Spring 2020

Generalization of Kullback-Leibler Divergence for Multi-Stage Diseases: Application to Diagnostic Test Accuracy and Optimal Cut-Points Selection Criterion

Chen Mo

Recommended Citation

Mo, Chen, "Generalization of Kullback-Leibler Divergence for Multi-Stage Diseases: Application to Diagnostic Test Accuracy and Optimal Cut-Points Selection Criterion" (2020). *Electronic Theses and Dissertations*. 2046.
https://digitalcommons.georgiasouthern.edu/etd/2046

GENERALIZATION OF KULLBACK-LEIBLER DIVERGENCE FOR MULTI-STAGE DISEASES: APPLICATION TO DIAGNOSTIC TEST ACCURACY AND OPTIMAL CUT-POINTS SELECTION CRITERION

by

CHEN MO

(Under the Direction of Hani M. Samawi)

ABSTRACT

The Kullback-Leibler (KL) divergence, which captures the disparity between two distributions, has been considered as a measure for determining the diagnostic performance of an ordinal diagnostic test. This study applies the Kullback-Leibler (KL) divergence and further generalizes it to comprehensively measure the diagnostic accuracy test for multi-stage ($K > 2$) diseases, named generalized total Kullback-Leibler (GTKL) divergence. Additionally, the GTKL can be used as an optimal cut-point selection criterion for discriminating subjects among different stages. Moreover, the study investigates a variety of applications of the GTKL divergence on measuring the rule-in/out potentials in the single-stage and multi-stage levels. Furthermore, the study compares the GTKL divergence with other diagnostic measures such as the generalized Youden index (GYI), hypervolume under the manifold (HUM), and maximum absolute determinant (MADET). Intensive simulation studies were conducted to investigate the performance of the proposed measure comparative to other methods in the literature. Finally, a comprehensive analysis of a real dataset was performed to illustrate the application of the proposed measure.

INDEX WORDS: Diagnostic test, Biomarker, Cut-point selection, ROC, AUC, Youden index, Kullback-Leibler divergence, Multi-stage, HUM, VUS, Generalized Youden index, MADET, Closed-to-perfection, Maximum volume, Predictive values, Likelihood ratio, Rule-in/out potentials.

GENERALIZATION OF KULLBACK-LEIBLER DIVERGENCE FOR MULTI-STAGE DISEASES: APPLICATION TO DIAGNOSTIC TEST ACCURACY AND OPTIMAL CUT-POINTS SELECTION CRITERION

by

CHEN MO

B.S., University of Oklahoma, 2011

M.P.H., Georgia Southern University, 2015

A Dissertation Submitted to the Graduate Faculty of Georgia Southern University in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PUBLIC HEALTH

GENERALIZATION OF KULLBACK-LEIBLER DIVERGENCE FOR MULTI-STAGE DISEASES: APPLICATION TO DIAGNOSTIC TEST ACCURACY AND OPTIMAL CUT-POINTS SELECTION CRITERION

by

CHEN MO

Major Professor: Hani M. Samawi

Committee: Jingjing Yin
Haresh D. Rochani
Xinyan Zhang

Electronic Version Approved:

May 2020

DEDICATION

To my parents, Xian-Zhong Mo, Chun-Lan Chen, and aunt Jessica Hightower. Your love and support are my most powerful strengths facing the challenges.

In memory of my paternal grandfather, Ding-You Mo,
maternal grandfather, Shang-Shu Chen, maternal grandmother, Feng-yue Huang and
uncle Willie Hightower.

ACKNOWLEDGMENTS

During my D.P.H. program, this dissertation would have never been completed without the assistance from my mentors, collegeaus, friends, and family.

First and foremost, I would like to thank my committee chair, Dr. Hani M. Samawi for his patient, illuminated, and comprehensive guidance and support. He offered a great deal of time and effort to guide me through the academic disquisitions. I very much appreciate his advice for helping me wipe off all bramble on the way toward the completion of this research.

With this opportunity, I would like to extend my sincere thanks to my dissertation committee members, Dr. Jingjing Yin, Dr. Haresh D. Rochani, and Dr. Xinyan Zhang, for investing their precious time on reading this dissertation and providing suggestions to improve the work.

I would love to give my special appreciation to Dr. Robert L. Vogel, who taught my first class in Biostatistics and gave me wise advice on my career; Dr. Tarasenko, who has provided numerous assistances in developing my skills in public health research; and Dr. Jingjing Yin, who enlightened the biostatistical research during my study here.

Also, plentiful thanks to all the other faculties, staffs and colleagues in the Jiann-Ping Hsu College of Public Health at Georgia Southern University for their generous assistance. Lastly, my thanks go to my parents and my best friends, Jingyu Qin, Su Ting, and Sarbesh Pandeya, for their love and support.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1
INTRODUCTION

Diagnostic tests play an essential role in health care, including medical diagnosis, screening tests, and research. There are three purposes of performing a diagnostic test: 1) to provide reliable information about a patient's health condition; 2) to influence the treatment plan for the patient from health providers; and 3) to understand disease mechanisms and natural history via research (McNeil & Adelstein, 1976; Sox Jr et al., 1989; Zhou, McClish, & Obuchowski, 2009). Thus, a good diagnostic test is essential to discriminate between diseased and non-diseased subjects and provides a strong understanding of patients' health condition. A diagnostic test that generates continuous results can also be referred to as a continuous biomarker. When using a continuous biomarker to discriminate subjects to diseased or non-diseased, a biomarker will be dichotomized with a specific value. The particular value of the continuous biomarker is the cut-point that separates subjects into diseased and non-diseased groups. In this case, the diagnostic test and biomarker can be used interchangeably. Choosing a cut-point that can best discriminate subjects is essential and challenging for clinicians to correctly identify subjects with the disease of interest and provide appropriate treatments. Consequently, a good criterion or procedure of an optimal cut-point selection is also necessary for medical diagnosis.

Studies have examined the efficiency of different diagnostic tests; however, the test results are not always accurate (Akobeng, 2007; Altman & Bland, 1994; Deeks & Altman, 2004; Pepe, 2003; Šimundić, 2009; Wong & Lim, 2011; Zhou et al., 2009). In the case of misclassification, a medical diagnostic test can give a positive result for a subject who does not have the disease (Pepe, 2003; Zhou et al., 2009). Vice versa, a diseased subject may be diagnosed as non-diseased. The test's diagnostic accuracy is the ability to discriminate among alternative states of health. Health providers need to assess the performance of the diagnostic tests on discriminating patients with and without the disease of interest to better determine the true stage of the patients. The four basic measures of diagnostic accuracy to evaluate the accuracy of the diagnostic test are sensitivity, specificity, false positive rate (FPR), and false negative rate (FNR). Among these measures, sensitivity and specificity are the correct classification rates, and the FPR and the FNR are the misclassification rates (Pepe, 2003; Zhou et al., 2009). Generally, a measure of diagnostic accuracy attempts to maximize the correct classification rates (i.e., sensitivity and specificity) and minimize the misclassification rates (i.e., the FPR and the FNR). A test's sensitivity is its ability to detect the disease when it is present, and a test's specificity is its ability to exclude the condition among subjects without the disease. A diagnostic test is used to determine the presence or absence of a specific disease when a subject shows significant symptoms of that disease. The diagnostic test is an important determinant for the health care providers to decide whether to give interventions of the disease, especially when the interventions are invasive or harmful such as chemotherapy (Gilbert, Logan, Moyer, & Elliott, 2001). A screening test is designed to identify asymptomatic subjects at sufficient risk of the disease to warrant further health interventions among the population who have not received medical attention (Gilbert et al., 2001). Usually, the diagnostic test is performed after a screening test to make a definite diagnosis. Different tests are carried out to discriminate subjects between diseased and non-diseased conditions based on sensitivity and specificity measures, including sensitivity and specificity. The Receiver Operating Characteristic curve (ROC) and the area under the ROC curve (AUC) provide summary measures associated with single sensitivity and specificity pairs by including all the decision thresholds. Additionally, some of the measures incorporate sensitivity and specificity into a single index like accuracy (also called diagnostic effectiveness), such as diagnostic odds ratio (OR), and Youden index, overlap measure (Samawi, Yin, Rochani, & Panchal, 2017), and KL divergence. Similar to sensitivity and specificities, the OR and the Youden index do not depend on disease prevalence yet are affected by the spectrum of a disease, such as a disease severity, phase, stage, and comorbidity (Zhou et al., 2009). On the other hand, the accuracy is affected by disease prevalence (Šimundić, 2009; Zhou et al., 2009). The accuracy of a test increases as the disease prevalence decreases with the same sensitivity and specificity. It means that the accuracy estimated from a population cannot be generalized to another population with different disease prevalence. In addition to the measures that have been mentioned above, predictive values and diagnostic likelihood ratios (LRs) are also measures of diagnostic accuracy (Šimundić, 2009; Zhou et al., 2009).. The predictive values give significant clinical implication about a diagnostic test. Although measures like sensitivity and specificity give an estimation of the probability of the disease in patients, they cannot answer how likely the patients would receive positive or negative test results. The predictive values are the measures that provide information about the probability that a test result gives the correct diagnosis. The positive predictive value (PPV) shows the probability of having the disease of interest in a subject given a positive test result, and the negative predictive value (NPV) gives the probability that a subject receives a negative result yet not having the disease of interest (Altman & Bland, 1994; Wong & Lim, 2011). These measures highly depend on the disease prevalence, which cannot be generalized among different populations with different disease prevalence. Compared to the predictive values, the LRs can also provide information about the probability that a subject can be correctly diagnosed; nevertheless, the LRs do not depend on prevalence as the predictive values, and they are applicable to other clinical settings for the same disease (Boyko, 1994; Deeks & Altman, 2004).

Also, the LRs are the best indicator for rule-in/out diagnosis (Boyko, 1994; Deeks & Altman, 2004; Gilbert et al., 2001). Particularly, a rule-in test assesses if the results from a diagnostic test will include the possibility that a subject has the disease of interest. A positive response from a specificity test makes the presence of the disease more likely since it is specific to that disease. Whereas, a rule-out test, based on sensitivity, emphasizes in assessing if the test results will exclude the possibility that a subject is non-diseased. In addition to the LRs, the KL divergence is another measure that has the rule-in/out potentials (Lee, 1999). Lee (1999) proposed the KL divergence, a weighted average of the $\log LR$, and suggested that the KL divergence is capable of predicting the fate of an average subject before the diagnostic test. Although the LRs can only estimate the rule-in/out potentials after the diagnostic test is conducted (i.e., after-test), the KL divergence predicts the rule-in/out potentials from another perspective (Lee, 1999). Furthermore, Samawi et al. (2019) suggested that the total sum of KL divergences, for potential rule-in/out, (TKL) as an overall measure of diagnostic test accuracy and an optimal cut-point selection criterion for two-stage diseases when the purpose of the using a biomarker is to predict for rule-in/out potentials into disease and non-disease stages.

Diseases are commonly classified into two stages, diseased or non-diseased. However, many diseases progress in more than two stages in nature, such as Alzheimer’s disease. Alzheimer’s disease has stages including preclinical stage, early stage (mild), middle stage (moderate), and late stage (severe) (Alzheimer’s Association, 2019; Johns Hopkins Medicine, 2019). For this type of disease, a measure which can discriminate among more than two stages is desired. Several measures from the binary setting have been extended to the multi-stage (i.e., $k > 2$) setting, such as the hypervolume under the manifold (HUM) which is naturally extended from AUC (Scurfield, 1996, 1998) and the generalized Youden index (GYI) in multi-stage setting as the extension of the Youden index in binary diseases (Nakas, Alonzo, & Yiannoutsos, 2010; Nakas, Dalrymple-Alford, Anderson, & Alonzo, 2013). Moreover, a new measure of multi-stage diseases, called the maximum absolute determinant (MADET), was proposed by Dong, Attwood, Hutson, Liu, and Tian (2017). The GYI and the MADET are intended to be used as criteria for optimal cut-point selection in diagnostic tests (Dong et al., 2017; Nakas et al., 2010; Nakas et al., 2013). All of these studies have discussed these methods for three-stage diseases, and further extended them to higher dimensions (Dong et al., 2017; Nakas et al., 2010; Nakas et al., 2013). However, the ROC, the AUC, and the HUM cannot be directly used for selecting optimal cut-point. Thus, other studies have developed criteria for optimal cut-point selection based on those measures, such as the northwest corner measure using the ROC curve for binary disease (Fawcett, 2006; Letón & Molanes, 2009; Perkins & Schisterman, 2006). The application of this approach has become the favored method for optimal cut-point selection. Additionally,   proposed two methods for optimal cut-point selection in the three-stage setting: 1) the closest-to-perfection (CP) is a measure that generalized from the NC; and 2) the maximum volume (MV) is a measure that built on the concept of the VUS (i.e., the special case of HUM in three-stage setting) for three-stage diseases (Attwood et al., 2014).

The research in optimal cut-point selection in multi-stage diseases is far limited compared to two-stage diseases, and much fewer studies have discussed the cut-point selection in diseases with more than three stages. Although Dong et al. (2017) proposed the MADET for evaluating the diagnostic accuracy and cut-point selection in multi-stage setting, this method only works on some cases and has limited clinical interpretation, such as the rule in/out potentials. Therefore, there is a need to demonstrate the stages' rule-in/out potentials for multi-stage diseases. The KL divergence emphasizes on rule-in/out potentials, which can be extended to the multi-stage setting. This method provides the clinical interpretation of the diagnostic tests, which tells how likely a subject will be diagnosed in different stages. In this dissertation, the KL divergence is generalized to the multi-stage setting and proposed as an optimal cut-point selection criterion when the purpose of the using certain biomarker to predict for rule-in/out subjects into disease and non-disease stages.

We generalize the TKL divergence, named the generalized total Kullback-Leibler (GTKL) divergence, as a comprehensive measure of accuracy as well as a criterion of optimal cut-point selection for multi-stage diseases. The measure sums up the rule-in/out information in all stages, and it comprehensively evaluates the correct classification rates in all stages of a multi-stage disease. Overall, the GTKL divergence combines the correct classification rates and misclassification rates based on the KL for diseases with more than two stages, and simultaneously emphasizes the rule-in/out potentials for diagnosis in all stages.

CHAPTER 2
LITERATURE REVIEW

Emerging studies of diagnostic accuracy for multi-stage diseases show a high demand in developing a more reliable diagnostic procedure to discriminate subjects in different diseased stages accurately (Attwood et al., 2014; Li & Fine, 2008; Nakas et al., 2010; Nakas et al., 2013; Xiong, van Belle, Miller, & Morris, 2006). For example, some chronic diseases, such as Alzheimer's disease, kidney disease, and cancers, have more than two stages in nature and require measures that can identify subjects among stages (Alzheimer's Association, 2019; Johns Hopkins Medicine, 2019). The traditional measures for binary classification cannot directly be used for multi-stage classification; however, some popular measures can be extended and are generalized to the multi-stage setting.

The ROC and Youden index are prevalent and essential measures for binary classification, and they describe different aspects of a biomarker. Both measures are built based on the basic measures in diagnostic accuracy, sensitivity, specificity, FPR, and FNR. The four basic measures are not affected by the prevalence of the disease of interest; however, they capture the intrinsic diagnostic accuracy of a diagnostic test (Zhou et al., 2009). In other words, these measures are influenced by the disease's spectrum, which is the range of clinical severity, or anatomic extent constitutes a disease. Also, these measures from a sample population are generalizable to other populations with different prevalence rates. Moreover, sensitivity and specificity are the correct classification rates of diseased and non-diseased populations in the true categorization of their real states, respectively. The sensitivity is the probability that the diseased subjects are correctly identified as diseased in a diagnostic test; and, the specificity is the probability that the non-diseased subjects are correctly identified as non-diseased in the diagnostic test (Zhou et al., 2009). The FPR and FNR, respectively, are the misclassification rates in non-diseased and diseased populations and are produced when subjects are not correctly classified in the corresponding groups (Zhou et al., 2009). In terms of accuracy of the test, type I error (α) rate, which is the probability of rejecting the null hypothesis when the null hypothesis is true in reality, is analogous to FPR (i.e., 1-specificity) (Zhou et al., 2009). Type II error (β) rate, which is the probability of failing to reject the null hypothesis when the alternative hypothesis is true in reality, is analogous to FNR (i.e., 1-sensitivity). Additionally, the statistical power, 1 - type II error rate, is similar to sensitivity. Instead of using the standard type I error rate at 0.05 (5%), the particular clinical application dictates the allowable error rates in diagnostic tests.

The ROC was first introduced in the analysis of radar signals before it was employed in signal detection theory during World War II; afterwards, new research to increase the prediction of correctly detected Japanese aircraft from their radar signals after the attack on Pearl Harbor in 1941 (Egan, 1975; Green & Swets, 1966). Later, the ROC was applied to radiological, psychophysical, and epidemiological studies (H. Aoki, Watanabe, Furuichi, & Tsuda, 1997; Hsiao et al., 1989; Metz, 1989). The potential of the ROC in medical diagnostic testing was recognized as early as the 1960s (Lusted, 1960). The application and evaluation of diagnostic accuracy using the ROC have been systematically reviewed and illustrated in previous studies (Pepe, 2003; Swets & Pickett, 1982; Zweig & Campbell, 1993). The ROC is constructed by plotting the false positive rate (i.e., $1 - specificity$) against sensitivity. The graph geometrically summarizes the entire set of possible true and false positive rates with different thresholds, and it serves as a device to describe the range of trade-off between true positive and true negative rates that can be achieved by a diagnostic test.

Sometimes it is not feasible to construct the ROC, and a summary index becomes a critical measure to summarize the information from the ROC. The AUC is the most widely used summary statistic of the ROC and is computed by taking the integral of the ROC statistics in the range of 0 to 1 (Fawcett, 2006). The value of the AUC ranges from 0.5 to 1, indicating an uninformative diagnostic test (AUC = 0.5) and a perfect diagnostic test (AUC = 1). The AUC is a global measure of diagnostic accuracy, and it does not provide information about sensitivity and specificity as a summary index. For instance, in some situations, two diagnostic tests can have identical AUC with different sensitivity and specificity for the two different tests. That is, the first test can have higher sensitivity compared to the second one; however, the second test has higher specificity. Moreover, the ROC curve and the AUC have no information about predictive values, nor rule-in/out information of a test in medical diagnostics.

The Youden index is another prevalent measure for binary classification in diagnostics, and it is also a global measure and was first proposed by Youden in 1950 (Youden, 1950). The Youden index (J) is a statistic that maximizes the correct classification rates (i.e., sensitivity and specificity) and achieves the maximum discrimination between two stages. The Youden index also encounters the same issues as the ROC and the AUC as two diagnostics with same Youden index value having different sensitivity and specificity, and it does not characterize the rule-in/out information in diagnosis.

The Youden index can be directly used as a criterion to select optimal cut-point (c) for a biomarker; however, neither the ROC nor the AUC can be directly applied to select the optimal cut-point. Nonetheless, methods have been developed based on the ROC and the AUC for optimal cut-point selection. The most popular criterion that was developed according to the ROC for selecting optimal cut-point is the northwest corner (NC), also named the closest-to-perfection, in binary setting (Fawcett, 2006; Letón & Molanes, 2009; Perkins & Schisterman, 2006). The NC method also incorporates the correct classification rates as in the Youden index, and it measures the distances from (0,1) to the point $(1-p_{2,2}(c), p_{1,1}(c))$ on the ROC, where $p_{1,1}(c)$ and $p_{2,2}(c)$ are the correct classification rates of stage 1 (non-disease) and 2 (disease), respectively. Compare to the Youden index, when minimizing the statistics, the NC method has an additional term which is the average of the squared correct classification rates that the Youden index does not include (Perkins & Schisterman, 2006). Although there is no justification for this term in practice, the results from the NC generally have higher specificity than that from the Youden index and produces lower FPRs (Fawcett, 2006; Letón & Molanes, 2009; Perkins & Schisterman, 2006).

In clinical practice, after estimating an optimal cut-point, we need to understand the implication of the results, such as how likely the test would give the correct diagnosis. The measures that can answer this question are the predictive values (i.e., the PPV and the NPV) and the LRs, which approach the data from an aspect different from sensitivity and specificity (Altman & Bland, 1994). The PPV was defined as the proportion of subjects with positive test results which were correctly diagnosed (i.e., true positive results) (Fletcher, Fletcher, & Fletcher, 2012). Similarly, the NPV is the proportion of the cases giving negative test results which are truly non-diseased (i.e., true negative results) (Fletcher et al., 2012). Although the PPV and the NPV are commonly used in clinical decision making, they depend on the prevalence of the disease as they differ in different populations of the same diagnostic test (Altman & Bland, 1994). When the sensitivity and specificity are fixed, the PPV increases as the prevalence of the disease increases, whereas the NPV decreases (Wong & Lim, 2011). Therefore, the PPV and the NPV of a population cannot be generalized to a different population.

Compare to the PPV and the NPV, the LRs do not depend on the prevalence of the disease, and the LRs of the same diagnostic test can be generalized to different populations.

Additionally, the LRs provide information about rule-in/out of a diagnostic test (Boyko, 1994; Deeks & Altman, 2004). The rule-in/out tests are important for different medical purposes. For example, the rule-in principle (specificity) is useful when a toxic treatment of the disease will be initiated if the diagnosis is confirmed, such as chemotherapy (Lee, 1999). The rule-out principle (sensitivity) is helpful when there is a significant penalty for missing the disease, and the initial treatment is relatively safe, like screening tests for tuberculosis (Lee, 1999). The LRs can be calculated for either positive or negative test results particularly. A positive LR tells how likely a diseased subject will receive a positive test result compared to a non-diseased subject; whereas a negative LR shows how likely a non-diseased subject will receive a negative test result compared to a diseased subject (Šimundić, 2009).

A biomarker that can discriminate subjects from diseased and non-diseased populations is efficient for diagnosing a disease. However, some diseases have several distinct ordinal stages which cannot be recognized by existing measures in diagnostics. Dichotomize biomarker to binary stages generally combines diseased stages which results in the delay of diagnosing patients in the early disease stage. Missing to diagnose patients in the early stage of the disease will delay the appropriate treatments and cause serious health problems in the future. Therefore, being able to diagnose a patient in the early disease stage will allow physicians to provide early interventions and decrease the progression of the disease. The medical community has demonstrated high interest in the ability to discriminate diseased population into different stages to provide better treatment strategies, such as the identification of mild cognitive impairment of Parkinson disease and the early diagnosis of Alzheimer’s disease (Aarsland & Kurz, 2010; DAFFNEr & Scinto, 2000). Thus, having the appropriate methods to discriminate among different stages is imperative for early clinical interventions, such as early interventions for breast cancer (Early Breast Cancer Trialists' Collaborative Group, 2005; Richards, Westcombe, Love, Littlejohns, & Ramirez, 1999). Moreover, some frontier studies proposed measures that generalized from binary classification to multi-stage classification using the ROC, AUC, and Youden index (Nakas et al., 2010; Nakas et al., 2013; Scurfield, 1996, 1998; Xiong et al., 2006). Furthermore, Dong et al. (2017) introduced the maximum absolute determinant (MADET) as a new measure for diagnostic accuracy suggesting better disease diagnostics procedure, in some cases, compared to the other existing measures.

Although limited studies have been done to investigate the accuracy measures in multi-stage setting, some researchers have provided breakthrough evidence for the need to expand on multi-stage disease diagnosis. For instances, Scurfield (1996, 1998) proposed the concept of hypervolume under the manifold (HUM) by extending the AUC. Similarly, Nakas and Yiannoutsos (2004) introduced the non-parametric measure of the HUM, and Li and Fine (2008) proposed the inference procedures and methods that correspondent to the measure of the HUM for estimation of classification probabilities and calculating the HUM. Additionally, Nakas et al. (2010) extended the Youden index for binary classification to the GYI for multi-stage classification. Moreover, Dong et al. (2017) suggested the maximum absolute determinant (MADET) for general multi-stage classification, which embraces the correct classification rates and the misclassification rates. In the special case of multi-stage diseases, all the measures that mentioned above were discussed in the case of three-stage setting, such as the volume under the ROC surface (VUS), the GYI and the MADET (Dong et al., 2017; Nakas et al., 2010; Xiong et al., 2006).

Furthermore, the GYI can be used as a criterion to select optimal cut-points for multi-stage diseases. Also, the MADET proposed by Dong et al. (2017) is capable of providing information on optimal cut-points selection. However, similar to the ROC and the AUC, the HUM cannot be directly used in the selection of cut-points. Nevertheless, Attwood et al. (2014) proposed two methods to select optimal cut-points for three-stage diseases. The first method, known as the closest-to-perfection (CP), is a criterion of selecting cut-points in the multi-stage setting similar to the concept of the ROC (Attwood et al., 2014). To avoid confusion, we use the NC for binary diseases and the CP for three-stage diseases in this study. The second method, called the maximum volume (MV) method, is also a criterion of selecting cut-points in the multi-stage setting; however, the MV measures the volume under the surface curve that covered by all the possible cut-points which has the similar concept of AUC (Attwood et al., 2014). Attwood et al. (2014) compared the correlation of the cut-points that were selected by different methods, including the GYI, the CP, and the MV. In addition, they proposed a statistic, the loss of the total correct classification (LTCC), to compare the correct classification information that was measured by different selection criteria (Attwood et al., 2014). Dong et al. (2017) compared the MADET with the GYI, the CP, and the MV in their study. They also generalized the criteria to multi-stage diseases and evaluated the performance of optimal cut-points selection (Dong et al., 2017). These studies suggested that the new methods were comparable to the GYI and achieved better-balanced rates when minimizing the LTCC (Attwood et al., 2014; Dong et al., 2017). The GYI required a larger sample size within different diseased groups to accurately estimate the cut-points as compared to other methods (Attwood et al., 2014). This measure only incorporates the correct classification rates and consequently loses some unignorable information in the classification process. Dong et al. (2017) compared four methods by simulation using power analysis and the LTCC to evaluate their performance for optimal cut-point selection. Their simulation studies showed that the proposed measures performed comparatively well in different distributions (Dong et al., 2017). However, the results do not provide clinical interpretation of rule-in/out potentials of diagnostic biomarkers in the multi-stage setting. Therefore, further investigation is needed for rule-in/out potentials in clinical studies.

Recently, the KL divergence from information theory has been applied to medical diagnostics, and it nicely characterized rule-in and rule-out potentials of a diagnostic test (Lee, 1999). The KL divergence estimates the separation between two probability distributions, which are the probability distributions of diseased and non-diseased populations. Lee (1999) illustrated the application of the KL divergence on measuring the diagnostic performance of biomarkers with a given optimal cut-point for binary diseases. Hughes and Bhattacharya (2013) constructed information graphs, based on Lee's application, which provided a diagrammatic interpretation of the KL divergence. The information graph demonstrates a visual basis for the evaluation and comparison of binary diagnostic tests and makes the application of the KL divergence more appealing to clinicians (Hughes & Bhattacharya, 2013). Additionally, Samawi et al. (2019) investigated the applications of the KL divergence in measuring the performance of a diagnostic test of dichotomized continuous biomarkers. Furthermore, they suggested the total KL (TKL) as a comprehensive measure of rule-in/out information, as well as a criterion for optimal cut-point selection in the binary setting (Samawi et al., 2019). While the number of stages of disease increases, the information from misclassification rates becomes more massive, and disregarding the information will lead to loss of information in diagnostic accuracy. In the binary setting, the KL divergence incorporates correct classification rates and misclassification rates, thus covering more information than the Youden index. In this dissertation, the TKL divergence in the binary setting is generalized to the multi-stage setting by summing up the rule-in/out information in all stages. Likewise, the GTKL was used for optimal cut-point selection for multi-stage diseases.

Simultaneously, this measure integrated the before-test rule-in/out potentials for diagnosis in different stages. Lastly, the predictive values were generalized to assess the performance of the GTKL and other existing measures for multi-stage diseases based on the methods suggested by (Samawi, 2019).

CHAPTER 3
METHODS

This chapter provides an overview of some related measures that have been used in medical diagnostics, including sensitivity (Se), specificity (Sp), FPR, FNR and all related measures of diagnostics test accuracy and criteria of optimal cut-points selection.

