Spring 2022

Generalization of Net Benefit of Diagnostic Tests into Multi-stage Clinical Conditions: A Collapsing Approach

Ferdous Ahmed

GENERALIZATION OF NET BENEFIT OF DIAGNOSTIC TESTS INTO MULTI-STAGE CLINICAL CONDITIONS: A COLLAPSING APPROACH

by

FERDOUS AHMED

(Under the Direction of Hani M. Samawi)

ABSTRACT

Using accuracy measures alone to compare diagnostic tests may be unconvincing to clinicians. The diagnostic test accuracy is commonly evaluated in clinical performance based on its classification accuracy (specificity, sensitivity, negative and positive likelihood ratio) or its predictive values (negative and positive predictive value). However, these accuracy measures do not entirely account for the clinical and health economic consequences of diagnostic errors. The limitation of these measures is that one test may have a better sensitivity and worse specificity than another test. Comparing tests on benefit-risk is another approach where benefits and risks are put on the same scale to determine test benefits and clinical consequences of the diagnostic errors. Consequently, evaluating diagnostic tests based on benefit-risk involves both the tests' accuracy and the clinical implications of the diagnostic errors. Diagnostic tests are commonly classified into two stages: either positive or negative for a clinical condition (diseased or non-diseased). However, some diseases have more than two stages, such as Alzheimer's. In diseases with more than two stages, the benefits and risks of the clinical consequences could differ from stage to stage. I could not find any investigations to account for the difference in benefits and risks of tests with more than two stages in the literature. The benefit to cost values for each stage of the disease could be different. This dissertation extends the net benefit approach of evaluating diagnostic tests in binary disease cases to multi-stage clinical conditions. Consequently, I extend the diagnostic yield table to multi-stage clinical conditions. I develop a decision process based on net benefit for evaluating diagnostic tests. The decision process provides additional interpretation for rule-in or rule-out clinical conditions and their adverse consequences from unnecessary workups in multi-stage diseases. Numerical examples, as well as real data, are provided to illustrate the proposed measures.

INDEX WORDS: Loss function, Diagnostic yield table, Relative net benefit, Clinical utility, Benefit-risk, Medical diagnostics yield, Decision theory, Alzheimer’s disease.

GENERALIZATION OF NET BENEFIT OF DIAGNOSTIC TESTS INTO MULTI-STAGE CLINICAL CONDITIONS: A COLLAPSING APPROACH

by

FERDOUS AHMED

B.D.S., Sapporo Dental College, University of Dhaka, Bangladesh, 2011
M.P.H., Georgia Southern University, 2018

A Dissertation Submitted to the Graduate Faculty of Georgia Southern University

Fulfillment of the Requirements for the Degree of

DOCTOR OF PUBLIC HEALTH

GENERALIZATION OF NET BENEFIT OF DIAGNOSTICS TEST INTO MULTI-STAGE CLINICAL CONDITIONS: A COLLAPSING APPROACH

by

FERDOUS AHMED

Major Professor: Hani M. Samawi

Committee: Jingjing Yin
Lili Yu

Electronic Version Approved:

May 2022

ACKNOWLEDGMENTS

Pursuing a doctoral degree is like climbing a high peak, step by step, accompanied by encouragement, challenge, faith, assurance, frustration, and hardship. When enjoying the striking scenery at the top of the peak, I recognize that teamwork has allowed me to reach my goal. This doctoral study has helped me build the skills and resilience to tackle complex problems and persevere, despite any obstacles that may have come my way. My gratitude is not enough, but I would still like to thank many people for their guidance and support throughout my study.

My profound appreciation goes to my committee chair, Dr. Hani M. Samawi, whose expertise has been invaluable in formulating my research ideas, questions and methodology. I am also thankful for his comprehensive guidance and support, thoughtful comments, and recommendations. I appreciate his patience and insightful feedback that have pushed me to sharpen my thinking and bring my work to a higher level. Even though Dr. Samawi is very busy with his teaching, research, and chairing of the department, he has still enthusiastically and continually encouraged me. He has been willing to assist me throughout my study.

As I acknowledge my dissertation committee members, Dr. Jingjing Yin, Dr. Lili Yu, and Dr. Jing Kersey, all of whom have invested valuable time to read my dissertation and provide their thoughtful suggestions, their help has been invaluable in shaping my methods and critique my results.

And finally, I want to thank my friends, lab mates, and Ms. David Parker for their friendship and help, which enabled me to present my dissertation more clearly and fluently in English. I definitely express my profound gratitude to my parents and sister for giving me continuous encouragement while researching and writing this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Diagnostic tests play a significant role in patients’ health care, including medical diagnosis, screening tests, appropriate therapy, research, and health policy. There are mainly three purposes of performing a diagnostic test, namely, to provide reliable information about a patient's health condition, to influence the treatment plan for the patient from health care providers, and to understand that diseases have mechanisms and natural history of progress within the body via research (McNeil & Adelstein, 1976; Sox Jr et al., 1989; Zhou, McClish, & Obuchowski, 2009). A test can serve these purposes only if the health care provider knows the parameters and conditions of the disease and how to interpret them. This information is acquired by assessing the test's diagnostic accuracy, which is simply the ability of a test to discriminate between subjects, between diseased or non-diseased, providing a solid understanding of patients’ health conditions. When a diagnosis is accurate and made promptly, a patient has the best opportunity for a positive health outcome. Clinical decision-making has been tailored to a correct understanding of the patient's health problem, ultimately improving healthcare for all patients (Holmboe & Durning, 2014). Also, diagnostic information often influences public policy decisions, such as setting payment policies, resource allocation, and research priorities (Jutel & illness, 2009; Rosenberg, 2002; World Health Organization [WHO], 2012).

While a perfect diagnostic test discriminates between diseased and non-diseased subjects completely, in reality, it is hard to achieve. A diagnostic test can only partially distinguish between subjects with or without the disease. Furthermore, the clinical consequences of diagnostic error are required for a diagnostic test or continuous biomarker in making a diagnostic decision in the case of binary or ordinal disease stages.

A diagnostic test result does not accurately represent the patient’s condition because diagnostic tests rarely have perfect accuracy. Accuracy refers to the probability of the test result that ultimately reflects the actual disease state. Developing quantitative methods to measure diagnostic accuracy and clinical consequences is essential. Some of the well-established techniques of diagnostic accuracy before test measures, such as sensitivity (Se or TPR) and specificity (Sp or TNR), Youden index, the area under the ROC curve (AUC), and diagnostic odds ratio (DOR) are used to assess the discriminative property of the test. Other methods of diagnostic accuracy used after test measures are likelihood ratios (LRs) and predictive values (PPV and NPV), which help assess its predictive ability.

Different measures of diagnostic accuracy have various aspects of use based on the purpose of diagnostic procedures. Some diagnostic accuracy measures assess the ability to differentiate between the non-diseased and the diseased, and others measure its predictive ability. Diagnostic accuracy measures are also susceptible to the spectrum of diseases and the tested population. It is essential to know which measure to use under what conditions and interpret these measures carefully. Some studies have examined the efficiency of different diagnostic tests, but it is hard to find test results that are always accurate in reality (Akobeng, 2007; Altman & Bland, 1994; Deeks & Altman, 2004; Margaret Sullivan Pepe, 2003; A.-M. Šimundić, 2009; Wong & Lim, 2011; Zhou et al., 2009). In misclassification, a medical diagnostic test positively affects a subject who does not have the disease, and a diseased subject may be diagnosed as non-diseased (Margaret Sullivan Pepe, 2003; Zhou et al., 2009). Diagnostic accuracy is the ability to discriminate between non-diseased and diseased or non-diseased and different stages of a particular disease state. Health care professionals need to assess better the performance of diagnostic tests on discriminating patients with and without the disease of interest to determine the actual stage of the patient's condition and make treatment plans for their patients. It is essential to use accurate medical tests and thus avoid error, unnecessary suffering, and expense. Many measures for diagnostic accuracy have been developed to describe the performance of a biomarker for binary scale disease. Sensitivity and specificity are the correct classification rates, and false-positive rate (FPR) and false-negative rate (FNR) are the misclassification rates among these measures (Margaret Sullivan Pepe, 2003; Zhou et al., 2009). While sensitivity and specificity are used to maximize correct classification rates, the FPR and FNR minimize misclassification rates.

In medical diagnosis, the sensitivity of a test is the ability of a test to accurately detect when an individual has a disease. The specificity of a test is the ability to accurately detect an individual as disease-free who has no disease. A high sensitivity test helps rule out disease if a person tests negative. A high specificity test helps rule a condition if a person tests positive. A diagnostic test identifies the presence or absence of a specific disease when a subject shows significant disease symptoms. The diagnostic test is an essential determinant for health care providers to decide whether to give treatments for the disease, especially when the treatments are invasive or harmful procedures, such as chemotherapy and radiotherapy (Gilbert, Logan, Moyer, & Elliott, 2001). Sometimes the diagnostic test itself has consequences, including an invasive procedure, such as a biopsy, or introducing energy into the body, as with radiation, using an X-ray. A screening test is designed to identify asymptomatic subjects at sufficient risk of the disease who have not received medical attention or who do not warrant further health interventions among the population (Gilbert et al., 2001). Commonly, the diagnostic test is performed after a screening test to make a confirmed diagnosis. Different tests are carried out to discriminate between diseased and non-diseased conditions based on sensitivity and specificity measures.

The Receiver Operating Characteristic curve (ROC) and the area under the ROC curve (AUC) provide brief measures associated with single sensitivity and specificity pairs by including all the decision thresholds. Mo (2020) states that some of the steps integrate sensitivity and specificity into a single index-like accuracy called diagnostic effectiveness. For example, diagnostic odds ratio (OR) and Youden index overlap measure (H. M. Samawi, Yin, Rochani, & Panchal, 2017) and KL divergence (Lee, 1999). The OR and the Youden index do not depend on disease prevalence like sensitivity and specificity. These measures could quickly transfer results from one study to another with a different disease prevalence in the population. But these measures 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, accuracy is affected by disease prevalence (A.-M. J. E. 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 other populations with different disease prevalence. In addition to these measures introduced above, another type of diagnostic test accuracy measure is predictive accuracy, including the negative and positive predictive value (NPV and PPV), respectively, and diagnostic likelihood ratios (LRs) (A.-M. Šimundić, 2009; Zhou et al., 2009). Predictive values have significant clinical implications for a diagnostic test. Although measures like sensitivity and specificity estimate the probability of the disease in patients, they cannot answer how likely it would be for patients to receive positive or negative test results. Predictive values are the measures that provide information about the probability that a test result gives a correct diagnosis. Given a positive test result, a positive predictive value (PPV) shows the probability of having the state or disease of interest in a subject. That means PPV represents a proportion of patients with positive test results among total subjects with positive results. A negative predictive value (NPV) is the probability that a subject receives a negative effect yet does not have the disease of interest (Altman & Bland, 1994; Wong & Lim, 2011). It means that NPV represents a proportion of subjects without the disease, having a negative test result in a total of subjects with negative test results. Unlike sensitivity and specificity, predictive values are highly dependent on the disease prevalence, which cannot be generalized among different populations with different disease prevalence. Compared to predictive values, LRs can also provide information about the probability that a subject can be correctly diagnosed; nevertheless, LRs do not depend on prevalence as the predictive values, and they apply to other clinical settings for the same disease (Boyko, 1994; Deeks & Altman, 2004). LRs are also the best indicator for rule-in or rule-out of the diagnosis (Boyko, 1994; Deeks & Altman, 2004; Gilbert et al., 2001). Mainly, 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 high specificity test's positive response makes the patient more likely to have the condition since it is specific. With more significant, more considerable sensitivity, a rule-out test emphasizes assessing if test results will exclude the possibility that a subject is non-diseased. A high sensitivity test's negative response makes the patient more likely not to have the disease since it is sensitive.

Traditionally, medical diagnostic tests are evaluated using accuracy measures of the discrimination ability of the test and its predictive ability. However, these accuracy measures do not entirely account for clinical and health economic contexts. Gail and Pfeiffer (2005) argue that decision theory methods could provide more relevant clinical application outcomes. H.

Samawi, Chen, Ahmed, and Kersey (2021) state that these methods evaluate risk models in treatment decisions by conveying costs and benefits on the same scale that involves the accuracy of diagnostic tests and clinical consequences of diagnostic errors using utilities (Rapsomaniki, White, Wood, Thompson, & Emerging Risk Factors, 2012).

Furthermore, as G. Pennello, N. Pantoja-Galicia, and S. Evans (2016a) argue, evaluating a diagnostic test based on benefit-risk involves both the test's accuracy and the clinical consequences of diagnostic errors. Three things, clinical setting, the intended use of diagnostic tests, and the characteristics of the population on whom they will be used, may affect the evaluation of the clinical consequences of false positive and false negative test errors. The diagnostic test may have clinical effects (e.g., biopsy). Other health economics investigators may consider the test's cost and the consequential costs of treating positive test subjects (Tsalik et al., 2016).

In the literature, I find some studies that compare diagnostic tests based on benefit-risk in two-stage diseases. They describe methods for evaluating the benefit-risk of a binary diagnostic test based on its diagnostic accuracy from a clinical performance study and external information on clinical consequences (Pennello et al., 2016a).

Pennello et al. (2016a) use a benefit-risk approach based on a decision-theoretic framework to compare diagnostic tests or biomarkers. Their method assigns losses to misclassifications (i.e., false-positive and negative) and gives utilities to correct classifications (true positive and true negative). Their theory provides different interpretations of quantities in the diagnostic yield table. It indicates that a weighted accuracy measure proposed previously (Evans et al., 2016) could be interpreted as a relative utility measure. They define comparable utility measure as the test's expected utility close to a perfect test. They also describe expected benefits from testing and net benefits relative to a perfect test. Evans et al. (2016) indicate that the expected benefit of a test and the net benefit close to an ideal test are similar to the expected benefit measures proposed for risk prediction models (S. G. Baker, Cook, Vickers, & Kramer, 2009; S. G. Baker & Kramer, 2012; S. G. Baker, Van Calster, & Steyerberg, 2012; S. G. J. J. J. o. t. N. C. I. Baker, 2009; Gail & Pfeiffer, 2005; Vickers & Elkin, 2006) and (Margaret S Pepe et al., 2016).

Most medical diagnostic tests are commonly classified into two stages: either positive or negative for a clinical condition (diseased or non-diseased). However, some diseases have more than two stages in clinical practice, such as Alzheimer's. Alzheimer's disease has four stages, including preclinical stage, early-stage (mild), middle stage (moderate), and late-stage (severe) (Alzheimer's Association, 2019; Johns Hopkins Medicine, 2019a). A measure that can discriminate among more than two stages is desired for this type of disease. But, I could not find any studies comparing diagnostic tests based on the benefit-risk of multi-stage disease tests in the literature. Clinical utility values for each stage of the disease could be different. Therefore, this dissertation extends the net benefit approach of evaluating diagnostic tests to multi-stage clinical conditions. Consequently, I extend the diagnostic yield table to multi-stage clinical conditions. I develop a decision theory based on the net benefit of evaluating diagnostic tests that provide additional interpretation for rule-in or rule-out clinical needs and their adverse consequences from unnecessary workups for multi-stage diseases.

In this research, we generalize the net benefit of a diagnostic test, from two-stage to multi-stage diseases, as an evaluation of a diagnostic test that involves both the test's accuracy and the clinical consequences of diagnostic error. This evaluation sums up the rule-in or rule-out information in all stages, and it comprehensively evaluates the correct classification rates in all phases of a multi-stage disease. Overall, the generalization of a diagnostic test based on benefit-risk combines correct classification rates and misclassification rates based on benefit-risk for conditions with more than two stages and simultaneously emphasizes the rule-in, rule-out potentials for diagnosis in all stages.

CHAPTER 2

LITERATURE REVIEW

Traditionally, medical diagnostic tests are evaluated using accuracy measures based on the discrimination ability of the test. However, using accuracy measures alone to compare diagnostic tests or biomarkers may be unconvincing for clinicians. Comparing tests based on benefit-risk may be more conclusive because it involves the accuracy of the test and the clinical consequences of diagnostic error. Many factors, such as clinical setting, the intended use of the test, and the population on whom it will be used, play an essential role in evaluating the clinical consequences of false positive and false negative test errors. Sometimes, the test itself has a clinical implication. For example, it may involve an invasive procedure (e.g., biopsy) or introduce energy into the body, such as radiation (e.g., X-ray computed tomography (CT) scan). A health economics analysis may also consider the decision-analytics framework and the potential cost-effectiveness of working up positive test subjects (Tsalik et al., 2016).

To compare diagnostic tests or biomarkers on accuracy measures alone may not be convincing to clinicians. The diagnostic test accuracy is commonly evaluated in a clinical performance based on its classification accuracy (specificity, sensitivity, negative and positive likelihood ratio) or its predictive values (negative and positive predictive value). However, these evaluations can sometimes be insufficient for comparing the clinical consequences of the two tests. The primary constraint of these measures is that one test may have a better sensitivity and worse specificity than another test.

ROC and Youden indices are prevalent, essential measures and the most popular tools for binary classification. They describe different aspects of a biomarker in diagnostic studies for test evaluation. Both measures are built based on the four fundamental estimates of diagnostic accuracy: sensitivity, specificity, FPR, and FNR. Diagnostic accuracy measures are susceptible to the characteristics of the population in which the test accuracy is evaluated. Some estimates of diagnostic accuracy highly depend on the prevalence of the disease of interest, while others are susceptible to the spectrum of the disease in the studied population (A.-M. J. E. Šimundić, 2009). The four basic measures (Sensitivity (TPR), Specificity (TNR), False Positive Rate (FPR), and False Negative Rate (FNR)) are not affected by the prevalence of the disease of interest. Still, their values are intrinsic to the 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 that constitutes a disease. Moreover, sensitivity and specificity measures are transferrable from a sample population to other populations with different prevalence rates. Additionally, sensitivity and specificity are the correct classification rates of diseased and non-diseased people, correctly categorizing their actual states. They provide a holistic picture of a diagnostic test.

The ROC Curve was first introduced in the analysis of radar signals. Later on, it was employed in signal detection theory during World War II. It opened the door to new research to increase the prediction of correctly detecting Japanese aircraft from their radar signals after the attack on Pearl Harbor in 1941(Egan & Egan, 1975; Green & Swets, 1966). Later, ROC was applied to radiological, psychophysical, and epidemiological studies (Aoki, Watanabe, Furuichi, & Tsuda, 1997; Hsiao et al., 1989; Metz, 1989). ROC curve potentials in medical diagnostics were recognized as early as the 1960s (Lusted, 1960). Previous studies have systematically reviewed and illustrated the application and evaluation of diagnostic accuracy using ROC (Margaret Sullivan Pepe, 2003; Swets & Pickett, 1982; Zweig & Campbell, 1993).

Sometimes, it is not feasible to construct ROC, and a summary index becomes a critical measure to summarize its information. AUC is the most widely used summary statistic of ROC and is computed by taking the integral of ROC statistics from 0 to 1 (Fawcett, 2006). AUC is a global measure of diagnostic accuracy that summarizes the test's overall diagnostic accuracy, and it does not provide information about sensitivity and specificity as a summary index. Moreover, the ROC curve and AUC have no information about predictive values of rule-in or rule-out of a test in medical diagnostics.

The Youden index is another prevalent measure for binary classification in diagnostic accuracy. It is also a global measure, which Youden first proposed in 1950 (Youden, 1950). The Youden index (J) is a statistic that maximizes correct classification rates (i.e., sensitivity and specificity) and achieves the maximum discrimination between two stages of the disease. The Youden index also encounters the same issues as ROC and AUC as two diagnostics, with the same Youden index value having different sensitivity and specificity. It is mostly used to determine the overall performance, and it does not characterize the rule-in or rule-out information in diagnosis.

After estimating an optimal cut-point in clinical practice, we need to understand the implication of the results, such as how likely it would be that the test would provide 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). PPV of the test is defined as the proportion of subjects with positive test results who are correctly diagnosed (i.e., true positive results) (Fletcher, Fletcher, & Fletcher, 2012). Similarly, 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 PPV and 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 of the test increases as the underlying prevalence of the disease of interest increases, whereas the NPV decreases (Wong & Lim, 2011). When the sensitivity and specificity are fixed, the PPV of the test increases as the underlying prevalence of the disease of interest increases, whereas the NPV decreases (Wong & Lim, 2011). Similarly, the PPV of the test decreases as the underlying prevalence of the disease of interest decreases, whereas the NPV increases. Therefore, the PPV and the NPV of a population cannot be generalized to a different population.

LRs are another statistical tool to understand diagnostic tests. Compared to PPV and NPV, LRs do not depend on the prevalence of the disease, and LRs of the same diagnostic test can be generalized from one population to another population. Additionally, LRs provide information about the rule-in or rule-outs of a diagnostic test (Boyko, 1994; Deeks & Altman, 2004). Rule-in or rule-out tests are essential for different health care purposes. For example, the rule-in principle (specificity) is functional when a toxic treatment of the disease will be initiated if the diagnosis is confirmed, such as in the use of chemotherapy or combination chemotherapy for malignancies (Lee, 1999). The rule-out principle (sensitivity) is also helpful when there is a significant penalty for missing the disease, and the initial treatment is relatively safe, like in the use of screening tests for tuberculosis or hypothyroidism (Lee, 1999). LRs can be calculated for either positive or negative test results. It allows health care professionals to determine how much the utilization of a particular test will alter the probability. A positive LR tells how likely it will be that a diseased subject will receive a positive test result compared to a non-diseased subject. In contrast, a negative LR shows how likely it will be that a non-diseased subject will receive a negative test result compared to a diseased subject (A.-M. Šimundić, 2009). Note that these measures depend on the disease prevalence. However, these measures are sometimes not enough to evaluate the clinical consequences relative to other tests. The shortcoming of using these measures is that a test may have better sensitivity than another test but have worse specificity. In this case, we can use a benefit-risk approach where benefits and risks are put on the same scale to decide which test has better, worse, or about the same benefit-risk trade-off when considering the clinical consequences of a test.

G. Pennello, N. Pantoja-Galicia, and S. J. J. o. b. s. Evans (2016b) describes benefit-risk as another approach to determine whether a diagnostic test has better, worse, or the same outcomes when assessing a test's clinical consequences. Consequently, evaluating diagnostic tests based on benefit-risk involves both test accuracy and clinical implications of diagnostic errors. Diagnostic tests are commonly classified into two stages: either positive or negative for a clinical condition (diseased or non-diseased). A biomarker that can discriminate between subjects into diseased and non-diseased populations is efficient for diagnosing a disease. However, some diseases have distinct ordinal stages that existing measures cannot recognize in diagnostics. Dichotomizing biomarker to binary stages generally combines disease stages, resulting in the delay in diagnosing patients in the early stage. Failing to diagnose patients in the early stage of the disease will delay appropriate treatments and cause serious health problems in the future. Therefore, diagnosing 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 populations into different stages to provide better treatment strategies, such as the identification of mild cognitive impairment of Parkinson’s 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 of a disease is imperative for early clinical interventions, such as early interventions for breast cancer (Abe et al., 2005; Richards, Westcombe, Love, Littlejohns, & Ramirez, 1999). A decision-theoretic approach is vital for screening a population for disease and deciding whether to administer a preventive intervention with adverse or beneficial effects.

Moreover, some frontier studies propose measures generalizing from binary to multi-stage classification using the ROC, AUC, and Youden index (Nakas, Alonzo, & Yiannoutsos, 2010; Nakas, Dalrymple-Alford, Anderson, & Alonzo, 2013; Brian K. Scurfield, 1996; Brian K Scurfield, 1998; Xiong, van Belle, Miller, & Morris, 2006). A clinical utility study can be another approach to evaluate clinical outcomes, which can be improved when the test influences subject management. But clinical utility studies can be expensive to conduct, take lengthy follow-up time on subjects, and be challenging to design. A poorly designed clinical utility study can be inefficient and may not even permit these studies to reveal whether the evaluation of a test affects clinical consequences (Bossuyt, Lijmer, & Mol, 2000; Hoering, Leblanc, & Crowley, 2008; Simon, 2010). Pennello et al. (2016b) use the benefit-risk approach based on a decision-theoretic framework to compare diagnostic tests for binary classification as a new measure for diagnostic accuracy, suggesting it is a better disease diagnostic procedure, in some cases, compared to the other existing measures. This theory indicates a weighted accuracy measure, proposed previously by Evans, which can be interpreted as a relative utility measure, with expected utility close to that for a perfect test (sensitivity = specificity = 1) (Evans et al., 2016). The comparisons of binary-valued tests based on benefit-risk evaluation (Pennello et al., 2016a) are similar to recent work in the review of new markers for risk prediction (S. G. Baker et al., 2009; S. G. Baker & Kramer, 2012; S. G. Baker et al., 2012; S. G. J. J. J. o. t. N. C. I. Baker, 2009; Gail & Pfeiffer, 2005; Vickers & Elkin, 2006) and diagnosis (Margaret S Pepe et al., 2016). For example, the relative utility curve (S. G. Baker et al., 2009) is a fraction of the expected net benefit of perfect prediction, and a risk prediction model obtains it at the optimal cut point. Similarly, weighted accuracy (Evans et al., 2016) is a fraction of the expected utility of a perfect diagnostic test that is obtained by an investigational diagnostic test.

Emerging studies of diagnostic accuracy for multi-stage diseases show high demand for developing a more reliable diagnostic procedure to discriminate among subjects in different diseased stages accurately (Attwood, Tian, & Xiong, 2014; Li & Fine, 2008; Nakas et al., 2010; Nakas et al., 2013; Xiong et al., 2006). For example, some chronic diseases, such as Alzheimer’s disease, kidney disease, and cancers (prostate, lung, colorectal, and ovarian), have more than two stages in nature and require measures that can identify subjects among stages (Alzheimer's Association, 2019; Johns Hopkins Medicine, 2019a). Ovarian cancer ranks the fifth leading cause of cancer death among women in developed countries (Chudecka-Głaz, 2015). It generally presents in advanced stages with a high case fatality ratio (CFR) but has favorable survival rates if diagnosed earlier. Additionally, clinical symptoms are not well manifested in the early stages of the disease, resulting in late diagnosis and poor prognosis (Cramer et al., 2011). Some traditional binary tests cannot directly be used for multi-stage diseases. However, some popular measures can be extended and are generalized to the multi-stage setting.

