intervals I_j on the real-line, where here the range of j is restricted to $j = 1, \ldots, n_k + 1$, corresponding to the intervals within IC_k. Each IC_k has an interval $I_1 = (y_{n_{k-1}}^{k-1}, y_1^k)$ to its left and an interval $I_{n_k+1} = (y_{n_k}^k, y_1^{k+1})$ to its right, where future observations in these intervals may or may not be assigned to IC_k. Of course, only one interval can be assigned to IC_1 or IC_K. Categories with $n_k = 0$ do not have any I_j intervals assigned. These left and right intervals are referred to as boundary intervals. The ordinal data structure is shown in Figure 2.1, with shaded intervals representing these boundary intervals. This highlights a distinction, as the overall count of boundary intervals for IC_k with $2 \leq k \leq K - 1$ differs from that of IC_1 and IC_K, which have only one boundary interval each. Consequently, for the event $M_k = 1$, it is necessary to consider both of these cases separately, one involving M_1 and M_K, and the other involving M_k for $2 \leq k \leq K-1$.

Next, the NPI lower probability for the event $M_k = 1$ is introduced, first for the cases $k = 1$ and $k = K$, followed by the case $2 \leq k \leq K-1$. To derive the NPI lower probability for the event $M_k = 1$, the orderings of the next future observation among n_k data observations, in the latent representation, are counted, and then multiplied by the total number of orderings for the other future observation for which the event $M_k = 1$ must hold. This product then is divided by the total number of orderings, which is $\binom{n+m}{m}$.

The case $\boldsymbol{k = 1}$ or $\boldsymbol{k = K}$

First, consider the cases $k = 1$ and $k = K$. Figure 2.1 illustrates the right boundary interval for IC_1 and the left boundary interval for IC_K. The NPI lower probability is derived by counting all the orderings of the future observations for which the event $M_k = 1$ must hold. This involves counting the orderings of future observations within the past data points of IC_k excluding the data point that generates the boundary interval for IC_1 or IC_K, as future observations in these intervals cannot lead to the event $M_k = 1$. Therefore, if $k = 1$ or $k = K$, the number of past data points that are involved in the ordering of future data is $n_k - 1$, leading to n_k intervals I_j that must be assigned to IC_k. Similarly, the total number of the remaining past data points that are involved in the ordering is $n - n_k - 1$, while the remaining I_j intervals, excluding those assigned to IC_k for $k = 1$ or $k = K$, is $n - n_k$. To derive the NPI lower probability, the orderings for the next future observation within the $n_k - 1$ past data points are first counted. This count is then multiplied by the total number of orderings for the other future observation within the $n - n_k - 1$ remaining past data points, for which the event $M_k = 1$ must hold. Thus, there are $(n_k - 1 + 1)(n - n_k - 1 + 1)$ such orderings. This product is then divided by $\binom{n+m}{m}$.

The case $\boldsymbol{2 \leq k \leq K - 1}$

If $2 \leq k \leq K-1$ and $n_k > 1$, there are two data points that generate the two boundary intervals on either side of IC_k, the left and right intervals, as illustrated in Figure 2.1. Therefore, for $2 \leq k \leq K-1$, the number of past data points involved in the ordering of future data is $n_k - 2$. By treating the two data points

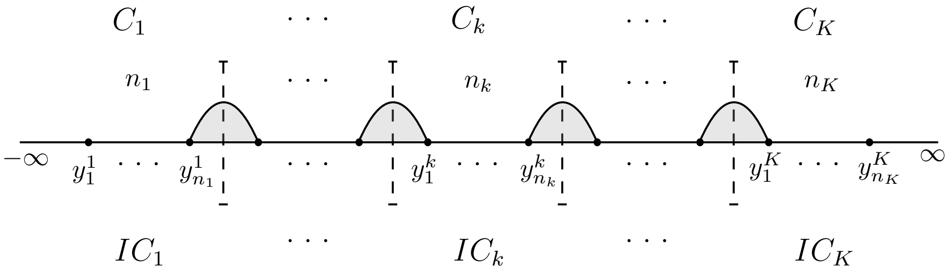


Figure 2.1: Data structure with the boundary intervals

on either side of IC_k as a single data point, the total number of the remaining past data points involved in the ordering is $n - n_k - 1$. Consequently, the total number of different orderings that must lead to the event $M_k = 1$, if $n_k > 1$, is $(n_k - 2 + 1) \times (n - n_k - 1 + 1)$. The NPI lower probability for the event $M_k = 1$ with $m = 2$ is

$$\underline{P}(M_k = 1) = \begin{cases} \binom{n+2}{2}^{-1} n_k(n - n_k) & \text{if } k = 1 \text{ or } k = K \\ \binom{n+2}{2}^{-1} (n_k - 1)(n - n_k) & \text{if } 2 \leq k \leq K - 1 \end{cases} \tag{2.1}$$

In the case $n_k = 0$ for any k, $\underline{P}(M_k = 1) = 0$. Similarly, if $n_k = 1$ for all $2 \leq k \leq K - 1$, $\underline{P}(M_k = 1) = 0$, because there is no I_j interval that can be assigned to IC_k.

The corresponding NPI upper probability for the event $M_k = 1$ is derived by counting all the different orderings of the two future observations among the n data observations, in the latent representation, that can lead to the event $M_k = 1$. This involves counting the orderings where the future observations can be in the boundary intervals of IC_k.

The case $2 \leq k \leq K - 1$

For the case where $2 \leq k \leq K - 1$, three possible scenarios are considered. Firstly, the next future observation could be in any of the I_j intervals of IC_k, including the boundary intervals, resulting in $n_k + 1$ orderings. The ordering of the other future observation must then be within the remaining past data points, treating the two data points on either side of IC_k as a single data point to exclude the boundary intervals, leaving $n - n_k - 1$ data points. Thus, there are $(n - n_k - 1) + 1$ orderings. Consequently, the total number of orderings for this scenario is $(n_k + 1)(n - n_k)$. Secondly, both future observations could be in IC_k, where one of them must be in a boundary interval. Since there are 2 boundary intervals for the middle categories, there are $2(n_k - 2 + 1)$ such orderings. Thirdly, both future observations could be in the two boundary intervals, so there are 3 such orderings, as the two future observations could both be in the right or the left boundary intervals, or there could be one in each. Adding all these together, the number of orderings of the two future observations for which it is possible that precisely one is in IC_k is equal to $(n_k + 1)(n - n_k) + 2(n_k - 1) + 3$.

The case $k = 1$ or $k = K$

For the case where $k = 1$ or $k = K$, the process for deriving the NPI upper probability is similar but slightly different due to the presence of only one boundary interval. The next future observation could be in any of the n_k intervals of IC_k, including the boundary interval, resulting in $n_k + 1$ orderings. The other future observation must be in one of the remaining I_j intervals, leaving $n - n_k$ possible orderings. Thus, the total number of orderings for this scenario is $(n_k + 1)(n - n_k)$. Additionally, both future observations could be in IC_k, with one in the boundary interval, leading to $1(n_k - 1 + 1) = n_k$ such orderings, since there is only one boundary interval. Finally, there is 1 ordering where both future observations are in the only boundary interval. Adding all these together, the total number of orderings of the two future observations for which it is possible that precisely one future observation is in IC_k is $(n_k + 1)(n - n_k) + n_k + 1$. So, the NPI upper probability for $M_k = 1$ is

$$\overline{P}(M_k = 1) = \begin{cases} \binom{n+2}{2}^{-1} [(n_k + 1)(n - n_k) + n_k + 1] & \text{if } k = 1 \text{ or } k = K \\ \binom{n+2}{2}^{-1} [(n_k + 1)(n - n_k) + 2(n_k - 1) + 3] & \text{if } 2 \leq k \leq K - 1 \end{cases} \tag{2.2}$$

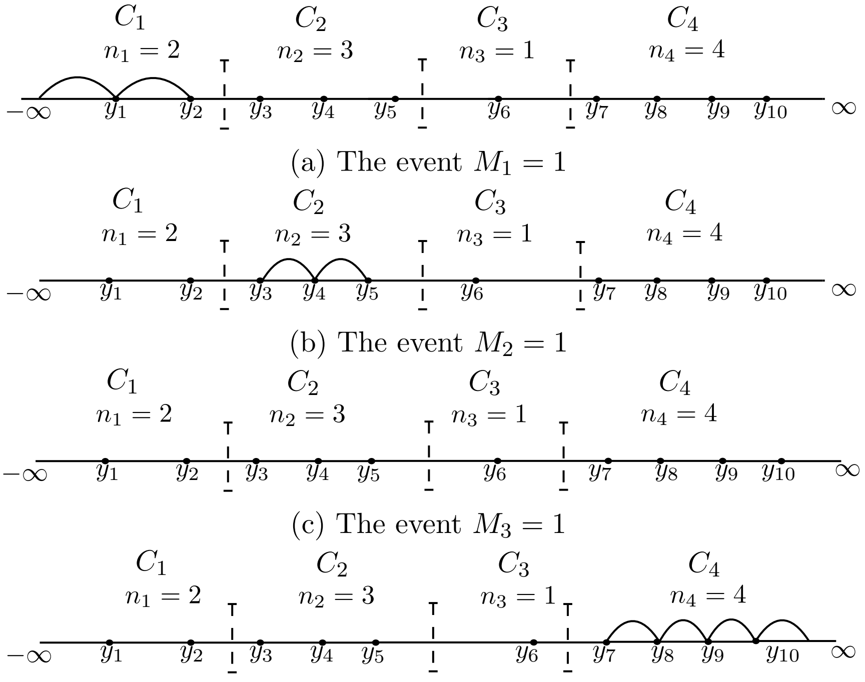
The NPI lower and upper probabilities, given in Equations (2.1) and (2.2), for the event $M_k = 1$ are illustrated in the following example. Note that the examples presented in this chapter aim only to illustrate the application of NPI lower and upper probabilities for each event. Inferences and more detailed scenarios involving these events, along with examples with data from the literature, will be presented and discussed in later chapters of the thesis.

Example 2.1. Consider four ordered categories, $C_1 < C_2 < C_3 < C_4$, and $m = 2$ future observations, and $n = 10$ data observations as follows, $n_1 = 2$, $n_2 = 3$, $n_3 = 1$ and $n_4 = 4$. Figure 2.2 illustrates the ordinal data representation corresponding to the NPI lower probabilities for the events $M_k = 1$ for $k = 1, 2, 3, 4$. Figure 2.3 shows the data representation corresponding to the NPI upper probabilities for the same events . In Figures 2.2 and 2.3, the non-shaded intervals represent the intervals that must be assigned to IC_k, while the shaded intervals in Figure 2.3 represent the boundary intervals.

It should be noted that in the case where $n_3 = 1$, no I_j must be assigned to IC_3, as shown in Figure 2.2, resulting in $\underline{P}(M_3 = 1) = 0$. Category C_1 is considered first. Equations (2.1) and (2.2) are applied with $k = 1$ to derive the NPI lower and upper probabilities for the event $M_1 = 1$, using the values $n = 10$, $n_1 = 2$, and $m = 2$. For the event where precisely one of the two future observations is in C_1, the NPI lower and upper probabilities are 0.2424 and 0.4091 respectively.

The NPI lower and upper probabilities for the other categories are presented in Table 2.1. The results show how these NPI lower and upper probabilities vary based on the number of observations in each category.

To further illustrate how changes in the data affect the NPI lower and upper probabilities, consider the following additional scenarios. First, keep the total number of observations at $n = 10$, but adjust the values of n_k as $n_1 = 1$, $n_2 = 4$, $n_3 = 3$, and $n_4 = 2$. Using Equations (2.1) and (2.2), the corresponding NPI lower and upper probabilities for the events $M_k = 1$ for $k = 1, 2, 3, 4$ in this scenario are presented in Table 2.2. Next, vary the total number of observations to $n = 12$ while using the following values, $n_1 = 2$, $n_2 = 4$, $n_3 = 3$, and $n_4 = 3$. The corresponding NPI lower and upper probabilities are given in Table 2.3.



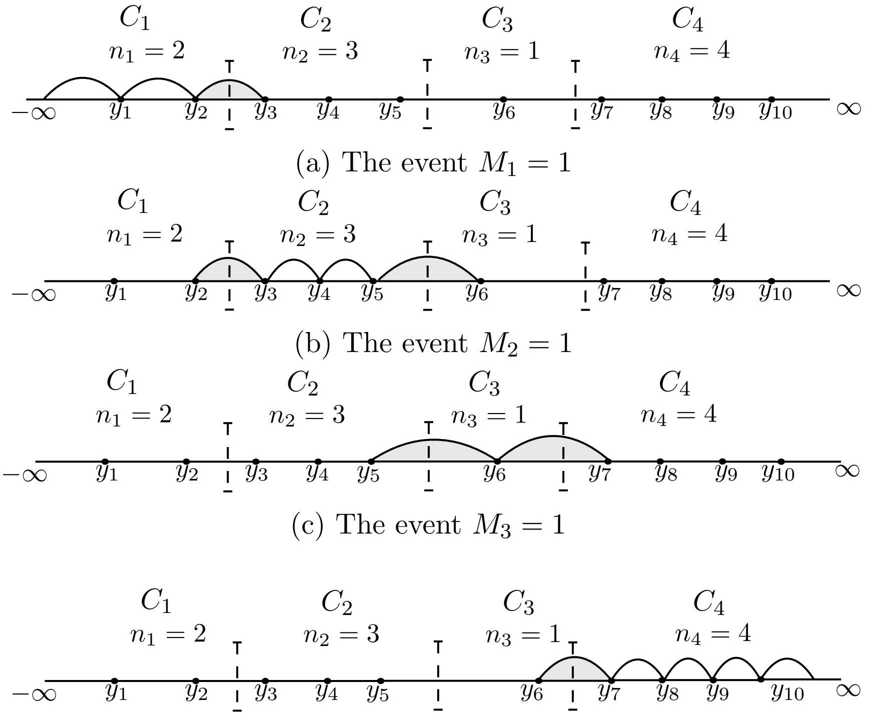
(a) The event $M_1 = 1$

(b) The event $M_2 = 1$

(c) The event $M_3 = 1$

(d) The event $M_4 = 1$

Figure 2.2: Ordinal data representation corresponding to the NPI lower probability for the event $M_k = 1$ for $k = 1, 2, 3, 4$



(a) The event $M_1 = 1$

(b) The event $M_2 = 1$

(c) The event $M_3 = 1$

(d) The event $M_4 = 1$

Figure 2.3: Ordinal data representation corresponding to the NPI upper probability for the event $M_k = 1$ for $k = 1, 2, 3, 4$

k	1	2	3	4
$\underline{P}(M_k = 1)$	0.2424	0.2121	0	0.3636
$\overline{P}(M_k = 1)$	0.4091	0.5303	0.3182	0.5303

Table 2.1: NPI lower and upper probabilities for the event $M_k = 1$ for $k = 1, 2, 3, 4$ with $n = 10$

k	1	2	3	4
$\underline{P}(M_k = 1)$	0.1364	0.2727	0.2121	0.2424
$\overline{P}(M_k = 1)$	0.3030	0.5909	0.5303	0.4091

Table 2.2: NPI lower and upper probabilities for the event $M_k = 1$ for $k = 1, 2, 3, 4$ with $n = 10$ and modified n_k

k	1	2	3	4
$\underline{P}(M_k = 1)$	0.2198	0.2637	0.1978	0.2967
$\overline{P}(M_k = 1)$	0.3626	0.5385	0.4725	0.4396

Table 2.3: NPI lower and upper probabilities for the event $M_k = 1$ for $k = 1, 2, 3, 4$ with $n = 12$

Comparing Tables 2.1, 2.2, and 2.3 illustrates how the NPI lower and upper probabilities change when n_k values are modified while keeping $n = 10$, and how they vary when n is increased. $\diamond$

The second event considered is that both future observations are in C_k, so the event is $M_k = 2$. To begin, note that the NPI lower and upper probabilities for the event $M_k = 2$ require considering the same two cases related to the value of k as for the first event.

The case $k = 1$ or $k = K$

If $k = 1$ or $k = K$, the NPI lower probability is derived by counting all the orderings of the future observations for which the event $M_k = 2$ must hold. This involves counting the orderings of both future observations within the past data points of IC_k excluding the data point that generates the boundary interval for IC_1 or IC_K, as future observations in these intervals cannot lead to the event $M_k = 2$. Therefore, if $k = 1$ or $k = K$, the number of past data points involved in the ordering of future data is $n_k - 1$. If $n_k = 0$, no interval I_j is assigned to IC_k. For $k = 1$ or $k = K$ with

$n_k \geq 1$, the total number of different orderings that must lead to the event $M_k = 2$ is equal to $\binom{n_k-1+2}{2}$. This is then divided by $\binom{n+m}{m}$ to obtain the NPI lower probability for the cases $k = 1$ and $k = K$.

The case $2 \leq k \leq K - 1$

For the case $2 \leq k \leq K - 1$ with $n_k > 1$, the total number of different orderings of the two future observations within each IC_k which must lead to $M_k = 2$ is equal to $\binom{n_k-2+2}{2}$, since there are two boundary intervals that are next to IC_k, as shown in Figure 2.1. If $n_k \leq 1$, no interval is assigned to IC_k. The NPI lower probability for the event $M_k = 2$ is

$$\underline{P}(M_k = 2) = \begin{cases} \binom{n+2}{2}^{-1}\binom{n_k+1}{2} & \text{if } k = 1 \text{ or } k = K \\ \binom{n+2}{2}^{-1}\binom{n_k}{2} & \text{if } 2 \leq k \leq K-1 \end{cases} \tag{2.3}$$

The corresponding NPI upper probability for the event $M_k = 2$ is derived by counting all the different orderings of the two future observations among the n data observations, in the latent representation, that can lead to the event $M_k = 2$. This involves counting the orderings where the intervals I_j can be assigned to IC_k, including the boundary intervals presented in Figure 2.1. If $k = 1$ or $k = K$, the number of intervals I_j that can be assigned to IC_k is equal to n_k. Similarly, the total number of the I_j intervals for the case $2 \leq k \leq K - 1$ is equal to n_k. Therefore, for $1 \leq k \leq K$, there are $\binom{n_k+2}{2}$ orderings such that both future observations are in IC_k. So, for $n_k \geq 0$, the NPI upper probability for the event $M_k = 2$ is

$$\overline{P}(M_k = 2) = \binom{n+2}{2}^{-1}\binom{n_k+2}{2} \qquad \text{for } 1 \leq k \leq K \tag{2.4}$$

The NPI lower and upper probabilities for the event $M_k = 2$, given in Equations (2.3) and (2.4), are illustrated in the following example.

Example 2.2. As in Example 2.1, consider four ordered categories, $C_1 < C_2 < C_3 < C_4$, $m = 2$, and $n = 10$ data observations: $n_1 = 2$, $n_2 = 3$, $n_3 = 1$, and $n_4 = 4$. The NPI lower and upper probabilities for the events $M_k = 2$, for $k = 1, 2, 3, 4$, are

k	1	2	3	4
$\underline{P}(M_k = 2)$	0.0455	0.0455	0	0.1515
$\overline{P}(M_k = 2)$	0.0909	0.1515	0.0455	0.2273

Table 2.4: NPI lower and upper probabilities for the event $M_k = 2$ for $k = 1, 2, 3, 4$ with $n = 10$

k	1	2	3	4
$\underline{P}(M_k = 2)$	0.0152	0.0909	0.0152	0.0909
$\overline{P}(M_k = 2)$	0.0455	0.2273	0.0909	0.1515

Table 2.5: NPI lower and upper probabilities for the event $M_k = 2$ for $k = 1, 2, 3, 4$ with $n = 10$ and modified n_k

k	1	2	3	4
$\underline{P}(M_k = 2)$	0.0659	0.0110	0.0659	0.0659
$\overline{P}(M_k = 2)$	0.1099	0.0659	0.1648	0.1099

Table 2.6: NPI lower and upper probabilities for the event $M_k = 2$ for $k = 1, 2, 3, 4$ with $n = 12$

presented in Table 2.4. The results show how these NPI lower and upper probabilities vary based on the number of observations in each category. For example, although C_2 has 3 data observations and C_1 has 2 data observations, the NPI lower probabilities are the same for both categories. This occurs because IC_1 has only one boundary interval, resulting in $\binom{n_1-1+2}{2}$ possible orderings for the two future observations within IC_1, while IC_2 has two boundary intervals, so there are $\binom{n_2-2+2}{2}$ orderings. Consequently, both categories yield the same total number of orderings with $n_1 = 2$ and $n_2 = 3$ leading to the same value for the NPI lower probability.

To further illustrate how changes in the data affect the NPI lower and upper probabilities, consider the following additional scenarios. First, keep the total number of observations at $n = 10$, but adjust the values of n_k: $n_1 = 1$, $n_2 = 4$, $n_3 = 2$, and $n_4 = 3$. Using Equations (2.3) and (2.4), the corresponding NPI lower and upper probabilities for the events $M_k = 2$ for $k = 1, 2, 3, 4$ are presented in Table 2.5.

Next, vary the total number of observations to $n = 12$ while using the following values: $n_1 = 3$, $n_2 = 2$, $n_3 = 4$, and $n_4 = 3$. The corresponding NPI lower and upper probabilities are given in Table 2.6.

Comparing Tables 2.4, 2.5, and 2.6 illustrates how the NPI lower and upper probabilities change when n_k values are modified while keeping $n = 10$, and how they vary when n is increased. $\diamond$

This section presented the NPI lower and upper probabilities for two events for $m = 2$. First, the event where one of the two future observations is in a specific category and the other is in any of the remaining categories was considered. The second event presented was that both future ordinal observations are in a specific category. It was emphasized that there are two different cases for deriving the NPI lower probability: $k = 1$ or $k = K$, or $2 \leq k \leq K - 1$. This distinction arises because of the boundary intervals, as future observations in these intervals may or may not be assigned to IC_k. The IC_1 and IC_K have only one boundary interval, while the other IC_k have two boundary intervals. The NPI lower probability is derived by counting all the orderings of the future observations among the n data observations, in the latent representation, where the event must hold, while the NPI upper probability is derived by counting all the orderings where the event can hold. The results obtained in this section extend to derive the NPI lower and upper probabilities for events involving $m \geq 2$ future observations. This will be presented in the next section.

2.3 NPI for multiple future ordinal observations

This section generalises NPI for ordinal data to $m \geq 2$ future observations, considering two different events of interest. For ordinal data, as introduced in Section 2.1, there are K ordered categories, denoted by $C_1 < C_2 < \ldots < C_K$. Let the random quantity M_k represent the number of the m future observations in category C_k, for $k = 1, \ldots, K$. The first event considered is that there is a particular number of future observations in each one of the categories, so the event is $\bigcap\limits_{k=1}^{K} \{M_k = m_k\}$, where $\sum_{k=1}^{K} m_k = m$.

In the latent variable representation, as explained in Section 1.2, the category C_k is assumed to be represented by the interval IC_k for $k = 1, \ldots, K$, and the n observations are assumed to be represented by $y_1 < \ldots < y_n$, of which n_k are in IC_k, these are also denoted by y_i^k for $i = 1, \ldots, n_k$. As explained in Section 2.2, there are $I_j = (y_{j-1}, y_j)$, for $j = 1, \dots, n+1$, and with $n_k \geq 1$ for all k, each IC_k has an interval I_j to its left and an interval I_j to its right where future observations in these intervals may or may not be assigned to IC_k. Of course, only one interval is assigned to IC_1 or IC_K. Categories with $n_k = 0$ do not have any intervals assigned. First, the NPI lower probability for the event $\bigcap\limits_{k=1}^{K} \{M_k = m_k\}$ is derived. Throughout this section, the case involving $k = 1$ and $k = K$, and the case involving $2 \leq k \leq K-1$, are considered separately.

In NPI, as presented in Section 2.2, the lower probability for an event of interest involving m future observations is derived by counting all the different orderings of the m future observations among the n data observations, in the latent representation, for which the event of interest must hold. So, for $k = 1$, this involves counting the orderings of future observations within the I_j that must be assigned to IC_1, excluding the right boundary interval, as shown in Figure 2.1. Similarly, for IC_K with the left boundary interval as illustrated in Figure 2.1. So, if $k = 1$ or $k = K$, the total number of different orderings of m_k within IC_k is equal to $\binom{n_k-1+m_k}{m_k}$. If $2 \leq k \leq K-1$, there are 2 boundary intervals for IC_k, see Figure 2.1. Thus, the total number of different orderings of m_k within IC_k is equal to $\binom{n_k-2+m_k}{m_k}$. Let

$$h_k = \begin{cases} 1 & \text{if } k = 1 \text{ or } k = K \\ 0 & \text{otherwise} \end{cases}$$

Then, the NPI lower probability is

$$\underline{P}\left(\bigcap_{k=1}^{K}\{M_k = m_k\}\right) = \binom{n+m}{m}^{-1} \prod_{k=1}^{K} \binom{n_k - 2 + h_k + m_k}{m_k} \tag{2.5}$$

The corresponding NPI upper probability for the event $\bigcap\limits_{k=1}^{K} \{M_k = m_k\}$ is derived by counting all the different orderings of the m future observations among the n data observations, in the latent representation, for which the event of interest must hold. This involves counting the orderings where the intervals of the data can be assigned to IC_k, including the boundary intervals presented in Figure 2.1. If $k = 1$ or $k = K$, the total number of different orderings that can lead to $M_k = m_k$ is equal to $\binom{n_k+m_k}{m_k}$. If $2 \leq k \leq K-1$, the total number of different orderings that can lead to $M_k = m_k$ is equal to $\binom{n_k+m_k}{m_k}$ as well. For $k = 1, \dots, K$, the NPI upper probability for the event $\bigcap_{k=1}^{K} \{M_k = m_k\}$ is

$$\overline{P}\left(\bigcap_{k=1}^{K}\{M_k = m_k\}\right) = \binom{n+m}{m}^{-1} \prod_{k=1}^{K} \binom{n_k + m_k}{m_k} \tag{2.6}$$

The NPI lower and upper probabilities for the event $\bigcap_{k=1}^{K} \{M_k = m_k\}$, given by Equations (2.5) and (2.6), are illustrated in the following example. Note that, as mentioned in Section 2.2, the examples presented in this chapter aim only to illustrate how to apply the NPI lower and upper probabilities for each event. Inferences and more detailed scenarios involving these events, along with examples with data from the literature, will be presented and discussed in later chapters of the thesis.

Example 2.3. Suppose that there are $K = 3$ ordered categories $C_1 < C_2 < C_3$, with data $(n_1, n_2, n_3) = (2,\ 2,\ 3)$. Consider $m = 3$ future observations with the aim to derive the NPI lower and upper probabilities of the event $\bigcap_{k=1}^{3} \{M_k = 1\}$. Table 2.7 displays the NPI lower and upper probabilities for this event and for all different combinations of m_k where $\sum_{k=1}^{3} m_k = 3$.

The results in Table 2.7 show how these NPI lower and upper probabilities vary based on the number of data observations and the number of m_k future observations in each category. For example, although IC_1 and IC_2 have the same number of data observations, the NPI lower probabilities differ for these categories with the same number of future observations. For example, with $(m_1, m_2, m_3) = (2, 1, 0)$, the NPI lower probability for the event is 0.0250, but when $(m_1, m_2, m_3) = (1, 2, 0)$, the NPI lower probability for the event is 0.0167. This occurs because IC_1 has only one boundary interval, resulting in $\binom{n_1-1+m_1}{m_1}$ possible orderings for the future observations, while IC_2 has two boundary intervals, so there are $\binom{n_2-2+m_2}{m_2}$ orderings. Consequently, the two categories yield different numbers of orderings despite having the same number of data observations.

The results also show how NPI lower or upper probabilities vary based on the number of data observations in each category and how a specific category with more data observations has larger NPI lower or upper probabilities when all future observations are in that specific category. For example, the NPI upper probability is larger for the event with all 3 future observations in C_3, it is 0.1667, compared

(m_1, m_2, m_3)	$\underline{P}\left(\bigcap_{k=1}^{K}\{M_k = m_k\}\right)$	$\overline{P}\left(\bigcap_{k=1}^{K}\{M_k = m_k\}\right)$
$(2, 1, 0)$	0.0250	0.1500
$(2, 0, 1)$	0.0750	0.2000
$(0, 2, 1)$	0.0250	0.2000
$(0, 1, 2)$	0.0500	0.2500
$(1, 2, 0)$	0.0167	0.1500
$(1, 0, 2)$	0.1000	0.2500
$(3, 0, 0)$	0.0333	0.0833
$(0, 3, 0)$	0.0083	0.0833
$(0, 0, 3)$	0.0833	0.1667
$(1, 1, 1)$	0.0500	0.3000

Table 2.7: NPI lower and upper probabilities for all combinations of future observations with $K = 3$ and $m = 3$

to the NPI upper probabilities when either all future observations are in C_1 or C_2. This difference is due to the larger number of the data observations in category C_3, which results in more possible orderings for future observations and larger NPI lower or upper probabilities.

To further explore how the lower and upper NPI probabilities change, consider the scenario where the data is modified to $(n_1, n_2, n_3) = (4, 4, 2)$. In this case, the combination $(m_1, m_2, m_3) = (2, 1, 0)$ yields a larger NPI lower probability equal to 0.1049 and an NPI upper probability equal to 0.2622, reflecting the impact of more observations in the first two categories. Similarly, when all three future observations are in the first category $(3, 0, 0)$, the values increase to 0.0699 and 0.1224, compared to the original scenario with $(n_1, n_2, n_3) = (2, 2, 3)$, where the values were 0.0333 and 0.0833, respectively. This comparison highlights how the number of the data in each category influences the NPI lower and upper probabilities. $\diamond$

The second event of interest is that precisely m_k of the m future observations are in a specific category C_k, while the remaining $m - m_k$ observations are in the remaining categories. Let the random quantity M_k represent the number of future observations belonging to category C_k for $k = 1, \ldots, K$, so the event is $M_k = m_k$. To begin, note that the NPI lower and upper probabilities for the event $M_k = m_k$ require considering the same two cases related to the value of k as in the first event.

For the event $M_k = m_k$, the NPI lower probability is derived first for the cases where $k = 1$ or $k = K$, followed by the case where $2 \leq k \leq K - 1$.

The case $k = 1$ or $k = K$

For the cases $k = 1$ and $k = K$, the NPI lower probability is derived by counting the orderings of the m_k future observations among n_k data observations, in the latent representation, and then multiplied by the total number of orderings for the $m - m_k$ future observation for which the event $M_k = m_k$ must hold. This product then is divided by the total number of orderings, which is $\binom{n+m}{m}$. This involves, as explained in Section 2.2, counting the orderings of the future observations within the past data points of IC_k, excluding the data point that generates the right boundary interval for IC_1 or the left boundary interval for IC_K. As a result, for $k = 1$ or $k = K$, the number of past data points involved in the ordering of future data is $n_k - 1$. Therefore, the number of different orderings of m_k future observations within these past data points is equal to $\binom{n_k-1+m_k}{m_k}$. If $n_k = 0$ for $k = 1$ or $k = K$, then IC_k does not have any past data points assigned to it. Similarly, the total number of remaining past data points, excluding those assigned to that specific IC_k, is equal to $n - n_k - 1$. Thus, the number of orderings for $m - m_k$ future observations among the remaining past data points is $\binom{n-n_k-1+m-m_k}{m-m_k}$. Using Equation (1.2), the NPI lower probability for the event $M_k = m_k$, where $k = 1$ or $k = K$, is

$$\underline{P}(M_k = m_k) = \binom{n+m}{m}^{-1} \binom{n_k - 1 + m_k}{m_k} \binom{n - n_k - 1 + m - m_k}{m - m_k} \tag{2.7}$$

The case $2 \leq k \leq K - 1$

If $2 \leq k \leq K - 1$ with $n_k > 1$, the number of different orderings of the m_k future observations within the past data points of IC_k, excluding the two data points that generate the boundary intervals, is equal to $\binom{n_k-2+m_k}{m_k}$. No I_j intervals must be assigned to IC_k if $n_k \leq 1$. By treating the two data points on either side of IC_k as a single data point, the total number of the remaining data points, excluding those assigned to that specific IC_k is equal to $n - n_k - 1$. So, the number of orderings for $m - m_k$ future observations is equal to $\binom{n-n_k-1+m-m_k}{m-m_k}$. Using Equation (1.2), the

NPI lower probability for the event $M_k = m_k$, where $2 \leq k \leq K-1$, is

$$\underline{P}(M_k = m_k) = \binom{n+m}{m}^{-1} \binom{n_k - 2 + m_k}{m_k} \binom{n - n_k - 1 + m - m_k}{m - m_k} \tag{2.8}$$

The corresponding NPI upper probability is derived by counting all the different orderings of the m future observations among the n data observations, in the latent representation, that can lead to the event $M_k = m_k$. This involves considering the orderings of the future observations within the boundary intervals presented in Figure 2.1. For the event $M_k = m_k$, the NPI upper probability is derived first for the case where $2 \leq k \leq K-1$, followed by the cases where $k = 1$ and $k = K$.

The case $2 \leq k \leq K-1$

For the case where $2 \leq k \leq K-1$, the objective is to maximise the number of orderings for the future observations, ensuring that m_k future observations can be in IC_k. As in the case of the NPI lower probability, all orderings where m_k observations must be in that specific IC_k and $m - m_k$ future observations must be in the remaining IC_k are counted. It was previously shown that there are $\binom{n_k-2+m_k}{m_k}\binom{n-n_k-1+m-m_k}{m-m_k}$ such orderings. However, the two boundary intervals (shaded intervals in Figure 2.1) are now also considered. Any future observations that are in one of the boundary intervals may be counted either as belonging to that specific IC_k or as not belonging. This implies that, to determine the NPI upper probability, orderings with one or more observations in the boundary intervals need to be included. Let W represent the total number of future observations in the boundary intervals, where W ranges from 1 to m. For $W = 1$, there are two possible ways in which W can be in the two boundary intervals, as the observation could be in either the left or the right boundary interval. By similar reasoning, for $W = 2$, there are three possible ways.

Generally, there are $W + 1$ possible ways for each value of W. Let D be an integer representing the number of future observations in the boundary intervals that can be counted as belonging to that specific IC_k such that $D \leq m_k$ and $W - D \leq m - m_k$ where $W - D$ represents the number of future observations in the boundary intervals that can be assigned to IC_k other than that specific IC_k. Then, there may be $m_k - D$ future observations in the I_j intervals of the specific IC_k, which is equal to $n_k - 2$, $(m - m_k) - (W - D)$ future observations in the remaining I_j intervals, which is equal to $n - n_k - 1$, and W observations in the boundary intervals, where D ranges from $W - m - m_k$ to m_k. Therefore, the total number of orderings that can lead to $M_k = m_k$ with one or more observations in the boundary intervals is equal to $\sum_{W=1}^{m} \sum_{D=(W-(m-m_k))^+}^{\min\{m_k,W\}} (W+1)\binom{n_k - 2 + (m_k - D)}{m_k - D}\binom{n - n_k - 1 + (m - m_k) - (W - D)}{m - m_k - (W - D)}$.

The NPI upper probability for $2 \leq k \leq K - 1$ is:

$$\overline{P}(M_k = m_k) = \binom{n+m}{m}^{-1}\Bigg[\binom{n_k - 2 + m_k}{m_k}\binom{n - n_k - 1 + m - m_k}{m - m_k} + \sum_{W=1}^{m} \sum_{D=(W-(m-m_k))^+}^{\min\{m_k,W\}} (W+1)\binom{n_k - 2 + m_k - D}{m_k - D} \times \binom{n - n_k - 1 + m - m_k - (W - D)}{m - m_k - (W - D)}\Bigg] \tag{2.9}$$

The case $k = 1$ or $k = K$

Now, for the cases $k = 1$ and $k = K$, all orderings in which m_k observations must be in IC_k and $m - m_k$ future observations must be in the remaining IC_k intervals are counted. There are $\binom{n_k-1+m_k}{m_k}\binom{n-n_k-1+m-m_k}{m-m_k}$ such orderings, as shown in the case of the NPI lower probability. If $k = 1$ or $k = K$, only one boundary interval needs to be considered. However, for $k = 1$ or $k = K$, the presence of only one boundary interval means that W represents the total number of future observations in the single boundary interval. Since there is only one boundary interval, each observation in W does not have $W + 1$ ways to consider (unlike the two boundary intervals in the $2 \leq k \leq K - 1$ case). Therefore, the term $W + 1$ is not used. The NPI upper probability for $k = 1$ or $k = K$ is

$$\overline{P}(M_k = m_k) = \binom{n+m}{m}^{-1}\Bigg[\binom{n_k - 1 + m_k}{m_k}\binom{n - n_k - 1 + m - m_k}{m - m_k} + \sum_{W=1}^{m} \sum_{D=(W-(m-m_k))^+}^{\min\{m_k,W\}} \binom{n_k - 1 + m_k - D}{m_k - D} \times \binom{n - n_k - 1 + m - m_k - (W - D)}{m - m_k - (W - D)}\Bigg] \tag{2.10}$$

The NPI lower and upper probabilities for the event $M_k = m_k$ are illustrated in the following example.

Example 2.4. Consider $K = 3$ ordered categories, $C_1 < C_2 < C_3$, with the data $(n_1, n_2, n_3) = (2, 3, 4)$. The data representation is presented in Figure 2.4. Consider $m = 3$ future observations with the aim to derive the NPI lower and upper probabilities for the event $M_2 = 2$. First, Equations (2.7) and (2.8) will be applied in order to derive the NPI lower probability with the values $n = 9$, $n_2 = 3$ and $m = 3$. Thus,

$$\underline{P}(M_2 = 2) = \binom{9+3}{3}^{-1} \binom{3-2+2}{2} \binom{9-3-1+3-2}{3-2} = \frac{18}{220} = 0.0818$$

The NPI lower probability for the event $M_2 = 2$ is equal to $\frac{18}{220}$, indicating 18 different orderings for which M_2 must be 2. From Figure 2.4, it is clear that the past data points of IC_2, in which the 2 future observations must be ordered within, excluding the two data points that generate the boundary intervals, is equal to $n_2 - 2 = 1$ data point. There are 3 orderings for the 2 future observations that must be in IC_2. Similarly, the total number of the remaining past data points involved in the ordering, excluding those assigned to IC_2, is equal to $n - n_2 - 1 = 5$. There are 6 orderings for the 1 future observation within these remaining past data points. Multiplying these orderings 3 × 6 leads to a total of 18 orderings.

Equation (2.10) is applied to derive the NPI upper probability for the event $M_2 = 2$ with the values $n = 9$, $n_2 = 3$ and $m = 3$, leading to

$$\begin{aligned}\overline{P}(M_2 = 2) = \binom{9+3}{3}^{-1} \Bigg[& \binom{3}{2}\binom{6}{1} + \sum_{D=0}^{1} (2) \binom{1+(2-D)}{2-D} \binom{6-(1-D)}{1-(1-D)} + \\ & \sum_{D=1}^{2} (3) \binom{1+(2-D)}{2-D} \binom{6-(2-D)}{1-(2-D)} + \\ & \sum_{D=2}^{2} (4) \binom{1+(2-D)}{2-D} \binom{6-(3-D)}{1-(3-D)} \Bigg] \\ = & \frac{1}{220}[18 + 30 + 24 + 4] = \frac{76}{220} = 0.3455.\end{aligned}$$

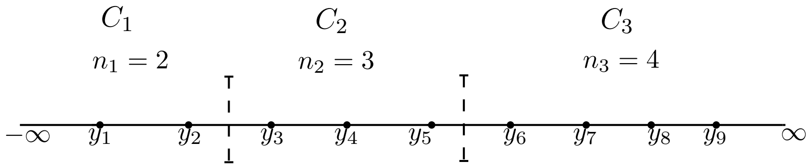


Figure 2.4: Ordinal Data structure for Example 2.4

The corresponding NPI upper probability is equal to $\frac{76}{220}$, indicating that there are 76 different orderings for which the event $M_2 = 2$ is possible. These 76 orderings are presented in the Appendix in Figures A1 to A4, with detailed explanations provided. The NPI lower and upper probabilities for the events $M_1 = 2$ and $M_3 = 2$ can be derived similarly. For $M_1 = 2$, using Equations (2.7) and (2.8), the NPI lower probability is 0.0955 and the NPI upper probability is 0.2182. For $M_3 = 2$, using the same equations, the NPI lower probability is 0.2273 and the NPI upper probability is 0.4091.

$\diamond$

Instead of solely focusing on the precise number of future observations belonging to a specific category, attention is extended to a further event of interest, which is presented in the next section. This event is that at least m_k out of the m future observations are in a specific category. To present this, a path counting technique will be introduced, which can also be used to obtain the NPI lower and upper probabilities for other events of interest.

2.4 NPI for multiple future ordinal observations using path counting

In this section, the event considered is that at least m_k of the m future observations are in a specific category. This event will be used in the next chapter. Let the random quantity M_k represent the number of future observations belonging to category C_k for $k = 1, \ldots, K$, so the event is $M_k \geq m_k$. The NPI lower and upper probabilities will be derived using a path counting technique to avoid double counting of orderings of future observations.

The path counting method was introduced by Aboalkhair [1] for NPI with Bernoulli random quantities. The method uses the $A_{(\cdot)}$ assumptions for inference on m future random quantities given n observations and a latent variable representation with Bernoulli quantities represented by observations on the real line, with a threshold such that successes are on one side and failures on the other side of the threshold. Under the $A_{(\cdot)}$ assumptions, the $\binom{n+m}{n}$ different orderings of these observations, when not distinguishing between the n observed values nor between the m future observations, are all equally likely. For each such ordering, the success-failure threshold can be in any of the intervals of the partition of the real line created by the $n+m$ values of the latent variables.

These outcomes can be visualized as paths on a rectangular lattice grid from $(0,0)$ to (n,m). In this grid, a move to the right corresponds to a data observation, while an upward move represents a future observation. Consequently, a path from $(0,0)$ to (n,m) on this grid consists of n steps to the right and m steps upward. Each path on the lattice directly corresponds to one of the possible orderings of the n data and m future observations. For example, let $n=2$ and $m=2$, possible orderings might include "Data, Data, Future, Future," "Data, Future, Data, Future," "Future, Data, Data, Future," and so on. The corresponding paths of these orderings on the lattice are "Right, Right, Up, Up," "Right, Up, Right, Up," "Up, Right, Right, Up," etc. There are $\binom{n+m}{n}$ such paths, which correspond directly to the number of all possible orderings. This correspondence allows for the application of path counting techniques to derive NPI lower and upper probabilities, focusing specifically on counting these paths. Therefore, the $\binom{n+m}{n}$ different orderings, which are all equally likely, correspond to the $\binom{n+m}{n}$ different right-upwards paths from $(0,0)$ to (n,m), and hence the NPI lower and upper probabilities can also be derived by counting paths.