3.1 Binary (two-stage) diseases

A diagnostic cut point, *c*, is generally required to classify a subject either as a diseased or non-diseased for clinical decision making with diagnostic biomarkers. Let X_1 and X_2 denote the marker values for non-diseased and diseased subjects, with c.d.fs $F_1(.)$ and $F_2(.)$ respectively. The probabilities classification matrix $\mathbf{P}$ in the binary setting, given $F_1(.)$ and $F_2(.)$ with threshold *c*, can be expressed as

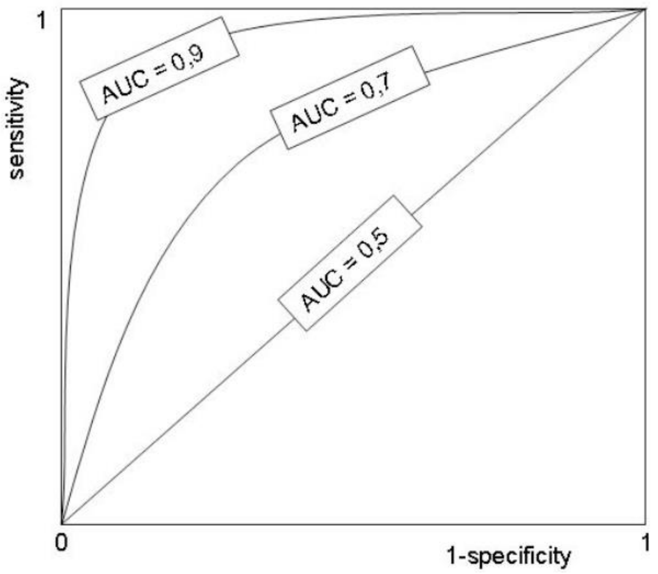
$$\mathbf{P} = \begin{array}{cc} T^{-} & T^{+} \\ \end{array} \\ \begin{pmatrix} F_1(c) & 1-F_1(c) \\ F_2(c) & 1-F_2(c) \end{pmatrix} \begin{array}{l} S=1 \\ S=2. \end{array} \tag{3. 1}$$

where T^{-} and T^{+} respectively are the negative and positive test results of a test, and $S = i;\ i = 1,\ 2$ is the disease stages, imply non-disease and disease, respectively.

Without the loss of generality, in most circumstances, higher marker values indicate greater severity of the disease. This assumption of directionality is important for the ROC analysis to guarantee valid values of ROC indices. The ROC curve is a graph of true positive rate or sensitivity ($Se(c) = P(X_2 > \text{c}) = 1 - F_2(c)$) versus false positive rate or 1- specificity ($FPR = 1 - Sp(c) = 1 - F_1(c)$), where $Sp(c) = P(X_1 \le \text{c}) = F_1(c)$, over all possible thresholds of the marker. On the other hand, the false negative rate is given by $FNR = 1 - Se(c) = F_2(c)$.

The following graph shows an example of visual interpretation of the ROC.

Figure 3.1. ROC curve (source: Šimundić, 2009).



In practice, it is common to summarize the information of the ROC curve into a single global value or index, such as the AUC. The AUC has a range of [0.5, 1] and can be measured by the following equation:

$$AUC = P(X_2 > X_1) = \int_{-\infty}^{\infty} f_1(x_1)[1 - F_2(x_1)]dx_1 \ . \qquad (3.\ 2)$$

The AUC evaluates the discriminatory ability of a marker, where $ROC(q) = 1 - F_2[F_1^{-1}(1-q)]$ and $q = 1 - Sp(c)$.

The Youden index (J) is another measure that summarizes the sensitivity and specificity that is frequently used, and it is a criterion of selecting an optimal diagnostic cut-point, (c), (K. Aoki, Misumi, Kimura, Zhao, & Xie, 1997). The Youden index has a range of [0, 1], and it is defined as

$$J = \underset{c}{Max}(Se(c) + Sp(c) - 1) = \underset{c}{Max}(F_1(c) - F_2(c)) \ , \qquad (3.\ 3)$$

where

$$c = \arg \underset{c}{Max}(Se(c) + Sp(c) - 1) = \arg \underset{c}{Max}(F_1(c) - F_2(c)) \ . \qquad (3.\ 4)$$

The OR that measures the diagnostic test accuracy at the threshold c can be calculated as

$$OR = \frac{Se(c) \times Sp(c)}{[1 - Se(c)] \times [1 - Sp(c)]} = \frac{F_1(c) \times [1 - F_2(c)]}{F_2(c) \times [1 - F_1(c)]}, \qquad (3.5)$$

The range of OR is from 0 to infinity with higher values indicating higher discriminative power in diagnostic tests. A test is improper when the value of OR is less than 1, which means there are more negative tests among the diseased population. A value of 1 indicates that a test has no discriminative power to identify patients with the disease and those without the disease (Glas, Lijmer, Prins, Bonsel, & Bossuyt, 2003).

The PPV and the NPV are the after-test performance measures of diagnostic tests. These measures show the probability that a subject will receive the correct diagnostic test result with its true stage. When the diagnostic tests are binary (i.e., non-diseased $(S = 1)$ or diseased $(S = 2)$ vs. test positive (T^+) or test negative (T^-)), the PPV and the NPV are calculated based on the 2 x 2 classification matrix $\mathbf{P}$ (3.1) as follows:

$$\begin{aligned} PPV &= \frac{Se(c).P(S = 2)}{Se(c).P(S = 2) + P(S = 1).(1 - Sp(c))} \\ &= \frac{[1 - F_2(c)].P(S = 2)}{[1 - F_2(c)].P(S = 2) + P(S = 1).[1 - F_1(c)]Sp(c))} \end{aligned} \qquad (3.6)$$

and

$$\begin{aligned} NPV &= \frac{Sp(c).P(S = 1)}{Sp(c).P(S = 1) + P(S = 2).(1 - Se(c))} \\ &= \frac{F_1(c).P(S = 1)}{F_1(c).P(S = 1) + P(S = 2).F_2(c)} \end{aligned}. \qquad (3.7)$$

Similar to the PPV and the NPV, the LRs are other after-test performance measures. Generally speaking, LRs are defined as a ratio of the probability that a test result is correct to the probability that the test result is incorrect. In particular, sensitivity and specificity of a test are used to calculate the LRs, which is calculated for both positive and negative test results and is expressed as ‘ LR_+ ’ and ‘ LR_- ’, respectively. The calculations are based on the following formulas in the binary setting:

$$LR_+ = \ sensitivity \ / \ 1- \ specificity = \frac{Se(c)}{1-Sp(c)} = \frac{1-F_2(c)}{1-F_1(c)}, \tag{3. 8}$$

and

$$LR_- = \ 1- \ sensitivity \ / \ specificity = \frac{1-Se(c)}{Sp(c)} = \frac{F_2(c)}{F_1(c)}. \tag{3. 9}$$

The LR with a value greater than 1 indicates a test result that is associated with the presence of the disease, whereas a value less than 1 indicates a test result that is associated with the absence of the disease (Deeks & Altman, 2004). The LR with a value that is further from 1 shows stronger evidence for the presence or absence of the disease.

Lee (1999) suggested the KL divergence as the before-test measure of diagnostic performance. For a binary test, the proportions of the non-diseased and diseased populations are denoted by functions g_1 and g_2 respectively, the KL divergence of $D(g_1, g_2)$ and $D(g_2, g_1)$ can be interpreted as before-test potentials of rule-out and rule-in disease, respectively. The equations of $D(g_1, g_2)$ and $D(g_2, g_1)$ are as follows:

using the non-diseased distribution as the reference,

$$\begin{aligned} D(g_2, g_1) &= (1-Se)\cdot \log\frac{1-Se}{Sp} + Se\cdot\log\frac{Se}{1-Sp} \\ &= (1-Se)\cdot\log LR_- + Se\cdot\log LR_+; \end{aligned} \tag{3. 10}$$

and using the diseased distribution as the reference,

$$\begin{aligned} D(g_1, g_2) &= Sp \cdot \log \frac{Sp}{(1-Se)} + (1-Sp) \cdot \log \frac{1-Sp}{Se} \\ &= Sp \cdot \log(1/LR_{-}) + (1-Sp) \cdot \log(1/LR_{+}). \end{aligned} \qquad (3.\ 11)$$

A diagnostic test with a larger $D(g_1, g_2)$, will on an average make diseased subjects more likely to have positive diagnosis results. A subject with a negative diagnosis resulting from a test with a large $D(g_1, g_2)$ value will more likely be ruled out from the disease group in which case, the potential of rule-out disease is higher. Similarly, a diagnostic test with greater $D(g_2, g_1)$, will on average make non-disease subjects more likely to have a negative diagnosis. A subject with a positive diagnosis resulting from a test with larger $D(g_2, g_1)$ value will become more likely to be ruled-in to the disease group in which case, the potential of rule-in disease is higher.

On the other hand, the KL divergence is a measure for continuous biomarkers as well, and it is an indicator of the rule-in/out potentials. Likewise, the KL in the continuous case can be computed using different reference levels as in the discrete case, either diseased or non-diseased populations. In the continuous case, the KL divergence that uses the non-diseased population as the reference is denoted as $D(f_2, f_1)$ and can be computed as

$$D(f_2, f_1) = \int_{-\infty}^{\infty} f_2(x) \ln\left(\frac{f_2(x)}{f_1(x)} \right) dx, \qquad (3.\ 12)$$

similarly, the KL divergence that uses the diseased population as the reference is denoted as $D(f_1, f_2)$ and can be computed as

$$D(f_1, f_2) = \int_{-\infty}^{\infty} f_1(x) \ln\left(\frac{f_1(x)}{f_2(x)} \right) dx, \qquad (3.\ 13)$$

where $f_1(.)$ and $f_2(.)$ are the underlying probability density functions (p.d.f.) of random variables X_1 (non-disease values) and X_2 (disease values). Then the TKL measure suggested by Samawi et al. (2019), which is defined by $TKL = D(f_1, f_2) + D(f_2, f_1)$, as a measure of overall diagnostics test accuracy. Samawi et al. (2019a) showed the relationship between the TKL and some diagnostic accuracy indices for the binary diagnostic test at a given cut-point (c) as follows:

$$\begin{aligned} TKL &= D(f_1, f_2) + D(f_2, f_1) \\ &= [Se(c) + Sp(c) - 1]\ln\left(OR(c)\right) + R(c), \end{aligned} \tag{3. 14}$$

where $[Se(c) + Sp(c) - 1]\ln\left(OR(c)\right)]$ is the discrete portions, denotes by $TKL_{discrete}$, *R(c)* is the remainder, which is the loss of information from dichotomizing the continuous tests. Furthermore, they proposed a criterion of optimal cut-point selection, for binary diseases, as

$$c = \arg \underset{c}{Max}(TKL_{discrete}(c)) \,. \tag{3. 15}$$

To optimize the diagnostic accuracy, the $TKL_{discrete}$ can be maximized with respect to the cut-point value (c) across all possible values of c; simultaneously. Consequently, the reminder $R(c)$ will be minimized. They showed that the discrete portion, the $TKL_{discrete}$, can be considered as a scaled measure of the Youden index (*J*) by the logarithm of the OR at a given threshold c, when c is selected based on the statistic *J*.

3.2 Multi-stage (k > 2) diseases

Some of the measures of binary classification have been extended to multi-stage (k stages) classification as follows: Define a class of probabilities $p_{i,j}$ for $i, j = 1, 2, ..., k$ that classifying a randomly selected subject in j*th* test class given the subject is in the i*th* stage of the disease. To make a diagnosis for a disease with k -stages when we have a continuous biomarker *X*, cut-points $\mathbf{c}' = (c_1, c_2, ..., c_{k-1})$ are needed. If $c_{j-1} < X \le c_j$, $j = 1, 2, ..., k-1,$ this subject is classified into ordinal stage *j*, with larger *j* corresponds to severer condition. If we let X_i denote the marker values for the ith disease stage with p.d.f. and c.d.f. $f_i(x)$ and $F_i(x)$ respectively. Then, the corresponding conditional probability $p_{i,j}$ can be defined as

$$p_{i,j} = P(c_{j-1} < X_i \le c_j \mid S = i) = F_i(c_j) - F_i(c_{j-1}) = P(T = j \mid S = i) \text{ , \quad for } i, j = 1, 2, ..., k \text{ .} \qquad (3.\ 16)$$

$$\mathbf{P} = \begin{array}{c} \begin{array}{cccc} T_{S=i}^{1} & T_{S=i}^{2} & \cdots & T_{S=i}^{k} \end{array} \\ \left[\begin{array}{cccc} p_{1,1} & p_{1,2} & \cdots & p_{1,k} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,k} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ p_{k,1} & p_{k,2} & \cdots & p_{k,k} \end{array} \right] \end{array} \begin{array}{l} \\ \mathrm{S} = 1 \\ S = 2 \\ . \\ . \\ . \\ S = k \end{array} \qquad (3.\ 17)$$

The next subsequent discussion will be for k = 3 for simplicity. The generalization to k > 3 is straight forward and will be discussed at the end of the section.

For the case when k = 3, the $\mathbf{P}$ matrix corresponding with all possible pairs of thresholds c_1 and c_2 is defined as

$$\mathbf{P} = \begin{array}{c} \begin{array}{ccc} T_{S=i}^{1} & T_{S=i}^{2} & T_{S=i}^{3} \end{array} \\ \left[\begin{array}{ccc} p_{1,1} & p_{1,2} & p_{1,3} \\ p_{2,1} & p_{2,2} & p_{2,3} \\ p_{3,1} & p_{3,2} & p_{3,3} \end{array} \right] \end{array} \begin{array}{l} \\ S{=}1 \\ S{=}2 \\ S{=}3 \end{array} = \left[\begin{array}{ccc} F_1(c_1) & F_1(c_2) - F_1(c_1) & 1 - F_1(c_2) \\ F_2(c_1) & F_2(c_2) - F_2(c_1) & 1 - F_2(c_2) \\ F_3(c_1) & F_3(c_2) - F_3(c_1) & 1 - F_3(c_2) \end{array} \right], \qquad (3.\ 18)$$

Nakas et al. (2010) discussed the GYI in the three-stage setting which is denoted as J_3. The J_3 is defined as the maximum of the sum of the correct classification rates over all possible pairs of thresholds c_1 and c_2. Then, the J_3 can be computed by maximizing the sum of correct classification rates in all stages with thresholds c_1 and c_2 as

$$\begin{aligned} J_3 &= \underset{c_1<c_2}{Max}[p_{1,1} + p_{2,2} + p_{3,3} - 1] \\ &= \underset{c_1<c_2}{Max}[F_1(c_1) - F_2(c_1) + F_2(c_2) - F_3(c_2)], \end{aligned} \qquad (3.\ 19)$$

The GYI for three-stage diseases, J_3, is the unification of two-stage and three-stage analysis approaches Nakas et al. (2013). That is, J_3 is the sum of the statistics J_2 of the pairwise comparison between adjacent stages (i.e., stage 1 vs. 2 and stage 2 vs. 3). The unification can be mathematically presented as follows:

$$\begin{aligned} J_3 &= \underset{c_1<c_2}{Max}[F_1(c_1) - F_2(c_1) + F_2(c_2) - F_3(c_2)] \\ &= \underset{c_1<c_2}{Max}[F_1(c_1) - F_2(c_1)] + \underset{c_1<c_2}{Max}[F_2(c_2) - F_3(c_2)] \\ &= J_{2;(1,2)} + J_{2;(2,3)}, \end{aligned} \tag{3. 20}$$

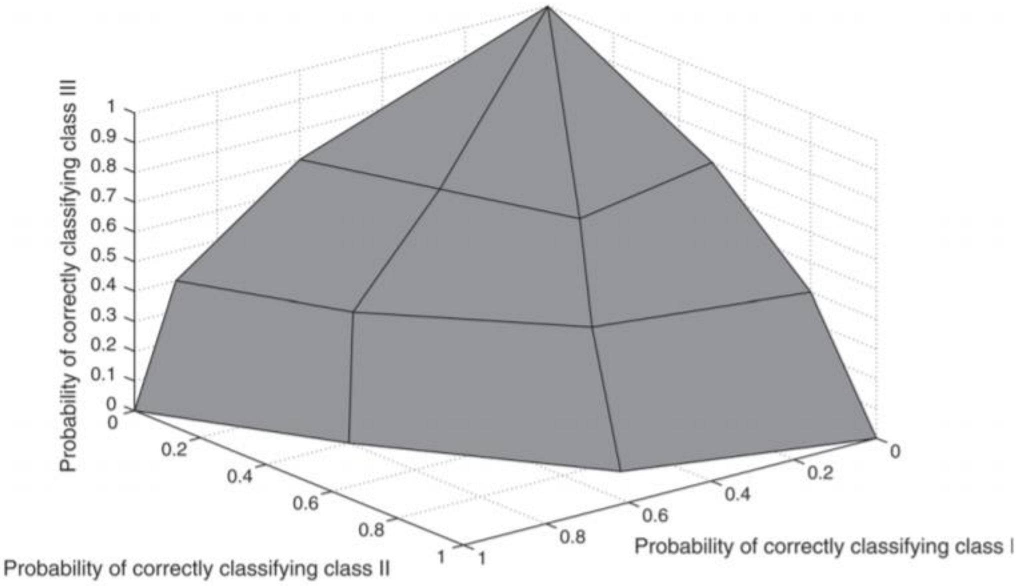
where $J_{2;(1,2)}$ and $J_{2;(2,3)}$ are the Youden indices corresponding to stage 1 vs. 2 and stage 2 vs. 3, respectively.

A three-dimensional ROC surface can be plotted by using all possible pairs of thresholds c_1 and c_2. Similar to the ROC in the binary setting, the ROC of the adjacent pairs of the three stages summarizes all possible thresholds between two stages (stage 1 vs. 2 and stage 2 vs. 3). The two-dimensional plots are comprehensively interpreted in one three-dimensional graph. Figure 2 shows an example of a hypothetical three-dimensional graph of a three-stage disease. A statistic calculated from the three-dimensional ROC surface, called the volume under the ROC surface (VUS), is another measure of three-stage classification, which is a special case of the hypervolume under manifold (HUM) (Scurfield, 1996, 1998; Xiong et al., 2006). The VUS is mathematically defined as

$$VUS = \int_0^1 \int_0^{f_1(t_1)} f_2(t_1, t_2) dt_2 dt_1, \tag{3. 21}$$

where t_i with $i = 1, 2, 3$ is the correct classification rate for the i^{th} stage (i.e., the $p_{i,i}$ as defined by the $\mathbf{P}$ matrix) and f_{i-1}'s are the recursive equations when $t_i = f_{i-1}(t_1, \ldots, t_{i-1}), i = 2, 3$.

Figure 3.2. A hypothetical example of 3-class ROC analysis (Source: Li & Fine, 2008).



The MADET is a measure that uses the determinant of the $\mathbf{P}$ matrix for disease classification with more than two stages (Dong et al., 2017). The measure incorporates both correct and false classification rates in different stages of the disease. For three-stage diseases, the MADET statistic is computed from the determinant of the $\mathbf{P}$ matrix shown in (3.18), by the following equation

$$MADET = \underset{c_1,c_2}{Max}\left|det(\mathbf{P})\right|, \tag{3. 22}$$

where $det(\mathbf{P})$ is the determinant of the $\mathbf{P}$ matrix which can be calculated using either formula as follows:

1). $\underset{c_1,c_2}{Max}|\, p_{1,1}p_{2,2}p_{3,3} + p_{1,2}p_{2,3}p_{3,1} + p_{1,3}p_{3,2}p_{2,1} - p_{1,1}p_{2,3}p_{3,2} - p_{1,2}p_{2,1}p_{3,3} - p_{1,3}p_{3,1}p_{2,2}\,|$, (3. 23)

2). $\underset{c_1,c_2}{Max}|1-[p_{1,1}(1-p_{2,2})+p_{2,2}(1-p_{3,3})+p_{3,3}(1-p_{1,1})]-(p_{1,2}p_{2,1}+p_{1,3}p_{3,1}+p_{2,3}p_{3,2})|$. (3. 24)

The MADET ranges from 0 to 1, with 0 indicates no discriminative power, and 1 indicates the perfect discrimination of a diagnostic test. A larger MADET indicates a better diagnostic test.

There are three possible conditions that the MADET is 0: 1) all three probability vectors in the three stages coincide; 2) any two of the probability vectors coincide; and, 3) all three probability vectors fall on the same two-dimensional plane (Dong et al., 2017). The graphical interpretation of the MADET is similar to VUS; however, they may have different possible thresholds c_1 and c_2. Note that in the binary setting, the MADET is the same as the Youden index.

Finally, for diseases with more stages (i.e., $k > 3$), those measures discussed above can be generalized as follows:

GYI:

$$\begin{aligned} J(c_1, c_2, \ldots, c_{k-1}) &= \underset{c_1<c_2<\ldots<c_{k-1}}{Max} (p_{1,1} + p_{2,2} + \cdots + p_{k,k} - 1) \\ &= J_2(c_1, c_2) + J_2(c_2, c_3) + \cdots + J_2(c_{k-2}, c_{k-1}); \end{aligned} \tag{3. 25}$$

HUM:

$$HUM = \int_0^1 \int_0^{f_1(t_1)} \cdots \int_0^{f_{k-2}(t_1,\ldots,t_{k-2})} f_{k-1}(t_1, \ldots, t_{k-1}) dt_{k-1} \cdots dt_2 t_1, \tag{3. 26}$$

where t_i with $i = 1, 2, \ldots, k$ and $t_i = f_{i-1}(t_1, \ldots, t_{i-1})$ with $i = 2, 3, \ldots, k$;

MADET:

$$MADET(c_1, c_2, \ldots, c_{k-1}) = \underset{c_1<c_2<\ldots<c_{k-1}}{Max} \left| det(\mathbf{P}) \right|. \tag{3. 27}$$

In the discrete case, the KL divergence that proposed by Lee (1999) can be generalized to multi-stage discrete biomarkers. The equations of rule-in/out potentials are defined as follows: using the non-diseased distribution as the reference,

$$D(g_2, g_1) = \sum_{i=1}^{k} g_{2,i} \cdot \log \frac{g_{2,i}}{g_{1,i}} = \sum_{i=1}^{k} g_{2,i} \cdot \log LR_i\,; \tag{3. 28}$$

and using the diseased distribution as the reference,

$$D(g_1, g_2) = \sum_{i=1}^{k} g_{1,i} \cdot \log \frac{g_{1,i}}{g_{2,i}} = \sum_{i=1}^{k} g_{1,i} \cdot \log(1/LR_i), \tag{3. 29}$$

where $\{g_{1,i}, i = 1, 2, ..., k\}$ and $\{g_{2,i}, i = 1, 2, ..., k\}$ are the proportions of non-diseased and diseased subjects in the i^{th} testing category, respectively.

3.3 Optimal Cut-point Selection

3.3.1 Criteria for optimal cut-point selection in binary (k=2) setting

In the binary setting, the popular criteria for selecting the optimal cut-point are the Youden index and the NC method. As mentioned above, the optimal cut-point that selected based on the Youden index is

$$c = \arg\underset{c}{Max}(Se(c) + Sp(c) - 1) = \arg\underset{c}{Max}(F_1(c) - F_2(c)), \tag{3. 30}$$

which gives the largest value of the statistics *J*. Based on the ROC, the NC selects the optimal threshold based on the point on the ROC that closest to (0,1). The 'optimal' diagnostic cut-point (*c*) is the ideal point closest to perfection where $p_{2,2}(c) = 1$ and $p_{1,1}(c) = 1$. The cut-point selected minimizes the distance (D) from (0,1) to $(1 - p_{2,2}(c), p_{1,1}(c))$ (Perkins & Schisterman, 2006). Compared to those thresholds on the ROC curve that are further from (0,1), the optimal threshold is considered to be more accurate (Fawcett, 2006). The distance (D) is defined as

$$D = \underset{c}{Min}\left[\sqrt{(1 - p_{2,2}(c))^2 + (1 - p_{1,1}(c))^2}\right], \tag{3. 31}$$

where

$$\begin{aligned} c &= \arg\underset{c}{Min}\, D \\ &= \arg\underset{c}{Min}\left[\sqrt{(1 - p_{2,2}(c))^2 + (1 - p_{1,1}(c))^2}\right]. \end{aligned} \tag{3. 32}$$

The range of D is from 0 to $\sqrt{0.5}$, indicating perfection and complete lack of discrimination, respectively. The D can also be expressed in the following form,

$$D = [(1 - p_{2,2}(c)) + (1 - p_{1,1}(c)) + 0.5(p_{2,2}(c)^2 + p_{1,1}(c)^2)], \tag{3. 33}$$

which minimizes the total misclassification rates and a third term (i.e., the average of the squared correct classification rates) (Perkins & Schisterman, 2006), and the cut-point can be expressed as

$$\begin{aligned} c &= \arg \underset{c}{Min}\, D \\ &= \arg \underset{c}{Min}[(1 - p_{2,2}(c)) + (1 - p_{1,1}(c)) + 0.5(p_{2,2}(c)^2 + p_{1,1}(c)^2)]. \end{aligned} \tag{3. 34}$$

3.3.2 Criteria for optimal cut-point selection in three-stage (k=3) setting

For three-stage diseases, the optimal cut-points are a pair of a cut-point (c_1) between non-diseased and early-diseased, and a cut-point (c_2) between early-diseased and fully-diseased. In the continuous case, the disease stages are ordinal, and the cut-points for later stages are larger than cut-points for early stages thus $c_1 < c_2$. In the following equations, $p_{1,1}(c_1)$, $p_{2,2}(c_1, c_2)$, and $p_{3,3}(c_2)$ are the correct classification rates in stage 1, 2, and 3, respectively, at the optimal cut-points (c_1, c_2) selected based on each diagnostic accuracy criterion.

The GYI used as a criterion for selecting optimal cut-points for three-stage diseases (Naka et al., 2010). The optimal cut-points are produced numerically by using constrained maximization based on the J_3 statistic (Nakas et al., 2010). Thus, the 'optimal' cut-points are selected by maximizing the J_3 statistic, where

$$(c_1, c_2) = \arg \underset{c_1 < c_2}{Max}[F_1(c_1) - F_2(c_1) + F_2(c_2) - F_3(c_2)]. \tag{3. 35}$$

Alternatively, the criterion is equivalent to minimizing the total misclassification rate in the following equation

$$\underset{c_1 < c_2}{Min}[([1 - p_{3,3}(c_1)] + (1 - p_{2,2}(c_1, c_2)) + 1 - p_{1,1}(c_2)], \tag{3. 36}$$

where

$$(c_1,c_2)=\arg\underset{c_1<c_2}{Min}[(1-p_{3,3}(c_1))+(1-p_{2,2}(c_1,c_2))+(1-p_{1,1}(c_2))]\,. \tag{3. 37}$$

For the CP in the three-stage setting, the perfect discrimination is at point (1,1,1). The distance (D_3) is minimized at the optimal cut-points and can be found numerically using constrained minimization with the following definition:

$$D_3=\underset{c_1<c_2}{Min}\left[\sqrt{(1-p_{3,3}(c_1))^2+(1-p_{2,2}(c_1,c_2))^2+(1-p_{1,1}(c_2))^2}\,\right], \tag{3. 38}$$

where

$$\begin{aligned}(c_1,c_2)&=\arg\underset{c_1<c_2}{Min}D_3\\ &=\arg\underset{c_1<c_2}{Min}\left[\sqrt{(1-p_{3,3}(c_1))^2+(1-p_{2,2}(c_1,c_2))^2+(1-p_{1,1}(c_2))^2}\,\right].\end{aligned} \tag{3. 39}$$

Same as the binary setting, a distance of 0 indicates perfect discrimination while increasing the distance means weaker power in discrimination among the three stages. This method also minimizes the misclassification rates and a third term in the three-stage setting as follows

$$\begin{aligned}&\underset{c_1<c_2}{Min}[(1-p_{3,3}(c_1))+(1-p_{2,2}(c_1,c_2))+(1-p_{1,1}(c_2))\\ &\quad+0.5(p_{3,3}(c_1)^2+p_{2,2}(c_1,c_2)^2+p_{1,1}(c_2)^2)],\end{aligned} \tag{3. 40}$$

Alternatively, the optimal cut-points can be expressed using (3.37) as

$$\begin{aligned}(c_1,c_2)=\arg&\underset{c_1<c_2}{Min}[(1-p_{3,3}(c_1))+(1-p_{2,2}(c_1,c_2))+(1-p_{1,1}(c_2))\\ &+0.5(p_{3,3}(c_1)^2+p_{2,2}(c_1,c_2)^2+p_{1,1}(c_2)^2)];\end{aligned} \tag{3. 41}$$

The MV is a criterion of selecting cutpoints for multi-stage diseases built on the concept of the VUS (Attwood et al., 2014). The VUS is defined as

$$V_3=\underset{c_1<c_2}{Max}[p_{3,3}(c_1)\times p_{2,2}(c_1,c_2)\times p_{1,1}(c_2)], \tag{3. 42}$$

Additionally, the maximization is equivalent to minimization of the following equation:

$$\underset{c_1<c_2}{Min}[-\log(p_{3,3}(c_1))-\log(p_{2,2}(c_1,c_2))-\log(p_{1,1}(c_2))], \tag{3. 43}$$

The ‘optimal’ cut-points can be obtained using equations (3.19) and (3.20) as follows:

$$\begin{aligned}(c_1,c_2) &= \arg\underset{c_1<c_2}{Max}[p_{3,3}(c_1)\times p_{2,2}(c_1,c_2)\times p_{1,1}(c_2)] \\ &= \arg\underset{c_1<c_2}{Min}[-\log(p_{3,3}(c_1))-\log(p_{2,2}(c_1,c_2))-\log(p_{1,1}(c_2))].\end{aligned} \tag{3. 44}$$

The statistic V_3 has a range of $\frac{1}{27}$ to 1, indicating the least discrimination and the perfect discrimination, respectively.