To my knowledge, no studies have investigated accuracy measures and the clinical consequences of medical diagnostic test errors in multi-stage disease settings. Also, the clinical implications of treating or not treating patients at a different stage of the disease have different benefit-risk consequences. Therefore, this dissertation intends to expand the net benefit approach for evaluating medical diagnostic tests for a multi-stage clinical condition. And consequently, this study will provide additional interpretation, using the net benefit approach for rule-in or rule-out clinical conditions and their adverse consequences from unnecessary workups in multi-stage diseases.

CHAPTER 3

METHODS

This chapter provides an overview of some related methods for evaluating the benefit-risk of a binary diagnostic test based on its diagnostic accuracy from a clinical performance study and external information related to clinical consequences (see Pennello et al. (2016a)).

3.1 Introduction

For most medical diagnostic testing, biomarkers are dichotomized to classify subjects in a binary manner, either positive or negative for a clinical condition (positive for diseased or negative for non-diseased). The test is evaluated for its diagnostic accuracy by comparing test negative and positive results (T = 0,1) for agreement with the absence and presence of the clinical condition (D = 0,1), as determined by a clinical reference standard or best available method.

Pennello et al. (2016a) indicate that comparing diagnostic tests based on accuracy alone could be inconclusive. They propose that comparing tests based on benefit-risk may be more conclusive because clinical consequences of diagnostic error are considered. For benefit-risk evaluation, they present diagnostic yield as the expected distribution of subjects with true positive, false positive, true negative, and false-negative test results in a hypothetical population. They construct a table of diagnostic yield, which indicates the number of false-positive subjects experiencing adverse consequences from unnecessary workup. Then they develop a decision theory for evaluating tests. Pennello et al. (2016a) argue that their progressive approach provides additional interpretation to quantities in the diagnostic yield table. This approach also indicates that the expected utility of a test relative to a perfect test is a weighted accuracy measure. The average sensitivity and specificity weighted for prevalence and relative importance of false-positive and false-negative testing errors are also interpretable as the cost-benefit ratio of treating non-diseased and diseased subjects. These researchers also propose plots of diagnostic yield, weighted accuracy, and relative net benefit of tests as functions of prevalence or cost-benefit ratio. For example, they illustrate these concepts with hypothetical screening tests for colorectal cancer, with positive test subjects referred to colonoscopy.

Furthermore, Pennello et al. (2016a) argue that the benefit-risk evaluation of a diagnostic test involves not just the accuracy of the test but the clinical consequences of diagnostic error. They explain that assessing the clinical implications of false positive and false negative test errors depends on the clinical setting, the intended use of the test, and the population on whom it will be used. Sometimes, the test itself has clinical consequences. For example, these consequences may involve an invasive procedure, such as a biopsy, or introducing energy into the body as with radiation, using an X-ray. These researchers argue that a health economics analysis might also consider the cost of testing and downstream costs of working up positive test subjects (Tsalik et al., 2016).

A diagnostic test classifies subjects as either positive or negative for a clinical condition (e.g., disease absence or presence). Also, a diagnostic test predicts a future binary state (e.g., susceptibility or resistance of a microbe to an antimicrobial drug). Diagnostic accuracy is evaluated in a clinical performance study for its classification accuracy (e.g., specificity, sensitivity, negative and positive likelihood ratio) or its predictive accuracy (e.g., negative and positive predictive value – NPV, PPV). However, these evaluations can sometimes be insufficient for examining the clinical consequences of the test relative to other tests. For example, one test may have better sensitivity but actually have worse specificity than another test. Thus, based on such accuracy measures alone, a determination of whether a test has better, worse, or about the same benefit-risk tradeoff as another test can be equivocal.

A clinical utility study should be performed to evaluate the clinical consequences of a test in order to show whether clinical outcomes can be improved when the test is used in order to influence subject management. However, there could be limitations to performing clinical utility studies. They could be expensive, require lengthy subject monitoring, and design could be complex. A poorly designed clinical utility study can be inefficient and may not even permit an evaluation of test effect on clinical outcome (Bossuyt et al., 2000; Hoering et al., 2008; Simon, 2010). Additionally, such clinical utility data are usually not available to a regulatory agency when deciding whether or not to approve a test for the market.

3.2. Test Accuracy

Pennello et al. (2016a) describe hypothetically comparing a standard test (S) and a new test (T) used to screen for colorectal cancer (CRC) as follows: consider a new diagnostic test that indicates subjects as test negative or positive for a clinical condition, (e.g., disease). The test is evaluated for its diagnostic accuracy by comparing test negative and positive results (T=0, 1) for agreement with the absence and presence of the clinical condition (D = 0, 1), as determined by a clinical reference standard or best available method. These researchers consider comparing the new test with a standard test, indicating subjects as negative or positive (S=0, 1). They provide the following example: the new test has better sensitivity (0.90 vs. 0.75) but worse specificity (0.85 vs. 0.95) than the standard. However, one of the tests could still be declared better than the other if its negative and positive predictive values (NPV, PPV) are better. PPV is monotone since it increases the positive diagnostic likelihood ratio PLR = Se/(1−Sp). NPV is monotone, decreasing the negative diagnostic likelihood ratio NLR = (1 − Se)/Sp. Thus, the new test would have better NPV and PPV for the same prevalence as the standard test if its NLR is smaller and its PLR are larger. Pennello et al. (2016a) suggest using the following graph that shows an example of visual interpretation of the likelihood ratio graph (Figure 3.1), a helpful display proposed originally by Biggerstaff (2000). The graph has the same axes as the ROC plot. The coordinate of both the true and false-positive fractions of the standard test is plotted in the graph, with two lines drawn through it to the points (0,0) and (1, 1). The slope of the lines through (0,0) and (1,1) are PLR and NLR, respectively. These two lines define four regions in which the coordinate of the new test could lie. In this case, the new test falls in region A, indicating that it is better at detecting the absence of CRC than the standard test. However, it is worse than the standard test at detecting the presence of CRC because, respectively, its PLR is worse (smaller), while its NLR is better (smaller). In summary, evaluating which test is better based on test accuracy alone is equivocal.

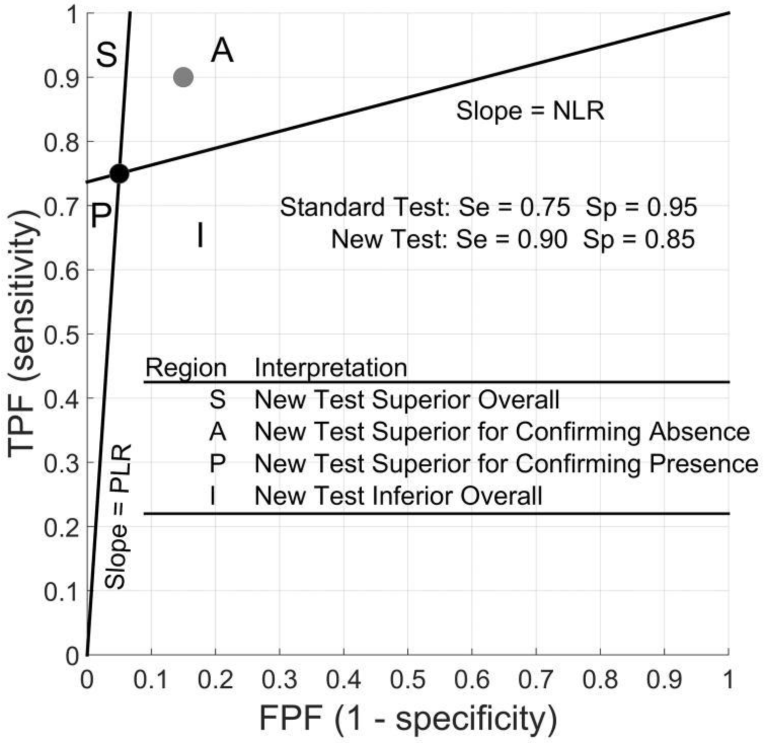


Figure 3.1. *Likelihood Ratio Graph: Regions of Comparison (source: Pennello et al., 2016)*

3.3 Diagnostic Yield

The measures of diagnostic accuracy are all based on conditional probabilities. Classification accuracy measures (Se, Sp, NLR, PLR) are based on the probabilities of test results conditional on disease status. In contrast, measures of predictive accuracy (NPV, PPV) are based on probabilities of disease status, conditional on the test result. Likewise, Pennello et al. (2016a) indicate that they consider their diagnostic yield table to compare two tests based on benefit-risk, which can provide knowledge on the clinical significance of a test and relate directly to formal decision-theoretic evaluations of the benefit-risk. They consider the distribution of false-negative (FN), true positive (TP), true negative (TN), and false-positive (FP) results in the screening population as the joint probability distribution of disease status and test results.

Table 3.1. *Equivalent Loss and Utility Functions*

(Loss due to the act of testing is assumed to be 0)

Loss Functions		
Test	**D=0**	**D =1**
T=0	r_{00}	r_{01}
T=1	r_{10}	r_{11}
Utility Functions		
Test	**D=0**	**D =1**
T=0	$-r_{00}$	$-r_{01}$
T=1	$-r_{10}$	$-r_{11}$

The parameters of the diagnostic yield table (3.1) designated by joint probabilities' distribution are then given as

$$\mathbf{Y} = \begin{array}{cc} & \begin{array}{cc} D=0 & D=1 \end{array} \\ \begin{array}{c} T=0 \\ T=1 \end{array} & \begin{bmatrix} \rho_0\tau_{00} & \rho_1\tau_{01} \\ \rho_0\tau_{10} & \rho_1\tau_{11} \end{bmatrix} \end{array} , \qquad (3.1)$$

with Disease D= d*=0 or 1*(disease negative or disease positive), Test T=*t=0 or 1* (test negative or test positive), Prevalence $\rho_d = \Pr(D = d)$ and $\tau_{td} = \Pr(T = t \mid D=d)$.

In general, when comparing the diagnostic yield tables of some tests, we can quantify the number of FP subjects harmed from unnecessary additional workup involving an invasive procedure (e.g., colonoscopy) and the number of FN subjects harmed by lack of further workup. The harm associated with an FN result includes not receiving necessary treatment for a disease, which then may progress unattended. The disease is typically aggressive in some settings, and all FN subjects are harmed by lack of detection. In other settings, the disease is typically slowly progressing, and harm from delay in detection may be weighed against competing risks, as may occur with older men with early-stage prostate cancer, who may die of other causes (Pennello, 2016).

Based on test accuracy alone, the new test may seem more worthwhile than a standard test because of its superior sensitivity to detecting the disease. Yet its inferior specificity of the new test may have clinical consequences for many subjects in the intended use population who will falsely test positive. The diagnostic yield table facilitates a quantitative discussion of questions related to these consequences.

3.4 Decision-Theoretic Evaluation

A diagnostic yield table and plots (Table 3.1) provide information about the clinical significance of test results. This information will directly impact formal decision-theoretic evaluations of benefit-risk.

3.4.1 Expected Loss

If d= 0, 1 indicates disease absence, presence, and t= 0, 1 indicates the binary test results (negative, positive), Pennello et al. (2016) express the loss of a test as follows,

$$L(t,d) = L_0(d) + t(L_1(d) - L_0(d)), \qquad (3.2)$$

where, $L_0(d)$ and $L_1(d)$ are the losses ascribed to negative and positive test results, respectively. This is related to the incorrect classification of stages 0, 1, and 2 of a disease condition. We consider ascribing a loss r_{td} to the binary test result T = t on a subject with disease state D = d with d= 0, 1 (absent and present disease stages), and t = 0, 1 (negative and positive test result). General loss functions follow;

$$L_0(d) = (1-d)r_{00} + dr_{01}$$

$$L_1(d) = (1-d)r_{10} + dr_{11},$$

With the loss of a test (3.2)

$$\begin{aligned} L(t,d) &= L_0(d) + t[(1-d)(r_{10} - r_{00}) - d(r_{11} - r_{01}))] \\ &= L_0(d) + t[(1-d)C - dB], \end{aligned} \qquad (3.3)$$

$$L(t,d) = B\{L_0(d)/B + t[(1-d)r - d]\} \qquad (3.4)$$

In expression (3.3), the loss depends on test result t only through C and B, with $B = r_{01} - r_{11}$ and $C = r_{10} - r_{00}$. In expression (3.4), the loss function is proportional to one, which depends on test result *t* only through *r*, where $r = \frac{C}{B} = FP : FN$ loss ratio=TN:TP utility ratio.

3.4.2 Expected Utility

Losses $L_0(d)$ and $L_1(d)$ are defined based on incorrect binary test classifications (false negative, false positive), and utilities $U_0(d)$ and $U_1(d)$ are credited to correct test classifications (true negative, true positive). Pennello et al. (2016a) present the "utility" of the test as:

$$\begin{aligned} U(t,d) &= (1-t)U_0(d) + tU_1(d) \\ &= U_0(d) + t(U_1(d) - U_0(d)). \end{aligned} \qquad (3.5)$$

The utility function is simply the negative of the loss. So, the expected loss is the negative of the expected utility ($U_1(d) - U_0(d) = L_0(d) - L_1(d)$). This equality occurs if, for instance $U_1(d) = L_0(d) = d$ and $U_0(d) = L_1(d) = (1-d)r$. Thus, r is interpretable as the relative loss ratio of false-positive to false-negative test results and the relative utility ratio of true negative to true positive test results. Under these utility functions, Pennello et al. (2016) express the utility functions as

$$U(t,d) = (1-t)(1-d)r + td, \qquad (3.6)$$

and hence the expected utility for the test as

$$E \equiv EU(t,d) = \rho_0 \tau_{00} r + \rho_1 \tau_{11} \qquad (3.7)$$

Upon examination of utility and loss functions, respectively, B and C have been interpreted as the overall net benefit and net cost of treating test positive subjects with and without disease, respectively (S. G. Baker et al., 2012; Pauker & Kassirer, 1975; Margaret S Pepe et al., 2016; Vickers & Elkin, 2006). Upon examining the utility function, r may be interpreted as the TN: TP utility ratio and the FP: FN loss ratio.

3.4.3 Net Benefit

The net benefit of a test compared with a random test with test positive probability τ is defined as the difference in expected utility between the test (E) and the random test (E_τ).

$$NB_\tau = E - E_\tau \qquad (3.8)$$

The net benefit of the test over never treating a subject is NB_0 a difference in expected utility from the always negative test. The net benefit over always treating a subject is NB_1 a difference in expected utility from the always positive test. The relative net benefit of a test is defined as

$$RNB_\tau = (E - E_\tau) / (E_{perf} - E_\tau) \qquad (3.9)$$

which scales the net benefit to have a maximum of 1 relative to the net benefit of a perfect test. The following graphs (Figure 3.2 and Figure 3.3) show a visual interpretation of the relative net benefit over never treat and always treat policies as a function of relative importance ratio r, which provide an overall comparison of the new and standard tests (Pennello et al., 2016a). Findings from these plots indicate that relative net benefit over the never treat policy is noticeably worse for the new test than the standard test over an extensive range of r values. However, the relative net benefit over the always treat policy is slightly worse than the standard test over an extensive range of r values. Thus, the two tests can be considered comparable in settings where prophylactic treatment is practiced in place of testing.

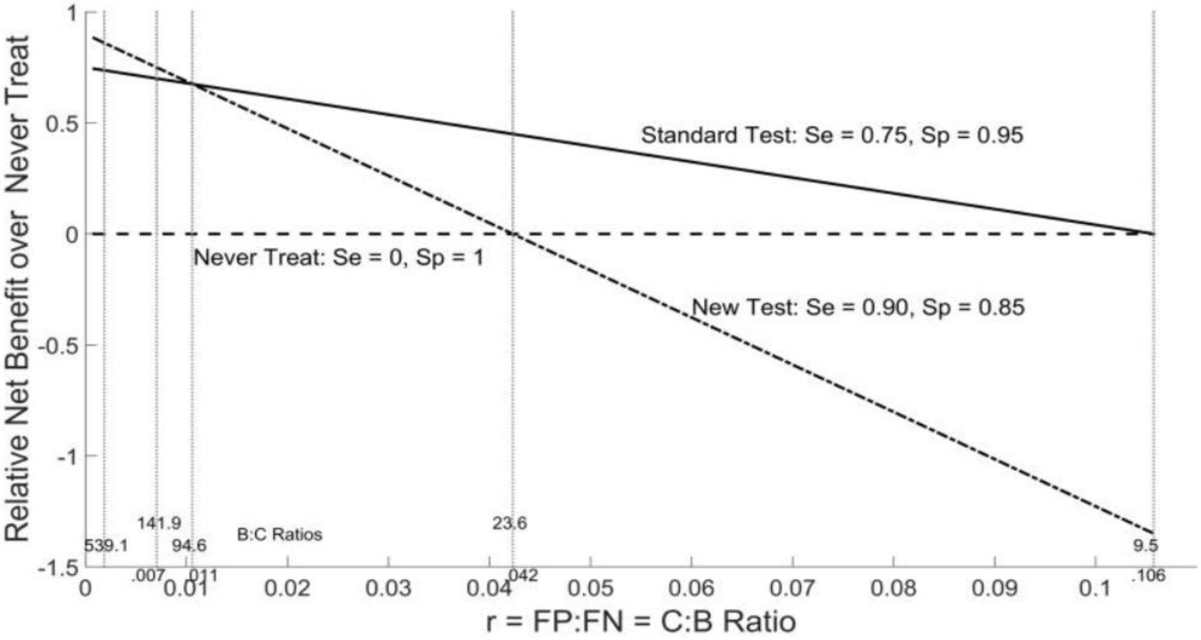


Figure 3.2. *Relative Net Benefit over Never Treat by Cost-Benefit Ratio r (source: Pennello et al., 2016)*

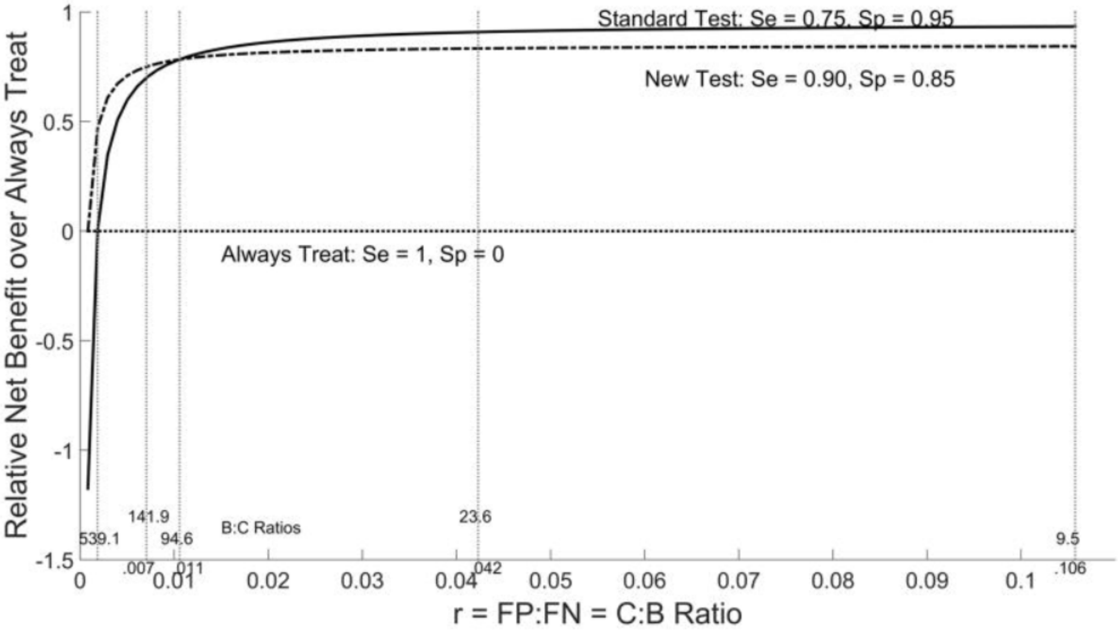


Figure 3.3. *Relative Net Benefit over Always Treat by Cost-Benefit Ratio r (source: Pennello et al., 2016)*

3.4.4 Choosing *r*

Pennello et al. (2016a) indicate that, at a minimum, the expected utility of the new test should be greater than the expected utility of any non-informative test that renders a positive test result at random with probability $\tau(0 \le \tau \le 1)$. The difference in the expected utility of a test compared with a random test is called net benefit. The test has a positive net benefit compared with any random test if the FP: FN relative importance ratio $r < \frac{P_1}{(1-P_1)} \equiv \theta_1$.

It is used where $\mathrm{p}_t = \Pr(D=1|T=t)$ is the predictive value of test results $T=t$ for the disease. In other words, the test is valid (better than any random test) only for choices when $r < \theta_1$.

The note $\theta_1 = \frac{\rho_1 \tau_{11}}{\rho_0(1-\tau_{00})}$ is the reciprocal of the FP to TP ratio. Thus, information should be acquired based on choices of r that are acceptable for the test. Equivalently,

$$\left[(\tau_{11}/(1-\tau_{00}) > (r/\theta)\right],$$

Where, $\theta_1 = \frac{\rho_1}{\rho_0}$ is the pre-test odds of disease. Noting that a test is informative only if the ratio of its true to false-positive fraction $(\tau_{11}/1-\tau_{00}) <$ 1, we find that $r > \theta$ is an additional constraint on valid choices of r. Thus, in terms of reducing expected loss relative to the trivial test, the new test is valid only for FP: FN loss ratios

$$r \in (\theta, \theta_1)$$

Thus, $\theta_1 = \frac{\rho_1 \tau_{11}}{\rho_0(1-\tau_{00})}$ is the reciprocal of the FP to TP ratio. Also, θ is the reciprocal of the FP to TP ratio for a trivial test that classifies everyone as test positive.

This dissertation proposes extending the net benefit approach of evaluating diagnostic tests to multi-stage clinical conditions. This new approach uses the diagnostic yield table, all the classification information, and both correct and incorrect classification probabilities. Also, this approach aims to demonstrate the application of the net benefit approach to evaluating diagnostic tests for multi-stage clinical conditions based on their diagnostic accuracy from a clinical performance study, along with external information on clinical consequences. More details of the proposed criterion in the multi-stage setting are discussed in Chapter 4.

CHAPTER 4

NET-BENEFIT APPROACH FOR COMPARING DIAGNOSTIC TESTS OF MULTI-STAGE DISEASES

This dissertation proposes extending the net benefit approach of evaluating diagnostic tests to multi-stage clinical conditions (*k>2*). Consequently, I extend the diagnostic yield table presented by Pennello et al. (2016a) to multi-stage clinical conditions. I develop a decision theory based on net benefit for evaluating diagnostic tests that provide additional interpretation for rule-in or rule-out clinical needs and their adverse consequences from unnecessary workup in multi-stage diseases.

4.1 Introduction and Preliminaries

As in Samawi et al. (2021), we define a class of probabilities p_{bd} $b=0,1,...,k-1; d=0,1,...,k-1$ for classifying a randomly selected subject in the b^{th} test class, given the subject is in the d^{th} stage of the disease (In general, when the number of the test outcomes is equal to the number of the disease stages). Based on the continuous biomarker *X*, cut-points $\mathbf{c}'=(c_{-1}=-\infty, c_0, c_1, ..., c_{k-2}, c_{k-1}=\infty)$ are needed to diagnose the disease's k-stage (Patients with biomarker values within the range of $c_{b-1}<X_d\le c_b, b=0,1,...,k\;-1,$ diagnosed as stage *b*). Let X_d denote the d^{th} disease stage (determined by the gold standard) with pdf and CDF $f_d(x)$, $F_d(x)$ respectively. Then τ_{bd} defines as

$$\tau_{b,d}=P(c_{b-1}<X_d\le c_b \mid D=d)=F_d(c_b)-F_d(c_{b-1})=P(T=b \mid D=d), \qquad (4.1)$$

$b=0,1,...,k-1; d=0,1,...,k-1$where T is the random variable of the test results, and D is the random variable for the disease stage's true classification. Now we can define the probability classification matrix as follows:

$$\mathbf{P}=\begin{array}{c}\begin{array}{cccc}\mathrm{D}=0 & D=1 & ... & D=k-1\end{array}\\ \left[\begin{array}{cccc}\tau_{00} & \tau_{01} & \cdots & \tau_{0(k-1)}\\ \tau_{10} & \tau_{11} & \cdots & \tau_{1(k-1)}\\ . & . & . & .\\ . & . & . & .\\ . & . & . & .\\ \tau_{(k-1)0} & \tau_{(k-1)1} & \cdots & \tau_{k-1(k-1)}\end{array}\right]\begin{array}{c}T=0\\ T=1\\ .\\ .\\ .\\ T=k-1\end{array}\end{array}. \qquad (4.2)$$

Noting that $\sum_{b=1}^{k-1}\tau_{bd}=1, d=0,1,2,...,k-1.$ Also, let $\rho_d=p(D=d); d=0,1,2,...,k-1$be the prevalences of the dth disease stage.

To show how to calculate the classification matrix (4.2), when we have a continuous biomarker, we introduce the case for the number of three stages of disease condition to simplify the discussion. When cut-points are specified, one can use the empirical or kernel smoothing approach to estimate the distribution function to find the classification matrix (4.2). For *k=3*, the classification matrix in (4.2) reduces to

$$\mathbf{P}=\begin{array}{c}\begin{array}{ccc}D=0 & \mathrm{D=1} & \mathrm{D=2}\end{array}\\ \left[\begin{array}{ccc}\tau_{00} & \tau_{01} & \tau_{02}\\ \tau_{10} & \tau_{11} & \tau_{12}\\ \tau_{20} & \tau_{2,1} & \tau_{22}\end{array}\right]\begin{array}{c}T=0\\ T=1\\ T=2\end{array}\end{array}=\left[\begin{array}{ccc}F_0(c_1) & F_1(c_1) & F_2(c_1)\\ F_0(c_2)-F_0(c_1) & F_1(c_2)-F_1(c_1) & F_2(c_2)-F_2(c_1)\\ 1-F_0(c_2) & 1-F_1(c_2) & 1-F_2(c_2)\end{array}\right].$$

4.2 Three-stage Diseases Diagnostic Yield Tables

Following Pennello et al.'s (2016) definitions, we define the diagnostic yield table, which can provide knowledge about the clinical significance and relate these findings directly to formal decision-theoretic evaluations of the benefit-risk. I assign losses to misclassification or utilities to correct classification as in table 4.1.