Extending this approach to ordinal data involves considering the $n \times m$ lattice for $K \geq 2$ ordered categories. For ordinal data, the total number of paths from $(0,0)$ to (n,m), with movements going either one to the right or one upwards, is $\binom{n+m}{m}$. The $A_{(\cdot)}$ assumptions lead to each path being equally likely. Figure 2.5, following the notation in Figure 2.1, illustrates all possible paths from $(0,0)$ to (n,m), with

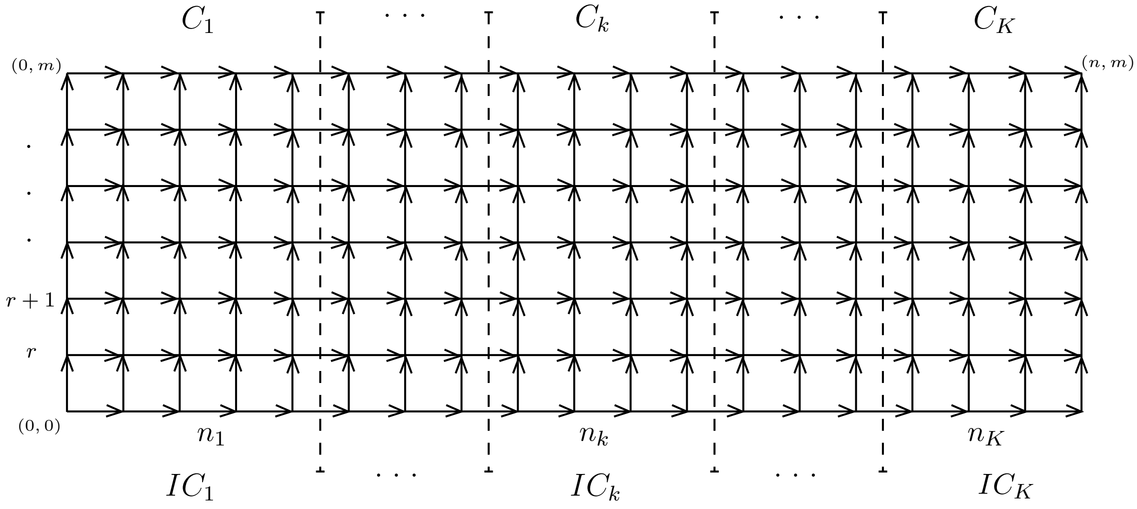


Figure 2.5: Data structure with all possible paths from $(0, 0)$ to (n, m)

movements going either one to the right or one upwards. The NPI lower probability for the event $M_k \geq m_k$ with $k = 1$ or $k = K$ is first considered.

The case $k = 1$ or $k = K$

The derivation of the NPI lower probability involves counting all paths that pass through the two points $(n_k - 1, r)$ and (n_k, r), respectively, where $r \geq m_k$, but not through any point (n_k, r) where $r < m_k$. Any path passing through (n_k, r) with $r < m_k$ would indicate that the number of future observations does not meet the required m_k, meaning the event $M_k \geq m_k$ cannot hold. Counting the paths in this way ensures that the event $M_k \geq m_k$ holds. The number of these paths equals the number of paths from $(0, 0)$ to (n, m) through at least one of $(n_k - 1, r), (n_k - 1, r + 1)$, $(n_k - 1, r + 2), \ldots, (n_k - 1, m)$. Each possible value of r is considered separately in order to avoid counting any path more than once. For a given value of r, there are $\binom{n_k-1+r}{r}$ different paths from $(0, 0)$ to (n_k, r) that must pass through the point $(n_k - 1, r)$. The number of paths from (n_k, r) to (n, m) of the remaining $m - r$ future observations is equal to $\binom{n-n_k+m-r}{m-r}$. Therefore, the NPI lower probability is

$$\underline{P}(M_k \geq m_k) = \binom{n+m}{m}^{-1} \sum_{r=m_k}^{m} \binom{n_k - 1 + r}{r} \binom{n - n_k + m - r}{m - r} \qquad (2.11)$$

The case $2 \le k \le K-1$

For $2 \le k \le K-1$, the derivation of the NPI lower probability involves counting all paths from $(0,0)$ to (n,m) that pass through the points $(n_{1:k-1}, v)$ and (n_k, r), with $m_k \le r \le m$, $0 \le v \le m-r$, and $n_{1:k-1} = \sum_{i=1}^{k-1} n_i$. Initially, all paths from $(0,0)$ to $(n_{1:k-1}, v)$ are counted, where v ranges from 0 to $m-r$. This corresponds to counting all paths in IC_1 to IC_{k-1}. The number of these paths is $\binom{n_{1:k-1}+v}{v}$. Next, paths are counted from $(n_{1:k-1}, v)$ to (n_k, r), corresponding to IC_k with $2 \le k \le K-1$, where r ranges from m_k to m. Within this specific middle IC_k, the paths must follow the constraint of entering and exiting IC_k horizontally. This constraint ensures no upward movements occur on either side of IC_k, ensuring the event $M_k \ge m_k$ holds. For a given value of r, the number of these paths is $\binom{n_k-2+r}{r}$. Finally, all paths from (n_k, r) to (n,m) are counted, corresponding to IC_{k+1} to IC_K. With $n_{k+1:K} = \sum_{l=k+1}^{K} n_l$, the number of these paths is $\binom{n_{k+1:K}+m-r-v}{m-r-v}$. This method of counting ensures that the event $M_k \ge m_k$ must hold. The NPI lower probability for the event $M_k \ge m_k$ with $2 \le k \le K-1$ is

$$\underline{P}(M_k \ge m_k) = \binom{n+m}{m}^{-1} \left[\sum_{r=m_k}^{m} \binom{n_k-2+r}{r} \sum_{v=0}^{m-r} \binom{n_{1:k-1}+v}{v} \times \right. \\ \left. \binom{n_{k+1:K}+m-r-v}{m-r-v} \right] \tag{2.12}$$

The corresponding NPI upper probability for $1 \le k \le K$ can be derived by counting all paths that go through at least one point (n_k, r) with $r \ge m_k$. To avoid that any path is counted more than once, the number of these paths can be computed by counting all paths from $(0,0)$ to (n,m) via (n_k, r), in addition to paths from $(0,0)$ to (n,m) via at least one of $(n_k-1, r+1), (n_k-1, r+2), \ldots, (n_k-1, m)$. This ensures that the event $M_k \ge m_k$ can hold. The NPI upper probability is

$$\overline{P}(M_k \ge m_k) = \binom{n+m}{m}^{-1} \times \left[\binom{n_k+r}{r} \binom{n-n_k+m-r}{m-r} + \right. \\ \left. \sum_{j=r+1}^{m} \binom{n_k-1+j}{j} \binom{n-n_k+m-j}{m-j} \right] \tag{2.13}$$

Example 2.5 illustrates the NPI lower and upper probabilities for the event $M_k \geq m_k$ and provides further explanation of the counting of the paths.

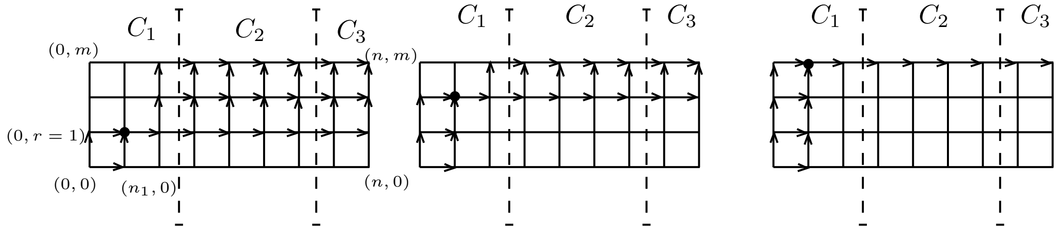
Example 2.5. Consider an ordinal data set with 3 ordered categories, $C_1 < C_2 < C_3$, where the data are $(n_1, n_2, n_3) = (2,\ 4,\ 2)$. Consider $m = 3$ future observations with the aim to derive the NPI lower and upper probabilities for the event that C_1 contains at least one of the three future observations, so $M_1 \geq 1$. Equations (2.11) and (2.13) are applied to derive the NPI lower and upper probabilities, respectively. The NPI lower probability for $M_1 \geq 1$ with the values $n = 8$, $n_1 = 2$ and $m = 3$ is

$$
\begin{aligned}
\underline{P}(M_1 \geq 1) &= \binom{n+m}{m}^{-1} \sum_{r=m_1}^{m} \binom{n_1 - 1 + r}{r} \binom{n - n_1 + m - r}{m - r} \\
&= \binom{8+3}{3}^{-1} \sum_{r=1}^{3} \binom{2 - 1 + r}{r} \binom{8 - 2 + 3 - r}{3 - r} \\
&= \binom{8+3}{3}^{-1} [2 \times 28 + 3 \times 7 + 4 \times 1] = 0.4909.
\end{aligned}
$$

This quantity $\sum_{r=1}^{3} \binom{2-1+r}{r}\binom{8-2+3-r}{3-r}$ represents the number of paths from $(0,0)$ to (n,m) that have to pass through at least one of $(n_1 - 1, r), (n_1 - 1, r+1)$ or $(n_1 - 1, m)$, visualized by the black dots in Figure 2.6. The total number of orderings for the event $M_1 \geq 1$ is $2 \times 28 + 3 \times 7 + 4 \times 1$, where the 2×28, 3×7 and 4×1 paths are shown in Figure 2.6(a), (b), and (c), respectively. The corresponding NPI upper probability is derived by counting all paths that go through at least one point (n_1, r) with $r \geq m_1$, in addition to paths from $(0,0)$ to (n,m) via at least one of $(n_1 - 1, r+1)$ or $(n_1 - 1, m)$. The NPI upper probability for the event $M_1 \geq 1$ is

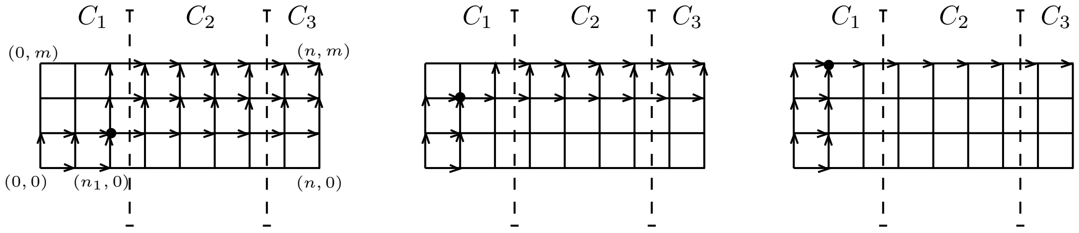
$$
\begin{aligned}
\overline{P}(M_1 \geq 1) &= \binom{8+3}{3}^{-1} \times \Bigg[\binom{2+1}{1} \binom{8 - 2 + (3-1)}{3-1} + \\
&\quad \sum_{j=2}^{3} \binom{2 - 1 + j}{j} \binom{8 - 2 + (3-j)}{m - j} \Bigg] \\
&= \binom{8+3}{3}^{-1} [3 \times 28 + 3 \times 7 + 4 \times 1] = 0.6606.
\end{aligned}
$$

The total number of orderings for the event $M_1 \geq 1$ is equal to 109, where these



(a) $r = 1$ (b) $r = 2$ (c) $r = 3$

Figure 2.6: All paths for which the event $M_1 \geq 1$ has to hold



(a) $r = 1$ (b) $r + 1$ (c) $r + 2$

Figure 2.7: All paths for which the event $M_1 \geq 1$ is possible

paths are presented in Figure 2.7. If the aim is to derive the NPI lower and upper probabilities for the event that C_2 contains at least one of the three future observations, so $M_2 \geq 1$. Equations (2.12) and (2.13) are applied to derive the NPI lower and upper probabilities, respectively. The NPI lower probability for $M_2 \geq 1$ with the values $n = 8$, $n_1 = 4$ and $m = 3$ is

$$\begin{aligned}
\underline{P}\left(M_2 \geq 1\right) &= \binom{n+m}{m}^{-1}\left[\sum_{r=m_2}^{m}\binom{n_2-2+r}{r}\sum_{v=0}^{m-r}\binom{n_1+v}{v}\binom{n_3+m-r-v}{m-r-v}\right] \\
&= \binom{11}{3}^{-1}\left[\sum_{r=1}^{3}\binom{2+r}{r}\sum_{v=0}^{3-r}\binom{2+v}{v}\binom{2+3-r-v}{3-r-v}\right] \\
&= \binom{11}{3}^{-1}\left[3\left(1\times 6+3\times 3+6\times 1\right)+6\left(1\times 3+3\times 1\right)+10\left(1\times 1\right)\right] \\
&= \binom{11}{3}^{-1}\left[3\times 21+6\times 6+10\times 1\right] = 0.6606
\end{aligned}$$

The corresponding NPI upper probability is derived by counting all paths that go through at least one point (n_2, r) with $r \geq m_2$, in addition to paths from $(0,0)$ to (n,m) via at least one of $(n_2 - 1, r + 1)$ and $(n_2 - 1, m)$. The NPI upper probability for the event $M_2 \geq 1$ is

$$\begin{aligned}\overline{P}(M_2 \geq 1) &= \binom{8+3}{3}^{-1} \times \left[\binom{4+1}{1}\binom{8-4+(3-1)}{3-1}+\right.\\ &\left.\sum_{j=2}^{3}\binom{4-1+j}{j}\binom{8-4+(3-j)}{m-j}\right]\\ &= \binom{8+3}{3}^{-1}[5 \times 15 + 10 \times 5 + 20 \times 1] = 0.8788.\end{aligned}$$

For the event $M_3 \geq 1$, the NPI lower and upper probabilities are equal to those for the event $M_1 \geq 1$ as the $n_1 = n_3 = 2$. The results in this example show how NPI lower and upper probabilities vary specifically with the number of observations in each category, as shown by the different calculations for the events $M_1 \geq 1$, $M_2 \geq 1$, and $M_3 \geq 1$. Inferences using this event will be presented in the next chapter. ◇

The event $M_k \geq m_k$ can be extended to the case in which the future observations are in adjoining categories instead of a specific category. This case involving adjoining categories, which is presented next, is effectively the same as reducing the total number of categories by considering these adjoining categories together as one combined category. Inferences using this event are presented in Chapters 3 and 4.

Consider the event that at least a specific number out of the m future observations are in adjoining categories. To derive the NPI lower and upper probabilities for this event, we introduce new notation, following Coolen et al. [32] and Elkhafifi [47]. Let $\mathcal{C}_T = \bigcup_{k \in T} C_k$ and $T \subset \{1, \ldots, K\}$, where $\mathcal{C}_T$ consists of adjoining categories. In order to derive the NPI lower and upper probabilities, the corresponding latent variable $Y_{n+i} \in \mathcal{IC}_T$ for $i = 1, \ldots, m$ is considered, where $\mathcal{IC}_T = \bigcup_{k \in T} IC_k$. Suppose that $T = \{s, \ldots, t\}$ with $s, t \in \{1, \ldots, K\}, s \leq t$. Let $\mathcal{C}_{s,t} = \bigcup_{k=s}^{t} C_k, \mathcal{IC}_{s,t} = \bigcup_{k=s}^{t} IC_k$ and $n_{s,t} = \sum_{k=s}^{t} n_k$. Let $M_{s,t}$ represent the number of m future observations that belong to the combined category $\mathcal{C}_{s,t}$, so the event is $M_{s,t} \geq m_{s,t}$. The case with $s = 1$ and $t = K$ will be excluded, as both NPI lower and upper probabilities for the event $M_{s,t} \geq m_{s,t}$ are equal to 1.

For the event $M_{s,t} \geq m_{s,t}$, two different situations should be considered to derive the NPI lower and upper probabilities. These two situations are whether the category $\mathcal{C}_{s,t}$ has one of the first or last categories, so $s = 1$ or $t = K$, or not. The NPI lower probability for the event $M_{s,t} \geq m_{s,t}$ for the case where $s = 1$ or $t = K$ is

$$\underline{P}(M_{s,t} \geq m_{s,t}) = \binom{n+m}{m}^{-1} \sum_{r=m_{s,t}}^{m} \binom{n_{s,t}-1+r}{r} \binom{n-n_{s,t}-1+m-r}{m-r} \tag{2.14}$$

For $1 < s \leq t < K$, let $n_{1,s-1} = \sum_{k=1}^{s-1} n_k$ and $n_{t+1,K} = \sum_{k=t+1}^{K} n_k$. The NPI lower probability for the event $M_{s,t} \geq m_{s,t}$ for the case where $1 < s \leq t < K$ is

$$\begin{aligned} \underline{P}(M_{s,t} \geq m_{s,t}) = \binom{n+m}{m}^{-1} \Bigg[\sum_{r=m_{s,t}}^{m} \binom{n_{s,t}-2+r}{r} \sum_{v=0}^{m-r} \binom{n_{1,s-1}+v}{v} \times \\ \binom{n_{t+1,K}+m-r-v}{m-r-v} \Bigg] \end{aligned} \tag{2.15}$$

The corresponding NPI upper probability for the event $M_{s,t} \geq m_{s,t}$ with $1 \leq s \leq t \leq K$ is

$$\begin{aligned} \overline{P}(M_{s,t} \geq m_{s,t}) = \binom{n+m}{m}^{-1} \times \Bigg[\binom{n_{s,t}+r}{r} \binom{n-n_{s,t}+(m-r)}{m-r} + \\ \sum_{j=r+1}^{m} \binom{n_{s,t}-1+j}{j} \binom{n-n_{s,t}+(m-j)}{m-j} \Bigg] \end{aligned} \tag{2.16}$$

This section has presented the path counting technique for NPI with ordinal data to obtain the NPI lower and upper probabilities for the events of interest. This technique can similarly be applied to other events of interest, offering a useful approach for future analyses. The derivation of NPI lower and upper probabilities has been presented first for the event $M_k \geq m_k$, and then for the event $M_{s,t} \geq m_{s,t}$. The total number of paths from $(0, 0)$ to (n, m) and their implications on the derivation of NPI lower and upper probabilities have been explained. Inferences and more detailed scenarios involving the events presented in this section, along with examples from the literature, will be presented throughout this thesis.

2.5 Concluding remarks

In this chapter, NPI lower and upper probabilities for several events of interest involving multiple future ordinal observations were presented. These results will enable a variety of statistical inferences, several of which are presented in the following chapter.

The initial focus was on two future observations to provide a clear understanding of the methodology involved in counting the orderings of future observations. Two specific events were presented: one where one of the future observations is in a specific category while the other is in any of the remaining categories, and another where both future observations are in a specific category. The distinction between boundary intervals of first or last categories and middle categories was highlighted, as it impacts the derivation of the NPI lower and upper probabilities.

The methodology was then generalised to $m \geq 2$ future observations, considering two events of interest. The first event considered is that there is a particular number of future observations in each one of the categories. The second event of interest is that precisely m_k out of the m future observations belong to a specific category. The results show how these NPI lower and upper probabilities vary based on the number of data observations and the number of future observations.

Attention was then directed towards the event where at least m_k out of m future observations are in a specific category. To derive the NPI lower and upper probabilities for this event, a path counting technique was introduced to avoid double counting of the orderings of future observations by representing all possible orderings as paths in an $n \times m$ lattice. The path counting technique was extended to the event involving adjoining categories, where the corresponding union of intervals forms a single interval on the real-line. The flexibility of the path counting method makes it suitable for application to more complex events. For instance, the method can be used to events where a specific number of future observations are in non-adjoining categories. This is left as a topic for future research.

Chapter 3

Inferences involving multiple future ordinal observations

3.1 Introduction

Many statistical applications involve identifying the optimal choice or choices from a set of possibilities, such as determining the most effective treatment in a clinical study. Methods specifically designed to select the best treatment or the optimal member of some group are known as selection procedures [53]. For example, in the context of multinomial data, the objective may be to select the category with the largest probability of occurrence [12, 13]. Selection procedures have important applications in a wide range of fields, including social sciences [20], medicine [25] and marketing [65]. Moreover, comparing two or more independent groups of data, such as those from different treatments in a clinical study, is a common problem. Pairwise comparison can be used to determine which group performs better or is more likely to lead to a desired outcome [7]. This chapter presents statistical methods for category selection, subset selection of categories and pairwise comparison problems.

The structure of the chapter is as follows. Section 3.2 introduces a method for selecting a single category based on multiple future ordinal observations. Section 3.3 presents a method for selecting a subset of categories. In Section 3.4, a pairwise comparison method based on multiple future observations from two groups is presented. Finally, concluding remarks are provided in Section 3.5.

3.2 NPI-based category selection

Generally, selection methods have important applications in a wide range of fields. In social science research, applying tools like the Likert scale allows for the analysis of questionnaire data, enabling the identification of key factors influencing people's opinions [20, 63]. The Likert scale, commonly used in social research, was developed in 1932 by Rensis Likert to measure attitudes [20]. For example, in social sciences, an ordinal scale may be used to analyze questionnaire data when presenting five options to respond to a statement in a survey of opinions, such as: strongly disagree, disagree, neutral, agree, and strongly agree. Similarly, in medical research, these methods play a crucial role in identifying factors contributing to the development of effective prevention and treatment strategies for diseases [25]. In market research, these methods can assist companies in making informed decisions regarding product development, marketing strategies, pricing, and enhancing sales management efficacy [65]. Methods for selection problems, have been extensively studied in the statistics literature [16, 17, 56]. Similar NPI-based methods for selection problems, with some important variations, have been developed for real-valued data [36], proportions data [30, 31], and lifetime data, including right-censored observations [42, 44], as discussed in Section 1.3.

For category selection, NPI methods have been introduced for situations involving multiple future observations, aiming to select either a single category with the largest lower or upper probability of occurrence or the smallest subset of categories that meets a specified probability requirement for multinomial data [12, 13]. In this section, NPI-based method for category selection for an ordinal data set will be presented, where the inferences involve m future observations. The method aims to determine which category should be selected based on the NPI lower and upper probabilities for future observations and a specified criterion.

The ordinal data, as discussed in Chapter 2, are represented on the real-line with n data observations that partition the real-line into $n+1$ intervals. Data and future observations are linked via this assumed underlying data representation, with latent observations on the real-line falling into intervals which represent the categories, and the use of the $A_{(\cdot)}$ assumptions. Consider $K \geq 2$ categories with the ordering between these categories indicated by $C_1 < C_2 < \ldots < C_K$. Assume that n observations are available and that n_k is the number of observations in category C_k, for $k = 1, \ldots, K$, so $\sum_{k=1}^{K} n_k = n$, where $n_k \geq 0$. Let the random quantity M_k represent the number of future observations that belong to category C_k, for $k = 1, \ldots, K$.

The event that will be considered for category selection is based on the event of interest that has been presented in Section 2.4. The event is that at least m_k of the m future observations are in a specific category C_k, denoted by $M_k \geq m_k$. To select a category based on this event, the NPI lower and upper probabilities for each category are derived. For a specified p^*, categories that satisfy either $\underline{P}(M_k \geq m_k) \leq p^*$ or $\overline{P}(M_k \geq m_k) \leq p^*$ are identified. The criterion $\leq p^*$ is used to exclude any category that has an NPI lower or upper probability larger than p^* and to focus on those that meet the criterion. Among the remaining categories, the one with the largest NPI lower or upper probability is chosen. For example, in loan risk assessment, where categories represent levels of repayment risk, p^* could represent the confidence level for accepting a borrower; then, the category with the largest NPI lower or upper probability can be selected to ensure it is the closest to p^*, making it the most appropriate choice among those meeting the criterion. This approach ensures that, while we are only considering categories that meet the criterion, we are still selecting the best one from those remaining, such that the selected category has the largest NPI lower or upper probability for the event that at least m_k future observations belong to that selected category. This method is relevant in practice when selecting categories, especially when the goal is to ensure a specific level of confidence that the selected category does not exceed this level, or to avoid selecting categories that exceed the desired selection criterion. If one aims to select a category that has NPI lower or upper probabilities greater than or equal to p^*, the criterion $\geq p^*$ can be used. In the example later in this section, a scenario is presented to show how the category selected using $\geq p^*$ differs from the one selected with $\leq p^*$.

The practical importance of this event and criterion in real-life contexts can, for example, be illustrated by a situation involving ordered categories representing the severity level of a condition or disease. Importantly, these events are not confined solely to medical contexts; their applicability can be extended to various fields.

Consider a scenario in which healthcare providers are required to categorize a disease based on severity levels, from mild to severe, using the World Health Organization (WHO) ordinal scale. The WHO ordinal scale is a useful tool to assess illness severity in Coronavirus disease (COVID-19) patients [81]. It categorizes patients into different stages based on their clinical status, ranging from mild illness to severe disease and death. The scale provides a standardized framework for evaluating and monitoring patients, aiding in treatment decisions and predicting outcomes. In the context of COVID-19, the WHO ordinal scale has been widely used to guide clinical management and research efforts. It helps healthcare professionals classify patients according to their need for hospitalization, oxygen therapy, and intensive care.

Suppose that the objective here is to address the problem of determining the severity category to which at least a certain number of future patients will belong. NPI helps in predicting which category future observations are likely to be within, such as whether at least m_k future patients will be in mild or death category. One can use the NPI lower probability $\underline{P}(M_k \geq m_k)$ as a minimum boundary for ensuring at least a number of future observations are in a specific category C_k. Alternatively, the NPI upper probability $\overline{P}(M_k \geq m_k)$ can be used as a maximum boundary for ensuring at least a number of future observations are in a specific category C_k. By setting a specific criterion p^*, healthcare administrators can ensure a specific level of confidence that the selected category, meeting this criterion, will contain at least m_k future cases. The choice of m is essential in this selection process. When m is set to a small number, it may reflect a scenario where decision-makers only focus on a limited number of future observations, such as during a period of decreasing COVID-19 cases, or m can be set to a large value to prepare for a scenario with a larger number of severe cases, such as during a peak in the pandemic. Whether increasing the number of future observations, m, will affect the selection of the category depends on how the NPI lower and upper probabilities for each category respond to changes in m. This influence of m on NPI lower and upper probabilities will be illustrated via an example using data from the literature.

The category selection is demonstrated in a study on the availability of personal protective equipment (PPE) in NHS (National Health Service) hospitals during

COVID-19, as presented by Mantelakis et al. [66]. The PPE items included eye and face protection, surgical masks, filtering facepiece class 3 (FFP3) respirator, gloves, and plastic aprons. This study investigates PPE availability, using social media for survey distribution to UK healthcare professionals and a Likert scale to assess respondents' perceived protection against infections.

Example 3.1. The study distributed a survey via social media to various UK COVID-19 healthcare professional groups, involving 121 participants who were healthcare workers employed by the NHS. This distribution strategy enabled researchers to collect responses from a large number of healthcare workers employed by the NHS while they were working within the hospital. Over a period of 3 weeks, the survey collected a total of 121 replies from 35 hospitals across England. Participants were asked the following question: "Are the aforementioned PPEs available as needed?" Their responses were categorized using the Likert scale into five categories: C_1: always; C_2: usually; C_3: occasionally; and C_4: almost never; and C_5: never. Table 3.1 illustrates the numbers of participants who selected each response category. The availability of personal protective equipment (PPE) during a pandemic such as COVID-19 can be crucial not only for healthcare workers, but also for controlling the spread of the disease, and assessing the availability of sufficient PPE may contribute to public health efforts. For example if the category selected is either C_3: occasionally; or C_4: almost never; or C_5: never, it highlights the need for immediate attention and resource allocation to meet the demand, while if the category selected is C_1: always or C_2: usually, it suggests that there is a consistent or frequent availability of PPE. This can be encouraging as it implies that healthcare workers have the necessary protective equipment for their safety.

Consider inferences about the next 4 healthcare workers, so $m = 4$ future observations, with the objective to select the category that satisfies either $\underline{P}(M_k \geq m_k) \leq 0.75$ or $\overline{P}(M_k \geq m_k) \leq 0.75$, which can be particularly relevant in scenarios where decisions must be based on the responses of a small group. The assessment of PPE availability during the COVID-19 pandemic requires a careful consideration of the value of m as it affects both the NPI lower ($\underline{P}$) and upper ($\overline{P}$) probabilities. One can set a high criterion that is enough to reflect a concern

Category	C_1	C_2	C_3	C_4	C_5
Observations	36	49	23	10	3

Table 3.1: Self-reported PPE availability

C_k	$M_k \geq 1$		$M_k \geq 2$		$M_k \geq 3$		$M_k \geq 4$	
	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$
C_1	0.7481	0.7594	0.3404	0.3547	0.0833	0.0896	0.0085	0.0094
C_2	0.8604	0.8746	0.5122	0.5401	0.1754	0.1944	0.0258	0.0302
C_3	0.5438	0.5787	0.1540	0.1783	0.0222	0.0281	0.0013	0.0018
C_4	0.2612	0.3115	0.0319	0.0457	0.0020	0.0034	0	0.0001
C_5	0.0937	0.1234	0.0045	0.0075	0.0001	0.0002	0	0

Table 3.2: NPI lower and upper probabilities for the event $M_k \geq m_k$ where $k = 1, \ldots, 5$ and $m = 4$

for events that could impact PPE availability, ensuring that attention is directed toward scenarios with considerable potential for such events to occur. By setting a decision-making criterion at $p^* = 0.75$, one can focus on identifying categories where either $\underline{P}$ or $\overline{P}$ is below this criterion. The optimal category is then selected based on the largest NPI lower or upper probability among those that meet this criterion.

The NPI lower and upper probabilities presented in Section 2.4, particularly Equations 2.11, 2.12, and 2.13, are applied. Table 3.2 presents the NPI lower and upper probabilities for the events $M_k \geq m_k$ where $k = 1, \ldots, 5$ and $m = 4$. Table 3.3 extends this analysis by increasing the number of future observations to $m = 8$, while considering the same events presented in Table 3.2. In this example, how m values affect the NPI lower and upper probabilities for the events $M_k \geq m_k$ will be discussed first. The optimal category selection is then introduced considering the condition $\underline{P}(M_k \geq m_k) \leq 0.75$, followed by the selection based on the condition $\overline{P}(M_k \geq m_k) \leq 0.75$. Finally, the selection is explored with different scenarios to show how the selection of the optimal category differs depending on whether the criterion is based on $\underline{P}(M_k \geq m_k) \leq p^*$ or $\overline{P}(M_k \geq m_k) \leq p^*$.

For the event $M_k \geq 1$ with $m = 4$, as presented in Table 3.2, the NPI lower probabilities for categories $k = 1$ through $k = 5$ are 0.7481, 0.8604, 0.5438, 0.2612, and 0.0937, respectively. The corresponding NPI upper probabilities are 0.7594, 0.8746, 0.5787, 0.3115, and 0.1234. When m is increased to 8, as shown in Table 3.3, the NPI lower probabilities the event $M_k \geq 1$ for $k = 1$ through $k = 5$ become

C_k	$M_k \geq 1$		$M_k \geq 2$		$M_k \geq 3$		$M_k \geq 4$	
	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$
C_1	0.9332	0.9389	0.7263	0.7425	0.4350	0.4553	0.1918	0.2064
C_2	0.9789	0.9829	0.8789	0.8962	0.6644	0.6979	0.3929	0.4300
C_3	0.7862	0.8171	0.4344	0.4826	0.1673	0.2012	0.0452	0.0591
C_4	0.4487	0.5201	0.1179	0.1622	0.0207	0.0337	0.0025	0.0049
C_5	0.1760	0.2283	0.0191	0.0308	0.0015	0.0029	0.0001	0.0002

Table 3.3: NPI lower and upper probabilities for the event $M_k \geq m_k$ where $k = 1, \ldots, 5$ and $m = 8$

0.9332, 0.9789, 0.7862, 0.4487, and 0.1760, and the NPI upper probabilities are 0.9389, 0.9829, 0.8171, 0.5201, and 0.2283. As the number of future observations increases to 8, the NPI lower probabilities for all categories generally increase, and the NPI upper probabilities also increase. For the other events, such as $M_k \geq 2$, $M_k \geq 3$, and $M_k \geq 4$, similar increases in both NPI lower and upper probabilities are observed with more observations. These changes indicate that as m increases, both the NPI lower and upper probabilities increase. Next, the optimal category selection is introduced by first considering the condition $\underline{P}(M_k \geq m_k) \leq 0.75$, followed by the selection based on the condition $\overline{P}(M_k \geq m_k) \leq 0.75$.

For the event $M_k \geq 1$ with $m = 4$, the NPI lower probability for $k = 1$ is 0.7481, which is just below the criterion $p^* = 0.75$. For $k = 2$, the NPI lower probability is 0.8604, which exceeds the criterion, while the NPI lower probability for $k = 3$ is 0.5438, which satisfies the criterion but is not the largest. Among the categories that satisfy the criterion for the event $M_k \geq 1$, C_1 is selected as the optimal choice due to its larger NPI lower probability. When the number of future observations is increased to $m = 8$, the NPI lower probabilities for $k = 1$ and $k = 2$ increase to 0.9332 and 0.9789, respectively, both of which no longer satisfy the criterion. While the NPI lower probability for $k = 3$ is 0.7862, which still does not meet the criterion, the NPI lower probability for $k = 4$ is 0.4487. Therefore, C_4 becomes the optimal choice when $m = 8$. This demonstrates that as m increases, the optimal choice can change.

For the event $M_k \geq 2$ with $m = 4$, the NPI lower probabilities for $k = 1$ and $k = 2$ are 0.3404 and 0.5122, respectively, both of which satisfy the criterion. Among these, C_2 is selected as the optimal choice due to its larger NPI lower probability. When the number of future observations is increased to $m = 8$, the NPI lower probabilities for $k = 1$ and $k = 2$ change, with C_2 no longer meeting the criterion at 0.8789, while $k = 1$ still meets it at 0.7263, making C_1 the optimal category.

For the event $M_k \geq 3$ with $m = 4$, the NPI lower probabilities for $k = 1$ and $k = 2$ are 0.0833 and 0.1754, respectively, both of which satisfy the criterion. C_2 is selected as the optimal choice. When the number of future observations is increased to $m = 8$, the NPI lower probabilities for $k = 1$ and $k = 2$ are increased to 0.4350 and 0.6644, respectively. Both still satisfy the criterion, with C_2 remaining the optimal category due to its larger NPI lower probability.

For the event $M_k \geq 4$ with $m = 4$, all categories have small NPI lower probabilities, with $k = 1$ and $k = 2$ showing values of 0.0085 and 0.0258, respectively, which satisfy the criterion $p^* = 0.75$. C_2 is considered optimal due to its larger NPI lower probability. When m is increased to 8, the NPI lower probabilities for $k = 1$ and $k = 2$ increase to 0.1918 and 0.3929, respectively, but they still remain below the criterion, making C_2 the optimal category.

When considering NPI upper probabilities for the event $M_k \geq 1$ with $m = 4$, the NPI upper probabilities for $k = 1$ and $k = 2$ are 0.7594 and 0.8746, respectively, which are above the $p^* = 0.75$ criterion and do not meet it. However, $k = 3$ satisfies the criterion with the NPI upper probability equal to 0.5787, making C_3 the optimal choice at $m = 4$. When $m = 8$, the NPI upper probabilities for $k = 1$ and $k = 2$ increase to 0.9389 and 0.9829, respectively, both no longer meeting the criterion $p^* = 0.75$. Also, $k = 3$ no longer meets the criterion as the NPI upper probability is 0.8171. However, the NPI upper probability with $k = 4$ is 0.5201, which meets the criterion, making C_4 the optimal choice when $m = 8$.

For the event $M_k \geq 2$ with $m = 4$, $k = 1$ and $k = 2$ meet the criterion with NPI upper probabilities of 0.3547 and 0.5401, respectively, making C_2 the optimal choice. When m is increased to 8, the NPI upper probabilities for $k = 1$ and $k = 2$ change to values that no longer meet the criterion, and C_3 is selected as optimal based on its maximum NPI upper probability among those that satisfy the criterion.

When analyzing $M_k \geq 3$ with $m = 4$, both $k = 1$ and $k = 2$ satisfy the criterion with NPI upper probabilities of 0.0896 and 0.1944, respectively, making C_2 the optimal choice. When m is increased to 8, the NPI upper probability for $k = 2$ still meets the criterion keeping C_2 as the optimal choice based on its maximum NPI upper probability. Similarly, for $M_k \geq 4$ with $m = 4$, C_2 is the optimal choice, and it continues to meet the criterion as m increases to 8.

The optimal category is selected as the one with the maximum NPI lower or upper probability among those that satisfy the criterion for the event. The category selection, however, can differ depending on whether the criterion is based on $\underline{P}(M_k \geq m_k) \leq p^*$ or $\overline{P}(M_k \geq m_k) \leq p^*$. For example, in the event $M_k \geq 1$ with $m = 4$ and a criterion of 0.75, C_1 is considered optimal based on $\underline{P}$ because its NPI lower probability is 0.7481, the largest value that meets the criterion, while C_3 is optimal based on $\overline{P}$. Another example is for the event $M_k \geq 1$ with $m = 8$ and a criterion of 0.45; here, C_4 is optimal for $\underline{P}$ because its NPI lower probability is 0.4487, which meets the criterion, and C_5 is optimal for $\overline{P}$. Similarly, for the event $M_k \geq 3$ with $m = 8$ and a criterion of 0.45, C_1 is optimal for $\underline{P}$ because its NPI lower probability is 0.4350, the largest value within the criterion, while C_3 is selected for $\overline{P}$. In another scenario, one might be interested in adjusting the NPI upper criterion and selecting the category based on the criterion $\overline{P}(M_k \geq m_k) \geq p^*$. For instance, in the event $M_k \geq 1$ with $m = 4$, the NPI upper probabilities for categories C_1 and C_2 are above the criterion $p^* = 0.75$, making them suitable. Among these, C_2 is the optimal category because its NPI upper probability is larger than that of C_1. This selection differs from using the criterion $\overline{P}(M_k \geq 1) \leq p^*$. As discussed earlier in this section, if one aims to select a category that has NPI lower or upper probabilities greater than or equal to p^*, the criterion $\geq p^*$ can be used instead of $\leq p^*$, depending on the specific objectives of the analysis. This highlights the importance of carefully considering both the chosen criterion and the condition depending on one's analysis objectives. $\diamond$

Overall, the selection of the optimal category can vary with the number of future observations considered. As m increases, the NPI lower and upper probabilities typically become larger, which may cause categories that were optimal at a smaller m to exceed the criterion and thus no longer qualify. This can lead to different optimal categories being selected as m changes. Consequently, the choice between $\underline{P}(M_k \geq m_k) \leq p^*$ or $\overline{P}(M_k \geq m_k) \leq p^*$, along with the number of future observations, can result in different selections, highlighting the importance of carefully considering these in decision-making. The next section presents how NPI can be used to select subsets of categories from an ordinal dataset, moving beyond the selection of just one category.

3.3 NPI-based selection of a subset of categories

Subset selection is a common problem in statistical analysis. Classical methods for subset selection, such as Gupta's subset selection method [56], focused on identifying a subset of treatments that includes the best treatment with a certain confidence level. These methods do not focus on selecting one or more subsets of categories most likely to occur for a single variable, as discussed by Bechhofer et al. [16]. These methods are non-predictive, relying solely on hypothesis testing. This section introduces a nonparametric predictive approach, called NPI, for selecting subsets of categories with multiple future ordinal observations. An NPI-based method has been developed for selecting a subset of categories, particularly in scenarios where no prior knowledge of the relationships between the categories is available [12]. This section presents an NPI-based method where the knowledge about ordering of categories is taken into account to select a subset of ordered categories. The inferences about these future observations utilize the event of interest introduced in Section 2.4. For this subset selection, making inferences about m future observations will require the introduction of some new notation.

Recall that the combined ordered category $\mathcal{C}_{s,t} = \bigcup_{k=s}^{t} C_k$ and $n_{s,t} = \sum\limits_{k=s}^{t} n_k$. If $s = t$ then the m future observations are in a specific category. Let the selected subset of categories be denoted by S, where $S = \{s, \ldots, t\}$ with $s, t \in \{1, \ldots, K\}, s \leq t$. Let the number of future observations that are in S be represented by the random quantity M_S. The focus of this section is on the event that $M_S \geq m_S$. Note that, the case with $s = 1$ and $t = K$ will be excluded, as both NPI lower and upper probabilities for the event $M_S \geq m_S$ are equal to 1 for all $m_S \in \{0, 1, \ldots, m\}$. Instead of focusing on a specific value of M_S, it may be more practical to consider the event $M_S \geq m_S$, which provides a clear criterion (e.g., ensuring that subset S contains at least a certain number of future observations). Thinking about having at least a number of future observations in S seems more practical when dealing with future observations. Gupta [56] introduced the p^* condition, a probability criterion used to ensure that a selected subset meets or exceeds a probability threshold for a desired outcome. This criterion p^* is chosen to achieve a specific level of confidence or assurance that the selected subset meets a desired outcome. In this section, a similar idea is applied to identify a subset S such that the NPI lower or upper probability of the event $M_S \geq m_S$ meets or exceeds a specified p^*. The objective here is to identify a subset S such that $\underline{P}(M_S \geq m_S) \geq p^*$ or $\overline{P}(M_S \geq m_S) \geq p^*$, while ensuring that S is of minimal size by evaluating all possible subsets and selecting the one with the smallest number of categories that still satisfies the criterion. If several such subsets exist, the one with the maximum NPI lower or upper probability is selected. This subset is then referred to as the optimal subset S. This method will be presented using the NPI lower and upper probabilities in Section 2.4.

The selection of a minimal-sized subset consisting of adjoining categories with a specified criterion could be of interest in various fields. For instance, healthcare providers might need to select a subset of categories containing specific adjoining severity levels. According to Serlin et al. [84], cancer pain is categorized into 10 severity levels, with C_1 to C_4 representing mild pain, C_5 and C_6 representing moderate pain, and C_7 to C_{10} representing severe pain. In healthcare, particularly in cancer pain management, it is crucial to identify a subset of severe pain categories such as C_7 to C_{10} that are likely to include at least a certain number of future patients. With limited resources, healthcare providers need a method to focus on the most critical cases, ensuring that those in severe pain categories receive timely and adequate treatment. The NPI-based subset selection method can help achieve this by identifying subsets of pain severity levels likely to contain at least a specific number of future severe pain cases. Suppose that one aims to select a minimal-sized subset containing adjoining categories that represent severe levels of cancer pain, ensuring that $\underline{P}(M_S \geq m_S) \geq p^*$ or $\overline{P}(M_S \geq m_S) \geq p^*$. This criterion p^* can be set to achieve a specific level of confidence that the selected subset will contain at least m_S of future severe pain cases. In healthcare, meeting such a criterion could be beneficial in directing resources toward future cases likely to be severe, potentially enabling more effective and targeted interventions. The NPI-based subset selection method allows healthcare providers to choose m based on their specific objectives; for instance, selecting a larger m to prepare for a scenario where a larger number of severe cases is likely, or a smaller m when the focus is only on the immediate severe cases.

NPI method of selecting subsets is illustrated with data sets from the literature, for example, a study conducted by Wang et al. [89], which investigated the clinical severity and outcomes of adult patients hospitalized with laboratory-confirmed seasonal influenza A or B virus infections. The subset selection method is demonstrated using data related to the severity of illness caused by the influenza A virus.

Example 3.2. The study was conducted by Wang et al. [89] between October 2016 and June 2018, involving adult patients aged 18 years and older diagnosed with laboratory-confirmed seasonal influenza A or B infections. There were two types of influenza virus infections among the patients. The research concluded that influenza A infection demonstrated more severe clinical outcomes compared to influenza B infection among hospitalized adults with laboratory-confirmed seasonal influenza. The assessment of clinical improvement was based on a 7-category ordinal scale, reflecting the patient's condition at discharge. According to the study, the rate of clinical improvement assessed by the ordinal scale may be a reasonable endpoint for patients who are hospitalized with influenza infection. The study's results highlighted the potential utility of a 7-category ordinal scale in assessing the severity of influenza infections and predicting clinical outcomes. This ordinal scale was used to assess the clinical status of patients at fixed time points, categorized into seven categories, where C_1 indicated the best outcome and C_7 indicated the worst outcome. On day 14, the study reported the number of patients in each category of the 7-category ordinal scale, with 363 identified as influenza A cases.

These categories are C_1: Discharged with resumption of normal activities, C_2: Discharged without resumption of normal activities, C_3: Non-ICU (intensive care unit), not requiring oxygen, C_4: Non-ICU, requiring oxygen, C_5: ICU, not requiring IMV (invasive mechanical ventilation), C_6: ICU, requiring IMV, C_7: Death. Table 3.4 illustrates the number of patients for each response category.

Consider inferences about four future observations with the objective of selecting subsets of minimal size that satisfy the criterion $\underline{P}(M_S \geq m_S) \geq p^*$ or

Category	C_1	C_2	C_3	C_4	C_5	C_6	C_7
Observations	201	3	81	18	16	23	21

Table 3.4: Ordinal scale data at day 14 for influenza A patients

$\overline{P}(M_S \geq m_S) \geq p^*$. We are interested in four events: having at least 1 future observation in S, at least 2, 3, and all 4 future observations in S. In this example, an increasing sequence of subsets will be considered, starting with the individual categories, and gradually expanding by one category at a time. The subset containing all of the categories, $\{C_1, \ldots, C_7\}$, will be excluded, as the NPI lower and upper probabilities are equal to one. The NPI lower and upper probabilities for several events $M_S \geq m_S$ are derived with S consisting of a single category or a subset of categories by applying Equations (2.14) and (2.15) for the NPI lower probability and Equation (2.16) for the NPI upper probability, considering m_S values ranging from 1 to 4. The composition of each individual subset and these NPI lower and upper probabilities are presented in Table 3.5.