The MADET also is a criterion for optimal cut-points selection, and the cut-points selected according to its statistics is defined as

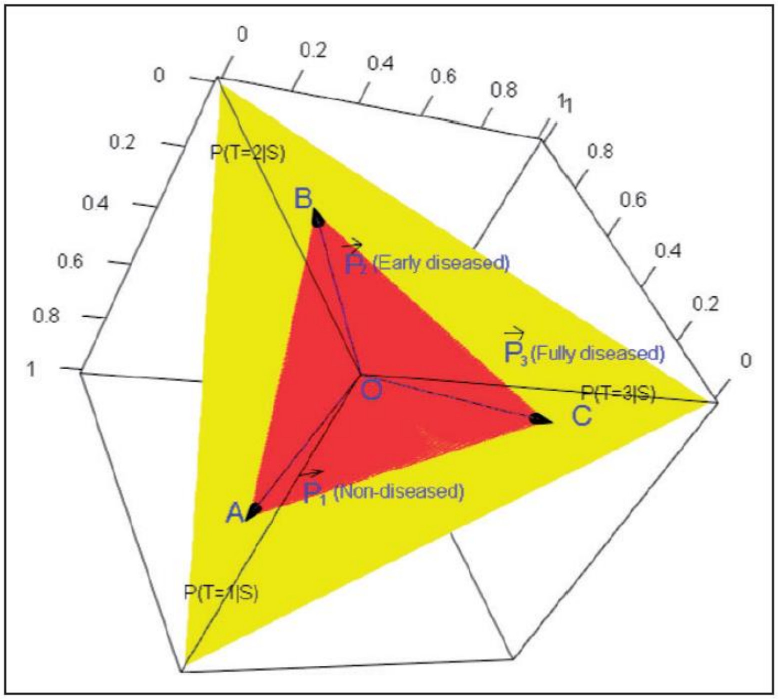
$$(c_1,c_2) = \arg\underset{c_1,c_2}{Max}\left|det(\mathbf{P})\right|, \tag{3. 45}$$

using equation (3.22) shown above. The $\mathbf{P}$ matrix can be expressed by three vectors denoted as $\overrightarrow{\mathbf{P}_1}$, $\overrightarrow{\mathbf{P}_2}$ and $\overrightarrow{\mathbf{P}_3}$, defined as follows:

$$\overrightarrow{\mathbf{P}_i} = (P_{i,1} \quad P_{i,2} \quad P_{i,3}), \tag{3. 46}$$

where $i=1,2,3$ is the disease stage, and we know that $P_{i,1}+P_{i,2}+P_{i,3}=1$ (Dong et al., 2017). The conditional probability, $P(T=j\mid S=i)$, where $i, j=1,2,3$ respectively are the disease stage and test result, is the probability that a subject with the true i^{th} disease condition receives a j test result. The values in the $\mathbf{P}$ matrix depend on the cut-points that are selected by the measure. The larger MADET value indicates the better diagnostic ability of the diagnostic biomarker with the larger difference among the classification rates vectors $\overrightarrow{\mathbf{P}_1}$, $\overrightarrow{\mathbf{P}_2}$ and $\overrightarrow{\mathbf{P}_3}$ (Dong et al., 2017). The ‘optimal’ cut-points would give the largest MADET value. The MADET can be geometrically interpreted by a three-dimensional graph, as shown in Figure 3. In Figure 3, the yellow part is the area that a volume formed by $\overrightarrow{\mathbf{P}_1}$, $\overrightarrow{\mathbf{P}_2}$ and $\overrightarrow{\mathbf{P}_3}$ can fall in. The red tetrahedron, which is calculated from $\overrightarrow{\mathbf{P}}_1$, $\overrightarrow{\mathbf{P}}_2$ and $\overrightarrow{\mathbf{P}}_3$, represents the value of the MADET with the selected cut-points geometrically. A perfect diagnostic biomarker will achieve the maximum value of the MADET as 1, and $\overrightarrow{\mathbf{P}}_1$, $\overrightarrow{\mathbf{P}}_2$ and $\overrightarrow{\mathbf{P}}_3$ will be (1, 0, 0), (0, 1, 0) and (0, 0, 1) (Dong et al., 2017). On the contrary, the MADET is 0 when the classification rate vectors fall in the same plane, such as all three vectors or two of them overlap (Dong et al., 2017).

Figure 3.3. Illustration of the MADET for diseases with three-stage.
The yellow part is the area that the MADET can fall in. The red tetrahedron OABC is the actual value of the MADET obtained from the statistics. The larger the red tetrahedron, the better the diagnostic biomarker can correctly discriminate subjects into their true stages. (Source: Dong et al., 2017)


0
0.2
0.4
0.6
0.8
1
0
0.2
0.4
0.6
0.8
1
1
0.8
0.6
0.4
0.2
0
P(T=2|S)
P(T=3|S)
P(T=1|S)
B
O
C
A
$\overrightarrow{P}_2$ (Early diseased)
$\overrightarrow{P}_3$ (Fully diseased)
$\overrightarrow{P}_1$ (Non-diseased)

3.3.3 Criteria for optimal cut-point selection generalized to multi-stage (k>3) setting

The measures discussed above can be generalized to diseases with more than three stages. The extension of the measures had been discussed for the GYI, the CP, the MV, and the MADET to select ‘optimal’ cut-points in the setting with more than three stages (Dong et al., 2017; Nakas et al., 2010). Based on the $\mathbf{P}$ matrix defined in (3.16) and (3.17), the selection criteria are defined as follows:

$$\text{criterion 1: } (c_1, c_2, \ldots, c_{k-1}) = \underset{c_1 < c_2 < \ldots < c_{k-1}}{\arg Max} (p_{1,1} + p_{2,2} + \cdots + p_{k,k} - 1) \;; \tag{3. 47}$$

$$\text{criterion 2: } (c_1, c_2, \ldots, c_{k-1}) = \underset{c_1 < c_2 < \ldots < c_{k-1}}{\arg Min} (\sqrt{(1 - p_{1,1})^2 + (1 - p_{2,2})^2 + \cdots + (1 - p_{k,k})^2}) \tag{3. 48}$$

$$\text{criterion 3: } (c_1, c_2, \ldots, c_{k-1}) = \underset{c_1 < c_2 < \ldots < c_{k-1}}{\arg Max} (p_{1,1} \times p_{2,2} \times \cdots \times p_{k,k}) \,; \tag{3. 49}$$

$$\text{criterion 4: } (c_1, c_2, \ldots, c_{k-1}) = \underset{c_1 < c_2 < \ldots < c_{k-1}}{\arg Max} \left| det(\mathbf{P}) \right| . \tag{3. 50}$$

The four criteria have different properties and perform comparatively well in different distributions. The statistics of the first three criteria incorporate the correct classification rates only; however, the MADET, the last criterion, incorporates both correct and false classification rates. In this dissertation, we proposed a criterion based on the generalized KL measure, which has the characteristics of ruling in/out of a multi-stage disease and incorporates both correct and false classification rates. More details of the proposed criterion in the multi-stage setting are discussed in Chapter 4.

CHAPTER 4
PROPOSED MEASURE

This chapter introduces the generalized total Kullback-Leibler divergence (GTKL) as the diagnostic accuracy measure and optimal cut-points selection criterion for multi-stage (*k>2*) diseases.

4.1. Rule-in and Rule-out information for multi-stage diseases

The Kullback-Leibler (KL) divergence is applied to estimate the rule-in/out information in a diagnostic accuracy test by measuring the distance between two probability distributions (Lee, 1999; Samawi et al., 2019). In the multi-stage setting, the KL can also be generalized to measure the distance between each pair of stages for the estimation of rule-in/out information.

4.1.1. Rule-in information

A continuous biomarker *X* of a multi-stage disease has *k-1* cut-points, where $c' = (c_1, c_2, \ldots, c_{k-1})$ and $c_{j-1} < X \le c_j,\ j = 1, 2, \ldots, k-1$. The generalized KL divergence summarizes the rule-in information of adjacent pairs of stages using the lower stage as the reference. The overall rule-in information is denoted as $D_{in}(f_{i+1}, f_i; i = 1, 2, .., k-1)$, where *i* pertains to the disease stage. From (3.16) and (3.17), we can define the generalized adjacent KL divergence of rule-in as:

$$\begin{aligned} D_{in}(f_{i+1}, f_i; i = 1, 2, .., k-1) &= D_{in}(f_2, f_1) + D_{in}(f_3, f_2) + \cdots + D_{in}(f_k, f_{k-1}) \\ &= \sum_{i=1}^{k-1} \int_{-\infty}^{\infty} f_{i+1}(x) \ln\left(\frac{f_{i+1}(x)}{f_i(x)}\right) dx \\ &= \sum_{i=1}^{k-1} \sum_{j=1}^{k} \int_{c_{j-1}}^{c_j} f_{i+1}(x) \ln\left(\frac{f_{i+1}(x)}{f_i(x)}\right) dx, \{c_0 = -\infty, c_k = \infty\}; \end{aligned} \tag{4. 1}$$

where i denotes the i^{th} stage and j denotes the test in the j^{th} stage.

Using similar arguments as in (Samawi et al., 2019), we have

$$D_{in}(f_{i+1},f_i;i=1,2,..,k-1)=\sum_{i=1}^{k-1}\sum_{j=1}^{k}\int_{c_{j-1}}^{c_j}\frac{F_{i+1}(c_j)-F_{i+1}(c_{j-1})}{F_{i+1}(c_j)-F_{i+1}(c_{j-1})}f_{i+1}(x)\ln\left(\frac{\frac{F_{i+1}(c_j)-F_{i+1}(c_{j-1})}{F_{i+1}(c_j)-F_{i+1}(c_{j-1})}f_{i+1}(x)}{\frac{F_i(c_j)-F_i(c_{j-1})}{F_i(c_j)-F_i(c_{j-1})}f_i(x)}\right)dx$$

$$=\sum_{i=1}^{k-1}\sum_{j=1}^{k}[F_{i+1}(c_j)-F_{i+1}(c_{j-1})]\ln\left(\frac{F_{i+1}(c_j)-F_{i+1}(c_{j-1})}{F_i(c_j)-F_i(c_{j-1})}\right)$$

$$+\sum_{i=1}^{k-1}\sum_{j=1}^{k}\int_{c_{j-1}}^{c_j}[F_{i+1}(c_j)-F_{i+1}(c_{j-1})]f_{i+1(c_j-c_{j-1})}(x)\ln\left(\frac{f_{i+1(c_j-c_{j-1})}(x)}{f_{i(c_j-c_{j-1})}(x)}\right)dx$$

$$=\sum_{i=1}^{k-1}\sum_{j=1}^{k}P_{i+1,j}\ln\left(\frac{p_{i+1,j}}{p_{i,j}}\right)+\sum_{i=1}^{k-1}\sum_{j=1}^{k}\int_{c_{j-1}}^{c_j}P_{i+1,j}\,f_{i+1(c_j-c_{j-1})}(x)\ln\left(\frac{f_{i+1(c_j-c_{j-1})}(x)}{f_{i(c_j-c_{j-1})}(x)}\right)dx$$

$$=D_{in-discrete}(f_{i+1},f_i;i=1,2,..,k-1)+R(c_1,c_2,...,c_{k-1}). \qquad (4.\ 2)$$

Note that as the specific rule-in information between two adjacent disease's stages (say, i+1, i), which denoted by $D_{in(i+1,i)}(f_{i+1},f_i);i=1,2,..,k-1$, is given by

$$D_{in(i+1,i)}(f_{i+1},f_i)=\sum_{j=1}^{k}P_{i+1,j}\ln\left(\frac{p_{i+1,j}}{p_{i,j}}\right)+\sum_{j=1}^{k}\int_{c_{j-1}}^{c_j}P_{i+1,j}\,f_{i+1(c_j-c_{j-1})}(x)\ln\left(\frac{f_{i+1(c_j-c_{j-1})}(x)}{f_{i(c_j-c_{j-1})}(x)}\right)dx$$
$$=D_{in(i+1,i)-discrete}(f_{i+1},f_i)+R_{(i+1,i)}(c_1,c_2,...,c_k)i=1,2,..,k-1.$$

Consequently, (4. 2) can be written as

$$D_{in}(f_{i+1},f_i;i=1,2,..,k-1)=\sum_{i=1}^{k-1}D_{in(i+1,i)-discrete}(f_{i+1},f_i)+\sum_{i=1}^{k-1}R_{(i+1,i)}(c_1,c_2,...,c_{k-1}). \qquad (4.\ 3)$$

4.1.2. Rule-out information

Similar to the rule-in information, the KL divergence can be generalized in the multi-stage setting to estimate the overall adjacent rule-out information using the higher stage as the reference level. The generalized adjacent KL divergence of rule-out is denoted as $D_{out}(f_i,f_{i+1};i=1,2,..,k-1)$. It can be defined and computed as:

$$D_{out}(f_i,f_{i+1};i=1,2,..,k-1)=D_{out}(f_1,f_2)+D_{out}(f_2,f_3)+\cdots+D_{out}(f_{k-1},f_k)$$

$$
\begin{aligned}
&= \sum_{i=1}^{k-1} \int_{-\infty}^{\infty} f_i(x) \ln\left(\frac{f_i(x)}{f_{i+1}(x)}\right) dx \\
&= \sum_{i=1}^{k-1} \sum_{j=1}^{k} \int_{c_{j-1}}^{c_j} f_i(x) \ln\left(\frac{f_i(x)}{f_{i+1}(x)}\right) dx, \{c_0 = -\infty, c_k = \infty\} \\
&= \sum_{i=1}^{k-1} \sum_{j=1}^{k} P_{i,j} \ln\left(\frac{p_{i,j}}{p_{i+1,j}}\right) + \sum_{i=1}^{k-1} \sum_{j=1}^{k} \int_{c_{j-1}}^{c_j} P_{i,j}\, f_{i(c_j - c_{j-1})}(x) \ln\left(\frac{f_{i(c_j - c_{j-1})}(x)}{f_{i+1(c_j - c_{j-1})}(x)}\right) dx. \qquad (4.\ 4)
\end{aligned}
$$

However, as the specific rule-out information between two adjacent diseases' stages (say, *i*, *i+1*), which is denoted by $D_{out(i,i+1)}(f_i, f_{i+1}); i = 1, 2, .., k-1$, is given by

$$
\begin{aligned}
D_{out(i,i+1)}(f_i, f_{i+1}) &= \sum_{j=1}^{k} P_{i,j} \ln\left(\frac{p_{i,j}}{p_{i+1,j}}\right) + \sum_{j=1}^{k} \int_{c_{j-1}}^{c_j} P_{i,j}\, f_{i(c_j - c_{j-1})}(x) \ln\left(\frac{f_{i(c_j - c_{j-1})}(x)}{f_{i+1(c_j - c_{j-1})}(x)}\right) dx \\
&= D_{out(i,i+1)-discrete}(f_i, f_{i+1}) + R_{(i,i+1)}(c_1, c_2, ..., c_k) i = 1, 2, .., k-1.
\end{aligned}
$$

Consequently, (4. 4) can be written as

$$
D_{out}(f_i, f_{i+1}; i = 1, 2, .., k-1) = \sum_{i=1}^{k-1} D_{out(i,i+1)-discrete}(f_i, f_{i+1}) + \sum_{i=1}^{k-1} R_{(i,i+1)}(c_1, c_2, ..., c_{k-1}). \qquad (4.\ 5)
$$

*4.2. Generalized Total Kullback-Leibler divergence (GTKL): summary of rule-in and rule-***out** *information for multi-stage (k>2) diseases*

The GTKL divergence is similar to the total Kullback-Leibler (TKL) divergence for two-stage diseases proposed by Samawi et al. (2019). Using same arguments in Samawi et al. (2019), the comprehensive information measured using Kullback-Leibler divergence is the sum of the generalized rule-in and rule-out information in (4. 2) and (4. 4), and it is computed as

$$
\begin{aligned}
&D_{out}(f_i, f_{i-1}; i = 1, 2, .., k-1) + D_{in}(f_{i+1}, f_i; i = 1, 2, .., k-1) \\
&\qquad = \sum_{i=1}^{k-1} \sum_{j=1}^{k} (P_{i+1,j} - p_{i,j}) \ln\left(\frac{p_{i+1,j}}{p_{i,j}}\right) + R(\mathrm{P}), \qquad (4.\ 6)
\end{aligned}
$$

where the discrete part, $GTKL_{discrete}=\sum_{i=1}^{k-1}\sum_{j=1}^{k}(P_{i+1,j}-p_{i,j})\ln\left(\frac{p_{i+1,j}}{p_{i,j}}\right)$, is corresponding to the information of adjacent pairs based on the cut-points $c'=(c_1,c_2,\ldots,c_{k-1})$; and, the reminder $R(\mathrm{P})$ is computed as follows, using the P matrix defined in (3.17):

$$R(\mathrm{P})=\sum_{i=1}^{k-1}\sum_{j=1}^{k}\int_{c_{j-1}}^{c_j}\left[P_{i+1,j}\,f_{i+1(c_j-c_{j-1})}(x)\ln\left(\frac{f_{i+1(c_j-c_{j-1})}(x)}{f_{i(c_j-c_{j-1})}(x)}\right)+P_{i,j}\,f_{i(c_j-c_{j-1})}(x)\ln\left(\frac{f_{i(c_j-c_{j-1})}(x)}{f_{i+1(c_j-c_{j-1})}(x)}\right)\right]dx. \quad (4.\ 7)$$

4.2.1 Proposed diagnostic accuracy measure and optimal cut-points selection criterion

As dividing the continuous variable into discrete groups, some information would be lost during the process. The summary of rule-in and rule-out information uses a similar concept that the reminder, $R(\mathrm{P})$, is the loss of information when a continuous biomarker is divided into discrete groups when selecting cut-points. To reduce the loss of information, we want to select the optimal cut-points which can minimize the loss of information, $R(\mathrm{P})$, and maximize the information included in the discrete groups. The information from the discrete groups includes the information that a biomarker collected. Using similar arguments as the TKL in the two-stage setting, and the diagnostic accuracy measure in multi-stage setting can be defined as:

$$GTKL_{discrete}=\sum_{i=1}^{k-1}\sum_{j=1}^{k}(P_{i+1,j}-p_{i,j})\ln\left(\frac{p_{i+1,j}}{p_{i,j}}\right). \quad (4.\ 8)$$

When the $GTKL_{discrete}$ is the discrete part from (4. 6). The GTKL is approaching its maximum and reducing the reminder to its minimum in (4. 6), with corresponding optimal cut-points for a biomarker:

$$\begin{aligned}&D_{out}(f_i, f_{i-1}; i=1,2,..,k-1)+D_{in}(f_{i+1}, f_i; i=1,2,..,k-1)\\&\quad=\max_{c_1,c_2,\ldots,c_k}\left(\sum_{i=1}^{k-1}\sum_{j=1}^{k}(P_{i+1,j}-p_{i,j})\ln\left(\frac{p_{i+1,j}}{p_{i,j}}\right)\right)+\min_{c_1,c_2,\ldots,c_k}\left(R(\mathrm{P})\right) \qquad (4.\ 9)\\&\quad=GTKL_{discrete}(c_1,c_2,\ldots,c_k)+\min_{c_1,c_2,\ldots,c_k}\left(R(\mathrm{P})\right).\end{aligned}$$

Therefore, the corresponding optimal with the maximum information from GTKL is computed as:

$$\begin{aligned}\left(c_1, c_2, \ldots, c_{k-1}\right) &= \underset{c_1, c_2, \ldots, c_{k-1}}{\arg\max}\left(GTKL(c_1, c_2, ..., c_k)\right) \\ &= \underset{c_1, c_2, \ldots, c_{k-1}}{\arg\max}\left(\sum_{i=1}^{k-1}\sum_{j=1}^{k}(P_{i+1,j(c_j)} - p_{i,j(c_j)})\ln\left(\frac{p_{i+1,j(c_j)}}{p_{i,j(c_j)}}\right)\right).\end{aligned} \tag{4. 10}$$

4.3. Special case: three-stage (k=3) diseases

In the three-stage setting, using the P matrix in (3.18), the proposed measure using KL divergence and the corresponding GTKL with optimal cut-points can be computed as:

$$\begin{aligned}GTKL &= \left(p_{2,1} - p_{1,1}\right)\ln\left(\frac{p_{2,1}}{p_{1,1}}\right) + \left(p_{2,2} - p_{1,2}\right)\ln\left(\frac{p_{2,2}}{p_{1,2}}\right) + \left(p_{2,3} - p_{1,3}\right)\ln\left(\frac{p_{2,3}}{p_{1,3}}\right) \\ &\quad + \left(p_{3,1} - p_{2,1}\right)\ln\left(\frac{p_{3,1}}{p_{2,1}}\right) + \left(p_{3,2} - p_{2,2}\right)\ln\left(\frac{p_{3,2}}{p_{2,2}}\right) + \left(p_{3,3} - p_{2,3}\right)\ln\left(\frac{p_{3,3}}{p_{2,3}}\right) + R\left(P\right) \\ &= \left(p_{1,1} + p_{2,2} - 1\right)\ln\theta_{1,2} + \left(p_{2,2} + p_{3,3} - 1\right)\ln\theta_{2,3} + R\left(P\right).\end{aligned} \tag{4. 11}$$

Then, the discrete GTKL part, for the three-stage disease, is computed as:

$$GTKL = \left(p_{1,1} + p_{2,2} - 1\right)\ln\theta_{1,2} + \left(p_{2,2} + p_{3,3} - 1\right)\ln\theta_{2,3}; \tag{4. 12}$$

where $\ln\theta_{1,2}$ and $\ln\theta_{2,3}$ are the natural logarithm of the diagnostics odds ratios of pairs of stages 1 and 2, and stages 2 and 3, respectively. As discussed above, when selecting the optimal cut-points, the GTKL with the optimal cut-points (c_1, c_2) can be obtained by maximizing the measure as:

$$GTKL(c_1, c_2) = \max_{c_1 < c_2}\left\{\left(p_{1,1}(c_1) + p_{2,2}(c_1, c_2) - 1\right)\ln\theta_{1,2} + \left(p_{2,2}(c_1, c_2) + p_{3,3}(c_2) - 1\right)\ln\theta_{2,3}\right\}, \tag{4. 13}$$

and, the corresponding cut-points are:

$$(c_1, c_2) = \underset{c_1 < c_2}{\arg\max}\left\{\left(p_{1,1}(c_1) + p_{2,2}(c_1, c_2) - 1\right)\ln\theta_{1,2} + \left(p_{2,2}(c_1, c_2) + p_{3,3}(c_2) - 1\right)\ln\theta_{2,3}\right\}. \tag{4. 14}$$

Besides, when expanding the odds ratios of pairs of stages, the GTKL in three-stage setting can be written as:

$$\begin{aligned} GTKL_{discrete} &= \left(p_{1,1}(c_1)+p_{2,2}(c_1,c_2)-1\right)\ln\left(\frac{p_{1,1}(c_1)\times p_{2,2}(c_1,c_2)}{p_{1,2}(c_1)\times p_{2,1}(c_1,c_2)}\right) \\ &+\left(p_{2,2}(c_1,c_2)+p_{3,3}(c_2)-1\right)\ln\left(\frac{p_{2,2}(c_1,c_2)\times p_{3,3}(c_2)}{p_{3,2}(c_2)\times p_{2,3}(c_1,c_2)}\right). \end{aligned} \tag{4. 15}$$

The expression of GTKL in the three-stage setting can be generalized to k-stages by mathematical induction, and the odds ratio of each adjacent pair of stages can be expressed by two-by-two classification matrix, $\theta_{i,j} = OR\begin{pmatrix} p_{i,j} & p_{i,j+1} \\ p_{i+1,j} & p_{i+1,j+1} \end{pmatrix} = \left(\frac{p_{i,j}\times p_{i+1,j+1}}{p_{i,j+1}\times p_{i+1,j}}\right)$. This expression of $GTKL_{discrete}$ allows to estimate the cut-points for each pair of stages separately, which provides specific information for those two stages.

The Kullback-Leibler divergence is also generalized by summing all the information among pairs of stages 1 and 2, stages 2 and 3, and stages 1 and 3. However, compared to the $GTKL_{discrete}$ using the adjacent pairs (i.e., stages 1 and 2, stages 2 and 3), the performance from power analysis of comprehensive measure using all the pairs is more mediocre. The reason behind it may be that the information covered by the comprehensive measure is overlayed. The information caught from stages 1 and 3 may be overlapped by the other two adjacent pairs and results in over-estimation of the true information. Thus, the focus of the dissertation remains on evaluating the performance of the adjacent pairs. Studies on how to address the overlapped information are encouraged for future research.

CHAPTER 5
SIMULATION STUDY

5.1. Power analysis

Power analysis using the data simulated from normal and gamma distributions is conducted to evaluate the performance of the generalized total Kullback-Leibler divergence (GTKL) comparing with the existing methods, generalized Youden index (GYI), volume under the surface (VUS), known as the hypervolume under the manifold (HUM), and maximum absolute determinant (MADET), in the special case of multi-stage diseases when *k* is 3. The analysis is based on two scenarios: 1) whether a biomarker can discriminate subjects among stages, and 2) whether a biomarker performs the same as the gold standard. The simulation is conducted for the three-stage disease under settings with different values of parameters shown in Table 5.1 and Table 5.3 under different scenarios. For each setting, a random sample is simulated under the null hypothesis (H_0) for 2000 iterations with sample size (20, 20, 20), (50, 50, 50) and (100, 100, 100). The 95% quantile of the estimated statistics (GTKL, GYI, VUS, and MADET) is obtained, then the mean of the two-thousand 95% quantiles serves as the critical value that is used to determine whether the null hypothesis is rejected and further estimates the power of the measure. The statistics of different measures are also estimated under alternative hypothesis (H_a) based on the settings in Table 5.1 and Table 5.3 with 2000 iterations. H_0 is rejected if the statistics estimated under H_a is greater than the critical values. The power is the proportion out of 2000 iterations where the H_0 is rejected.

5.1.1 Scenario I: whether a biomarker can discriminate subjects among stages

In the first scenario, the three disease groups are assumed to have the same distributions under the null hypothesis:

$$N_1(\mu_1,\delta_1) = N_2(\mu_2,\delta_2) = N_3(\mu_3,\delta_3) \text{ or } \Gamma_1(\alpha_1,\beta_1) = \Gamma_2(\alpha_2,\beta_2) = \Gamma_3(\alpha_3,\beta_3)\,.$$

The distributions under the alternative distributions are different from those under the null hypothesis, and they are assumed as the settings in Table 5.1.