Table 4.1. *Equivalent Loss and Utility Functions*

(Loss due to the act of testing is assumed to be 0)

Test	**D=0**	**D =1**	**D =2**
T=0	r_{00}	r_{01}	r_{02}
T=1	r_{10}	r_{11}	r_{12}
T=2	r_{20}	r_{21}	r_{22}
Test	**D=0**	**D =1**	**D =2**
T=0	$-r_{00}$	$-r_{01}$	$-r_{02}$
T=1	$-r_{10}$	$-r_{11}$	$-r_{12}$
T=2	$-r_{20}$	$-r_{21}$	$-r_{22}$

However, the parameters of the diagnostic yield table (4.1) designated by the joint probabilities' distribution are then given as

$$\mathbf{Y} = \begin{array}{r} \\ T=0 \\ T=1 \\ T=2 \end{array} \begin{array}{c} \begin{array}{ccc} D=0 & D=1 & D=2 \end{array} \\ \begin{bmatrix} \rho_0\tau_{00} & \rho_1\tau_{01} & \rho_2\tau_{02} \\ \rho_0\tau_{10} & \rho_1\tau_{11} & \rho_2\tau_{12} \\ \rho_0\tau_{20} & \rho_1\tau_{21} & \rho_2\tau_{22} \end{bmatrix} \end{array}. \qquad (4.3)$$

I have two major categories for the test *t=0 or 1* (test negative or test positive for any stage of the disease). Also, I have two major categories: disease d*=0 or 1* (absence or presence, for any stage of the disease). I can identify the loss functions

1- $L_0(d) = (1-d)r_{00} + dr_{01} + dr_{02}$

2- $L_1(d) = (1-d)r_{10} + dr_{11} + dr_{12}$

3- $L_2(d) = (1-d)r_{20} + dr_{21} + dr_{22}$.

Similarly, I can extend the diagnostic yield tables to k-stage diseases.

4.3 Expected Loss Function

As in Pennello et al. (2016), we will express the loss of a test as follows.

$$L(t,d) = L_0(d_1) + t(L_1(d) + L_2(d) - L_0(d)), \qquad (4.4)$$

which is related directly to the incorrect classification of stages 0, 1, and 2 diseases. I consider ascribing a loss r_{td} to the three-stage disease test result T = t on a subject with disease state D = d where d= 0, 1 (absent and present disease stages), and t = 0, 1 (negative and stage 1 or 2 of a disease [positive test result]).

Therefore, from the table (4.1), I can get

$$\begin{aligned} L(t,d) &= L_0(d) + t[L_1(d) + L_2(d) - L_0(d)] \\ &= L_0(d) + t[(1-d)r_{10} + dr_{11} + dr_{12} + (1-d)r_{20} + dr_{21} + dr_{22} \\ &\quad -((1-d)r_{00} + dr_{01} + dr_{02})] \\ &= L_0(d) + t[(1-d)(r_{10} + r_{20} - r_{00}) + d(r_{11} + r_{12} + r_{21} + r_{22} - r_{01} - r_{02})] \\ &= L_0(d) + t[(1-d)(r_{10} + r_{20} - r_{00}) - d(r_{01} + r_{02} - (r_{11} + r_{12} + r_{21} + r_{22}))] \\ &\propto t(1-d)r - d]. \end{aligned} \qquad (4.5)$$

Thus, the cost to benefit ratio is given by

$r = \dfrac{C}{B} = \dfrac{(r_{10} + r_{20} - r_{00})}{(r_{01} + r_{02} - (r_{11} + r_{12} + r_{21} + r_{22}))} = FP : FN$ loss ratio=TN:TP utility ratio,

C =net cost (harm) of treating a subject without disease with stage one or two treatments,

B =net benefit of using stage 1 or 2 treatments to treat a subject at stage 1 or 2 of the disease. This ratio is for rule-out patients, and then $r = r^{RO}$. To distinguish between loss and utility ratio, we have $r_L^{RO} = \dfrac{r_{10} + r_{20}}{r_{01} + r_{02}}$ and $r_U^{RO} = \dfrac{r_{00}}{r_{11} + r_{12} + r_{21} + r_{22}}$.

Traditionally, clinicians choose the most sensitive diagnostic test to rule-out disease and the most specific diagnostic test to rule-in disease. In this dissertation, following Pennello (2019) (ENAR presentation for binary screening tests), I will examine the validity of these recommendations concerning the expected loss or expected utility of clinical consequences of diagnostic error in multi-stage clinical conditions. I must simultaneously capture the tradeoffs between sensitivity to each stage of the disease, specificity, disease probability, and utilities of correct and incorrect disease classifications by the diagnostic test to determine which strategy minimizes expected loss or maximizes expected clinical utility.

4.3.1 Expected Loss Function for Rule-In

In general, 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 stage 1 or 2) from a specific test (high correct classifications of stage 0 (non-diseased)) makes the presence of the disease (at stage 1 or 2) more likely since it is specific to that disease.

The counts in the diagnostic yield tables are the products of joint probabilities in (4.3) by N (population size), using tables 4.1 and (4.3), and for the rule-in patient where $r_L = \dfrac{1}{r_L^{RO}}$

$$\begin{aligned} E_L^{RI} &= E\ L(t,d) = r_L \rho_0 \tau_{10} + r_L \rho_0 \tau_{20} - \rho_1 \tau_{11} - \rho_2 \tau_{22} - \rho_1 \tau_{21} - \rho_2 \tau_{12} \\ &= r_L \rho_0 (\tau_{10} + \tau_{20}) - \rho_1 (\tau_{11} + \tau_{21}) - \rho_2 (\tau_{22} + \tau_{12}) \end{aligned}$$

[Note: Only applicable terms are $(t,d) = (1,0),\ (1,1)$].

4.3.2 Expected Loss Function for Rule-Out

Similarly, based on correct classifications of the stage (1 or 2), a rule-out test emphasizes assessing if test results will exclude the possibility that a subject is non-diseased or not in the lower stage.

I consider ascribing a loss r_{td} to the three-stage disease test result T = t on a subject with disease state D = d where d= 0, 1 (absent and present disease stages), and t = 0, 1 (negative and correct stage 1 and 2 of the diseases [positive test result]).

Hence, from the table (4.5), I can get

$$L(t,d) \propto t[(1-d) - r_L^{RO} d] \qquad (4.6)$$

where $r^{RO} = \frac{1}{r_L}$, and r_L are defined above, and

$$E_L^{RO} = E\ L(t,d) = \rho_0(\tau_{10}+\tau_{20}) - r_L^{RO}\{\rho_1(\tau_{11}+\tau_{21}) + \rho_2(\tau_{22}+\tau_{12})\}$$

[Note: Only applicable terms are $(t,d) = (1,0),\ (1,1)$].

Therefore, I can find the expected loss of rule-in and rule-out as follows:

$$E_L = \begin{cases} E_L^{RI} = r\rho_0(\tau_{10}+\tau_{20}) - \rho_1(\tau_{11}+\tau_{21}) - \rho_2(\tau_{22}+\tau_{12}), \\ E_L^{RO} = \rho_0(\tau_{10}+\tau_{20}) - r^{RO}\{\rho_1(\tau_{11}+\tau_{21}) + \rho_2(\tau_{22}+\tau_{12})\}. \end{cases} \quad (4.7)$$

Finally, (4.7) can be generalized for k-stages diseases as follows:

$$E_{GL} = \begin{cases} E_{GL}^{RI} = r_L \sum_{i=1}^{k-1} \rho_0 \tau_{0i} - \sum_{i=1}^{k-1} \rho_i (1-\tau_{0i}) = r_L \sum_{i=1}^{k-1} \rho_0 \tau_{0i} - \sum_{i=1}^{k-1} \rho_i \sum_{j=1}^{k-1} \tau_{ji}, \\ E_{GL}^{RO} = \sum_{i=1}^{k-1} \rho_0 \tau_{0i} - r_L^{RO} \sum_{i=1}^{k-1} \rho_i (1-\tau_{0i}) = \sum_{i=1}^{k-1} \rho_0 \tau_{0i} - r_L^{RO} \sum_{i=1}^{k-1} \rho_i \sum_{j=1}^{k-1} \tau_{ji}. \end{cases}$$

4.4 Expected Utility Function

Utility function tables are simply the negative of loss tables, assuming no testing cost. Like in Pennello et al. (2016), I express utility functions as

$$\begin{aligned} U(t,d) &= (1-t)U_0(d) + t(U_1(d) + U_2(d)) \\ &= U_0(d) + t([U_1(d) + U_2(d)] - U_0(d)) \end{aligned} \qquad (4.8)$$

which are related directly to correct test classifications (true negative, true positive in stages 1 and 2). Since *t* is in the right term of (4.5) and (4.8), then modulo a constant expected loss is the negative of the expected utility if $[(L_1(d)+L_2(d))-L_0(d)]=[U_0(d)-(U_1(d)+U_2(d)]$. This equality exists because, for example

$[U_1(d)+U_2(d)]=L_0(d)=d$ and $U_0(d)=[L_1(d)+L_2(d)]=(1-d)r.$

4.4.1 Expected Utility Functions for Rule-In

I define utility functions for rule-in as follows: from (4.8), we can get

$$U(t,d)=(1-t)(1-d)r_U+td\,,$$

where $r_U=\frac{1}{r_U^{RO}}$, and hence the expected utility is given by

$$\begin{aligned}EU(t,d)&=E[(1-t)(1-d)r_U+td]\\&=r_U\rho_0\tau_{00}+\rho_1(\tau_{11}+\tau_{21})+\rho_2(\tau_{22}+\tau_{12}),\end{aligned}\qquad(4.9)$$

[Note: Only applicable terms are $(t,d)=(0,0),\ (1,1)$].

4.4.2 Expected Utility Function for Rule-Out

Similarly, I define utility functions for rule-out as follows:

$$\begin{aligned}U_0(d)&=(1-d)\\U_1(d)+U_2(d)&=r_U^{RO}d.\end{aligned}$$

Therefore,

$$E\ U(t,d)=\rho_0\tau_{00}+r_U^{RO}[\rho_1(\tau_{11}+\tau_{21})+\rho_2(\tau_{22}+\tau_{12})],\quad(4.10)$$

[Note: Only applicable terms are $(t,d)=(0,0),\ (1,1)$].

Note that $r_U^{RO}=\frac{1}{r_U}$.

Therefore, I have the expected utility of rule-in and rule-out as follows

$$E_U=\begin{cases}E_U^{RI}=r_U\rho_0\tau_{00}+\rho_1(\tau_{11}+\tau_{21})+\rho_2(\tau_{22}+\tau_{12}),\\E_U^{RO}=\rho_0\tau_{00}+r_U^{RO}[\rho_1(\tau_{11}+\tau_{21})+\rho_2(\tau_{22}+\tau_{12})].\end{cases}\qquad(4.11)$$

Similarly, (4.11) can be generalized for the k-stage diseases as follows:

$$E_{GU}=\begin{cases} E_{GU}^{RI}=r_U\rho_0\tau_{00}+\sum_{i=1}^{k-1}\rho_i(1-\tau_{0i})=r_U\rho_0\tau_{00}+\sum_{i=1}^{k-1}\rho_i\sum_{j=1}^{k-1}\tau_{ji}, \\ E_{GU}^{RO}=\rho_0\tau_{00}+r_U^{RO}\sum_{i=1}^{k-1}\rho_i(1-\tau_{0i})=\rho_0\tau_{00}+r_U^{RO}\sum_{i=1}^{k-1}\rho_i\sum_{j=1}^{k-1}\tau_{ji}. \end{cases}$$

4.5 Expected Relative Net Benefit

4.5.1 Relative Net Benefit of Rule-In

Based on table 4.1, I can see that evaluating expected loss can be equivalently defined as a problem of assessing expected utility. I can say that utilities are set to the negatives of the losses to define utility functions corresponding to the three equivalent loss functions. Upon examining the utility and loss tables, $r=FP:FN$ loss ratio=TN:TP utility ratio, r has been interpreted as the overall net benefit and net cost of treating test positive subjects with and without any stage of the disease. Under the utility function, the expected utility for the random test is

$E=$ Expected utility of a test

$E_\tau=$ Expected utility of a random test with

$\text{P}(R=0)=1\text{-}\tau,\ P(R=1)=\tau_1, P(R=2)=\tau_2;\tau=\tau_1+\tau_2$

$E_p=$ Expected utility of perfect test ($\tau_{11}=1,\tau_{22}=1,\tau_{00}=1$).

Therefore, the expected net benefit for rule-in is given by

$NB_\tau=E-E_\tau$, where $E=r_U\rho_0\tau_{00}+\rho_1(\tau_{11}+\tau_{21})+\rho_2(\tau_{22}+\tau_{12})$ and $E_\tau=r_U\rho_0(1-\tau)+\tau_1\rho_1+\tau_2\rho_2$,

resulting in

$$\begin{aligned} NB_\tau &= r_U\rho_0\tau_{00}+\rho_1(\tau_{11}+\tau_{21})+\rho_2(\tau_{22}+\tau_{12})-[r_U\rho_0(1-\tau)+\tau_1\rho_1+\tau_2\rho_2] \\ &= r_U\rho_0(\tau_{00}-1+\tau)+\rho_1(\tau_{11}+\tau_{21}-\tau_1)+\rho_2(\tau_{22}+\tau_{12}-\tau_2) \end{aligned} \qquad (4.12)$$

For rule-in $\tau = 0 \Rightarrow \tau_1 = \tau_2 = 0$ (so we do not send the patient to treatment), then I have

$$NB^{RI} = NB_0 = \rho_1(\tau_{11} + \tau_{21}) + \rho_2(\tau_{22} + \tau_{12}) - r_U \rho_0 (1 - \tau_{00}). \quad (4.13)$$

Furthermore, when the test is perfect, we have $(\tau_{11} = 1, \tau_{22} = 1, \tau_{00} = 1, \tau_{21} = 0, \tau_{22} = 0).$ From (4.13); the expected utility for the perfect test would be

$$NB_P^{RI} = \rho_1 + \rho_2 .$$

Therefore, the relative net benefit for rule-in can be defined as the ratio of expected utility to perfect expected utility.

$$\begin{aligned} RNB^{RI} &= \frac{NB_0^{RI}}{NB_P^{RI}} = \frac{\rho_1(\tau_{11} + \tau_{21}) + \rho_2(\tau_{22} + \tau_{12}) - r_U \rho_0 (1 - \tau_{00})}{\rho_1 + \rho_2} \\ &= \frac{\rho_1}{\rho_1 + \rho_2}(\tau_{11} + \tau_{21}) + \frac{\rho_2}{\rho_1 + \rho_2}(\tau_{22} + \tau_{12}) - \frac{\rho_0}{\rho_1 + \rho_2} r(1 - \tau_{00}) \\ &= W_1(\tau_{11} + \tau_{21}) + W_2(\tau_{22} + \tau_{12}) - (1 - \tau_{00}) r O^{-1} = W_1(\tau_{11} + \tau_{21}) + W_2(\tau_{22} + \tau_{12}) - (1 - \tau_{00}) m_{RI}, \end{aligned}$$

(4.14)

where, $W_1 + W_2 = 1,\ W_1 = \frac{\rho_1}{\rho_1 + \rho_2},\ W_2 = \frac{\rho_2}{\rho_1 + \rho_2},\ m_{RI} = r_U O^{-1}$ and $O = \frac{\rho_1 + \rho_2}{\rho_0}.$

Like in Pennello et al. (2016), I notice that $RNB^{RI} > 0$ if and only if

$W_1(\tau_{11} + \tau_{21}) + W_2(\tau_{22} + \tau_{12}) > (1 - \tau_{00}) m_{RI}$, which implies that

$$\begin{aligned} & m_{RI} = \frac{W_1(\tau_{11} + \tau_{21}) + W_2(\tau_{22} + \tau_{12})}{(1 - \tau_{00})} = W_1 \frac{(\tau_{11} + \tau_{21})}{(1 - \tau_{00})} + W_2 \frac{(\tau_{22} + \tau_{12})}{(1 - \tau_{00})} \\ & \equiv \frac{\text{Weighted Function of } TP \text{ (stage 1 or 2) of the disease}}{FP} \end{aligned} \quad (4.15)$$

which I call the Generalized Weighted Positive Likelihood Ratio (GPLR)

and $W_1(\tau_{11} + \tau_{21}) + W_2(\tau_{22} + \tau_{12}) = (1 - \tau_{00}) m_{RI}$ is the line in ROC space.

Also, I can generalize to k-stage diseases as follows: from (4.14), I have

$$GRNB^{RI}=\sum_{i=1}^{k-1}W_i\sum_{j=1}^{k-1}\tau_{ji}-(1-\tau_{00})r_U O_G^{-1}=\sum_{i=1}^{k-1}W_i\sum_{j=1}^{k-1}\tau_{ji}-(1-\tau_{00})m_{RI}\,,$$

where, $\sum_{i=1}^{k-1}W_i=1,\ (W_i=\frac{\rho_i}{\sum_{i=1}^{k-1}\rho_i}, i=1,2,...,k-1),\ m_{RI}=r_U O_G^{-1}$and $O_G^{-1}=\frac{\rho_0}{\sum_{i=1}^{k-1}\rho_i}$.

Like in Pennello et al. (2016, 2019 ENAR), we notice that $GRNB^{RI}>0$ if and only if $\sum_{i=1}^{k-1}W_i\sum_{j=1}^{k-1}\tau_{ji}>(1-\tau_{00})m_{RI}$, which implies that

$$m_{RI}=\frac{\sum_{i=1}^{k-1}W_i\sum_{j=1}^{k-1}\tau_{ji}}{(1-\tau_{00})}\,.$$

$$\equiv\frac{\text{Weighted Function of } TP \text{ (of k-stages diseases)}}{FP}$$

I call the Generalized Weighted Positive Likelihood Ratio (GPLR) for the k-stage diseases and $Y_{RI}=\sum_{i=1}^{k-1}W_i\sum_{j=1}^{k-1}\tau_{ji}=(1-\tau_{00})m_{RI}$ is the line in ROC space.

4.5.2 Relative Net-Benefit of Rule-Out

Similarly, to find the relative net benefit for rule-out patients, I have

$$\begin{aligned}NB_\tau&=E-E_\tau\\&=\rho_0\tau_{00}+r_U^{RO}[\rho_1(\tau_{11}+\tau_{21})+\rho_2(\tau_{22}+\tau_{12})]-\rho_0(1-\tau)-r_U^{RO}\tau(\rho_1+\rho_2)\\&=\rho_0(\tau_{00}-1+\tau)-r_U^{RO}[\tau(\rho_1+\rho_2)-\rho_1(\tau_{11}+\tau_{21})-\rho_2(\tau_{22}+\tau_{12})]\end{aligned}$$

and

$$\begin{aligned}NB^{RO}&=NB_1=\rho_0\tau_{00}-r_U^{RO}\rho_1(1-\tau_{11}-\tau_{21})-r_U^{RO}\rho_2(1-\tau_{22}-\tau_{12})\\&=\rho_0\tau_{00}-r_U^{RO}[(\rho_1+\rho_2)-\rho_1(\tau_{11}+\tau_{21})-\rho_2(\tau_{22}+\tau_{12})].\end{aligned},$$

This occurs where $\tau = 1 \Rightarrow \{\tau_1 = 1 \text{ and } \tau_2 = 0\}$ or $\{\tau_1 = 0 \text{ and } \tau_2 = 1\}$ because we send the patient to treatment, despite the disease stage. Also, I have $NB_P^{RO} = \rho_0$, when I have $\tau_{00} = 1, \tau_{01} = \tau_{02} = 0$ (the false-negative rate at all stages =0). Additionally, I have

$$\begin{aligned} RNB^{RO} = \frac{NB_1^{RO}}{NB_P^{RO}} &= \frac{\rho_0\tau_{00} - r_U^{RO}[(\rho_1 + \rho_2) - \rho_1(\tau_{11} + \tau_{21}) - \rho_2(\tau_{22} + \tau_{12})]}{\rho_0} \\ &= \tau_{00} - \frac{r_U^{RO}[(\rho_1 + \rho_2) - \rho_1(\tau_{11} + \tau_{21}) - \rho_2(\tau_{22} + \tau_{12})]}{\rho_0} \\ &= \tau_{00} - r_U^{RO}\frac{(\rho_1 + \rho_2)}{\rho_0}[1 - \frac{\rho_1}{(\rho_1 + \rho_2)}(\tau_{11} + \tau_{21}) - \frac{\rho_2}{(\rho_1 + \rho_2)}(\tau_{22} + \tau_{12})] \\ &= \tau_{00} - m_0^{-1}[1 - W_1(\tau_{11} + \tau_{21}) - W_2(\tau_{22} + \tau_{12})] \end{aligned} \qquad (4.16)$$

where $[m_0 = r_U O_1,\ O_1 = \frac{\rho_0}{\rho_1 + \rho_2}]$.

Finally, I have $RNB^{RO} = \tau_{00} - m_0^{-1}[1 - W_1(\tau_{11} + \tau_{21}) - W_2(\tau_{22} + \tau_{12})]$. Again, I notice that $RNB^{RO} > 0$ if and only if $\tau_{00} > m_0^{-1}[1 - W_1(\tau_{11} + \tau_{21}) - W_2(\tau_{22} + \tau_{12})]$. Hence, when

$$\begin{aligned} m_0\tau_{00} &= [1 - W_1(\tau_{11} + \tau_{21}) - W_2(\tau_{22} + \tau_{12})] \\ W_1(\tau_{11} + \tau_{21}) + W_2(\tau_{22} + \tau_{12}) &= 1 - m_0\tau_{00} \\ &= 1 - m_0(1 - 1 + \tau_{00}), \end{aligned}$$

where

$$m_0 = \frac{[1 - W_1(\tau_{11} + \tau_{21}) - W_2(\tau_{22} + \tau_{12})]}{\tau_{00}} \equiv \frac{\text{Weighted Function of } FN \text{ from stage 1 and 2 of the disease}}{TN},$$

calling it the Generalized Weighted Negative Likelihood Ratio (GNLR), I find,

$$W_1(\tau_{11} + \tau_{21}) + W_2(\tau_{22} + \tau_{12}) = 1 - m_0 + m_0(1 - \tau_{00}). \qquad (4.17)$$

Also, I can generalize to k-stage diseases as follows: from (4.16), I have

$$\begin{aligned} GRNB^{RO} &= \tau_{00} - \frac{r^{RO}}{\rho_0}\sum_{i=1}^{k-1}\rho_i\tau_{0i} \\ &= \tau_{00} - m_0^{-1}[\tau_{01} + \sum_{i=2}^{k-1}O_i\tau_{0i}] \end{aligned}$$

where $[m_0 = r_u O_1,\ O_1 = \frac{\rho_0}{\rho_1}, O_i = \frac{\rho_i}{\rho_1}; i = 2,3,...,k-1]$.

Finally, I have $GRNB^{RO} = \tau_{00} - m_0{}^{-1}[\tau_{01} + \sum_{i=2}^{k-1} O_i \tau_{0i}]$. Again, I notice that $GRNB^{RO} > 0$ if and only if $\tau_{00} > m_0{}^{-1}[\tau_{01} + \sum_{i=2}^{k-1} O_i \tau_{0i}]$, and hence, when

$$m_0 \tau_{00} = [\tau_{01} + \sum_{i=2}^{k-1} O_i \tau_{0i}]$$

$$m_0(\tau_{00} - 1 + 1) = \tau_{01} + \sum_{i=2}^{k-1} O_i \tau_{0i}$$

$$m_0 - m_0(1-\tau_{00}) = (1-(\tau_{11}+\tau_{21})) + \sum_{i=2}^{k-1} O_i (1 - \sum_{j=1}^{k-1} \tau_{ij}),$$

where $m_0 = \frac{[\tau_{01} + \sum_{i=2}^{k-1} O_i \tau_{0i}]}{\tau_{00}} \equiv \frac{\text{Weighted Function of } FN \text{ in the k-stages disease}}{TN}$,

I call it the Generalized Weighted Negative Likelihood Ratio (GNLR), with the result being

$(\tau_{11}+\tau_{21}) + \sum_{i=2}^{k-1} O_i \sum_{j=1}^{k-1} \tau_{ij} = 1 + \sum_{i=2}^{k-1} O_i - m_0 + m_0(1-\tau_{00})$ is in the line in ROC space.