Suppose the objective is to select subsets containing the category 'Discharged with resumption of normal activities', denoted as C_1, that satisfy a specified criterion where one or more of the future observations belong to a category in that subset, so the event is $M_S \geq 1$. Considering the criterion is equal to 0.7, for the NPI lower probability, $\underline{P}(M_S \geq 1) \geq 0.7$, subsets are evaluate based on their respective lower probabilities derived from Equation (2.14). First, a subset of minimal size is chosen such that the NPI lower probability for the event that one or more of the future observations belong to a category in that subset is at least 0.7. As a result, $S = \{C_1\}$ is selected as the initial subset satisfying this criterion. Moving to the second event of interest, aiming for at least two future observations within the subset that meet the criterion, $\underline{P}(M_S \geq 2) \geq 0.7$, we select $S = \{C_1\}$ from Table 3.5. Similarly, the goal remains to select a minimal-size subset where three, $\underline{P}(M_S \geq 3) \geq 0.7$, and eventually all future observations, $\underline{P}(M_S \geq 4) \geq 0.7$, satisfy the criterion. A larger subset needs to be selected now to achieve this minimally required probability. From Table 3.5, the smallest subset which satisfies $\underline{P}(M_S \geq 3) \geq 0.7$ is $S = \{C_1, C_2, C_3\}$. The smallest subset which meet the required criterion based on their respective NPI lower probabilities for $\underline{P}(M_S \geq 4) \geq 0.7$ is $S = \{C_1, \ldots, C_6\}$.

S	$M_S \geq 1$		$M_S \geq 2$		$M_S \geq 3$		$M_S \geq 4$	
	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$
$\{C_1\}$	0.9590	0.9600	0.7603	0.7639	0.3953	0.3998	0.0942	0.0961
$\{C_1, C_2\}$	0.9619	0.9628	0.7711	0.7746	0.4089	0.4134	0.0999	0.1019
$\{C_1, C_2, C_3\}$	0.9976	0.9978	0.9649	0.9662	0.7918	0.7961	0.3775	0.3828
$\{C_1, \ldots, C_4\}$	0.9991	0.9992	0.9830	0.9837	0.8658	0.8696	0.4817	0.4881
$\{C_1, \ldots, C_5\}$	0.9997	0.9998	0.9928	0.9932	0.9217	0.9248	0.5912	0.5986
$\{C_1, \ldots, C_6\}$	0.9999	1	0.9990	0.9992	0.9791	0.9808	0.7801	0.7892
$\{C_2\}$	0.0217	0.0430	0.0003	0.0009	0	0	0	0
$\{C_2, C_3\}$	0.6431	0.6531	0.2259	0.2349	0.0402	0.0429	0.0029	0.0031
$\{C_2, C_3, C_4\}$	0.7258	0.7340	0.3092	0.3188	0.0687	0.0725	0.0062	0.0067
$\{C_2, \ldots, C_5\}$	0.7863	0.7931	0.3863	0.3961	0.1020	0.1067	0.0110	0.0118
$\{C_2, \ldots, C_6\}$	0.8551	0.8602	0.4977	0.5072	0.1632	0.1693	0.0225	0.0238
$\{C_2, \ldots, C_7\}$	0.9039	0.9058	0.6002	0.6047	0.2361	0.2397	0.0400	0.0410
$\{C_3\}$	0.6277	0.6380	0.2126	0.2215	0.0363	0.0389	0.0025	0.0027
$\{C_3, C_4\}$	0.7131	0.7216	0.2950	0.3045	0.0633	0.0669	0.0055	0.0059
$\{C_3, C_4, C_5\}$	0.7758	0.7829	0.3718	0.3815	0.0952	0.0997	0.0100	0.0107
$\{C_3, \ldots, C_6\}$	0.8472	0.8525	0.4833	0.4929	0.1543	0.1602	0.0206	0.0218
$\{C_3, \ldots, C_7\}$	0.8981	0.9001	0.5866	0.5911	0.2254	0.2289	0.0372	0.0381
$\{C_4\}$	0.1735	0.1923	0.0129	0.0159	0.0005	0.0006	0	0
$\{C_4, C_5\}$	0.3151	0.3314	0.0444	0.0495	0.0030	0.0035	0	0
$\{C_4, C_5, C_6\}$	0.4858	0.4990	0.1155	0.1229	0.0134	0.0148	0.0006	0.0007
$\{C_4, \ldots, C_7\}$	0.6172	0.6225	0.2039	0.2082	0.0338	0.0351	0.0022	0.0024
$\{C_5\}$	0.1543	0.1735	0.0102	0.0129	0.0003	0.0005	0	0
$\{C_5, C_6\}$	0.3554	0.3710	0.0576	0.0632	0.0045	0.0052	0.0001	0.0002
$\{C_5, C_6, C_7\}$	0.5119	0.5183	0.1304	0.1342	0.0163	0.0170	0.0008	0.0009
$\{C_6\}$	0.2199	0.2379	0.0209	0.0246	0.0009	0.0012	0	0
$\{C_6, C_7\}$	0.4014	0.4088	0.0752	0.0783	0.0068	0.0072	0.0002	0.0003
$\{C_7\}$	0.2108	0.2199	0.0192	0.0209	0.0008	0.0009	0	0

Table 3.5: NPI lower and upper probabilities for several events $M_S \geq m_S$ with $m = 4$

For the NPI upper probability, a subset of minimal size is selected such that the NPI upper probability for the event that one or more of the future observations belong to a category in that subset is at least 0.7. The subsets meeting this criterion for all events of interest are similar to those selected for the NPI lower probability.

If the objective is to select subsets containing C_2, meeting a predefined criterion of 0.7, subsets are chosen where there exists at least a 0.7 NPI lower or upper probability that at least one of the future observations will belong to a category within that subset. The first subset which meet the required criterion based on their respective NPI lower and upper probabilities presented in the table for the event $M_S \geq 1$ is $S = \{C_2, C_3, C_4\}$. The use of Table 3.5 depends on the objective and interest in selecting the subset. For instance, suppose the objective is to select a minimal-sized subset containing a category of patients with a clinical status of non-ICU, not requiring oxygen. This subset should meet the criterion that the NPI lower or upper probability for the event, where at least half of the future observations belong to a category within that subset, is at least 0.25. Looking at Table 3.5 for the event $M_S \geq 2$, the first subset satisfying the criterion is $S = \{C_2, C_3, C_4\}$. However, as the goal is to select a minimal-size subset, we see that $S = \{C_3, C_4\}$ is the subset that meets this requirement and is therefore selected. In the event $M_S \geq 4$ with the same criterion, the first minimal-size subset meeting the required criterion is $S = \{C_1, C_2, C_3\}$.

A further interest could be in selecting a minimal-sized subset containing adjoining high-risk categories of patients that require oxygen or ICU but do not require IMV, which are C_4 and C_5. The subset should meet the criterion that the NPI lower or upper probability for the event, where at least one of the future observations belong to a category within that subset, is at least 0.60. Looking at Table 3.5 for the event $M_S \geq 1$, the minimal-size subset which satisfies the criterion is $S = \{C_3, C_4, C_5\}$. Suppose a high criterion is set for the event that all 4 future patients are in a category in that subset, with the aim to select a subset of minimal size such that the NPI lower probability for this event is at least 0.8. One might set a high criterion for this event if there is an interest in indicating a critical condition that requires immediate attention or intervention. In such a scenario, medical professionals may be able to concentrate on subsets that satisfy the criterion that the NPI lower or upper probability exceeds this high threshold, indicating a potentially serious or urgent situation. The results in Table 3.5 indicate that no strict subset has at least an 80% lower or upper probability that all of the future observations will belong to a category within that subset.

Last but not least for $m = 4$, suppose there is an interest in selecting a minimal-sized subset containing adjoining high-risk patient categories that require ICU (C_5, C_6). This selection can be particularly useful for healthcare providers needing to ensure that a certain threshold of future cases belonging to a category in the selected subset, satisfying a specified criterion, is considered for treatment planning, or preparing healthcare facilities. Assuming $p^* = 0.5$ for the event $M_S \geq 1$, the first subset that meets the required criterion is $S = \{C_5, C_6, C_7\}$. Next, to investigate how increasing m from 4 to 8 affects the optimal subset selection, the effect of this change on the minimal subset required to meet the specified criterion is presented.

Table 3.6 presents the NPI lower and upper probabilities for $m = 8$, considering the same events presented in Table 3.5 for $m = 4$. These NPI lower and upper probabilities for $M_S \geq m_S$ are derived with S consisting of a single category or a subset of categories by applying Equations (2.14) and (2.15) for the NPI lower probability and Equation (2.16) for the NPI upper probability. We notice that as m increases, both the NPI lower and upper probabilities generally increase. When comparing the selection of minimal subsets based on NPI lower probabilities for $m = 4$ and $m = 8$, different optimal subsets are selected as the increased number of observations influences both the NPI lower and upper probabilities.

For the objective of selecting subsets that contain C_1, the analysis with $m = 4$ identified $\{C_1, C_2, C_3\}$ and $\{C_1, \ldots, C_6\}$ as the optimal subsets that meet the required criterion of 0.7 for $\underline{P}(M_S \geq 3)$ and $\underline{P}(M_S \geq 4)$, respectively. However, with $m = 8$, the subset $\{C_1\}$ is the optimal for the same events, indicating that a smaller subset can meet the criterion as the number of observations increases.

S	$M_S \geq 1$		$M_S \geq 2$		$M_S \geq 3$		$M_S \geq 4$	
	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$	$\underline{P}$	$\overline{P}$
$\{C_1\}$	0.9982	0.9983	0.9814	0.9821	0.9112	0.9138	0.7416	0.7467
$\{C_1, C_2\}$	0.9985	0.9985	0.9835	0.9841	0.9187	0.9211	0.7569	0.7619
$\{C_1, C_2, C_3\}$	0.9999	1	0.9998	0.9998	0.9978	0.9980	0.9841	0.9849
$\{C_1, \ldots, C_4\}$	1	1	0.9999	1	0.9994	0.9995	0.9948	0.9951
$\{C_1, \ldots, C_5\}$	1	1	1	1	0.9998	0.9999	0.9986	0.9988
$\{C_1, \ldots, C_6\}$	1	1	1	1	1	1	0.9999	1
$\{C_2\}$	0.0427	0.0838	0.0012	0.0039	0	0.0001	0	0
$\{C_2, C_3\}$	0.8710	0.8781	0.5736	0.5882	0.2689	0.2821	0.0879	0.0945
$\{C_2, C_3, C_4\}$	0.9235	0.9280	0.6947	0.7067	0.3910	0.4049	0.1576	0.1667
$\{C_2, \ldots, C_5\}$	0.9534	0.9563	0.7817	0.7912	0.5014	0.5150	0.2367	0.2476
$\{C_2, \ldots, C_6\}$	0.9784	0.9799	0.8739	0.8802	0.6496	0.6615	0.3714	0.3839
$\{C_2 \ldots C_7\}$	0.9904	0.9908	0.9312	0.9332	0.7687	0.7734	0.5111	0.5175
$\{C_3\}$	0.8597	0.8673	0.5512	0.5662	0.2496	0.2625	0.0784	0.0847
$\{C_3, C_4\}$	0.9164	0.9212	0.6762	0.6886	0.3703	0.3841	0.1445	0.1532
$\{C_3, C_4, C_5\}$	0.9488	0.9519	0.7669	0.7768	0.4809	0.4946	0.2208	0.2314
$\{C_3, \ldots, C_6\}$	0.9761	0.9777	0.8639	0.8706	0.6314	0.6436	0.3528	0.3652
$\{C_3, \ldots, C_7\}$	0.9893	0.9897	0.9249	0.9270	0.7540	0.7589	0.4920	0.4984
$\{C_4\}$	0.3154	0.3460	0.0524	0.0636	0.0054	0.0073	0.0004	0.0006
$\{C_4, C_5\}$	0.5289	0.5510	0.1609	0.1768	0.0310	0.0361	0.0040	0.0049
$\{C_4, C_5, C_6\}$	0.7336	0.7469	0.3547	0.3718	0.1139	0.1236	0.0247	0.0278
$\{C_4, \ldots, C_7\}$	0.8517	0.8558	0.5359	0.5436	0.2369	0.2432	0.0725	0.0754
$\{C_5\}$	0.2835	0.3154	0.0420	0.0524	0.0039	0.0054	0.0002	0.0004
$\{C_5, C_6\}$	0.5824	0.6022	0.2012	0.2177	0.0444	0.0505	0.0065	0.0078
$\{C_5, C_6, C_7\}$	0.7597	0.7660	0.3889	0.3974	0.1336	0.1387	0.0311	0.0328
$\{C_6\}$	0.3897	0.4174	0.0820	0.0951	0.0108	0.0136	0.0009	0.0013
$\{C_6, C_7\}$	0.6395	0.6483	0.2514	0.2599	0.0639	0.0676	0.0109	0.0117
$\{C_7\}$	0.3755	0.3897	0.0757	0.0820	0.0095	0.0108	0.0008	0.0009

Table 3.6: NPI lower and upper probabilities for several events $M_S \geq m_S$ with $m = 8$

For the objective involving the category with the smallest data observations (C_2), the $m = 4$ analysis selected $\{C_2, C_3, C_4\}$ for $\underline{P}(M_S \geq 1) \geq 0.7$, but the $m = 8$ analysis selected $\{C_2, C_3\}$ as the optimal subset. For subsets containing C_3 (non-ICU, not requiring oxygen), the subset $\{C_3, C_4\}$ was selected for $\underline{P}(M_S \geq 2) \geq 0.25$ with $m = 4$, but $\{C_3\}$ alone was selected for $m = 8$, demonstrating that a smaller subset could be selected when considering more future observations.

With the objective concerning high-risk categories, which are C_4 and C_5, with $m = 4$, no subset met the criterion $p^* = 0.8$ for $M_S \geq 4$. However, when considering $m = 8$, the subset $\{C_1, \ldots, C_5\}$ met this criterion. Finally, for the objective involving ICU-requiring categories (C_5, C_6), where the criterion was $\underline{P}(M_S \geq 1) \geq 0.5$, the selection of $\{C_5, C_6, C_7\}$ for $M_S \geq 1$ with $m = 4$ changed to $\{C_5, C_6\}$ with $m = 8$. This reflects that a smaller subset could still meet the required criterion of 0.5 as the number of observations increased. This change illustrates that in some cases, increasing m allows for a smaller subsets to be selected due to the increase in the NPI lower and upper probabilities.

The subset selection can differ depending on whether the criterion is based on $\underline{P}(M_S \geq m_S) \geq p^*$ or $\overline{P}(M_S \geq m_S) \geq p^*$. For example, if the objective of selecting subsets containing the category that has the largest data observations is to meet a criterion of 0.1 for $\underline{P}(M_S \geq 1)$ with $m = 4$, the subset $\{C_1, C_2, C_3\}$ is selected. However, if $\overline{P}(M_S \geq 1) \geq 0.1$ is selected as the optimal subset, $\{C_1, C_2\}$ meets the criterion. In another scenario, where the objective is to select subsets containing the category with the smallest data observations and exclude the category with the largest data, with a requirement that at least 2 future observations belong to a selected subset with $m = 4$ and $p^* = 0.5$, the subset $\{C_2, \ldots, C_7\}$ is selected for $\underline{P}(M_S \geq 2) \geq 0.5$, while the subset $\{C_2, \ldots, C_6\}$ is selected for $\overline{P}(M_S \geq 2) \geq 0.5$. Furthermore, in the case where the objective is to select a minimal-sized subset containing adjoining high-risk patient categories that require ICU, which are C_5 and C_6, with $m = 8$, $p^* = 0.65$ and ensuring that at least 3 future observations belong to the subset, the selection differs again. The subset $\{C_2, \ldots, C_7\}$ is selected for $\underline{P}(M_S \geq 3) \geq 0.65$ as it is the subset that meets the criterion. However, the NPI upper probability criterion allows for different subset to be selected which is $\{C_2, \ldots, C_6\}$. These cases illustrate how the choice of $\underline{P}(M_S \geq m_S) \geq p^*$ or $\overline{P}(M_S \geq m_S) \geq p^*$ can lead to different subset selections, depending on the specific criterion and objectives, particularly in cases involving high-risk patient categories and larger numbers of future observations. $\diamond$

Overall, the comparison highlights how the choice of m affect the selection of subsets. This indicates the importance of clearly defining the objective when selecting subsets, as the chosen value of m will directly influence the results and, consequently, the decision-making process. The selection of the minimal size subset that consists of adjoining categories with a specified criterion could be of interest in different ways depending on one's belief or aim, of which subset is more important to be selected. The next section presents a different type of inference involving multiple future ordinal observations, focusing on pairwise comparison.

3.4 Pairwise comparison

This section introduces a pairwise comparison method for two groups of ordinal data. The goal is to compare the number of future observations within categories of the first group to those within the same categories of the second group. This comparison will be conducted using the sampling methodology outlined in Section 1.1.

Many statistical inference applications involve comparing two or more independent groups of data, such as those resulting from different treatments. Classical statistical methods typically involve testing the equality of parameters within assumed parametric models or utilizing rank-based approaches like Wilcoxon's or Kruskal-Wallis tests for comparing two or more independent populations [52]. These methods assume that each population's random quantities are independent and identically distributed. The fundamental distinction between comparison in NPI and classical statistical approaches lies in the formulation of the question of interest. Classical tests typically begin with the hypothesis that both groups originate from the same distribution, which may not always be practical. The NPI method uses a direct approach based solely on future observations without requiring the formulation of specific hypotheses. This enables a natural manner of comparison, particularly well-suited for making decisions, such as determining the best treatment for future units or individuals [29].

Suppose that there are two independent groups, A and B. In this context, 'independent' means that knowledge about a random quantity in one group is not influenced by information about a random quantity in the other group. Assuming the same K ordered categories for each group, the same setting and data notation from Section 2.1 are followed. Let n^a and n^b represent the number of data observations for groups A and B, respectively, with n_k^a and n_k^b indicating the observations within category C_k, for $k = 1, \ldots, K$. The analysis focuses on m future observations from each group, as it seems logical to consider the same number of future observations for each group when comparing future ordinal observations. The $A_{(\cdot)}$-based inferences are applied per group to consider m future observation from each group. So, attention is focused on the future observations from group A, represented by $X^A_{n^a+l}$ and their corresponding latent observations $Y^A_{n^a+l}$, for $l = 1, \ldots, m$. Similarly, for group B, $X^B_{n^b+l}$ and $Y^B_{n^b+l}$ denote the future observations and the corresponding latent observations, respectively.

Suppose the aim is to compare the number of future observations in $A_{C_k} = \bigcup_{j=1}^{k} C_j$ from group A, denoted as MA_{C_k}, to the number of future observations in $B_{C_k} = \bigcup_{j=1}^{k} C_j$ from group B, denoted as MB_{C_k}, where k ranges from 1 to $K-1$, by considering the events $MA_{C_k} > MB_{C_k}$ and $MA_{C_k} \geq MB_{C_k}$. By examining whether the number of future observations in one group consistently exceeds those in another across multiple categories, one can identify significant differences that might warrant further investigation or inform decision-making processes. For instance, if group A consistently shows more future observations than group B, it might indicate a higher activity level, or effectiveness of a treatment or intervention applied to group A. The comparison of future observations between two groups, A and B, based on their cumulative occurrences within ordered categories, reflect a concept in statistics known as stochastic ordering [85]. This concept can help us understand which group is likely to have more future observations. Specifically, if the number of future observations in A_{C_k} is consistently greater than the number of future observations in B_{C_k} for all k using the NPI-based lower and upper probabilities for this event, then we can say that group A 'dominates' group B in terms of future observations.

In this section, the sampling of orderings method [34], introduced in Section 1.1, will be applied to estimate the NPI lower and upper probabilities, along with corresponding confidence intervals for these estimates, for the event that the number of future observations in A_{C_k} is greater than (or equal to) the number of future observations in B_{C_k}. Section 1.1 explains that calculating exact NPI lower and upper probabilities is a computationally intensive process, requiring the evaluation of a large number of possible orderings of future observations. The sampling of orderings method offers a practical solution by allowing for the estimation of these NPI lower and upper probabilities through sampling. This approach is particularly useful when closed-form expressions are unavailable for specific events of interest, but where, for each ordering, it is easily verified if the event of interest must hold, can hold, or cannot hold.

The NPI pairwise comparison will be introduced using the sampling of orderings method for both group A and group B for the event $MA_{C_k} > MB_{C_k}$, for all $k = 1, \ldots, K-1$. Rather than analytically determining the NPI lower and upper probabilities by considering all possible orderings, a select number of orderings are sampled. To calculate estimates of NPI lower and upper probabilities for the event $MA_{C_k} > MB_{C_k}$, the following notation is introduced.

Recall that n^a and n^b represent the number of observations for groups A and B, respectively. Using the latent variable representation, explained in Chapter 2, let the n^a data observations be represented by $y^a_1 < \ldots < y^a_{n^a}$ for group A. These n^a observations divide the real-line into $n^a + 1$ intervals, $I^a_{j_a} = (y^a_{j_a-1}, y^a_{j_a})$ for $j_a = 1, \ldots, n^a + 1$. Similarly, for group B, let the n^b be represented by $y^b_1 < \ldots < y^b_{n^b}$ and these n^b divide the real-line into n^b+1 intervals, $I^b_{j_b} = (y^b_{j_b-1}, y^b_{j_b})$ for $j_b = 1, \ldots, n^b+1$.

Considering m and n^a, there are $\binom{n^a+m}{m}$ different orderings of the m future observations among the n^a data observations for group A. Each specific ordering can be represented by $\left(S^A_1, \ldots, S^A_{n^a+1}\right)$. More explicitly, $S^A_{j_a}$ is the number of the future observations in the interval $(y^a_{j_a-1}, y^a_{j_a})$ for $j_a = 1, \ldots, n^a + 1$, where the following conditions are satisfied: $S^A_{j_a} \geq 0$ and $\sum_{j_a=1}^{n^a+1} S^A_{j_a} = m$. Similarly, for group B, there are $\binom{n^b+m}{m}$ different orderings of the m future observations among the n^b data observations. Each specific ordering can be denoted by $\left(S^B_1, \ldots, S^B_{n^b+1}\right)$, where $S^B_{j_b}$ represents the number of future observations in the interval $(y^b_{j_b-1}, y^b_{j_b})$ for $j_b = 1, \ldots, n^b + 1$, with the conditions $S^B_{j_b} \geq 0$ and $\sum_{j_b=1}^{n^b+1} S^B_{j_b} = m$.

The sampling method for determining the orderings is based on simple random sampling (SRS). The process, as explained in Section 1.1, ensures that each possible ordering is equally likely to be chosen during each selection and that each selection is conducted independently of others. Using this sampling method for groups A and B, each generated ordering of the future observations for group A is characterized by $(S^A_1, \ldots, S^A_{n^a+1})$, while for group B, it is characterized by $(S^B_1, \ldots, S^B_{n^b+1})$.

For the event $MA_{C_k} > MB_{C_k}$, the focus is on K ordered categories from each group, denoted by $C_1, C_2, \ldots, C_K$, and the ordering between them is indicated by $C_1 < C_2 < \ldots < C_K$. For a given number of categories K, the number of future observations in group A will be compared to the number of future observations in group B. Specifically, the number of future observations in the first category of group A, $\left(S^A_1, \ldots, S^A_{n^a_1+1}\right)$, will be compared to the number of future observations in the first category of group B, $\left(S^B_1, \ldots, S^B_{n^b_1+1}\right)$. Additionally, for each k from 2 up to $K-1$, the number of future observations within the combined k categories of group A, represented as $\left(S^A_1, \ldots, S^A_{n^a_{1:k}+1}\right)$, will be compared to those in group B, represented as $\left(S^B_1, \ldots, S^B_{n^b_{1:k}+1}\right)$, where $n^a_{1:k} = \sum_{i=1}^{k} n^a_i$ and $n^b_{1:k} = \sum_{i=1}^{k} n^b_i$, thus quantifying the cumulative number of observations across up to $K-1$ categories for both groups. The conditions for these comparisons need to be satisfied together across the $K-1$ categories, so the logical "and" denoted by $\bigwedge$, will be used later in this section to combine all these conditions into a single expression, indicating that all of them need to hold simultaneously.

This can be done by comparing each possible ordering of m and n^a observations with each possible orderings of m and n^b observations. As the sample size increases, the number of all orderings to consider increases rapidly. To derive the NPI lower and upper probabilities with such comparison, $\binom{n^a+m}{m}\binom{n^b+m}{m}$ orderings need to be considered. For example, when $m = 10$, $n^a = 15$, and $n^b = 8$, there are $\binom{25}{10} \times \binom{18}{10} = 32687600 \times 43758 = 1.43 \times 10^{11}$ possible orderings. Calculating the NPI lower and upper probabilities becomes computationally expensive, or even impossible due to the large number of possible orderings. The sampling of orderings method offers a practical solution, as it allows for the estimation of the NPI lower and upper probabilities by sampling of orderings, reducing computational time. Therefore, estimates for the NPI lower and upper probabilities will be determined using the sampling of orderings method. Define n^* as the desired number of orderings sampled to generate for each group (i.e., $n^*_a = n^*_b = n^*$).

Let the estimates of the NPI lower and upper probabilities for the event $MA_{C_k} > MB_{C_k}$ be denoted by $\underline{P}^{>}$ and $\overline{P}^{>}$, respectively, and by $\underline{P}^{\geq}$ and $\overline{P}^{\geq}$ for the event $MA_{C_k} \geq MB_{C_k}$. Generally, in NPI, if one wants to compare the number of future observations in the intervals $(y^a_{j_a-1}, y^a_{j_a})$ for $j_a = 1, \dots, n^a + 1$ in group A to the number of future observations in the intervals $(y^b_{j_b-1}, y^b_{j_b})$ for $j_b = 1, \dots, n^b + 1$ in group B, by considering the scenario where $S^A_{j_a}$ are greater than $S^B_{j_b}$, the $\underline{P}^{>}$ is obtained by assigning all probability masses for $S^A_{j_a}$ corresponding to the intervals $(y^a_{j_a-1}, y^a_{j_a})$ to the right endpoints of these intervals, and assigning all probability masses for $S^B_{j_b}$ corresponding to the intervals $(y^b_{j_b-1}, y^b_{j_b})$ to the left endpoints of these intervals to minimise the chance of $S^A_{j_a} > S^B_{j_b}$[35]. Similarly, to obtain the $\overline{P}^{>}$, the probability masses for $S^A_{j_a}$ are assigned to the left endpoints of the intervals $(y^a_{j_a-1}, y^a_{j_a})$, and the probability masses for $S^B_{j_b}$ are assigned to the right endpoints of the intervals $(y^b_{j_b-1}, y^b_{j_b})$ to maximise the chance of $S^A_{j_a} > S^B_{j_b}$.

Assigning probability masses to the left or right endpoint of an interval means treating all future observations within an interval as occurring at the endpoint of that interval. For example, assigning the probability mass to the right endpoint $y^a_{j_a}$ of the interval $(y^a_{j_a-1}, y^a_{j_a})$ in group A means treating all future observations in that interval as if they occur at $y^a_{j_a}$. Similarly, assigning the probability mass to the left endpoint $y^b_{j_b-1}$ of the interval $(y^b_{j_b-1}, y^b_{j_b})$ in group B means treating all future observations in that interval as if they occur at $y^b_{j_b-1}$. This placement is performed to calculate the lower and upper probability estimates. This can also be performed in the case with ordered categories that partition the real-line. As there are intervals which are already within the category and there are boundary intervals, as explained in Section 2.2, the focus will only be on the boundary intervals.

For the NPI lower probability estimate $\underline{P}^{>}$, the probability mass for $S^A_{n^a_1+1}$ in group A, corresponding to the boundary interval $(y^a_{n^a_1}, y^a_{n^a_1+1})$, is placed at the right endpoint

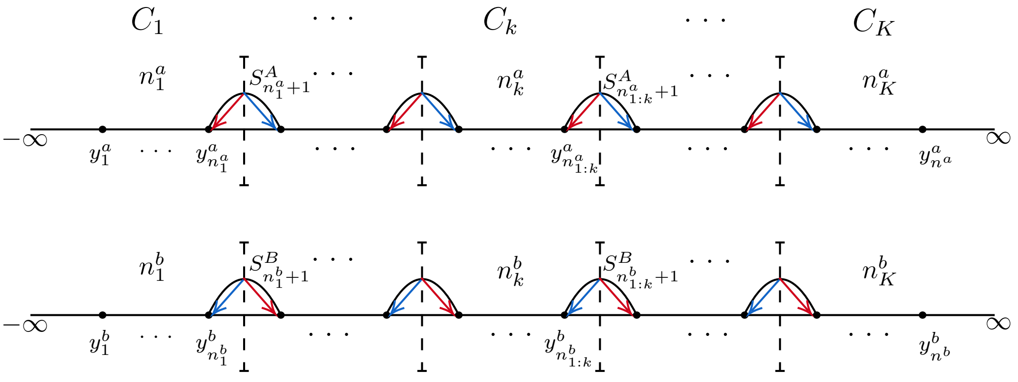


Figure 3.1: Locations of probability masses corresponding to the NPI lower and upper probabilities for $S^A_{j_a} > S^B_{j_b}$, represented by blue and red arrows respectively

of this interval. So, all $S^A_{n_1^a+1}$ future A observations in the interval $(y^a_{n_1^a}, y^a_{n_1^a+1})$ are assigned to the right endpoint $y^a_{n_1^a+1}$, meaning they are considered to belong to C_2. Similarly, for $S^B_{n_1^b+1}$ in group B, the probability mass corresponding to the interval $(y^b_{n_1^b}, y^b_{n_1^b+1})$ is placed at the left endpoint. So, all $S^B_{n_1^b+1}$ future B observations in the interval $(y^b_{n_1^b}, y^b_{n_1^b+1})$ are assigned to the left endpoint $y^b_{n_1^b}$, meaning they are considered to belong to C_1. For the NPI lower probability estimate $\underline{P}^{\geq}$, the same placement of probability masses is applied as in the $\underline{P}^{>}$ case.

For the NPI upper probability estimate $\overline{P}^{>}$, the locations of the probability masses are reversed. Here, the probability mass for $S^A_{n_1^a+1}$ is placed at the left endpoint of its interval, while for $S^B_{n_1^b+1}$ is placed at the right endpoint. Similarly, for the NPI upper probability estimate $\overline{P}^{\geq}$, the placement of probability masses is done in the same way as for $\overline{P}^{>}$. These locations of all probability masses are similarly applied to $S^A_{n^a_{1:k}+1}$ and $S^B_{n^b_{1:k}+1}$ for combined categories, with their respective boundary intervals $(y^a_{n^a_{1:k}}, y^a_{n^a_{1:k}+1})$ and $(y^b_{n^b_{1:k}}, y^b_{n^b_{1:k}+1})$. This is illustrated in Figure 3.1. This comparison is performed for n^* orderings. The cases where the event must hold are counted and divided by n^* to obtain $\underline{P}^{>}$ or $\underline{P}^{\geq}$. Similarly, the cases where the event can hold are counted and divided by n^* to obtain $\overline{P}^{>}$ or $\overline{P}^{\geq}$. Then, a 95% confidence interval will be calculated for both $\underline{P}^{>}$ and $\overline{P}^{>}$ in each replication. Confidence intervals (CI) come from the binomial properties; they are calculated using the standard result based on the Normal approximation [69],

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n^*} \tag{3.1}$$

Equation (3.1) is used to calculate confidence intervals for the estimated NPI lower and upper probabilities, where $\hat{p}$ is the estimated value of the NPI lower and upper probabilities, $z_{\alpha/2}$ is the quantile of the standard Normal distribution corresponding to the confidence level, and $\sqrt{\hat{p}(1-\hat{p})/n^*}$ is the standard error of $\hat{p}$. The parameter α determines the confidence level of the interval, which directly affects the width of the confidence interval. For example, $\alpha = 0.01$ corresponds to a 99% confidence interval, resulting in a wider interval that provides greater confidence that the interval includes the true probability. However, using a larger value of α, such as $\alpha = 0.10$, corresponds to a 90% confidence interval, which narrows the interval but reduces the confidence level, making it less likely that the interval includes the true probability. The choice of $\alpha = 0.05$ in this work, widely used in statistical inference, reflects standard practice in statistical analysis [69].

The Normal approximation assumes that n^*, the number of sampled orderings, is large enough for the binomial distribution to approximate the Normal distribution closely and that $\hat{p}$ is not near 0 or 1. However, when n^* is small or when $\hat{p}$ is close to 0 or 1, the approximation may not provide accurate coverage. In these situations, alternative methods such as the Wilson interval can be used to improve accuracy [22].

Algorithm 1 describes the calculation of $\underline{P}^{>}$ and $\overline{P}^{>}$ using the sampling method for the event $MA_{C_k} > MB_{C_k}$ with K ordered categories. The same steps can be applied to the event $MA_{C_k} \geq MB_{C_k}$ by replacing $>$ with $\geq$ in the relevant comparisons, specifically affecting Steps 3(b) and 4(b) of Algorithm 1. This modification allows the algorithm to handle the event and compute $\underline{P}^{\geq}$ and $\overline{P}^{\geq}$ using the sampling method with K ordered categories.

Algorithm 1 $\underline{P}^{>}$ and $\overline{P}^{>}$ for the event $MA_{C_k} > MB_{C_k}$ with K categories

1. **Sampling**: Sample n^* orderings for both groups A and B.
2. **Boundary intervals**: For each k from 1 to $K-1$, identify boundary intervals:
    - (a) Group A: $\left(y^a_{n^a_{1:k}}, y^a_{n^a_{1:k}+1}\right)$, with $S^A_{n^a_{1:k}+1}$ representing the number of future observations in the interval.
    - (b) Group B: $\left(y^b_{n^b_{1:k}}, y^b_{n^b_{1:k}+1}\right)$, with $S^B_{n^b_{1:k}+1}$ representing the number of future observations in the interval.
3. **Check if the event can occur**:
    - (a) **Condition setup**:
        - i. Set the locations of all probability masses for $S^A_{n^a_{1:k}+1}$ at the left endpoint of their boundary intervals $\left(y^a_{n^a_{1:k}}, y^a_{n^a_{1:k}+1}\right)$.
        - ii. Set the locations of all probability masses for $S^B_{n^b_{1:k}+1}$ at the right endpoint of their boundary intervals $\left(y^b_{n^b_{1:k}}, y^b_{n^b_{1:k}+1}\right)$.
    - (b) **Comparison**:
        - i. Let the condition in which the event can occur be represented by D_u where $D_u = \bigwedge_{k=1}^{K-1}\left(\sum_{i=1}^{n^a_{1:k}+1} S_i^A > \sum_{j=1}^{n^b_{1:k}} S_j^B\right)$.
        - ii. For each ordering, check if D_u holds and count how many of the sampled orderings satisfy this condition.
        - iii. Calculate the NPI upper probability estimate $\overline{P}^{>}$ by dividing the total number of sampled orderings that satisfy D_u by n^*.
4. **Check if the event must occur**:
    - (a) **Condition setup**:
        - i. Set the locations of all probability masses for $S^A_{n^a_{1:k}+1}$ at the right endpoint of their boundary intervals $\left(y^a_{n^a_{1:k}}, y^a_{n^a_{1:k}+1}\right)$.
        - ii. Set the locations of all probability masses for $S^B_{n^b_{1:k}+1}$ at the left endpoint of their boundary intervals $\left(y^b_{n^b_{1:k}}, y^b_{n^b_{1:k}+1}\right)$.
    - (b) **Comparison**:
        - i. Let the condition in which the event must occur be represented by $D_l = \bigwedge_{k=1}^{K-1}\left(\sum_{i=1}^{n^a_{1:k}} S_i^A > \sum_{j=1}^{n^b_{1:k}+1} S_j^B\right)$.
        - ii. For each ordering, check if D_l holds and count how many satisfy this condition.
        - iii. Calculate the NPI lower probability estimate $\underline{P}^{>}$ by dividing the total number of sampled orderings that satisfy D_l by n^*.

The events presented in this section will be illustrated using the following examples, based on data sets from the literature.

Example 3.3. The study conducted by Campochiaro et al. [23] focused on severe COVID-19 patients, comparing the effectiveness of Tocilizumab (TCZ) therapy versus standard care alone. TCZ is a monoclonal antibody that targets the interleukin-6 (IL-6) receptor, a key molecule in the body's inflammatory response. IL-6 is involved in the immune response and is known to play a significant role in the severe inflammatory reactions seen in some COVID-19 patients. A total of 65 severe COVID-19 pneumonia patients were included in the study, with 32 patients in Group A treated with TCZ according to the study protocol and 33 patients in Group B receiving only the institutional standard of care. Clinical improvement after 28 days of treatment was assessed using a 6-category ordinal scale that evaluated the patients' clinical status, where C_1 represents the best outcome: discharge from hospital, and C_6 the worst outcome: death.

The categories used to assess clinical improvement were as follows: C_1: discharged from hospital, C_2: hospitalized, not requiring supplemental oxygen, C_3: hospitalized, requiring supplemental low-flow oxygen, C_4: hospitalized, requiring high-flow oxygen and/or Non-Invasive Ventilation, C_5: hospitalized, requiring Extracorporeal Membrane Oxygenation and/or Invasive Mechanical Ventilation, and C_6: death. Table 3.7 illustrates the number of patients for each response category.

Using the sampling of orderings method, the estimates of the NPI lower and upper probabilities, together with 95% confidence intervals (CIs), are presented in Tables 3.8 and 3.9. Each table corresponds to a different value of m, representing the number of future observations considered: $m = 10$ and $m = 25$ in Tables 3.8 and 3.9, respectively. These tables illustrate how varying m affects the events $MA_{C_k} > MB_{C_k}$ and $MA_{C_k} \geq MB_{C_k}$, with sampling of orderings of size $n^* = 1000, 2000, 5000, 10000, 20000, 50000, 100000$.

Treatment Group	C_1	C_2	C_3	C_4	C_5	C_6	Total
Group A	20	2	2	3	0	5	32
Group B	16	2	2	1	1	11	33

Table 3.7: Number of patients in each clinical improvement category on day 28 for group A (Tocilizumab) and group B (Standard Care)

n^*	$\underline{P}^{>}$	CI(0.95)	$\overline{P}^{>}$	CI(0.95)	$\underline{P}^{\geq}$	CI(0.95)	$\overline{P}^{\geq}$	CI(0.95)
1000	0.4430	(0.4122 , 0.4738)	0.5480	(0.5172 , 0.5788)	0.6170	(0.5869 , 0.6471)	0.7090	(0.6808 , 0.7372)
2000	0.4415	(0.4197 , 0.4633)	0.5495	(0.5277 , 0.5713)	0.6140	(0.5927 , 0.6353)	0.7125	(0.6927 , 0.7323)
5000	0.4534	(0.4396 , 0.4672)	0.5586	(0.5448 , 0.5724)	0.6252	(0.6118 , 0.6386)	0.7224	(0.7100 , 0.7348)
10000	0.4506	(0.4408 , 0.4604)	0.5592	(0.5495 , 0.5689)	0.6255	(0.6160 , 0.6350)	0.7218	(0.7130 , 0.7306)
20000	0.4516	(0.4447 , 0.4584)	0.5569	(0.5501 , 0.5638)	0.6234	(0.6167 , 0.6302)	0.7191	(0.7129 , 0.7254)
50000	0.4525	(0.4481 , 0.4568)	0.5578	(0.5534 , 0.5621)	0.6236	(0.6193 , 0.6278)	0.7185	(0.7146 , 0.7225)
100000	0.4517	(0.4486 , 0.4548)	0.5579	(0.5548 , 0.5610)	0.6247	(0.6217 , 0.6277)	0.7189	(0.7161 , 0.7217)

Table 3.8: Estimated NPI lower and upper probabilities for the events $MA_{C_k} > MB_{C_k}$ and $MA_{C_k} \geq MB_{C_k}$, and 95% CIs with $m = 10$ and increasing values of n^*

When $m = 10$, the $\underline{P}^{>}$ for the event $MA_{C_k} > MB_{C_k}$ starts at 0.4430 with a CI of (0.4122, 0.4738) for $n^* = 1000$. As n^* increases to 20000, this value slightly increases to 0.4516 with a CI of (0.4447, 0.4584), and then becomes 0.4517 at $n^* = 100000$, with a CI of (0.4486, 0.4548). The narrowing of the CIs as n^* increases indicates precision in the estimates. For $\overline{P}^{>}$, it starts at 0.5480 with a CI of (0.5172, 0.5788) at $n^* = 1000$. As n^* increases to 20000, this value slightly increases to 0.5569 with a CI of (0.5501, 0.5638), and then becomes 0.5579 at $n^* = 100000$, with a CI of (0.5548, 0.5610). Similarly, for the event $MA_{C_k} \geq MB_{C_k}$, the $\underline{P}^{\geq}$ starts at 0.6170 with a CI of (0.5869, 0.6471) at $n^* = 1000$, slightly decreases to 0.6234 at $n^* = 20000$ with a CI of (0.6167, 0.6302), and then becomes 0.6247 at $n^* = 100000$ with a CI of (0.6217, 0.6277). Finally, the $\overline{P}^{\geq}$ starts at 0.7090 with a CI of (0.6808, 0.7372) at $n^* = 1000$, slightly decreases to 0.7191 at $n^* = 20000$ with a CI of (0.7129, 0.7254), and then becomes 0.7189 at $n^* = 100000$ with a CI of (0.7161, 0.7217).

When comparing the results for $m = 10$ with $m = 25$, increasing m leads to increased values for $\underline{P}^{>}$, $\overline{P}^{>}$, $\underline{P}^{\geq}$, and $\overline{P}^{\geq}$, as well as narrower confidence intervals. For $m = 25$, the $\underline{P}^{>}$ starts at 0.5580 with a confidence interval of (0.5272, 0.5888) at $n^* = 1000$, and decreases slightly to 0.5535 with a confidence interval of (0.5504, 0.5566) at $n^* = 100000$. Similarly, the $\overline{P}^{>}$ starts at 0.6980 with a confidence

n^*	$\underline{P}^{>}$	CI(0.95)	$\overline{P}^{>}$	CI(0.95)	$\underline{P}^{\geq}$	CI(0.95)	$\overline{P}^{\geq}$	CI(0.95)
1000	0.5580	(0.5272 , 0.5888)	0.6980	(0.6695 , 0.7265)	0.6700	(0.6409 , 0.6991)	0.7790	(0.7533 , 0.8047)
2000	0.5680	(0.5463 , 0.5897)	0.7065	(0.6865 , 0.7265)	0.6765	(0.6560 , 0.6970)	0.7855	(0.7675 , 0.8035)
5000	0.5628	(0.5491 , 0.5765)	0.6942	(0.6814 , 0.7070)	0.6584	(0.6453 , 0.6715)	0.7702	(0.7585 , 0.7819)
10000	0.5538	(0.5441 , 0.5635)	0.6931	(0.6841 , 0.7021)	0.6499	(0.6406 , 0.6592)	0.7735	(0.7653 , 0.7817)
20000	0.5500	(0.5431 , 0.5569)	0.6923	(0.6859 , 0.6987)	0.6452	(0.6385 , 0.6518)	0.7738	(0.7680 , 0.7795)
50000	0.5513	(0.5470 , 0.5557)	0.6908	(0.6868 , 0.6949)	0.6451	(0.6409 , 0.6493)	0.7716	(0.7680 , 0.7753)
100000	0.5535	(0.5504 , 0.5566)	0.6920	(0.6891 , 0.6949)	0.6464	(0.6435 , 0.6494)	0.7710	(0.7684 , 0.7736)

Table 3.9: Estimated NPI lower and upper probabilities for the events $MA_{C_k} > MB_{C_k}$ and $MA_{C_k} \geq MB_{C_k}$, and 95% CIs with $m = 25$ and increasing values of n^*

interval of (0.6695, 0.7265) at $n^* = 1000$ and decreases slightly to 0.6920 with a confidence interval of (0.6891, 0.6949) at $n^* = 100000$. For $\underline{P}^{\geq}$, it starts at 0.6700 with a confidence interval of (0.6409, 0.6991) at $n^* = 1000$ and decreases slightly to 0.6464 with a confidence interval of (0.6435, 0.6494) at $n^* = 100000$. Finally, $\overline{P}^{\geq}$ at $m = 25$ starts at 0.7790 with a confidence interval of (0.7533, 0.8047) at $n^* = 1000$ and decreases slightly to 0.7710 with a confidence interval of (0.7684, 0.7736) at $n^* = 100000$. These results show that as m increases, the values for $\underline{P}^{>}$, $\overline{P}^{>}$, $\underline{P}^{\geq}$, and $\overline{P}^{\geq}$ do increase, and the corresponding confidence intervals become narrower, indicating an increase in precision in the estimates as more orderings are sampled.