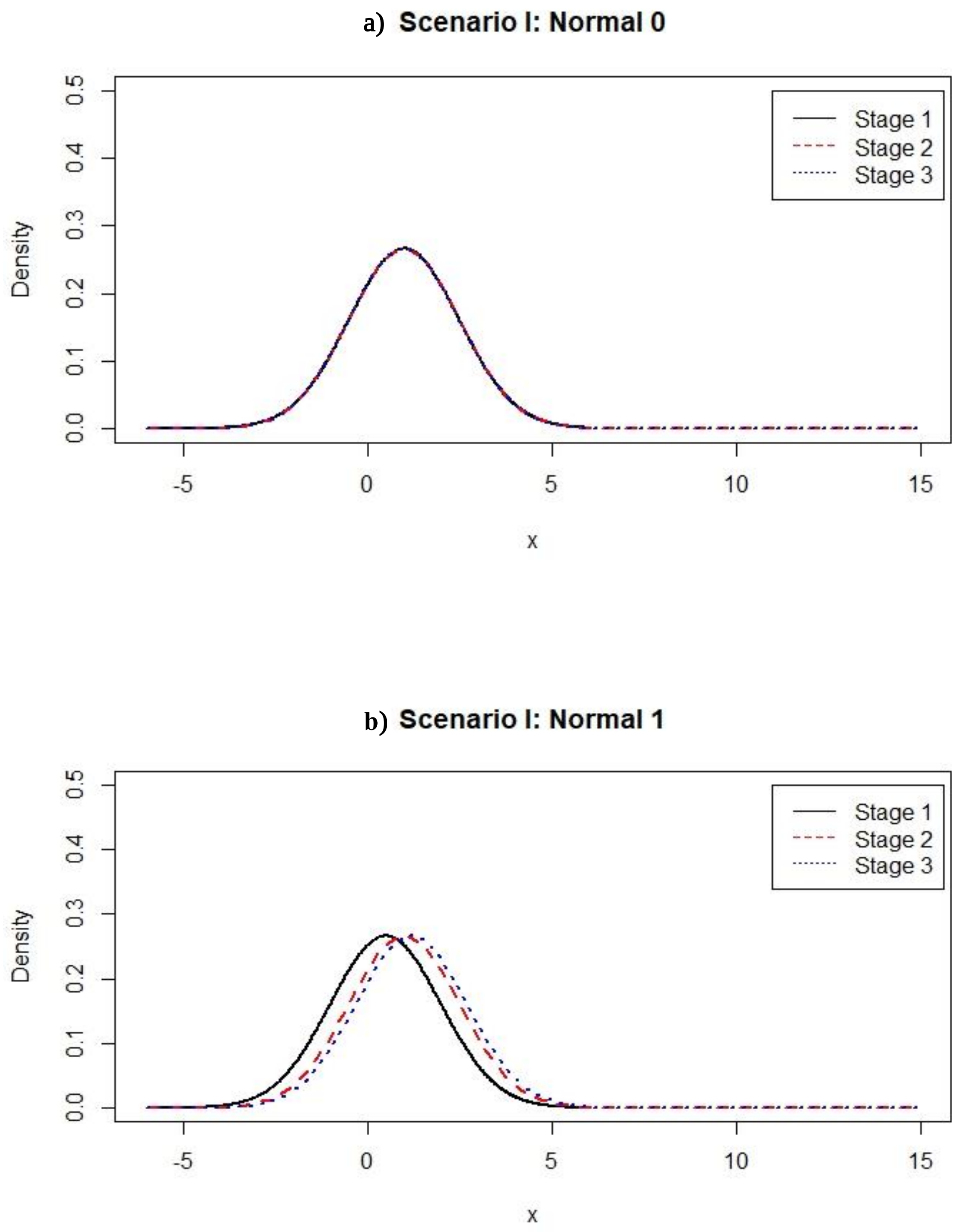
Table 5.1. Distribution settings for power analysis in Scenario I.

	Distribution under H_0			Distribution under H_a		
Setting	Non-disease	Early diseased	Fully Diseased	Non-disease	Early diseased	Fully Diseased
Normal 0	$N_1(1.0, 1.5)$	$N_2(1.0, 1.5)$	$N_3(1.0, 1.5)$	$N_1(1.0, 1.5)$	$N_2(1.0, 1.5)$	$N_3(1.0, 1.5)$
Normal 1				$N_1(0.5, 1.5)$	$N_2(1.0, 1.5)$	$N_3(1.2, 1.5)$
Normal 2				$N_1(1.0, 1.5)$	$N_2(1.0, 2.0)$	$N_3(1.0, 2.5)$
Normal 3				$N_1(1.0, 1.5)$	$N_2(1.0, 1.0)$	$N_3(1.0, 0.8)$
Gamma 0	$\Gamma_1(5.0, 3.0)$	$\Gamma_2(5.0, 3.0)$	$\Gamma_3(5.0, 3.0)$	$\Gamma_1(5.0, 3.0)$	$\Gamma_2(5.0, 3.0)$	$\Gamma_3(5.0, 3.0)$
Gamma 1				$\Gamma_1(5.0, 3.0)$	$\Gamma_2(6.0, 3.0)$	$\Gamma_3(7.0, 3.0)$
Gamma 2				$\Gamma_1(4.5, 3.0)$	$\Gamma_2(5.0, 3.2)$	$\Gamma_3(5.0, 3.5)$
Gamma 3				$\Gamma_1(5.0, 3.0)$	$\Gamma_2(5.0, 4.0)$	$\Gamma_3(5.0, 4.2)$

Normal Distribution: $N(\mu, \delta)$; Gamma Distribution: $\Gamma(\alpha, \beta)$.

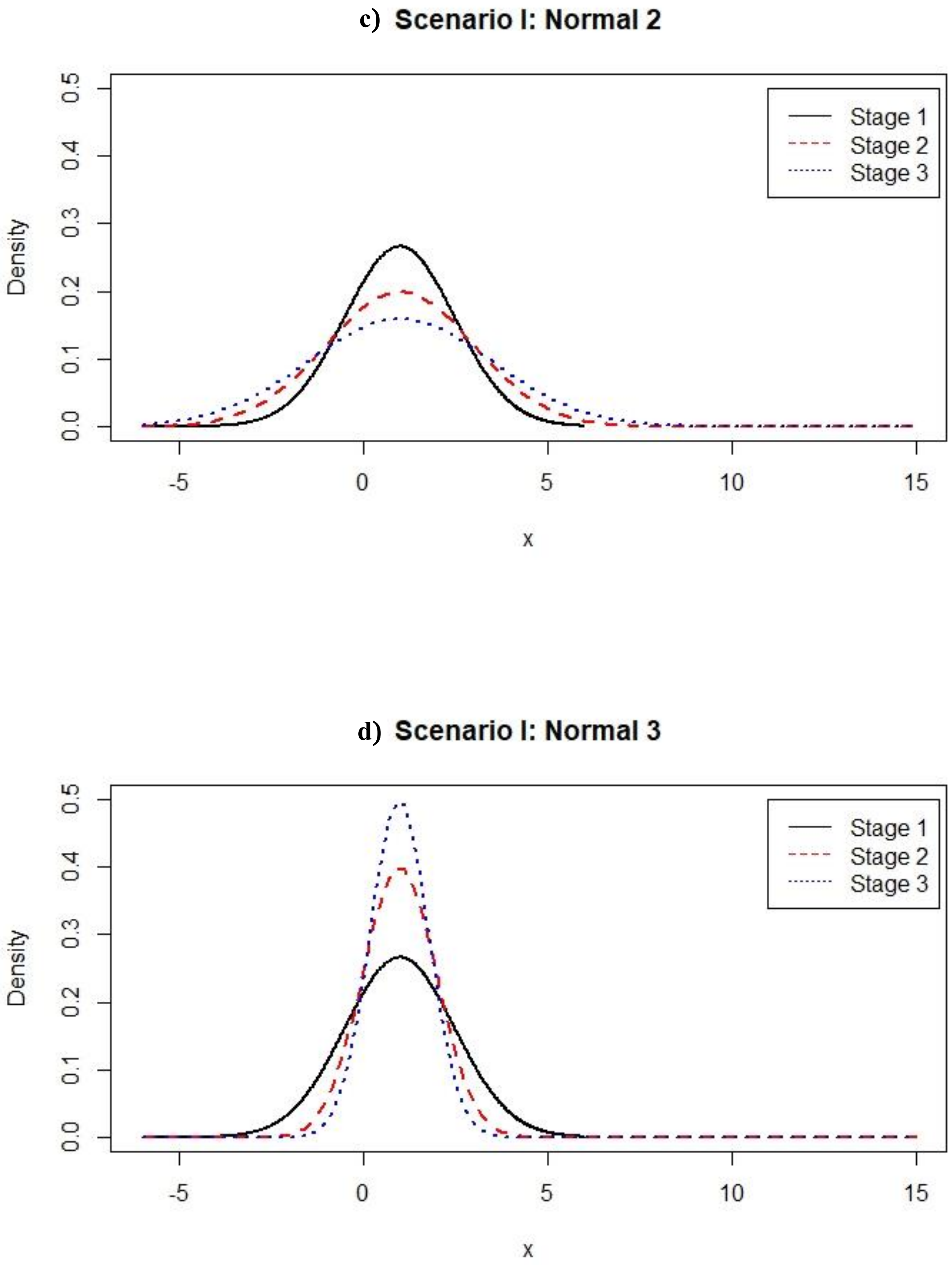
The following figures show density plots for different distributions assumed in the Scenario I (see Figure 5.1). The estimated power under the Scenario I is then summarized in Table 5.2. In this case, when the estimated measure under the alternative hypothesis is higher than the estimated measure under the null hypothesis, it indicates that the null hypothesis is rejected, and the measure is able to assess a biomarker’s capability to discriminate subjects among stages.

Figure 5.1. Density plots of the distributions assumed in Scenario I.



a) Scenario I: Normal 0

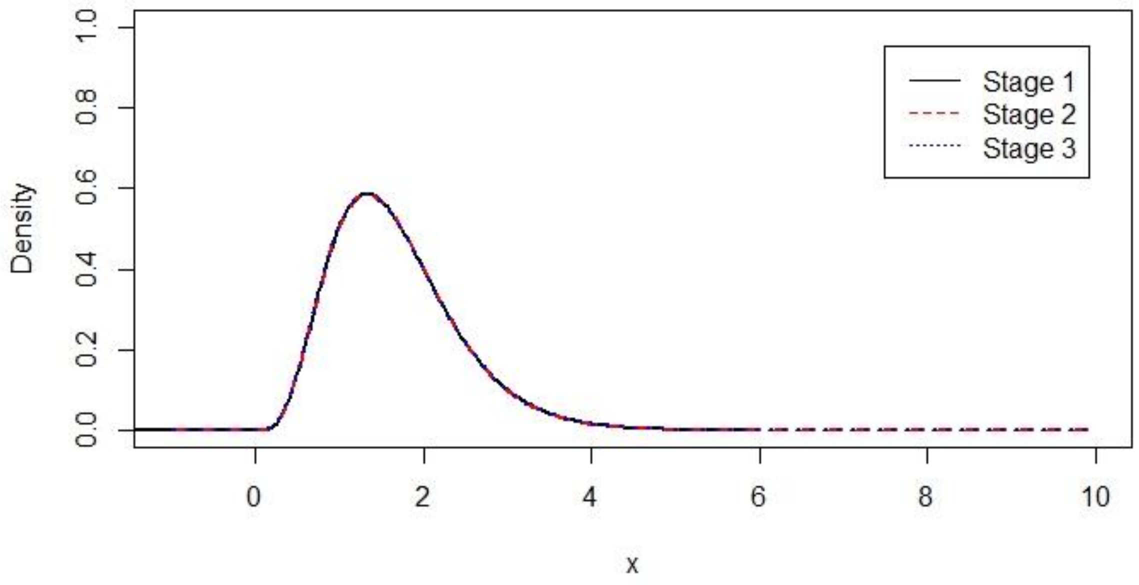
b) Scenario I: Normal 1



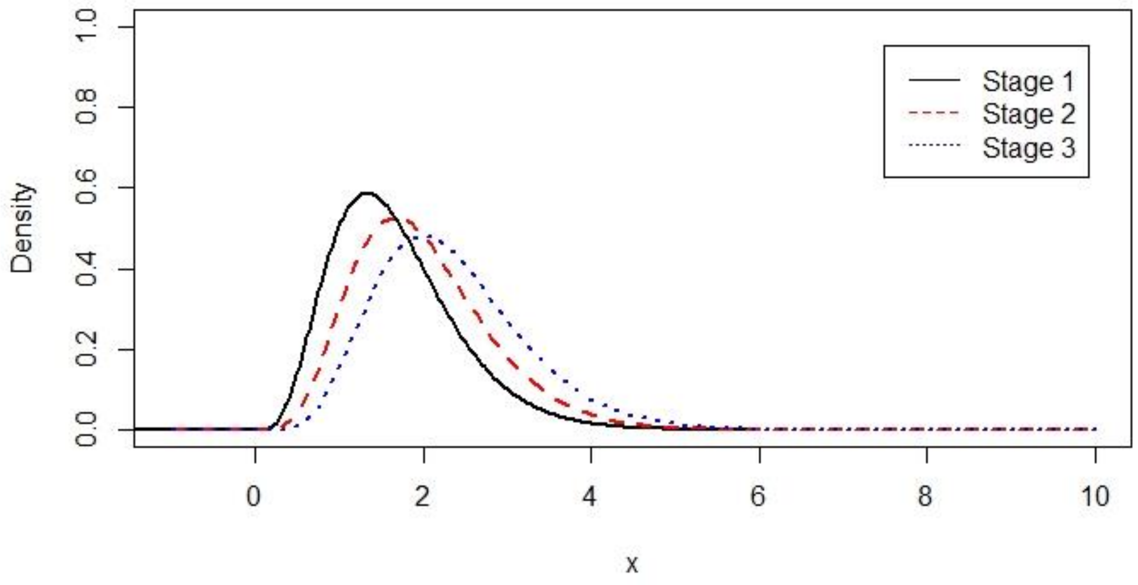
c) Scenario I: Normal 2

d) Scenario I: Normal 3

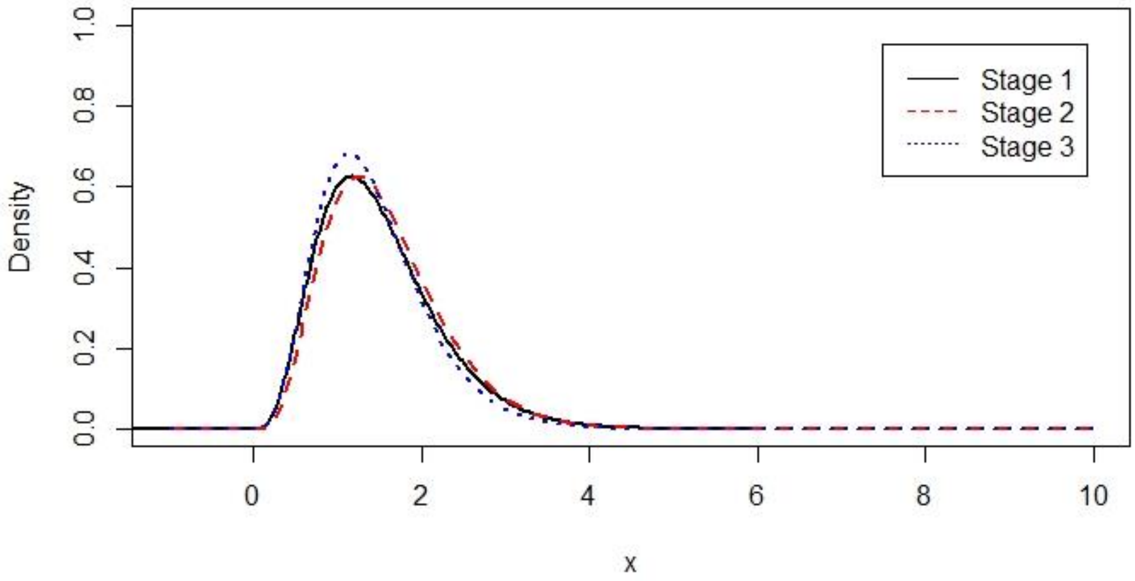
e) **Scenario I: Gamma 0**



f) **Scenario I: Gamma 1**



g) Scenario I: Gamma 2



h) Scenario I: Gamma 3

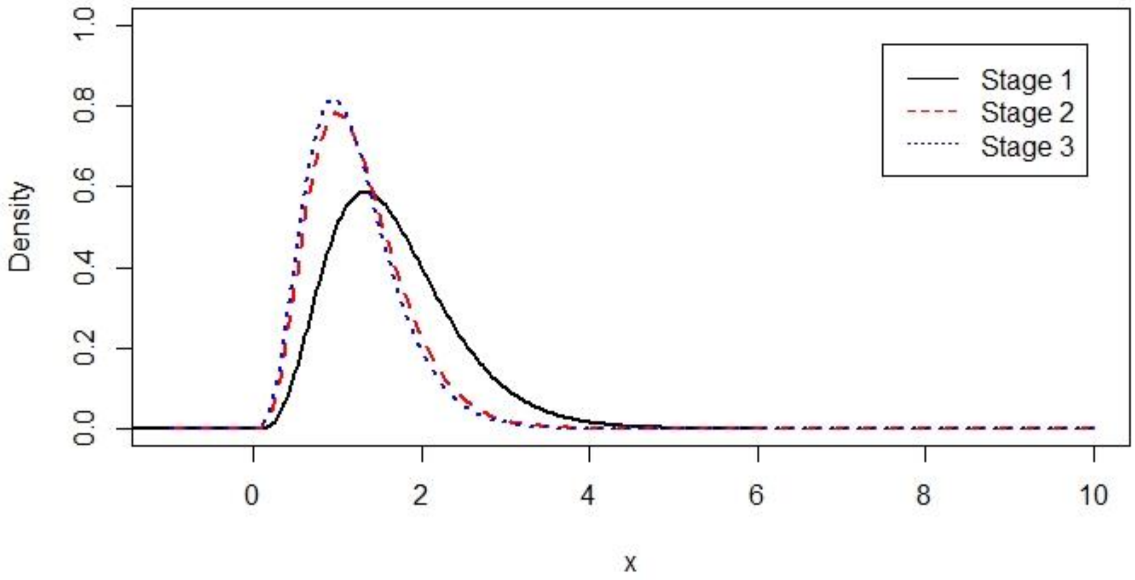


Table 5.2. Simulated power in Scenario I.

Setting	Sample Size	GTKL	GYI	VUS	MADET
Normal 0	20	0.0630	0.0565	0.0635	0.0640
	50	0.0360	0.0505	0.0515	0.0590
	100	0.0535	0.0455	0.0490	0.0475
Normal 1	20	0.0870	0.3010	0.3435	0.1630
	50	0.1540	0.5160	0.6380	0.2985
	100	0.3050	0.8350	0.9130	0.4705
Normal 2	20	0.1190	0.0815	0.0485	0.1445
	50	0.2860	0.1330	0.0405	0.2815
	100	0.7315	0.3730	0.0505	0.4925
Normal 3	20	0.1765	0.1255	0.0875	0.1905
	50	0.6010	0.2965	0.1160	0.3980
	100	0.9330	0.5405	0.1370	0.6100
Gamma 0	20	0.0715	0.0630	0.0590	0.0810
	50	0.0610	0.0575	0.0455	0.0715
	100	0.0525	0.0600	0.0490	0.0850
Gamma 1	20	0.1880	0.6540	0.7620	0.4120
	50	0.4740	0.9630	0.9815	0.7145
	100	0.9110	1.0000	1.0000	0.9125
Gamma 2	20	0.0870	0.0390	0.0300	0.0690
	50	0.1610	0.0505	0.0235	0.1240
	100	0.2115	0.0630	0.0170	0.1865
Gamma 3	20	0.2990	0.0060	0.0000	0.3570
	50	0.6325	0.0045	0.0000	0.6435
	100	0.9445	0.0030	0.0000	0.8535

Under Scenario I, none of the measures has a dominant well-performance in all kinds of distributions. Data with distribution as Normal 1 should consider GYI and VUS as the measures to assess the accuracy of the diagnostic test; and, data with distribution as Normal 2, Normal 3 and Gamma 3 can use GTKL as the accuracy measure of the diagnostic test. Data with distribution as Gamma 1 can use any of the measures to test the accuracy since the power is similar when the sample size is 100; however, GYI and VUS provide more reliable results as both of them also perform well with smaller sample sizes. None of the measures perform well with the distribution as Gamma 2 though GTKL gives the highest power. In general, GTKL is more inferior to other measures when sample means are different among groups, with the same variance. In contrast, GTKL outperforms the others when the sample variance gets more diverse, while the sample mean remains the same among groups. The results provide hints for diagnostic accuracy measures that our proposed measure, the GTKL, can catch the distinctions among stages when the group sample variance is diverse but has a similar group sample mean.

5.1.2 Scenario II: compare the performance of two biomarkers

In the second scenario, the three stages are assumed to have different distributions under the null hypothesis:

$$N_1(\mu_1,\delta_1) \neq N_2(\mu_2,\delta_2) \neq N_3(\mu_3,\delta_3) \text{ or } \Gamma_1(\alpha_1,\beta_1) \neq \Gamma_2(\alpha_2,\beta_2) \neq \Gamma_3(\alpha_3,\beta_3) .$$

The distributions of the three stages have the settings assumed in Table 3, and Figure 4 displays the distributions with density plots. Table 5.3 shows a summary of the estimated power under those settings.

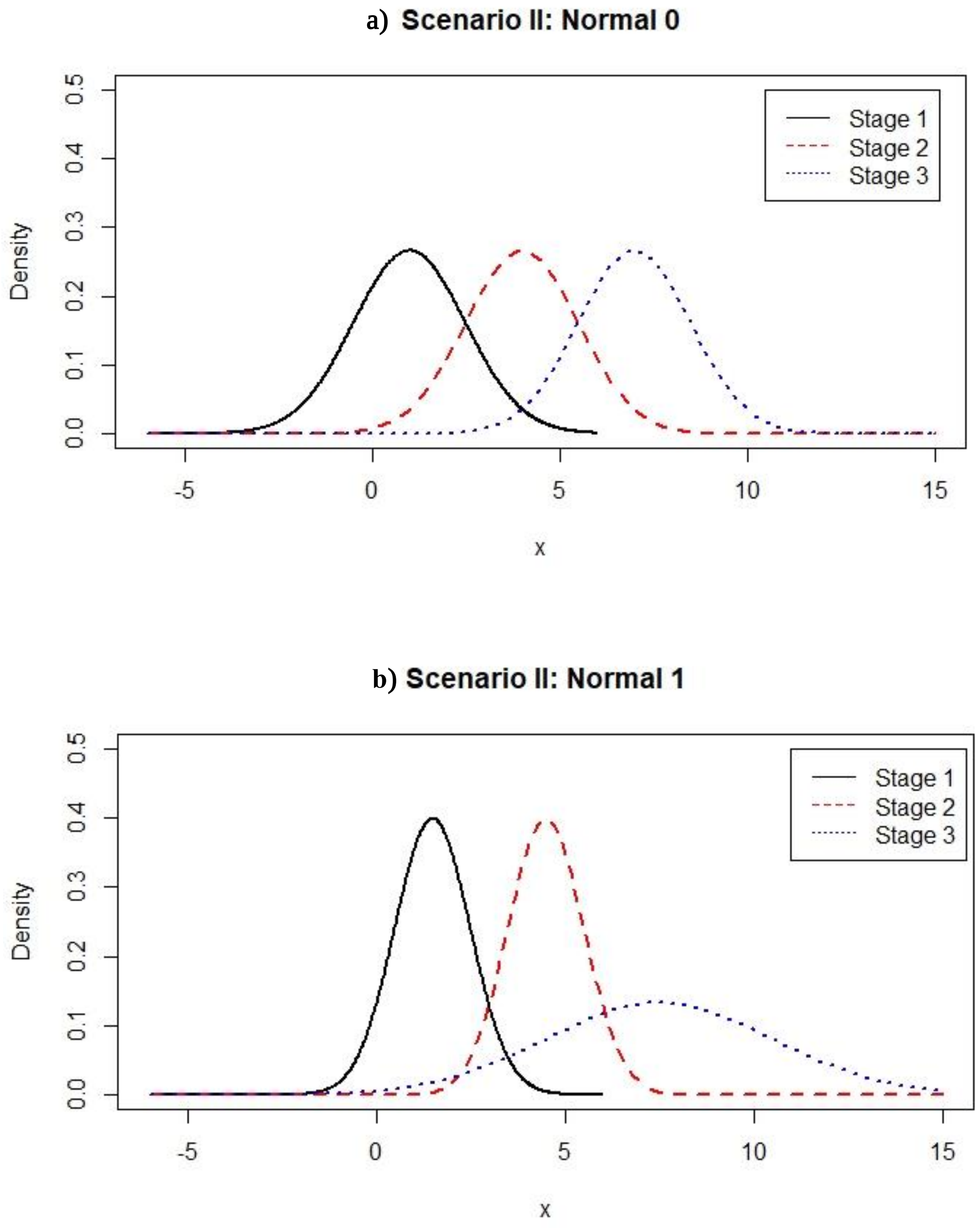
Table 5.3. Distribution settings for power analysis in Scenario II.

	Distribution under H_0			Distribution under H_a		
Setting	Non-disease	Early diseased	Fully Diseased	Non-disease	Early diseased	Fully Diseased
Normal 0	N_1(1.0, 1.5)	N_2(4.0, 1.5)	N_3(7.0, 1.5)	N_1(1.0, 1.5)	N_2(4.0, 1.5)	N_3(7.0, 1.5)
Normal 1				N_1(1.5, 1.0)	N_2(4.5, 1)	N_3(7.5, 3)
Normal 2				N_1(1.5, 1.5)	N_2(4.9, 1)	N_3(7.5, 2)
Normal 3				N_1(1.5, 1.5)	N_2(6, 2.5)	N_3(7.1, 2.5)
Gamma 0	Γ_1(2, 2)	Γ_2(5, 2)	Γ_3(7, 2)	Γ_1(2, 2)	Γ_2(5, 2)	Γ_3(7, 2)
Gamma 1				Γ_1(2, 2)	Γ_2(5.5, 2)	Γ_3(7., 2)
Gamma 2				Γ_1(3, 3)	Γ_2(6.5, 2.5)	Γ_3(9, 2.2)
Gamma 3				Γ_1(3, 3)	Γ_2(6.9, 2.5)	Γ_3(7.3, 2)

Normal Distribution: N (μ, δ); Gamma Distribution: Γ (α, β).

The density plots shown in Figure 5.2 display the distribution under settings assumed in Table 5.3. The estimated power under Scenario II is then shown in Table 5.4. In this case, when the estimated measure under the alternative hypothesis is higher than the estimated measure under the null hypothesis, it indicates that the measure can assess the difference between a biomarker and the gold standard.

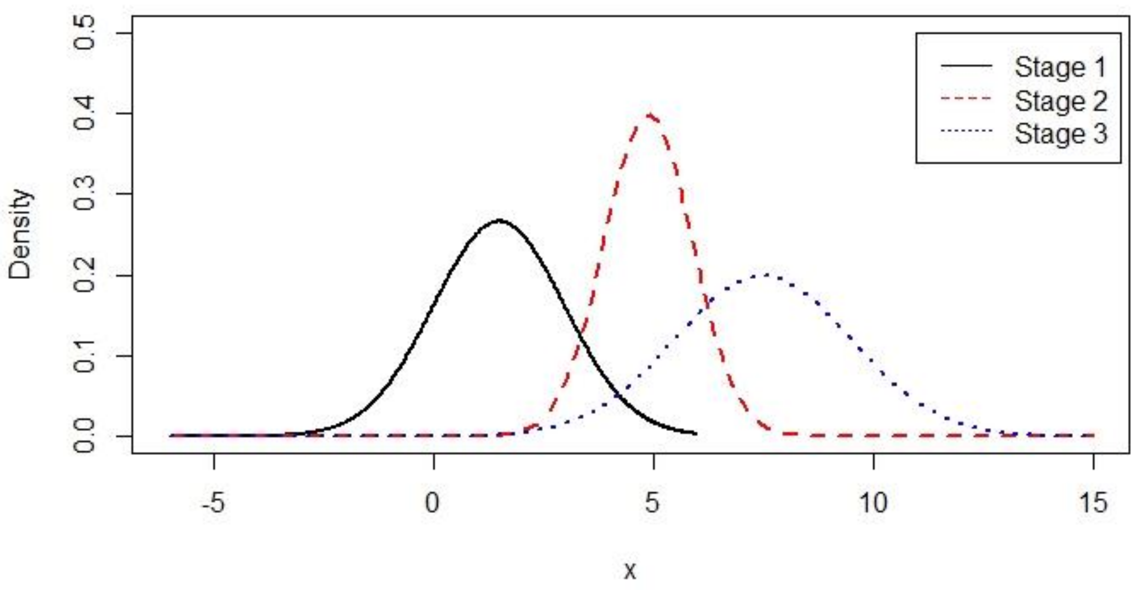
Figure 5.2. Density plots of the distributions assumed in Scenario II.