4.5.3 The Valid Choice of Utility Ratio *r*

Using (4.12), I have

$$\begin{aligned} NB_\tau = E - E_\tau &= r_U \rho_0(\tau_{00} - 1 + \tau) + \rho_1(\tau_{11} + \tau_{21} - \tau_1) + \rho_2(\tau_{22} + \tau_{12} - \tau_2) \\ &= -r_U \rho_0(1 - \tau_{00} - \tau) + \rho_1(\tau_{11} + \tau_{21} - \tau_1) + \rho_2(\tau_{22} + \tau_{12} - \tau_2). \end{aligned}$$

For a test to have a positive net benefit $NB_\tau > 0$ when it is

$$r_U \rho_0(1-\tau_{00}-\tau) < \rho_1(\tau_{11}+\tau_{21}-\tau_1) + \rho_2(\tau_{22}+\tau_{12}-\tau_2) .$$

Depending on the order, $\tau_{00}, \tau, (\tau_{11}+\tau_{21})$, and $(\tau_{22}+\tau_{12})$, I have

$$r_U < \frac{\rho_1(\tau_{11}+\tau_{21}-\tau_1) + \rho_2(\tau_{22}+\tau_{12}-\tau_2)}{\rho_0(1-\tau_{00}-\tau)}, \text{ if } \tau < (1-\tau_{00}), \tau_1 < (\tau_{11}+\tau_{21}) \text{ and/or } \tau_2 < (\tau_{22}+\tau_{12}), \quad (4.18)$$

provided that $\rho_1(\tau_{11}+\tau_{21}-\tau_1)+\rho_2(\tau_{22}+\tau_{12}-\tau_2)>0$;

$$r_U > \frac{\rho_1(\tau_1-\tau_{11}-\tau_{21})+\rho_2(\tau_2-\tau_{22}-\tau_{12})}{\rho_0[\tau-(1-\tau_{00})]},\ \text{if } (1-\tau_{00})<\tau, \tau_1>(\tau_{11}+\tau_{21}) \text{ and / or } \tau_2>(\tau_{22}+\tau_{12})$$

(8.19) provided that $\rho_1(\tau_{11}+\tau_{21}-\tau_1)+\rho_2(\tau_{22}+\tau_{12}-\tau_2)<0$. The upper bound inequality is minimized when $\tau, \tau_1, \tau_2,$ and $\tau=\tau_1+\tau_2$ are chosen, and such that

$$\frac{(\tau_{11}+\tau_{21}-\tau_1)+(\tau_{22}+\tau_{12}-\tau_2)}{[(1-\tau_{00})-\tau)} = \frac{(\tau_{11}+\tau_{21}+\tau_{22}+\tau_{12}-\tau)}{[(1-\tau_{00})-\tau)} = 1+\frac{(\tau_{11}+\tau_{21}+\tau_{22}+\tau_{12}-(1-\tau_{00}))}{[(1-\tau_{00})-\tau)}.$$

We can achieve this when $\tau=0 \Rightarrow \tau_1=\tau_2=0$ (by never sending a patient to treatment at any stage). This conclusion is similar to Pennello et al. (2016), so we need a trivial test to find the right choice of r_u,where the NB for a test relative to any random test remains positive. Therefore, from (4.18), the constraint on the upper bound of r_u is

$$r_U < \frac{\rho_1(\tau_{11}+\tau_{21})+\rho_2(\tau_{22}+\tau_{12})}{\rho_0(1-\tau_{00})}.$$

However, when the constraint for the lower bound of *r* for an informative test is $(\tau_{11}+\tau_{21}+\tau_{22}+\tau_{12})>(1-\tau_{00})$, I require that the pre-test odds at any stage of the disease $\left(\frac{\rho_1+\rho_2}{\rho_0}\right)$by the lower bound $r>\frac{\rho_1+\rho_2}{\rho_0}$ to eliminate random tests from consideration. Therefore, as in Pennello et al. (2016), (4.19), no additional constraint is required. Furthermore, for the general case, k-stage disease, the restriction on the upper bound of *r* is

$$r_U < \frac{\sum_{i=1}^{k-1}\rho_i\sum_{j=1}^{k-1}\tau_{ij}}{\rho_0(1-\tau_{00})},\ \text{and on the lower bound is } \left(\frac{\sum_{i=1}^{k-1}\rho_i}{\rho_0}\right).$$

4.5.3.1 Parametric Expressions for Utility Ratio r

As I discussed in (4.5), we derive r_u (loss ratio or the utility) as follows: For rule-out patients

$$r_U^{RO} = \frac{C}{B} = \frac{(r_{10}+r_{20}-r_{00})}{(r_{01}+r_{02}-(r_{11}+r_{12}+r_{21}+r_{22}))} = FP:FN \text{ loss ratio=TN:TP utility ratio,}$$

C =net cost (harm) of treating a subject without disease with stage one or two treatments,

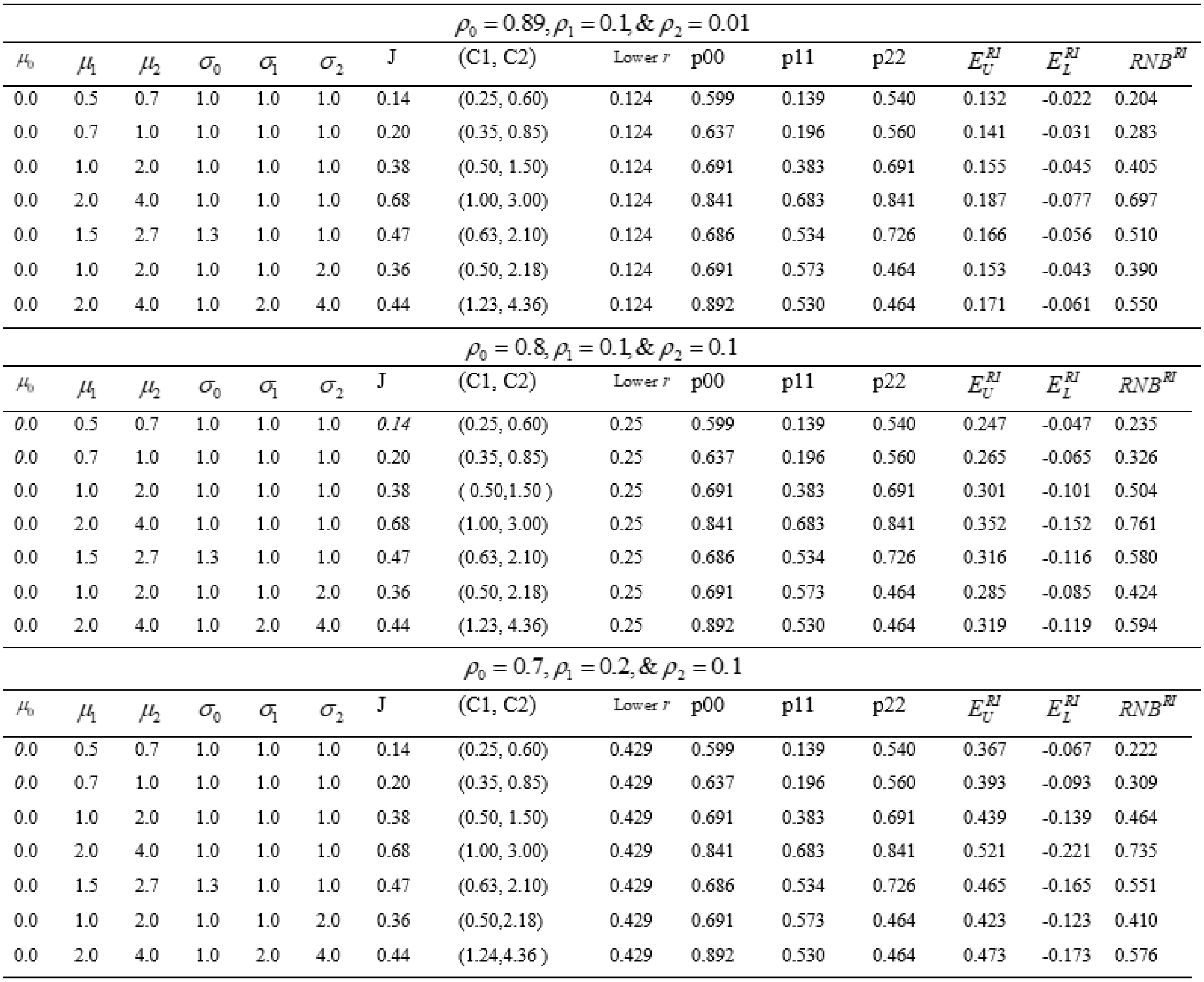
B =net benefit of using stage 1 or 2 treatments to treat a subject at stage 1 or 2 of the disease. Therefore, we identify r as a loss ratio by defining it $r_L^{RO} = FP:FN$ loss ratio $= \frac{(r_{10}+r_{20})}{(r_{01}+r_{02})}$.

Similarly, we represent $r_U^{RI} =$ TN : TP utility ratio = $\frac{r_{00}}{(r_{11}+r_{12}+r_{21}+r_{22})}$. Consequently, from (4.3), we have the parametric expression r_L^{RO} as follows: $r_L^{RO} = \frac{(\rho_0\tau_{10}+\rho_0\tau_{20})}{(\rho_1\tau_{01}+\rho_2\tau_{02})}$ and for r_U^{RO} we have $r_U^{RO} = \frac{\rho_0\tau_{00}}{\rho_1(\tau_{11}+\tau_{21})+\rho_2(\tau_{22}+\tau_{12})}$.

4.6 Numerical Examples

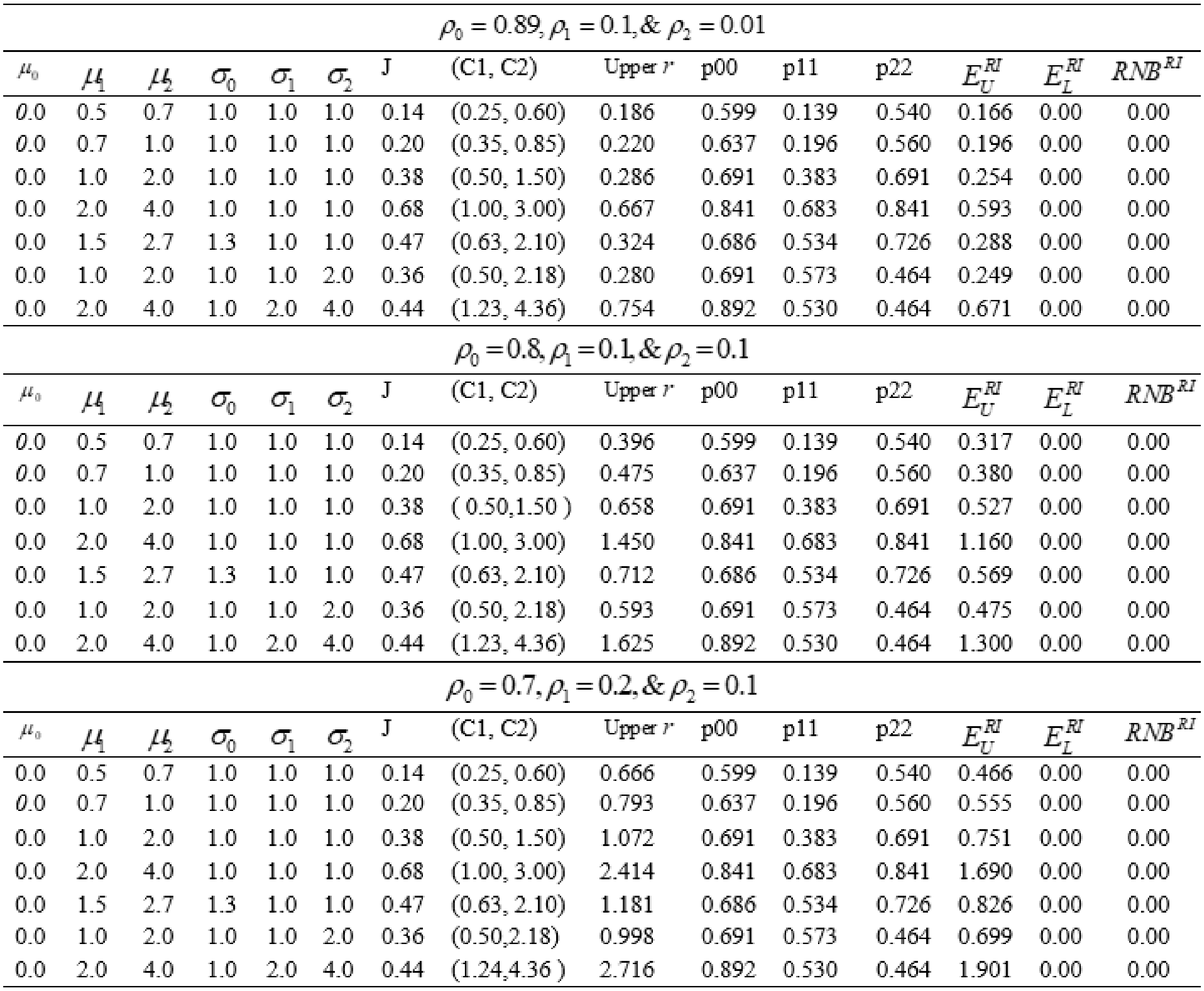
My research considers a disease with three stages (i.e., k=3) (non-diseased, stage 1, and stage 2 diseased). Numerical examples are conducted for the three-stage diseases, assuming different prevalence settings, with varying parameters for the diseased underlying distributions as shown in Table 4.2, Table 4.3, Table 4.4, and Table 4.5. To illustrate, we $X_1, X_2,$ and X_3 denote biomarker values for non-diseased, stage 1, and stage 2 diseased subjects with pdfs $f_1(.)$, $f_2(.)$, and $f_3(.)$, respectively. In Tables 4.2-4.5, I discuss symmetric distributions and assume that $X_1{\sim}N(\mu_1,\sigma_1)$, $X_2{\sim}N(\mu_2,\sigma_2)$ and $X_3{\sim}N(\mu_3,\sigma_3)$. For illustration purposes, we use one objective function, the Youden index (i.e., $J(C)$), to select the cut-points $(C_1$ and $C_2)$ and then calculate the values of the lower bound, the upper bound, the average, and the proposed form of r. For each prevalence set, I consider seven different parameters for three underlying disease distributions for various means and variances.

Table 4.2. *Settings of Parameters and Prevalences by Using Lower Bound of r for Rule-In*



μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Lower r	p00	p11	p22	E_U^{RI}	E_L^{RI}	RNB^{RI}
							$\rho_0 = 0.89, \rho_1 = 0.1, \& \rho_2 = 0.01$							
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	0.124	0.599	0.139	0.540	0.132	-0.022	0.204
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	0.124	0.637	0.196	0.560	0.141	-0.031	0.283
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	0.124	0.691	0.383	0.691	0.155	-0.045	0.405
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	0.124	0.841	0.683	0.841	0.187	-0.077	0.697
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	0.124	0.686	0.534	0.726	0.166	-0.056	0.510
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	0.124	0.691	0.573	0.464	0.153	-0.043	0.390
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	0.124	0.892	0.530	0.464	0.171	-0.061	0.550
							$\rho_0 = 0.8, \rho_1 = 0.1, \& \rho_2 = 0.1$							
μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Lower r	p00	p11	p22	E_U^{RI}	E_L^{RI}	RNB^{RI}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	0.25	0.599	0.139	0.540	0.247	-0.047	0.235
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	0.25	0.637	0.196	0.560	0.265	-0.065	0.326
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50,1.50)	0.25	0.691	0.383	0.691	0.301	-0.101	0.504
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	0.25	0.841	0.683	0.841	0.352	-0.152	0.761
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	0.25	0.686	0.534	0.726	0.316	-0.116	0.580
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	0.25	0.691	0.573	0.464	0.285	-0.085	0.424
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	0.25	0.892	0.530	0.464	0.319	-0.119	0.594
							$\rho_0 = 0.7, \rho_1 = 0.2, \& \rho_2 = 0.1$							
μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Lower r	p00	p11	p22	E_U^{RI}	E_L^{RI}	RNB^{RI}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	0.429	0.599	0.139	0.540	0.367	-0.067	0.222
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	0.429	0.637	0.196	0.560	0.393	-0.093	0.309
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	0.429	0.691	0.383	0.691	0.439	-0.139	0.464
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	0.429	0.841	0.683	0.841	0.521	-0.221	0.735
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	0.429	0.686	0.534	0.726	0.465	-0.165	0.551
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50,2.18)	0.429	0.691	0.573	0.464	0.423	-0.123	0.410
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.24,4.36)	0.429	0.892	0.530	0.464	0.473	-0.173	0.576

Table 4.3. *Settings of Parameters and Prevalences by Using Upper Bound of r for Rule-In*



$\rho_0 = 0.89, \rho_1 = 0.1, \& \ \rho_2 = 0.01$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Upper r	p00	p11	p22	E_U^{RI}	E_L^{RI}	RNB^{RI}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	0.186	0.599	0.139	0.540	0.166	0.00	0.00
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	0.220	0.637	0.196	0.560	0.196	0.00	0.00
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	0.286	0.691	0.383	0.691	0.254	0.00	0.00
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	0.667	0.841	0.683	0.841	0.593	0.00	0.00
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	0.324	0.686	0.534	0.726	0.288	0.00	0.00
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	0.280	0.691	0.573	0.464	0.249	0.00	0.00
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	0.754	0.892	0.530	0.464	0.671	0.00	0.00

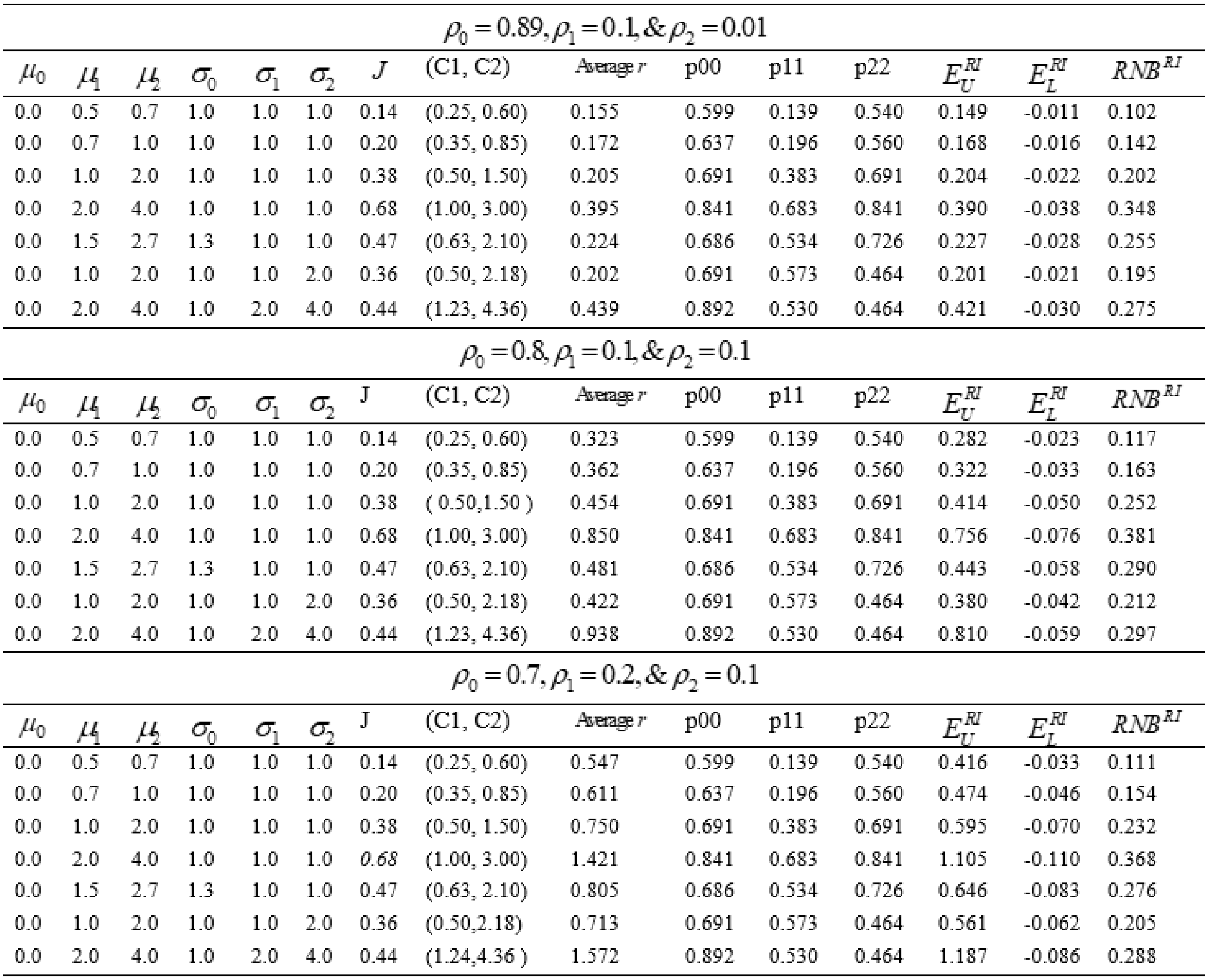
$\rho_0 = 0.8, \rho_1 = 0.1, \& \rho_2 = 0.1$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Upper r	p00	p11	p22	E_U^{RI}	E_L^{RI}	RNB^{RI}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	0.396	0.599	0.139	0.540	0.317	0.00	0.00
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	0.475	0.637	0.196	0.560	0.380	0.00	0.00
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50,1.50)	0.658	0.691	0.383	0.691	0.527	0.00	0.00
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	1.450	0.841	0.683	0.841	1.160	0.00	0.00
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	0.712	0.686	0.534	0.726	0.569	0.00	0.00
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	0.593	0.691	0.573	0.464	0.475	0.00	0.00
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	1.625	0.892	0.530	0.464	1.300	0.00	0.00

$\rho_0 = 0.7, \rho_1 = 0.2, \& \ \rho_2 = 0.1$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Upper r	p00	p11	p22	E_U^{RI}	E_L^{RI}	RNB^{RI}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	0.666	0.599	0.139	0.540	0.466	0.00	0.00
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	0.793	0.637	0.196	0.560	0.555	0.00	0.00
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	1.072	0.691	0.383	0.691	0.751	0.00	0.00
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	2.414	0.841	0.683	0.841	1.690	0.00	0.00
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	1.181	0.686	0.534	0.726	0.826	0.00	0.00
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50,2.18)	0.998	0.691	0.573	0.464	0.699	0.00	0.00
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.24,4.36)	2.716	0.892	0.530	0.464	1.901	0.00	0.00

Table 4.4. *Settings of Parameters and Prevalences by Using the average of r for Rule-In*



$\rho_0 = 0.89, \rho_1 = 0.1, \& \rho_2 = 0.01$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Average r	p00	p11	p22	E_U^{RI}	E_L^{RI}	RNB^{RI}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	0.155	0.599	0.139	0.540	0.149	-0.011	0.102
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	0.172	0.637	0.196	0.560	0.168	-0.016	0.142
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	0.205	0.691	0.383	0.691	0.204	-0.022	0.202
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	0.395	0.841	0.683	0.841	0.390	-0.038	0.348
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	0.224	0.686	0.534	0.726	0.227	-0.028	0.255
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	0.202	0.691	0.573	0.464	0.201	-0.021	0.195
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	0.439	0.892	0.530	0.464	0.421	-0.030	0.275

$\rho_0 = 0.8, \rho_1 = 0.1, \& \rho_2 = 0.1$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Average r	p00	p11	p22	E_U^{RI}	E_L^{RI}	RNB^{RI}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	0.323	0.599	0.139	0.540	0.282	-0.023	0.117
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	0.362	0.637	0.196	0.560	0.322	-0.033	0.163
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50,1.50)	0.454	0.691	0.383	0.691	0.414	-0.050	0.252
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	0.850	0.841	0.683	0.841	0.756	-0.076	0.381
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	0.481	0.686	0.534	0.726	0.443	-0.058	0.290
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	0.422	0.691	0.573	0.464	0.380	-0.042	0.212
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	0.938	0.892	0.530	0.464	0.810	-0.059	0.297

$\rho_0 = 0.7, \rho_1 = 0.2, \& \rho_2 = 0.1$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Average r	p00	p11	p22	E_U^{RI}	E_L^{RI}	RNB^{RI}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	0.547	0.599	0.139	0.540	0.416	-0.033	0.111
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	0.611	0.637	0.196	0.560	0.474	-0.046	0.154
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	0.750	0.691	0.383	0.691	0.595	-0.070	0.232
0.0	2.0	4.0	1.0	1.0	1.0	*0.68*	(1.00, 3.00)	1.421	0.841	0.683	0.841	1.105	-0.110	0.368
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	0.805	0.686	0.534	0.726	0.646	-0.083	0.276
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50,2.18)	0.713	0.691	0.573	0.464	0.561	-0.062	0.205
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.24,4.36)	1.572	0.892	0.530	0.464	1.187	-0.086	0.288

Table 4.5. *Settings of Parameters and Prevalences by Using Cost-Benefit Ratio r for Rule-In*

$\rho_0 = 0.89, \rho_1 = 0.1, \& \rho_2 = 0.01$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	r_U	r_L	p00	p11	p22	E_U^{RI}	E_L^{RI}	RNB^{RI}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	0.125	0.121	0.599	0.139	0.540	0.133	-0.023	0.200
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	0.125	0.120	0.637	0.196	0.560	0.142	-0.032	0.278
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	0.128	0.115	0.691	0.383	0.691	0.157	-0.047	0.395
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	0.126	0.112	0.841	0.683	0.841	0.188	-0.078	0.694
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	0.148	0.069	0.686	0.534	0.726	0.181	-0.071	0.446
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	0.125	0.121	0.691	0.573	0.464	0.154	-0.044	0.387
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	0.091	0.391	0.892	0.530	0.464	0.145	-0.035	0.579