For every value of m and n^*, the $\underline{P}^{\geq}$ and $\overline{P}^{\geq}$ are larger than those for the event $MA_{C_k} > MB_{C_k}$. In other words, the equality condition consistently increases the estimates of the NPI lower and upper probabilities. The CIs narrow as n^* increases, reflecting an increase in precision in the estimates. For example, when comparing the estimates of the NPI lower and upper probabilities and their CIs for n^* values of 2000 and 100000 at $m = 25$, both tend to decrease. At $n^* = 2000$, for example, the $\underline{P}^{>}$ is 0.5680 with a CI of (0.5463, 0.5897), which decreases slightly to 0.5535 with a CI of (0.5504, 0.5566) at $n^* = 100000$. Furthermore, as m increases, there is a consistent increase in the estimates $\underline{P}^{>}$, $\underline{P}^{\geq}$, $\overline{P}^{>}$ and $\overline{P}^{\geq}$ across all n^* values.

Overall, the increase in the estimates with larger m values suggests that considering more future observations may highlight the potential for better outcomes with TCZ. As m increases to 25, the estimates of the NPI lower and upper probabilities increase, further supporting the conclusion that TCZ may lead to improved clinical outcomes compared to standard care. This analysis suggests that TCZ may be more effective than standard care alone for severe COVID-19 patients, as shown by the increasing estimates of the NPI lower and upper probabilities. $\diamond$

To gain insight into how the dataset size influences the results, the following example introduces a 7-category ordinal scale for assessing clinical improvement in critically ill influenza patients, considering unbalanced sample sizes.

Example 3.4. The study conducted by Wang et al. [90] focused on critically ill patients with influenza virus infection, comparing the effectiveness of combined Favipiravir and Oseltamivir therapy to Oseltamivir monotherapy. A total of 168 patients were included in the study, with 40 patients receiving the combination therapy and 128 patients receiving monotherapy. Clinical improvement after 14 days of treatment was assessed using a seven-category ordinal scale that evaluated the patients' respiratory function and overall recovery status. The categories used to assess clinical improvement were as follows: C_1: Not hospitalized, resumption of normal activities, C_2: Not hospitalized, unable to resume normal activities, C_3: Hospitalized, not requiring supplemental oxygen, C_4: Hospitalized, requiring supplemental oxygen, C_5: Hospitalized, requiring High-Flow Nasal Cannula and/or non-Invasive Mechanical Ventilation, C_6: Hospitalized, requiring Extracorporeal Membrane Oxygenation and/or Invasive Mechanical Ventilation, and C_7: Death. Table 3.10 illustrates the number of patients for each response category.

Using the sampling of orderings method, the estimates of the NPI lower and upper probabilities, together with 95% confidence intervals (CIs), are presented in Tables 3.11 and 3.12. Each table corresponds to a different value of m, representing the number of future observations considered: $m = 10$ and $m = 25$ in Tables 3.11 and 3.12, respectively. These tables illustrate how varying m affects the events $MA_{C_k} > (\geq)\ MB_{C_k}$, with sampling of orderings of size $n^* = 1000, 2000, 5000, 10000, 20000, 50000, 100000$.

For $m = 10$, the $\underline{P}^{>}$ and $\underline{P}^{\geq}$ in this example are smaller compared to those in Example 3.3. Specifically, $\underline{P}^{>}$ in this example is 0.2760 with a CI of (0.2483, 0.3037) for $n^* = 1000$ and 0.2760 with a CI of (0.2733, 0.2788) for $n^* = 100000$. However, Example 3.3 shows $\underline{P}^{>}$ values of 0.4430 with a CI of (0.4122, 0.4738) and 0.4517 with a CI of (0.4486, 0.4548). Similarly, $\overline{P}^{>}$ in this example is 0.3460 with a CI of (0.3165,

Treatment Group	C_1	C_2	C_3	C_4	C_5	C_6	C_7	Total
Group A	11	7	7	3	1	7	4	40
Group B	11	27	13	15	17	21	24	128

Table 3.10: Number of patients in each clinical improvement category on day 14 for group A (Favipiravir + Oseltamivir) and group B (Oseltamivir only)

n^*	$\underline{P}^{>}$	CI(0.95)	$\overline{P}^{>}$	CI(0.95)	$\underline{P}^{\geq}$	CI(0.95)	$\overline{P}^{\geq}$	CI(0.95)
1000	0.2760	(0.2483 , 0.3037)	0.3460	(0.3165 , 0.3755)	0.4990	(0.4680 , 0.5300)	0.5760	(0.5454 , 0.6066)
2000	0.2865	(0.2667 , 0.3063)	0.3535	(0.3325 , 0.3745)	0.5115	(0.4896 , 0.5334)	0.5790	(0.5574 , 0.6006)
5000	0.2806	(0.2681 , 0.2931)	0.3496	(0.3364 , 0.3628)	0.5036	(0.4897 , 0.5175)	0.5726	(0.5589 , 0.5863)
10000	0.2807	(0.2719 , 0.2895)	0.3484	(0.3391 , 0.3577)	0.5059	(0.4961 , 0.5157)	0.5805	(0.5708 , 0.5902)
20000	0.2767	(0.2705 , 0.2828)	0.3467	(0.3401 , 0.3532)	0.5054	(0.4985 , 0.5124)	0.5819	(0.5751 , 0.5887)
50000	0.2732	(0.2693 , 0.2771)	0.3412	(0.3371 , 0.3454)	0.5032	(0.4988 , 0.5076)	0.5788	(0.5745 , 0.5831)
100000	0.2760	(0.2733 , 0.2788)	0.3442	(0.3412 , 0.3471)	0.5049	(0.5018 , 0.5080)	0.5792	(0.5761 , 0.5823)

Table 3.11: Estimated NPI lower and upper probabilities for the events $MA_{C_k} > MB_{C_k}$ and $MA_{C_k} \geq MB_{C_k}$, and 95% CIs with $m = 10$ and increasing values of n^*

0.3755) and 0.3442 with a CI of (0.3412, 0.3471), while in Example 3.3, it is 0.5480 and 0.5579. The difference between the NPI upper probability estimate $\overline{P}^{>}$ and the NPI lower probability estimate $\underline{P}^{>}$ is called imprecision, and it provides insight into the link between these estimates and the amount of information available [31, 88]. In general, larger sample sizes result in smaller imprecision because more information is available. At $n^* = 1000$, the imprecision in this example is 0.0700, which is smaller than the imprecision of 0.1050 in Example 3.3. Similarly, at $n^* = 100000$, the imprecision in this example is 0.0682, which is again smaller than the imprecision of 0.1062 in Example 3.3. This reflects the fact that large numbers of observations, as in this example, lead to small imprecision.

For the event $MA_{C_k} \geq MB_{C_k}$, a similar pattern is observed for $\underline{P}^{\geq}$, with this example showing values of 0.4990 with a CI of (0.4680, 0.5300) at $n^* = 1000$ and 0.5049 with a CI of (0.5018, 0.5080) at $n^* = 100000$, while Example 3.3 shows values of 0.6170 with a CI of (0.5869, 0.6471) and 0.6247 with a CI of (0.6217, 0.6277). The imprecision decreases from 0.0770 to 0.0741 in this example and increases from 0.0920 to 0.0942 in Example 3.3. For the $\overline{P}^{\geq}$, this example showing values of 0.5760 with a CI of (0.5454, 0.6066) at $n^* = 1000$ and 0.5792 with a CI of (0.5761, 0.5823) at $n^* = 100000$. However, Example 3.3 presents larger values, with $\overline{P}^{\geq}$ starting at

n^*	$\underline{P}^{>}$	CI(0.95)	$\overline{P}^{>}$	CI(0.95)	$\underline{P}^{\geq}$	CI(0.95)	$\overline{P}^{\geq}$	CI(0.95)
1000	0.4300	(0.3993 , 0.4607)	0.5420	(0.5111 , 0.5729)	0.5530	(0.5222 , 0.5838)	0.6650	(0.6357 , 0.6943)
2000	0.4120	(0.3904 , 0.4336)	0.5210	(0.4991 , 0.5429)	0.5395	(0.5177 , 0.5613)	0.6440	(0.6230 , 0.6650)
5000	0.4108	(0.3972 , 0.4244)	0.5184	(0.5046 , 0.5322)	0.5414	(0.5276 , 0.5552)	0.6420	(0.6287 , 0.6553)
10000	0.4172	(0.4075 , 0.4269)	0.5208	(0.5110 , 0.5306)	0.5450	(0.5352 , 0.5548)	0.6483	(0.6389 , 0.6577)
20000	0.4090	(0.4022 , 0.4158)	0.5185	(0.5116 , 0.5255)	0.5416	(0.5347 , 0.5485)	0.6462	(0.6395 , 0.6528)
50000	0.4101	(0.4058 , 0.4145)	0.5187	(0.5144 , 0.5231)	0.5413	(0.5370 , 0.5457)	0.6457	(0.6415 , 0.6499)
100000	0.4106	(0.4075 , 0.4136)	0.5187	(0.5156 , 0.5218)	0.5421	(0.5391 , 0.5452)	0.6454	(0.6424 , 0.6483)

Table 3.12: Estimated NPI lower and upper probabilities for the events $MA_{C_k} > MB_{C_k}$ and $MA_{C_k} \geq MB_{C_k}$, and 95% CIs with $m = 25$ and increasing values of n^*

0.7090 with a CI of (0.6808, 0.7372) and slightly decreasing to 0.7189 with a CI of (0.7161, 0.7217). The imprecision decreases from 0.0306 to 0.0031 in this example, while in Example 3.3, decreasing from 0.0564 to 0.0056.

For $m = 25$, the $\underline{P}^{>}$ in this example is 0.4300 with a CI of (0.3993, 0.4607) at $n^* = 1000$ and 0.4106 with a CI of (0.4075, 0.4136) at $n^* = 100000$. Example 3.3 shows larger values, from 0.5580 with a CI of (0.5272, 0.5888) to 0.5535 with a CI of (0.5504, 0.5566). The $\overline{P}^{>}$ in this example is 0.5420 with a CI of (0.5111, 0.5729) and 0.5187 with a CI of (0.5156, 0.5218), while in Example 3.3, it is 0.7065 and 0.6920. The imprecision in this example decreases from 0.1120 at $n^* = 1000$ to 0.1081 at $n^* = 100000$, compared to a decrease from 0.1400 to 0.1385 in Example 3.3. The $\overline{P}^{\geq}$ in this example begins at 0.6650 with a CI of (0.6357, 0.6943) at $n^* = 1000$ and slightly decreases to 0.6454 with a CI of (0.6424, 0.6483) at $n^* = 100000$. In Example 3.3, the corresponding $\overline{P}^{\geq}$ values are consistently larger, starting at 0.7790 with a CI of (0.7533, 0.8047) and decreasing to 0.7710 with a CI of (0.7684, 0.7736).

A similar pattern is observed for the $\underline{P}^{\geq}$, where this example shows values of 0.5530 with a CI of (0.5222, 0.5838) at $n^* = 1000$ and 0.5421 with a CI of (0.5391, 0.5452) at $n^* = 100000$, while Example 3.3 has values of 0.6700 with a CI of (0.6409, 0.6991) and 0.6464 with a CI of (0.6435, 0.6494). The imprecision in this example decreases from 0.1120 to 0.1031, and in Example 3.3 from 0.1090 to 0.1046, indicating patterns of reduction in imprecision for larger datasets. $\diamond$

Overall, the comparison between the two studies shows that treatment with TCZ for severe COVID-19 patients in Example 3.3 generally has larger estimates compared to the combined Favipiravir and Oseltamivir therapy for influenza patients in Example 3.4. In both examples, as the number of future observations increases, the estimates also increase, which may indicate better expected outcomes with larger m. When comparing the imprecision, it is clear from the results that the effect of increasing the sample size leads to decreasing the imprecision, because more information is available.

3.5 Concluding remarks

This chapter presented NPI for ordinal data in selection problems involving selecting a specific category or a subset of categories based on m future observations. Pairwise comparison of future observations from two independent groups is also presented. The methods were illustrated and discussed via examples with data from the literature.

One objective was to present the selection of a specific category, achieving a specified criterion for the event that at least m_k of the m future observations are in that category. The results showed that the category fulfilling this criterion varied depending on the number of future observations taken into account. Another aim was to present the selection of a minimal-sized subset of adjoining categories, achieving a specified criterion for the event that at least a certain number of future observations within that subset meet the criterion. The results indicated that the selected subset could change based on the number of future observations considered within that subset. The idea of subset selection can be further developed. An example of this would be to develop a method for selecting a subset with non-adjoining categories. The derivation of the NPI lower and upper formulae for this method requires consideration of whether the first and last categories are included in the subset, as well as all pairs of neighbouring categories included in the subset. This is left as a future research topic. The chapter also presented the NPI pairwise comparison using the sampling of orderings method, highlighting the influence of the data size and the number of future observations on the results.

Chapter 4

Optimal threshold selection in two-group classification

4.1 Introduction

Measuring the accuracy of diagnostic tests is crucial in many application areas, including medicine, machine learning, and credit scoring. In medical applications, to completely define a diagnostic test and accurately assess its performance, it is required to select an appropriate threshold for classifying whether a patient is diseased or healthy based on their diagnostic test results [38, 40]. Test results may take two values (binary tests), real values (continuous tests), or values within a finite number of categories which can be ordered or not; if ordered the test outcome is ordinal. Diagnostic tests that yield ordinal results are the focus of this chapter. The optimal diagnostic test threshold will be selected such that the categories on the right side of the threshold indicate disease, and the categories on the left side indicate non-disease.

Diagnostic test accuracy is determined by selecting the optimal threshold that can distinguish between diseased and healthy individuals. The specificity and sensitivity, defined in Section 1.4, of a test have an inverse relationship [67]. This means that adjusting the threshold to increase one will decrease the other. A test with higher specificity generally has lower sensitivity, and vice versa. Misclassification can occur in two ways, healthy individuals may be classified as diseased, and diseased individuals may be classified as healthy. It is ideal to choose a threshold based on the relative importance of correctly diagnosing one group over another. In the literature, researchers have used the utility concept when choosing the optimal threshold [57]. For example, Hand [57] discussed selecting the optimal threshold in scenarios where misclassifying a diseased individual as healthy is considered a more significant error than misclassifying a healthy person as diseased, or vice versa.

This chapter introduces two NPI-based methods for selecting the optimal diagnostic test threshold in two-group classification settings, considering inference based on multiple future individuals. The first method is based on the product of the NPI lower or upper probabilities of correct classification for both groups, while the second method is based on the sum of these lower or upper probabilities. Criteria based on each group's target proportion of successful diagnoses are presented to reflect the relative importance of correct classification of members of one group over members of the other group.

This chapter is organised as follows. In Section 4.2, a brief introduction to diagnostic tests for two groups of ordinal data is given. Section 4.3 presents NPI for selecting the optimal threshold for two-group diagnostic tests considering a fixed number of multiple future individuals per group and is based on maximising the product of the NPI lower or upper probabilities of the correct classification for both groups. In Section 4.4, an NPI method inspired by the Youden index is presented, in the sense that the criterion maximises the sum of the NPI lower or upper probabilities of correct classification with multiple future individuals. The practical application of the two proposed NPI methods to a real dataset is provided in Section 4.5, followed by an investigation of their predictive performance via simulations in Section 4.6, alongside comparisons with the classical methods. Finally, some concluding remarks are given in Section 4.7.

4.2 Diagnostic tests for two groups

In a two-group classification study, the objective is to assess how well a diagnostic test can distinguish between individuals who have the disease and those who do not.

To measure the accuracy of a diagnostic test, it is imperative to determine an optimal threshold for identifying positive and negative results. Alabdulhadi [4] and Coolen-Maturi et al. [40] introduced NPI to determine the optimal threshold for two-group classification problems based on tests that yield real-valued results for a given number of future observations from each group. Elkhafifi and Coolen [47, 48] presented NPI for the accuracy of diagnostic tests with ordinal outcomes for a disease group and a non-disease group based on a single future observation. In this chapter, NPI-based methods for selecting the optimal threshold for two-group diagnostic tests with ordinal outcomes are presented, with inferences based on multiple future individuals.

Diagnostic tests with ordinal outcomes appear in many medical applications and other areas [2, 14]. In this chapter, diagnostic tests with ordinal results are considered. This means that each individual's test outcome indicates one of $K \geq 2$ ordered categories, denoted by C_1 to C_K, representing an increasing severity level related to their indication of having the condition of interest [47, 48].

Suppose that ordinal data from a diagnostic test are available on individuals categorized into two groups based on their disease status, where G^0 indicates the absence of the disease ('healthy group') and G^1 indicates the presence of the disease ('disease group'). Throughout this chapter, the healthy (disease) group is indicated by superscript 0 (1), and it is assumed that these two groups are fully independent, meaning that any information regarding one group does not provide any information about the other group. Table 4.1 provides notation for the number of individuals for each combination of condition status and test result. Throughout this chapter, the definitions and notation are similar to those used by Alabdulhadi [4], Coolen-Maturi et al. [40], and Elkhafifi and Coolen [47, 48].

Assume that there is a threshold $k \in \{1, \ldots, K\}$, such that a test result in categories $\{C_1, \ldots, C_k\}$ indicates an absence of the disease, called a negative test result, while a test result in categories $\{C_{k+1}, \ldots, C_K\}$ indicates a presence of the disease, called a positive test result [91, 96]. A main goal for statistical inference in this scenario is the study of the best choice for the value k, referred to as the 'optimal threshold' k'.

Condition status	Diagnostic test result						Total
	C_1	$\dots$	C_k	C_{k+1}	$\dots$	C_K	
G^0	n_1^0	$\dots$	n_k^0	n_{k+1}^0	$\dots$	n_K^0	n^0
G^1	n_1^1	$\dots$	n_k^1	n_{k+1}^1	$\dots$	n_K^1	n^1

Table 4.1: Ordinal test data

As the NPI-based inferences are in terms of future observations, the optimal threshold k' will be selected based on the number of future observations considered from each group. This raises the question of how to choose the value of k that maximises the correct classification of diseased and healthy individuals. To this end, the next section introduces the first NPI method for selecting the optimal threshold k' for two-group diagnostic tests.

4.3 Optimal threshold selection for two groups diagnostic tests

Assume that ordinal data from a diagnostic test are available on individuals from two groups, and that there are n^0 observations from the healthy group G^0 and n^1 observations from the disease group G^1. Let the number of future individuals from the healthy group be denoted by m^0, with diagnostic test results $T^0_{n^0+q}$ for $q = 1, \dots, m^0$ and let the number of future individuals from the disease group be denoted by m^1, with diagnostic test results $T^1_{n^1+v}$ for $v = 1, \dots, m^1$. The value k that yields the best correct classification is selected based on the m^0 and m^1 future individuals since the NPI-based inferences are based on future observations. To this end, the NPI results presented in Section 2.4 will be used, but new notation needs to be introduced first.

For a specific value of a threshold $k \in \{1, \dots, K\}$, let W_k^0 denote the number of correctly classified future individuals from the healthy group, that is those with test results $T^0_{n^0+q} \in \bigcup_{j=1}^{k} C_j$ for $q = 1, \dots, m^0$, and let W_k^1 denote the number of correctly classified future individuals from the disease group, that is those with test results $T^1_{n^1+v} \in \bigcup_{j=k+1}^{K} C_j$ for $v = 1, \dots, m^1$. Assume that α and β are two values in $(0, 1)$ that are selected to reflect the relative importance of correct classification of members of one group over members of the other group. The event of interest that will be considered here is that the number of correctly classified future individuals from the healthy group is at least αm^0 and the number of correctly classified future individuals from the disease group is at least βm^1, so $W_k^0 \geq \alpha m^0$ and $W_k^1 \geq \beta m^1$.

Depending on one's perspective regarding the relative importance of accurately diagnosing the two groups, values of α and β can be varied to gain intuitive insight [4]. For example, if giving medication to diseased patients is crucial, then it might be appropriate to assign more weight to the correct classification of the diseased group than to the healthy group. In this scenario, the proportion of correctly diagnosed patients as diseased is expected to increase, while the proportion of correctly classified healthy individuals is expected to decrease. There is, of course, the option of setting α and β to be equal if one prefers to give the same importance to both groups in regard to correct classifications of future individuals. It should be noted that α and β represent target proportions per group, and their values are not constrained, with the exception of falling within the range of $(0, 1)$.

Considering that the two groups are assumed to be independent, the joint NPI lower and upper probabilities for the events $W_k^0 \geq \alpha m^0$ and $W_k^1 \geq \beta m^1$ can be derived as the products of the corresponding NPI lower and upper probabilities for the individual events,

$$\underline{P}\big(W_k^0 \geq \alpha m^0, W_k^1 \geq \beta m^1\big) = \underline{P}\big(W_k^0 \geq \alpha m^0\big) \times \underline{P}\big(W_k^1 \geq \beta m^1\big) \tag{4.1}$$

$$\overline{P}\big(W_k^0 \geq \alpha m^0, W_k^1 \geq \beta m^1\big) = \overline{P}\big(W_k^0 \geq \alpha m^0\big) \times \overline{P}\big(W_k^1 \geq \beta m^1\big) \tag{4.2}$$

The NPI lower and upper probabilities for selecting the optimal diagnostic test threshold in two-group classification settings with ordinal data, given in Equations (4.1) and (4.2), will be denoted as NPI-2G-L and NPI-2G-U, respectively. The method in general will be referred to as NPI-2G.

The NPI results provided in Section 2.4, particularly Equations (2.14) and (2.16), will be used to derive the NPI-2G-L and NPI-2G-U. First, the results for the healthy group will be presented and then those for the disease group. Let $n_{1:k}^0 = \sum\limits_{i=1}^{k} n_i^0$ and let $\lceil \alpha m^0 \rceil$ denote the smallest integer greater than or equal to αm^0. The NPI lower and upper probabilities for the event $W_k^0 \geq \alpha m^0$ with $k \in \{1, \ldots, K\}$ are given by

$$\underline{P}\big(W_k^0 \geq \alpha m^0\big) = \binom{n^0+m^0}{m^0}^{-1} \sum_{r=\lceil \alpha m^0 \rceil}^{m^0} \binom{n_{1:k}^0 - 1 + r}{r} \binom{n^0 - n_{1:k}^0 + m^0 - r}{m^0 - r} \tag{4.3}$$

$$\overline{P}\big(W_k^0 \geq \alpha m^0\big) = \binom{n^0+m^0}{m^0}^{-1} \times \Bigg[\binom{n_{1:k}^0 + \lceil \alpha m^0 \rceil}{\lceil \alpha m^0 \rceil} \binom{n^0 - n_{1:k}^0 + m^0 - \lceil \alpha m^0 \rceil}{m^0 - \lceil \alpha m^0 \rceil} + \sum_{s=\lceil \alpha m^0 \rceil + 1}^{m^0} \binom{n_{1:k}^0 - 1 + s}{s} \binom{n^0 - n_{1:k}^0 + m^0 - s}{m^0 - s} \Bigg] \tag{4.4}$$

Similarly, the NPI lower and upper probabilities are derived for the disease group, for the event that the number of correctly classified future individuals from the disease group is at least βm^1. With $n_{k+1:K}^1 = \sum_{j=k+1}^{K} n_j^1$ and $\lceil \beta m^1 \rceil$ the smallest integer greater than or equal to βm^1, the NPI lower and upper probabilities for the event $W_k^1 \geq \beta m^1$ with $k \in \{1, \ldots, K\}$ are given by

$$\underline{P}\big(W_k^1 \geq \beta m^1\big) = \binom{n^1+m^1}{m^1}^{-1} \sum_{r=\lceil \beta m^1 \rceil}^{m^1} \binom{n_{k+1:K}^1 - 1 + r}{r} \binom{n^1 - n_{k+1:K}^1 + m^1 - r}{m^1 - r} \tag{4.5}$$

$$\overline{P}\big(W_k^1 \geq \beta m^1\big) = \binom{n^1+m^1}{m^1}^{-1} \times \Bigg[\binom{n_{k+1:K}^1 + \lceil \beta m^1 \rceil}{\lceil \beta m^1 \rceil} \binom{n^1 - n_{k+1:K}^1 + m^1 - \lceil \beta m^1 \rceil}{m^1 - \lceil \beta m^1 \rceil} + \sum_{s=\lceil \beta m^1 \rceil + 1}^{m^1} \binom{n_{k+1:K}^1 - 1 + s}{s} \binom{n^1 - n_{k+1:K}^1 + m^1 - s}{m^1 - s} \Bigg] \tag{4.6}$$

Using Equations (4.3) and (4.5), the NPI-2G-L can be derived. Similarly, Equations (4.4) and (4.6) will be applied to derive the NPI-2G-U. The optimal diagnostic threshold k' is selected by maximisation of Equation (4.1) for the NPI-2G-L or Equation (4.2) for the NPI-2G-U. Note that the NPI-2G-L and NPI-2G-U are different criteria which means they may yield different optimal thresholds.

In the next section, NPI-based inference related to the two-group Youden index considering a fixed number of multiple future individuals per group is introduced. Unlike the methodology presented in this section, which focuses on the product of the

NPI lower or upper probabilities of correct classification, the next section introduces a NPI method based on the sum of the NPI lower or upper probabilities. In Section 4.5, an example is presented to illustrate the proposed NPI-based methods using data from the literature.

4.4 Optimal threshold selection for two groups Youden index

The NPI method for two-group classification has been developed with a focus on continuous diagnostic tests for multiple future individuals, inspired by the sum-based approach of the Youden index [4, 40], and for ordinal outcomes with a single future individual [47, 48]. This section introduces an NPI-based method for two-group classification with ordinal outcomes for multiple future individuals, also inspired by the sum-based approach of the Youden index.

The NPI results presented in Section 4.3, in particular Equations (4.3), (4.4), (4.5), and (4.6), are applied using the idea of maximising the sum of the probabilities of the correct classification for the two groups, similar to the Youden index method. While the classical Youden index method does not use target proportions or m values, this NPI-based method maximises the sum of the lower and upper probabilities to determine the optimal threshold for the two groups using these values. This optimal threshold is the point at which the sum of the NPI lower or upper probabilities of correct classification is maximised. Let the approaches that use the sum of the NPI lower and upper probabilities for the two-group classification, inspired by the Youden index, be denoted by NPI-2G-Y-L and NPI-2G-Y-U, respectively. The method in general will be referred to as NPI-2G-Y. The NPI-2G-Y-L and NPI-2G-Y-U are given by

$$\text{NPI-2G-Y-L} = \underline{P}\big(W_k^0 \geq \alpha m^0\big) + \underline{P}\big(W_k^1 \geq \beta m^1\big) - 1 \tag{4.7}$$

$$\text{NPI-2G-Y-U} = \overline{P}\big(W_k^0 \geq \alpha m^0\big) + \overline{P}\big(W_k^1 \geq \beta m^1\big) - 1 \tag{4.8}$$

The NPI-2G-Y-L and NPI-2G-Y-U are derived as explained in Section 4.3. To determine the optimal diagnostic threshold, the NPI-2G-Y-L in Equation (4.7) or the NPI-2G-Y-U in Equation (4.8) is maximised. The NPI-2G-Y-L and NPI-2G-Y-U can result in different optimal thresholds. The methods introduced in Sections 4.3 and 4.4, namely the NPI-2G and NPI-2G-Y, will be illustrated by an example in the following section.

4.5 Example of optimal threshold selection

This section introduces a detailed example using a dataset from the literature to illustrate the NPI-2G and NPI-2G-Y methods. Additionally, the example presents the optimal thresholds obtained from the classical methods, including the Youden index and the Liu index methods, explained in Section 1.4.

Table 4.2 presents the test results of a study with outcomes on an ordinal scale with five categories, conducted to assess the accuracy of "Cine" MRI (magnetic resonance imaging) for the detection of thoracic aortic dissection [96]. The study used the following confidence scale: 1: definitely not dissection, 2: probably not dissection, 3: possible dissection, 4: probable dissection, and 5: definite dissection.

In the literature, researchers have discussed the choice of α and β, as mentioned in Section 4.3. Consider a scenario where a diagnostic decision needs to be made, and the outcomes are classified on an ordinal scale. One might prefer to set equal values of α and β and apply the NPI-2G method to balance the importance of avoiding misdiagnosis for both healthy and diseased individuals. However, in situations where the treatment has severe side effects for diseased individuals, but a misdiagnosis of a healthy person as diseased only leads to moderate consequences, one might prefer to focus on correctly classifying more diseased individuals by choosing a larger value for β. Any weighting regarding the importance of misdiagnosis should be reflected in the choice of the target proportions in the two methods. It is important to note that the choices of α and β are crucial because they directly influence the NPI lower and upper probabilities in the NPI-2G-Y method. Poor choices of α and β can lead to very small values for the NPI lower and upper probabilities, which may cause the

Dissection status	Diagnostic test results					Total
	1	2	3	4	5	
Absent (G^0)	39	19	9	1	1	69
Present (G^1)	7	7	3	5	23	45

Table 4.2: The test results

NPI-2G-Y-L or NPI-2G-Y-U to be less than zero. Therefore, careful consideration must be given to selecting α and β. In this example, the NPI-based methods are applied to select the optimal threshold k' based on different values of α, β, and $m^0 = m^1 = m$. We are interested in determining how the optimal threshold may vary in relation to these values.

Table 4.3 is introduced to illustrate the derivation of the results presented in Table 4.4 for the NPI-2G and NPI-2G-Y methods. Table 4.3 displays, for just one case, $\alpha = \beta = 0.8$ with $m^0 = m^1 = m = 5$, the NPI lower and upper probabilities of correct classification for the event $W_k^0 \geq \alpha m^0$, presented in Equations (4.3) and (4.5) and for the event $W_k^1 \geq \beta m^1$, presented in Equations (4.4) and (4.6), along with the results of the NPI-2G and NPI-2G-Y methods. For each possible threshold k, Equations (4.3) and (4.5) are used to derive the NPI-2G-L and NPI-2G-Y-L, while Equations (4.4) and (4.6) are applied to derive the NPI-2G-U and NPI-2G-Y-U.

In Table 4.3, the NPI lower probability for the event $W_k^0 \geq \alpha m^0$ is 0.2752, whereas for the event $W_k^1 \geq \beta m^1$, it is 0.7839. Consequently, the NPI-2G-L is 0.2157. However, the NPI-2G-Y-L is much smaller; it is equal to 0.0590. This difference arises because the NPI-2G-Y method, unlike the NPI-2G, is based on the sum of the NPI lower or upper probabilities for correct classification rather than their product. The NPI-2G-Y method can sometimes yield very small or even negative values, making it challenging to achieve higher target proportions of correctly classified individuals. Moreover, the NPI-2G-Y method may lead to unbalanced classification rates because with $k = 1$, the $\underline{P}(W_k^1 \geq \beta m^1)$ and $\overline{P}(W_k^1 \geq \beta m^1)$ are large, while they are small for the event $W_k^0 \geq \alpha m^0$.

Table 4.4 gives all possible thresholds and corresponding values of NPI-2G-L, NPI-2G-U, NPI-2G-Y-L, and NPI-2G-Y-U, with $m = 5$. There have been different scenarios considered for α and β. It should be noted that the optimal threshold

	k				
	1	2	3	4	5
$\underline{P}(W_k^0 \geq \alpha m^0)$	0.2752	0.7905	0.9790	0.9892	1
$\overline{P}(W_k^0 \geq \alpha m^0)$	0.2972	0.8173	0.9892	0.9963	1
$\underline{P}(W_k^1 \geq \beta m^1)$	0.7839	0.4815	0.3624	0.2004	0
$\overline{P}(W_k^1 \geq \beta m^1)$	0.8241	0.5239	0.4005	0.2286	0
NPI-2G-L= $\underline{P}(W_k^0 \geq \alpha m^0) \times \underline{P}(W_k^1 \geq \beta m^1)$	0.2157	**0.3806**	0.3547	0.1982	0
NPI-2G-U= $\overline{P}(W_k^0 \geq \alpha m^0) \times \overline{P}(W_k^1 \geq \beta m^1)$	0.2450	**0.4282**	0.3962	0.2278	0
NPI-2G-Y-L= $\underline{P}(W_k^0 \geq \alpha m^0) + \underline{P}(W_k^1 \geq \beta m^1) - 1$	0.0590	0.2720	**0.3414**	0.1896	0
NPI-2G-Y-U= $\overline{P}(W_k^0 \geq \alpha m^0) + \overline{P}(W_k^1 \geq \beta m^1) - 1$	0.1214	0.3412	**0.3897**	0.2249	0

Table 4.3: Corresponding value of NPI-2G-L, NPI-2G-U, NPI-2G-Y-L, and NPI-2G-Y-U with $m = 5$ and $\alpha{=}\beta = 0.8$

selection for the NPI-2G and NPI-2G-Y methods depends on the values of α and β, as well as the value of m, and may vary based on the considered scenario. As shown in Table 4.4, both NPI-based methods provide the same optimal threshold for $\alpha = \beta = 0.4$. Similarly, when α and β values are increased, both NPI-based methods yield the same optimal threshold for $\alpha = \beta = 0.6$. However, when α and β are increased further to 0.8, the optimal thresholds for the NPI-2G and NPI-2G-Y methods differ. The optimal threshold that maximises both NPI-2G-L and NPI-2G-U is $k' = 2$, while the optimal threshold that maximises both NPI-2G-Y-L and NPI-2G-Y-U is $k' = 3$. As a result, for the NPI-2G-Y method, the optimal diagnostic test is such that an outcome in categories C_4 and C_5 indicates disease while a result in categories C_1 to C_3 indicates no disease, whereas for the NPI-2G method, the optimal diagnostic test should result in a diagnosis of disease in categories C_3 to C_5 and no disease in categories C_1 and C_2.

The optimal threshold also changes when α and β are set at different values. For instance, when $\alpha = 0.6$ and $\beta = 0.3$, the optimal threshold is $k' = 2$ for NPI-2G, but the NPI-2G-Y method yields different optimal thresholds, with NPI-2G-Y-L at $k' = 3$ and NPI-2G-Y-U at $k' = 2$. However, when $\alpha = 0.5$ and $\beta = 0.8$, the optimal thresholds decrease to $k' = 1$ for NPI-2G and to $k' = 2$ for NPI-2G-Y-L. It is clear from the scenario with $\alpha = 0.5$, $\beta = 0.8$ that the optimal thresholds for the NPI-2G

	NPI-2G method		NPI-2G-Y method	
k	NPI-2G-L	NPI-2G-U	NPI-2G-Y-L	NPI-2G-Y-U
$\alpha = \beta = 0.4$				
1	0.8621	0.8775	0.8613	0.8770
2	**0.9448**	**0.9563**	**0.9446**	**0.9562**
3	0.9084	0.9241	0.9084	0.9241
4	0.7996	0.8257	0.7996	0.8257
5	0	0	0	0
$\alpha = \beta = 0.6$				
1	0.5743	0.6066	0.5551	0.5936
2	**0.7559**	**0.7915**	**0.7467**	**0.7852**
3	0.6890	0.7247	0.6886	0.7245
4	0.4997	0.5390	0.4994	0.5389
5	0	0	0	0
$\alpha = \beta = 0.8$				
1	0.2157	0.2450	0.0590	0.1214
2	**0.3806**	**0.4282**	0.2720	0.3412
3	0.3547	0.3962	**0.3414**	**0.3897**
4	0.1982	0.2278	0.1896	0.2249
5	0	0	0	0
$\alpha = 0.6, \beta = 0.3$				
1	0.5996	0.6261	0.5971	0.6245
2	**0.9081**	**0.9261**	0.9059	**0.9247**
3	0.9071	0.9235	**0.9070**	0.9235
4	0.7991	0.8255	0.7990	0.8255
5	0	0	0	0
$\alpha = 0.5, \beta = 0.8$				
1	**0.4730**	**0.5181**	0.3873	0.4528
2	0.4604	0.5054	**0.4377**	**0.4886**
3	0.3618	0.4003	0.3609	0.3999
4	0.2003	0.2286	0.1998	0.2285
5	0	0	0	0

Table 4.4: Selecting the optimal threshold using the NPI-based methods with $m = 5$ and different values of α and β

and NPI-2G-Y methods decrease compared to the scenario with $\alpha = 0.6$, $\beta = 0.3$, as this scenario with $\alpha = 0.5$, $\beta = 0.8$ requests to put more weight on the number of correctly classified future individuals from the disease group over those from the healthy group.

When the number of future observations increases to $m = 10$, the optimal thresholds for the NPI-2G and NPI-2G-Y methods differ compared to when $m = 5$. Table 4.5 presents all possible thresholds and the corresponding values of NPI-2G-L, NPI-2G-U, NPI-2G-Y-L, and NPI-2G-Y-U, with $m = 10$. The scenarios for α and β

	NPI-2G method		NPI-2G-Y method	
k	NPI-2G-L	NPI-2G-U	NPI-2G-Y-L	NPI-2G-Y-U
$\alpha = \beta = 0.4$				
1	0.8898	0.9059	0.8897	0.9058
2	**0.9711**	**0.9794**	**0.9711**	**0.9793**
3	0.9337	0.9492	0.9337	0.9492
4	0.8040	0.8375	0.8040	0.8375
5	0	0	0	0
$\alpha = \beta = 0.6$				
1	0.5069	0.5469	0.4925	0.5383
2	**0.7619**	**0.8059**	**0.7565**	**0.8026**
3	0.6436	0.6923	0.6436	0.6923
4	0.3886	0.4388	0.3885	0.4388
5	0	0	0	0
$\alpha = \beta = 0.8$				
1	0.0920	0.1136	-0.1326	-0.0583
2	**0.2550**	**0.3087**	0.0920	0.1815
3	0.2047	0.2454	**0.1943**	**0.2411**
4	0.0729	0.0924	0.0676	0.0910
5	0	0	0	0
$\alpha = 0.6, \beta = 0.3$				
1	0.5227	0.5577	0.5226	0.5577
2	0.9688	0.9775	0.9687	0.9774
3	**0.9815**	**0.9868**	**0.9815**	**0.9868**
4	0.9267	0.9427	0.9266	0.9427
5	0	0	0	0
$\alpha = 0.5, \beta = 0.8$				
1	**0.5501**	**0.6136**	**0.4834**	**0.5670**
2	0.3365	0.3887	0.3330	0.3864
3	0.2074	0.2468	0.2074	0.2468
4	0.0733	0.0925	0.0733	0.0925
5	0	0	0	0

Table 4.5: Selecting the optimal threshold using the NPI-based methods with $m = 10$ and different values of α and β

considered here are similar to those presented in Table 4.4. Based on the comparison of Tables 4.4 and 4.5, the two NPI-based methods provide different optimal thresholds in some cases. For example, with $\alpha = 0.6$ and $\beta = 0.3$, the optimal threshold shifts compared to $m = 5$. The NPI-2G method now selects $k' = 3$ as the optimal threshold compared to $k' = 2$ when $m = 5$. The NPI-2G-Y method yields same optimal thresholds with $k' = 3$ for NPI-2G-Y-L and NPI-2G-Y-U compared to $k' = 3$ and $k' = 2$ when $m = 5$. Also, when $\alpha = \beta = 0.8$, the optimal threshold now is $k' = 1$ for the NPI-2G-Y compared to $k' = 2$ when $m = 5$.

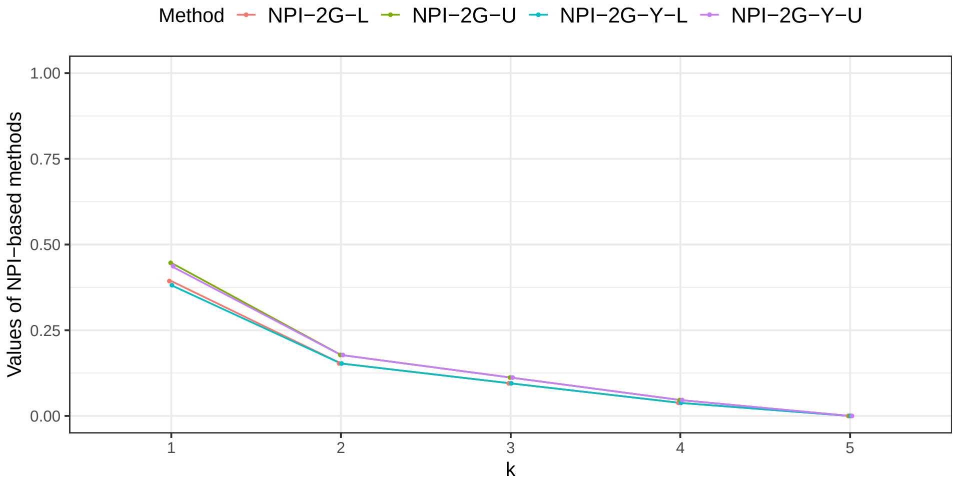


Figure 4.1: All possible thresholds and corresponding value of NPI-2G-L, NPI-2G-U, NPI-2G-Y-L, and NPI-2G-Y-U with $m = 5$ and $\alpha = 0.1, \beta = 0.9$

It is important to note that the NPI-2G-Y method may yield values lower than zero, as seen in Table 4.5 with $\alpha = \beta = 0.8$, when compared to the NPI-2G method. This is due to the very small NPI lower and upper probabilities obtained from Equations (4.3)–(4.6) when considering large target proportions. This is in line with what we have discussed earlier in this section regarding the NPI-2G-Y method, which, based on the sum of the NPI lower probabilities for correct classification or the sum of the NPI upper probabilities rather than their product, can sometimes yield very small or even negative values.

Figure 4.1 shows thresholds and values for NPI-2G-L, NPI-2G-U, NPI-2G-Y-L, and NPI-2G-Y-U with $\alpha = 0.1$ and $\beta = 0.9$ for different values of m. The optimal thresholds are less than those for the corresponding cases with $\alpha = \beta$. This decrease is due to the higher weight given to correctly classifying diseased individuals. However, when more weight is given to the number of correctly classified future individuals from the healthy group than from the diseased group ($\alpha = 0.9$ and $\beta = 0.1$), the optimal thresholds for the NPI-2G and NPI-2G-Y methods are greater than those for the corresponding cases with $\alpha = \beta$ and $\alpha = 0.1$ and $\beta = 0.9$, as seen in Figure 4.2. In this scenario, with $\alpha = 0.9$ and $\beta = 0.1$, it is clear that the optimal thresholds from the NPI-based methods are high. Therefore, a higher optimal threshold in this case can be effective if one believes that classifying healthy individuals is considered more important than classifying diseased individuals.