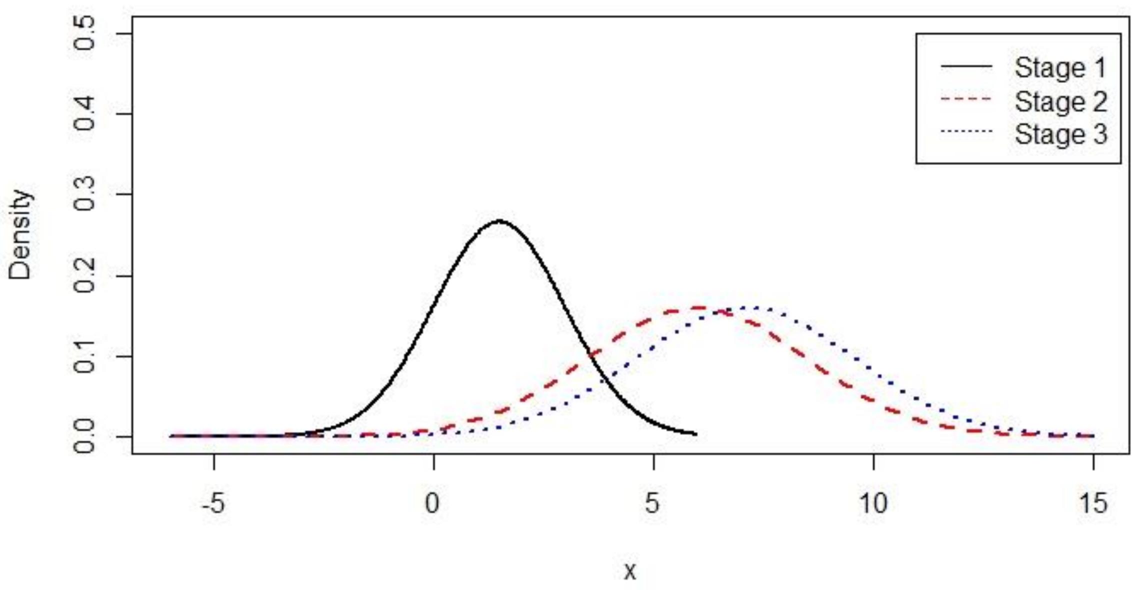
a) Scenario II: Normal 0

b) Scenario II: Normal 1

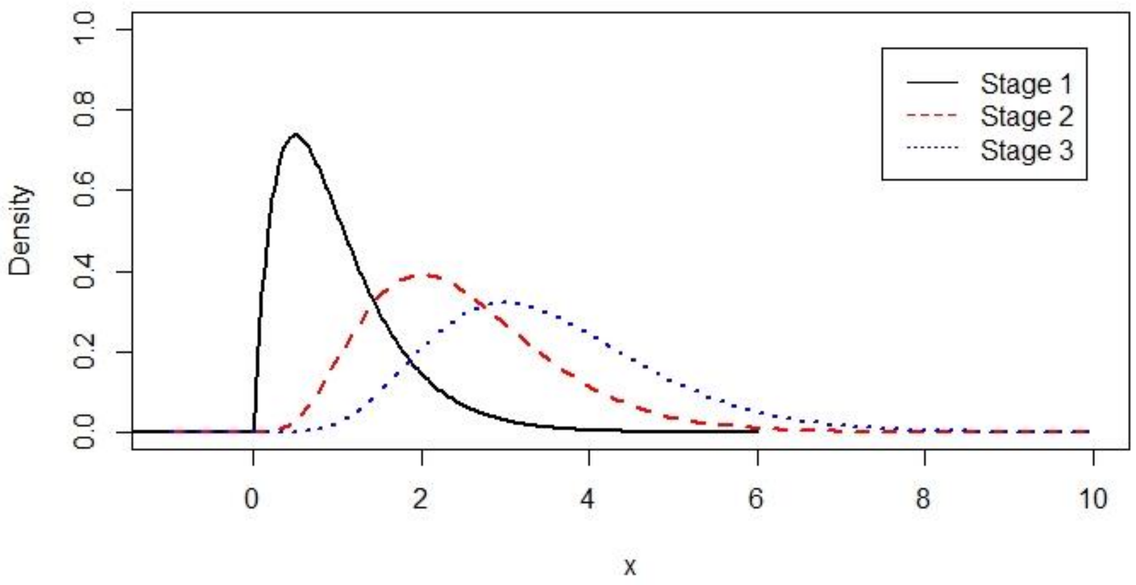
c) Scenario II: Normal 2



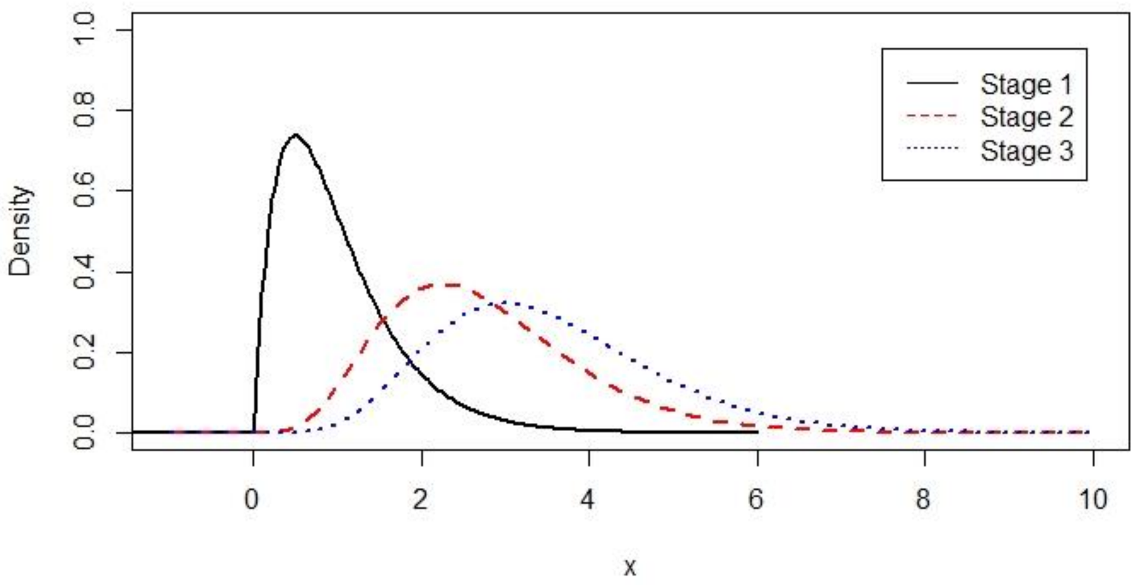
d) Scenario II: Normal 3



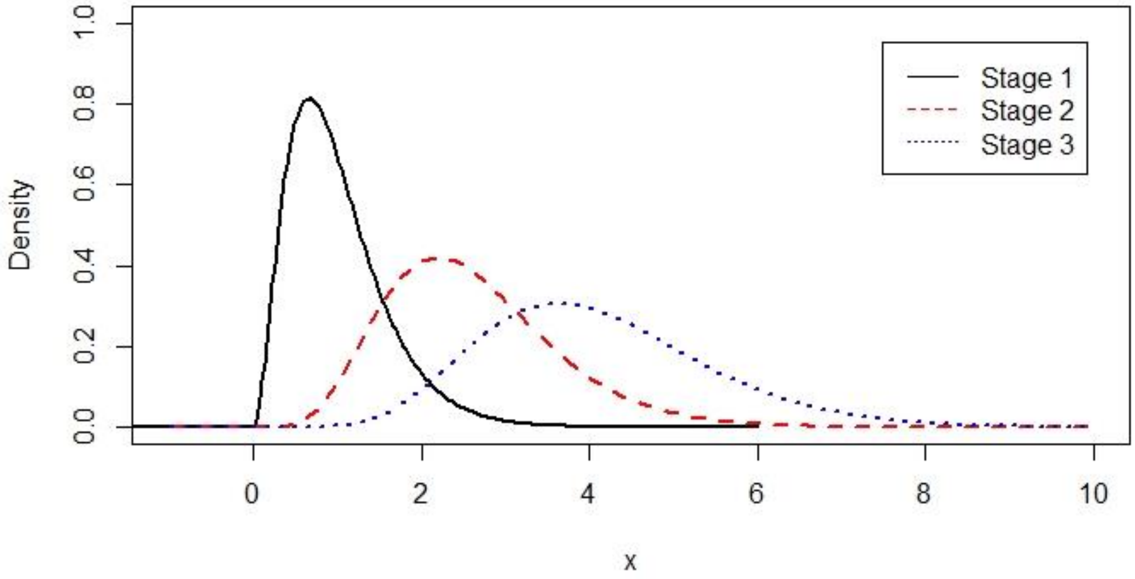
e) Scenario II: Gamma 0



f) Scenario II: Gamma 1



g) **Scenario II: Gamma 2**



h) **Scenario II: Gamma 3**

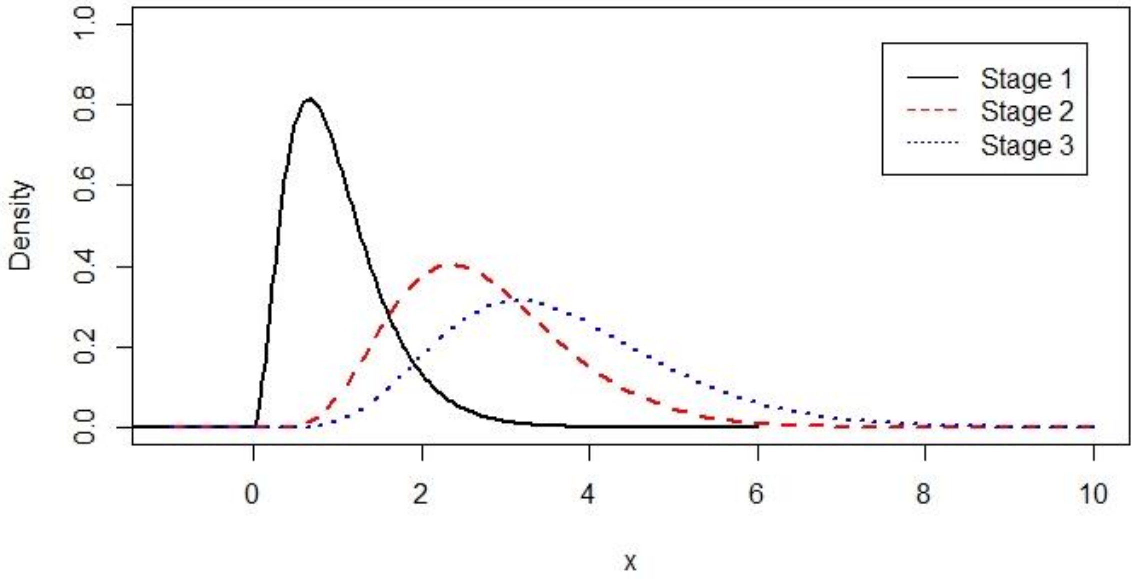


Table 5.4. Simulated power in Scenario II.

Setting	Sample Size	GTKL	GYI	VUS	MADET
Normal 0	20	0.0435	0.0530	0.0785	0.0510
	50	0.0685	0.0565	0.0470	0.0535
	100	0.0510	0.0535	0.0505	0.0530
Normal 1	20	0.6020	0.5790	0.4690	0.5880
	50	0.9145	0.9325	0.7480	0.9330
	100	0.9895	0.9965	0.9555	0.9950
Normal 2	20	0.4800	0.3335	0.2215	0.3190
	50	0.8370	0.6285	0.3660	0.5740
	100	0.9555	0.8710	0.5520	0.8265
Normal 3	20	0.4385	0.0015	0	0.0010
	50	0.8135	0	0	0
	100	0.9430	0	0	0
Gamma 0	20	0.0690	0.0715	0.0630	0.0660
	50	0.0705	0.0700	0.0470	0.0595
	100	0.0560	0.0735	0.0555	0.0725
Gamma 1	20	0.1825	0.1130	0.0815	0.0990
	50	0.2330	0.1830	0.1065	0.1420
	100	0.3405	0.2235	0.0980	0.1395
Gamma 2	20	0.4125	0.5090	0.5020	0.4735
	50	0.6085	0.8260	0.8305	0.7680
	100	0.9010	0.9815	0.9885	0.9660
Gamma 3	20	0.3855	0.1900	0.1050	0.1410
	50	0.6205	0.3135	0.1415	0.1980
	100	0.8155	0.4280	0.1575	0.2365

Under Scenario II, the results also show no dominant measure in all kinds of distributions. Data with distributions as Normal 1 and Gamma 2 have apparent differences among the three stages. Any of the measures can be used in such situations as the power of all the measures is high. GTKL, GYI and MADET show solid performance with data that has distribution as Normal 2; however, VUS shows the worse performance. In the distributions shown in Normal 3 and Gamma 3, GTKL has much better performance than the other measures. The data with such distributions shows substantial overlap between the middle and the last stages. In this case, GTKL is suggested as the measure for accuracy tests. Lastly, all measures have low power with the distribution shown in Gamma 1. Although GTKL outperforms the others, its power is 0.3405 which remains low even the sample size is increased to 100. All in all, GTKL shows better performance than the others when the middle and last stages have heavy overlaps.

5.2. Optimal cut-points selection

In the diagnostic study, the optimal cut-points are estimated for continuous biomarkers to diagnose whether a subject having the disease of interest. Of multi-stage diseases, more than one optimal cut-points are expected to identify the staging of the subjects. Among the measures that we have discussed above, the GTKL, GYI, and MADET can be directly used as the selection criteria; however, HUM, known as the VUS for three-stage diseases, cannot be used as the criteria yet have to use the closet-to-perfection (CP) and the maximum volume (MV) measures that are derived based on the HUM. In this section, we provide the estimation of the GTKL, GYI, VUS, and MADET in the three-stage setting, as well as the corresponding optimal cut-points selected by the GTKL, GYI and MADET, and the CP and MV with respect to VUS in the simulation.

5.2.1 Indices to evaluate the selection criteria

To compare the optimal cut-point criteria in three-stage setting, the relative bias (RBias), the normalized root-mean-square error (NRMSE), the total correct classification rate (TCCR), the percentage of loss of the total correct classification rate (LCCR%), and the maximum-minimum difference (MMDIF) are obtained in the simulation.

The NRMSE is provided in the simulation as an example to demonstrate the estimation by different selection criteria. Although the root-mean-square error (RMSE) is frequently used for measuring the power of the estimation of a model, the RMSE cannot be used in comparing values with different scales. In this study, the five diagnostic accuracy measures have different scales that the values can be very diverse from each other, so the NRMSE is a more appropriate index to measure the accuracy of the estimation among the measures in this study. Additionally, the ranges of the measures are widely different, and the RMSE is therefore normalized by dividing the RMSE by the range (i.e., the difference of the maximum and minimum estimates) of a measure. In the simulation, in order to obtain the NRMSE, the RBias of the optimal statistics was estimated from N rounds of iteration, and its variance is captured by the following formulas:

$$\text{RBias } (\hat{T}) = \left[\left(\frac{1}{N} \sum_{i=1}^{N} \hat{T}_i - T \right) \Big/ T \right],$$

$$\text{Variance } (\hat{T}) = \sum_{i}^{N} (\hat{T}_i - T)^2 \Big/ (N-1) .$$

Then, the RMSE is computed as:

$$RMSE = \sqrt{Bias^2 + Variance(\hat{T})} \, ;$$

where $Bias = \hat{T}_i - T$. $\hat{T}_i$ and T denote estimated value and the true value of the measure, respectively. The RMSE is then normalized by dividing the RMSE by the difference between maximum and minimum as:

$$NRMSE = \frac{RMSE}{\max(\hat{T}) - \min(\hat{T})} .$$

The TCCR is defined as the summation of all the correct classification rates in all stages, and it is computed as:

$$TCCR = \sum_{i}^{k} p_{i,i} .$$

The LCCR% is an index to compare the TCCR among the criteria, using the Youden index as reference (Attwood et al., 2014; Dong et al., 2017). The LCCR% using the GYI as the reference is computed as:

$$LCCR\% = \frac{TCCR_{GYI} - TCCR_{M}}{TCCR_{GYI}} \times 100\% \,,$$

where $TCCR_{GYI}$ is the total correct classification rate calculated based on the optimal cut-points selected by the GYI, while $TCCR_{M}$ is based on the method *M* (i.e., GTKL, CP, MADET, and MV). A positive LCCR% indicates that a measure loses the information of the total correct classification rate compared to the GYI; in contrast, a negative LCCR indicates a measure having more information on the total correct classification rate compared to the GYI. A larger LCCR indicates a measure loses more correct information in contrast to the reference measure. In addition to the LCCR%, the maximum-minimum difference (MMDIF) is proposed by Dong et al. (2017) to measure the balance of correct classification rates among disease stages, and it is computed as:

$$MMDIF = \frac{\max(p_{11}, \ldots, p_{k,k}) - \min(p_{11}, \ldots, p_{k,k})}{\min(p_{11}, \ldots, p_{k,k})} \,.$$

A smaller MMDIF indicates a more balanced measure in selecting optimal cut-points.

5.2.2 Simulation

Simulation is conducted under the three-stage setting to assess the performance of the optimal cut-points selection criteria. The data is simulated for three stages using normal and gamma distribution, and the settings for the parameters are shown in Table 5.5. For each setting, the random sample is simulated for 10000 iterations (*N*) with the sample sizes (20, 20, 20), (50, 50, 50), (100, 100, 100) and (100, 50, 30). The optimal statistics and optimal cut-points are estimated using the smoothed kernel approach (Simonoff, 2012; Wand & Jones, 1994).

Table 5.6 – 5.11 summarized the estimated optimal statistics of the diagnostic measure and the corresponding optimal cut-points selected by each measure. The estimation is conducted

using kernel estimation as an example. The results give a reasonable estimation, and Rbias and NRMSE allow to compare the results among all the measures as they normalized the estimates by the range of the estimates from each measure.

As shown in Table 5.12 - 5.15, the GTKL is not as balanced as the other measures. The GYI is the most balanced measure among all. The CP, MV, and MADET have similar results. Some settings show that the GTKL has the highest CCR in the middle stage, such as Normal 3 and Gamma 6 from cut-points selection simulation. In setting Normal 1 and Gamma 4, the GTKL performs better with the smaller sample size, and the MADET catches up in the larger sample size. Generally, the GYI has the highest TCCR with highest CCRs in the first and last stages.

Table 5.5. Distribution settings for optimal cut-point selection.

	Distributions		
Settings	Non-disease	Early diseased	Fully Diseased
Normal 1	$N_1(2.0, 2.0)$	$N_2(4.0, 2.0)$	$N_3(6.0, 2.0)$
Normal 2	$N_1(2.0, 2.0)$	$N_2(4.0, 2.5)$	$N_3(6.0, 3.5)$
Normal 3	$N_1(2.0, 2.0)$	$N_2(3.0, 2.0)$	$N_3(5.0, 2.0)$
Normal 4	$N_1(2.0, 2.0)$	$N_2(4.5, 2.0)$	$N_3(6.0, 1.5)$
Normal 5	$N_1(2.0, 2.0)$	$N_2(5.0, 2.0)$	$N_3(6.0, 1.5)$
Normal 6	$N_1(2.0, 2.0)$	$N_2(4, 1.5)$	$N_3(5.0, 1.5)$
Gamma 1	$\Gamma_1(2.0, 2.0)$	$\Gamma_2(4.0, 2.0)$	$\Gamma_3(6.0, 2.0)$
Gamma 2	$\Gamma_1(2.0, 2.0)$	$\Gamma_2(5.0, 2.0)$	$\Gamma_3(6.0, 2.0)$
Gamma 3	$\Gamma_1(2.5, 2.0)$	$\Gamma_2(4.0, 2.0)$	$\Gamma_3(5.5, 2.0)$
Gamma 4	$\Gamma_1(2.0, 2.0)$	$\Gamma_2(5.0, 2.5)$	$\Gamma_3(7.0, 3.0)$
Gamma 5	$\Gamma_1(2.0, 2.0)$	$\Gamma_2(5.0, 2.5)$	$\Gamma_3(5.5, 3.0)$
Gamma 6	$\Gamma_1(2.0, 2.0)$	$\Gamma_2(5.5, 2.0)$	$\Gamma_3(6.5, 2.5)$

Normal Distribution: $N(\mu, \delta)$; Gamma Distribution: $\Gamma(\alpha, \beta)$

Table 5.6. Relative bias and normalized root-mean-square error of the estimated optimal statistics in the normal distribution.

Sample size			n_1= n_2= n_3 20		n_1= n_2= n_3 50		n_1= n_2= n_3 100		n_1=100; n_2=50; n_3=30	
Setting	Measure	Stat*	RBias	NRMSE	RBias	NRMSE	RBias	NRMSE	RBias	NRMSE
Normal 1	GTKL	1.5236	0.3444	0.1256	0.0743	0.0913	0.0031	0.1141	0.0842	0.1144
	MADET	0.1040	0.0444	0.1378	-0.0448	0.1441	-0.0630	0.1405	-0.0455	0.1287
	GYI	0.7658	-0.0042	0.1294	-0.0291	0.1470	-0.0322	0.1383	-0.0275	0.1299
	CP	0.7253	0.0288	0.1363	0.0293	0.1459	0.0261	0.1477	0.0290	0.1378
	MV	0.1971	-0.0316	0.1271	-0.0500	0.1419	-0.0493	0.1437	-0.0490	0.1260
Normal 2	GTKL	1.1784	0.5043	0.1682	0.1822	0.1348	0.0846	0.1184	0.1722	0.1332
	MADET	0.0783	0.0779	0.1504	-0.0284	0.1417	-0.0440	0.1251	-0.0356	0.1451
	GYI	0.6411	0.0154	0.1319	-0.0188	0.1300	-0.0252	0.1256	-0.0133	0.1353
	CP	0.8033	0.0161	0.1368	0.0198	0.1465	0.0180	0.1284	0.0182	0.1395
	MV	0.1550	-0.0123	0.1307	-0.0389	0.1421	-0.0400	0.1259	-0.0331	0.1350
Normal 3	GTKL	0.9794	0.5165	0.1403	0.1376	0.1344	0.0350	0.1077	0.1686	0.1208
	MADET	0.0435	0.4716	0.1603	0.1600	0.1482	0.0577	0.1500	0.0987	0.1316
	GYI	0.5803	0.0405	0.1427	-0.0068	0.1300	-0.0185	0.1338	-0.0113	0.1312
	CP	0.8348	0.0127	0.1470	0.0169	0.1367	0.0154	0.1460	0.0188	0.1343
	MV	0.1392	-0.0062	0.1413	-0.0345	0.1346	-0.0364	0.1414	-0.0395	0.1259
Normal 4	GTKL	1.9673	0.2834	0.1461	0.0741	0.1382	0.0081	0.1235	0.0836	0.1373
	MADET	0.1044	0.0460	0.1451	-0.0583	0.1485	-0.0828	0.1211	-0.0567	0.1494
	GYI	0.8159	-0.0126	0.1326	-0.0283	0.1388	-0.0326	0.1247	-0.0284	0.1442
	CP	0.7146	0.0317	0.1454	0.0303	0.1448	0.0285	0.1347	0.0309	0.1435
	MV	0.2045	-0.0379	0.1364	-0.0514	0.1351	-0.0535	0.1258	-0.0522	0.1378
Normal 5	GTKL	2.1420	0.2684	0.0951	0.0642	0.1164	0.0037	0.1267	0.0597	0.1120
	MADET	0.0815	0.2113	0.1308	-0.0045	0.1334	-0.0526	0.1353	0.0167	0.1449
	GYI	0.7952	-0.0073	0.1299	-0.0268	0.1318	-0.0277	0.1339	-0.0235	0.1341
	CP	0.7363	0.0263	0.1362	0.0276	0.1460	0.0236	0.1358	0.0256	0.1437
	MV	0.1922	-0.0315	0.1281	-0.0501	0.1376	-0.0468	0.1318	-0.0451	0.1351
Normal 6	GTKL	1.9305	0.3244	0.1484	0.1315	0.1315	0.0646	0.1376	0.1347	0.1269
	MADET	0.1116	0.0130	0.1263	-0.0625	0.1425	-0.0693	0.1132	-0.0486	0.1424
	GYI	0.7864	-0.0099	0.1232	-0.0268	0.1336	-0.0288	0.1240	-0.0265	0.1297
	CP	0.7300	0.0279	0.1287	0.0275	0.1461	0.0243	0.1274	0.0266	0.1436
	MV	0.1955	-0.0313	0.1197	-0.0473	0.1371	-0.0465	0.1170	-0.0454	0.1386

Stat*: optimal statistics of the measure; RBias: relative bias; NRMSE: normalized root-mean-square error.

Table 5.7. Relative bias and normalized root-mean-square error of the estimated optimal statistics in the gamma distribution.

Sample size			$n_1= n_2= n_3$ 20		$n_1= n_2= n_3$ 50		$n_1= n_2= n_3$ 100		n_1=100; n_2=50; n_3=30	
Setting	Measure	Stat*	RBias	NRMSE	RBias	NRMSE	RBias	NRMSE	RBias	NRMSE
Gamma 1	GTKL	1.8816	0.3534	0.1331	0.0678	0.1158	-0.0270	0.1186	0.0142	0.1158
	MADET	0.1301	-0.0023	0.1351	-0.0778	0.1327	-0.0846	0.1471	-0.0611	0.1409
	GYI	0.8312	-0.0089	0.1304	-0.0307	0.1365	-0.0327	0.1334	-0.0275	0.1413
	CP	0.6866	0.0362	0.1304	0.0368	0.1467	0.0319	0.1500	0.0338	0.1402
	MV	0.2208	-0.0385	0.1215	-0.0574	0.1402	-0.0545	0.1461	-0.0517	0.1323
Gamma 2	GTKL	2.5377	0.3152	0.1518	0.0653	0.1368	-0.0224	0.1295	-0.0120	0.1159
	MADET	0.0827	0.2596	0.1497	0.0391	0.1439	-0.0146	0.1438	0.1033	0.1383
	GYI	0.7927	0.0119	0.1340	-0.0174	0.1427	-0.0215	0.1245	-0.0092	0.1162
	CP	0.7216	0.0339	0.1467	0.0337	0.1497	0.0280	0.1351	0.0296	0.1343
	MV	0.2018	-0.0396	0.1403	-0.0568	0.1389	-0.0512	0.1319	-0.0475	0.1258
Gamma 3	GTKL	1.0411	0.5546	0.1455	0.1251	0.1223	0.0038	0.1082	0.0849	0.0882
	MADET	0.0607	0.2179	0.1293	0.0229	0.1470	-0.0354	0.1132	0.0379	0.1458
	GYI	0.6268	0.0225	0.1333	-0.0151	0.1294	-0.0248	0.1200	-0.0123	0.1255
	CP	0.8048	0.0193	0.1290	0.0211	0.1380	0.0198	0.1235	0.0198	0.1301
	MV	0.1536	-0.0186	0.1219	-0.0411	0.1341	-0.0441	0.1140	-0.0379	0.1278
Gamma 4	GTKL	1.7228	0.2428	0.0992	-0.0314	0.1084	-0.1019	0.0996	-0.0693	0.1071
	MADET	0.0563	0.3862	0.1526	0.0541	0.1388	-0.0485	0.1395	0.1542	0.1482
	GYI	0.6766	0.0114	0.1362	-0.0240	0.1205	-0.0301	0.1364	-0.0117	0.1327
	CP	0.7833	0.0303	0.1513	0.0312	0.1350	0.0269	0.1504	0.0256	0.1427
	MV	0.1651	-0.0434	0.1395	-0.0627	0.1258	-0.0584	0.1466	-0.0489	0.1366
Gamma 5	GTKL	1.4112	0.2669	0.1157	-0.0349	0.0971	-0.1137	0.1352	-0.0764	0.1257
	MADET	0.0282	1.2330	0.1644	0.5818	0.1417	0.2976	0.1573	0.7345	0.1749
	GYI	0.5635	0.0568	0.1297	0.0086	0.1294	-0.0102	0.1380	0.0190	0.1314
	CP	0.8424	0.0219	0.1440	0.0218	0.1371	0.0206	0.1498	0.0187	0.1499
	MV	0.1358	-0.0325	0.1268	-0.0486	0.1267	-0.0513	0.1404	-0.0399	0.1411
Gamma 6	GTKL	3.0462	0.3135	0.1457	0.0813	0.1368	-0.0061	0.1116	-0.0045	0.1219
	MADET	0.0394	1.2899	0.1815	0.6447	0.1727	0.3720	0.1551	0.7573	0.1933
	GYI	0.6753	0.0419	0.1317	-0.0013	0.1142	-0.0161	0.1292	0.0014	0.1237
	CP	0.8392	0.0171	0.1371	0.0198	0.1398	0.0188	0.1500	0.0163	0.1380
	MV	0.1400	-0.0280	0.1258	-0.0489	0.1234	-0.0490	0.1481	-0.0390	0.1307

Stat*: optimal statistics of the measure; RBias: relative bias; NRMSE: normalized root-mean-square error.