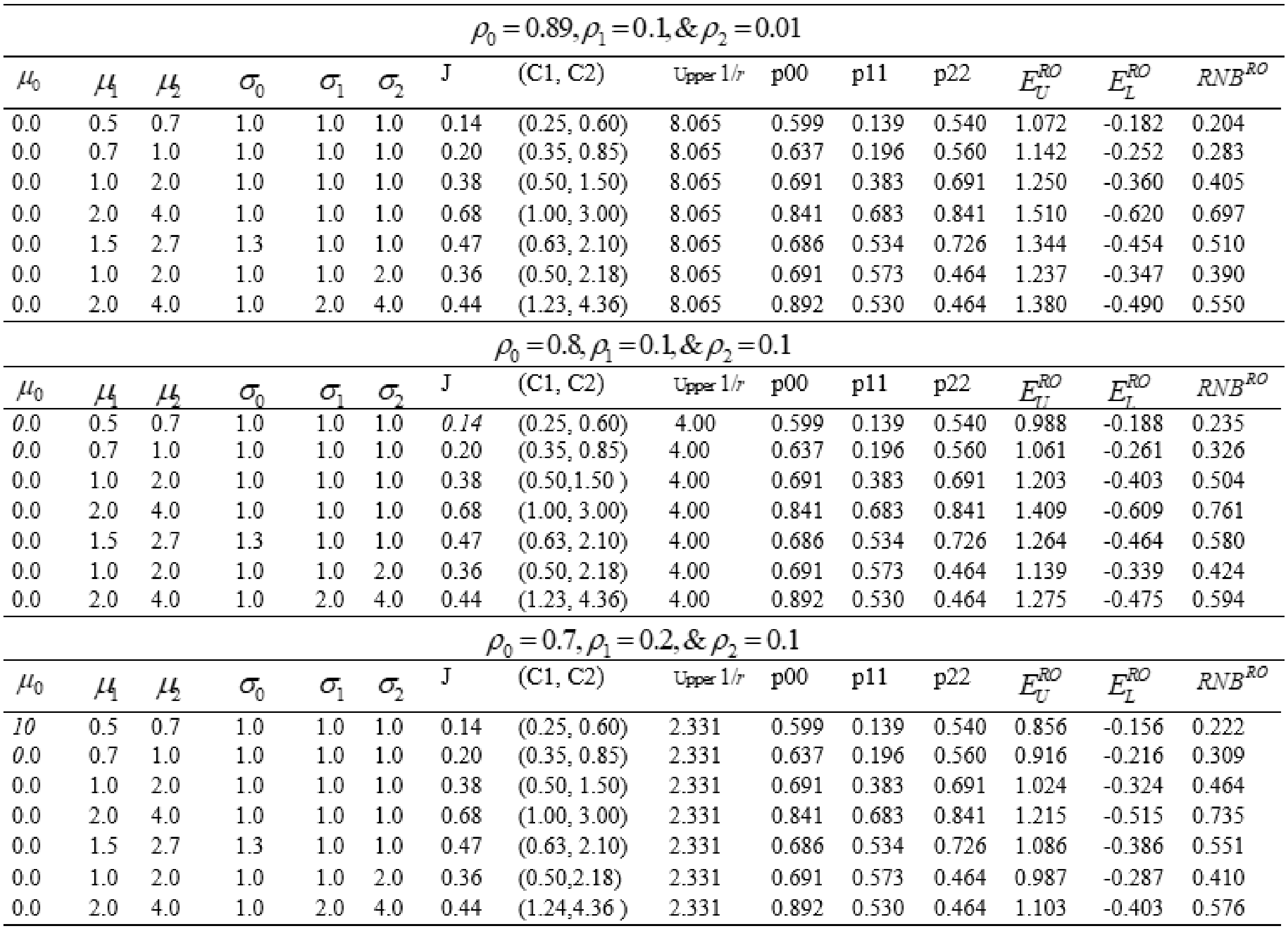
$\rho_0 = 0.8, \rho_1 = 0.1, \& \rho_2 = 0.1$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	r_U	r_L	p00	p11	p22	E_U^{RI}	E_L^{RI}	RNB^{RI}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	0.266	0.227	0.599	0.139	0.540	0.254	-0.054	0.210
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	0.271	0.214	0.637	0.196	0.560	0.276	-0.076	0.296
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50,1.50	0.294	0.152	0.691	0.383	0.691	0.325	-0.125	0.450
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	0.273	0.126	0.841	0.683	0.841	0.368	-0.168	0.747
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	0.326	0.084	0.686	0.534	0.726	0.358	-0.158	0.485
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	0.265	0.217	0.691	0.573	0.464	0.293	-0.093	0.406
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	0.197	0.691	0.892	0.530	0.464	0.281	-0.081	0.617

$\rho_0 = 0.7, \rho_1 = 0.2, \& \rho_2 = 0.1$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	r_U	r_L	p00	p11	p22	E_U^{RI}	E_L^{RI}	RNB^{RI}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	0.446	0.402	0.599	0.139	0.540	0.374	-0.074	0.206
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	0.452	0.387	0.637	0.196	0.560	0.403	-0.103	0.289
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	0.479	0.317	0.691	0.383	0.691	0.463	-0.163	0.428
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	0.455	0.287	0.841	0.683	0.841	0.536	-0.236	0.725
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	0.541	0.183	0.686	0.534	0.726	0.519	-0.219	0.469
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	0.445	0.391	0.691	0.573	0.464	0.431	-0.131	0.398
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.24, 4.36	0.329	1.255	0.892	0.530	0.464	0.410	-0.110	0.601

Table 4.6. *Settings of Parameters and Prevalences by Using Upper Bound of r for Rule-Out*



$\rho_0 = 0.89, \rho_1 = 0.1, \& \rho_2 = 0.01$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Upper 1/r	p00	p11	p22	E_U^{RO}	E_L^{RO}	RNB^{RO}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	8.065	0.599	0.139	0.540	1.072	-0.182	0.204
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	8.065	0.637	0.196	0.560	1.142	-0.252	0.283
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	8.065	0.691	0.383	0.691	1.250	-0.360	0.405
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	8.065	0.841	0.683	0.841	1.510	-0.620	0.697
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	8.065	0.686	0.534	0.726	1.344	-0.454	0.510
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	8.065	0.691	0.573	0.464	1.237	-0.347	0.390
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	8.065	0.892	0.530	0.464	1.380	-0.490	0.550

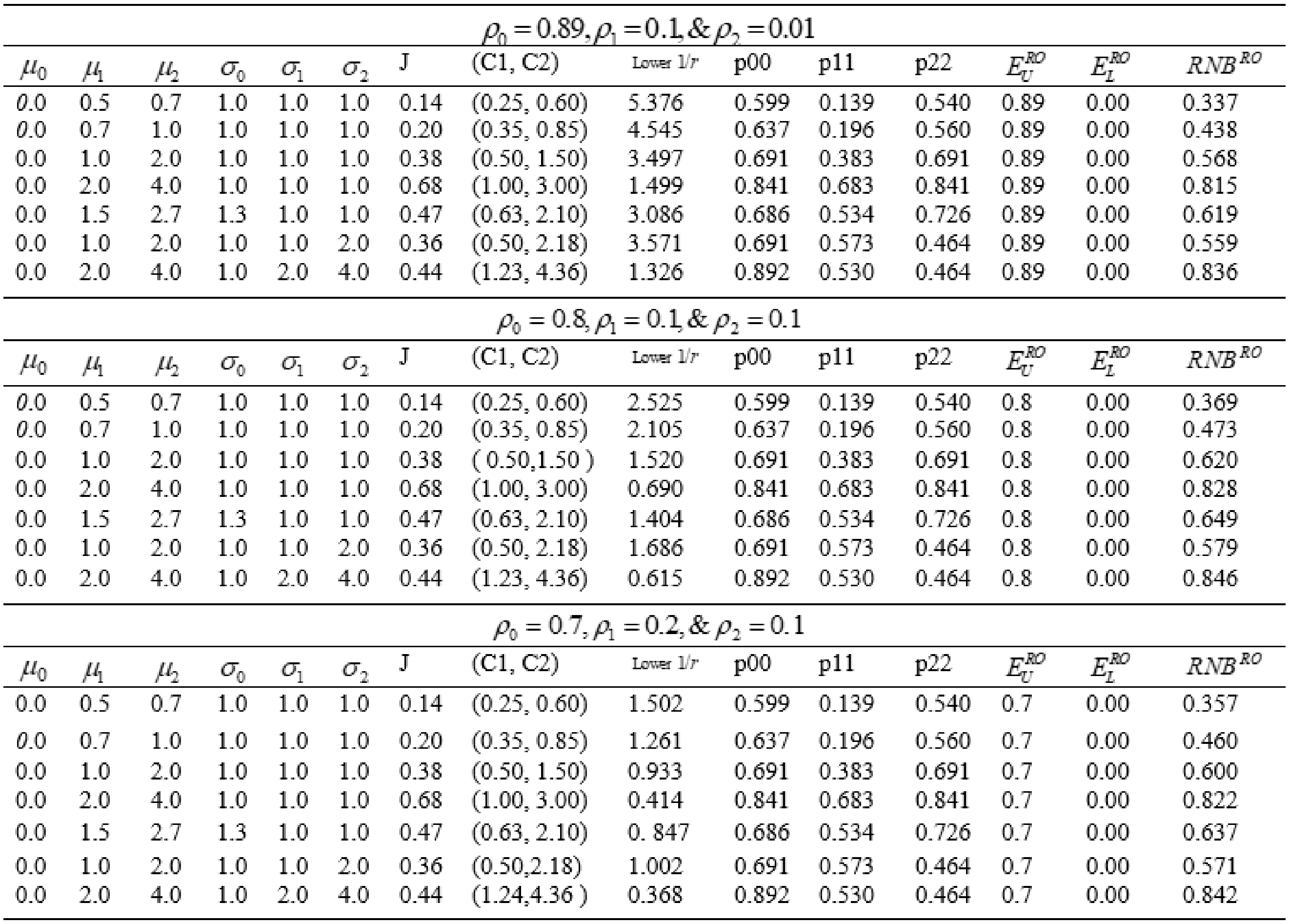
$\rho_0 = 0.8, \rho_1 = 0.1, \& \rho_2 = 0.1$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Upper 1/r	p00	p11	p22	E_U^{RO}	E_L^{RO}	RNB^{RO}
0.0	0.5	0.7	1.0	1.0	1.0	*0.14*	(0.25, 0.60)	4.00	0.599	0.139	0.540	0.988	-0.188	0.235
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	4.00	0.637	0.196	0.560	1.061	-0.261	0.326
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50,1.50)	4.00	0.691	0.383	0.691	1.203	-0.403	0.504
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	4.00	0.841	0.683	0.841	1.409	-0.609	0.761
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	4.00	0.686	0.534	0.726	1.264	-0.464	0.580
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	4.00	0.691	0.573	0.464	1.139	-0.339	0.424
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	4.00	0.892	0.530	0.464	1.275	-0.475	0.594

$\rho_0 = 0.7, \rho_1 = 0.2, \& \rho_2 = 0.1$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Upper 1/r	p00	p11	p22	E_U^{RO}	E_L^{RO}	RNB^{RO}
10	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	2.331	0.599	0.139	0.540	0.856	-0.156	0.222
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	2.331	0.637	0.196	0.560	0.916	-0.216	0.309
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	2.331	0.691	0.383	0.691	1.024	-0.324	0.464
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	2.331	0.841	0.683	0.841	1.215	-0.515	0.735
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	2.331	0.686	0.534	0.726	1.086	-0.386	0.551
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50,2.18)	2.331	0.691	0.573	0.464	0.987	-0.287	0.410
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.24,4.36)	2.331	0.892	0.530	0.464	1.103	-0.403	0.576

Table 4.7. *Settings of Parameters and Prevalences by Using Lower Bound of r for Rule-Out*



$\rho_0 = 0.89, \rho_1 = 0.1, \& \rho_2 = 0.01$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Lower $1/r$	p00	p11	p22	E_U^{RO}	E_L^{RO}	RNB^{RO}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	5.376	0.599	0.139	0.540	0.89	0.00	0.337
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	4.545	0.637	0.196	0.560	0.89	0.00	0.438
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	3.497	0.691	0.383	0.691	0.89	0.00	0.568
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	1.499	0.841	0.683	0.841	0.89	0.00	0.815
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	3.086	0.686	0.534	0.726	0.89	0.00	0.619
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	3.571	0.691	0.573	0.464	0.89	0.00	0.559
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	1.326	0.892	0.530	0.464	0.89	0.00	0.836

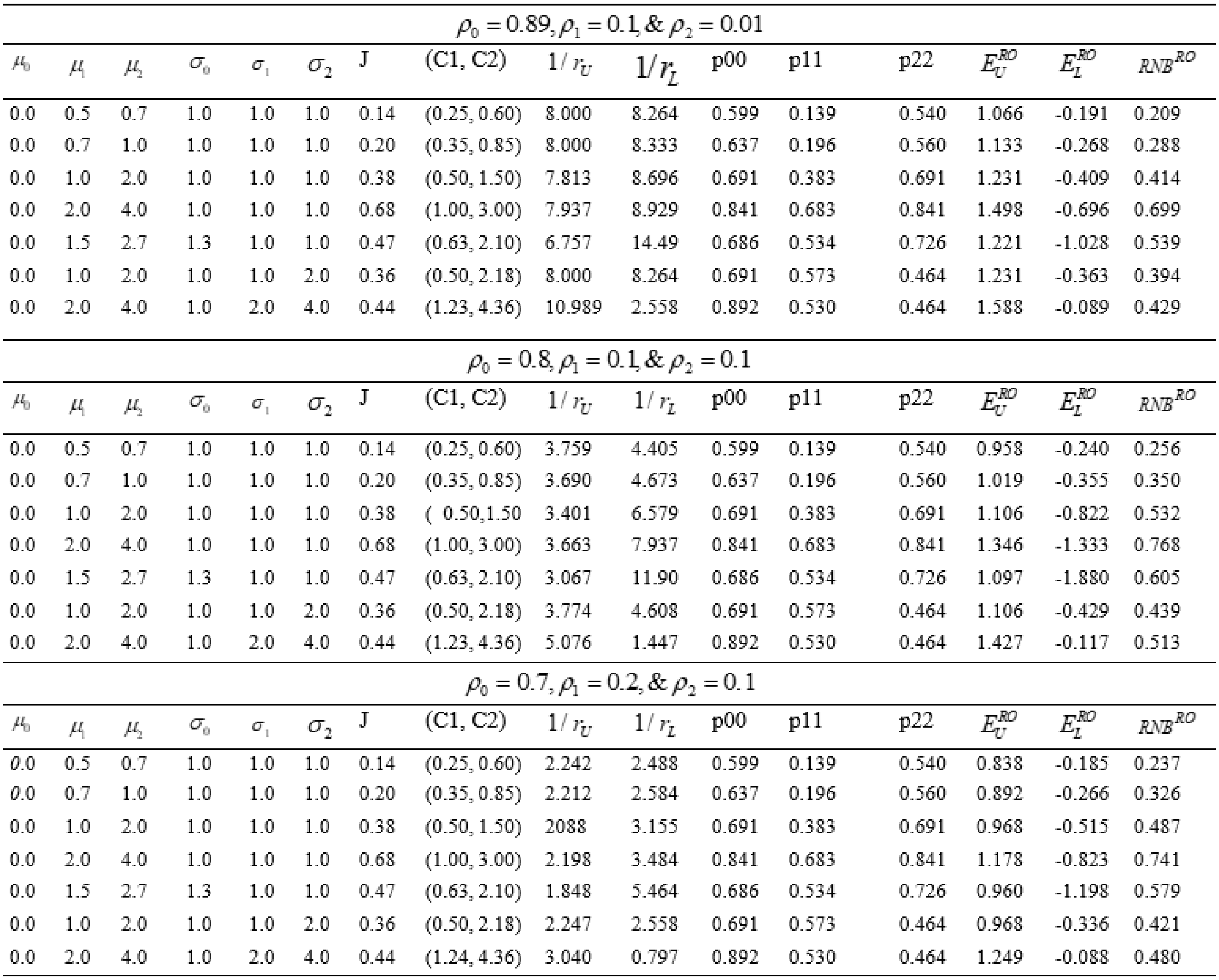
$\rho_0 = 0.8, \rho_1 = 0.1, \& \rho_2 = 0.1$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Lower $1/r$	p00	p11	p22	E_U^{RO}	E_L^{RO}	RNB^{RO}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	2.525	0.599	0.139	0.540	0.8	0.00	0.369
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	2.105	0.637	0.196	0.560	0.8	0.00	0.473
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50,1.50)	1.520	0.691	0.383	0.691	0.8	0.00	0.620
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	0.690	0.841	0.683	0.841	0.8	0.00	0.828
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	1.404	0.686	0.534	0.726	0.8	0.00	0.649
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	1.686	0.691	0.573	0.464	0.8	0.00	0.579
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	0.615	0.892	0.530	0.464	0.8	0.00	0.846

$\rho_0 = 0.7, \rho_1 = 0.2, \& \rho_2 = 0.1$

μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	Lower $1/r$	p00	p11	p22	E_U^{RO}	E_L^{RO}	RNB^{RO}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	1.502	0.599	0.139	0.540	0.7	0.00	0.357
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	1.261	0.637	0.196	0.560	0.7	0.00	0.460
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	0.933	0.691	0.383	0.691	0.7	0.00	0.600
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	0.414	0.841	0.683	0.841	0.7	0.00	0.822
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	0. 847	0.686	0.534	0.726	0.7	0.00	0.637
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50,2.18)	1.002	0.691	0.573	0.464	0.7	0.00	0.571
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.24,4.36)	0.368	0.892	0.530	0.464	0.7	0.00	0.842

Table 4.8. *Settings of Parameters and Prevalences by Using Cost-Benefit Ratio r for Rule-Out*



$\rho_0 = 0.89, \rho_1 = 0.1, \& \rho_2 = 0.01$															
μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	$1/r_U$	$1/r_L$	p00	p11	p22	E_U^{RO}	E_L^{RO}	RNB^{RO}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	8.000	8.264	0.599	0.139	0.540	1.066	-0.191	0.209
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	8.000	8.333	0.637	0.196	0.560	1.133	-0.268	0.288
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	7.813	8.696	0.691	0.383	0.691	1.231	-0.409	0.414
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	7.937	8.929	0.841	0.683	0.841	1.498	-0.696	0.699
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	6.757	14.49	0.686	0.534	0.726	1.221	-1.028	0.539
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	8.000	8.264	0.691	0.573	0.464	1.231	-0.363	0.394
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	10.989	2.558	0.892	0.530	0.464	1.588	-0.089	0.429
$\rho_0 = 0.8, \rho_1 = 0.1, \& \rho_2 = 0.1$															
μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	$1/r_U$	$1/r_L$	p00	p11	p22	E_U^{RO}	E_L^{RO}	RNB^{RO}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	3.759	4.405	0.599	0.139	0.540	0.958	-0.240	0.256
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	3.690	4.673	0.637	0.196	0.560	1.019	-0.355	0.350
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50,1.50	3.401	6.579	0.691	0.383	0.691	1.106	-0.822	0.532
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	3.663	7.937	0.841	0.683	0.841	1.346	-1.333	0.768
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	3.067	11.90	0.686	0.534	0.726	1.097	-1.880	0.605
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	3.774	4.608	0.691	0.573	0.464	1.106	-0.429	0.439
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.23, 4.36)	5.076	1.447	0.892	0.530	0.464	1.427	-0.117	0.513
$\rho_0 = 0.7, \rho_1 = 0.2, \& \rho_2 = 0.1$															
μ_0	μ_1	μ_2	σ_0	σ_1	σ_2	J	(C1, C2)	$1/r_U$	$1/r_L$	p00	p11	p22	E_U^{RO}	E_L^{RO}	RNB^{RO}
0.0	0.5	0.7	1.0	1.0	1.0	0.14	(0.25, 0.60)	2.242	2.488	0.599	0.139	0.540	0.838	-0.185	0.237
0.0	0.7	1.0	1.0	1.0	1.0	0.20	(0.35, 0.85)	2.212	2.584	0.637	0.196	0.560	0.892	-0.266	0.326
0.0	1.0	2.0	1.0	1.0	1.0	0.38	(0.50, 1.50)	2088	3.155	0.691	0.383	0.691	0.968	-0.515	0.487
0.0	2.0	4.0	1.0	1.0	1.0	0.68	(1.00, 3.00)	2.198	3.484	0.841	0.683	0.841	1.178	-0.823	0.741
0.0	1.5	2.7	1.3	1.0	1.0	0.47	(0.63, 2.10)	1.848	5.464	0.686	0.534	0.726	0.960	-1.198	0.579
0.0	1.0	2.0	1.0	1.0	2.0	0.36	(0.50, 2.18)	2.247	2.558	0.691	0.573	0.464	0.968	-0.336	0.421
0.0	2.0	4.0	1.0	2.0	4.0	0.44	(1.24, 4.36)	3.040	0.797	0.892	0.530	0.464	1.249	-0.088	0.480

Tables 4.2-4.8 contain different numerical parameters, results of the lower bound, the upper bound, the average, and the proposed form of r. Also, they include correct classification rate, expected utility of rule-in and rule-out, expected loss of rule-in and rule-out, and the relative net benefit of the rule-in and rule-out under different prevalence settings. The proposed *r* values decrease when increasing τ_{00} but decrease τ_{11}, and τ_{22}. This observation implies that the test is more specific, which means that this test is better at detecting the absence of diseases and has a positive net benefit. On the other hand, the proposed *r* values increase when decreasing τ_{00} but increasing τ_{11}, and τ_{22}. This observation implies that the test is more sensitive, indicating that the test has more power to detect the presence of diseases and still has a positive net benefit.

Figures 4.1, 4.2, and 4.3 show the plots of relative net benefit for rule-in by utility ratio (*r*) under different prevalence settings to gain insight into decision-theoretic benefit-risk. These figures indicate that all chosen values of *r* for the lower bound, upper bound, average, and proposed *r* show a similar pattern of relative net benefit (RNB) values for different prevalence settings. However, the values of RNB are different depending on prevalence settings and the underlying distribution set of parameters of the stages of a disease.

For the upper bound of *r* with different prevalence settings, *r* values are greater with larger τ_{00} but smaller τ_{11}, and τ_{22}, implying the test is more specific indicating that the test is better at detecting the absence of disease. Also, I notice that the test has zero net benefits, which is fixed for every case at the upper bound of *r*. The test at the upper bound of *r* balances benefit to risk, and RNB will be negative if *r* is larger than the upper bound of r. The value of *r* is increasing $\tau_{td}(0.599,\ 0.139,\ 0.540)$ if we compare it with $\tau_{td}(0.637,\ 0.196,\ 0.560)$. This means that larger τ_{00} but smaller τ_{11}, and τ_{22}, implying the test has better specificity, better at detecting the absence of disease. I can also see that when all $\tau_{00}, \tau_{11},$ and τ_{22} are increasing at every point, RNB increases.

For the lower bound, the average, and the proposed *r* with different prevalence settings, relative net benefit values are maximized for larger $\tau_{00}, \tau_{11},$ and τ_{22}, implying the test has high correct classification rates of disease stages. In all prevalence settings, RNB is higher for the lower bound of *r*, except when the proposed value of *r* is smaller than the lower bound of *r*. The value of r increases $\tau_{td}(0.599,\ 0.139,\ 0.540)$ if I compare it with $\tau_{td}(0.637,\ 0.196,\ 0.560)$ in terms of net benefit, which means larger τ_{00}, but smaller $\tau_{11},$ and τ_{22}, implying the test is more specific and has a better ability to detect the absence of disease. More specific tests enable the health care provider not to send a patient for treatment, thus lowering the possibility of adverse events of treating a non-diseased subject. I can also see that when all $\tau_{00}, \tau_{11},$ and τ_{22} are increasing at every point, the relative net benefit increases for every prevalence setting. Hence, the test with higher correct classifications rates has better RNB, reducing adverse events when treating the non-diseased and treating the correct stage of the disease.

Tables 4.6-4.8 present rule-out as the proposed utility ratio (1/*r)*= TN: TP values decrease when increasing τ_{00} but decreasing $\tau_{11},$ and τ_{22}. This observation implies that the test is more sensitive, which means that this test is better at detecting the presence of disease and has a positive net benefit. On the other hand, 1/*r* values increase when they τ_{00} decrease but $\tau_{11},$ and τ_{22} increase. This observation implies that the test is more specific, indicating that the test has more power to detect the absence of disease and still has a positive net benefit.

Figures 4.4, 4.5, and 4.6 show the plots of relative net benefit for rule-out by utility ratio (1/*r*)=TN: TP under different prevalence settings to gain insight into decision-theoretic benefit-risk. These figures indicate that all choices of the values of 1/*r* for the lower bound, upper bound, average, and proposed utility ratio (1/*r*)=TN: TP, which showed a similar pattern of relative net benefit (RNB) values for different prevalence settings. However, the values of RNB are different depending on prevalence settings and the underlying distribution set of parameters of disease stages.

For the upper bound of *r* with different prevalence settings, the 1/*r* value is greater and constant with larger τ_{00} but smaller τ_{11}, and τ_{22}, implying the test is more sensitive indicating that the test is better at detecting the presence of disease. Also, I notice that the test has a constant value of 1/r, which is fixed for every case at the upper bound of *r*, which means the test at the upper bound of 1/*r* balances the benefit to risk. Also, the RNB is changing with a constant value of 1/*r*. Additionally, when all τ_{00}, τ_{11}, and τ_{22} are increasing at every point, the implication is that RNB increases.

For the lower bound, the average, and the proposed 1/*r* with different prevalence settings, relative net benefit values are maximized for larger τ_{00}, τ_{11}, and τ_{22}, implying the test has high correct classification rates of disease stages. In all prevalence settings, RNB is higher for the lower bound of 1/*r*. The value of the relative net benefit increases $\tau_{td}(0.599,\ 0.139,\ 0.540)$ if I compare it with $\tau_{td}(0.637,\ 0.196,\ 0.560)$ terms of the net benefit. This results in larger τ_{00} but smaller τ_{11}, and τ_{22}, implying the test is more sensitive and has a better ability to detect the presence of disease. More sensitive tests are most likely to send a patient to treatment, lowering the adverse events of not treating a diseased subject. When all τ_{00}, τ_{11}, and τ_{22} are increasing at every point, relative net benefit increases for every prevalence setting. Hence, the test with higher correct classification rates has better RNB, reducing the adverse events of not treating the diseased while at the same time treating the proper stage of the disease.

Accordingly, my proposed measures have the advantage of indicating which biomarker to be used based on the diagnostic purpose to identify rule-in or rule-out patients.