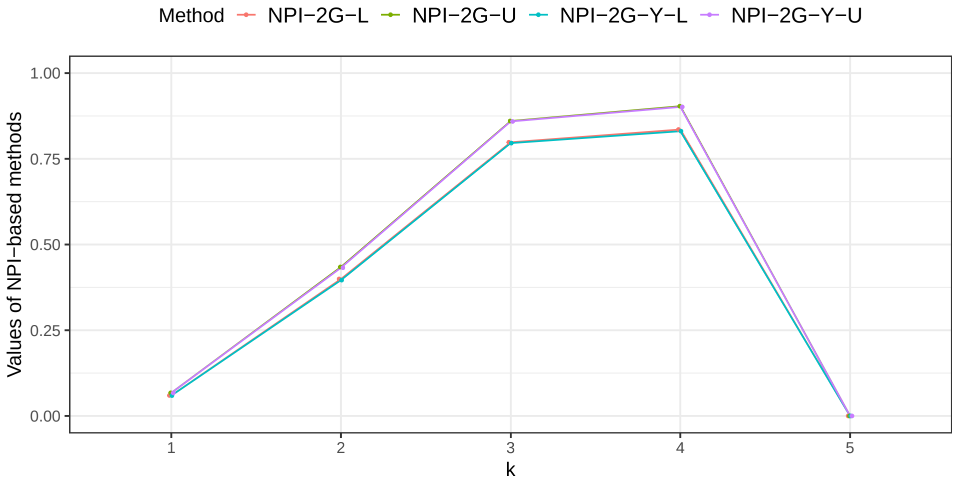


Figure 4.2: All possible thresholds and corresponding value of NPI-2G-L, NPI-2G-U, NPI-2G-Y-L, and NPI-2G-Y-U with $m = 5$ and $\alpha = 0.9, \beta = 0.1$

	k				
	1	2	3	4	5
EYI	0.4097	0.5295	**0.5932**	0.4966	0
ELI	0.4773	0.5791	**0.6042**	0.5037	0

Table 4.6: The empirical estimate of Youden's index and Liu's index

Table 4.6 presents the maximum values of the empirical Youden index (EYI) together with the empirical estimator for the Liu index (ELI), using the data shown in Table 4.2. These are all maximal for $k = 3$. The maximum values of the empirical Youden index and Liu index are equal to 0.5932 and 0.6042, respectively, and the optimal threshold for both methods is $k' = 3$. This leads to the optimal diagnostic test being such that an outcome in categories C_1 to C_3 indicates non-disease while categories C_4 and C_5 indicate disease.

Overall, the NPI-2G and NPI-2G-Y methods demonstrate that the optimal threshold selection can vary depending on the values of α, β, and m. A researcher should carefully choose whether to use the NPI-2G or NPI-2G-Y method for their analysis. One should have a careful argument for choosing between the sum or product versions. Any weighting regarding the importance of misdiagnosis should be reflected in the choice of the target proportions in our methods. Therefore, one should not just apply both methods and choose whichever seems to align better with ideas about the importance of avoiding specific misdiagnoses.

To provide further insight into the predictive performance of the proposed NPI methods along with classical methods, simulation studies will be conducted in the following section.

4.6 Predictive performance evaluation

This section presents simulation studies evaluating the performances of the proposed NPI methods compared with classical methods in the case of two-group classification. Two different scenarios are considered using Beta distributions, where the probability density function (PDF) for the Beta distribution is given by [55]

$$f(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{\mathrm{B}(a, b)}, \quad 0 \leq x \leq 1, \quad a > 0, \quad b > 0 \tag{4.9}$$

where $\mathrm{B}(a, b)$ is the Beta function, and a and b are the shape parameters of the distribution.

The two scenarios considered are constructed to represent different levels of overlap between the G^0 and G^1 groups, where the groups in Scenario 1 overlap more than in Scenario 2. The first scenario uses Beta distributions $\mathrm{B}(0.7, 2.1)$ and $\mathrm{B}(3.5, 3.5)$ for groups G^0 and G^1, respectively. The second scenario is simulated using Beta distributions $\mathrm{B}(1.2, 4.5)$ and $\mathrm{B}(4.5, 4.5)$ for groups G^0 and G^1, respectively. The degree of overlap between the two groups is quantified by calculating the overlapping area, which is found by integrating the minimum of the two PDFs over the interval $[0, 1]$, so that $\int_0^1 \min(f_{G^0}(x), f_{G^1}(x))\, dx$, where $f_{G^0}(x)$ and $f_{G^1}(x)$ denote the probability density functions of the Beta distributions for groups G^0 and G^1, respectively. This calculation was implemented using the R programming language. The overlap was found to be 0.485 for Scenario 1 and 0.364 for Scenario 2, indicating that less overlap occurs in Scenario 2. These two scenarios will also be used in Section 5.8.

For each scenario, $K = 5$ categories are considered. For categorizing the simulated values from the Beta distributions, the cut-points $0.2, 0.4, 0.6$ and 0.8 are used. This is similar to the approach presented by Coolen-Maturi [38], where ordinal outcomes were categorized based on specific cut-points. The Beta distributions are used to simulate n^0 and n^1 observations for groups G^0 and G^1. Using these simulated data observations, the optimal thresholds are determined using the methods presented in this chapter for different values of the target proportions α and β. The next step is to simulate the m^0 and m^1 future observations from the same underlying Beta distributions as the n^0 and n^1 simulated data observations. Then, classify the m^0 and m^1 simulated future observations using the optimal thresholds, so that the number of correctly classified future observations per group can be obtained. That is, for the two-group scenario, the number of future observations out of m^0 with the simulated test results in $\{C_1, \dots, C_k\}$ and out of m^1 with the simulated test results in $\{C_{k+1}, \dots, C_K\}$ are obtained.

The predictive performances of all methods have been studied with regard to the number of correctly classified future observations achieved using the desired criteria, namely, when the number of correctly classified future observations from groups G^0 and G^1 is greater than or equal to αm^0 and βm^1, respectively. Using the notation introduced by Alabdulhadi [4] and Coolen-Maturi et al. [40], denote by "+" when the desired criterion is achieved, and "−" otherwise. In the simulation, we assume that $n^0 = n^1 = n$ and $m^0 = m^1 = m$. For unbalanced cases for the data and future observations, further evaluation will be presented later in this section. For each scenario and each method, the results are based on 10,000 simulations.

Simulations have been run for each scenario and each distribution, with different α and β values selected for $n = 100$ and $m = 5, 10$. The predictive performance of the methods presented in this chapter is compared with the classical Youden index and Liu index methods, using the criterion that at least αm^0 and βm^1 of future observations from the healthy and disease groups, respectively, are correctly classified. Tables 4.7 and 4.8 display the results of the predictive performance for $m = 5$ for Scenario 1 and Scenario 2, respectively, while Tables 4.9 and 4.10 present the results for $m = 10$. To give the same importance to both groups in regard to correct classifications of future individuals, the performances have been studied for $\alpha = \beta = 0.2, 0.5, 0.7, 0.8$ for the proposed NPI-based methods, NPI-2G and NPI-2G-Y, along with empirical estimates of the Youden index and the Liu index methods, which will be denoted by EYI-2G and ELI-2G, respectively.

G^0	G^1	NPI-2G-L	NPI-2G-U	NPI-2G-Y-L	NPI-2G-Y-U	EYI-2G	ELI-2G
$\alpha\!=\!\beta = 0.2$							
+	-	27	27	27	27	8	17
-	+	11	11	11	11	153	66
-	-	0	0	0	0	0	0
+	+	9962	9962	9962	9962	9839	9917
$\alpha\!=\!\beta = 0.5$							
+	-	1486	1486	1473	1479	546	1120
-	+	872	862	918	906	3205	1733
-	-	138	139	136	137	61	109
+	+	7504	7513	7473	7478	6188	7038
$\alpha\!=\!\beta = 0.8$							
+	-	2977	2990	1258	1466	1210	2366
-	+	2149	2125	5449	5083	5568	3306
-	-	1538	1547	708	807	678	1253
+	+	3336	3338	2585	2644	2544	3075

Table 4.7: Simulation results for Scenario 1 for $m = 5$

As an example, consider Table 4.7, in which "+ +" indicates that the desired criteria have been achieved for both groups, whereas "$-$ $-$" indicates that the desired criteria have not been achieved for both groups. The desired criteria, for example, for the NPI-based methods with $\alpha\!=\!\beta$=0.2, have been achieved in 9962 out of 10,000 simulations. This means that at least one ($\alpha m = 0.2 \times 5$ and $\beta m = 0.2 \times 5$) future observation is correctly classified from each of the healthy and disease groups in the simulation. The similar performance of the NPI-based methods can be related to the fact that the proposed methods return the same optimal thresholds. On the other hand, consider Table 4.7 for $\alpha\!=\!\beta$=0.8, out of the 10,000 simulations, there are 1538 cases for NPI-2G-L in which both groups fail to meet the desired criteria.

It is clear from Tables 4.7–4.10 that the NPI-2G method outperforms all other methods in achieving the desired criteria in both groups for all the settings that have been considered. For small values of the target proportions, where $\alpha\!=\!\beta = 0.2$, it appears that the NPI-2G and NPI-2G-Y perform similarly, as the desired criteria are easily met. These tables also illustrate that, for $\alpha\!=\!\beta = 0.2$, 0.5, the performance of the methods is better for $m = 10$ than for $m = 5$. While the NPI-2G and NPI-2G-Y methods demonstrate similar performance when $\alpha = \beta = 0.2$, the NPI-2G-Y method

G^0	G^1	NPI-2G-L	NPI-2G-U	NPI-2G-Y-L	NPI-2G-Y-U	EYI-2G	ELI-2G
$\alpha = \beta = 0.2$							
+	-	16	16	16	16	13	15
-	+	4	4	4	4	35	18
-	-	0	0	0	0	0	0
+	+	9980	9980	9980	9980	9952	9967
$\alpha = \beta = 0.5$							
+	-	1273	1273	1273	1273	1058	1191
-	+	207	207	209	207	831	415
-	-	28	28	28	28	22	26
+	+	8492	8492	8490	8492	8089	8368
$\alpha = \beta = 0.8$							
+	-	3638	3642	3408	3461	3030	1221
-	+	886	876	1320	1206	2003	3458
-	-	551	551	523	529	466	526
+	+	4925	4931	4749	4804	4501	4795

Table 4.8: Simulation results for Scenario 2 for $m = 5$

shows poorer performance for larger values of α and β. This is because, unlike the NPI-2G method, the NPI-2G-Y method is based on the sum of the NPI lower or upper probabilities for correct classification (Equations 4.3 to 4.6) rather than their product, as explained in Section 4.4. When α and β are large, the values in Equations 4.3 to 4.6 can sometimes be very small. Summing these small values and then subtracting 1, as indicated in Equations 4.7 and 4.8, may yield very small or even negative results. According to Youden [95], as mentioned in Section 1.4, the Youden index ranges from zero to one, where a value of zero indicates a completely ineffective test with no ability to discriminate between individuals with and without the condition. Therefore, if the NPI-2G-Y-L or NPI-2G-Y-U value is zero or less, it may not be ideal for correctly classifying individuals at higher target proportions.

Interestingly, as shown in Tables 4.7–4.10, the ELI-2G method is generally the closest to the NPI-2G method in terms of performance. We have already discussed that summing the NPI lower or upper probabilities of correct classification may not be ideal when considering prediction performance, so it is not surprising that the ELI-2G method performs better than the EYI-2G method.

In the case of large α and β value, all methods have a better performance for $m = 5$ than for $m = 10$. For example, in Table 4.7 for NPI-2G with $\alpha = \beta = 0.8$, the

G^0	G^1	NPI-2G-L	NPI-2G-U	NPI-2G-Y-L	NPI-2G-Y-U	EYI-2G	ELI-2G
				$\alpha = \beta = 0.2$			
+	-	2	2	2	2	0	0
-	+	0	0	0	0	41	15
-	-	0	0	0	0	0	0
+	+	9998	9998	9998	9998	9959	9985
				$\alpha = \beta = 0.5$			
+	-	476	476	476	476	172	365
-	+	145	145	145	145	2077	902
-	-	7	7	7	7	4	5
+	+	9372	9372	9372	9372	7747	8728
				$\alpha = \beta = 0.7$			
+	-	2872	2878	2576	2667	1075	2177
-	+	1341	1335	1976	1758	5426	2941
-	-	666	666	609	622	253	509
+	+	5121	5121	4839	4953	3246	4373
				$\alpha = \beta = 0.8$			
+	-	3447	3456	1343	1156	1274	2621
-	+	1871	1847	7220	7129	6434	3588
-	-	2683	2692	524	704	1062	2084
+	+	1999	2005	913	1011	1230	1707

Table 4.9: Simulation results for Scenario 1 for $m = 10$

desired criteria have been achieved for both groups in 3336 out of 10000 simulations with $m = 5$, compared to 1999 out of 10000 with $m = 10$, as shown in Table 4.9. This shows that the different values of the target proportions, as well as the number of future observations may have an impact on the performance of the methods. In general, we notice that, as the values of α and β increase, the NPI-2G-Y method starts to perform poorly. For example, in Table 4.7 with $\alpha = \beta = 0.8$, the NPI-2G-Y-U prefers to reach the desired criterion for group G^1 with 5083 cases. The EYI-2G method is not close in terms of performance to the NPI-2G-Y method in some settings. For example, in Table 4.9 with $\alpha = \beta = 0.7$, for NPI-2G-Y-U, the desired criterion is achieved for both groups in 4953 out of 10000 simulations, while the EYI-2G method prefers to reach the desired criterion for group G^1 with 5426 out of 10000 simulations. We notice that with $\alpha = \beta = 0.8$, other methods tend to meet the desired criterion for either group G^0 or group G^1. Based on Tables 4.7–4.10, it can be seen that for $\alpha = \beta = 0.8$, the NPI-2G method outperforms all the other methods. Finally, Scenario 2 (less overlap) provides better performance for all methods due

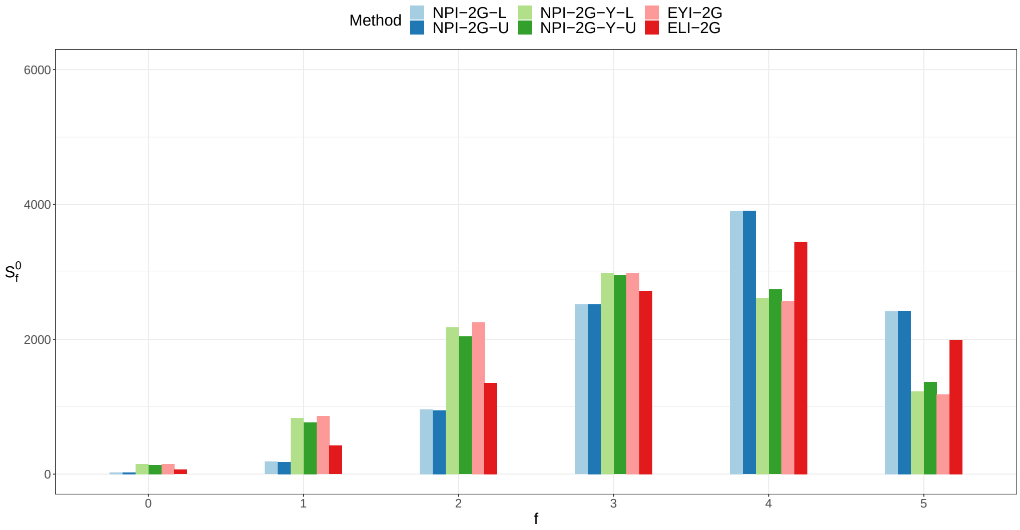
G^0	G^1	NPI-2G-L	NPI-2G-U	NPI-2G-Y-L	NPI-2G-Y-U	EYI-2G	ELI-2G
$\alpha\!=\!\beta = 0.2$							
+	-	0	0	0	0	0	0
-	+	0	0	0	0	7	1
-	-	0	0	0	0	0	0
+	+	10000	10000	10000	10000	9993	9999
$\alpha\!=\!\beta = 0.5$							
+	-	349	349	349	349	302	336
-	+	29	29	29	29	479	179
-	-	0	0	0	0	0	0
+	+	9622	9622	9622	9622	9219	9485
$\alpha\!=\!\beta = 0.7$							
+	-	2803	2803	2797	2801	2266	2628
-	+	286	286	317	302	1556	719
-	-	109	109	109	109	85	100
+	+	6802	6802	6777	6788	6093	6553
$\alpha\!=\!\beta = 0.8$							
+	-	4727	4727	4154	4350	3849	4456
-	+	681	681	1662	1330	2155	1154
-	-	726	726	631	667	577	671
+	+	3866	3866	3553	3653	3419	3719

Table 4.10: Simulation results for Scenario 2 for $m = 10$

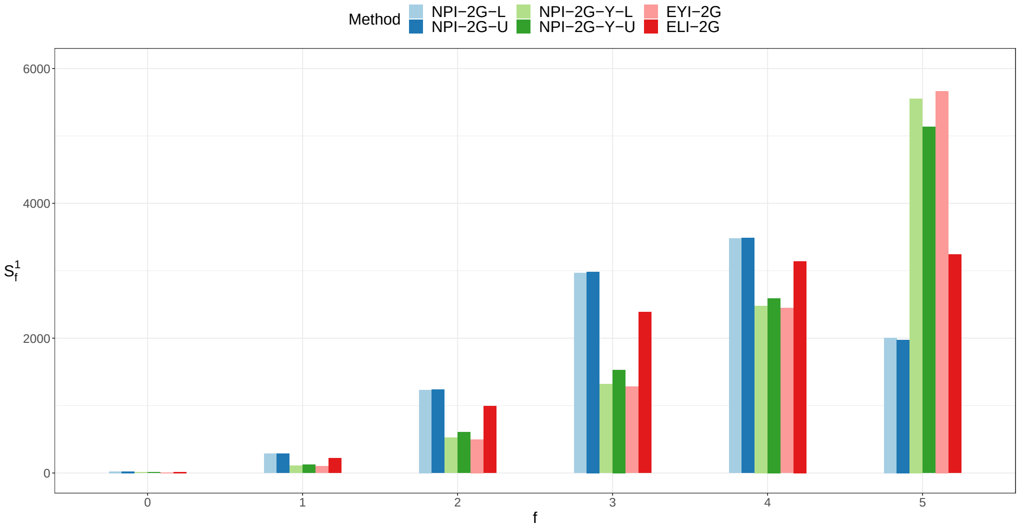
to the increased separation between the two groups in this scenario, compared to Scenario 1. To gain further insight into the predictive performance of the methods for this case with $\alpha\!=\!\beta = 0.8$, additional predictive investigation via simulation will be provided later in this section.

The number of correctly classified future observations in all simulations from groups G^0 and G^1 has been summarized using bar plots. Let the number of correctly classified future observations from group G^0 with regard to the event $W^0_k \geq \alpha m^0$ be denoted by $S^0_{f^0}$, where $f^0 \in \{0, 1, \dots, m^0\}$. Similarly, let the number of correctly classified future observations from group G^1 with regard to the event of interest, which include β, be denoted by $S^1_{f^1}$, where $f^1 \in \{0, 1, \dots, m^1\}$. Given that $n^0 = n^1 = n$ and $m^0 = m^1 = m$, therefore, $f^0 = f^1 = f$ and $f \in \{0, 1, \dots, m\}$.

Figures 4.3 and 4.4 show the distributions of the numbers of future observations, out of $m = 5$ future individuals, correctly classified in all 10,000 simulations for all methods for the case $\alpha\!=\!\beta = 0.8$, considering the two scenarios of group overlap.



(a) G^0

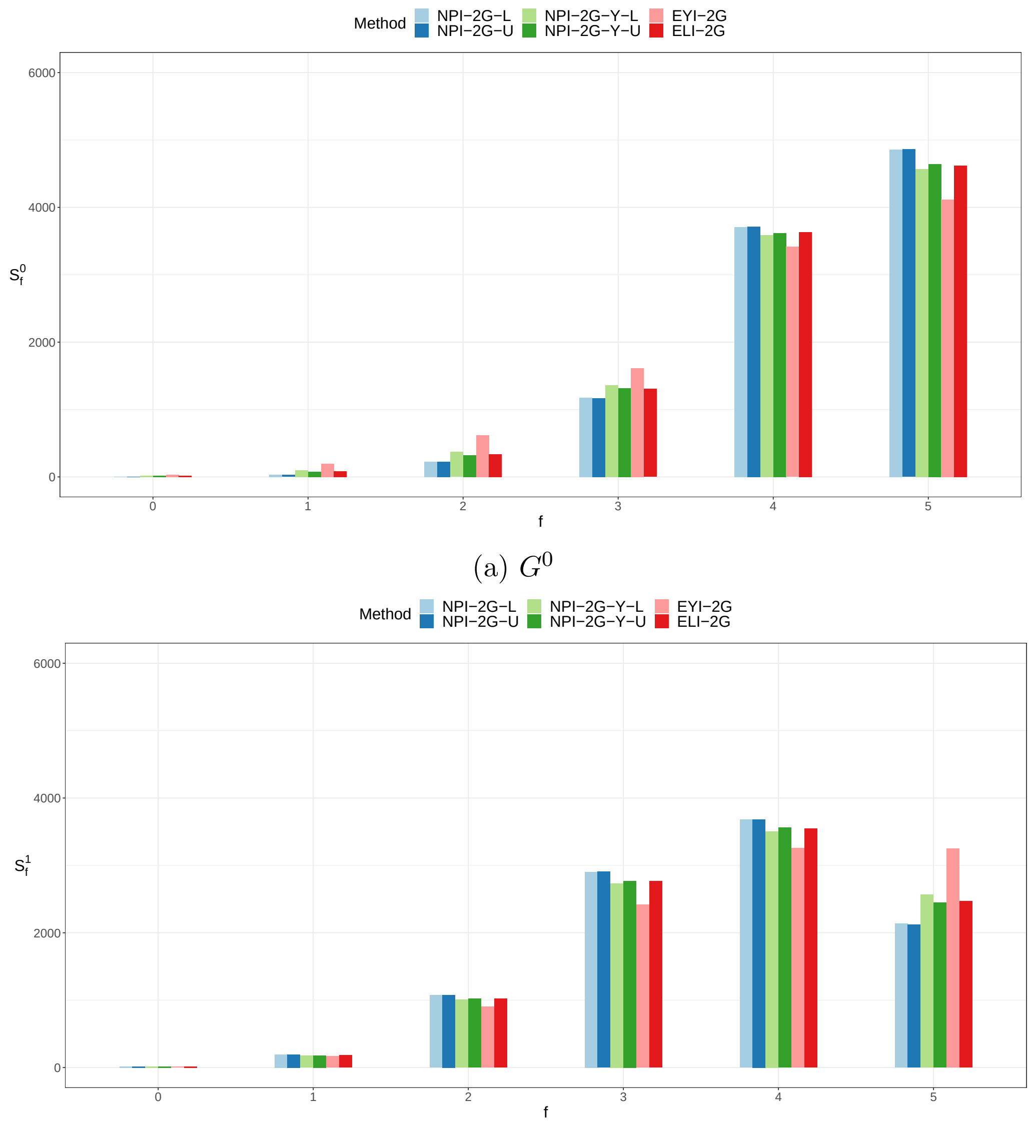


(b) G^1

Figure 4.3: Simulation results for Scenario 1 with $m = 5$ and $\alpha = \beta = 0.8$

The predictive performance is clearly better in Scenario 2, where there is less overlap between groups, compared to Scenario 1.

When α and β are large, the NPI-2G method performs better than the NPI-2G-Y method. Figure 4.3 clearly illustrates how the NPI-2G-Y method tends to meet the desired criterion for group G^1 when it fails to achieve the criterion for both groups. As shown in Figure 4.3, the NPI-2G-Y and EYI-2G methods correctly classified group G^1 more than 5000 times out of 10,000, compared with the number of correct classifications made by the methods for group G^0, while in the NPI-2G method, correct classification appears to be balanced between the two groups.



(a) G^0

(b) G^1

Figure 4.4: Simulation results for Scenario 2 with $m = 5$ and $\alpha = \beta = 0.8$

Figure 4.4 shows that in both the NPI-2G and ELI-2G methods, correct classification appears to be balanced between the two groups. In terms of the performance of the empirical methods, the ELI-2G method is generally the closest to the NPI-2G method. The ELI-2G performs better than the EYI-2G. These results should not be surprising, as it was already mentioned in Section 1.4 that summing up probabilities of correct classification rather than the product may not be ideal when attempting to achieve a higher proportion of correctly classified individuals. This may result in unbalanced classification rates, as discussed in Section 1.4, leading to the correct classification of more future individuals from either group G^0 or G^1. In both scenarios considered, the NPI-2G method clearly outperforms all other methods.

The results of this section show that, all methods perform poorly as the number of future observations increases to $m = 10$, except when α and β are not large. When they are large, the NPI-2G method performs better than the other methods. In the case that the NPI-2G-Y method has poor predictive performance, the NPI-2G method can overcome this and provide balanced classification for the two groups. The overall results show that the optimal thresholds for a given diagnostic test are dependent on the values of α and β, as well as the number of future observations considered. This highlights the importance of taking these values into account when selecting the optimal thresholds for a given diagnostic test, as they have an impact on predictive performance. This section has presented the evaluation of the performance of the proposed NPI methods compared with classical empirical methods for different values of m, α, and β. However, the poor performance of the NPI-2G-Y with larger values of α and β raises the question of whether the number of categories considered and the sample size impact the method's performance. Therefore, further investigation is presented next, specifically focusing on the case where $\alpha = \beta = 0.8$.

It may be of interest to investigate an unbalanced scenario for n^0 and n^1, since it is possible that the numbers in the groups may differ considerably in practice [39]. For example, individuals in group G^1 may present severe problems, but there would likely be few such individuals in the study. Therefore, the question arises as to whether or not the sample size influences the performances of the methods presented in this chapter. A further question that needs addressing is whether the numbers of future observations or the numbers of categories considered have an impact on the performances of the methods. To gain insight into the predictive performance of the proposed NPI methods compared with empirical methods, further investigation is presented, specifically focusing on the case where $\alpha = \beta = 0.8$.

Table 4.11 presents different cases for Scenario 1, where both the data and future observations are simulated from the same underlying Beta distributions. To investigate the effect of the number of categories, the first case considered is where $K = 8$ with $n = 100$ and $m = 5$. When comparing this case with $\alpha = \beta = 0.8$ presented in Table 4.7, the same performances are observed for all methods, with

G^0	G^1	NPI-2G-L	NPI-2G-U	NPI-2G-Y-L	NPI-2G-Y-U	EYI-2G	ELI-2G
$n = 100,\ m = 5,\ K = 8$							
+	-	1937	1955	862	957	761	1391
-	+	3186	3157	5119	4953	5257	4199
-	-	1288	1297	783	819	733	1029
+	+	3589	3591	3236	3271	3249	3381
$n = 100,\ m = 10,\ K = 8$							
+	-	2158	2181	771	681	743	1394
-	+	3300	3252	7043	6820	6537	5014
-	-	2240	2261	714	874	965	1580
+	+	2302	2306	1472	1625	1755	2012
$n^0 = 100,\ n^1 = 35,\ m^0 = 5,\ m^1 = 10,\ K = 8$							
+	-	1550	1762	811	944	1201	1740
-	+	4231	3913	5749	5312	4869	3930
-	-	1114	1223	624	773	900	1236
+	+	3105	3102	2816	2971	3030	3094

Table 4.11: Simulation results for Scenario 1 with $\alpha = \beta = 0.8$

the NPI-2G method outperforming all the other methods. However, the NPI-2G-Y method performs slightly better than with $K = 5$ due to the increased number of the categories. For NPI-2G-Y-L with $K = 8$, the desired criteria have been achieved for both groups in 3236 out of 10,000 simulations, that is, at least 4 future observations are correctly classified from each of the disease and non-disease groups. This compares to 2585 out of 10,000 simulations with $K = 5$. Similarly, for NPI-2G-Y-U with $K = 8$, the desired criteria have been achieved for both groups in 3271 out of 10,000 simulations, compared to 2644 out of 10,000 simulations with $K = 5$.

Next, an increase in m to 10 is implemented. When comparing this case in Table 4.11 with $\alpha = \beta = 0.8$ presented in Table 4.9, again, all methods perform similarly to the case presented in Table 4.9 with a slight increase in the numbers indicating that the desired criteria have been achieved for both groups. Finally, we consider the case with $n^0 = 100$, $n^1 = 35$, and $m^0 = 5$, $m^1 = 10$, with $K = 8$. It is observed that the NPI-2G method outperforms all the other methods

Similarly, Table 4.12 presents cases for Scenario 2, where the data and future observations are simulated from the same underlying Beta distributions. When comparing the case with $\alpha = \beta = 0.8$ presented in Table 4.8, we observe similar behaviours with $n = 100$ and $m = 5$. By increasing m to 10 with more categories,

G^0	G^1	NPI-2G-L	NPI-2G-U	NPI-2G-Y-L	NPI-2G-Y-U	EYI-2G	ELI-2G
$n = 100,\ m = 5,\ K = 8$							
+	-	2337	2337	1959	2029	1503	1930
-	+	1713	1713	2290	2170	3062	2334
-	-	510	510	443	457	373	437
+	+	5440	5440	5308	5344	5062	5299
$n = 100,\ m = 10,\ K = 8$							
+	-	3019	3040	2424	2564	1908	2433
-	+	1711	1669	2775	2533	3735	2779
-	-	719	728	579	607	484	587
+	+	4551	4563	4222	4296	3873	4201
$n^0 = 100,\ n^1 = 35,\ m^0 = 5,\ m^1 = 10,\ K = 8$							
+	-	1909	2201	1519	1788	1939	2287
-	+	2991	2594	3524	3149	2990	2508
-	-	444	500	363	421	424	510
+	+	4656	4705	4594	4642	4647	4695

Table 4.12: Simulation results for Scenario 2 with $\alpha{=}\beta = 0.8$

$K = 8$, all the methods achieve the desired criterion for both groups and do not tend to reach the desired criterion for the healthy group, as seen in the case presented in Table 4.10. In the unbalanced scenario of n and m, all methods perform better than in the case presented in Table 4.11. Finally, and not surprisingly, all methods perform much better over all settings in Scenario 2 than in Scenario 1, as the groups in Scenario 2 have less overlap.

The overall results presented in this section highlight the impact of the number of future observations and the values of α and β on achieving the required criteria for correctly classifying the number of future observations from groups G^0 and G^1. The NPI-2G outperforms all the other methods and for all the settings that have been considered when considering either a different number of categories or an unbalanced scenario of n and m observations. However, we notice that in such scenarios, the NPI-2G-Y method might perform slightly better. This raises the question of how to choose these values in practical applications or investigate the settings where the NPI-2G-Y method can work best. However, researchers should always make careful consideration when selecting these values for their analysis.

While α, β and the values of m for the methods presented in this section should be set by medical professionals, considering individual patient preferences for setting these values based on personal circumstances and potential consequences of misdiagnosis could enhance the applicability of these methods. For example, if giving medication to a healthy person by mistake could result in severe illness for a year, but giving it to a diseased person could cure a disease that would otherwise likely be fatal in ten years, preferences in choosing these values might vary. A young, healthy individual may not be concerned about the risk of being misdiagnosed and taking medication, since they will have a longer life expectancy and will have more time to recover. In contrast, an older individual may prioritize avoiding the severe illnesses caused by the medication if they are healthy, given their shorter life expectancy and the effect the medication has on their quality of life. This aligns with patient-centered approaches in medical decision-making, where individual preferences and values are critical. According to Elwyn et al. [49], shared decision-making plays an important role in clinical practice, emphasizing the importance of patient preferences when making treatment decisions. Barry and Edgman-Levitan [15] also emphasize the importance of shared decision-making in patient-centered care, which includes the consideration of patient values and preferences. This topic is left for future research regarding guidance on the choice of these numbers in practical situations.

4.7 Concluding remarks

This chapter has presented novel NPI methods based on considering multiple future individuals, explicitly in the form of a predictive problem to select the optimal diagnostic test threshold for two-group classification with ordinal outcomes. The NPI approach is used for selecting the optimal thresholds by taking into account a given number of future observations and criteria based on each group's target proportion of successful diagnoses. This chapter analyses the cases $n^0 = n^1$ and $m^0 = m^1$ with $K = 5$, also considering large α and β for unbalanced n and m with $K = 8$. Although in practice one would not know a specific number of future observations, the main objective is to investigate how the optimal threshold might vary in relation to the number of future observations. The methods presented in this chapter, however, would be straightforward to apply if there is a scenario involving specific values for m^0 and m^1.

The proposed methods have been illustrated with an example based on data from the literature, considering different scenarios of the target proportions α and β. Their performances have been evaluated via simulations. The proposed methods have been compared with classical methods, including the Youden index and the Liu index. The results show that, the two NPI-based methods perform similarly when α and β are not large, but when they are large, the NPI-2G method performs better than the other methods, and the NPI-2G-Y-L method starts to perform poorly. Moreover, the overall results indicate that the optimal threshold for a given diagnostic test depends on the values of α, β, and m. The results indicate that changing the number of future individuals may affect the optimal threshold selection. This highlights the importance of taking these values into account when selecting the optimal threshold for a given diagnostic test, as they have an impact on the predictive performances.

It would be of interest to develop a method for searching for the optimal values of these target proportions with ordinal outcomes. Recently, NPI-based methods have been presented for the optimal choice of α and β with real-valued data in the context of classification tree developments using a machine learning method that involves a two-loop process [6]. Although the implementation of these methods typically requires substantial data sets, it could be interesting to adjust the methods for ordinal data, which is an interesting topic for future research. The methods for optimal threshold selection in two-group classification problems presented in this chapter will be extended in the next chapter to include classification problems involving three ordinal groups.

Chapter 5

Optimal thresholds selection in three-group classification

5.1 Introduction

This chapter extends the two-group NPI methods presented in Chapter 4 to three-group classification problems with ordinal outcomes. Diagnostic tests with ordinal outcomes occur in many medical applications and other fields, in which the test yields a result in one of several ordered categories [2, 14]. For instance, a diagnostic test result with ordered categories may represent the level of severity of a disease. Traditionally, diagnostic studies have measured the accuracy of diagnostic tests by categorizing individuals into two binary groups: healthy or diseased. However, many diagnostic tests in practice deal with cases involving ordinal outcomes, where individuals are classified into multiple ordered groups based on known condition status [38, 96]. In medicine, there are some situations in which it is necessary to classify individuals into multiple groups, such as three groups based on the stage of their chronic disease or three groups based on their risk of developing the disease. Therefore, it is important to develop methods that accurately measure the performance of diagnostic tests in cases of ordinal outcomes.

In many medical diagnostic situations, there is an intermediate or transitional stage (group) between a non-diseased status and a diseased status for some diseases, such as liver cancer (LC) and Alzheimer's disease (AD) [9, 86, 93]. It is therefore common for LC and AD diseases to be classified into three groups based on disease status: non-diseased, early diseased, and fully diseased status. Detecting the intermediate stage of AD progress, for example, is crucial to prevent severe disease developments in the future. When AD is detected at an intermediate stage, new medications can be used to slow the disease's progression. Therefore, good diagnostic tests which can discriminate between disease status are essential. Consequently, the development of methodologies for measuring diagnostic accuracy for such tests is needed.

For three-group classification problems, the receiver operating characteristic (ROC) surface is a common tool for evaluating diagnostic test performance, introduced by Mossman [70] and further developed by Nakas and Yiannoutsos [75]. In this setting, two threshold values, k'_a and k'_b, are required, where $k'_a < k'_b$. Nakas et al. [72] extended the Youden index to three-group problems, maximising the sum of the probabilities of correctly classifying individuals into each group. These probabilities, also referred to as classification rates in the literature, represent the proportion of correctly classified individuals for each group. The maximum volume (MV-3G) method, introduced by Attwood et al. [9], uses the product of these probabilities, aiming to balance classification rates between the groups. Balancing here means that no group has a much larger probability of being correctly classified than the others. By multiplying these probabilities, the method reduces the overall product if one group has a very small probability. These methods are detailed in Section 5.3.

Coolen-Maturi [38] introduced a nonparametric predictive inference (NPI) approach for three-group classification problems with ordinal outcomes to assess a diagnostic test's ability to discriminate among the three ordered groups, focusing on a single future observation. This chapter presents three NPI-based methods for selecting optimal diagnostic test thresholds for three-group classification settings with ordinal outcomes, where the inference is based on multiple future individuals.

This chapter is organised as follows. Section 5.2 provides a brief introduction to ordinal diagnostic tests in three-group classification. Section 5.3 introduces an overview of classical diagnostic test threshold methods. In Section 5.4, a pairwise approach is presented using the NPI-2G method from Section 4.3, to independently select the optimal thresholds k'_a and k'_b. In Section 5.5, NPI is proposed for selecting the optimal thresholds for three-group classification problems with multiple future observations, extending the two-group NPI method presented in Section 4.3. Section 5.6 presents an NPI method inspired by the Youden index, in the sense that the criterion maximises the sum of the NPI lower or upper probabilities of correct classification with multiple future individuals for the three groups. A detailed example is given in Section 5.7 to illustrate and discuss the new methods. Simulation studies for three-group settings will be conducted in Section 5.8 to provide insight into the predictive performance of the proposed methods in comparison with the classical methods. Finally, Section 5.9 concludes with some concluding remarks.

5.2 Diagnostic tests for three groups

This section considers diagnostic accuracy when there are three ordered groups of diseases. Essentially, a diagnostic test with ordinal results is being considered, meaning that each individual's test outcome indicates one of $K \geq 2$ ordered categories, representing increasing severity levels related to their indication of having the condition of interest. These categories are denoted by C_1 to C_K.

Data are assumed to be available on individuals who are classified into three groups based on their known condition status, such as minor, moderate, and major conditions, which are denoted by G^0, G^1, and G^2, respectively. It should be noted that, throughout this chapter, superscript 0 indicates the first group, superscript 1 indicates the second group, while superscript 2 indicates the third group. The definitions and notation presented in this chapter are similar to those introduced in Chapter 4, as well as those by Alabdulhadi [4], Coolen-Maturi [38], Coolen-Maturi et al. [40], and Elkhafifi and Coolen [48].

In a diagnostic decision in the case of ordinal data, two ordered thresholds $k_a < k_b$ in $\{1, \ldots, K\}$ are required to classify individuals into one of three ordered groups of disease status based on their diagnostic test results. For such a decision, test results in categories $\{C_1, \ldots, C_{k_a}\}$ are interpreted as indicating the least severity of the condition, "minor" or mild condition, meaning the individual belongs to G^0. Test results in categories $\{C_{k_a+1}, \ldots, C_{k_b}\}$ are interpreted as indicating a "moderate"

Condition status	Diagnostic test result							Total
	C_1	$\dots$	C_{k_a}	$\dots$	C_{k_b}	$\dots$	C_K	
G^0	n_1^0	$\dots$	$n_{k_a}^0$	$\dots$	$n_{k_b}^0$	$\dots$	n_K^0	n^0
G^1	n_1^1	$\dots$	$n_{k_a}^1$	$\dots$	$n_{k_b}^1$	$\dots$	n_K^1	n^1
G^2	n_1^2	$\dots$	$n_{k_a}^2$	$\dots$	$n_{k_b}^2$	$\dots$	n_K^2	n^2

Table 5.1: Ordinal test data for three groups classification problems

condition, meaning the individual belongs to G^1. Finally, test results in categories $\{C_{k_b+1}, \dots, C_K\}$ are interpreted as indicating the most "severe" level of the condition, meaning the individual belongs to G^2. The goal in this scenario is to select the optimal values for k_a and k_b referred to as the 'optimal thresholds' k_a' and k_b', with $k_a' < k_b'$. Table 5.1 provides notation for the number of individuals for each combination of condition status and test result.

For a pair of thresholds (k_a, k_b), the probability of correct classification for each group is defined as follows [38]. Let T^0, T^1, and T^2 denote the diagnostic test results for individuals from groups G^0, G^1, and G^2, respectively. The probability of correct classification for a subject from group G^0 is $p_0(k_a) = P(T^0 \in \{C_1, \dots, C_{k_a}\})$. Similarly, the probability of correct classification for a subjects from group G^1 is $p_1(k_a, k_b) = P(T^1 \in \{C_{k_a+1}, \dots, C_{k_b}\})$, and the probability of correct classification for a subject from group G^2 is $p_2(k_b) = P(T^2 \in \{C_{k_b+1}, \dots, C_K\})$. The empirical estimators of these probabilities $p_0(k_a)$, $p_1(k_a, k_b)$ and $p_2(k_b)$ are $\hat{p_0}(k_a) = \frac{1}{n^0}\sum_{i=1}^{k_a} n_i^0$, $\hat{p_1}(k_a, k_b) = \frac{1}{n^1}\sum_{i=k_a+1}^{k_b} n_i^1$ and $\hat{p_2}(k_b) = \frac{1}{n^2}\sum_{i=k_b+1}^{K} n_i^2$, respectively [38]. Next, a brief overview of existing methods for selecting thresholds for three groups, based on receiver operating characteristic (ROC) surface analysis, is provided.

5.3 Methods for selecting optimal thresholds for three groups

In many applications, including medicine, healthcare, and machine learning, measuring diagnostic test accuracy is essential. The receiver operating characteristic (ROC) surface is a widely used tool for evaluating how well a diagnostic test discriminates between three ordered groups. Three-group ROC surfaces generalise the popular two-group ROC curve, as introduced in Section 1.4. To construct ROC surface, the probabilities of correct classification $(p_0(k_a), p_1(k_a, k_b), p_2(k_b))$ for all $k_a < k_b$ in $\{1, \ldots, K\}$ are plotted [38]. Based on these probabilities of correct classification for three groups, Mossman [70], Nakas and Alonzo [71] and Nakas and Yiannoutsos [75] introduced the construction of the ROC surface. For measuring the overall accuracy of the test, they also considered the volume under the ROC and its relation to the probability of correctly ordered observations within each of the three groups. For a recent overview and more details about ROC surface analysis and its applications, see Nakas et al. [73].

Making diagnosis decisions and classifying patients requires selecting optimal thresholds to classify individuals, based on their diagnostic test results, into one of three groups. Selecting the optimal thresholds is an important part of defining a diagnostic test and evaluating its quality. A popular approach that considers two thresholds is the generalisation of the Youden index, proposed by Nakas et al. [72], which is an extension of the Youden index for two groups (continuous) diagnostic tests. The three-group Youden index is defined as the sum of the probabilities of correct classification for the three groups.

For three-group threshold selection, Attwood et al. [9] introduced a method called the maximum volume (MV-3G), which is an extension of the maximum area method discussed in Section 1.4. The optimal thresholds determined by Attwood et al. [9] were compared with those determined by the three-group Youden index method proposed by Nakas et al. [72]. According to Attwood et al. [9], the maximisation problem in the three-group Youden index can be formulated as the sum of two maximisations: one between healthy and intermediate groups, and another between intermediate and diseased groups, essentially maximising two two-group problems, with the constraint that the first threshold is smaller than the second. This may result in imbalanced classification rates, favouring the identification of healthy and diseased groups but with poor identification of the intermediate group. To overcome this limitation, Attwood et al. [9] introduced the MV-3G approach, which maximises the product of the probabilities for correct classification of the three groups.

In a three-group setting, Nakas et al. [74] proposed using pairwise analysis with the Youden index method for Parkinson's disease patients. Through pairwise Youden index analysis, two thresholds were derived. The first threshold was identified as the optimal threshold for the two-group comparison between the first and second groups (patients with dementia and patients with mild cognitive impairment). The second threshold was selected as the optimal threshold for the two-group comparison between the second and third groups (patients with mild cognitive impairment and healthy individuals). Nakas et al. [74] also compared the pairwise approach with the three-group Youden index and found that, in their specific data example, both methods yielded the same optimal thresholds. Additionally, the value of the Youden index for the three-group problem equaled the sum of the values of the Youden index for the two pairwise comparisons.