Table 5.8. Relative bias and normalized root-mean-square error of the estimated C_1 in the normal distribution.

Sample Size			$n_1= n_2= n_3$ 20		$n_1= n_2= n_3$ 50		$n_1= n_2= n_3$ 100		$n_1=100; n_2=50; n_3=30$	
Setting	Measure	$\hat{C}_1$	Rbias	NRMSE	Rbias	NRMSE	Rbias	NRMSE	Rbias	NRMSE
Normal 1	GTKL	2.5800	-0.1874	0.1479	-0.1284	0.1462	-0.0934	0.1483	-0.2295	0.1600
	MADET	2.4020	0.0078	0.1370	-0.0123	0.1301	-0.0139	0.1156	-0.0163	0.1310
	GYI	3.0000	-0.0030	0.1170	-0.0008	0.1165	-0.0007	0.1222	0.0075	0.1146
	CP	2.5779	-0.0118	0.1266	-0.0132	0.1407	-0.0124	0.1381	-0.0094	0.1351
	MV	2.6140	-0.0139	0.1353	-0.0146	0.1487	-0.0138	0.1401	-0.0114	0.1395
Normal 2	GTKL	4.0826	-0.2705	0.1808	-0.1112	0.1356	-0.0454	0.1225	-0.1757	0.1545
	MADET	3.0097	-0.1329	0.1140	-0.0472	0.1181	-0.0041	0.1145	-0.0849	0.1352
	GYI	3.4155	-0.0091	0.0624	0.0108	0.1268	0.0154	0.1227	0.0174	0.1098
	CP	2.6664	-0.0059	0.1285	-0.0062	0.1342	-0.0035	0.1315	0.0013	0.1333
	MV	2.7468	-0.0120	0.1316	-0.0102	0.1353	-0.0061	0.1357	-0.0031	0.1344
Normal 3	GTKL	2.3579	-0.3278	0.1706	-0.2313	0.1487	-0.1676	0.1525	-0.3383	0.1580
	MADET	1.8024	0.0086	0.1303	0.0107	0.1286	0.0055	0.1384	-0.0663	0.1228
	GYI	2.5000	-0.0698	0.0660	-0.0264	0.0479	-0.0119	0.1458	-0.0122	0.1100
	CP	2.0717	-0.0281	0.1225	-0.0253	0.1134	-0.0203	0.1251	-0.0182	0.1138
	MV	2.0576	-0.0237	0.1268	-0.0238	0.1275	-0.0198	0.1414	-0.0194	0.1243
Normal 4	GTKL	2.2670	-0.1062	0.1487	-0.1237	0.1471	-0.1279	0.1389	-0.1709	0.1347
	MADET	2.3030	0.1034	0.1325	0.0035	0.1232	-0.0343	0.1189	-0.0205	0.1248
	GYI	3.2500	-0.0054	0.1200	-0.0010	0.1349	0.0006	0.1156	0.0060	0.1218
	CP	2.7168	-0.0142	0.1346	-0.0157	0.1349	-0.0134	0.1227	-0.0106	0.1195
	MV	2.7682	-0.0188	0.1431	-0.0190	0.1397	-0.0164	0.1206	-0.0145	0.1190
Normal 5	GTKL	2.2438	-0.0634	0.0546	-0.0902	0.1711	-0.0908	0.1697	-0.1557	0.1528
	MADET	2.3180	0.2409	0.1697	0.1136	0.1433	0.0303	0.1282	0.0171	0.1223
	GYI	3.5000	-0.0091	0.1391	-0.0035	0.1455	-0.0003	0.1275	0.0013	0.1234
	CP	2.8766	-0.0119	0.1169	-0.0136	0.1322	-0.0113	0.1325	-0.0093	0.1185
	MV	2.9644	-0.0189	0.1091	-0.0188	0.1212	-0.0153	0.1307	-0.0143	0.1219
Normal 6	GTKL	1.6298	-0.1255	0.0951	-0.2462	0.2000	-0.2435	0.1953	-0.1137	0.1608
	MADET	1.9740	0.1419	0.1581	0.0116	0.1419	-0.0383	0.1127	-0.0386	0.1072
	GYI	2.7355	-0.0164	0.1189	-0.0155	0.1065	-0.0134	0.1349	-0.0103	0.1276
	CP	2.4664	-0.0232	0.1326	-0.0235	0.1307	-0.0206	0.1482	-0.0184	0.1220
	MV	2.4681	-0.0249	0.1276	-0.0255	0.1309	-0.0226	0.1490	-0.0211	0.1233

RBias: relative bias; NRMSE: normalized root-mean-square error.

Table 5.9. Relative bias and normalized root-mean-square error of the estimated c_2 in the normal distribution.

Sample size			$n_1= n_2= n_3$ 20		$n_1= n_2= n_3$ 50		$n_1= n_2= n_3$ 100		n_1=100; n_2=50; n_3=30	
Setting	Measure	$\hat{c}_2$	Rbias	NRMSE	Rbias	NRMSE	Rbias	NRMSE	Rbias	NRMSE
Normal 1	GTKL	5.4200	0.0875	0.1482	0.0678	0.1518	0.0502	0.1480	0.0210	0.1445
	MADET	5.5980	-0.0015	0.1101	0.0049	0.1222	0.0081	0.1218	0.0017	0.1190
	GYI	5.0000	0.0039	0.0599	0.0019	0.1341	0.0007	0.1263	0.0049	0.1258
	CP	5.4221	0.0072	0.1172	0.0077	0.1282	0.0064	0.1350	0.0096	0.1265
	MV	5.3860	0.0085	0.1253	0.0088	0.1331	0.0072	0.1379	0.0101	0.1297
Normal 2	GTKL	7.5285	0.0239	0.1320	0.0578	0.1470	0.0645	0.1586	0.0527	0.1467
	MADET	6.9484	-0.0625	0.1322	-0.0099	0.1363	0.0127	0.1070	-0.0354	0.1436
	GYI	6.0512	0.0101	0.0541	0.0131	0.1271	0.0142	0.1211	0.0144	0.1253
	CP	6.0435	0.0186	0.1185	0.0178	0.1401	0.0162	0.1305	0.0208	0.1142
	MV	6.0724	0.0151	0.1186	0.0158	0.1370	0.0150	0.1334	0.0176	0.1421
Normal 3	GTKL	5.0232	0.0607	0.1614	0.0472	0.1509	0.0326	0.1600	0.0094	0.1658
	MADET	4.8421	-0.0320	0.1142	-0.0226	0.1292	-0.0145	0.1219	-0.0484	0.1177
	GYI	4.0000	0.0193	0.0909	0.0079	0.1404	0.0036	0.1173	0.0086	0.1204
	CP	4.5491	0.0069	0.1301	0.0081	0.1293	0.0065	0.1376	0.0091	0.1038
	MV	4.5219	0.0098	0.1405	0.0104	0.1331	0.0082	0.1393	0.0103	0.1338
Normal 4	GTKL	4.6715	0.1310	0.1583	0.0612	0.1321	0.0271	0.1317	0.0229	0.1202
	MADET	5.1744	0.0510	0.1300	0.0133	0.1163	-0.0013	0.0901	0.0029	0.1297
	GYI	4.8348	0.0007	0.0404	-0.0079	0.1270	-0.0081	0.1180	-0.0034	0.1272
	CP	5.4573	0.0028	0.1054	0.0031	0.1340	0.0025	0.1243	0.0055	0.1126
	MV	5.4014	0.0048	0.1274	0.0044	0.1363	0.0035	0.1280	0.0069	0.1301
Normal 5	GTKL	4.7606	0.1619	0.1679	0.1069	0.1743	0.0704	0.1709	0.0449	0.1550
	MADET	5.1169	0.1200	0.1612	0.0618	0.1364	0.0239	0.1240	0.0162	0.1349
	GYI	4.8572	0.0185	0.0532	-0.0061	0.1231	-0.0084	0.1303	-0.0016	0.1348
	CP	5.5927	0.0046	0.1112	0.0027	0.1234	0.0023	0.1345	0.0054	0.1111
	MV	5.5738	0.0056	0.1327	0.0035	0.1351	0.0031	0.1388	0.0063	0.1264
Normal 6	GTKL	3.5300	0.1103	0.1737	-0.0047	0.1032	-0.0279	0.0831	-0.0011	0.1030
	MADET	4.2286	0.0790	0.1573	0.0233	0.1225	-0.0004	0.0910	0.0017	0.0902
	GYI	4.0972	-0.0044	0.0421	-0.0098	0.1078	-0.0101	0.1301	-0.0063	0.1241
	CP	4.5833	0.0007	0.1135	0.0002	0.1069	0.0002	0.1324	0.0016	0.1206
	MV	4.5201	0.0025	0.1179	0.0011	0.1101	0.0006	0.1342	0.0026	0.1297

RBias: relative bias; NRMSE: normalized root-mean-square error.

Table 5.10. Relative bias and normalized root-mean-square error of the estimated c_1 in the gamma distribution.

Sample size			n_1= n_2= n_3 20		n_1= n_2= n_3 50		n_1= n_2= n_3 100		n_1=100; n_2=50; n_3=30	
Setting	Measure	$\hat{c}_1$	Rbias	NRMSE	Rbias	NRMSE	Rbias	NRMSE	Rbias	NRMSE
Gamma 1	GTKL	0.8469	0.5556	0.1627	0.4612	0.1760	0.3733	0.2416	0.3832	0.1080
	MADET	0.9894	0.2027	0.1142	0.1409	0.0950	0.1042	0.1504	0.1293	0.1057
	GYI	1.2247	0.1014	0.1225	0.0798	0.1327	0.0621	0.1380	0.0758	0.1607
	CP	1.0912	0.0574	0.1328	0.0413	0.1354	0.0323	0.1466	0.0394	0.1583
	MV	1.1097	0.0658	0.1414	0.0481	0.1283	0.0375	0.1465	0.0455	0.1531
Gamma 2	GTKL	0.7443	0.7448	0.1184	0.6764	0.2872	0.5934	0.2760	0.5307	0.2475
	MADET	1.1250	0.2243	0.1514	0.1608	0.1662	0.1064	0.1227	0.1375	0.1144
	GYI	1.4422	0.0664	0.1438	0.0521	0.1386	0.0412	0.1454	0.0489	0.1416
	CP	1.2274	0.0412	0.1291	0.0297	0.1296	0.0226	0.1447	0.0290	0.1300
	MV	1.2809	0.0432	0.1289	0.0320	0.1307	0.0250	0.1419	0.0311	0.1309
Gamma 3	GTKL	0.9713	0.4742	0.1504	0.3697	0.1160	0.2986	0.1465	0.3456	0.1180
	MADET	1.0485	0.2504	0.1465	0.1956	0.1316	0.1527	0.1195	0.1745	0.1219
	GYI	1.3656	0.0880	0.0409	0.0768	0.1470	0.0641	0.1560	0.0728	0.1295
	CP	1.2113	0.0549	0.1087	0.0376	0.1369	0.0305	0.1518	0.0372	0.1331
	MV	1.2148	0.0580	0.1422	0.0400	0.1371	0.0325	0.1539	0.0384	0.1319
Gamma 4	GTKL	0.5911	0.8134	0.0827	0.6413	0.0772	0.5121	0.0823	0.5053	0.1862
		0.8584								
	GYI	1.2181	0.0787	0.1417	0.0602	0.1303	0.0484	0.1553	0.0578	0.1433
	CP	1.0619	0.0525	0.1305	0.0328	0.1300	0.0255	0.1366	0.0346	0.1421
	MV	1.0789	0.0515	0.1294	0.0334	0.1301	0.0262	0.1338	0.0345	0.1395
Gamma 5	GTKL	0.6936	0.8355	0.1522	0.6494	0.1854	0.5355	0.1986	0.5692	0.1927
	MADET	1.1264	0.3177	0.1682	0.2794	0.1832	0.2435	0.1830	0.2431	0.1497
	GYI	1.5386	0.0730	0.1200	0.0566	0.1291	0.0476	0.1446	0.0544	0.1390
	CP	1.3505	0.0574	0.1401	0.0356	0.1360	0.0282	0.1435	0.0367	0.1443
	MV	1.3646	0.0518	0.1300	0.0336	0.1274	0.0272	0.1457	0.0337	0.1291
Gamma 6	GTKL	0.7484	0.7252	0.1904	0.6778	0.2571	0.6105	0.2642	0.5510	0.2188
	MADET	1.4041	0.0246	0.1042	0.0232	0.1258	0.0196	0.1226	-0.0168	0.1096
	GYI	1.4865	0.0682	0.1139	0.0513	0.1309	0.0415	0.1462	0.0435	0.1159
	CP	1.2591	0.0396	0.1277	0.0224	0.1156	0.0169	0.1254	0.0209	0.1213
	MV	1.3140	0.0299	0.1337	0.0183	0.1357	0.0152	0.1251	0.0164	0.1291

RBias: relative bias; NRMSE: normalized root-mean-square error.

Table 5.11. Relative bias and normalized root-mean-square error of the estimated c_2 in the gamma distribution.

Sample size			n_1= n_2= n_3 20		n_1= n_2= n_3 50		n_1= n_2= n_3 100		n_1=100; n_2=50; n_3=30	
Setting	Measure	$\hat{c}_2$	Rbias	NRMSE	Rbias	NRMSE	Rbias	NRMSE	Rbias	NRMSE
Gamma 1	GTKL	2.0562	0.5137	0.1177	0.4520	0.2224	0.3531	0.2271	0.3441	0.0923
	MADET	2.4558	0.1214	0.1044	0.0911	0.0883	0.0695	0.1247	0.0828	0.1081
	GYI	2.2361	0.0745	0.0930	0.0495	0.1225	0.0401	0.1190	0.0603	0.1070
	CP	2.4654	0.0435	0.1353	0.0340	0.1302	0.0278	0.1464	0.0377	0.1405
	MV	2.4417	0.0509	0.1388	0.0399	0.1407	0.0325	0.1498	0.0432	0.1551
Gamma 2	GTKL	1.9101	0.6378	0.1298	0.6160	0.2276	0.5472	0.2453	0.3922	0.2017
	MADET	2.6847	0.1562	0.1496	0.1189	0.1329	0.0868	0.1254	0.0900	0.1406
	GYI	2.5000	0.1009	0.0634	0.0670	0.0648	0.0496	0.0449	0.0736	0.0604
	CP	2.6840	0.0438	0.0955	0.0339	0.0814	0.0265	0.1187	0.0363	0.0970
	MV	2.6954	0.0451	0.1300	0.0356	0.1343	0.0282	0.1262	0.0377	0.1061
Gamma 3	GTKL	2.1222	0.5241	0.1183	0.4272	0.1236	0.3291	0.2248	0.3626	0.1015
	MADET	2.4288	0.1350	0.1389	0.1101	0.1270	0.0957	0.1345	0.0980	0.1008
	GYI	2.1189	0.1039	0.0562	0.0645	0.1255	0.0483	0.1239	0.0740	0.0874
	CP	2.3980	0.0465	0.0970	0.0347	0.1208	0.0303	0.1515	0.0381	0.1192
	MV	2.3910	0.0501	0.1144	0.0382	0.1491	0.0333	0.1529	0.0406	0.1347
Gamma 4	GTKL	1.4346	0.8818	0.2069	0.7042	0.2658	0.5216	0.2129	0.4350	0.2047
	MADET	1.9553	0.2887	0.1636	0.2283	0.1602	0.1676	0.1465	0.1881	0.1601
	GYI	1.8279	0.2192	0.0853	0.1165	0.0743	0.0638	0.0574	0.1208	0.0678
	CP	2.1161	0.0462	0.0980	0.0314	0.1068	0.0251	0.1445	0.0361	0.0980
	MV	2.1183	0.0461	0.1197	0.0339	0.1330	0.0274	0.1584	0.0377	0.1056
Gamma 5	GTKL	1.6936	0.9166	0.2489	0.7197	0.2538	0.5567	0.2270	0.5320	0.1898
	MADET	2.6095	0.2002	0.1364	0.1781	0.1385	0.1491	0.1467	0.1401	0.1351
	GYI	2.3783	0.2644	0.1206	0.1561	0.0951	0.1037	0.0751	0.1443	0.0734
	CP	2.6189	0.0580	0.0461	0.0382	0.1096	0.0298	0.1262	0.0406	0.1019
	MV	2.6275	0.0486	0.0988	0.0361	0.1409	0.0291	0.1394	0.0379	0.1401
Gamma 6	GTKL	1.9524	0.5832	0.1400	0.5954	0.2342	0.5518	0.3045	0.4049	0.2656
	MADET	3.1856	-0.0194	0.1068	-0.0149	0.1205	-0.0092	0.1252	-0.0559	0.1358
	GYI	1.4865	1.1745	0.1601	0.8959	0.1873	0.6505	0.1594	0.6922	0.1514
	CP	2.5423	0.0594	0.0386	0.0334	0.1022	0.0246	0.1102	0.0362	0.0307
	MV	2.5760	0.0420	0.1232	0.0300	0.1291	0.0241	0.1413	0.0321	0.1242

RBias: relative bias; NRMSE: normalized root-mean-square error.

Table 5.12. Summary of the correct classification rates and total correct classification rate in the normal distribution.

Sample size		n_1= n_2= n_3 20				n_1= n_2= n_3 50				n_1= n_2= n_3 100				n_1=100; n_2=50; n_3=30			
Setting	Measure	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR
Normal 1	GTKL	0.5407	0.5493	0.5424	1.6323	0.5582	0.5446	0.5525	1.6553	0.5701	0.5399	0.5648	1.6748	0.5070	0.5404	0.6022	1.6497
	MADET	0.5919	0.5401	0.5895	1.7215	0.5774	0.5545	0.5784	1.7103	0.5741	0.5623	0.5721	1.7085	0.5725	0.5518	0.5857	1.7100
	GYI	0.6903	0.3845	0.6878	1.7626	0.6853	0.3738	0.6845	1.7436	0.6847	0.3719	0.6846	1.7412	0.6882	0.3718	0.6849	1.7448
	CP	0.6098	0.5115	0.6079	1.7292	0.6044	0.5119	0.6036	1.7198	0.6041	0.5132	0.6037	1.7209	0.6056	0.5115	0.6035	1.7205
	MV	0.6159	0.5045	0.6137	1.7341	0.6105	0.5035	0.6095	1.7235	0.6102	0.5042	0.6097	1.7240	0.6113	0.5029	0.6102	1.7243
Normal 2	GTKL	0.6641	0.4939	0.3515	1.5094	0.7528	0.4512	0.3197	1.5237	0.7954	0.4289	0.3089	1.5332	0.7136	0.4673	0.3295	1.5104
	MADET	0.6191	0.5178	0.4622	1.5990	0.6609	0.5204	0.4173	1.5986	0.6828	0.5231	0.3956	1.6015	0.6390	0.5109	0.4397	1.5896
	GYI	0.7479	0.3922	0.5109	1.6510	0.7550	0.3774	0.4967	1.6290	0.7582	0.3759	0.4909	1.6249	0.7579	0.3737	0.5010	1.6326
	CP	0.6282	0.4926	0.4945	1.6152	0.6238	0.4888	0.4886	1.6012	0.6240	0.4893	0.4874	1.6006	0.6264	0.4889	0.4890	1.6043
	MV	0.6399	0.4860	0.4929	1.6188	0.6364	0.4820	0.4866	1.6049	0.6372	0.4822	0.4849	1.6043	0.6387	0.4814	0.4876	1.6076
Normal 3	GTKL	0.4551	0.5177	0.4787	1.4515	0.4839	0.5046	0.4775	1.4661	0.5038	0.4965	0.4824	1.4827	0.4355	0.5117	0.5192	1.4664
	MADET	0.4837	0.4862	0.5687	1.5386	0.4819	0.4925	0.5560	1.5304	0.4770	0.5044	0.5475	1.5289	0.4528	0.4948	0.5804	1.5280
	GYI	0.5909	0.3328	0.6802	1.6039	0.5936	0.3025	0.6802	1.5764	0.5945	0.2923	0.6829	1.5696	0.5934	0.2976	0.6829	1.5738
	CP	0.5160	0.4586	0.5881	1.5627	0.5103	0.4547	0.5808	1.5459	0.5087	0.4544	0.5815	1.5446	0.5094	0.4522	0.5817	1.5433
	MV	0.5141	0.4574	0.5908	1.5624	0.5081	0.4542	0.5841	1.5464	0.5062	0.4538	0.5853	1.5452	0.5062	0.4520	0.5860	1.5442
Normal 4	GTKL	0.5249	0.4688	0.6585	1.6522	0.5096	0.4372	0.7272	1.6740	0.5021	0.4216	0.7638	1.6875	0.4849	0.4241	0.7723	1.6812
	MADET	0.6070	0.4873	0.6417	1.7360	0.5653	0.4847	0.6855	1.7355	0.5467	0.4852	0.7022	1.7342	0.5523	0.4821	0.7017	1.7361
	GYI	0.7260	0.3085	0.7711	1.8056	0.7264	0.2890	0.7773	1.7928	0.7261	0.2851	0.7781	1.7893	0.7283	0.2892	0.7752	1.7927
	CP	0.6318	0.4852	0.6374	1.7544	0.6292	0.4847	0.6340	1.7479	0.6285	0.4853	0.6334	1.7472	0.6300	0.4844	0.6332	1.7475
	MV	0.6387	0.4747	0.6487	1.7621	0.6368	0.4720	0.6459	1.7547	0.6363	0.4717	0.6456	1.7536	0.6372	0.4719	0.6454	1.7545

$p_{1,1}$, $p_{2,2}$ and $p_{3,3}$: the correct classification rate for stage 1, 2 and 3, respectively; TCCR: total correct classification rates.

Table 5.12. Summary of the correct classification rates and total correct classification rate in the normal distribution (continued).

Sample size		$n_1= n_2= n_3$ 20				$n_1= n_2= n_3$ 50				$n_1= n_2= n_3$ 100				n_1=100; n_2=50; n_3=30			
Setting	Measure	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR
Normal 5	GTKL	0.5395	0.4512	0.5977	1.5883	0.5211	0.4265	0.6538	1.6014	0.5167	0.4092	0.6963	1.6222	0.4895	0.3974	0.7251	1.6119
	MADET	0.6569	0.4651	0.5735	1.6955	0.6059	0.4459	0.6395	1.6913	0.5731	0.4349	0.6840	1.6920	0.5680	0.4340	0.6952	1.6972
	GYI	0.7586	0.2795	0.7513	1.7894	0.7608	0.2437	0.7695	1.7739	0.7635	0.2362	0.7736	1.7732	0.7634	0.2466	0.7665	1.7765
	CP	0.6586	0.4659	0.6050	1.7295	0.6560	0.4618	0.6018	1.7196	0.6576	0.4625	0.6016	1.7217	0.6579	0.4640	0.6004	1.7223
	MV	0.6699	0.4576	0.6089	1.7363	0.6683	0.4521	0.6057	1.7261	0.6706	0.4519	0.6054	1.7279	0.6703	0.4542	0.6043	1.7288
Normal 6	GTKL	0.4198	0.3716	0.7821	1.5735	0.3811	0.3138	0.8896	1.5845	0.3762	0.3031	0.9187	1.5981	0.4108	0.3091	0.9018	1.6217
	MADET	0.5545	0.4870	0.6563	1.6978	0.5042	0.4714	0.7332	1.7088	0.4842	0.4659	0.7676	1.7177	0.4840	0.4708	0.7662	1.7210
	GYI	0.6376	0.3333	0.8078	1.7786	0.6336	0.3175	0.8142	1.7653	0.6334	0.3145	0.8159	1.7637	0.6335	0.3208	0.8113	1.7656
	CP	0.5832	0.4894	0.6586	1.7312	0.5801	0.4873	0.6564	1.7238	0.5805	0.4883	0.6560	1.7247	0.5806	0.4882	0.6559	1.7247
	MV	0.5830	0.4788	0.6779	1.7397	0.5797	0.4748	0.6770	1.7316	0.5800	0.4749	0.6772	1.7321	0.5798	0.4762	0.6766	1.7325

$p_{1,1}$, $p_{2,2}$ and $p_{3,3}$: the correct classification rate for stage 1, 2 and 3, respectively; TCCR: total correct classification rates.

Table 5.13. Summary of the correct classification rates and total correct classification rate in the gamma distribution.

Sample size		$n_1= n_2= n_3$ 20				$n_1= n_2= n_3$ 50				$n_1= n_2= n_3$ 100				n_1=100; n_2=50; n_3=30			
Setting	Measure	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR
Gamma 1	GTKL	0.7116	0.5267	0.4638	1.7021	0.6879	0.5463	0.4876	1.7217	0.6575	0.5405	0.5411	1.7391	0.6560	0.5122	0.5653	1.7335
	MADET	0.6718	0.5555	0.5611	1.7884	0.6460	0.5603	0.5707	1.7771	0.6295	0.5658	0.5814	1.7767	0.6406	0.5583	0.5824	1.7813
	GYI	0.7420	0.4152	0.6667	1.8238	0.7313	0.4017	0.6726	1.8057	0.7229	0.4043	0.6768	1.8040	0.7292	0.4089	0.6701	1.8083
	CP	0.6579	0.5286	0.6051	1.7916	0.6502	0.5278	0.6042	1.7821	0.6462	0.5311	0.6071	1.7844	0.6505	0.5300	0.6049	1.7854
	MV	0.6721	0.5187	0.6079	1.7987	0.6632	0.5163	0.6079	1.7873	0.6582	0.5189	0.6116	1.7888	0.6627	0.5185	0.6092	1.7904
Gamma 2	GTKL	0.7123	0.5248	0.4487	1.6858	0.6910	0.5437	0.4518	1.6865	0.6660	0.5304	0.4881	1.6845	0.6437	0.4531	0.5780	1.6748
	MADET	0.7432	0.5438	0.4626	1.7495	0.7157	0.5434	0.4836	1.7426	0.6952	0.5450	0.5010	1.7411	0.7059	0.5292	0.5116	1.7466
	GYI	0.8076	0.4096	0.5850	1.8021	0.7977	0.3967	0.5845	1.7789	0.7939	0.3928	0.5889	1.7757	0.7970	0.4013	0.5871	1.7854
	CP	0.7135	0.5052	0.5352	1.7538	0.7059	0.5042	0.5330	1.7431	0.7043	0.5056	0.5361	1.7460	0.7075	0.5068	0.5346	1.7489
	MV	0.7372	0.4962	0.5298	1.7632	0.7294	0.4949	0.5279	1.7521	0.7277	0.4959	0.5310	1.7546	0.7304	0.4972	0.5298	1.7574
Gamma 3	GTKL	0.6324	0.4945	0.3762	1.5031	0.6011	0.5060	0.4200	1.5271	0.5743	0.5038	0.4676	1.5456	0.5870	0.4777	0.4716	1.5363
	MADET	0.5945	0.5034	0.4845	1.5824	0.5706	0.5127	0.4934	1.5767	0.5515	0.5250	0.4979	1.5744	0.5603	0.5103	0.5086	1.5792
	GYI	0.6785	0.3412	0.6213	1.6409	0.6703	0.3162	0.6309	1.6174	0.6643	0.3084	0.6386	1.6113	0.6685	0.3203	0.6303	1.6191
	CP	0.5859	0.4679	0.5498	1.6036	0.5758	0.4666	0.5483	1.5907	0.5727	0.4671	0.5491	1.5889	0.5770	0.4670	0.5486	1.5926
	MV	0.5899	0.4660	0.5489	1.6048	0.5791	0.4647	0.5480	1.5918	0.5757	0.4650	0.5492	1.5899	0.5795	0.4650	0.5491	1.5936
Gamma 4	GTKL	0.6110	0.5408	0.3827	1.5345	0.5600	0.5095	0.4664	1.5359	0.5184	0.4618	0.5582	1.5383	0.5155	0.4250	0.6072	1.5477
	MADET	0.6650	0.5166	0.4248	1.6063	0.6289	0.5088	0.4649	1.6026	0.5937	0.5019	0.5062	1.6017	0.6142	0.4989	0.5011	1.6143
	GYI	0.7286	0.3610	0.5947	1.6843	0.7191	0.3201	0.6212	1.6604	0.7137	0.2990	0.6436	1.6563	0.7180	0.3260	0.6247	1.6688
	CP	0.6396	0.4662	0.5313	1.6371	0.6307	0.4626	0.5308	1.6242	0.6281	0.4643	0.5334	1.6258	0.6328	0.4665	0.5330	1.6322
	MV	0.6484	0.4618	0.5283	1.6384	0.6399	0.4594	0.5273	1.6266	0.6372	0.4608	0.5303	1.6283	0.6413	0.4628	0.5302	1.6343

$p_{1,1}$, $p_{2,2}$ and $p_{3,3}$: the correct classification rate for stage 1, 2 and 3, respectively; TCCR: total correct classification rates.