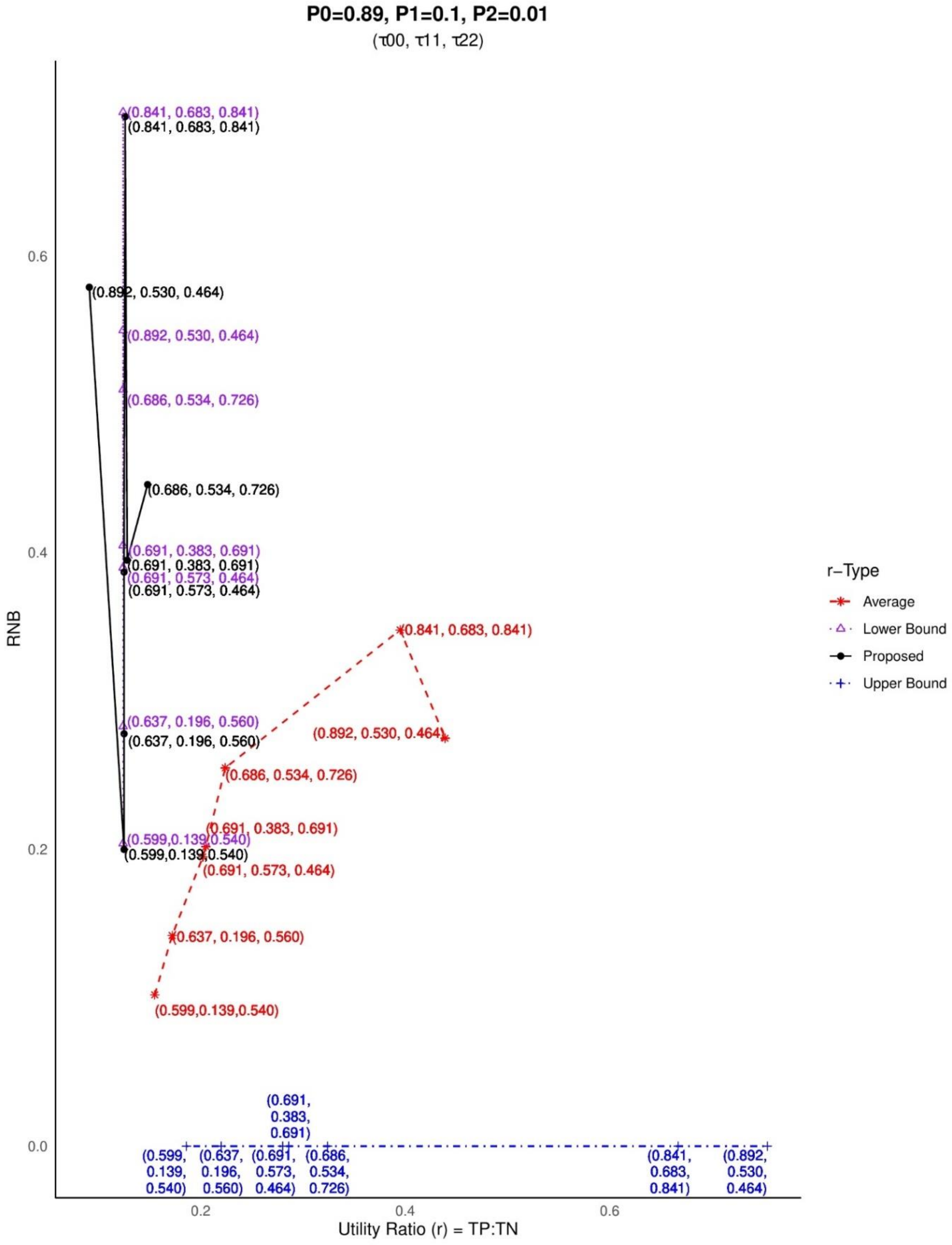


Figure 4.1. *Prevalence 1: Relative Net Benefit for Rule-In by Utility Ratio (r) = TP:TN*

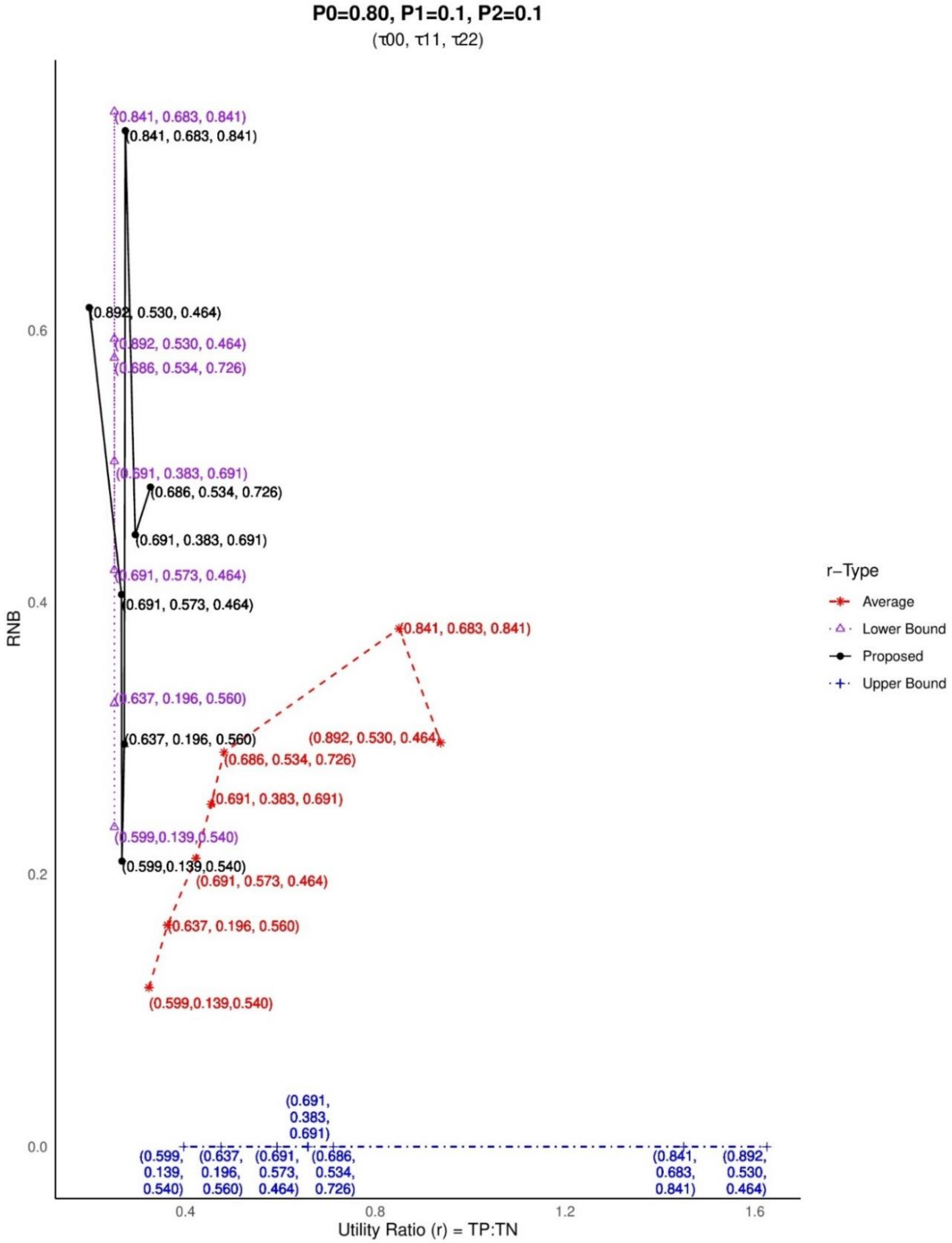


Figure 4.2. *Prevalence 2: Relative Net Benefit for Rule-In by Utility Ratio (r) = TP:TN*

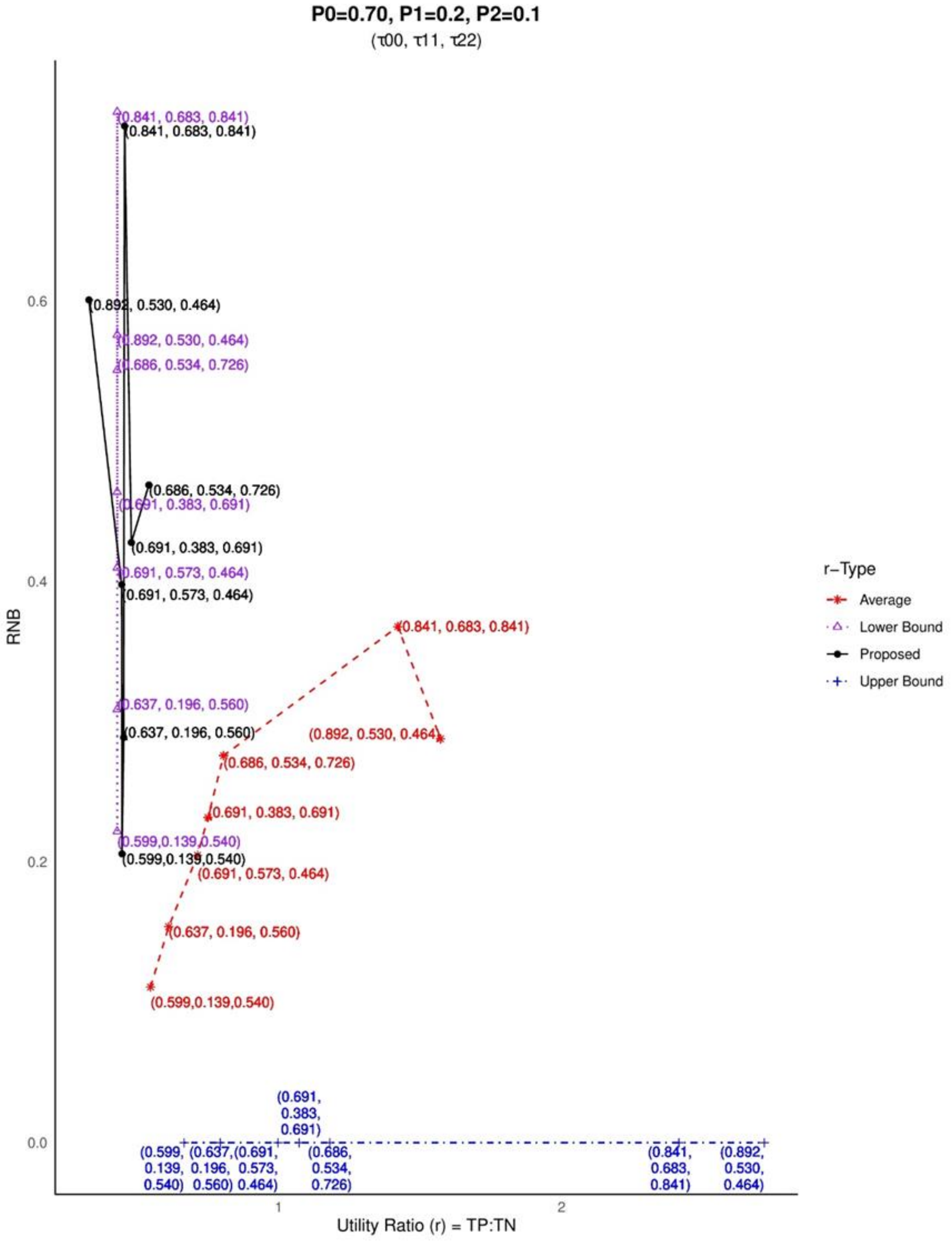


Figure 4.3. *Prevalence 3: Relative Net Benefit for Rule-In by Utility Ratio (r) = TP:TN*

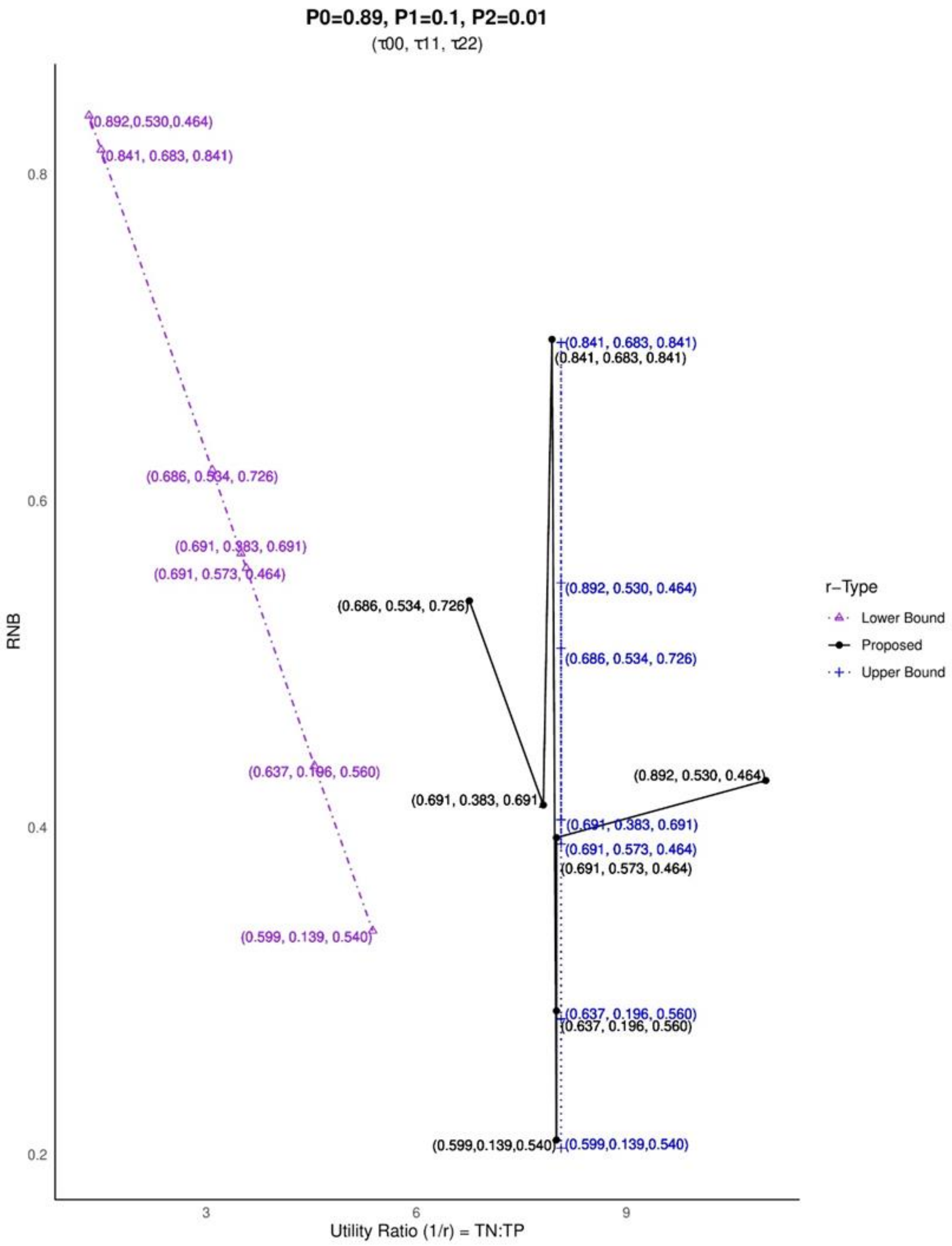


Figure 4.4. *Prevalence 1: Relative Net Benefit for Rule-Out by Utility Ratio (1/r)=TN:TP*

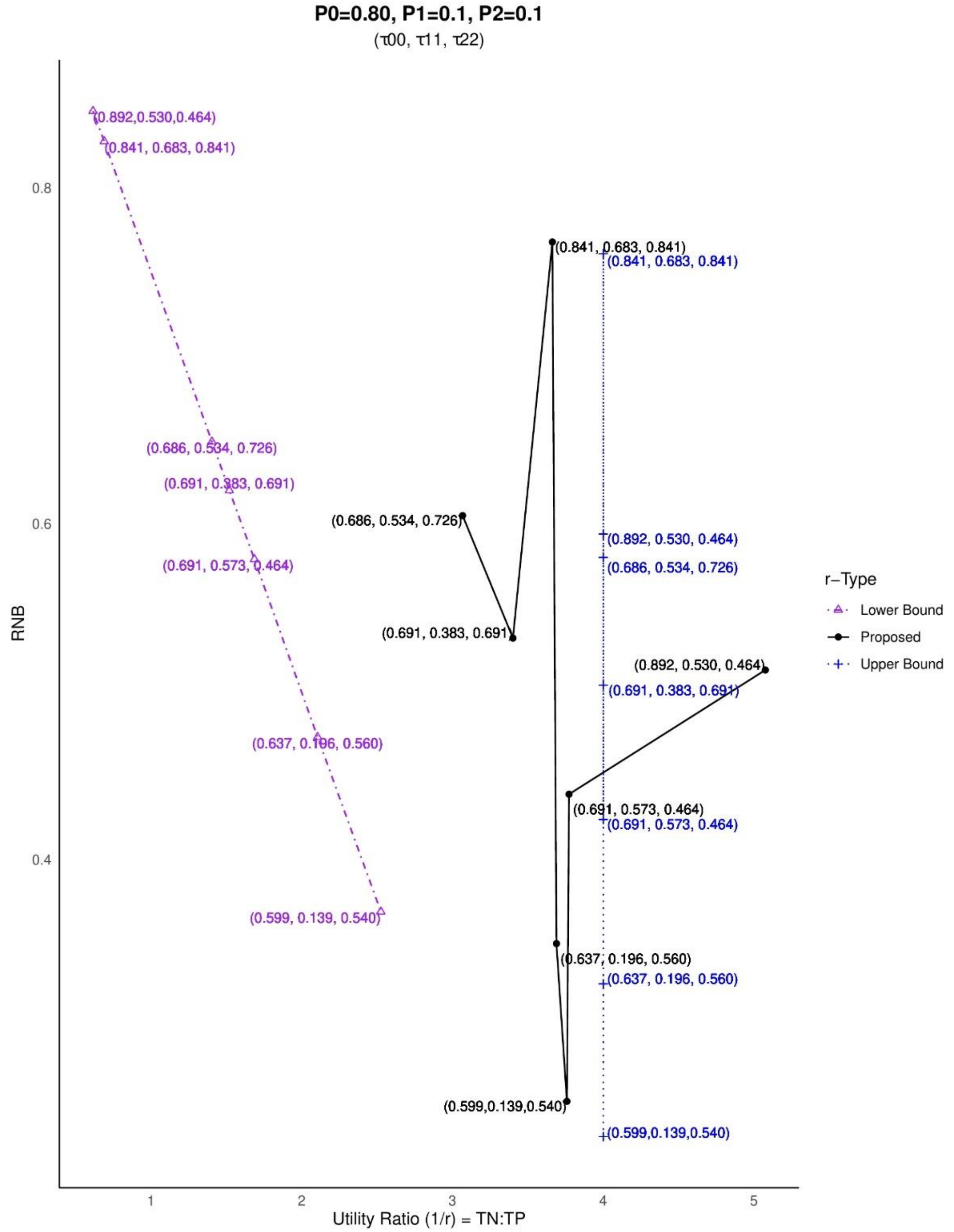


Figure 4.5. *Prevalence 2: Relative Net Benefit for Rule-Out by Utility Ratio (1/r)=TN:TP*

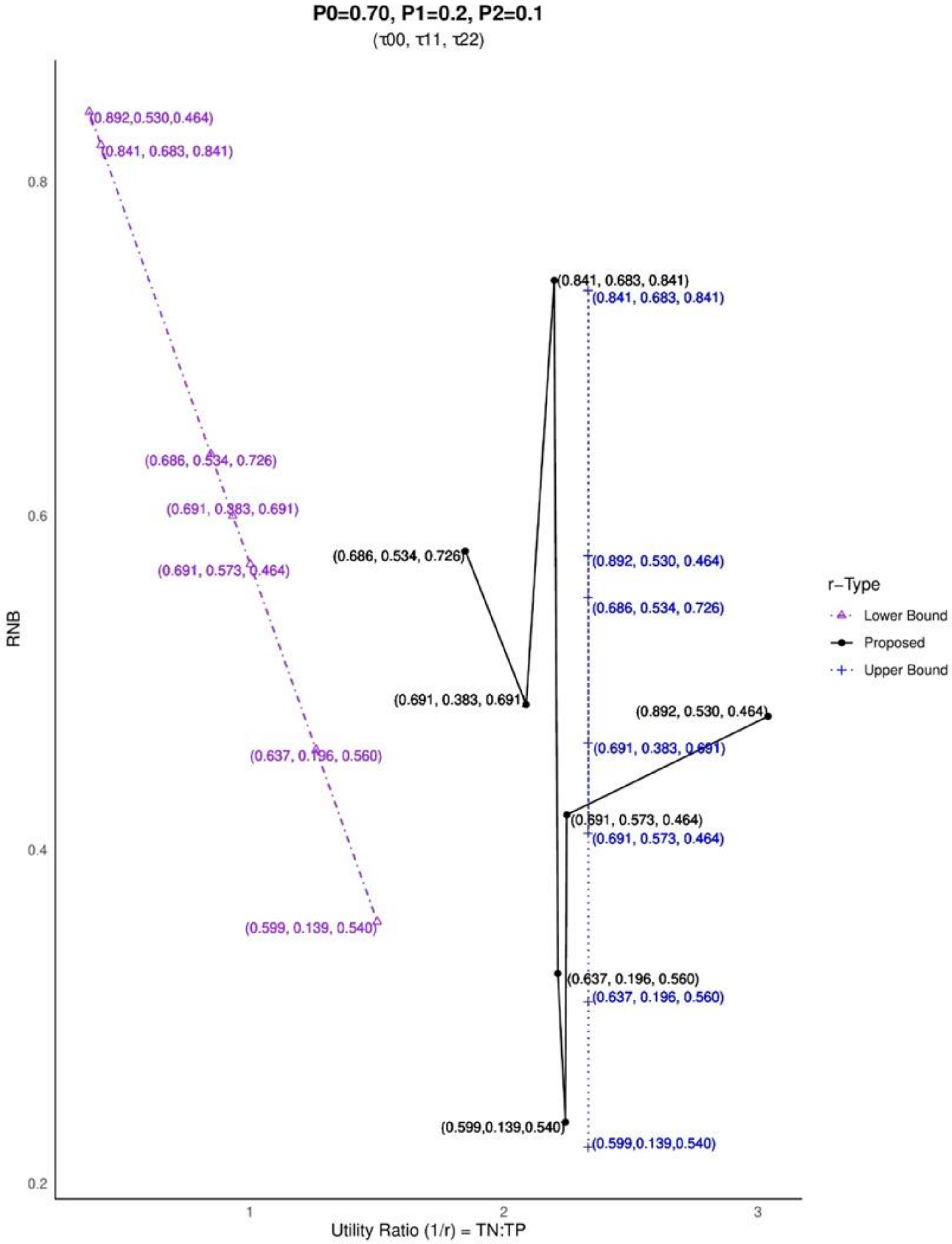


Figure 4 6. *Prevalence 3: Relative Net Benefit for Rule-Out by Utility Ratio (1/r)=TN:TP*

CHAPTER 5

REAL DATA ANALYSIS (ADNI DATA)

I use a dataset from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) to demonstrate the application of the net benefit approach to evaluating the biomarkers for the diagnosis of Alzheimer’s Disease (AD) on its diagnostic accuracy based on a clinical performance study, along with external information on clinical consequences. Consequently, I describe the diagnostic yield table for the multi-stage clinical condition of Alzheimer's Disease. I develop a decision theory based on net benefit for evaluating the biomarkers, which provides additional interpretation for rule-in or rule-out clinical needs, as well as their adverse consequences from unnecessary workup in multi-stage Alzheimer’s Disease.

5.1 Introduction of Alzheimer’s Disease and Dementia

Alzheimer's Disease (AD), a complex and progressive neurodegenerative disease, is the common form of dementia among seniors. AD damages mental and memory functions and eventually includes physical disability due to neurons' death and brain tissue deterioration. Based on 2022 Alzheimer’s Disease Facts and Figures information, AD is one of the most common causes of dementia, with 60% to 80% of cases occurring among 6.5 million Americans aged 65 and older. Approximately seventy-three percent are age 75 or older, and about 1 in 9 people (10.7%) age 65 and older have Alzheimer’s Dementia. The sixth-leading cause of death in the USA is AD. It is also predicted that every state in the United States will experience an increase of at least 6.7% in the number of people with Alzheimer’s between 2020 and 2025, with treatment costs increasing for patients during those years.

Dementia is a primary term that is used for describing memory impairment. Yet, dementia due to Alzheimer’s disease is characterized by noticeable memory, language, thinking, or behavioral symptoms that impair a person’s ability to function in daily life. When combined with biomarker evidence of Alzheimer's-related brain changes, these brain and behavioral symptoms yield a more accurate Alzheimer's diagnosis ("2022 Alzheimer's disease facts and figures," 2022). Individuals commonly experience multiple symptoms that change over time as Alzheimer’s progresses. These symptoms represent the degree of damage to neurons in different parts of the brain, as well as the pace at which dementia is advancing from mild to moderate to severe, person to person. ADNI provides data that tracks the progression of Alzheimer’s disease over time, using biomarkers and clinical measures (Alzheimer’s Disease Neuroimaging Initiative (ADNI), 2017).

Not all individuals with evidence of Alzheimer’s-related brain changes go on to develop symptoms of Mild Cognitive Impairment (MCI) or dementia due to Alzheimer’s. However, people with MCI tend to have a higher risk of developing Alzheimer's or other Dementia (Alzheimer's Association, 2020). According to Johns Hopkins Medicine (2019b), Alzheimer's disease typically develops slowly and gradually in four general stages; preclinical stage, mild (early stage), moderate (middle-stage), and severe (late-stage). In the preclinical stage, individuals may have measurable brain changes that indicate the earliest signs of Alzheimer’s disease, but they have not yet developed symptoms such as memory loss. In another study, MAYO Clinic (2020) presents a more transitional stage: mild cognitive impairment (MCI). They name five progressive stages; preclinical stage, mild cognitive impairment (MCI), mild Dementia, moderate Dementia, and severe Dementia. In this application for my proposed method, we consider three clinical disease stages for Alzheimer's disease: Cognitively Normal (non-diseased), MCI (early diseased), and dementia (fully diseased), as suggested by Alzheimer’s Disease Neuroimaging Initiative (ADNI) (2020). The gold standard to determine AD stages is based on the following global clinical dementia rating (CDGLOBAL). CDGLOBAL 0, 0.5, and 1, and greater than 1 (2 or 3) indicate cognitively normal (non-diseased), MCI (early diseased), and dementia (fully diseased), respectively.

5.2 Data Analysis

5.2.1 Data File from ADNI

The data files for this study are obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI is a global longitudinal multicenter study that unites researchers to collect, validate and utilize data, including MRI and PET images, genetics, cognitive tests, CSF, and blood biomarkers for early detection and tracking of the progression of AD. The main goals of the ADNI study are to detect and track the disease's progression with biomarkers of AD and to support AD intervention advancement, prevention, and treatment through the application of new diagnostic methods at the earliest possible stage of AD. It also provides all data to all scientists in the world without limitation. ADNI was started in 2004 under the leadership of Dr. Michael W. Weiner and funded as a private-public partnership with contributions from 20 companies and two foundations. The primary goal of ADNI is to determine the relationships between clinical, brain imaging measures, and biochemical biomarkers through the progress of AD. It also adds brain scans that detect tau protein tangles (tau PET), a vital indicator of the disease. ADNI also continues the discovery, optimization, standardization, and validation of clinical trial measures and biomarkers used in AD research.