For ordinal data, Coolen-Maturi [38] introduced the method for selecting the optimal thresholds k'_a and k'_b using the Youden's index for ordinal three-group diagnostic tests with ordered categories, based on one future observation. The Youden's index defined as the sum of these probabilities of correct classification $(p_0(k_a), p_1(k_a, k_b), p_2(k_b))$. The optimal thresholds k'_a and k'_b are found by maximising the Youden's index, with the constraint $k'_a < k'_b$. The Youden index for ordinal three-group diagnostic tests (YI-3G) is defined as

$$\text{YI-3G}\,(k_a, k_b) = p_0(k_a) + p_1(k_a, k_b) + p_2(k_b) \tag{5.1}$$

In order to obtain the empirical estimator for YI-3G, these probabilities of correct classification are replaced by their corresponding empirical estimators. The empirical estimator of the Youden index for ordinal data (EYI-3G) is

$$\text{EYI-3G}\,(k_a, k_b) = \left(\frac{1}{n^0}\sum_{i=1}^{k_a} n_i^0\right) + \left(\frac{1}{n^1}\sum_{i=k_a+1}^{k_b} n_i^1\right) + \left(\frac{1}{n^2}\sum_{i=k_b+1}^{K} n_i^2\right) \tag{5.2}$$

According to Attwood et al. [9], the MV-3G approach is defined as the product of the correct classification probabilities for three-group diagnostic tests. Similarly, the MV-3G for ordinal three-group diagnostic tests can be defined as

$$\text{MV-3G}\,(k_a, k_b) = p_0(k_a) \times p_1(k_a, k_b) \times p_2(k_b) \tag{5.3}$$

The optimal thresholds (k'_a and k'_b) are derived by maximising the MV-3G, with the constraint $k'_a < k'_b$. By replacing these probabilities of correct classification with their empirical estimators, the empirical estimator for MV-3G is obtained. The empirical estimate of the MV-3G for ordinal data (EMV-3G) is given by

$$\text{EMV-3G}\left(k_a, k_b\right) = \left(\frac{1}{n^0}\sum_{i=1}^{k_a} n_i^0\right) \times \left(\frac{1}{n^1}\sum_{i=k_a+1}^{k_b} n_i^1\right) \times \left(\frac{1}{n^2}\sum_{i=k_b+1}^{K} n_i^2\right) \tag{5.4}$$

This chapter compares the EYI-3G and EMV-3G methods with the proposed methods in Section 5.8. Next, a new NPI-based method for selecting the thresholds for three-group diagnostic tests is presented

5.4 Pairwise approach for selecting the optimal thresholds

For the three-group classification, one approach for selecting the two thresholds k'_a and k'_b is to base the selection on two-group settings: selecting the optimal threshold k'_a based on G^0 and G^1, and the optimal threshold k'_b based on G^1 and G^2. The NPI-2G method presented in Section 4.3 can be used here twice to independently select the optimal thresholds k'_a and k'_b.

By using the methodology presented in Section 4.3, in particular Equations (4.3)–(4.6), the optimal threshold k'_a is first determined based only on the groups G^0 and G^1. The NPI lower and upper probabilities used for selecting the optimal diagnostic test threshold k'_a are

$$\underline{P}\big(W^0_{k_a} \geq \alpha m^0, W^1_{k_a} \geq \beta m^1\big) = \underline{P}\big(W^0_{k_a} \geq \alpha m^0\big) \times \underline{P}\big(W^1_{k_a} \geq \beta m^1\big) \tag{5.5}$$

$$\overline{P}\big(W^0_{k_a} \geq \alpha m^0, W^1_{k_a} \geq \beta m^1\big) = \overline{P}\big(W^0_{k_a} \geq \alpha m^0\big) \times \overline{P}\big(W^1_{k_a} \geq \beta m^1\big) \tag{5.6}$$

The optimal diagnostic threshold k'_a is selected by maximising Equations (5.5) and (5.6). In order to determine the optimal threshold k'_b, the methodology presented in Section 4.3 can again be used, based only on groups G^1 and G^2, but first, further notation is introduced.

For group G^2, in addition to the notation introduced in Section 4.3 for groups G^0 and G^1, further notation is required. Let the number of future individuals from the group G^2 be denoted by m^2. For the optimal threshold k'_b, let the number of correctly classified future individuals from G^2 be denoted by $W^2_{k_b}$, and let γ denote the target proportion of correctly classified members of G^2 where γ is in $(0, 1)$. The values of α, β, and γ will vary depending on which group is considered more important for correct diagnosis. However, if one prefers to give equal importance to the correct classification of all future individuals, α, β, and γ can be set to the same value. The general event of interest here is that the number of correctly classified future individuals from G^1 is at least βm^1 and the number of correctly classified future individuals from G^2 is at least γm^2. The three groups are also assumed to be independent, meaning that any information regarding one group does not provide any information about any other group. Based on Equations (4.3)–(4.6), the optimal diagnostic test threshold k'_b for groups G^1 and G^2 is derived. Thus, the NPI lower and upper probabilities used for selecting the optimal diagnostic test threshold k'_b are

$$\underline{P}\big(W^1_{k_b} \geq \beta m^1, W^2_{k_b} \geq \gamma m^2\big) = \underline{P}\big(W^1_{k_b} \geq \beta m^1\big) \times \underline{P}\big(W^2_{k_b} \geq \gamma m^2\big) \tag{5.7}$$

$$\overline{P}\big(W^1_{k_b} \geq \beta m^1, W^2_{k_b} \geq \gamma m^2\big) = \overline{P}\big(W^1_{k_b} \geq \beta m^1\big) \times \overline{P}\big(W^2_{k_b} \geq \gamma m^2\big) \tag{5.8}$$

In order to select the optimal threshold k'_b, Equations (5.7) and (5.8) are maximised. In this chapter, the approach that uses the NPI lower probabilities in Equations (5.5) and (5.7) to obtain the optimal thresholds k'_a and k'_b is referred to as NPI-PW-L, and the approach that uses the NPI upper probabilities in Equations (5.6) and (5.8) is referred to as NPI-PW-U. The method in general will be referred to as NPI-PW. It is important to emphasize that the optimal thresholds k'_a and k'_b obtained by the NPI-PW method may not satisfy the condition $k'_a < k'_b$. This can occur in cases of high level of overlap between the groups, due to the fact that k'_a and k'_b are obtained independently. It is also possible for the pairwise analysis to result in poor identification of the intermediate group, as shown later in Section 5.8. The next section presents an alternative method for the three-group classification setting motivated by the above mentioned problems.

5.5 Optimal thresholds selection for three groups diagnostic tests

In this section, the NPI threshold selection method presented in Section 4.3 is extended to three-ordered groups with ordinal outcomes. The same notation as in Section 5.4 is used, following the latent observation setting and data notation presented in Table 5.1.

The values of thresholds k_a and k_b that provide the best classification will be selected based on multiple future individuals. This raises the question of how one can choose the optimal thresholds k'_a and k'_b that maximise the correct classification of patients from the three groups. For specific values of the optimal thresholds k'_a and k'_b, with $k'_a < k'_b$, let $W^0_{k_a}$ denote the number of correctly classified future individuals from G^0, let $W^1_{(k_a,k_b)}$ denote the number of correctly classified future individuals from group G^1, and let $W^2_{k_b}$ denote the number of correctly classified future individuals from G^2.

Assume that α, β, and γ are values in $(0, 1)$ that are selected to reflect the relative importance of correct classification of members of each group. The general event of interest here is that the number of correctly classified future individuals from G^0 is at least αm^0, the number of correctly classified future individuals from G^1 is at least βm^1, and the number of correctly classified future individuals from the group G^2 is at least γm^2. Considering that the three groups are assumed to be independent, the NPI lower and upper probabilities for the joint events $W^0_{k_a} \geq \alpha m^0$, $W^1_{(k_a,k_b)} \geq \beta m^1$, and $W^2_{k_b} \geq \gamma m^2$ can be derived as the products of the corresponding NPI lower and upper probabilities for the individual events. The NPI lower and upper probabilities for selecting the optimal diagnostic test threshold for three-group classification settings are given by

$$\begin{aligned}&\underline{P}\big(W^0_{k_a} \geq \alpha m^0, W^1_{(k_a,k_b)} \geq \beta m^1, W^2_{k_b} \geq \gamma m^2\big) = \\ &\underline{P}\big(W^0_{k_a} \geq \alpha m^0\big) \times \underline{P}\big(W^1_{(k_a,k_b)} \geq \beta m^1\big) \times \underline{P}\big(W^2_{k_b} \geq \alpha m^2\big)\end{aligned} \tag{5.9}$$

$$
\begin{aligned}
&\overline{P}\big(W^0_{k_a} \geq \alpha m^0, W^1_{(k_a,k_b)} \geq \beta m^1, W^2_{k_b} \geq \gamma m^2\big) = \\
&\overline{P}\big(W^0_{k_a} \geq \alpha m^0\big) \times \overline{P}\big(W^1_{(k_a,k_b)} \geq \beta m^1\big) \times \overline{P}\big(W^2_{k_b} \geq \alpha m^2\big)
\end{aligned}
\tag{5.10}
$$

The approaches that use the NPI lower and upper probabilities in Equations (5.9) and (5.10) are referred to as NPI-3G-L and NPI-3G-U, respectively, while the method in general is referred to as NPI-3G. Note that NPI-3G-L and NPI-3G-U are different criteria which means they may yield different optimal thresholds. The NPI lower and upper probabilities in Equations (5.9) and (5.10) can be computed by following the methodology explained in Section 4.3 as follows.

For group G^1, the NPI lower and upper probabilities of correct classification for the event $W^1_{(k_a,k_b)} \geq \beta m^1$ are obtained from Equations (2.14), (2.15) and (2.16) with $k_a < k_b$ in $\{1, \ldots, K\}$. This involves considering two scenarios using the path counting technique discussed in Section 2.4: either $k_a = 1$ or $k_b = K$, or the case where $1 < k_a < k_b < K$.

We propose an NPI-based method for the three-group classification problem, inspired by the Youden index procedure discussed in Section 5.6, to compare it with the NPI-based method introduced in this section. This is inspired by the Youden index criterion, which maximises the sum of the probabilities of correct classification for the three groups. NPI was introduced by Coolen-Maturi et al. [38] for a three-group Youden index based on one future individual for each group with ordinal data. This motivates an NPI-based method for the three-group Youden index taking into consideration a fixed number of multiple future individuals per group; this is presented next.

5.6 Optimal thresholds selection for three groups Youden index

An important consideration when designing a diagnostic method for three-group classification is the choice of decision thresholds k_a and k_b. This section introduces an NPI-based method for selecting the optimal thresholds in three-group classification setting, inspired by the sum-based approach of the Youden index, which maximises the sum of the probabilities of correct classification for the three groups [38]. This approach extends the two-group method presented in Section 4.4, taking into account a fixed number of multiple future individuals per group. By applying the NPI lower and upper probabilities of correct classifications from Section 4.3, the NPI lower and upper probabilities for the three-group classification can be obtained.

The approaches that use the sum of the NPI lower and upper probabilities for the three-group classification, inspired by the Youden index, are denoted by NPI-3G-Y-L and NPI-3G-Y-U, respectively, with NPI-3G-Y referring to the method in general. The NPI-3G-Y-L and NPI-3G-Y-U are

$$\text{NPI-3G-Y-L} = \underline{P}\big(W^0_{k_a} \geq \alpha m^0\big) + \underline{P}\big(W^1_{(k_a,k_b)} \geq \beta m^1\big) + \underline{P}\big(W^2_{k_b} \geq \alpha m^2\big) \tag{5.11}$$

$$\text{NPI-3G-Y-U} = \overline{P}\big(W^0_{k_a} \geq \alpha m^0\big) + \overline{P}\big(W^1_{(k_a,k_b)} \geq \beta m^1\big) + \overline{P}\big(W^2_{k_b} \geq \alpha m^2\big) \tag{5.12}$$

The terms in the right side of Equations (5.11) and (5.12) are derived as explained in Section 5.5. Optimal diagnostic thresholds can be determined by maximising either the NPI-3G-Y-L or the NPI-3G-Y-U. It should be noted that the NPI-3G-Y-L and NPI-3G-Y-U may result in different optimal thresholds. As an illustration of the three NPI-based methods proposed in this chapter, an example will be provided in Section 5.7. To provide further insight into the predictive performance of the proposed NPI methods along with classical methods, simulation studies will be conducted in Section 5.8.

5.7 Example of optimal thresholds selection

In this section, a detailed example is presented using a real medical dataset for Alzheimer's disease from the literature to illustrate the empirical Youden index method (EYI-3G) and the maximum volume (EMV-3G) method presented in Section 5.3, as well as the three NPI-based methods presented in Sections 5.4, 5.5, and 5.6, namely, NPI-PW, NPI-3G and NPI-3G-Y.

	C_1	C_2	C_3	C_4	Total
SCD (G^0)	15	0	13	3	31
MCI (G^1)	33	5	42	24	104
AD-Dementia (G^2)	4	1	16	61	82

Table 5.2: CSF biomarker risk categories for Alzheimer's progression in three clinical groups

For patients at preclinical stages of Alzheimer's disease (AD), cerebrospinal fluid (CSF) biomarkers can be evaluated in order to estimate the risk of developing dementia [50]. Filipek-Gliszczyńska et al. [50] considered participants from three ordered clinical groups, G^0 as subjective cognitive decline (SCD), G^1 as mild cognitive impairment (MCI) and G^2 as Alzheimer's disease (AD) dementia, with clinical follow-up averaging 14.33 months. In total, 217 patients were included in the study, 31 in the SCD group, 104 in the MCI group, and 82 in the AD-Dementia group. The cerebrospinal fluid (CSF) biomarkers of AD are ranked on an ordinal scale represented by the categories: C_1: None, C_2: Improbable, C_3: Possible, and C_4: Probable. This dataset is presented in Table 5.2.

The EYI-3G and EMV-3G methods

Table 5.3 presents the empirical estimators of the probabilities of correct classification for $k_a < k_b$ in $\{1, 2, 3, 4\}$, together with the EYI-3G and EMV-3G methods, derived using Equations (5.2) and (5.4), respectively. These are all maximal for $(k_a, k_b) = (1, 3)$. The maximum values of the EYI-3G and the EMV-3G methods are equal to 1.6797 and 0.1627, respectively, and the optimal thresholds for both methods are $k'_a = 1$ and $k'_b = 3$. This leads to the optimal diagnostic test being such that the test result in C_1 indicates individuals are assigned to the SCD (G^0 group), in C_2 and C_3 indicates individuals are assigned to the MCI (G^1 group), while observations in the final category, C_4, are assigned to the third group, the AD-Dementia (G^2 group).

(k_a, k_b)	$\hat{p_0}(k_a)$	$\hat{p_1}(k_a, k_b)$	$\hat{p_2}(k_b)$	EYI-3G	EMV-3G
(1,2)	0.4839	0.0481	0.9390	1.4710	0.0218
(1,3)	0.4839	0.4519	0.7439	**1.6797**	**0.1627**
(1,4)	0.4839	0.6827	0	1.1666	0
(2,3)	0.4839	0.4038	0.7439	1.6316	0.1454
(2,4)	0.4839	0.6346	0	1.1185	0
(3,4)	0.9032	0.2308	0	1.1340	0

Table 5.3: Empirical estimators of probabilities of correct classification and optimal thresholds for EYI-3G and EMV-3G methods

The NPI-PW method

The NPI lower and upper probabilities for the NPI-PW, as presented in Equations (5.5) to (5.8), can be derived as the products of the corresponding NPI lower and upper probabilities of correct classification for the individual events $W^0_{k_a} \geq \alpha m^0$ and $W^1_{k_a} \geq \beta m^1$ for groups G^0 and G^1, and $W^1_{k_b} \geq \beta m^1$ and $W^2_{k_b} \geq \gamma m^2$ for groups G^1 and G^2.

Table 5.4(a) displays the NPI lower probabilities of correct classification, while Table 5.4(b) shows the NPI upper probabilities of correct classification. These tables are provided to illustrate how the results for NPI-PW(G^0, G^1) and NPI-PW(G^1, G^2), presented in Table 5.5, are obtained. With $m^0 = m^1 = m^2 = m = 5$, two cases of target proportions have been considered. If equal importance is preferred for the correct classification of future individuals in all groups, values like $\alpha = \beta = \gamma = 0.2, 0.6, 0.8$ can be chosen, representing different weights of these target proportions. Alternatively, if more weight is given to the correct classification of individuals in group G^0, less weight in group G^1, and moderate weight in group G^2, one can choose $\alpha = 0.7, \beta = 0.3$, $\gamma = 0.5$.

k_a	$\underline{P}(W^0_{k_a} \geq \alpha m^0)$	$\underline{P}(W^1_{k_a} \geq \beta m^1)$	k_b	$\underline{P}(W^1_{k_b} \geq \beta m^1)$	$\underline{P}(W^2_{k_b} \geq \gamma m^2)$
$\alpha=\beta=\gamma=0.2$					
1	0.9460	0.9957	1	0.8419	0.9999
2	0.9460	0.9918	2	0.8886	0.9999
3	0.9998	0.7192	3	0.9990	0.9982
4	1	0	4	1	0
$\alpha=\beta=\gamma=0.6$					
1	0.4448	0.7994	1	0.1873	0.9969
2	0.4448	0.7268	2	0.2580	0.9952
3	0.9757	0.0869	3	0.9038	0.8738
4	1	0	4	1	0
$\alpha=\beta=\gamma=0.8$					
1	0.1688	0.4826	1	0.0400	0.9636
2	0.1688	0.3922	2	0.0654	0.9502
3	0.8680	0.0130	3	0.6575	0.6018
4	1	0	4	1	0
$\alpha = 0.7, \beta = 0.3, \gamma = 0.5$					
1	0.1688	0.9556	1	0.4985	0.9969
2	0.1688	0.9285	2	0.5903	0.9952
3	0.8680	0.3229	3	0.9850	0.8738
4	1	0	4	1	0

(a) NPI lower probabilities of correct classification

k_a	$\overline{P}(W^0_{k_a} \geq \alpha m^0)$	$\overline{P}(W^1_{k_a} \geq \beta m^1)$	k_b	$\overline{P}(W^1_{k_b} \geq \beta m^1)$	$\overline{P}(W^2_{k_b} \geq \gamma m^2)$
$\alpha=\beta=\gamma=0.2$					
1	0.9589	0.9963	1	0.8523	1
2	0.9589	0.9927	2	0.8964	1
3	0.9999	0.7357	3	0.9992	0.9986
4	1	0	4	1	0
$\alpha=\beta=\gamma=0.6$					
1	0.5000	0.8127	1	0.2006	0.9982
2	0.5000	0.7420	2	0.2732	0.9969
3	0.9873	0.0962	3	0.9131	0.8871
4	1	0	4	1	0
$\alpha=\beta=\gamma=0.8$					
1	0.2056	0.5015	1	0.0444	0.9751
2	0.2056	0.4097	2	0.0715	0.9636
3	0.9157	0.0150	3	0.6771	0.6267
4	1	0	4	1	0
$\alpha = 0.7, \beta = 0.3, \gamma = 0.5$					
1	0.2056	0.9600	1	0.5174	0.9982
2	0.2056	0.9346	2	0.6078	0.9969
3	0.9157	0.3425	3	0.9870	0.8871
4	1	0	4	1	0

(b) NPI upper probabilities of correct classification

Table 5.4: NPI lower and upper probabilities of correct classification for the NPI-PW method with $m = 5$ and different scenarios of α, β and γ

k_a	NPI-PW-L(G^0, G^1)	NPI-PW-U(G^0, G^1)	k_b	NPI-PW-L(G^1, G^2)	NPI-PW-U(G^1, G^2)
$\alpha = \beta = \gamma = 0.2$					
1	**0.9420**	**0.9553**	1	0.8418	0.8523
2	0.9382	0.9519	2	0.8885	0.8964
3	0.7191	0.7357	**3**	**0.9972**	**0.9977**
4	0	0	4	0	0
$\alpha = \beta = \gamma = 0.6$					
1	**0.3556**	**0.4064**	1	0.1867	0.2003
2	0.3233	0.3710	2	0.2568	0.2724
3	0.0848	0.0949	**3**	**0.7897**	**0.8100**
4	0	0	4	0	0
$\alpha = \beta = \gamma = 0.8$					
1	**0.0815**	**0.1031**	1	0.0386	0.0433
2	0.0662	0.0842	2	0.0621	0.0689
3	0.0113	0.0138	**3**	**0.3957**	**0.4243**
4	0	0	4	0	0
$\alpha = 0.7, \beta = 0.3, \gamma = 0.5$					
1	0.1613	0.1974	1	0.4970	0.5165
2	0.1568	0.1922	2	0.5875	0.6059
3	**0.2802**	**0.3136**	**3**	**0.8606**	**0.8756**
4	0	0	4	0	0

Table 5.5: Optimal thresholds and the NPI lower and upper probabilities for the NPI-PW (G^0, G^1) and NPI-PW (G^1, G^2) with $m = 5$ and different scenarios of α, β and γ

Table 5.6 presents the NPI-PW results for $m = 8$. Different scenarios for α, β, and γ have been considered. As shown in Tables 5.5 and 5.6, the optimal threshold for NPI-PW-L(G^0, G^1) and NPI-PW-U(G^0, G^1) is $k_a' = 1$, while for NPI-PW-L(G^1, G^2) and NPI-PW-U(G^1, G^2) is $k_b' = 3$ with $\alpha = \beta = \gamma$. However, the optimal thresholds change when α, β and γ are set at different values. For instance, for $\alpha = 0.7, \beta = 0.3$, $\gamma = 0.5$, the optimal thresholds for a decision are different than when α, β and γ are set to be equal, as it puts more emphasis on correctly

k_a	NPI-PW-L(G^0, G^1)	NPI-PW-U(G^0, G^1)	k_b	NPI-PW-L(G^1, G^2)	NPI-PW-U(G^1, G^2)
$\alpha = \beta = \gamma = 0.2$					
1	**0.9255**	**0.9453**	1	0.7618	0.7787
2	0.9219	0.9421	2	0.8381	0.8509
3	0.5698	0.5946	**3**	**0.9985**	**0.9989**
4	0	0	4	0	0
$\alpha = \beta = \gamma = 0.6$					
1	**0.2389**	**0.2918**	1	0.0770	0.0861
2	0.2077	0.2554	2	0.1284	0.1409
3	0.0222	0.0263	**3**	**0.7640**	**0.7902**
4	0	0	4	0	0
$\alpha = \beta = \gamma = 0.8$					
1	**0.0083**	**0.0125**	1	0.0023	0.0028
2	0.0056	0.0085	2	0.0052	0.0063
3	0.0003	0.0004	**3**	**0.1387**	**0.1594**
4	0	0	4	0	0
$\alpha = 0.7, \beta = 0.3, \gamma = 0.5$					
1	0.1310	0.1692	1	0.4820	0.5051
2	0.1284	0.1662	2	0.5944	0.6159
3	**0.2493**	**0.2812**	**3**	**0.9538**	**0.9613**
4	0	0	4	0	0

Table 5.6: Optimal thresholds and the NPI lower and upper probabilities for the NPI-PW (G^0, G^1) and NPI-PW (G^1, G^2) with $m = 8$ and different scenarios of α, β and γ

identifying future individuals from G^0 than those from G^1 and G^2. In this case, the optimal threshold $k'_a = k'_b$, this indicates that the group G^1 has no corresponding category (squeezing G^1), as the probability of correctly assigning individuals to that group would be zero. This aligns with the limitations and poor identification of the middle group discussed in Section 5.4.

It appears that the NPI-PW can perform well with small values of the target proportions, $\alpha = \beta = \gamma = 0.2$, since their NPI lower and upper probabilities are large, as shown in Table 5.4, and the optimal thresholds for both cases, $m = 5, 8$, are similar. For example, $\underline{P}(W^1_{k_a} \geq \beta m^1) = 0.9957$ and $\underline{P}(W^1_{k_b} \geq \beta m^1) = 0.8419$ reflect this. However, when the target proportions are set to be large, both the NPI lower and upper probabilities become small, as the required criteria become more difficult to meet than for the small target proportions scenario. For instance, with, $\alpha = \beta = \gamma = 0.8$, the NPI lower probabilities decrease, such as $\underline{P}(W^1_{k_b} \geq \beta m^1) = 0.0400$. This may result in imbalanced classification rates, particularly affecting the middle group, which is squeezed by the NPI lower probabilities for larger target proportions, as discussed in Section 5.4.

The NPI-3G method

Table 5.7 presents the NPI lower and upper probabilities, $\left[\underline{P}, \overline{P}\right]$, for correct classification of the events $W^0_{k_a} \geq \alpha m^0$, $W^1_{(k_a,k_b)} \geq \beta m^1$, and $W^2_{k_b} \geq \gamma m^2$ for each group with $m = 5$. The joint events $W^0_{k_a} \geq \alpha m^0$, $W^1_{(k_a,k_b)} \geq \beta m^1$, and $W^2_{k_b} \geq \gamma m^2$ are derived from the product of these NPI lower and upper probabilities for the individual events, as presented in Table 5.8. As before, two scenarios of the target proportions have been considered, one with $\alpha = \beta = \gamma$ with values 0.2, 0.6, and 0.8, and the other with $\alpha = 0.7, \beta = 0.3\ \gamma = 0.5$.

Table 5.8 shows the optimal thresholds k'_a and k'_b obtained for $m = 5, 8$, along with the NPI-3G-L and NPI-3G-U. Based on the comparison of Table 5.8 to Tables 5.5 and 5.6, both NPI-based methods, NPI-3G and NPI-PW, provide the same optimal thresholds for $\alpha = \beta = \gamma$ scenarios, regardless of the values of m. However, Table 5.8 shows that the NPI-3G method provides different optimal thresholds than the NPI-PW method for $\alpha = 0.7, \beta = 0.3$ and $\gamma = 0.5$.

(k_a, k_b)	$[\underline{P}, \overline{P}]\left(W^0_{k_a} \geq \alpha m^0\right)$	$[\underline{P}, \overline{P}]\left(W^1_{(k_a,k_b)} \geq \beta m^1\right)$	$[\underline{P}, \overline{P}]\left(W^2_{k_b} \geq \gamma m^2\right)$
$\alpha = \beta = 0.2$			
(1,2)	[0.9460, 0.9589]	[0.1735, 0.2507]	[0.9999, 1]
(1,3)	[0.9460, 0.9589]	[0.9398, 0.9491]	[0.9982, 0.9986]
(1,4)	[0.9460, 0.9589]	[0.9957, 0.9963]	[0, 0]
(2,3)	[0.9460, 0.9589]	[0.9108, 0.9235]	[0.9982, 0.9986]
(2,4)	[0.9460, 0.9589]	[0.9918, 0.9927]	[0, 0]
(3,4)	[0.9998, 0.9999]	[0.7192, 0.7357]	[0, 0]
$\alpha = \beta = \gamma = 0.6$			
(1,2)	[0.4448, 0.5000]	[0.0009, 0.0025]	[0.9952, 0.9969]
(1,3)	[0.4448, 0.5000]	[0.3871, 0.4215]	[0.8738, 0.8871]
(1,4)	[0.4448, 0.5000]	[0.7994, 0.8127]	[0, 0]
(2,3)	[0.4448, 0.5000]	[0.3045, 0.3369]	[0.8738, 0.8871]
(2,4)	[0.4448, 0.5000]	[0.7268, 0.7420]	[0, 0]
(3,4)	[0.9757, 0.9873]	[0.0869, 0.0962]	[0, 0]
$\alpha = \beta = \gamma = 0.8$			
(1,2)	[0.1688, 0.2056]	[0, 0.0001]	[0.9502, 0.9636]
(1,3)	[0.1688, 0.2056]	[0.1251, 0.1442]	[0.6018, 0.6267]
(1,4)	[0.1688, 0.2056]	[0.4826, 0.5015]	[0, 0]
(2,3)	[0.1688, 0.2056]	[0.0848, 0.0997]	[0.6018, 0.6267]
(2,4)	[0.1688, 0.2056]	[0.3922, 0.4097]	[0, 0]
(3,4)	[0.8680, 0.9157]	[0.0130, 0.0150]	[0, 0]
$\alpha = 0.7, \beta = 0.3, \gamma = 0.5$			
(1,2)	[0.1688, 0.2056]	[0.0161, 0.0324]	[0.9952, 0.9969]
(1,3)	[0.1688, 0.2056]	[0.7202, 0.7487]	[0.8738, 0.8871]
(1,4)	[0.1688, 0.2056]	[0.9556, 0.9600]	[0, 0]
(2,3)	[0.1607, 0.2056]	[0.6418, 0.6743]	[0.8738, 0.8871]
(2,4)	[0.1688, 0.2056]	[0.9285, 0.9346]	[0, 0]
(3,4)	[0.8680, 0.9157]	[0.3229, 0.3425]	[0, 0]

Table 5.7: NPI lower and upper probabilities of correct classification with $m = 5$ and different scenarios of α, β and γ.

Target proportions	k'_a	k'_b	NPI-3G-L	NPI-3G-U
		$m = 5$		
$\alpha=\beta=\gamma=0.2$	1	3	0.8875	0.9088
$\alpha=\beta=\gamma=0.6$	1	3	0.1504	0.1870
$\alpha=\beta=\gamma=0.8$	1	3	0.0127	0.0186
$\alpha = 0.7, \beta = 0.3, \gamma = 0.5$	1	3	0.1062	0.1366
		$m = 8$		
$\alpha=\beta=\gamma=0.2$	1	3	0.8537	0.8853
$\alpha=\beta=\gamma=0.6$	1	3	0.0666	0.0927
$\alpha=\beta=\gamma=0.8$	1	3	0.0002	0.0005
$\alpha = 0.7, \beta = 0.3, \gamma = 0.5$	1	3	0.0961	0.1301

Table 5.8: Optimal thresholds for the NPI-3G with the different scenarios of α, β and γ for $m = 5, 8$

Target proportions	k'_a	k'_b	NPI-3G-Y-L	NPI-3G-Y-U
		$m = 5$		
$\alpha=\beta=\gamma=0.2$	1	3	2.8840	2.9066
$\alpha=\beta=\gamma=0.6$	1	3	1.7057	1.8086
$\alpha=\beta=\gamma=0.8$	1	2	1.1190	1.1693
$\alpha = 0.7, \beta = 0.3, \gamma = 0.5$	1	3	1.7628	1.8414
		$m = 8$		
$\alpha=\beta=\gamma=0.2$	1	3	2.8479	2.8818
$\alpha=\beta=\gamma=0.6$	1	3	1.4171	1.5313
$\alpha=\beta=\gamma=0.8$	1	2	0.9211	0.9653
$\alpha = 0.7, \beta = 0.3, \gamma = 0.5$	1	3	1.8419	1.9200

Table 5.9: Optimal thresholds for the NPI-3G-Y with the different scenarios of α, β and γ for $m = 5, 8$

The NPI-3G-Y method

Table 5.9 shows the optimal thresholds k'_a and k'_b obtained from the NPI-3G-Y method for $m = 5, 8$. Two scenarios of the target proportions have been considered, $\alpha=\beta=\gamma$ with values 0.2, 0.6, and 0.8, and with $\alpha = 0.7, \beta = 0.3$ $\gamma = 0.5$. We notice that, as the NPI-3G-Y method is based on summing the individual NPI lower and upper probabilities of correct classification rather than taking the product, it squeezes one of the groups in order to maximise the NPI-3G-Y-L and NPI-3G-Y-U.

For example, for $\alpha=\beta=\gamma=0.8$, using Table 5.7, we have $\underline{P}(W^1_{(k_a,k_b)} \geq \beta m^1) = 0$, and the optimal thresholds are $(k'_a, k'_b) = (1, 2)$ from Table 5.9. In this case, the NPI-3G-Y method squeezes the middle group G^1 in order to maximise the sum of the probabilities of correct classification.

In this example, we notice that choosing large values for α, β and γ, as well as large m when using the NPI-3G-Y method, may lead to squeezing the middle group. Developing a method to guide the selection of these numbers in practical situations might be of interest, but we leave this as a topic for future research. In the following section, simulation studies are conducted to assess the predictive performance of the proposed NPI methods in comparison with the classical methods.

5.8 Predictive performance evaluation

In this section, simulation studies are presented for evaluating the performances of the proposed NPI methods compared with classical methods for three-group classification which is an extension of the simulation studies for two-group classification presented in Section 4.6. The third group, group G^2, follows the same simulation procedure used in Section 4.6 for groups G^0 and G^1.

Two scenarios are considered in which the data are simulated from Beta distributions. The two considered scenarios are constructed to represent different levels of overlap between the three groups, where the groups in Scenario 1 overlap more than in Scenario 2. The two scenarios are defined as follows. For the first scenario, with considerable overlap, the Beta distributions $\mathrm{B}(0.7, 2.1)$, $\mathrm{B}(3.5, 3.5)$ and $\mathrm{B}(2.1, 0.8)$ are used for groups G^0, G^1, and G^2, respectively. For the second scenario, with less overlap, the Beta distributions $\mathrm{B}(1.2, 4.5)$, $\mathrm{B}(4.5, 4.5)$ and $\mathrm{B}(4.5, 1.4)$ are used for groups G^0, G^1, and G^2, respectively. As in Section 4.6, the degree of overlap is quantified by integrating the minimum of the probability density functions (PDFs) over the interval $[0, 1]$, with the calculation extended to three groups. The overlap is determined using $\int_0^1 \min(f_{G^0}(x), f_{G^1}(x), f_{G^2}(x))\, dx$. This calculation was implemented in R, following the same procedure outlined in Section 4.6. The overlap was found to be 0.335 for Scenario 1 and 0.137 for Scenario 2.

Similar to Section 4.6, 5 categories ($K = 5$) are considered for each scenario. For categorizing the simulated values from Beta distributions into $K = 5$ categories, the cut-points $0.2, 0.4, 0.6$ and 0.8 are used. This is similar to the approach presented by Coolen-Maturi [38], where ordinal outcomes were categorized based on specific cut-points. The Beta distributions are used to simulate n^0, n^1 and n^2 observations from groups G^0, G^1 and G^2, respectively. Using these simulated data observations, the optimal thresholds are determined based on the methods presented in this chapter for specific values of the target proportions α, β and γ.

The next step is to simulate the m^0, m^1 and m^2 future observations from the same Beta distributions as the n^0, n^1 and n^2 simulated data observations in order to evaluate how the methods perform. Then, the simulated future observations are classified using the optimal thresholds, so that the number of correctly classified future observations per group can be obtained. That is, for the three-group classification, the number of future observations out of m^0, m^1 and m^2 with the simulated test results in $\{C_1, \ldots, C_{k_a}\}$, $\{C_{k_a+1}, \ldots, C_{k_b}\}$ and $\{C_{k_b+1}, \ldots, C_K\}$, respectively, are obtained. So, the m^0, m^1 and m^2 simulated future observations are compared with the optimal thresholds to obtain the number of correctly classified observations per group. The predictive performances of all methods have been studied 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 G^0, G^1 and G^2 exceed αm^0, βm^1 and γm^2, respectively. Using the notation introduced by Alabdulhadi [4] and Coolen-Maturi et al. [40], denote by "+" when the desired criterion is achieved, and "−" otherwise. In the simulation, it is assumed that $n^0 = n^1 = n^2 = n$ and $m^0 = m^1 = m^2 = m$. For unbalanced cases for the data and future observations, further evaluation will be presented later in this section.

Simulations have been run for $n = 100$ and $m = 5$, 10, for each scenario and for each distribution, and we have selected different α, β and γ values. It should be noted that the NPI-PW, $k'_a > k'_b$ may appear, since the threshold values are determined separately; in this case, k'_b is set to the same value as k'_a, so no classifications into G^1 occur. For each scenario and each method, the results in this section are based on 10,000 simulations.

Tables 5.10 and 5.11 display the results of the predictive performance for $m = 5$ for Scenario 1 and Scenario 2, respectively, while Tables 5.12 and 5.13 present the results for $m = 10$. We have studied the performance of the proposed NPI-based methods, namely, the NPI-PW, NPI-3G and NPI-3G-Y, along with the EYI-3G and EMV-3G methods. The performances have been studied in two scenarios of the target proportions, with $\alpha = \beta = \gamma = 0.2, 0.6,$ and 0.8, and with $\alpha = \beta = 0.4$ and $\gamma = 0.7$. One might prefer to set equal values of α, β, and γ to give the same importance of the correct classification of future individuals to all three groups. However, in situations where the treatment has severe side effects for severe cases individuals (G^2), one might prefer to give the same importance of the correct classification of future individuals to groups G^0 and G^1, and give a higher importance of the correct classification of future individuals to G^2, so γ can be set to be large. Any weighting regarding the importance of avoiding misdiagnosis should be reflected in the choice of the target proportions in the methods.

As an example, consider Table 5.10, in which "+ + +" indicates that the desired criteria have been achieved for all the three groups , whereas "$-$ $-$ $-$" indicates that the desired criteria have not been achieved for all groups. The desired criteria, for example, for the NPI-PW method with $\alpha = \beta = \gamma =$ 0.2, have been achieved in 9167 simulations out of 10,000 simulations. This means that at least one future observation is correctly classified from each of the three groups in the simulation. On the other hand, for $\alpha = \beta = \gamma = 0.8$, out of the 10,000 simulations, there are 1330 cases in which all groups fail to meet the desired criteria. The similar performance for the NPI-PW-L and NPI-PW-U can be related to the fact that the proposed method returns the same optimal thresholds, although this is not always the case.

Similar behaviour to the two-groups scenario, as presented in Section 4.6, is observed in Tables 5.10–5.13. The NPI-3G method performs better than other methods generally, however for small values of the target proportions, $\alpha = \beta = \gamma = 0.2$, all methods are equally effective since the desired criteria are easily met. It should be noted that for $\alpha = \beta = \gamma = 0.2$, the predictive performance of all methods is better for $m = 10$ than for $m = 5$.

G^0	G^1	G^2	NPI-PW-L	NPI-PW-U	NPI-3G-L	NPI-3G-U	NPI-3G-Y-L	NPI-3G-Y-U	EYI-3G	EMV-3G
$\alpha\!=\!\beta\!=\!\gamma\ =0.2$										
+	-	-	3	3	1	1	1	1	3	3
-	+	-	0	0	0	0	0	0	4	1
-	-	+	1	1	2	2	2	2	1	2
+	+	-	7	7	91	93	91	93	196	133
+	-	+	810	810	72	65	72	65	174	106
-	+	+	12	12	210	210	210	210	177	174
-	-	-	0	0	0	0	0	0	0	0
+	+	+	9167	9167	9624	9629	9624	9629	9445	9581
$\alpha\!=\!\beta\!=\!\gamma\ =0.6$										
+	-	-	816	817	457	451	552	543	434	538
-	+	-	33	33	404	399	398	389	925	471
-	-	+	560	559	638	644	574	585	422	563
+	+	-	380	378	1490	1474	1631	1618	1877	1659
+	-	+	5466	5474	1188	1221	1390	1384	1509	1333
-	+	+	292	286	2284	2276	2006	2031	1821	2023
-	-	-	70	70	132	130	121	122	83	120
+	+	+	2383	2383	3407	3405	3328	3328	2929	3293
$\alpha\!=\!\beta\!=\!\gamma\ =0.8$										
+	-	-	2384	2394	1427	1427	1619	1662	1289	1489
-	+	-	201	187	1285	1269	2475	2303	2476	1402
-	-	+	1840	1843	1852	1865	1107	1168	1316	1747
+	+	-	253	250	1029	1015	1064	1035	1139	1056
+	-	+	3392	3408	817	845	1730	1708	995	879
-	+	+	263	250	1367	1354	594	633	1068	1258
-	-	-	1330	1330	1736	1736	1093	1146	1313	1696
+	+	+	337	338	487	489	318	345	404	473
$\alpha = \beta = 0.4,\ \gamma = 0.7$										
+	-	-	1343	1338	221	211	219	230	393	321
-	+	-	44	44	566	576	575	570	601	437
-	-	+	46	46	76	70	58	62	22	33
+	+	-	2372	2353	3160	3204	3262	3211	5206	4865
+	-	+	2714	2770	470	407	370	417	456	315
-	+	+	45	43	846	855	856	843	421	563
-	-	-	19	19	32	31	30	31	17	20
+	+	+	3417	3387	4629	4646	4630	4636	2884	3446

Table 5.10: Prediction performance results for Scenario 1 for $m = 5$

G^0	G^1	G^2	NPI-PW-L	NPI-PW-U	NPI-3G-L	NPI-3G-U	NPI-3G-Y-L	NPI-3G-Y-U	EYI-3G	EMV-3G
$\alpha = \beta = \gamma = 0.2$										
+	-	-	0	0	0	0	0	0	0	0
-	+	-	0	0	0	0	0	0	0	0
-	-	+	1	1	0	0	0	0	1	0
+	+	-	1	1	61	64	59	65	24	55
+	-	+	539	539	52	42	57	42	413	219
-	+	+	2	2	147	146	147	146	41	71
-	-	-	0	0	0	0	0	0	0	0
+	+	+	9457	9457	9740	9748	9737	9747	9521	9655
$\alpha = \beta = \gamma = 0.6$										
+	-	-	223	223	198	200	219	214	228	235
-	+	-	3	3	90	94	75	77	36	60
-	-	+	116	116	463	463	415	420	198	341
+	+	-	148	148	815	856	808	791	392	739
+	-	+	5689	5691	1495	1412	2131	2102	4558	2825
-	+	+	71	71	2311	2337	1878	1923	639	1465
-	-	-	7	7	28	28	27	26	13	26
+	+	+	3743	3741	4600	4610	4447	4447	3936	4309
$\alpha = \beta = \gamma = 0.8$										
+	-	-	1447	1447	1049	1034	1457	1447	1339	1212
-	+	-	49	48	535	525	54	62	204	394
-	-	+	937	936	1953	1940	937	948	1217	1674
+	+	-	208	207	874	841	244	259	419	765
+	-	+	6059	6068	1994	2110	5982	5897	4866	2956
-	+	+	192	186	2011	1982	209	256	714	1527
-	-	-	233	232	563	551	244	253	336	496
+	+	+	875	876	1021	1017	873	878	905	976
$\alpha = \beta = 0.4$, $\gamma = 0.7$										
+	-	-	511	511	88	84	105	94	430	277
-	+	-	4	4	196	200	187	192	46	102
-	-	+	5	5	31	31	32	32	10	21
+	+	-	1404	1404	1631	1630	1623	1629	1820	2482
+	-	+	2131	2131	355	390	425	386	1635	923
-	+	+	15	15	865	872	836	857	189	418
-	-	-	1	1	8	8	8	8	2	6
+	+	+	5929	5929	6817	6819	6784	6802	5868	5771

Table 5.11: Prediction performance results for Scenario 2 for $m = 5$

For $\alpha = \beta = \gamma = 0.6$, as can be seen in Tables 5.10–5.13, the predictive performance of all methods is better for $m = 5$ than for $m = 10$. The NPI-3G method can achieve the desired criteria better than the other methods. As shown in these tables, the EMV-3G method is generally the closest to the NPI-3G method in terms of performance, yet the NPI method is better considering its predictive nature. A considerable squeeze for the middle group G^1 is observed with the NPI-PW method. For example, for NPI-PW-L in Table 5.10, the desired criterion is achieved for groups G^0 and G^2 in 5466 out of 10,000, indicating a squeeze on the middle group. This is due to the imbalanced classification rates between the three groups discussed in Section 5.3, that is, the classification rate for the G^0 and G^2 groups is high, but the classification rate for the intermediate group is poor.

In the case of large α, β and γ value ($\alpha = \beta = \gamma = 0.8$), all methods have a better performance for $m = 5$ than for $m = 10$, similar to $\alpha = \beta = \gamma = 0.6$. This shows that the number of future observations has an impact on the performance of the methods. As can be seen in Tables 5.10–5.13, there is difficulty in meeting the criteria for all methods with $\alpha = \beta = \gamma = 0.8$, especially in Scenario 1 where there is more overlap between the groups. A substantial squeeze for the middle group G^1 is observed for the NPI-3G-Y method. This due to the fact that the NPI-3G-Y method is based on summing up the probabilities of correct classification rather than the product, which may not be ideal when attempting to achieve a higher proportion of accurately classified individuals from the three groups simultaneously. In some occasions, the NPI-3G-Y tends to squeeze groups G^0 and G^2 and achieve the desired criterion for just group G^1. For example, in Table 5.12, for NPI-3G-Y-L, the desired criterion has been achieved for G^1 in 5092 out of 10,000 simulations, that is at least 8 future observations have been correctly classified. In both the NPI-3G and EMV-3G methods, classification appears to be balanced between the three groups. In some occasions, the NPI-3G tends to squeeze groups G^0 and G^1 and achieve the desired criteria for just group G^2, and EMV-3G tends to squeeze the middle group G^1. A substantial squeeze for the middle group G^1 is observed with the NPI-PW and EYI-3G methods.