Table 5.13. Summary of the correct classification rates and total correct classification rate in the gamma distribution (continued).

Sample size		$n_1= n_2= n_3$ 20				$n_1= n_2= n_3$ 50				$n_1= n_2= n_3$ 100				n_1=100; n_2=50; n_3=30			
Setting	Measure	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR	$p_{1,1}$	$p_{2,2}$	$p_{3,3}$	TCCR
Gamma 5	GTKL	0.5520	0.5151	0.3796	1.4467	0.4991	0.4780	0.4648	1.4419	0.4629	0.4407	0.5311	1.4348	0.4722	0.4229	0.5449	1.4399
	MADET	0.6324	0.5091	0.3762	1.5176	0.6165	0.5099	0.3918	1.5181	0.6006	0.5078	0.4086	1.5170	0.6005	0.4970	0.4228	1.5204
	GYI	0.6994	0.3906	0.5056	1.5956	0.6902	0.3571	0.5210	1.5684	0.6857	0.3402	0.5319	1.5578	0.6882	0.3507	0.5354	1.5742
	CP	0.6178	0.4582	0.4711	1.5472	0.6079	0.4523	0.4728	1.5330	0.6044	0.4518	0.4735	1.5297	0.6084	0.4534	0.4761	1.5379
	MV	0.6211	0.4523	0.4697	1.5431	0.6127	0.4500	0.4700	1.5327	0.6095	0.4503	0.4705	1.5303	0.6123	0.4511	0.4733	1.5367
Gamma 6	GTKL	0.7076	0.4861	0.3399	1.5336	0.6938	0.5163	0.3271	1.5372	0.6741	0.5084	0.3533	1.5359	0.6516	0.4243	0.4400	1.5158
	MADET	0.7632	0.4764	0.3119	1.5516	0.7613	0.4851	0.3052	1.5515	0.7627	0.4998	0.2963	1.5588	0.7402	0.4623	0.3449	1.5473
	GYI	0.8212	0.3369	0.5456	1.7036	0.8111	0.2534	0.6100	1.6744	0.8080	0.1793	0.6772	1.6645	0.8076	0.2254	0.6432	1.6762
	CP	0.7243	0.4229	0.4618	1.6090	0.7158	0.4097	0.4616	1.5872	0.7149	0.4075	0.4602	1.5825	0.7164	0.4116	0.4655	1.5935
	MV	0.7419	0.4122	0.4501	1.6042	0.7357	0.4071	0.4476	1.5904	0.7360	0.4071	0.4458	1.5889	0.7356	0.4090	0.4505	1.5951

$p_{1,1}$, $p_{2,2}$ and $p_{3,3}$: the correct classification rate for stage 1, 2 and 3, respectively; TCCR: total correct classification rates.

Table 5.14. The percentage of the loss of correct classification rate and the maximum-minimum difference in the normal distribution.

		MMDIF				LCCR%			
Setting	Measure	$n_1= n_2= n_3=20$	$n_1= n_2= n_3=50$	$n_1= n_2= n_3=100$	$n_1=100; n_2=50; n_3=30$	$n_1= n_2= n_3=20$	$n_1= n_2= n_3=50$	$n_1= n_2= n_3=100$	$n_1=100; n_2=50; n_3=30$
Normal 1	GTKL	1.3576	0.9696	0.7355	1.0019	7.3925	5.0642	3.8135	5.4505
	MADET	0.6965	0.4939	0.3741	0.5031	2.3318	1.9098	1.8780	1.9945
	GYI	1.2034	1.0229	0.9651	1.0458	--	--	--	--
	CP	0.3409	0.2600	0.2304	0.2654	1.8949	1.3650	1.1659	1.3927
	MV	0.3920	0.3062	0.2744	0.3127	1.6169	1.1528	0.9878	1.1749
Normal 2	GTKL	2.3690	2.0886	1.9636	2.1804	8.5766	6.4641	5.6434	7.4850
	MADET	0.9487	0.8586	0.8196	0.8575	3.1496	1.8662	1.4401	2.6338
	GYI	1.4461	1.2474	1.1366	1.2928	--	--	--	--
	CP	0.4553	0.3769	0.3439	0.3906	2.1684	1.7066	1.4955	1.7334
	MV	0.4869	0.4128	0.3829	0.4254	1.9503	1.4794	1.2678	1.5313
Normal 3	GTKL	2.0615	1.4264	1.0082	1.6241	9.5018	6.9970	5.5364	6.8242
	MADET	0.9821	0.8272	0.7117	0.8837	4.0713	2.9180	2.5930	2.9102
	GYI	1.8989	1.6645	1.5459	1.6772	--	--	--	--
	CP	0.4678	0.3575	0.3185	0.3575	2.5687	1.9348	1.5928	1.9380
	MV	0.4667	0.3627	0.3264	0.3647	2.5874	1.9031	1.5545	1.8808
Normal 4	GTKL	1.4101	1.2027	1.1142	1.2000	8.4958	6.6265	5.6894	6.2197
	MADET	0.8496	0.6712	0.5777	0.6820	3.8547	3.1961	3.0794	3.1572
	GYI	1.7566	1.7834	1.7842	1.7902	--	--	--	--
	CP	0.4281	0.3730	0.3516	0.3766	2.8356	2.5045	2.3529	2.5213
	MV	0.5030	0.4478	0.4254	0.4516	2.4092	2.1252	1.9952	2.1309
Normal 5	GTKL	1.7188	1.5071	1.3464	1.4001	11.2384	9.7243	8.5157	9.2654
	MADET	1.0642	0.9200	0.8087	0.8965	5.2476	4.6564	4.5793	4.4638
	GYI	2.2004	2.3823	2.4134	2.3642	--	--	--	--
	CP	0.5046	0.4556	0.4359	0.4549	3.3475	3.0611	2.9044	3.0509
	MV	0.5720	0.5241	0.5042	0.5238	2.9675	2.6946	2.5547	2.6851
Normal 6	GTKL	2.6171	2.4842	2.3836	2.3280	11.5315	10.2419	9.3894	8.1502
	MADET	0.9813	0.8575	0.7896	0.8196	4.5429	3.2006	2.6082	2.5261
	GYI	1.6070	1.5942	1.5987	1.5612	--	--	--	--
	CP	0.4290	0.3739	0.3526	0.3686	2.6650	2.3509	2.2113	2.3165
	MV	0.5054	0.4541	0.4354	0.4482	2.1871	1.9090	1.7917	1.8747

MMDIF: maximum minimum difference; LCCR%: percentage of loss of classification rate.

Table 5.15. The percentage of the loss of CCR and the maximum-minimum difference in the gamma distribution.

		MMDIF				LCCR%			
Setting	Measure	n_1= n_2= n_3=20	n_1= n_2= n_3=50	n_1= n_2= n_3=100	n_1=100; n_2=50; n_3=30	n_1= n_2= n_3=20	n_1= n_2= n_3=50	n_1= n_2= n_3=100	n_1=100; n_2=50; n_3=30
Gamma 1	GTKL	1.3203	0.9868	0.7616	0.8787	6.6729	4.6519	3.5976	3.5976
	MADET	0.6407	0.4336	0.3127	0.4124	1.9410	1.5839	1.5133	1.5133
	GYI	1.0568	0.9221	0.8453	0.8927	--	--	--	--
	CP	0.3379	0.2693	0.2370	0.2677	1.7655	1.3070	1.0865	1.0865
	MV	0.4040	0.3289	0.2932	0.3261	1.3762	1.0190	0.8426	0.8426
Gamma 2	GTKL	1.4465	1.2991	1.1949	1.1519	6.4536	5.1942	5.1360	5.1360
	MADET	1.0540	0.8632	0.7022	0.8464	2.9188	2.0406	1.9485	1.9485
	GYI	1.6734	1.4278	1.2616	1.4347	--	--	--	--
	CP	0.5404	0.4668	0.4312	0.4759	2.6802	2.0125	1.6726	1.6726
	MV	0.6139	0.5414	0.5064	0.5499	2.1586	1.5065	1.1883	1.1883
Gamma 3	GTKL	2.0230	1.3070	0.9050	1.1489	8.3978	5.5830	4.0775	4.0775
	MADET	0.9242	0.6774	0.5265	0.6487	3.5651	2.5164	2.2901	2.2901
	GYI	1.6047	1.3730	1.3021	1.3528	--	--	--	--
	CP	0.4195	0.3128	0.2709	0.3161	2.2731	1.6508	1.3902	1.3902
	MV	0.4349	0.3291	0.2869	0.3314	2.2000	1.5828	1.3281	1.3281
Gamma 4	GTKL	2.1656	1.7123	1.3910	1.2427	8.8939	7.4982	7.1243	7.1243
	MADET	1.1504	0.9391	0.7576	0.8605	4.6310	3.4811	3.2965	3.2965
	GYI	1.9853	1.8221	1.6868	1.7389	--	--	--	--
	CP	0.5124	0.4173	0.3762	0.4212	2.8024	2.1802	1.8415	1.8415
	MV	0.5322	0.4445	0.4059	0.4477	2.7252	2.0357	1.6905	1.6905
Gamma 5	GTKL	2.4481	1.7875	1.4180	1.3408	9.3319	8.0655	7.8958	7.8958
	MADET	1.2487	1.0922	0.9969	1.0095	4.8884	3.2071	2.6191	2.6191
	GYI	2.3452	2.0353	1.8339	2.0148	--	--	--	--
	CP	0.5970	0.4650	0.4104	0.4732	3.0333	2.2571	1.8038	1.8038
	MV	0.5676	0.4589	0.4138	0.4638	3.2903	2.2762	1.7653	1.7653
Gamma 6	GTKL	2.3027	2.2420	2.0106	1.8328	9.9789	8.1940	7.7260	7.7260
	MADET	2.0895	2.1045	2.1172	2.0502	8.9223	7.3399	6.3503	6.3503
	GYI	5.7566	7.6860	10.7359	6.5849	--	--	--	--
	CP	1.0641	0.9235	0.8565	0.9384	5.5529	5.2078	4.9264	4.9264
	MV	1.0186	0.9234	0.8785	0.9325	5.8347	5.0167	4.5419	4.5419

MMDIF: maximum minimum difference; LCCR%: percentage of loss of classification rate.

5.3. Summary

The simulation shows that the measures have strengths in different distributions. Also, they are beneficial for different study purposes. The GYI performs the best when the study does not target the middle stage. On the other hand, the GTKL and MADET performs better when the study focuses on identifying subjects in the middle stage. The CP and MV are more balanced compared to the other three measures as they have relatively smaller values than the others, according to the results of MMDIF and LCCR% in most settings. As shown in the power simulation under the Scenario I, the GTKL performs well when the group variance is diverse. In the simulation of optimal cut-points selection, the GTKL has higher CCR in the middle stage when the sample size is small (i.e., 20 and 50). A smaller sample size commonly produces a more significant variance. Thus, the GTKL has higher CCR in the middle stage in some settings. The exciting finding gives a clue for future studies about estimation of cut-points, primarily when the research focuses on detecting subjects in the middle stage.

The choice of diagnostic accuracy measures does not solely depend on the power of the tests but also depends on the target population and the goal of diagnosis. In other words, when a diagnostic test is designed to distinguish subjects in the first and last stages, GYI would give the highest correct classification rates for this purpose. If a diagnostic test is designed to separate subjects between the first and second stages, GTKL and MADET would be good choices as their correct classification rates are higher in the first two stages. VUS and its corresponding cut-point selection criteria (i.e., CP and MV) give similar correct classification rates across all stages; however, they may not be appropriate when a diagnostic test is targeting a specific stage. They can serve as validation measures and provide conventional correct classification rates when there is a strong difference among the distributions of all stages.

The simulation has been conducted in some other distributions that are not shown in the tables shown above since our research scope is to generalize Kullback-Leibler divergence to medical diagnosis. The results indicate that all measures have different kinds of characteristics that have been mentioned above. GTKL has the highest power when the distributions of stages strongly overlap and higher correct classification rates in the first and second stages, although its total correct classification rate is lower than others. Providing highest total correct classification rate, especially the high rates of the first and last stages, GYI is a robust measure when a diagnostic measure targets the population in the healthy and diseased population rather than the population in the transition stage. MADET is slightly similar to GTKL which provides higher power when the stages' distributions overlap. VUS, CP and MV are more balanced measures which do not give extreme results but generally similar results among all disease stages. Overall, VUS, CP and MV are suggested to evaluate a diagnostic test in the preliminary stage of the research and provides conventional results. GTKL, GYI and MADET can be used to refine the accuracy of a diagnostic test since they provide sharper evaluations for diagnosis in different stages and distributions.

In conclusion, when choosing a diagnostic accuracy measure, the distribution of the data gives an idea of which diagnostic accuracy measure would have the highest power. In addition, the target population is critical in choosing the measure based on the stage of diagnosis in which a test focuses.

CHAPTER 6
REAL DATA ANALYSIS

In this chapter, a dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI) is used to demonstrate the application of the generalized total Kullback-Leibler (GTKL) divergence, along with the comparison of existing measures including the generalized Youden index (GYI), volume under the surface (VUS), closest-to-perfection (CP), maximum volume (MV), and maximum absolute determinant (MADET).

6.1. Introduction of Alzheimer's Disease and Dementia

Alzheimer's disease is the most common form among the diseases related to dementia that can become worse over time. According to the Centers for Disease Control and Prevention, dementia is a prevalent brain impairment among adults at least 65 years of age, with 5.0 million adults suffering from dementia in 2014, and the prevalence of dementia is projected to be nearly 14 million by 2060 in the United States (CDC, 2019). Dementia is not a specific disease yet a general term of brain impairment with various kinds of symptoms including losing the ability to remember, think, and make decisions that affect daily life (CDC, 2019). ADNI provides data that tracks the progression of Alzheimer's disease over time to study this irreversible neurodegenerative disease using biomarkers and clinical measures (ADNI, 2017).

6.2. Data analysis

6.2.1 Data file from ADNI

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD).

The data file used in this study identifies the subjects into three clinical stages as the cognitive normal (CN), mild cognitive impairment (MCI), and dementia at the baseline. The staging of Alzheimer’s disease is based on the global clinical dementia rating (CDGLOBAL) with 0, 0.5, and greater than 1 indicating non-diseased, early diseased and fully diseased, respectively.

The data file provides test results from five potential biomarkers for Alzheimer’s disease including three biomarkers from core cerebrospinal fluid (CSF) and two biomarkers from magnetic resonance imaging (MRI). Blennow, Hampel, Weiner, and Zetterberg (2010) summarized the importance of early detection of Alzheimer’s disease and the clinical treatments related to potential CSF biomarkers based on previous studies (Das, Murphy, Younkin, Younkin, & Golde, 2001; Garcia-Alloza et al., 2009; Levites et al., 2006). The data file includes results of total tau (TAU), phosphorylated tau (PTAU), and the 42 amino acid form of amyloid-β (Abeta). These three CSF biomarkers reflect the pathology of Alzheimer’s disease and have been proposed as candidate markers for prediction of cognition decline as the progression indicator of dementia. Previous studies also discussed the other two potential biomarkers of Alzheimer’s disease measured from MRI: rate of volume change of hippocampus and whole-brain. Those studies discovered the potential relationship between the volume change and the initiative of Alzheimer’s disease, relating to the injury and death of neurons (Duthey, 2013; Grundman & Delaney, 2002; Shaffer et al., 2013).

6.2.2 Statistics obtained in the analysis

Data analysis is conducted based on the proposed measure and the five exiting measures using biomarkers, including Abeta, TAU, PTAU, rate of volume change of hippocampus, and whole-brain. Overall, we compute the optimal statistics of the GTKL, GYI, VUS, MADET, CP, and MV with their optimal cut-points. Additionally, the generalized PPV and NPV proposed by Samawi (2020) are calculated to assess the performance of optimal cut-point selection based on the GTKL, GYI, MADET, CP, and MV, as follows:

Generalized predictive values:

$$PPV_1 = \frac{p_{22} \times p_2}{p_{12} \times p_1 + p_{22} \times p_2 + p_{32} \times p_3} \ ,$$

$$PPV_2 = \frac{p_{33} \times p_3}{p_{13} \times p_1 + p_{23} \times p_2 + p_{33} \times p_3},$$

$$NPV = \frac{p_{11} \times p_1}{p_{11} \times p_1 + p_{21} \times p_2 + p_{31} \times p_3} \ ,$$

where, p_1, p_2 and p_3 are the prevalence of stages 1 (CN), stage 2 (MCI), and stage 3 (dementia), respectively.

Using the gold standard in the dataset (CDGLOBAL), the prevalence of Alzheimer's disease in different stages is approximated as 0.72 (p_1), 0.2 (p_2) and 0.08 (p_3) for stages 1, 2 and 3, respectively according to the estimation based on the estimations from Kantarci et al. (2009); Mitchell and Shiri‐Feshki (2009); (Roberts & Knopman, 2013).

6.2.3 Results

The dataset consists of 415 subjects with 114, 256, and 45 subjects for the CN, MCI, and dementia groups, respectively. Due to missing values, the actual sample sizes for each biomarker vary and are smaller than the group sizes. Table 6.1 provides a summary of the descriptive statistics of the biomarkers in the ADNI dataset.

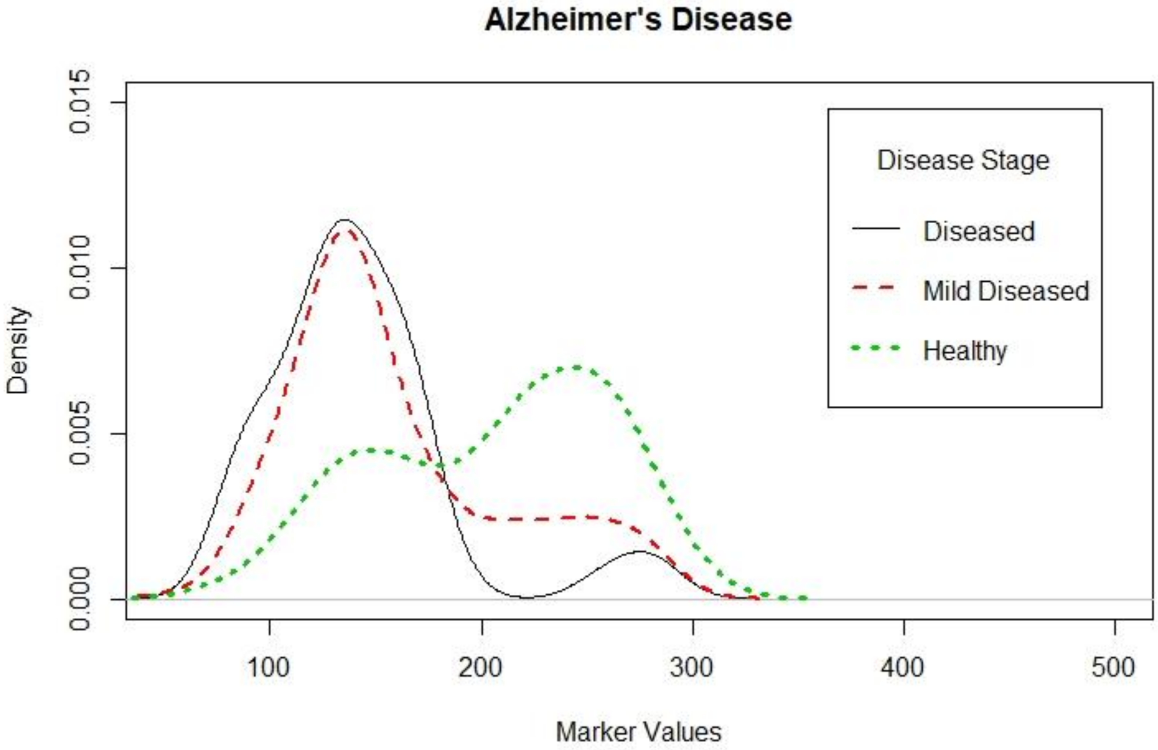
Table 6.1. Summary of descriptive statistics of five biomarkers of Alzheimer's disease.

	Biomarker	**Stage 1**			**Stage 2**			**Stage 3**		
		N_1	**$Mean_1$**	**SD_1**	**N_2**	**$Mean_2$**	**SD_2**	**N_3**	**$Mean_3$**	**SD_3**
Cerebrospinal Fluid	Abeta	114	205.59	55.09	255	159.27	52.19	45	141.67	44.38
	TAU	114	69.68	30.37	252	108.24	59.43	43	118.12	65.30
	PTAU	114	24.86	14.58	256	37.50	18.85	45	38.11	18.94
Brain Imaging	Hippocampus	100	7265.68	826.03	197	6298.47	1109.42	34	5301.94	929.43
	Whole-Brain	112	1,004,949	104,189	252	999,015	112,557	44	943,771	116,078

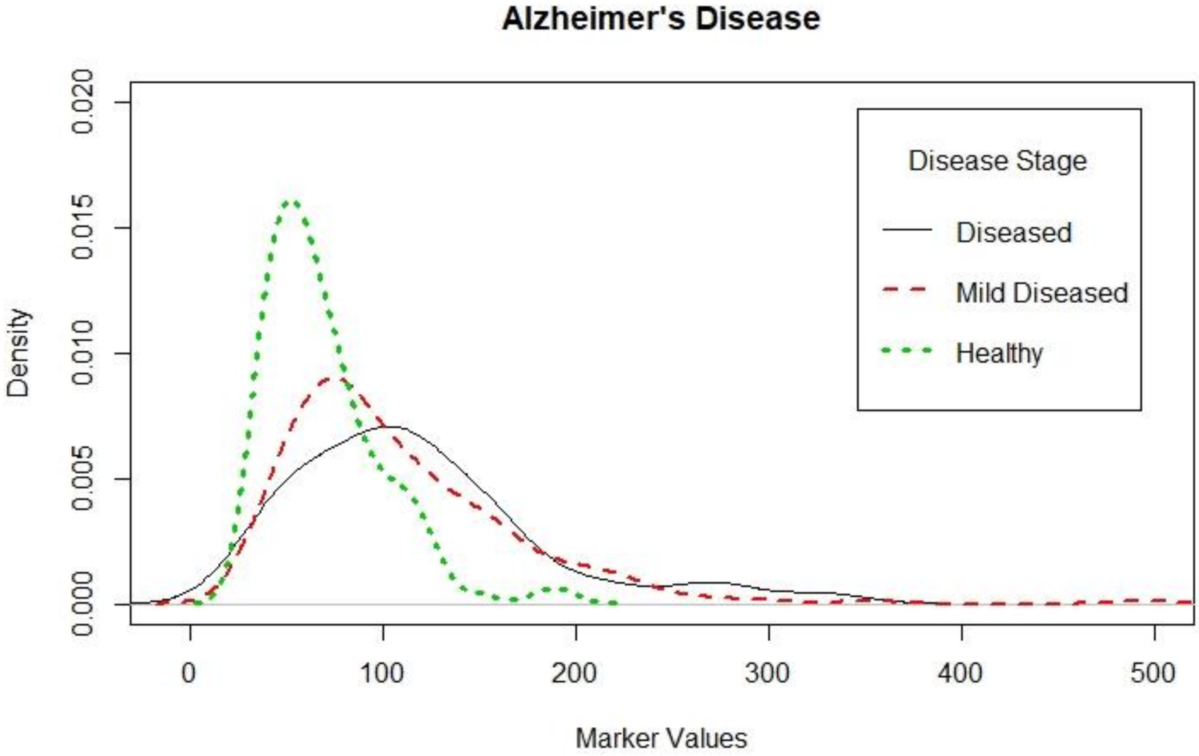
The distributions of the biomarkers in the three clinical stages are shown in density plots displayed in Figure 6.1. The optimal statistics of GTKL, GYI, VUS, and MADET are calculated and showed in Table 6.2, and the corresponding optimal cut-points are shown in Table 6.3. As mentioned above, the VUS is not directly used in optimal cut-point selection; instead, the optimal cutpoints estimated based on the measures derived from the VUS, the CP, and the MV are calculated (see Table 6.3).

Figure 6.1. Density plots of the distribution of biomarkers in different clinical stages.