To demonstrate the application of the net benefit approach to evaluating biomarkers for the diagnosis of multi-stage Alzheimer’s Disease based on its diagnostic accuracy and clinical consequences of diagnostic errors, ADNI-1 data collected between September 2005 to August 2007 are included. Kersey previously used the ADNI dataset to measure diagnostic accuracy with cut-point criteria for multi-stage diseases based on concordance and discordance (Kersey, Samawi, Yin, Rochani, & Zhang, 2022). I am using the same dataset to compare biomarkers of Alzheimer’s Disease by using our proposed approach. The dataset consists of 415 subjects with 114, 256, and 45 subjects for the non-diseased (Cognitively Normal, or CN), the early diseased (Mild Cognitive Impairment, or MCI), and the fully diseased (Dementia) groups, respectively.

The data file presents test results from five core biomarkers for Alzheimer's disease, including three biomarkers from core cerebrospinal fluid (CSF) and two from magnetic resonance imaging (MRI). Blennow, Hampel, Weiner, and Zetterberg (2010) summarize that the core CSF biomarkers reflect AD pathology, evaluate disease risk or prognosis, and have high diagnostic accuracy when diagnosing AD with dementia and prodromal AD in mild cognitive impairment cases, along with monitoring therapeutic interventions 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 (ABETA142). These three CSF biomarkers are the central pathogenic processes in AD and have been proposed as candidate markers for predicting cognition decline as the progression indicator of dementia. Previous studies also discuss the other two potential biomarkers of Alzheimer's disease measured from MRI: rate of volume change of the Hippocampus and whole brain. Imaging has a significant role in improving our understanding of this disease. Studies present the relationship between volume change and the initiative of Alzheimer's disease and how these biomarkers change over time, relating to the injury and death of neurons (Duthey, 2013; Grundman & Delaney, 2002; Shaffer et al., 2013).

5.2.2 Biomarker Selection

ADNI uses five core biomarkers to help predict the onset of AD over the progression of clinical disease stages. I am interested in biomarkers of Alzheimer's disease measured from CSF variables, including ABETA142, PTAU, and TAU. I also include in our analysis the other two potential biomarkers of Alzheimer's disease measured from MRI, including hippocampus volume and brain volume, because these biomarkers are related to the severity of cognitive impairment (Vijayakumar & Vijayakumar, 2013).

5.2.3 Analysis of ADNI Data

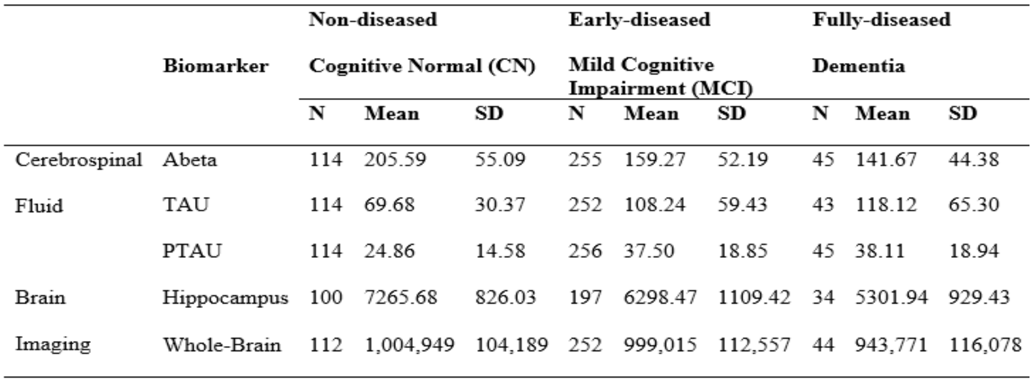
I apply the generalized Youden index (GYI) measure of diagnostic accuracy to ADNI-1 data. I also apply GYI criteria for cut-point selection to the dataset to find the corresponding optimal cut-points. Using the gold standard in the dataset (CDGLOBAL), the prevalence of Alzheimer’s disease in different stages is approximated as $0.72(p_1), 0.2(p_2)$, and $0.08(p_3)$ for stages 1, 2, and 3, respectively, where $p_1, p_2,$ and p_3 are the prevalence of stage 1 (CN), stage 2 (MCI), and stage 3 (Dementia), respectively, based on estimates from Kantarci et al. (2009); Mitchell and Shiri-Feshki (2009); Roberts and Knopman (2013).

My primary goal is to illustrate the application of the net benefit approach to evaluating the five biomarkers of interest: hippocampus volume (Hippocampus), brain volume (WholeBrain), Total tau (TAU), Aβ1-42 (ABETA142), and p-tau181 (PTAU181P) of Alzheimer's disease based on diagnostic accuracy and clinical consequences of diagnostic errors. Another goal is to compare those biomarkers using my proposed measure based on the diagnostic purpose to identify rule-in or rule-out patients. I use the correct classification rate (CCR) for each stage, misclassification rates over three stages, and the benefit and the loss to evaluate the performances of the biomarkers of interest using the above-proposed methods. I calculate the values of the lower bound, the upper bound, and the proposed form of the Cost-Benefit Ratio (r). I also calculate the expected utility for rule-in and rule-out, expected loss for rule-in and rule-out, and relative net benefit for rule-in and rule-out to evaluate the benefits of the five biomarkers of interest.

5.2.4 Results of ADNI Data

The dataset consists of a total of 415 subjects with 114, 256, and 45 subjects for the non-diseased (Cognitively Normal, or CN), the early diseased (Mild Cognitive Impairment, or MCI), and the fully diseased (Dementia) groups, respectively. The actual sample sizes for biomarkers in the dataset may vary and are smaller than the group sizes due to some missing values. Table 5.1 presents the summary descriptive statistics of the five interested biomarkers in the ADNI dataset. Table 5.1 shows that Hippocampus, WholeBrain, and ABETA142 average values decrease as the severity of the disease increases. This indicates that the lower the value of the biomarker is, the more severe the disease is. In contrast, the average values of the rest biomarkers, including PTAU181P and TAU, increase when the disease progresses to later stages.

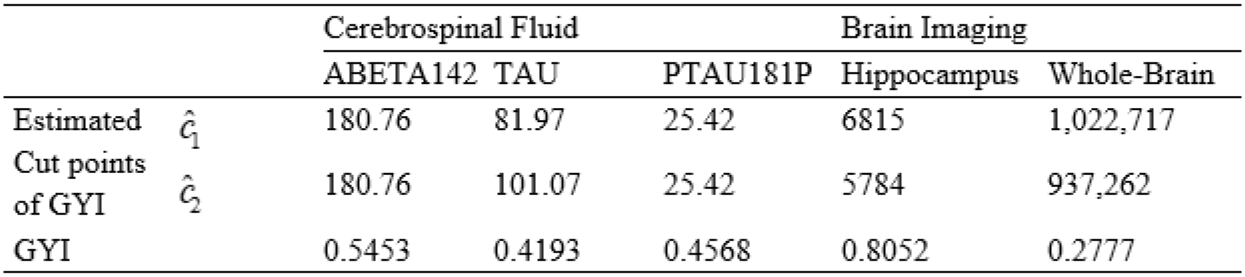
Table 5.1. *Summary of Descriptive Statistics of Five Biomarkers (source: Kersey et al., 2022)*



		Non-diseased			Early-diseased			Fully-diseased		
	Biomarker	Cognitive Normal (CN)			Mild Cognitive Impairment (MCI)			Dementia		
		N	Mean	SD	N	Mean	SD	N	Mean	SD
Cerebrospinal	Abeta	114	205.59	55.09	255	159.27	52.19	45	141.67	44.38
Fluid	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	Hippocampus	100	7265.68	826.03	197	6298.47	1109.42	34	5301.94	929.43
Imaging	Whole-Brain	112	1,004,949	104,189	252	999,015	112,557	44	943,771	116,078

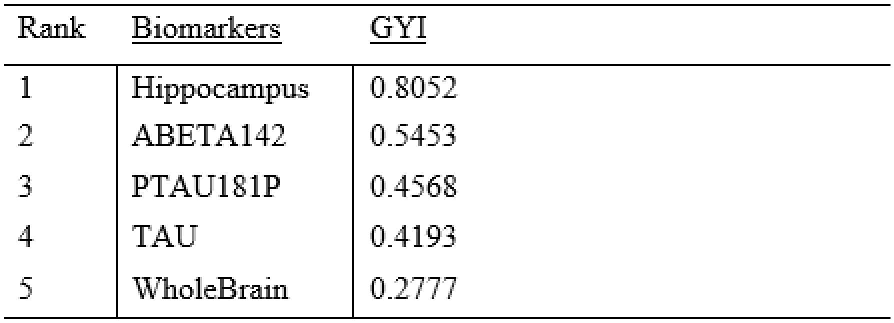
The optimal statistics of GYI are calculated, and corresponding optimal cut-points $(c_1$ and $c_2)$ are shown in Table 5.2. Based on estimated statistics, the five biomarkers are ranked in Table 5.3. The best biomarker is Hippocampus using the GYI measure of diagnostic accuracy, followed by ABETA142, PTAU181P, TAU, and WholeBrain. Hippocampus has the highest estimated GYI measure of diagnostic accuracy among the five biomarkers. Thus, Hippocampus is the best biomarker to use to discriminate subjects among the three clinical stages of Alzheimer's disease. WholeBrain is the least favorable biomarker to use to distinguish subjects among the three stages of Alzheimer's disease.

Table 5.2. *Estimated Optimal Statistics and Corresponding Cut-points for Five Biomarkers*



		Cerebrospinal Fluid			Brain Imaging	
		ABETA142	TAU	PTAU181P	Hippocampus	Whole-Brain
Estimated Cut points of GYI	$\hat{c}_1$	180.76	81.97	25.42	6815	1,022,717
	$\hat{c}_2$	180.76	101.07	25.42	5784	937,262
GYI		0.5453	0.4193	0.4568	0.8052	0.2777

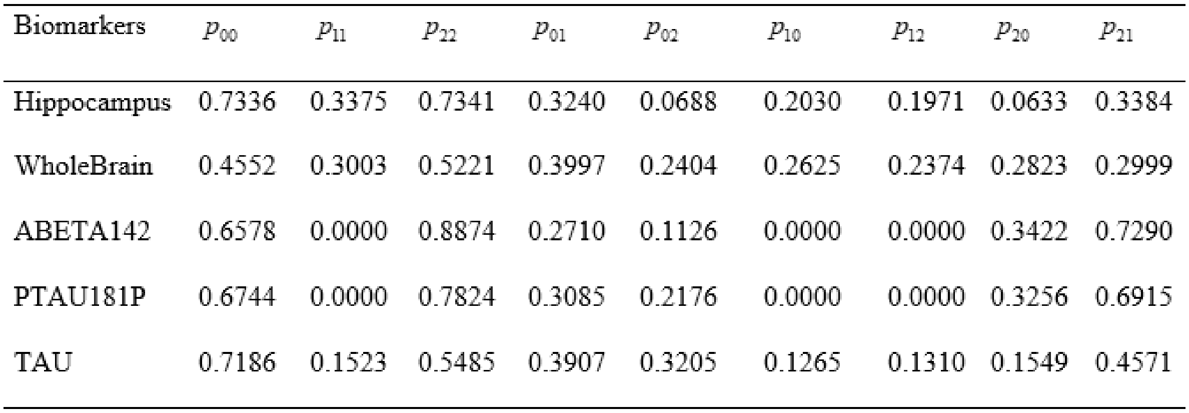
Table 5.3. *The Rank of Biomarkers with Different Diagnostic Measures*



Rank	Biomarkers	GYI
1	Hippocampus	0.8052
2	ABETA142	0.5453
3	PTAU181P	0.4568
4	TAU	0.4193
5	WholeBrain	0.2777

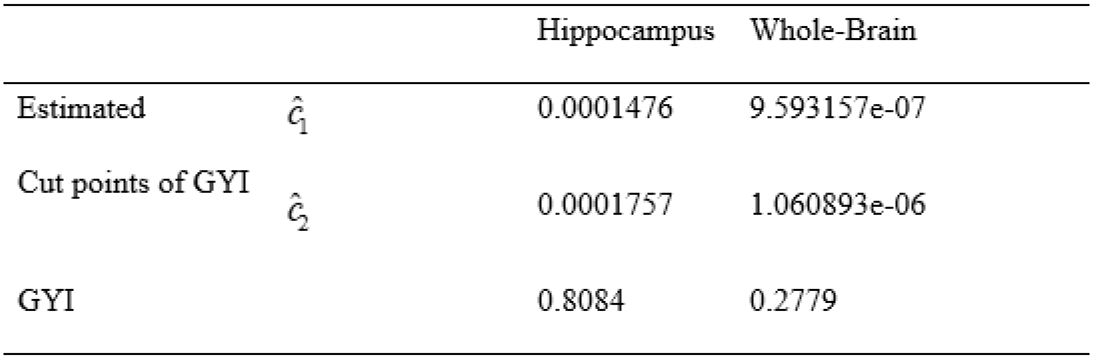
Correct classification rates $(p_{11}, p_{22},$ and $p_{33})$ and corresponding misclassification rates $(p_{01}, p_{02}, p_{10}, p_{12}, p_{20},$ and $p_{21})$ of the five interested biomarkers are calculated and presented in Table 5.4. The estimated optimal cut-points c_1 and c_2 of ABETA142 and PTAU181P using GYI measure criteria are identical. The corresponding correct classification rates for the second class (the early diseased or Mild Cognitive Impairment) are zeros. These results show no subject was correctly diagnosed with Mild Cognitive Impairment by the GYI criterion. Such results imply that ABETA142 and PTAU181P are not suitable biomarkers to distinguish early disease stage patients if using the GYI criterion. The nature of Hippocampus and WholeBrain biomarkers show the average values of these two biomarkers decrease as the severity of the disease increases. For this reason, the optimal statistics of the reciprocal of Hippocampus and WholeBrain using GYI are calculated, and corresponding optimal cut-points $(c_1 \text{ and } c_2)$ are shown in Table 5.5. Consequently, correct classification rates $(p_{11}, p_{22}, \text{ and } p_{33})$ and corresponding misclassification rates $(p_{01}, p_{02}, p_{10}, p_{12}, p_{20}, \text{ and } p_{21})$ of these two biomarkers are calculated and presented in Table 5.6. The correct classification rates, and corresponding misclassification rates of Hippocampus, WholeBrain, using GYI measure criteria, are the same as the correct classification rates and corresponding misclassification rates of the reciprocal Hippocampus and WholeBrain biomarkers using GYI measure criteria. I could see that the results are not affected by taking the reciprocal biomarkers. Thus, Hippocampus is the best biomarker, and WholeBrain is the least favorable biomarker to discriminate between subjects among the three stages of Alzheimer's disease.

Table 5.4. *The Corresponding Misclassification Rates for Five Biomarkers*



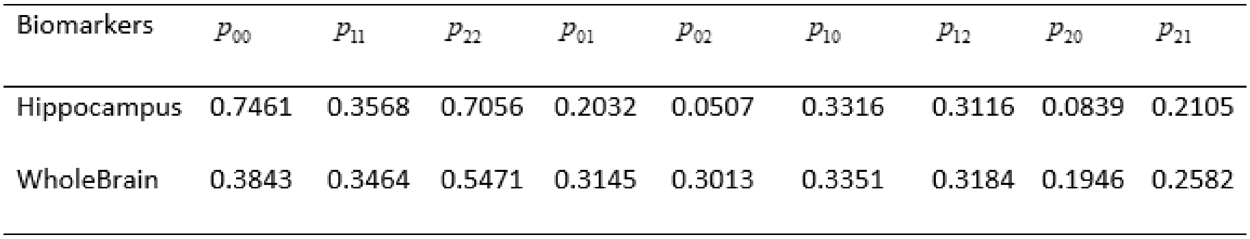
Biomarkers	p_{00}	p_{11}	p_{22}	p_{01}	p_{02}	p_{10}	p_{12}	p_{20}	p_{21}
Hippocampus	0.7336	0.3375	0.7341	0.3240	0.0688	0.2030	0.1971	0.0633	0.3384
WholeBrain	0.4552	0.3003	0.5221	0.3997	0.2404	0.2625	0.2374	0.2823	0.2999
ABETA142	0.6578	0.0000	0.8874	0.2710	0.1126	0.0000	0.0000	0.3422	0.7290
PTAU181P	0.6744	0.0000	0.7824	0.3085	0.2176	0.0000	0.0000	0.3256	0.6915
TAU	0.7186	0.1523	0.5485	0.3907	0.3205	0.1265	0.1310	0.1549	0.4571

Table 5.5. *Estimated Optimal Statistics and Corresponding Cut-points for Biomarkers*



		Hippocampus	Whole-Brain
Estimated	$\hat{c}_1$	0.0001476	9.593157e-07
Cut points of GYI	$\hat{c}_2$	0.0001757	1.060893e-06
GYI		0.8084	0.2779

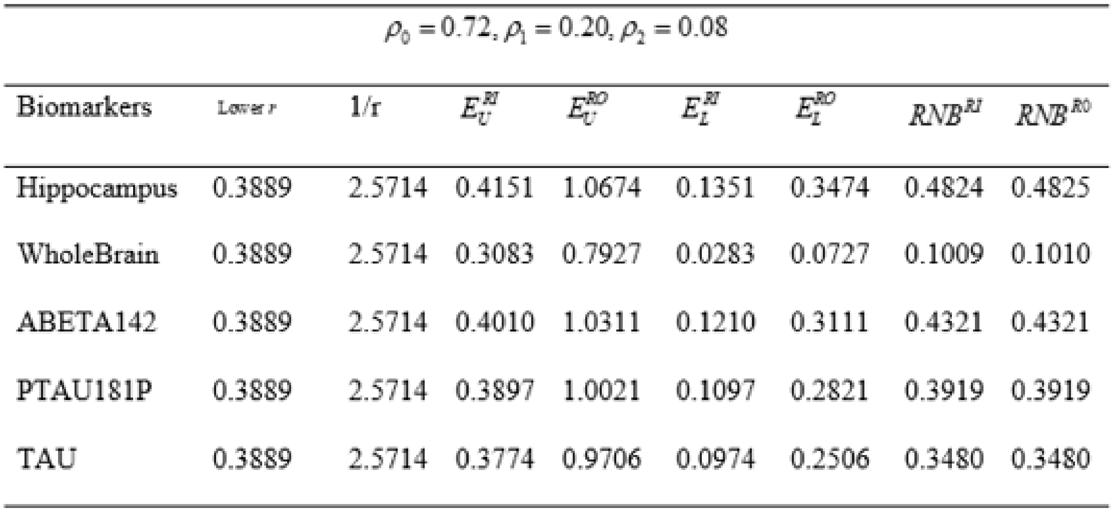
Table 5.6. *The Corresponding Misclassification Rates for Reciprocal Biomarkers*



Biomarkers	p_{00}	p_{11}	p_{22}	p_{01}	p_{02}	p_{10}	p_{12}	p_{20}	p_{21}
Hippocampus	0.7461	0.3568	0.7056	0.2032	0.0507	0.3316	0.3116	0.0839	0.2105
WholeBrain	0.3843	0.3464	0.5471	0.3145	0.3013	0.3351	0.3184	0.1946	0.2582

The expected utility of rule-in and rule-out, expected loss of rule-in and rule-out, the relative net benefit of rule-in and rule-out using the lower bound, upper bound, and proposed form of *r* for the five interested biomarkers are calculated and presented in Tables 5.7 to 5.9 respectively.

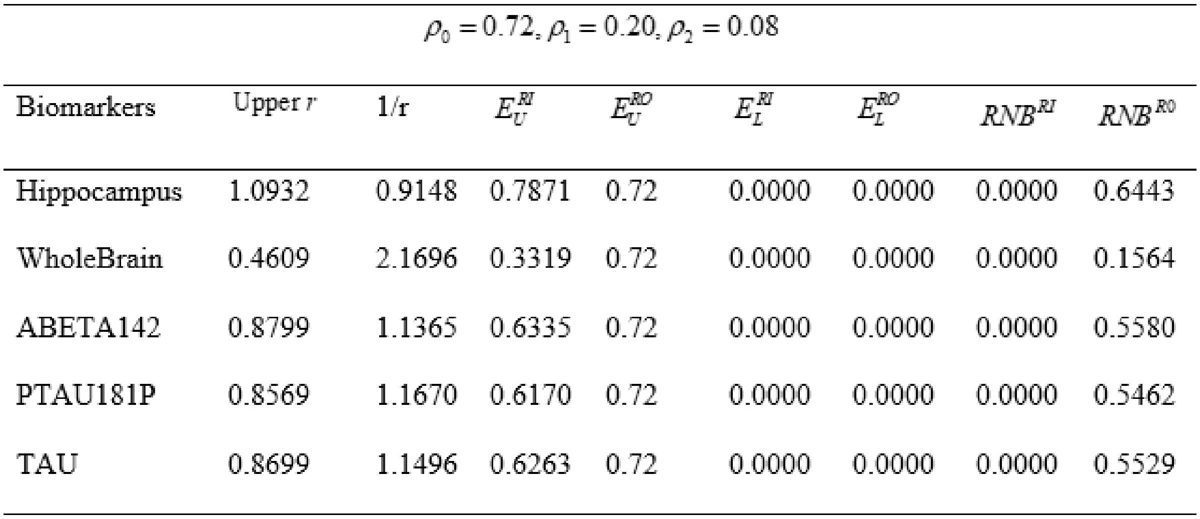
Table 5.7. *Expected Utility, Expected Loss, Relative Net Benefit Using Lower Bound of r*



$\rho_0 = 0.72, \rho_1 = 0.20, \rho_2 = 0.08$

Biomarkers	Lower r	1/r	E_U^{RI}	E_U^{RO}	E_L^{RI}	E_L^{RO}	RNB^{RI}	RNB^{R0}
Hippocampus	0.3889	2.5714	0.4151	1.0674	0.1351	0.3474	0.4824	0.4825
WholeBrain	0.3889	2.5714	0.3083	0.7927	0.0283	0.0727	0.1009	0.1010
ABETA142	0.3889	2.5714	0.4010	1.0311	0.1210	0.3111	0.4321	0.4321
PTAU181P	0.3889	2.5714	0.3897	1.0021	0.1097	0.2821	0.3919	0.3919
TAU	0.3889	2.5714	0.3774	0.9706	0.0974	0.2506	0.3480	0.3480

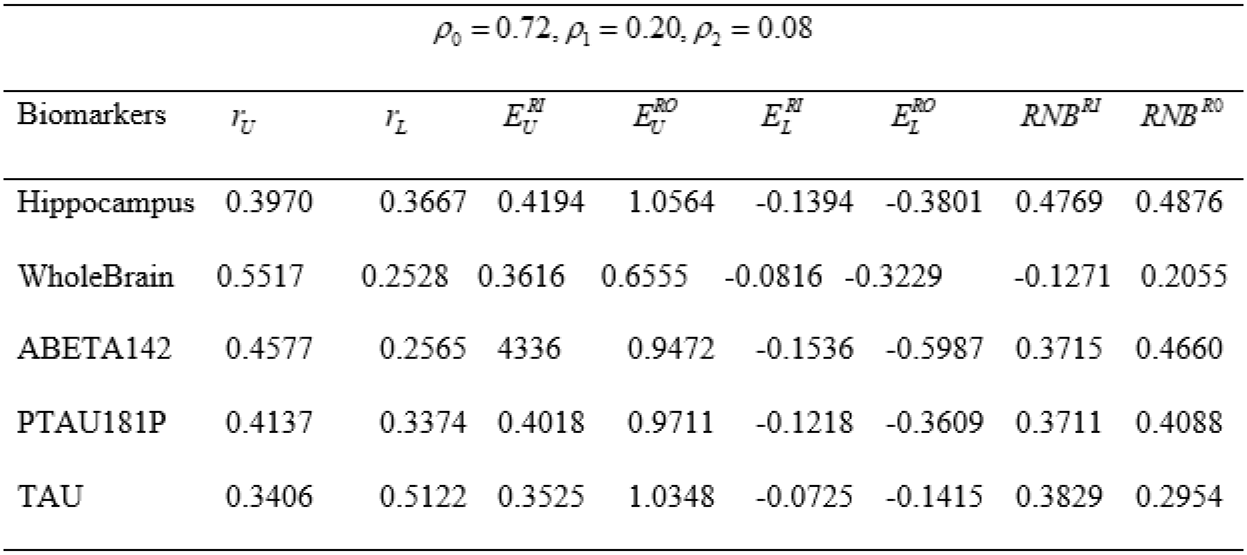
Table 5.8. *Expected Utility, Expected Loss, Relative Net Benefit Using Upper Bound of r*



$\rho_0 = 0.72, \rho_1 = 0.20, \rho_2 = 0.08$

Biomarkers	Upper r	1/r	E_U^{RI}	E_U^{RO}	E_L^{RI}	E_L^{RO}	RNB^{RI}	RNB^{R0}
Hippocampus	1.0932	0.9148	0.7871	0.72	0.0000	0.0000	0.0000	0.6443
WholeBrain	0.4609	2.1696	0.3319	0.72	0.0000	0.0000	0.0000	0.1564
ABETA142	0.8799	1.1365	0.6335	0.72	0.0000	0.0000	0.0000	0.5580
PTAU181P	0.8569	1.1670	0.6170	0.72	0.0000	0.0000	0.0000	0.5462
TAU	0.8699	1.1496	0.6263	0.72	0.0000	0.0000	0.0000	0.5529

Table 5.9. *Expected Utility, Expected Loss, Relative Net Benefit Using Proposed Form of r*

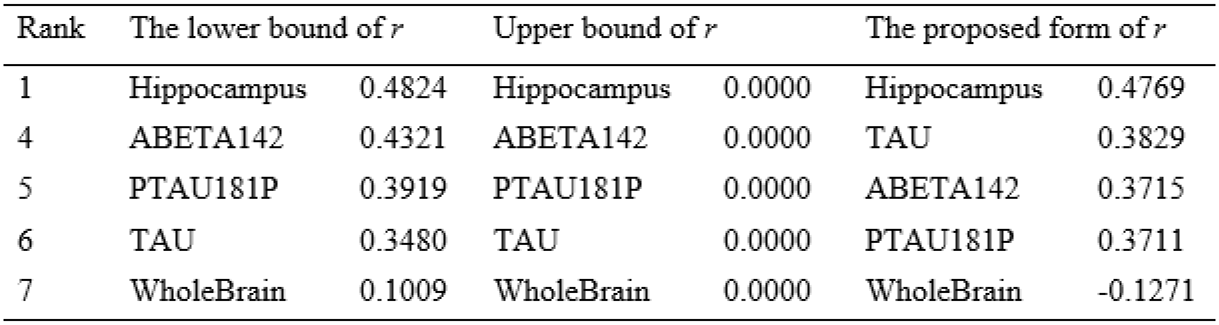


$\rho_0 = 0.72, \rho_1 = 0.20, \rho_2 = 0.08$

Biomarkers	r_U	r_L	E_U^{RI}	E_U^{RO}	E_L^{RI}	E_L^{RO}	RNB^{RI}	RNB^{RO}
Hippocampus	0.3970	0.3667	0.4194	1.0564	-0.1394	-0.3801	0.4769	0.4876
WholeBrain	0.5517	0.2528	0.3616	0.6555	-0.0816	-0.3229	-0.1271	0.2055
ABETA142	0.4577	0.2565	4336	0.9472	-0.1536	-0.5987	0.3715	0.4660
PTAU181P	0.4137	0.3374	0.4018	0.9711	-0.1218	-0.3609	0.3711	0.4088
TAU	0.3406	0.5122	0.3525	1.0348	-0.0725	-0.1415	0.3829	0.2954

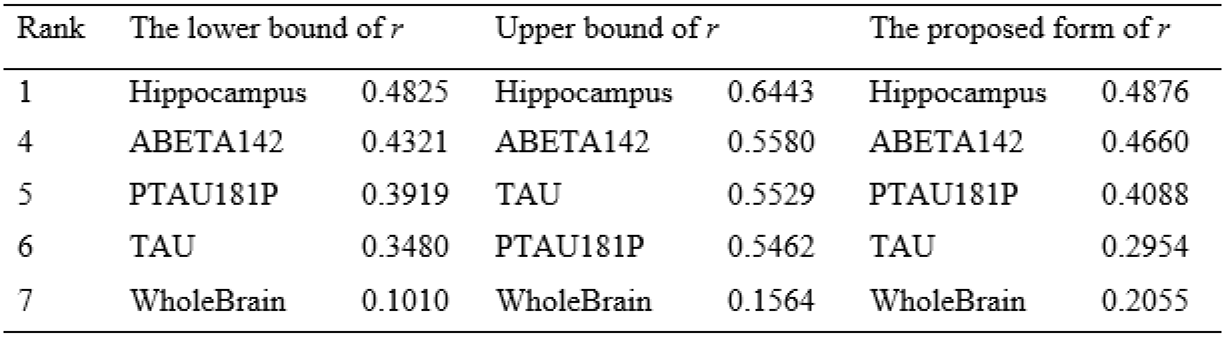
Based on the relative net benefit for rule-in and rule-out with different forms of *r*, the five biomarkers are ranked in Tables 5.10-5.11. For example, the relative net benefit for rule-in and rule-out with a lower bound of *r*, the best biomarker is Hippocampus, followed by ABETA142, PTAU181P, and TAU. Hippocampus has the highest relative net benefit among the five biomarkers for all three forms of *r*. Thus, Hippocampus is the best biomarker to discriminate between subjects among the three stages of Alzheimer's disease. All three forms of *r* agree that WholeBrain is the least favorable biomarker. In both rule-in and rule-out of the disease, ABETA142, PTAU181P, and TAU are ranked very differently using the three forms of r, while Hippocampus, ABETA142, PTAU181P, and TAU are consistent for the lower bound of r. The best biomarker for the proposed form of *r* is Hippocampus for rule-in, followed by TAU, ABETA142, PTAU181P, and WholeBrain. However, for the proposed form of a risk-benefit ratio (r), the best biomarker is Hippocampus for rule-out, followed by ABETA142, PTAU181P, TAU, and WholeBrain. In this sense, the proposed form of a risk-benefit ratio (r) performs better than other forms of *r* for deciding which biomarker to use based on the diagnostic purpose to identify rule-in or rule-out patients.