For $\alpha = \beta = 0.4$ and $\gamma = 0.7$, all methods meet the desired criteria more than those for $\alpha = \beta = \gamma = 0.6$. In Tables 5.10 and 5.12 (Scenario 1), the EYI-3G and EMV-3G methods tend to squeeze the group G^2 substantially with $\alpha = \beta = 0.4$ and $\gamma = 0.7$. Based on all the settings considered, the NPI-3G method clearly outperforms all the other methods.

G^0	G^1	G^2	NPI-PW-L	NPI-PW-U	NPI-3G-L	NPI-3G-U	NPI-3G-Y-L	NPI-3G-Y-U	EYI-3G	EMV-3G
$\alpha=\beta=\gamma=0.2$										
+	-	-	0	0	0	0	0	0	0	0
-	+	-	0	0	0	0	0	0	0	0
-	-	+	0	0	0	0	0	0	1	0
+	+	-	0	0	33	34	31	34	87	60
+	-	+	473	473	15	12	20	12	88	47
-	+	+	0	0	68	67	68	67	45	49
-	-	-	0	0	0	0	0	0	0	0
+	+	+	9527	9527	9884	9887	9881	9887	9779	9844
$\alpha=\beta=\gamma=0.6$										
+	-	-	844	844	490	494	651	617	503	624
-	+	-	12	12	447	442	390	397	1317	525
-	-	+	483	484	920	922	787	801	535	800
+	+	-	177	177	1591	1553	1755	1777	1983	1840
+	-	+	6977	6970	950	987	1573	1383	1664	1239
-	+	+	123	128	2769	2775	2245	2356	1889	2355
-	-	-	64	64	126	126	116	121	74	113
+	+	+	1320	1321	2707	2701	2483	2548	2035	2504
$\alpha=\beta=\gamma=0.8$										
+	-	-	2955	2959	1484	1468	1642	1756	1429	1606
-	+	-	89	86	1251	1274	5092	4369	3048	1476
-	-	+	1973	1969	2098	2085	757	883	1432	1903
+	+	-	52	51	524	529	564	577	528	545
+	-	+	2410	2425	297	276	764	934	532	355
-	+	+	58	51	746	756	56	87	504	682
-	-	-	2429	2424	3510	3520	1109	1373	2455	3345
+	+	+	34	35	90	92	16	21	72	88
$\alpha=\beta=0.4$, $\gamma=0.7$										
+	-	-	1092	1091	85	82	101	87	260	200
-	+	-	6	6	378	378	371	376	502	285
-	-	+	8	8	27	27	27	27	12	17
+	+	-	1712	1710	2377	2380	2370	2379	4918	4428
+	-	+	2868	2875	247	237	299	247	532	300
-	+	+	15	15	956	958	940	956	386	577
-	-	-	4	4	10	10	9	10	4	5
+	+	+	4295	4291	5920	5928	5883	5918	3386	4188

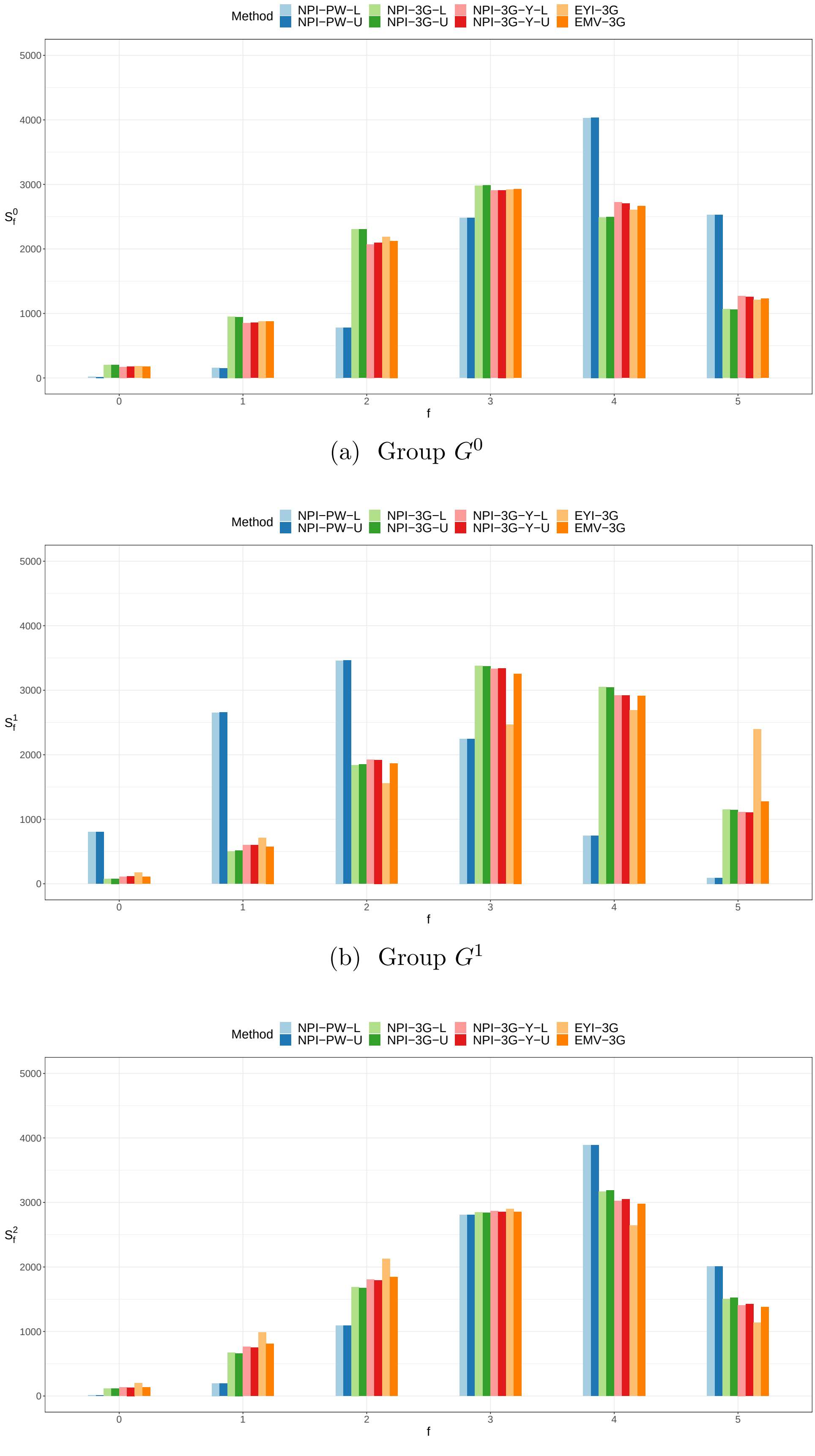
Table 5.12: Prediction performance results for Scenario 1 for $m = 10$

G^0	G^1	G^2	NPI-PW-L	NPI-PW-U	NPI-3G-L	NPI-3G-U	NPI-3G-Y-L	NPI-3G-Y-U	EYI-3G	EMV-3G
						$\alpha=\beta=\gamma=0.2$				
+	-	-	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	15	16	15	16	8	22
+	-	+	302	302	6	3	12	3	220	120
-	+	+	0	0	30	29	30	29	5	15
-	-	-	0	0	0	0	0	0	0	0
+	+	+	9698	9698	9949	9952	9943	9952	9767	9843
						$\alpha=\beta=\gamma=0.6$				
+	-	-	119	119	163	164	196	190	153	215
-	+	-	3	3	62	64	52	55	48	42
-	-	+	42	42	586	588	504	520	171	363
+	+	-	50	50	918	950	943	942	353	850
+	-	+	7367	7367	1296	1213	2104	1935	5738	3410
-	+	+	40	40	3047	3071	2502	2617	756	1767
-	-	-	4	4	10	11	8	8	5	4
+	+	+	2375	2375	3918	3939	3691	3733	2776	3349
						$\alpha=\beta=\gamma=0.8$				
+	-	-	1848	1850	1267	1266	1920	1905	1685	1551
-	+	-	15	14	593	581	16	18	240	415
-	-	+	1050	1046	3118	3091	997	1005	1587	2443
+	+	-	69	69	572	538	128	118	224	514
+	-	+	6469	6476	1443	1561	6422	6428	5031	2857
-	+	+	66	64	1671	1644	37	46	506	1154
-	-	-	324	322	1046	1022	319	321	528	812
+	+	+	159	159	290	297	161	159	199	254
						$\alpha=\beta=0.4$, $\gamma=0.7$				
+	-	-	210	210	23	21	28	23	175	109
-	+	-	0	0	60	60	59	60	22	31
-	-	+	1	1	7	8	7	8	3	4
+	+	-	495	495	620	622	616	620	975	1688
+	-	+	2585	2585	235	206	300	216	1969	1063
-	+	+	3	3	800	804	770	803	168	405
-	-	-	0	0	2	2	2	2	1	2
+	+	+	6706	6706	8253	8277	8218	8268	6687	6698

Table 5.13: Prediction performance results for Scenario 2 for $m = 10$

The number of correctly classified future observations in all simulations from groups G^0, G^1 and G^2 has been summarized by using bar plots. As shown in Figures 5.1–5.4, these numbers have been summarized from all methods together for each scenario. The bar plots from all methods provide a comprehensive overview of the performances of all methods. Let the number of correctly classified future observations from group G^0 with regard to the event of interest, which include α, be denoted by $S^0_{f^0}$, where $f^0 \in \{0, 1, \ldots, m^0\}$. The number of correctly classified future observations from group G^1 with regard to the event of interest, which include β, is represented by $S^1_{f^1}$, where $f^1 \in \{0, 1, \ldots, m^1\}$. The number of correctly classified future observations from group G^2 with regard to the event of interest, which include γ, is represented by $S^2_{f^2}$, where $f^2 \in \{0, 1, \ldots, m^2\}$. Throughout this section we assume that $n^0 = n^1 = n^2 = n$ and $m^0 = m^1 = m^2 = m$, therefore, $f^0 = f^1 = f^2 = f$ and $f \in \{0, 1, \ldots, m\}$. Figures 5.1–5.4 show the distributions of the numbers of future observations, out of a given number of future individuals in all 10,000 simulations, that are correctly classified for all methods together with the two scenarios of overlap for each group.

The number of future individuals considered in these figures is $m = 5$. The predictive performances for the first scenario are given in Figures 5.1 and 5.3 for $\alpha = \beta = \gamma = 0.6$ and 0.8, respectively, and in Figures 5.2 and 5.4 for the second scenario. With less overlap between groups, all methods clearly perform much better in Scenario 2 compared to Scenario 1. From Figures 5.1 and 5.2, for $\alpha = \beta = \gamma = 0.6$, the NPI-PW method clearly squeezes the G^1 group, leading to correct classification of more future individuals from G^0 and G^2. This is because the optimal thresholds k_a and k_b are equal or next to each other for most simulation runs due to the imbalanced classification, which indicates that there will be no or less future observations classified as belonging to G^1. The EMV-3G performs better than EYI-3G, which should not be surprising, since it has already mentioned that summing up probabilities of correct classification rather than the product may not be optimal when attempting to achieve a higher proportion of correctly classified individuals, which results in a squeeze on the middle group. Figure 5.3 provides a clear indication that, when the target proportions are set at large values ($\alpha = \beta = \gamma = 0.8$), the methods all struggle to meet the required criteria with the first scenario where the groups have more overlap. The behaviour of squeezing the middle group is now more obvious than for $\alpha = \beta = \gamma = 0.6$. With large values of α, β and γ it appears that the NPI-3G performs better than NPI-3G-Y method. Based on all the settings considered, the NPI-3G method clearly outperforms all the other methods.

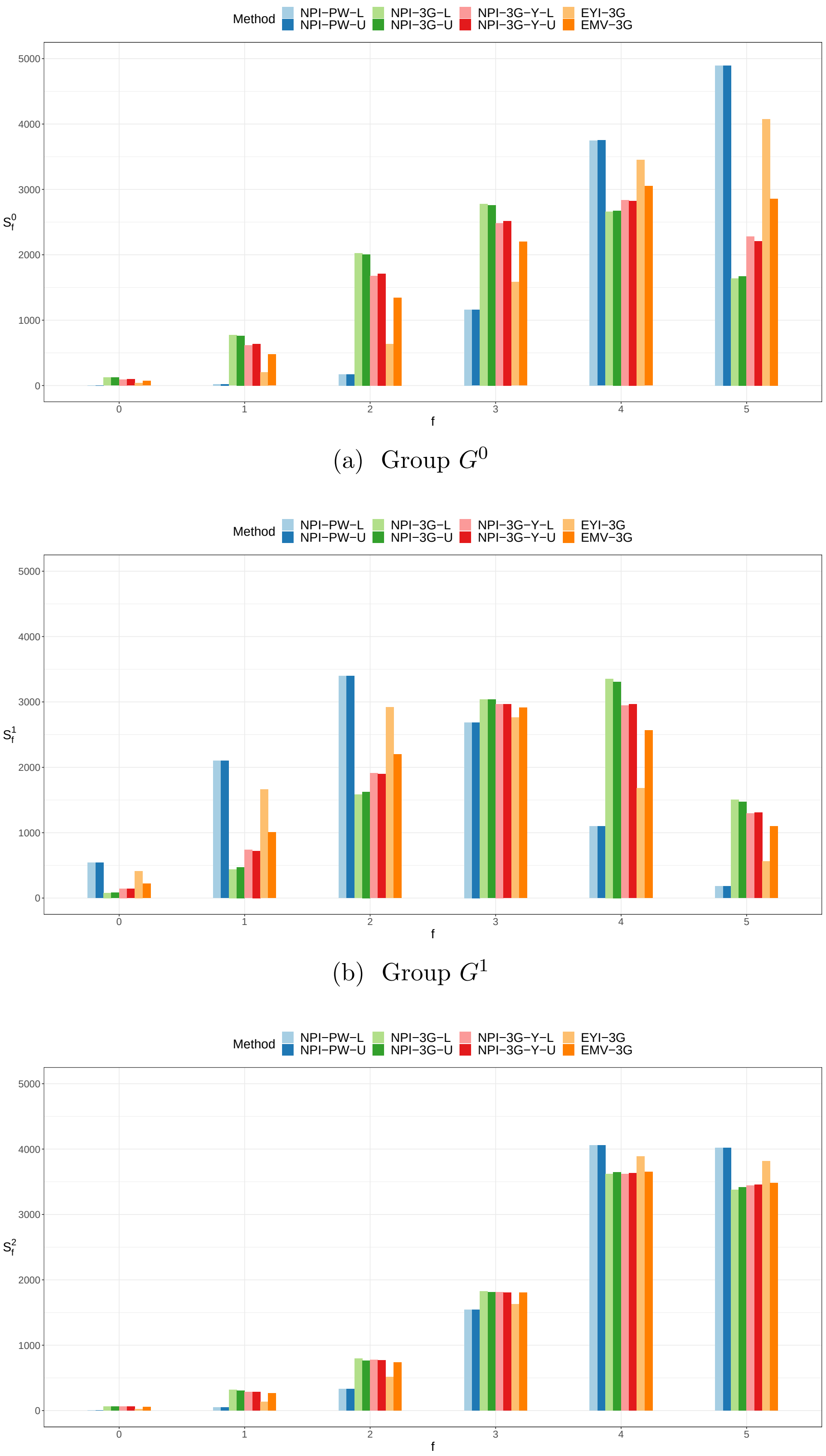


(a) Group G^0

(b) Group G^1

(c) Group G^2

Figure 5.1: Prediction performance results for Scenario 1 with $m = 5$ and $\alpha = \beta = \gamma = 0.6$

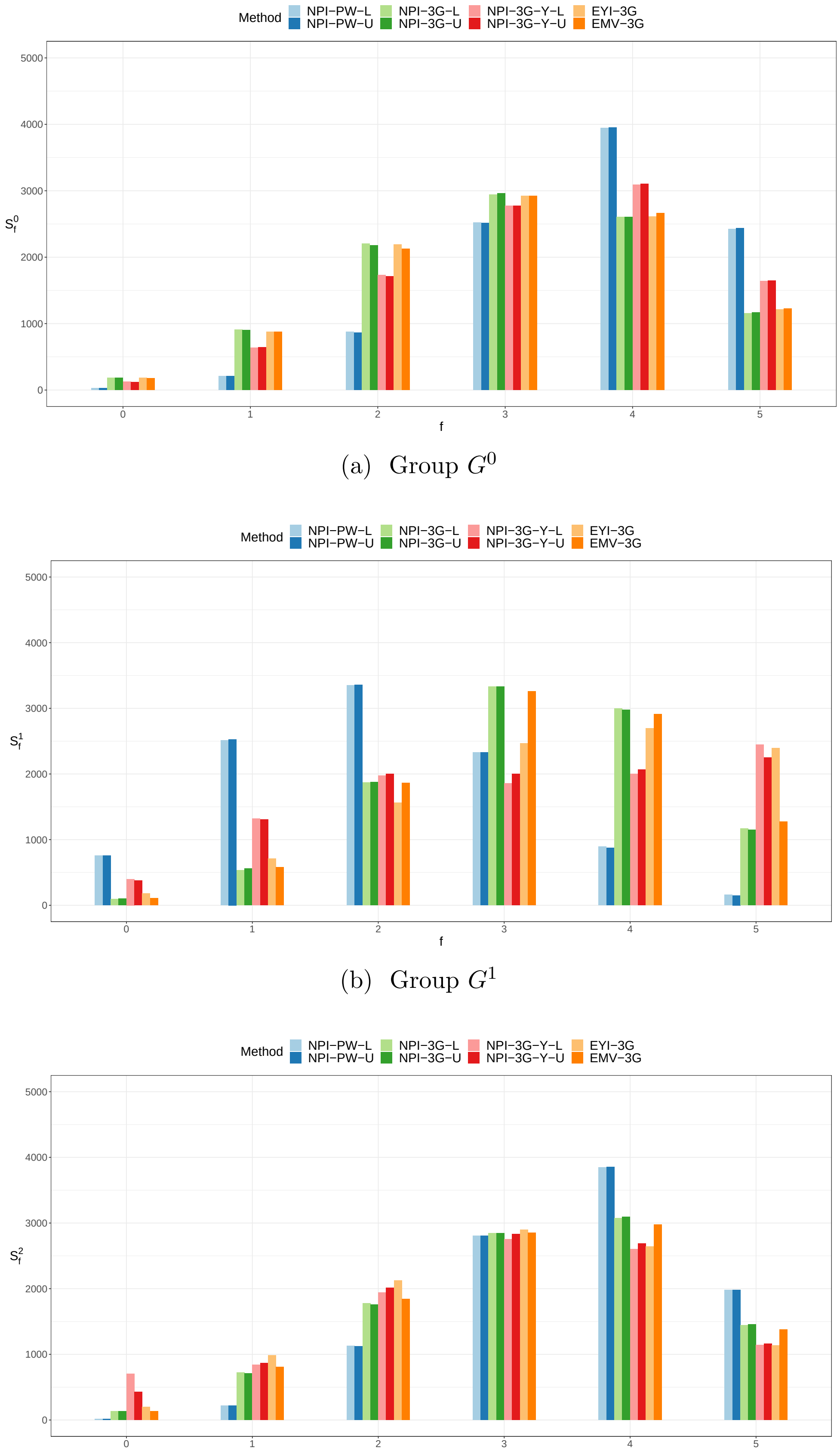


(a) Group G^0

(b) Group G^1

(c) Group G^2

Figure 5.2: Prediction performance results for Scenario 2 with $m = 5$ and $\alpha = \beta = \gamma = 0.6$

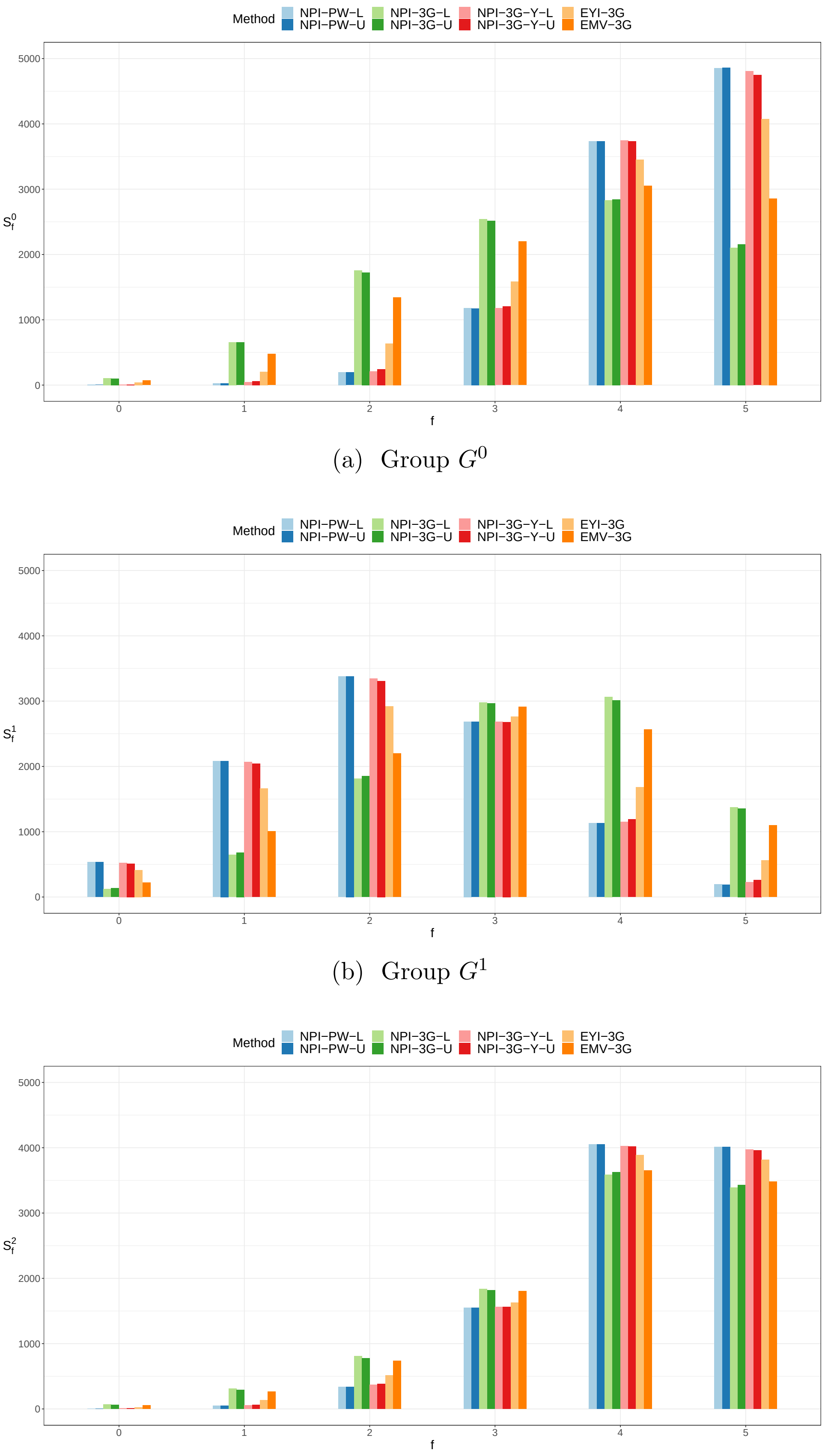


(a) Group G^0

(b) Group G^1

(c) Group G^2

Figure 5.3: Prediction performance results for Scenario 1 with $m = 5$ and $\alpha = \beta = \gamma = 0.8$



(a) Group G^0

(b) Group G^1

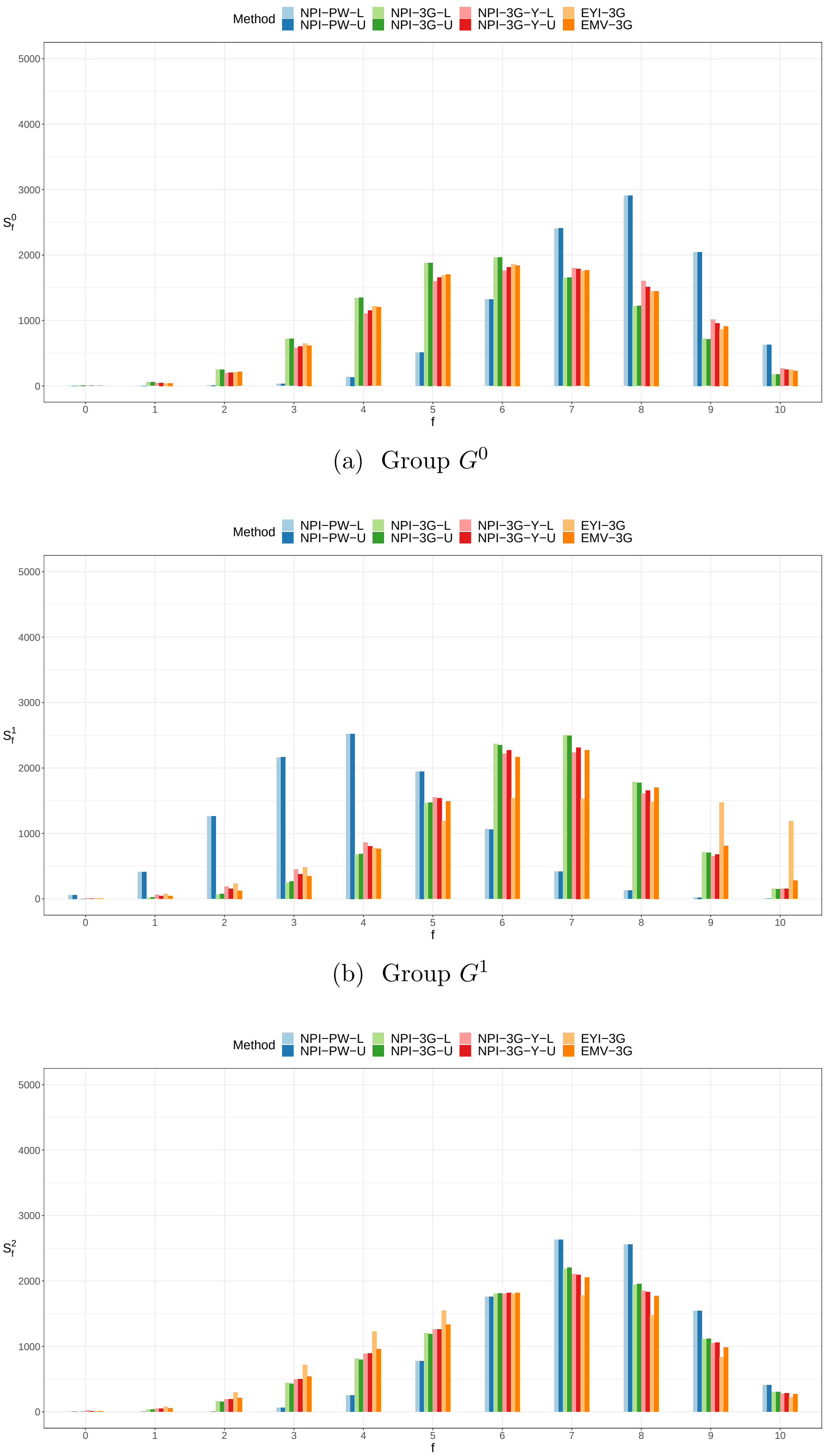
(c) Group G^2

Figure 5.4: Prediction performance results for Scenario 2 with $m = 5$ and $\alpha = \beta = \gamma = 0.8$

Figures 5.5–5.8 show the distributions of the numbers of future observations, out of m in all 10,000 simulations, that are correctly classified for all methods together for the two scenarios of overlap for each group. The number of future individuals considered in these figures is $m = 10$. The predictive performances for the first scenario are given in Figures 5.5 and 5.7 for $\alpha = \beta = \gamma = 0.6$ and 0.8, respectively, and in Figures 5.6 and 5.8 for the second scenario. The performance for all methods becomes better for Scenario 2 since the groups have less overlap. It is evident that all methods begin to perform poorly as α, β and γ are set at 0.8, and the number of future observations increases to 10, making the criteria more difficult to achieve.

For $\alpha = \beta = \gamma = 0.6$, a similar behaviour as with $m = 5$ has been observed. However, for $\alpha = \beta = \gamma = 0.8$, as shown in Figure 5.7, the performance of the NPI-3G-Y method becomes poor for G^2, the method classifies more future individuals correctly from group G^1. The figure shows that the NPI-2G-Y squeezes group G^2 more than 5000 out of 10,000 times. This indicates that for most simulation runs, the second optimal threshold, k'_b is equal to K ($k'_b = 5$). In this case, the group G^2 has no corresponding categories, as the probability of correctly assigning individual to that group would be zero, as explained in Section 5.5. This supports the conclusions we previously addressed, indicating that summing the probabilities of correct classification may not be ideal when considering prediction performance, and that using their product could be a more suitable approach.

In cases where α, β, and γ are large, the NPI-3G method performs better than other methods. In the case that the middle group has poor predictive performance with the NPI-3G-Y, the NPI-3G method can overcome this issue and provide a balanced classification for the three groups. The overall results show that the optimal thresholds for a given diagnostic test are dependent on the values of α, β and γ, as well as the number of future observations considered. Thus, when selecting thresholds for a diagnostic test, these values need to be taken into account, since they affect the predictive performance.

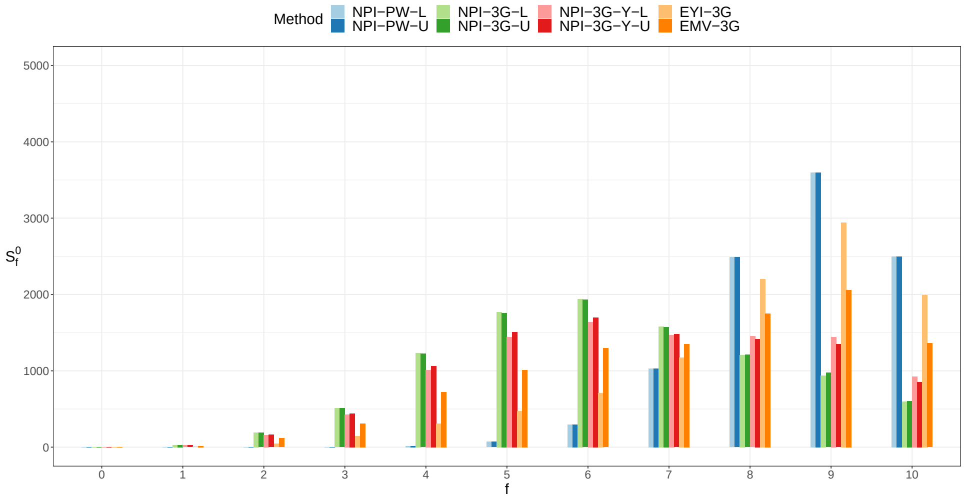


(a) Group G^0

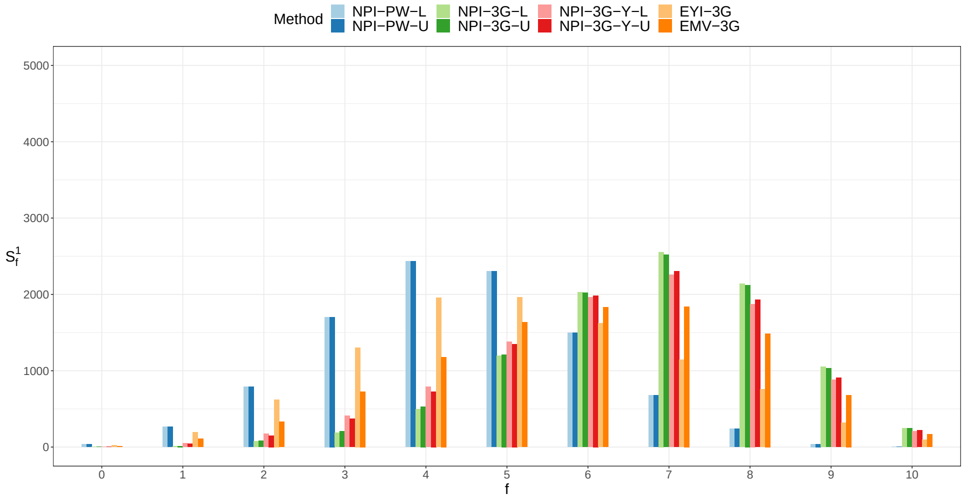
(b) Group G^1

(c) Group G^2

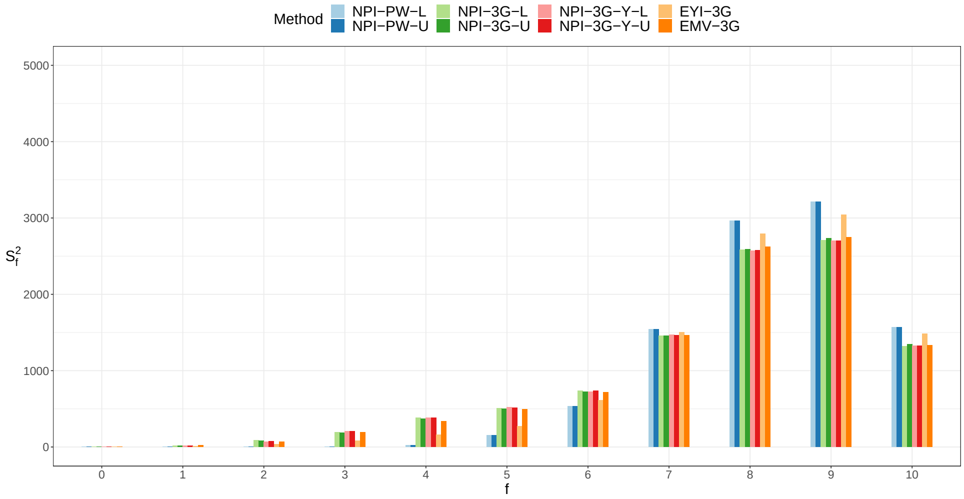
Figure 5.5: Prediction performance results for Scenario 1 with $m = 10$ and $\alpha = \beta = \gamma = 0.6$



(a) Group G^0

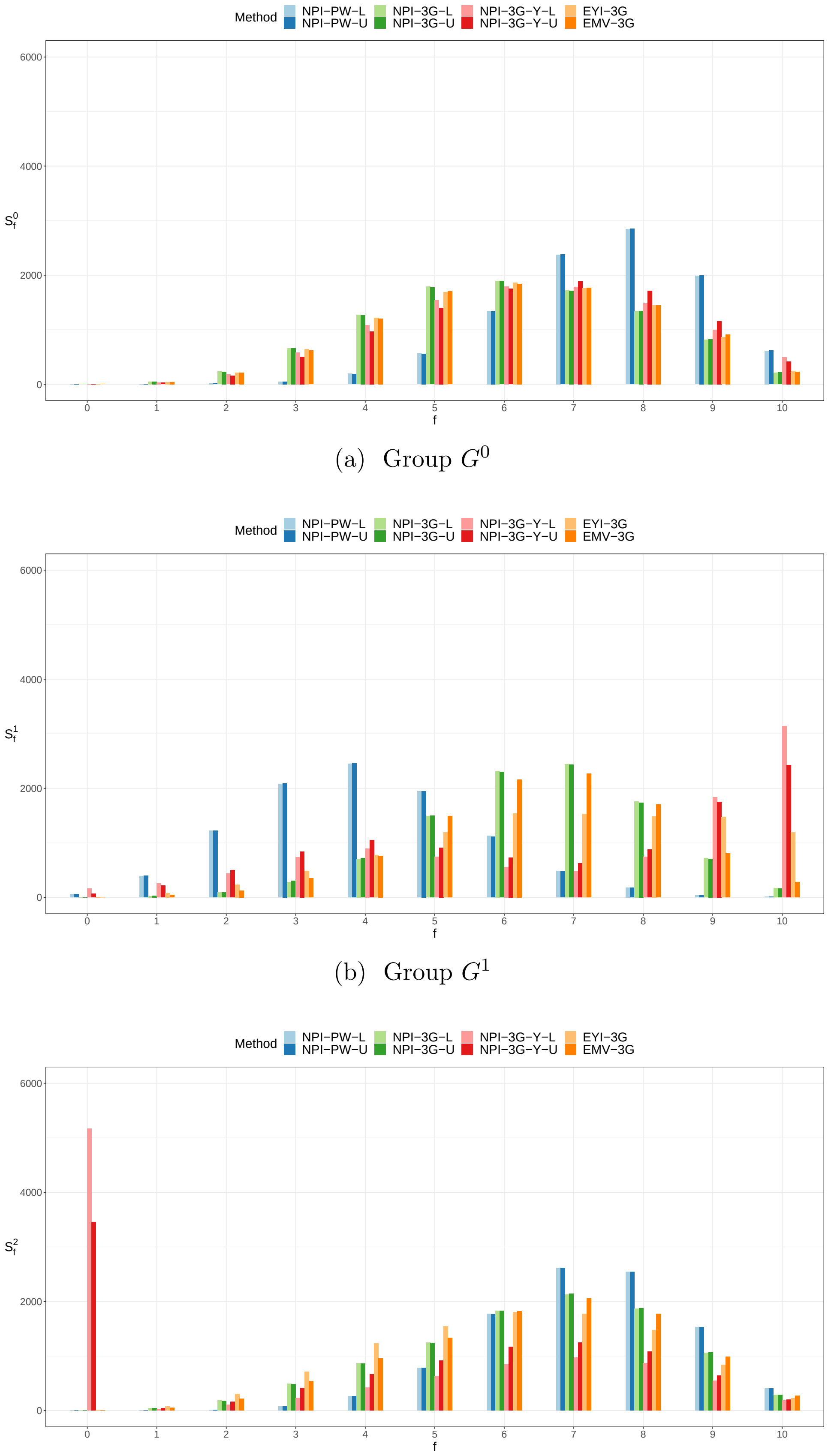


(b) Group G^1



(c) Group G^2

Figure 5.6: Prediction performance results for Scenario 2 with $m = 10$ and $\alpha = \beta = \gamma = 0.6$

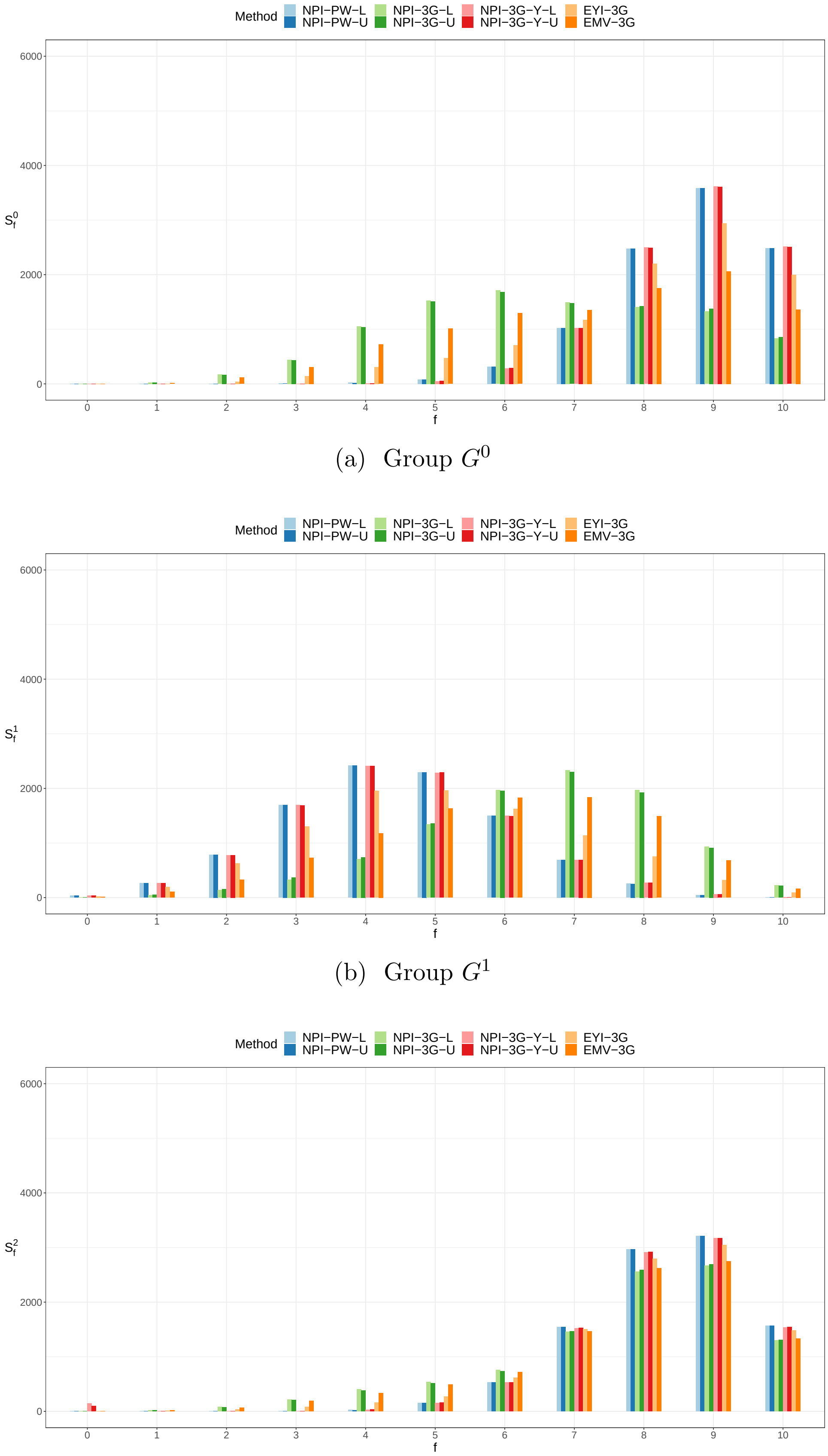


(a) Group G^0

(b) Group G^1

(c) Group G^2

Figure 5.7: Prediction performance results for Scenario 1 with $m = 10$ and $\alpha = \beta = \gamma = 0.8$



(a) Group G^0

(b) Group G^1

(c) Group G^2

Figure 5.8: Prediction performance results for Scenario 2 with $m = 10$ and $\alpha\!=\!\beta\!=\!\gamma\,=\!0.8$

The NPI-3G-Y method shows poor performance with $\alpha = \beta = \gamma = 0.8$, highlighting the need to explore whether the number of categories or the sample size influences its performance. Considering the results presented in this section, it appears that some methods face the squeezing problem of the intermediate group. In order to further investigate whether the methods might perform better in terms of reducing such squeezing when different scenarios are considered, a further investigation is presented next, focusing on the case where $\alpha = \beta = \gamma = 0.8$.

It may be of interest to investigate an unbalanced scenario for n, as practical considerations may result in considerable differences in group sizes [39]. For instance, individuals in group G^2 might experience severe problems, but their total number in the study could be small. This raises the question about how the performance of the methods presented in this chapter might be affected by the sample size. Additionally, an investigation is aimed at determining whether the predictive performance is influenced by the number of future observations or the number of categories considered.

Table 5.14 presents different cases for Scenario 1 and Scenario 2, where both the data and future observations are simulated from the same underlying Beta distributions for the two scenarios. As shown in Table 5.12, when comparing the results with those in Table 5.14 for $\alpha = \beta = \gamma = 0.8$, close performances are observed for all methods, where all methods struggle to meet the required criteria, especially in Scenario 1 where the groups have more overlap. To investigate the effect of the number of categories, all cases presented in Table 5.14 considered $K = 8$.