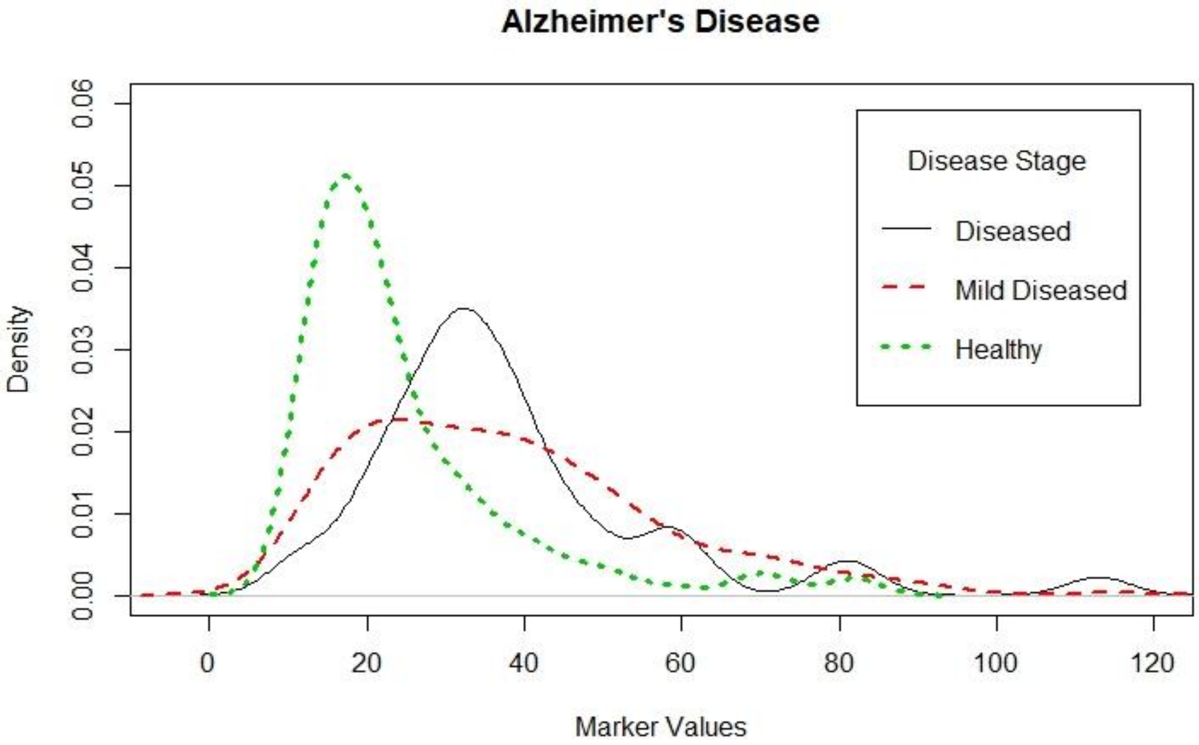
a) Abeta



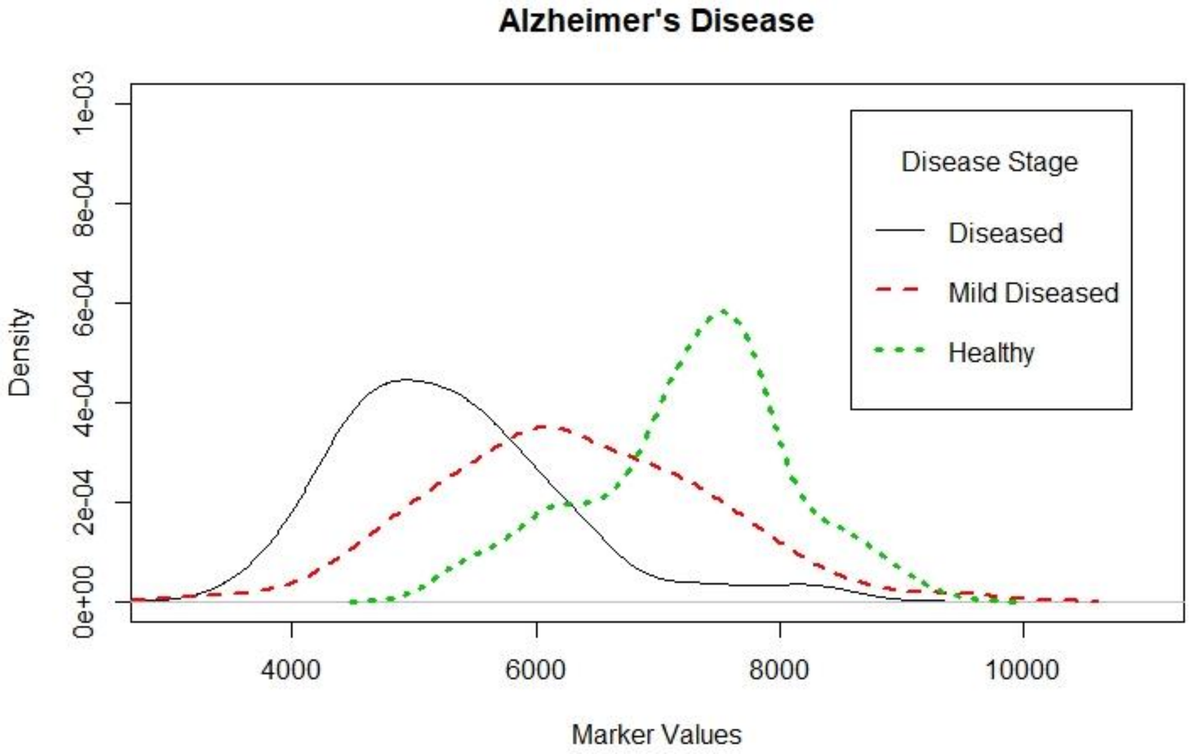
b) TAU



c) PTAU



d) Hippocampus



e) Whole-Brain

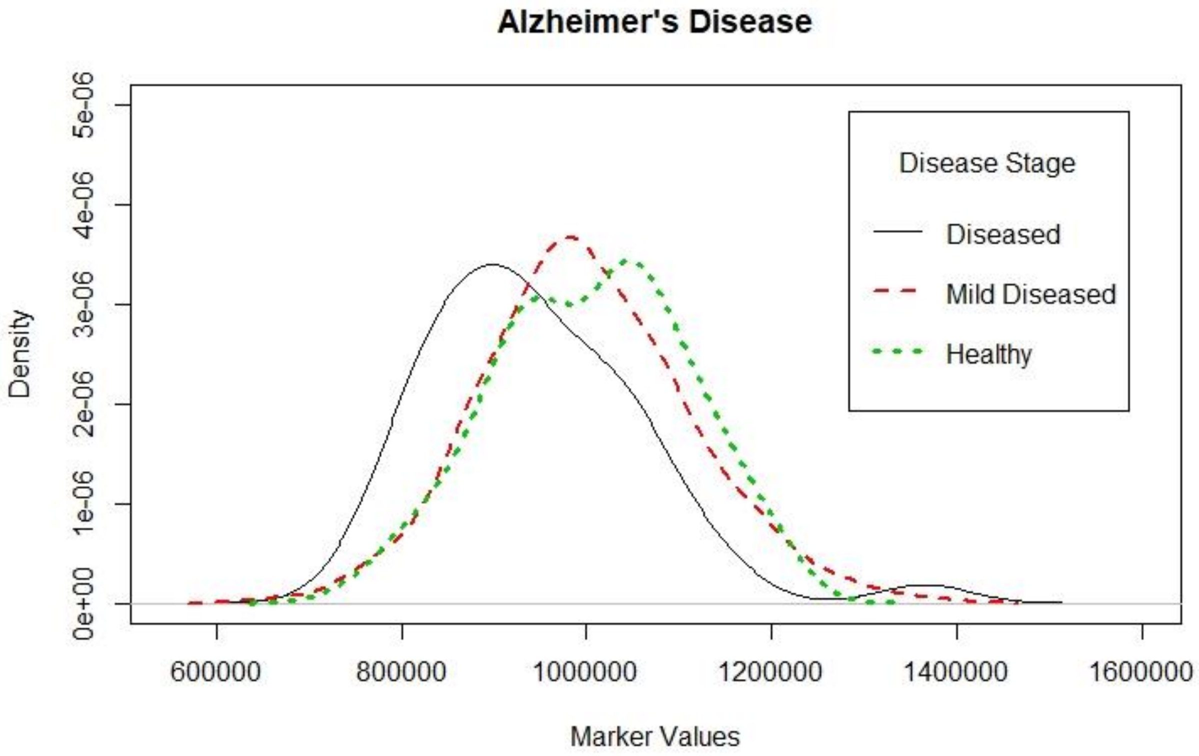


Table 6.2 The estimated optimal statistics of diagnostic accuracy measures.

Measure	Cerebrospinal fluid			Brain imaging	
	Abeta	TAU	PTAU	Hippocampus	Whole-Brain
GTKL	0.8801	0.7497	0.7676	1.7714	0.2464
GYI	0.5453	0.4193	0.4566	0.8052	0.2776
MADET	0.0126	0.0200	0.0514	0.1046	0.0097
VUS	0.3670	0.3333	0.3332	0.5431	0.2706

Table 6.3. The estimated optimal cut-points corresponding to the optimal statistics in Table 6.2.

	Estimated Cut-points	Cerebrospinal Fluid			Brain Imaging	
		Abeta	TAU	PTAU	Hippocampus	Whole-Brain
GTKL	$\hat{c}_1$	191.50	74.67	24.52	6649.28	1,023,917
	$\hat{c}_2$	158.59	133.24	39.74	4925.29	899,145
GYI	$\hat{c}_1$	181.46	81.97	25.58	6815.26	1,022,717
	$\hat{c}_2$	181.46	101.06	25.58	5784.19	937,262
MADET	$\hat{c}_1$	186.16	58.00	24.40	7025.60	1,025,544
	$\hat{c}_2$	154.07	101.56	41.13	5532.20	927,991
CP	$\hat{c}_1$	190.9	68.12	22.29	7033.32	1,032,416
	$\hat{c}_2$	131.24	114.75	32.83	5499.27	920,270
MV	$\hat{c}_1$	190.08	67.67	21.68	7002.24	1,033,179
	$\hat{c}_2$	130.93	116.22	35.24	5522.44	918,283

Table 6.4. The corresponding correct classification rates.

Biomarkers	Measures	p_{11}	p_{22}	p_{33}
Abeta	GTKL	0.6135	0.1504	0.7324
	GYI	0.6556	0	0.8897
	MADET	0.6359	0.1698	0.689
	CP	0.6161	0.4189	0.4386
	MV	0.6196	0.4200	0.4351
TAU	GTKL	0.6466	0.4032	0.3340
	GYI	0.7186	0.1523	0.5485
	MAD	0.4224	0.3643	0.5450
	CP	0.568	0.3646	0.4525
	MV	0.5622	0.3771	0.4424
PTAU	GTKL	0.6492	0.3094	0.3339
	GYI	0.6819	0	0.7749
	MAD	0.6454	0.3382	0.3016
	CP	0.5728	0.2205	0.5476
	MV	0.5484	0.2819	0.4654
Hippocampus	GTKL	0.7757	0.5117	0.3850
	GYI	0.7336	0.3375	0.7341
	MAD	0.6621	0.4732	0.6437
	CP	0.659	0.4847	0.6309
	MV	0.6712	0.4698	0.6399
Whole Brain	GTKL	0.4512	0.4128	0.3952
	GYI	0.4552	0.3003	0.5221
	MAD	0.4458	0.3385	0.4921
	CP	0.4228	0.3833	0.4665
	MV	0.4202	0.3914	0.4599

Table 6.5. Generalized predictive values (i.e., PPV and NPV) and likelihood ratios of biomarkers.

Biomarkers	Measures	PPV_1	PPV_2	NPV
Abeta	GTKL	0.2097	0.1624	0.8913
	GYI	0.1822	0.1530	0.8829
	MADET	0.2293	0.1641	0.8875
	CP	0.2744	0.1784	0.8909
	MV	0.2762	0.1786	0.8904
TAU	GTKL	0.2360	0.2525	0.8425
	GYI	0.2307	0.1778	0.8329
	MADET	0.1815	0.1787	0.8577
	CP	0.2114	0.2079	0.8502
	MV	0.2112	0.2118	0.8506
PTAU	GTKL	0.2336	0.1375	0.8642
	GYI	0.1729	0.1448	0.8585
	MADET	0.2386	0.1347	0.8647
	CP	0.1845	0.1518	0.8730
	MV	0.1896	0.1481	0.8748
Hippocampus	GTKL	0.3353	0.5586	0.8731
	GYI	0.2942	0.3414	0.8825
	MADET	0.2811	0.3995	0.8920
	CP	0.2814	0.4079	0.8922
	MV	0.2845	0.4020	0.8911
Whole-Brain	GTKL	0.2174	0.1603	0.7680
	GYI	0.2241	0.1369	0.7677
	MADET	0.2218	0.1421	0.7683
	CP	0.2189	0.1468	0.7697
	MV	0.2185	0.1480	0.7698

6.3. Discussion

Based on the distribution graphs in Figure 6.1, the biomarkers Abeta and PTAU are not functional biomarkers that can discriminate subjects among three stages, since the middle stage almost entirely overlaps with the other two stages. According to the optimal statistics in Table 6.2, hippocampus has the highest statistics compared to other biomarkers, and the results are consistent among all measures. Also, the plots show that hippocampus has the most distinct distribution curve among the three clinical stages.

The optimal cut-points of Abeta and PTAU selected by GYI are identical. Consequently, the correct classification rate at the middle stage is zero for both biomarkers. The results imply that Abeta and PTAU are not suitable biomarkers to distinguish subjects among three stages in this study although the optimal statistics of these two biomarkers are closed to the optimal statistics of TAU and slightly higher than the values of whole-brain. Except GYI, all four methods can identify two different optimal cut-points of Abeta and PTAU; however, the results are not promising.

The biomarkers TAU, hippocampus, and whole-brain show comparatively more distinct distributions compared to the other two. The optimal statistics of TAU lies between the values of Abeta and PTAU, and it is slightly higher than the statistics of whole-brain. Hippocampus has much higher statistics than others using all measures. Thus, hippocampus is the best biomarker that can discriminate subjects among three stages of Alzheimer's disease. Among all diagnostic accuracy measures, GYI is consistent and promising in identifying subjects in the first and the third stages. The other measures are comparatively well except GTKL has a lower correct classification rate of the last stage in some cases. However, the GTKL generally has the highest correct classification rate of the middle stage and provides more information when a diagnostic test aims to recognize subjects in the middle stage. Detecting subjects in the early stage of the disease is significant for Alzheimer's disease as it is an irreversible disease that progresses over a long period. The brain changes caused by Alzheimer's disease may begin 20 years or more before any symptoms appear (Gaugler, James, Johnson, Marin, & Weuve, 2019). Additionally, the generalized predictive values shown in Table 6.5 indicate that the GTKL has the highest PPV_2. The results suggest that the GTKL has predicted the true stage of the subjects in the middle stage better than the other measures, mainly when using biomarkers TAU, hippocampus, and whole-brain.

Early diagnosis and intervention are essential in slowing down the progression of the diseases. Gaugler et al. (2019) mentioned the reasons that seniors believe early diagnosis is important not just for early intervention of the disease but also helps them understand what is happening with the disease and allow the family to plan for the future. Compared to the existing measures, the proposed measure GTKL, has high correct classification rates of stage 1 and stage 2, which is the specificity of stages 1 and 2, respectively, using hippocampus as the biomarker. The specificity of stages 1 and 2 emphasizes the rule-out information and provides essential information for diagnosing Alzheimer's disease in the early stage.

Conclusively, the GTKL is a better criterion compared to others when using the hippocampus as the biomarker for early diagnosis for subjects in stage 1 and stage 2. Its classification rates in early diagnosis are the highest among all measures. When the diagnosis of early-stage is not the primary focus, GYI gives better overall correct classification rates. Additionally, the MADET, CP, and MV are more balanced in identifying subjects among three stages as their correct classification rates are reasonably well in all three stages. Lastly, hippocampus has been the best performance in the scenario of three-stage setting, and Abeta has decent performance in the situation of the two-stage setting.

CHAPTER 7
FINAL REMARKS, CONCLUSIONS, LIMITATIONS, AND FUTURE RESEARCH

7.1 Final remarks and conclusion

Accuracy has played a critical role in diagnostic tests. The accuracy of a diagnostic test is essential to placing patients with adequate treatment plans; thus, measuring the accuracy of a test has significant clinical implications. Traditionally, diagnostic accuracy tests, like Youden index and ROC curve, have been guiding clinical decisions, and further assisting in designation of clinical guidelines. Although plenty of studies have been conducted to improve the accuracy of diagnostic tests, limited studies have been performed to assess the accuracy of a diagnostic test for multi-stage diseases. Additionally, selecting optimal cut-points for multi-stage diseases is more challenging for multi-stage diseases compared to two-stage diseases.

In this innovative study, the Kullback-Leibler (KL) divergence, from information theory, is applied to measuring diagnostic accuracy and generalized to estimate the optimal cut-points for multi-stage diseases. Also, this study conducts massive simulation under three-stage setting (i.e., the special case of multi-stage diseases) to assess the performance of the proposed measure, the generalized total Kullback-Leibler (GTKL) divergence, the generalized Youden index (GYI), hypervolume under manifold (HUM), closest-to-perfection (CP), maximum volume (MV), and maximum absolute determinant (MADET) in terms of the power of detecting the difference among the distributions of different stages and selecting optimal cut-points for multi-stage diseases (Attwood et al., 2014; Dong et al., 2017; Nakas et al., 2010; Xiong et al., 2006). Additionally, this study adapts normalized statistics, like relative bias, normalized root-mean-squared error, maximum-minimum difference, and percentage of the loss of total correct classification rate, to compare the performance of diagnostic accuracy measures. The simulation results indicate that there is no dominant winner in general.

This study also provides an example of the application of the measures for multi-stage diseases, using a dataset from the Alzheimer’s Disease Neuroimaging Initiative (ADNI). The results concord the simulation study and highlight the strengths of different diagnostic accuracy measures. It provides clues for early diagnosis of Alzheimer's disease. For example, the GTKL is capable of detecting subjects in the first and the second stages with the highest correct classification rates in these two stages (1.2874); and, the GYI has the best performance in identifying subjects in the first and the last stages with the correct classification in these two stages (1.4678), using hippocampus as the biomarkers. The study provides exciting exploratory outcomes in improvement of diagnostic accuracy tests for multi-stage diseases, especially for those who want to discover biomarkers for early diagnosis.

7.2 Limitations and future research

In this study, the GTKL has shown some advantages in detecting subjects in the middle stage of a multi-stage disease and has been beneficial for early diagnosis; however, the total correct classification rate is slightly lower than the other measures, especially in the last stage of the disease. Also, although the performance of the measure is compared in terms of the balance among all disease stages, no existing standardized methods can be used to compare the performance of the measures since they have different properties. Even though the balance of a measure is evaluated based on percentage of the loss of total correct classification rate and the maximum-minimum difference, the question about whether the balance of a measure should be used as a criterion for comparison remains dubious. For two-stage diseases, high correct classification rates in both stages are desired. Nevertheless, for multi-stage diseases, the demand for a high correct classification rate of a specific stage relies on the interest of clinical needs. For instance, when a clinical test is designed to identify subjects in the early stages, the correct classification rates of the early stages are expected to be as high as needed. Alternatively, when a clinical test is considered to identify patients in the later stages, the correct classification rate of the last stage would be more critical compared to the others. It is ideal to have a balanced diagnostic test with high correct classification rates among all stages, yet it is hard to achieve in reality. Hence, a method that can evaluate performance of different diagnostic accuracy measures, besides the balance, is desired in future study.

Additionally, diagnostic accuracy measures range differently thus the accuracy of a biomarker is not comparable among different measures. The generalized predictive values proposed in an on-going project by Samawi et al. (2020) provides general statistics for the comparison among measures. However, the generalized predictive values depend highly on the prevalence of the disease and cannot be generalized to a different population. Consequently, future study is encouraged to compare the performance of various diagnostic accuracy measures using methods that do not depend on prevalence like diagnostic likelihood ratios.

Moreover, the estimation conducted in this study is restricted in using the kernel approach; so, simulation of estimation using other approaches is also encouraged for comparison of the performance of the diagnostic measures, as well as the estimation of their confidence intervals. Research about the properties and strengths of different measures under various distributions is desired in the future.

Furthermore, early diagnosis provides ample time for health practitioners to fight with severe diseases specifically for the diseases without a cure; however, the area of diagnostic tests for multi-stage diseases lacks real data applications. Therefore, there is a need to develop reliable and practical diagnostic accuracy measures, which further help improve diagnosis to assist in designing clinical treatments and guidelines.

Lastly, in a lot of cases that a disease does not have a gold standard or even the gold standard cannot be one-hundred percent accurate, a single biomarker is less than satisfactory for confirmation of clinical diagnosis. Also, when using genetic and epigenetic biomarkers in cancer diagnosis and treatment evaluation, a single biomarker is not enough to identify the subtypes of cancer or confirm the staging of cancer patients. As a result, generalization of the single biomarker measures to multiple biomarkers is an exciting topic in the future study of the diagnosis for multi-stage diseases.

REFERENCES

Aarsland, D., & Kurz, M. W. (2010). The epidemiology of dementia associated with Parkinson disease. *Journal of the neurological sciences*, *289*(1-2), 18-22.

Akobeng, A. K. (2007). Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta paediatrica, 96*(3), 338-341.

Altman, D. G., & Bland, J. M. (1994). Statistics Notes: Diagnostic tests 2: predictive values. *Bmj, 309*(6947), 102.

Alzheimer's Association. (2019). Alzheimer's and Dementia: Stages of Alzheimer's. Retrieved from https://www.alz.org/alzheimers-dementia/stages.

Alzheimer’s Disease Neuroimaging Initiative (ADNI). (2017). Study design. Retrieved from http://adni.loni.usc.edu/study-design/#background-container.

Aoki, H., Watanabe, T., Furuichi, M., & Tsuda, H. (1997). Use of alternative protein sources as substitutes for fish meal in red sea bream [Pagrus major] diets. *Suisanzoshoku (Japan).*

Aoki, K., Misumi, J., Kimura, T., Zhao, W., & 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*(3), 143-151.

Attwood, K., Tian, L., & Xiong, C. (2014). Diagnostic thresholds with three ordinal groups. *J Biopharm Stat, 24*(3), 608-633. doi:10.1080/10543406.2014.888437

Blennow, K., Hampel, H., Weiner, M., & Zetterberg, H. (2010). Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. *Nature Reviews Neurology, 6*(3), 131.

Boyko, E. J. (1994). Ruling out or ruling in disease with the most sensitiue or specific diagnostic test: Short cut or wrong turn?. *Medical Decision Making, 14*(2), 175-179.

Centers for Disease Control and Prevention (CDC). (2019). What is dementia?. Retrieved from https://www.cdc.gov/aging/dementia/index.html.

DAFFNEr, K. R., & Scinto, L. F. (2000). Early diagnosis of Alzheimer's disease *Early diagnosis of Alzheimer's disease* (pp. 1-27): Springer.

Das, P., Murphy, M. P., Younkin, L. H., Younkin, S. G., & Golde, T. E. (2001). Reduced effectiveness of Aβ1-42 immunization in APP transgenic mice with significant amyloid deposition. *Neurobiology of aging, 22*(5), 721-727.

Deeks, J. J., & Altman, D. G. (2004). Diagnostic tests 4: likelihood ratios. *Bmj, 329*(7458), 168-169.

Dong, T., Attwood, K., Hutson, A., Liu, S., & Tian, L. (2017). A new diagnostic accuracy measure and cut-point selection criterion. *Stat Methods Med Res, 26*(6), 2832-2852. doi:10.1177/0962280215611631

Duthey, B. (2013). Background paper 6.11: Alzheimer disease and other dementias. *A Public Health Approach to Innovation*, 1-74.

Early Breast Cancer Trialists' Collaborative Group. (2005). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. The Lancet, 365(9472), 1687-1717.

Egan, J. P. (1975). *Signal detection theory and ROC-analysis*: Academic press.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. doi:10.1016/j.patrec.2005.10.010

Fletcher, R. H., Fletcher, S. W., & Fletcher, G. S. (2012). *Clinical epidemiology: the essentials*: Lippincott Williams & Wilkins.

Garcia-Alloza, M., Subramanian, M., Thyssen, D., Borrelli, L. A., Fauq, A., Das, P., . . . Bacskai, B. J. (2009). Existing plaques and neuritic abnormalities in APP: PS1 mice are not affected by administration of the gamma-secretase inhibitor LY-411575. *Molecular neurodegeneration, 4*(1), 19.

Gaugler, J., James, B., Johnson, T., Marin, A., & Weuve, J. (2019). 2019 Alzheimer's disease facts and figures. *Alzheimers & Dementia, 15*(3), 321-387.

Gilbert, R., Logan, S., Moyer, V. A., & Elliott, E. J. (2001). Assessing diagnostic and screening tests: Part 1. Concepts. *The Western journal of medicine, 174*(6), 405-409.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. M. (2003). The diagnostic odds ratio: a single indicator of test performance. *Journal of Clinical Epidemiology, 56*(11), 1129-1135. doi:10.1016/s0895-4356(03)00177-x

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1): Wiley New York.

Grundman, M., & Delaney, P. (2002). Antioxidant strategies for Alzheimer's disease. *Proceedings of the Nutrition Society, 61*(2), 191-202.

Hsiao, K., Baker, H. F., Crow, T. J., Poulter, M., Owen, F., Terwilliger, J. D., . . . Prusiner, S. B. (1989). Linkage of a prion protein missense variant to Gerstmann–Sträussler syndrome. *Nature, 338*(6213), 342.

Hughes, G., & Bhattacharya, B. (2013). Symmetry Properties of Bi-Normal and Bi-Gamma Receiver Operating Characteristic Curves are Described by Kullback-Leibler Divergences. *Entropy, 15*(12), 1342-1356. doi:10.3390/e15041342

Kantarci, K., Weigand, S., Przybelski, S., Shiung, M., Whitwell, J. L., Negash, S., . . . Petersen, R. C. (2009). Risk of dementia in MCI: combined effect of cerebrovascular disease, volumetric MRI, and 1H MRS. *Neurology, 72*(17), 1519-1525.

Lee, W.-C. (1999). Selecting diagnostic tests for ruling out or ruling in disease: the use of the Kullback-Leibler distance. *International journal of epidemiology, 28*(3), 521-525.

Letón, E., & Molanes, E. M. (2009). Adjusted empirical likelihood estimation of the youden index and associated threshold for the bigamma model.

Levites, Y., Das, P., Price, R. W., Rochette, M. J., Kostura, L. A., McGowan, E. M., . . . Golde, T. E. (2006). Anti-Aβ 42–and anti-Aβ 40–specific mAbs attenuate amyloid deposition in an Alzheimer disease mouse model. *The Journal of clinical investigation, 116*(1), 193-201.

Li, J., & Fine, J. P. (2008). ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics, 9*(3), 566-576. doi:10.1093/biostatistics/kxm050

Lusted, L. B. (1960). Logical analysis in roentgen diagnosis: memorial fund lecture. *Radiology, 74*(2), 178-193.

McNeil, B. J., & Adelstein, S. J. (1976). Determining the value of diagnostic and screening tests. *Journal of Nuclear Medicine, 17*(6), 439-448.

Metz, C. E. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigative radiology, 24*(3), 234-245.

Mitchell, A. J., & Shiri ‐ Feshki, M. (2009). Rate of progression of mild cognitive impairment to dementia ‐ meta ‐ analysis of 41 robust inception cohort studies. *Acta Psychiatrica Scandinavica, 119*(4), 252-265.

Nakas, C. T., Alonzo, T. A., & Yiannoutsos, C. T. (2010). Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index. *Stat Med, 29*(28), 2946-2955. doi:10.1002/sim.4044

Nakas, C. T., Dalrymple-Alford, J. C., Anderson, T. J., & 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. *Stat Med, 32*(6), 995-1003. doi:10.1002/sim.5592

Nakas, C. T., & Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Stat Med, 23*(22), 3437-3449. doi:10.1002/sim.1917

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*: Medicine.

Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol, 163*(7), 670-675. doi:10.1093/aje/kwj063

Richards, M., Westcombe, A., Love, S., Littlejohns, P., & Ramirez, A. (1999). Influence of delay on survival in patients with breast cancer: a systematic review. *The Lancet, 353*(9159), 1119-1126.

Roberts, R., & Knopman, D. S. (2013). Classification and epidemiology of MCI. *Clinics in geriatric medicine, 29*(4), 753-772.

Samawi, H. M. (2020). *Generalization of predictive values and diagnostic likelihood ratios (inpress).*

Samawi, H. M., Yin, J., Rochani, H., & Panchal, V. (2017). Notes on the overlap measure as an alternative to the Youden index: How are they related? *Statistics in medicine, 36*(26), 4230-4240.

Samawi, H. M., Yin, J., Zhang, X., Rochani, H., Vogel, R. L., & Mo, C. (2019). *Kullback-Leibler Divergence for Medical Diagnostics Accuracy and Cut-point Selection Criterion: How it is related to Youden Index (inpress).*

Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology, 40*(3), 253-269.

Scurfield, B. K. (1998). Generalization of the Theory of Signal Detectability ton-Eventm-Dimensional Forced-Choice Tasks. *Journal of Mathematical Psychology, 42*(1), 5-31.

Shaffer, J. L., Petrella, J. R., Sheldon, F. C., Choudhury, K. R., Calhoun, V. D., Coleman, R. E., . . . Initiative, A. s. D. N. (2013). Predicting cognitive decline in subjects at risk for Alzheimer disease by using combined cerebrospinal fluid, MR imaging, and PET biomarkers. *Radiology, 266*(2), 583-591.

Simonoff, J. S. (2012). *Smoothing methods in statistics*: Springer Science & Business Media.

Šimundić, A.-M. (2009). Measures of diagnostic accuracy: basic definitions. *Ejifcc, 19*(4), 203.

Sox Jr, H. C., Koran, L. M., Sox, C. H., Marton, K. I., Dugger, F., & Smith, T. (1989). A medical algorithm for detecting physical disease in psychiatric patients. *Psychiatric Services, 40*(12), 1270-1276.

Swets, J., & Pickett, R. (1982). Methods from signal detection theory. *Evaluation of diagnostic systems. Academic Press, London*, 17-37.

Wand, M. P., & Jones, M. C. (1994). *Kernel smoothing*: Chapman and Hall/CRC.

Wong, H. B., & Lim, G. H. (2011). Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV. *Proceedings of Singapore healthcare, 20*(4), 316-318.

Xiong, C., van Belle, G., Miller, J. P., & Morris, J. C. (2006). Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Stat Med, 25*(7), 1251-1273. doi:10.1002/sim.2433

Youden, W. J. (1950). INDEX FOR RATING DIAGNOSTIC TESTS. *Cancer, 3*(1), 32-35.

Zhou, X.-H., McClish, D. K., & Obuchowski, N. A. (2009). *Statistical methods in diagnostic medicine* (Vol. 569): John Wiley & Sons.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry, 39*(4), 561-577.