For Scenario 1, the first case considered is when $K = 8$ with $n = 100$ and $m = 10$. The NPI-PW method in Table 5.12 tends to squeeze groups G^1 and G^2 and achieve the desired criterion only for group G^0, while in Table 5.14, the method mostly fails the desired criterion for each group. This is because, in most simulation runs with $K = 5$, the optimal thresholds are next to each other, such as $k'_a = 2$ and $k'_b = 3$, or $k'_a = 3$ and $k'_b = 4$, whereas with $K = 8$, they are not. As explained in Section 5.2, if, for example, the thresholds are $k'_a = 3$ and $k'_b = 4$, this implies that test results in categories $\{C_1, \ldots, C_3\}$ are interpreted as indicating the least severe condition, so individual belongs to G^0, test results in category $\{C_4\}$ correspond to a moderate condition (G^1), and results in $\{C_5\}$ indicate the most severe condition (G^2). Thus, it is clear that the NPI-PW method with $K = 5$ squeezes groups G^1 and G^2 and only meets the desired criterion for group G^0 compared to the case with $K = 8$ as there are more categories to correctly classify future individuals into G^0. The NPI-3G-Y method in Table 5.12 tends to squeeze groups G^0 and G^2 and achieve the desired criterion only for group G^1, while in Table 5.14, a squeeze for the middle group G^1 is observed. This is due to the fact that, in most simulation runs with $K = 5$, the optimal thresholds are $k'_a = 1$ and $k'_b = 3$, or $k'_a = 2$ and $k'_b = 5$ which leads to achieve the desired criterion only for group G^1 and squeeze groups G^0 and G^2, while with $K = 8$ the optimal thresholds are next to each other, such as $k'_a = 3$ and $k'_b = 4$, or $k'_a = 4$ and $k'_b = 5$ which leads to squeeze the middle group G^1.

The unbalanced case of the data and future observations is considered next for Scenario 1 where $n^0 = 100,\ n^1 = 60,\ n^2 = 30$ and $m^0 = 15,\ m^1 = 10,\ m^2 = 5$. Comparing this with the case where $n = 100$ and $m = 10$, similar conclusions for the performances have been observed for the NPI-PW, NPI-3G and EMV-3G methods with a very slight increase in achieving the desired criterion for each group as there are a slightly more categories with $K = 8$ compared to $K = 5$. The NPI-3G-Y method with $K = 5$, $n = 100$ and $m = 10$ tends to squeeze groups G^0 and G^2 and achieve the desired criterion for just group G^1, while with the unbalanced case, the method tends to squeeze groups G^1 and G^2 and achieve the desired criterion for just group G^0. This is because, in most simulation runs with $K = 5$ the optimal thresholds are $k'_a = 1$ and $k'_b = 3$, or $k'_a = 2$ and $k'_b = 5$ which leads to achieve the desired criterion only for group G^1, while with the unbalanced case and $K = 8$, in most simulation runs the optimal thresholds are $k'_a = 4$ and $k'_b = 5$, or $k'_a = 5$ and $k'_b = 6$ which leads to have more categories to correctly classify future individuals to G^0. This highlights the impact of the considered number of the data and future observations on achieving the required criterion.

Similarly, Scenario 2 considers the two cases presented for Scenario 1. When comparing the case where $n = 100$ and $m = 10$ in Table 5.13 for $\alpha{=}\beta{=}\gamma$ =0.8 to Table 5.14, similar performances have been observed for all methods. Interestingly, the squeezing cases for the middle group for the NPI-PW-L and NPI-PW-U methods

G^0	G^1	G^2	NPI-PW-L	NPI-PW-U	NPI-3G-L	NPI-3G-U	NPI-3G-Y-L	NPI-3G-Y-U	EYI-3G	EMV-3G
Scenario 1 with $n = 100,\ m = 10,\ K = 8$										
+	-	-	2580	2602	1430	1481	2258	2277	1155	1390
-	+	-	399	376	1806	1679	1484	1279	3141	2224
-	-	+	1789	1792	1443	1492	2152	2137	1002	1271
+	+	-	244	236	688	659	298	284	812	747
+	-	+	1555	1564	413	435	2610	2689	383	393
-	+	+	190	185	607	601	31	49	601	593
-	-	-	3150	3152	3454	3499	1163	1277	2769	3229
+	+	+	93	93	159	154	4	8	137	153
Scenario 1 with $n^0 = 100$, $n^1 = 60$, $n^2 = 30$, and $m^0 = 15$, $m^1 = 10$, $m^2 = 5$, $K = 8$										
+	-	-	1997	2008	1734	1823	3607	3027	845	932
-	+	-	708	678	1745	1605	911	1201	2518	2017
-	-	+	2081	2086	1296	1302	1150	1236	1948	2142
+	+	-	435	431	761	772	82	116	378	388
+	-	+	1433	1458	623	653	3126	2972	555	531
-	+	+	334	324	697	644	248	392	1161	1122
-	-	-	2812	2818	2893	2950	864	1033	2433	2695
+	+	+	200	197	251	251	12	23	162	173
Scenario 2 with $n = 100,\ m = 10,\ K = 8$										
+	-	-	2101	2107	1468	1475	984	991	1733	1689
-	+	-	161	158	896	848	13	38	953	876
-	-	+	1350	1342	2089	2102	1439	1466	1578	1810
+	+	-	300	295	912	887	14	31	774	887
+	-	+	4643	4670	1478	1552	7398	7230	2637	2022
-	+	+	299	290	1325	1301	21	51	822	1034
-	-	-	649	645	1142	1139	111	152	891	1021
+	+	+	497	493	690	696	20	41	612	661
Scenario 2 with $n^0 = 100$, $n^1 = 60$, $n^2 = 30$, and $m^0 = 15$, $m^1 = 10$, $m^2 = 5$, $K = 8$										
+	-	-	2135	2133	2300	2403	1692	1817	1390	1365
-	+	-	247	230	867	772	93	182	906	906
-	-	+	1184	1177	995	957	899	918	1837	1973
+	+	-	788	781	1626	1651	247	466	715	819
+	-	+	4048	4097	1629	1707	6701	5937	2630	2019
-	+	+	288	278	824	745	63	137	1114	1337
-	-	-	615	613	845	840	178	284	777	891
+	+	+	695	691	914	925	127	259	631	690

Table 5.14: Predictive performance evaluation with $K = 8$ and $\alpha = \beta = \gamma = 0.8$

decrease from 6469 cases to 4643 for NPI-PW-L, and from 6476 to 4670 cases for the NPI-PW-U as there are more categories with the case $K = 8$. Similarly, the squeezing cases for the middle group in the EYI-3G and EMV-3G methods decrease from 5031 and 2857 cases to 2637 and 2022, respectively. A substantial squeeze for the middle group G^1 is observed again with the NPI-3G-Y method due to the fact that the optimal thresholds are next to each other. Finally, for the unbalanced case in Scenario 2, similar overall performances to the case $n = 100$ and $m = 10$ have been observed with a slight increase in achieving the desired criterion for each group.

These results demonstrate how the value of m, and the values of α, β, and γ influence the performances of the methods. With large numbers of target proportions and $m = 10$, the NPI-3G-Y method demonstrated squeezing behaviour for the middle group. It might be of interest to study the optimal choices of these numbers in practical situations using this method, as more attention should be paid to it due to its potential to squeeze the middle group. This topic is left for future research.

5.9 Concluding remarks

In this chapter, the NPI-based methods presented in Chapter 4 are extended to three-group classification problems. The proposed methods for selecting optimal diagnostic thresholds for ordinal test outcomes are based on considering multiple future individuals. For each NPI approach, the optimal thresholds were determined by taking into account a given number of future observations and criteria according to the target proportion of successful diagnoses in each group. Throughout this chapter, both cases with equal numbers of future observations for each group and another with different numbers of future observations for each group were considered.

The proposed methods were illustrated with an example based on data from the literature, considering different scenarios of the target proportions α, β and γ, and their performances were evaluated through simulations. The proposed methods were compared with classical methods, including the EYI-3G and EMV-3G methods. The results showed that, the middle group may have poor predictive performance in the three-group scenario where the NPI-3G-Y method is used. It is possible to overcome this problem by using the NPI-3G method, since the method tries to balance the three groups. Additionally, the results showed that the optimal thresholds for a given diagnostic test is dependent on the values of α, β and γ, as well as the value of m.

This line of work provides opportunities for future research. For example, extending the methods presented in this chapter to $G > 3$ groups, so that there will be $G-1$ thresholds, $k_1 < k_2 < \ldots < k_{G-1}$ in $\{1, \ldots, K\}$. For more than three groups, Nakas et al. [74] introduced the Youden index method for G groups, which maximises the probabilities of correctly classifying individuals into each group. As this method involves optimizing multiple thresholds, the thresholds can end up very close to each other, making squeezing more likely. However, the generalisation of the NPI method may be more effective at reducing the impact of this squeezing problem. Coolen-Maturi [39] introduced NPI with more than three ordered groups for a single future observation and by considering the definitions and notation presented in the paper, the NPI-3G method presented in this chapter can be generalised to more than three ordered groups. In order to study this topic with more than three groups, an optimization method needs to be developed first to efficiently determine the optimal thresholds, rather than going through all possible combinations of thresholds. This optimization is expected to enhance computational efficiency and speed up the process of finding the optimal thresholds, especially when dealing with a large number of groups.

Chapter 6

Conclusions

This thesis presented contributions to statistical methods for ordinal data using the NPI method with multiple future observations. This chapter provides a brief overview of the main results presented in this thesis and highlights potential future research directions.

In Chapter 2, the NPI lower and upper probabilities for several events of interest were derived. Initially, an event involving two future observations was considered, one in a specific category and the other in the remaining categories. Next, the event that both observations are in a specific category was considered. The methodology, involving two future observations, was then generalised to m future observations. The focus was then directed towards the event where at least a given number out of m future observations are in a specific category, using a path counting method. Finally, this was extended to include adjoining categories, forming a single interval on the real-line. This path counting method can also be applied to derive NPI lower and upper probabilities for other events.

In Chapter 3, applications of NPI to selection problems with ordinal data were presented. NPI methods were introduced for selecting both a single category and choosing a subset of categories based on multiple future observations. Selection events of interest included selecting a specific category, achieving a specified criterion for the event that at least a given number out of m future observations are in that category. Another aim was to present the selection of a minimal-sized subset of adjoining categories, achieving a specified criterion for the event that at least a certain number of future observations within that subset meet the criterion. We have seen that the selection of a category or a subset of categories might change if the numbers of future observations change. Additionally, comparison of two groups of ordinal data using the sampling of orderings method was presented to estimate the NPI lower and upper probabilities, as their exact calculation can be computationally infeasible due to the large number of possible orderings of future observations. The aim was to compare the number of future observations within categories from the first group to the number within the same categories from the second group, with different ordering scenarios of future observations. The results of the category selection and pairwise comparison events vary depending on the number of future observations considered, illustrating how this number impacts the NPI lower and upper probabilities.

In Chapter 4, NPI-based methods were presented for selecting the optimal diagnostic test threshold for two-group classification, which were extended to a scenario with three-group classification in Chapter 5. These methods considered m future individuals in each group, along with criteria based on each group's target proportion of successful diagnoses. Results indicated that the optimal thresholds for a diagnostic test depended on both the target success proportions and the value of m. Decisions regarding the optimal selection of these values in real-world scenarios are left for future investigation. Examples from the literature were used to illustrate the methods, and their performances were assessed via simulations. In these evaluations, it was shown how the performances of the methods can vary depending on the number of successful diagnoses and the number of m individuals. Furthermore, when using the classical Youden's index approach on three-group situations, one of the groups may have poor predictive performance, but this can be avoided by applying the NPI-based methods presented in this chapter. A comparison of NPI-based methods presented in Chapters 4 and 5 has been conducted with the classical Youden index, Liu index and the maximum volume methods. It will be of interest to compare the NPI-based methods with those discussed in Section 1.4, such as the index of union or the close-to-(0,1) method, but this is left for future research.

There are many interesting and challenging topics based on the work presented in this thesis. One such challenge is the development of NPI-based methods, similar to those in Chapter 3, for selection problems and pairwise comparisons involving more complex events, using the path counting technique to derive NPI lower and upper probabilities. For example, developing an event where a specific, or at least a number, of future observations are in non-adjoining categories. The idea of subset selection can also be extended to other events of interest. An example would be the development of NPI-based methods using the path counting technique to select subsets of future observations for the event that the number of future observations is known for some categories within that subset but unknown for a particular category within that subset.

Coolen-Maturi [38] introduced the NPI approach for three-group ROC surfaces with ordinal outcomes, offering a novel perspective on evaluating the accuracy of diagnostic tests. Developing a similar method using ROC curves, while considering multiple future observations for each group will be an interesting research direction. Comparing the accuracy of two diagnostic tests, an important aim in medical research, traditionally involves metrics like specificity, sensitivity, or the area under the ROC curve [5]. However, such comparisons can be challenging, as one test might show higher specificity, while the other might demonstrate higher sensitivity, making the comparison less straightforward. Further study could investigate the application of NPI methods with ordinal data for comparing two diagnostic tests for two or three groups. NPI methods have been presented for selecting optimal thresholds for two- and three-group classification problems. These methods were shown to depend on the target success proportions, α, β, and γ, as well as the value of m. Further research could explore how these values can be selected in real-world applications. A method for selecting these values while taking misclassification costs into account would be particularly useful, as minimizing such costs is an important aspect of classification problems. These research directions, including consideration of multiple future observations, have the potential to provide new insights into ordinal data, offering substantial practical value from a predictive perspective.

Appendix: Orderings of future observations

This appendix provides detailed explanations for each ordering presented in Example 2.4. Recall that, in the latent variable representation, as explained in Section 2.3, the category C_k is assumed to be represented by the interval IC_k for $k = 1, \ldots, K$, and the n observations are assumed to be represented by $y_1 < \ldots < y_n$, of which n_k are in IC_k, these are also denoted by y_i^k for $i = 1, \ldots, n_k$. There are $I_j = (y_{j-1}, y_j)$, for $j = 1, \ldots, n+1$, and with $n_k \geq 1$ for all k, each IC_k has an interval I_j to its left and an interval I_j to its right where future observations in these intervals may or may not be assigned to IC_k. Of course, only one interval is assigned to IC_1 or IC_K. Categories with $n_k = 0$ do not have any intervals assigned.

Firstly, for the event $M_2 = 2$ with $m = 3$, the total number of future observations that are in the boundary intervals should be equal to 0. There are 3 orderings in which the 2 future observations (m_2) are in the intervals I_j that are assigned to IC_2 and 6 orderings where the 1 future observation is in the remaining I_j intervals. Figure A1 represents the 3 orderings for the 2 future observations in IC_2 (the blue dots). These orderings are as follows: firstly, one future observation could be in the interval (y_3, y_4) and one in (y_4, y_5); secondly, both future observations could be in (y_3, y_4); and thirdly, both observations could be in (y_4, y_5). For each of these orderings, there are 6 orderings for the 1 future observation among the remaining intervals. Multiplying these orderings 3 × 6 results in a total of 18 orderings, as shown in Figure A1.

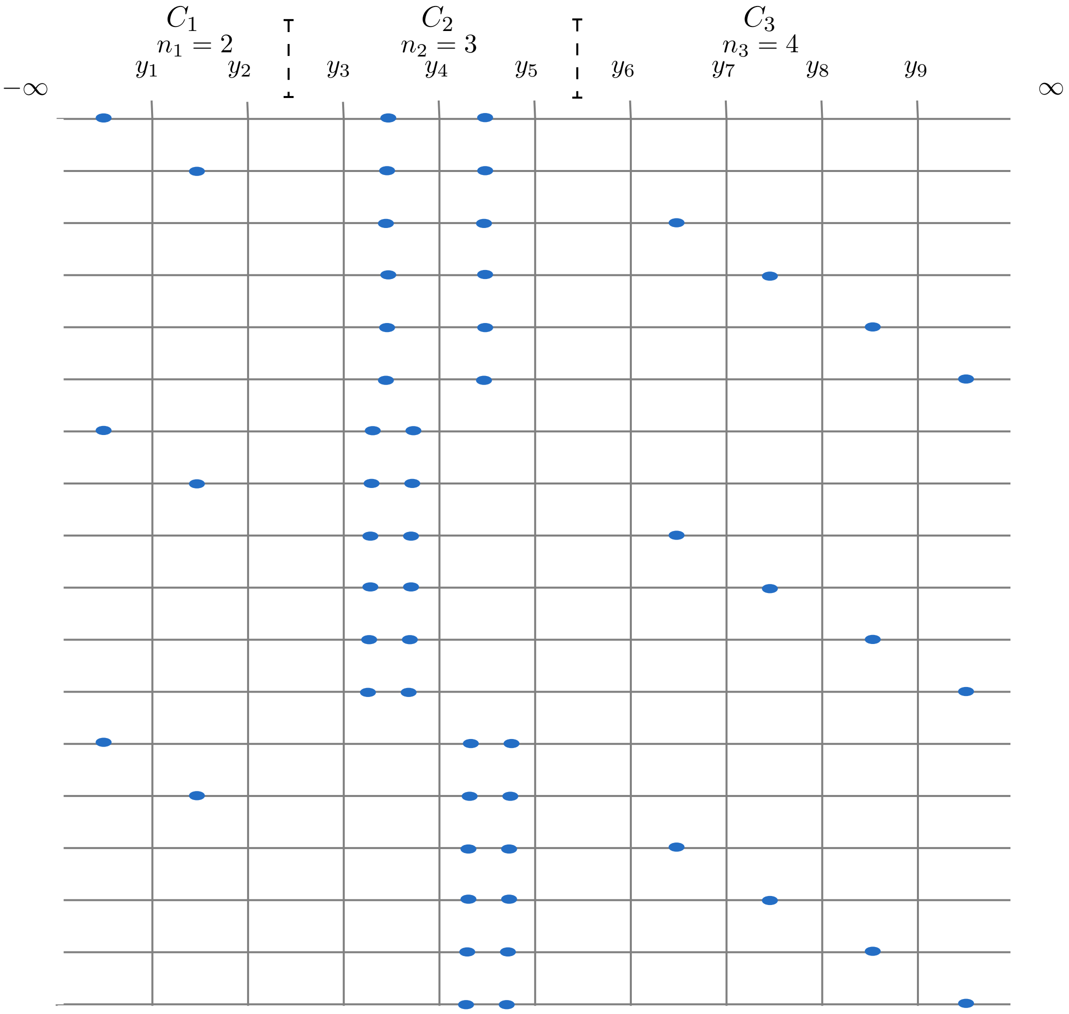
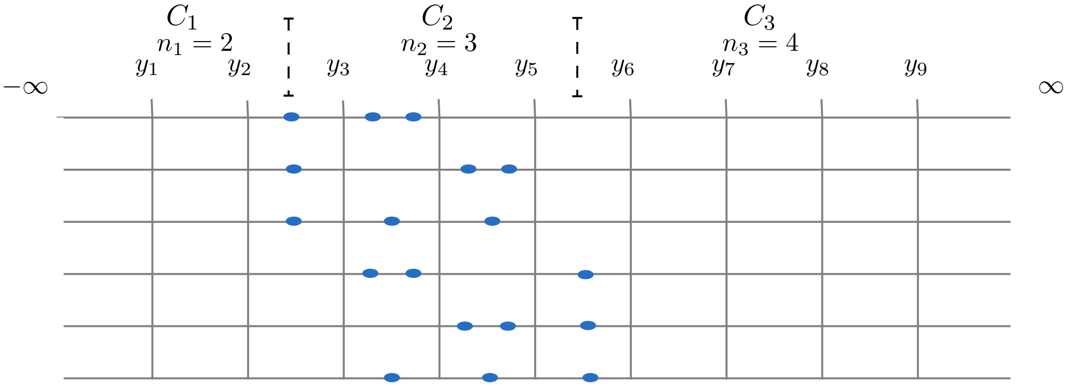


Figure A1: Orderings of future observations for which the event $M_2 = 2$ is possible when no future observations are in the boundary intervals

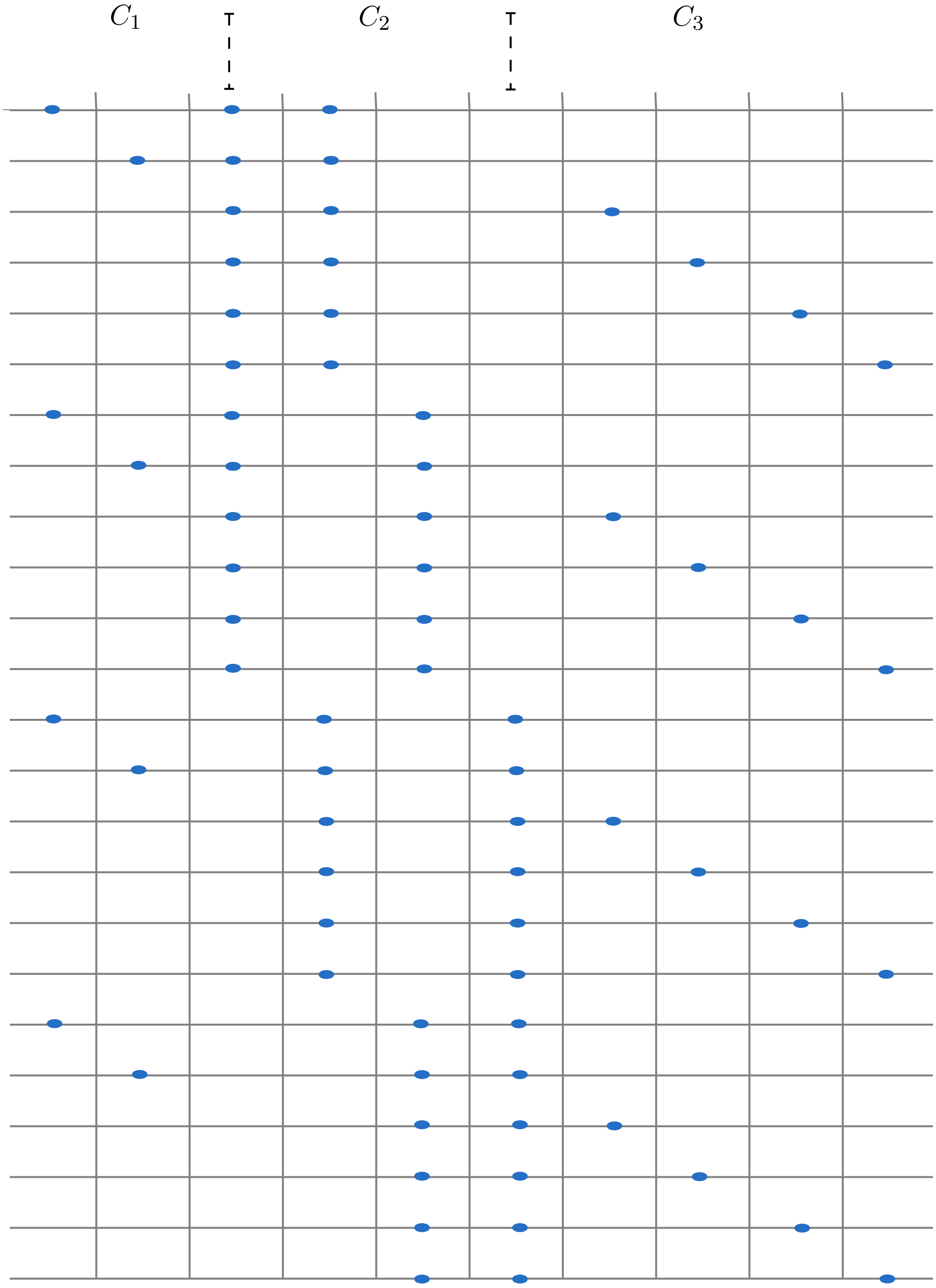
Secondly, recall that W represents the total number of future observations in the boundary intervals. There are $W + 1$ ways in which W can be in the boundary intervals. Also, recall that D represents the number of future observations in the boundary intervals that can be counted as belonging to IC_2. For the case with $D = 0$, this indicates that the two future observations (m_2) cannot be in the boundary intervals of IC_2 as $D = 0$. Consequently, when $W = 1$ and $D = 0$, this W in the boundary intervals can either belong to IC_1 or IC_3 depending on its specific location, whether it is on the right of IC_1 or on the left of IC_3. There are 6 orderings in total that can lead to $M_2 = 2$. These orderings are as follows: firstly, both future observations (m_2) could be in (y_3, y_4); secondly, both observations could be in (y_4, y_5); and thirdly, one future observation could be in the interval (y_3, y_4) and one in (y_4, y_5).

Regarding the W in the boundary intervals, there are two possible ways in which W can be in the boundary intervals. Multiplying 3 by 2 leads to a total of 6 orderings for the case $W = 1$ and $D = 0$, as can be seen in Figure A2(a). For the case $W = 1$ and $D = 1$, this indicates that one of the two future observation (m_2) can be in the boundary intervals as $W = D = 1$, resulting in two possible ways in which D can be in the boundary intervals. For $m_2 - D = 1$, this observation could be in the interval (y_3, y_4) or could be in (y_4, y_5), yielding two possible orderings. Additionally, there are 6 orderings where the $m - m_2 - (W - D) = 1$ future observation can be in the remaining I_j intervals. Multiplying these orderings $2 \times 2 \times 6$ leads to a total of 24 orderings for the case $W = 1$ and $D = 1$ as shown in Figure A2(b). Combining the total orderings for both cases of W when $W = 1$ sums up to 30 orderings.

Similarly, for $W = 2$ with $D = 1$ and $D = 2$, as presented in Figure A3(a) and (b), respectively, there are a total of 6 orderings for the case with $D = 1$. For the case with $D = 2$, there are 18 orderings in total. By summing the total orderings for both W cases with $W = 2$, the total number of orderings equals 24.

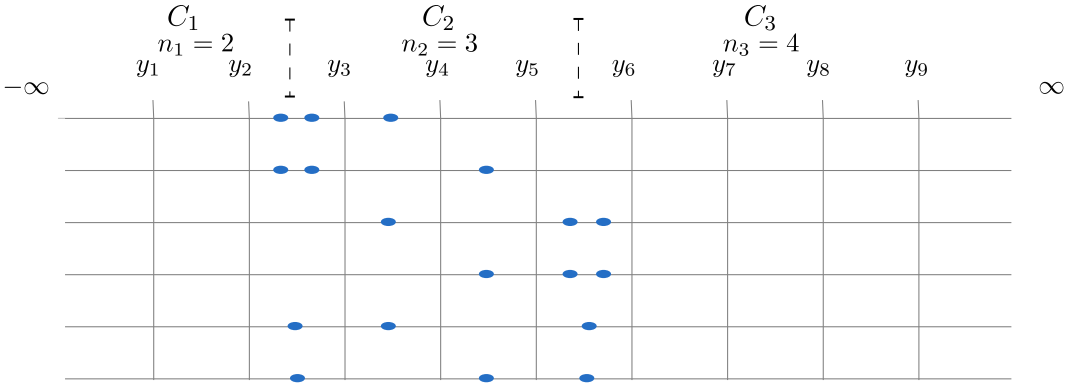


(a) $W = 1$ and $D = 0$

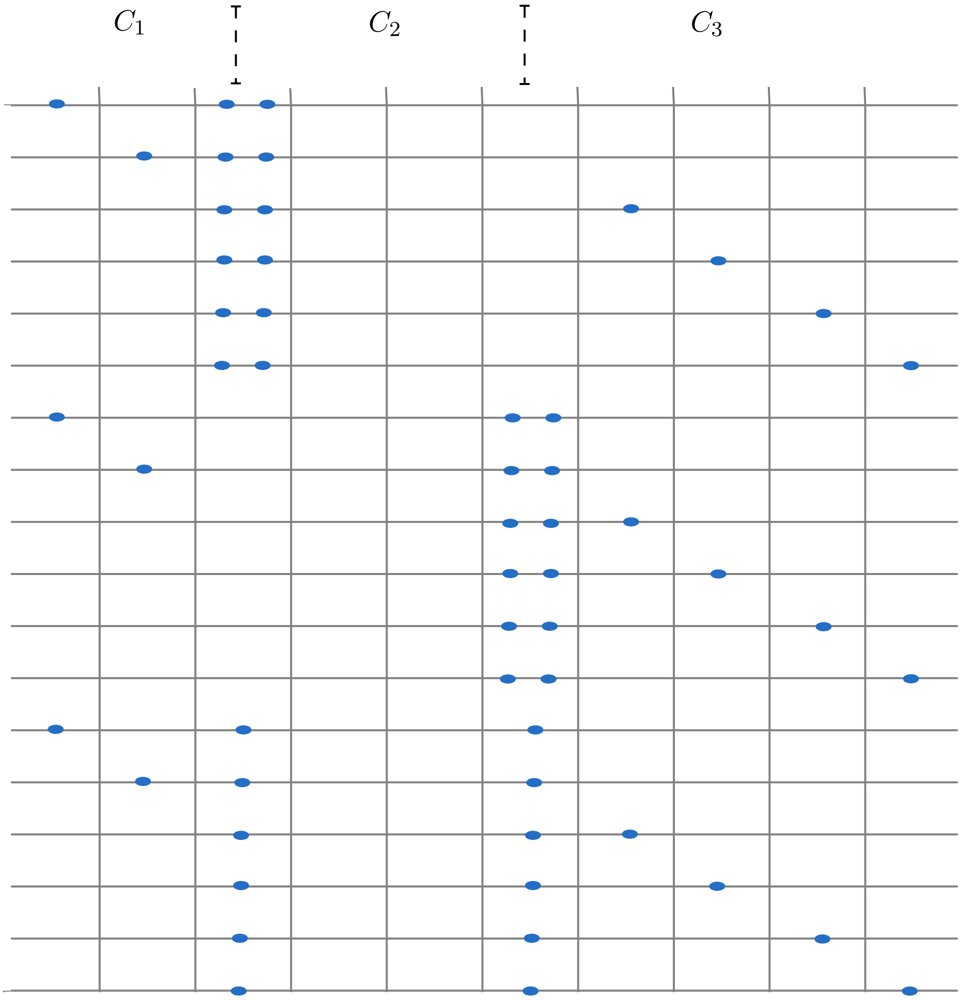


(b) $W = 1$ and $D = 1$

Figure A2: Orderings of future observations for which the event $M_2 = 2$ is possible when $W = 1$ and $D = 0, 1$



(a) $W = 2$ and $D = 1$



(b) $W = 2$ and $D = 2$

Figure A3: Orderings of future observations for which the event $M_2 = 2$ is possible when $W = 2$ and $D = 1, 2$

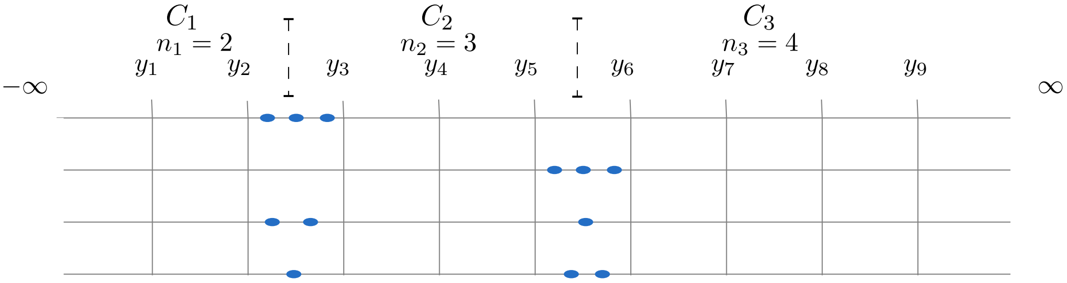


Figure A4: Orderings of future observations for which the event $M_2 = 2$ is possible when $W = 3$ and $D = 2$

Finally, for $W = 3$ with $D = 2$ as presented in Figure A4, there are 4 orderings in total. We could have all three future observations either in the left or right boundary interval. We could also have one future observation in the left boundary interval and the other two in the right boundary interval. Lastly, we could have one future observation in the right boundary interval and the other two in the left boundary interval. Adding all these orderings together results in a total of [18+30+24+4]=76 orderings. This is in line with the results derived in Example 2.4.

Bibliography

[1] Aboalkhair A. (2012). *Nonparametric Predictive Inference for System Reliability.* PhD thesis, Durham University, Durham, UK. Available at: `https://etheses.dur.ac.uk/3918/`.

[2] Agresti A. (2010). *Analysis of Ordinal Categorical Data.* Hoboken, NJ: Wiley.

[3] Agresti A. (2018). *An Introduction to Categorical Data Analysis.* Hoboken, NJ: Wiley.

[4] Alabdulhadi M. (2018). *Nonparametric Predictive Inference for Diagnostic Test Thresholds.* PhD thesis, Durham University, Durham, UK. Available at: `https://etheses.dur.ac.uk/12538/`.

[5] Alabdulhadi M., Coolen-Maturi T. and Coolen F.P.A. (2021). Nonparametric predictive inference for comparison of two diagnostic tests. *Communications in Statistics-Theory and Methods*, 50, 4470–4486.

[6] Alrasheedi M. (2023). *Optimal Thresholds for Classification Trees using Nonparametric Predictive Inference.* PhD thesis, Durham University, Durham, UK. Available at: `https://etheses.dur.ac.uk/14793/`.

[7] Altman D.G. (1991). *Practical Statistics for Medical Research.* London: Chapman and Hall.

[8] 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.

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

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

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

[12] Baker R.M. (2010). *Multinomial Nonparametric Predictive Inference: Selection, Classification and Subcategory Data.* PhD thesis, Durham University, Durham, UK. Available at: `https://etheses.dur.ac.uk/257/`.

[13] Baker R.M. and Coolen F.P.A. (2010). Nonparametric predictive category selection for multinomial data. *Journal of Statistical Theory and Practice*, 4, 509–526.

[14] Bamber D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12, 387–415.

[15] Barry M.J. and Edgman-Levitan S. (2012). Shared decision making-the pinnacle of patient-centered care. *New England Journal of Medicine*, 366, 780–781.

[16] Bechhofer R., Santner T. and Goldsman D. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening and Multiple Comparisons.* New York: Wiley.

[17] Bechhofer R.E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, 25, 16–39.

[18] Böhning D., Holling H. and Patilea V. (2011). A limitation of the diagnostic-odds ratio in determining an optimal cut-off value for a continuous diagnostic test. *Statistical Methods in Medical Research*, 20, 541–550.

[19] Boole G. (1854). *An Investigation of the Laws of Thought: On Which Are Founded the Mathematical Theories of Logic and Probabilities*. London: Walton and Maberly.

[20] Boone H.N. and Boone D.A. (2012). Analyzing Likert data. *Journal of Extension*, 50, 1–5.

[21] Bradley A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30, 1145–1159.

[22] Brown L.D., Cai T.T. and DasGupta A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16, 101–133.

[23] Campochiaro C., Della-Torre E., Cavalli G., De Luca G., Ripa M., Boffini N., Tomelleri A., Baldissera E., Rovere-Querini P. and Ruggeri A. (2020). Efficacy and safety of tocilizumab in severe COVID-19 patients: a single-centre retrospective cohort study. *European Journal of Internal Medicine*, 76, 43–49.

[24] Chen J., Coolen F.P.A. and Coolen-Maturi T. (2019). On nonparametric predictive inference for asset and European option trading in the binomial tree model. *Journal of the Operational Research Society*, 70, 1678–1691.

[25] Chowdhury M.Z.I. and Turin T.C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, 8, e000262.

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

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

[28] Coolen F.P.A. (2011). Nonparametric Predictive Inference. In: *International Encyclopedia of Statistical Science*, (Editor) M. Lovric, pp. 968–970. Berlin: Springer.

[29] Coolen F.P.A. and Augustin T. (2009). A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories. *International Journal of Approximate Reasoning*, 50, 217–230.

[30] Coolen F.P.A. and Coolen-Schrijner P. (2006). Nonparametric predictive subset selection for proportions. *Statistics and Probability Letters*, 76, 1675–1684.

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

[32] Coolen F.P.A., Coolen-Schrijner P., Coolen-Maturi T. and Elkhafifi F.F. (2013). Nonparametric predictive inference for ordinal data. *Communications in Statistics-Theory and Methods*, 42, 3478–3496.

[33] Coolen F.P.A., Coolen-Schrijner P. and Maturi T.A. (2010). On nonparametric predictive inference for ordinal data. In: *Computational Intelligence for Knowledge-Based Systems Design, Proceedings 13th International Conference on Information Processing and Management of Uncertainty*, (Editors) E. Hüllermeier, R. Kruse and F. Hoffmann, pp. 188–197. Berlin: Springer.

[34] Coolen F.P.A. and Marques F.J. (2020). Nonparametric Predictive Inference for Test Reproducibility by Sampling Future Data Orderings. *Journal of Statistical Theory and Practice*, 14, 62.

[35] Coolen F.P.A. and Maturi T.A. (2010). Nonparametric predictive inference for order statistics of future observations. In: *Combining Soft Computing and Statistical Methods in Data Analysis*, (Editors) C. Borgelt, G. González-Rodríguez, W. Trutschnig, M.A. Lubiano, M.A. Gil, P. Grzegorzewski and O. Hryniewicz, pp. 97–104. Berlin: Springer.

[36] 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.

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

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

[39] Coolen-Maturi T. and Coolen F.P.A. (2019). Nonparametric predictive inference for the validation of credit rating systems. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182, 1189–1204.

[40] Coolen-Maturi T., Coolen F.P.A. and Alabdulhadi M. (2020). Nonparametric predictive inference for diagnostic test thresholds. *Communications in Statistics-Theory and Methods*, 49, 697–725.

[41] 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.

[42] Coolen-Maturi T., Coolen-Schrijner P. and Coolen F.P.A. (2011). Nonparametric predictive selection with early experiment termination. *Journal of Statistical Planning and Inference*, 141, 1403–1421.

[43] 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.

[44] Coolen-Maturi T., Coolen-Schrijner P. and Coolen F.P.A. (2012). Nonparametric predictive multiple comparisons of lifetime data. *Communications in Statistics-Theory and Methods*, 41, 4164–4181.

[45] De Finetti B. (1974). *Theory of Probability.* Chichester: Wiley.

[46] 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.

[47] Elkhafifi F.F. (2012). *Nonparametric Predictive Inference for Ordinal Data and Accuracy of Diagnostic Tests*. PhD thesis, Durham University, UK. Available at: https://etheses.dur.ac.uk/3914/.

[48] 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.

[49] Elwyn G., Frosch D., Thomson R., Joseph-Williams N., Lloyd A., Kinnersley P., Cording E., Tomson D., Dodd C. and Rollnick S. (2012). Shared decision making: a model for clinical practice. *Journal of General Internal Medicine*, 27, 1361–1367.

[50] Filipek-Gliszczyńska A., Barczak A., Budziszewska M., Mandecka M., Gabryelewicz T. and Barcikowska M. (2018). The Erlangen score algorithm in the diagnosis and prediction of the progression from subjective cognitive decline and mild cognitive impairment to Alzheimer-type dementia. *Folia Neuropathologica*, 56, 88–96.

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

[52] Gibbons J.D. and Chakraborti S. (2014). *Nonparametric Statistical Inference: Revised and Expanded*. Boca Raton, FL: CRC Press.

[53] Gibbons J.D., Olkin I. and Sobel M. (1999). *Selecting and Ordering Populations: A New Statistical Methodology*. Philadelphia: Society for Industrial and Applied Mathematics (SIAM).

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

[55] Gupta A.K. and Nadarajah S. (2004). *Handbook of Beta Distribution and Its Applications*. Boca Raton, FL: CRC Press.

[56] Gupta S.S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics*, 7, 225–245.

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

[58] Hawass N. (1997). Comparing the sensitivities and specificities of two diagnostic procedures performed on the same group of patients. *The British Journal of Radiology*, 70, 360–366.

[59] He T., Coolen F.P.A. and Coolen-Maturi T. (2019). Nonparametric predictive inference for European option pricing based on the binomial tree model. *Journal of the Operational Research Society*, 70, 1692–1708.

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

[61] Horn S.D., Horn R.A. and Sharkey P.D. (1984). The severity of illness index as a severity adjustment to diagnosis-related groups. *Health Care Financing Review*, 1984, 33–45.

[62] Kersey J., Samawi H., Yin J., Rochani H. and Zhang X. (2023). On diagnostic accuracy measure with cut-points criterion for ordinal disease classification based on concordance and discordance. *Journal of Applied Statistics*, 50, 1772–1789.

[63] Likert R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22, 5–55.

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

[65] Lu C.J. (2014). Sales forecasting of computer products based on variable selection scheme and support vector regression. *Neurocomputing*, 128, 491–499.

[66] Mantelakis A., Spiers H.V., Lee C.W., Chambers A. and Joshi A. (2021). Availability of personal protective equipment in NHS hospitals during COVID-19: A national survey. *Annals of Work Exposures and Health*, 65, 136–140.

[67] Martin S. (1977). The evaluation of tests. *Canadian Journal of Comparative Medicine*, 41, 19–25.

[68] Mintoff D., Borg I. and Pace N.P. (2022). Serum immunoglobulin G is a marker of hidradenitis suppurativa disease severity. *International Journal of Molecular Sciences*, 23(22), 13800.

[69] Moore D.S., McCabe G.P. and Craig B.A. (2017). *Introduction to the Practice of Statistics.* New York, NY: W.H. Freeman and Company.

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

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

[72] 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.

[73] Nakas C.T., Bantis L.E. and Gatsonis C.A. (2023). *ROC Analysis for Classification and Prediction in Practice.* Milton: CRC Press.

[74] 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.

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

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

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

[78] Perkins N.J. and Schisterman E.F. (2006). The inconsistency of "optimal" cut-points obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, pp. 670–675.

[79] Powers D. and Xie Y. (2008). *Statistical Methods for Categorical Data Analysis*. Bingley, UK: Emerald Group Publishing.

[80] R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[81] Rubio-Rivas M., Mora-Luján J.M., Formiga F., Arévalo-Cañas C., Lebrón Ramos J.M., Villalba García M.V., Fonseca Aizpuru E.M., Díez-Manglano J., Arnalich Fernández F. and Romero Cabrera J.L. (2022). WHO ordinal scale and inflammation risk categories in COVID-19. Comparative study of the severity scales. *Journal of General Internal Medicine*, 37, 1980–1987.

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

[83] Schisterman E.F., Faraggi D., Reiser B. and Hu J. (2008). Youden index and the optimal threshold for markers with mass at zero. *Statistics in Medicine*, 27, 297–315.

[84] Serlin R.C., Mendoza T.R., Nakamura Y., Edwards K.R. and Cleeland C.S. (1995). When is cancer pain mild, moderate or severe? Grading pain severity by its interference with function. *Pain*, 61, 277–284.

[85] Shaked M. and Shanthikumar J.G. (2007). *Stochastic Orders*. New York, NY: Springer.

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

[87] Unal I. (2017). Defining an optimal cut-point value in ROC analysis: an alternative approach. *Computational and Mathematical Methods in Medicine*, 2017, 3762651.

[88] Walley P. (1991). *Statistical Reasoning with Imprecise Probabilities.* London: Chapman and Hall.

[89] Wang Y., Fan G., Horby P., Hayden F., Li Q., Wu Q., Zou X., Li H., Zhan Q., Wang C. and Network C.C. (2019). Comparative outcomes of adults hospitalized with seasonal influenza A or B virus infection: Application of the 7-category ordinal scale. *Open Forum Infectious Diseases*, 6, 1–9. Article ID: ofz053.

[90] Wang Y., Fan G., Salam A., Horby P., Hayden F.G., Chen C., Pan J., Zheng J., Lu B. and Guo L. (2020). Comparative effectiveness of combined favipiravir and oseltamivir therapy versus oseltamivir monotherapy in critically ill patients with influenza virus infection. *The Journal of Infectious Diseases*, 221, 1688–1698.

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

[92] Wickham H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Cham, Switzerland: Springer Cham.

[93] 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.

[94] Yerli H., Aydin E., Haberal N., Harman A., Kaskati T. and Alibek S. (2010). Diagnosing common parotid tumours with magnetic resonance imaging including diffusion-weighted imaging vs fine-needle aspiration cytology: a comparative study. *Dentomaxillofacial Radiology*, 39, 349–355.

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

[96] Zhou X.H., McClish D.K. and Obuchowski N.A. (2009). *Statistical Methods in Diagnostic Medicine.* Hoboken, NJ: Wiley.