Table 5.10. The Rank of Biomarkers for Rule-In with Different Forms of r



Rank	The lower bound of *r*		Upper bound of *r*		The proposed form of *r*	
1	Hippocampus	0.4824	Hippocampus	0.0000	Hippocampus	0.4769
4	ABETA142	0.4321	ABETA142	0.0000	TAU	0.3829
5	PTAU181P	0.3919	PTAU181P	0.0000	ABETA142	0.3715
6	TAU	0.3480	TAU	0.0000	PTAU181P	0.3711
7	WholeBrain	0.1009	WholeBrain	0.0000	WholeBrain	-0.1271

Table 5.11. *The Rank of Biomarkers for Rule-Out with Different Forms of r*



Rank	The lower bound of *r*		Upper bound of *r*		The proposed form of *r*	
1	Hippocampus	0.4825	Hippocampus	0.6443	Hippocampus	0.4876
4	ABETA142	0.4321	ABETA142	0.5580	ABETA142	0.4660
5	PTAU181P	0.3919	TAU	0.5529	PTAU181P	0.4088
6	TAU	0.3480	PTAU181P	0.5462	TAU	0.2954
7	WholeBrain	0.1010	WholeBrain	0.1564	WholeBrain	0.2055

Figures 5.1 and 5.2 show the relative net benefit of the biomarkers of interest for rule-in and rule-out by the proposed risk-benefit ratio. In this case, Hippocampus is the best biomarker that can discriminate between subjects among the three stages of Alzheimer's disease, and WholeBrain is the least favorable biomarker.

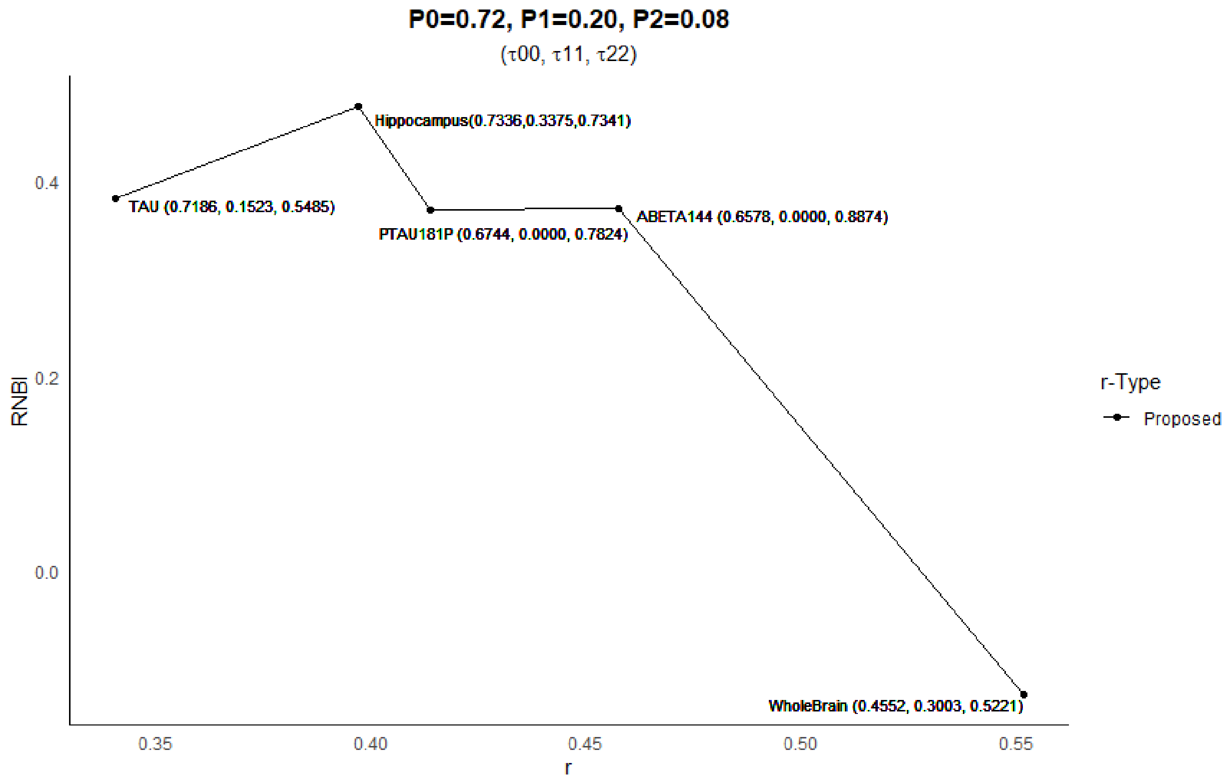


Figure 5.1. *Relative Net Benefit of Biomarkers for Rule-In by Proposed Utility Ratio (r)*

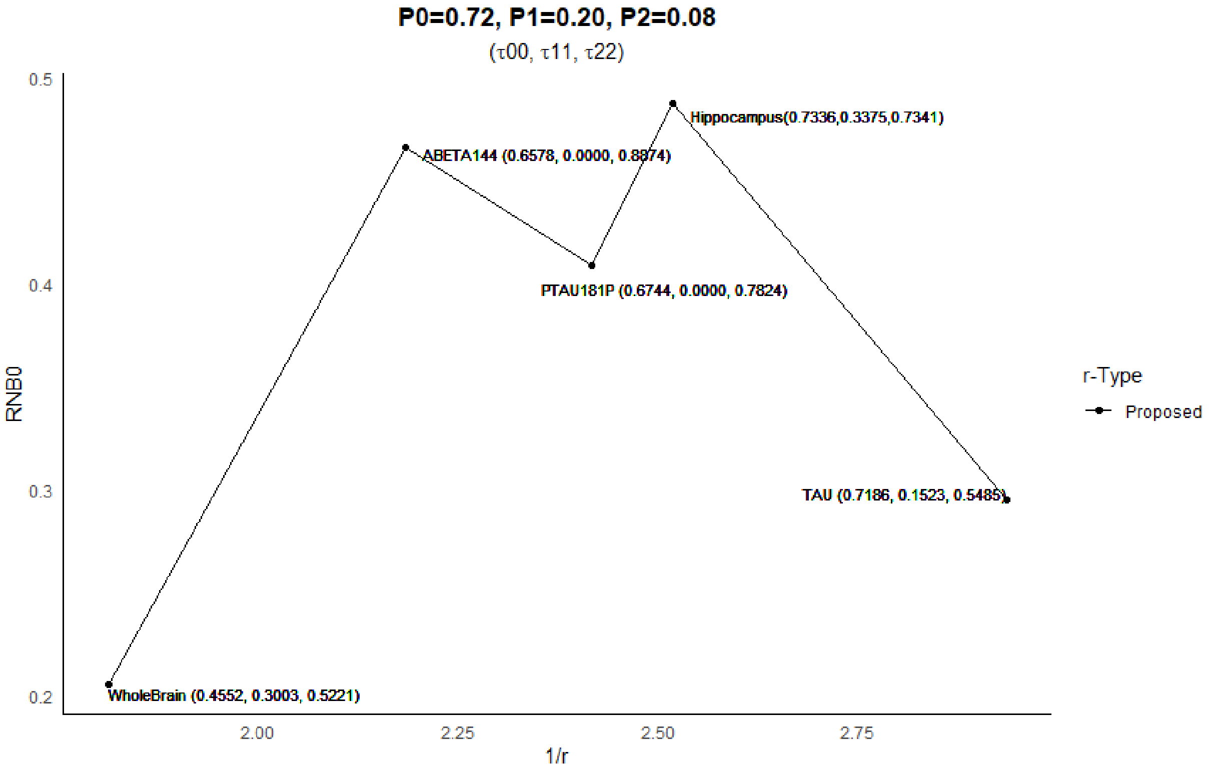


Figure 5. 2. *Relative Net Benefit of Biomarker for Rule-Out by Proposed Utility Ratio (1/r)*

The relative net benefit over never-treat and always-treat policies can be plotted as relative importance (risk-benefit) ratio *r* to compare the biomarkers (Figures 5.3–5.4). These plots indicate that the relative net benefit over the never-treat policy is noticeably worse for the WholeBrain than the Hippocampus over an extensive range of *r* values. In contrast, the relative net benefit over the always-treat policy for WholeBrain is still worse than Hippocampus over a comprehensive range of *r* values. Thus, the two tests can be considered comparable in settings where prophylactic treatment is practiced in place of testing.

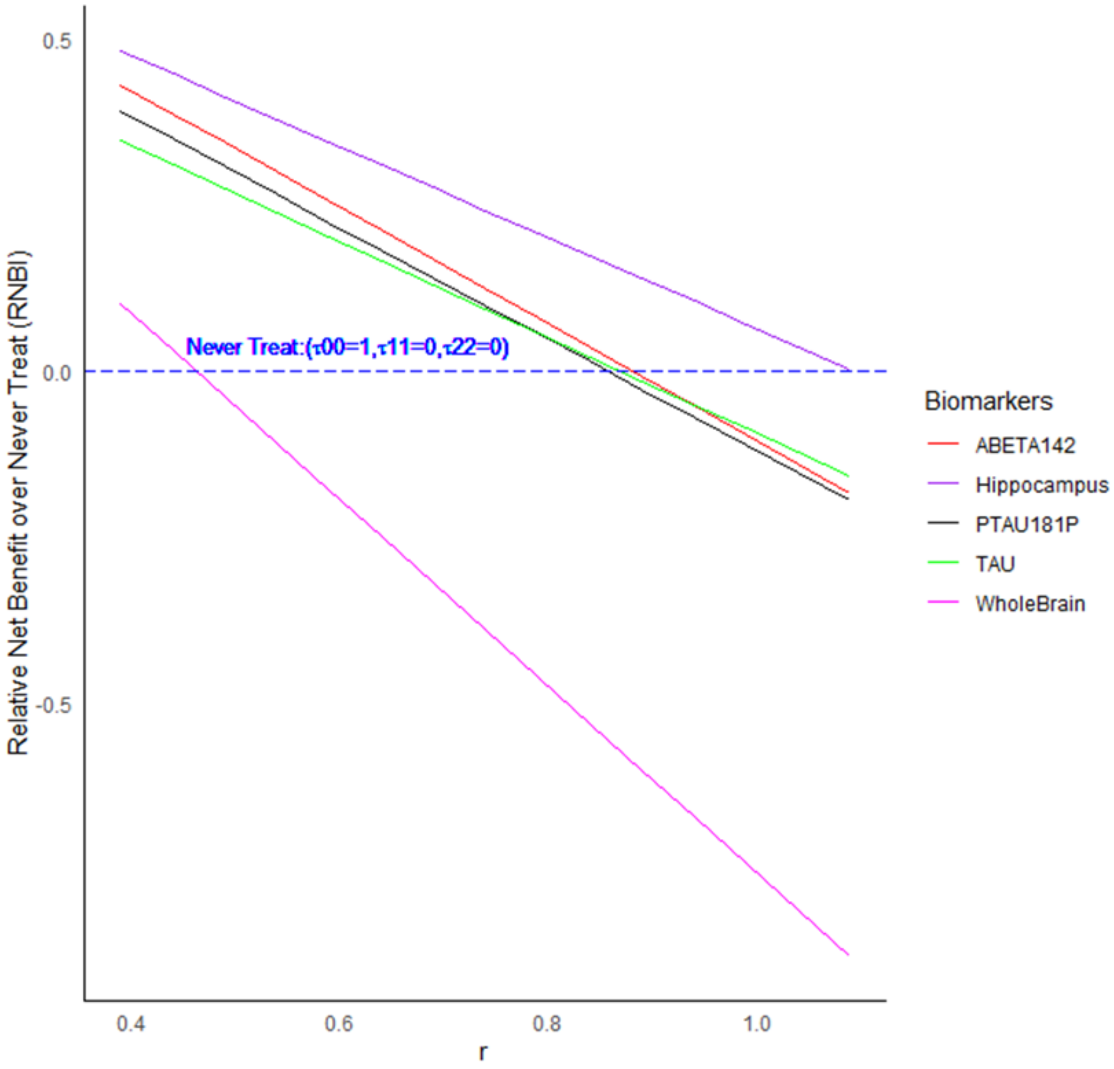


Figure 5.3. *Relative Net Benefit over Never-Treat Utility Ratio (r)*

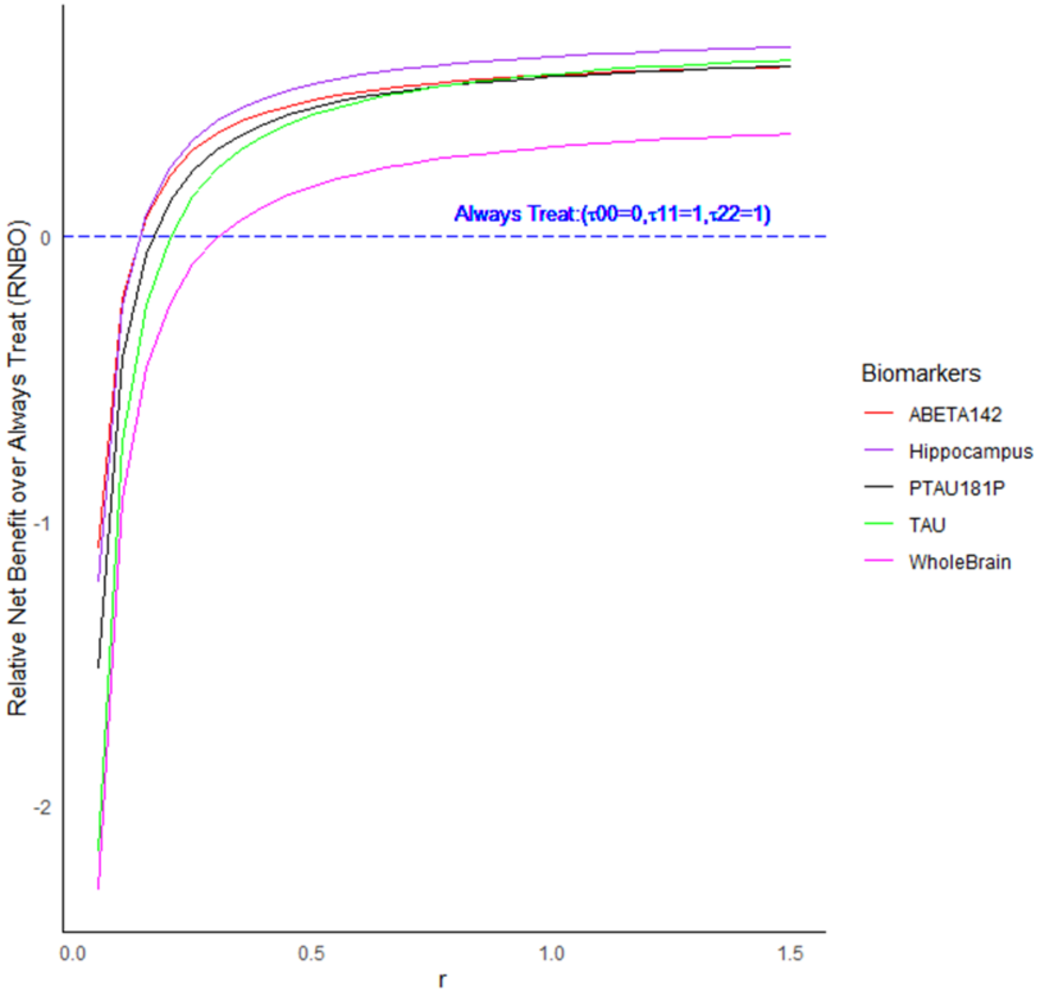


Figure 5.4. *Relative Net Benefit over-Always Treat Utility Ratio (r)*

Based on the proposed form of *r*, Hippocampus is the best biomarker to discriminate between subjects among the three stages of Alzheimer's disease. I illustrate these results with the Likelihood Ratio Graph (Figure 5.5), a helpful display proposed by Biggerstaff (2000). The graph has similar axes to the ROC plot. The coordinate of the biomarker's true and false-positive fractions is plotted in the graph. The biomarker is plotted with two lines on the ROC plot with PLR and NLR, respectively. The two lines define four regions in which the coordinate of the biomarker could lie.

Consequently, the biomarker coordinates of TPR and FPR fall in region A, indicating that the biomarker is better at detecting the absence of Alzheimer's disease than the Hippocampus but worse at detecting its presence because its PLR is worse (smaller). Still, its NLR is better (smaller). The evaluation of which biomarker is better based on test accuracy alone is equivocal.

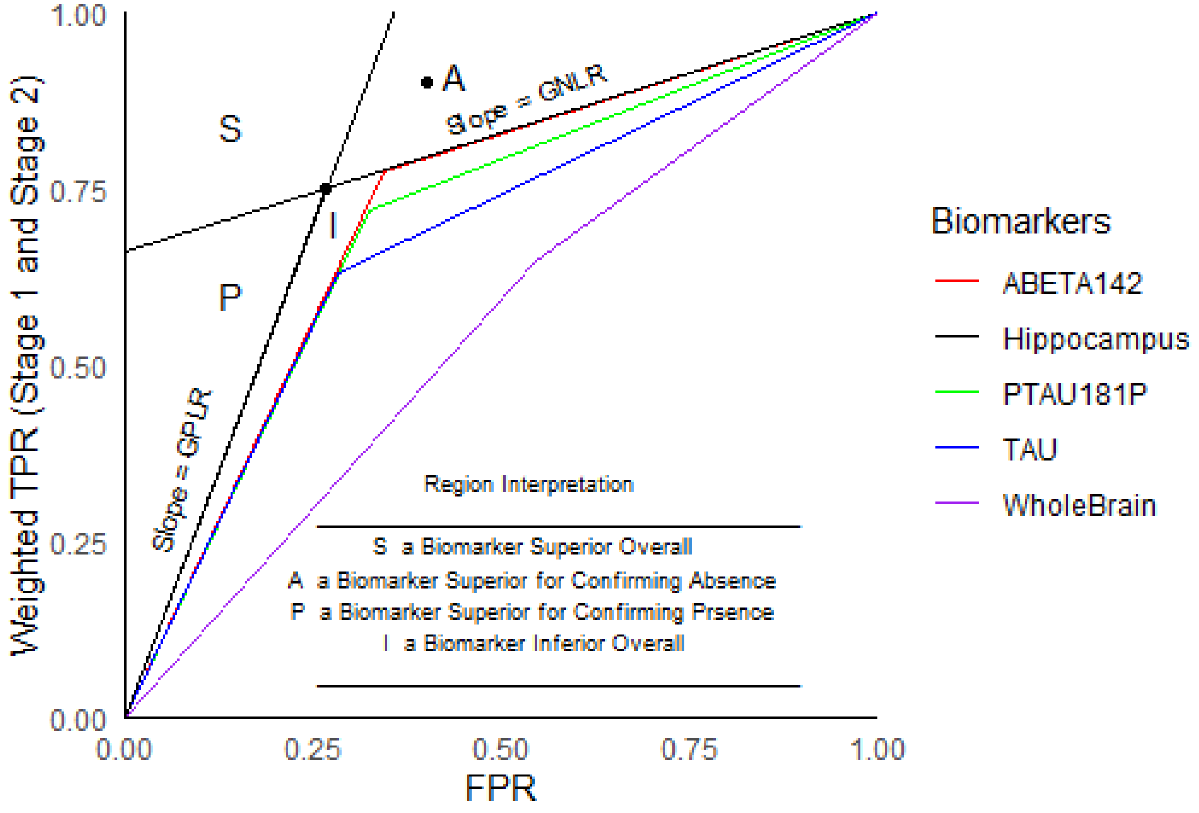
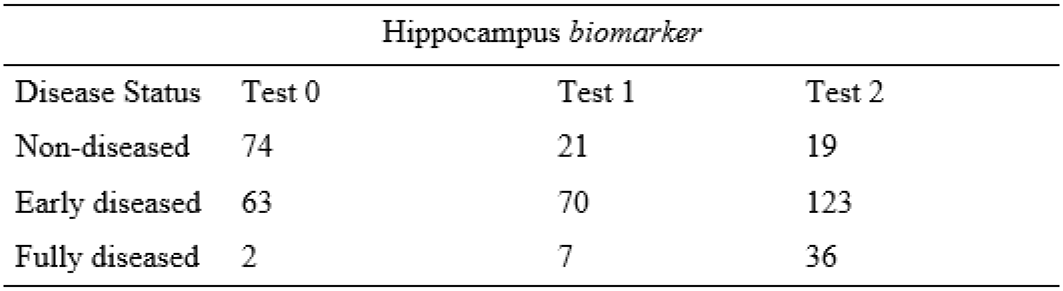


Figure 5.5. *Likelihood Ratio Graph: Regions of Comparison*

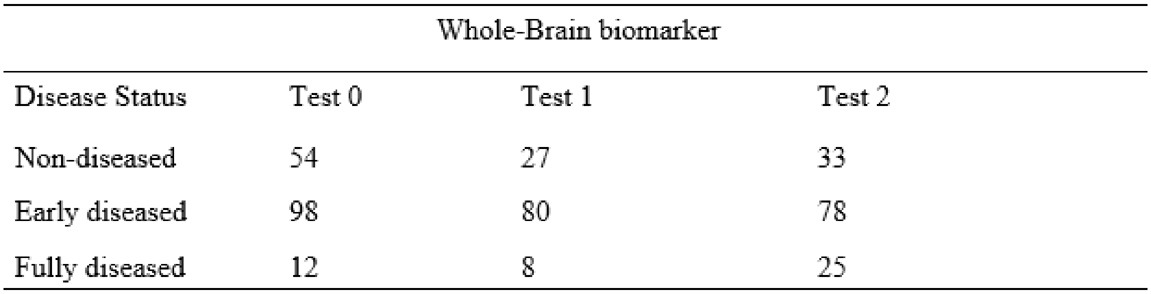
Comparing the diagnostic yield of biomarkers reveals several insights (Table 5.12- Table 5.16). The results represent positive test counts by disease status. In the table, clinical consequences are explored, assuming that a subject testing positive would be referred to an additional procedure that puts him or her at risk for adverse events. Also, I compare tests based on benefit-risk in a decision-theoretic framework. I assign losses to test misclassifications or utilities to correct classifications equivalently. The theory provides additional interpretations of quantities in the diagnostic yield table. In addition to the number of FP subjects harmed from unnecessary additional workup involving an invasive procedure, I can quantify the number of FN subjects harmed by lack of further workup. The harm associated with an FN result includes not receiving necessary treatment for a disease that may progress unattended. The disease is typically aggressive in some settings, and all FN subjects are harmed by lack of detection.

Table 5.12. *Diagnostic Yield of Hippocampus biomarker for Alzheimer's disease*



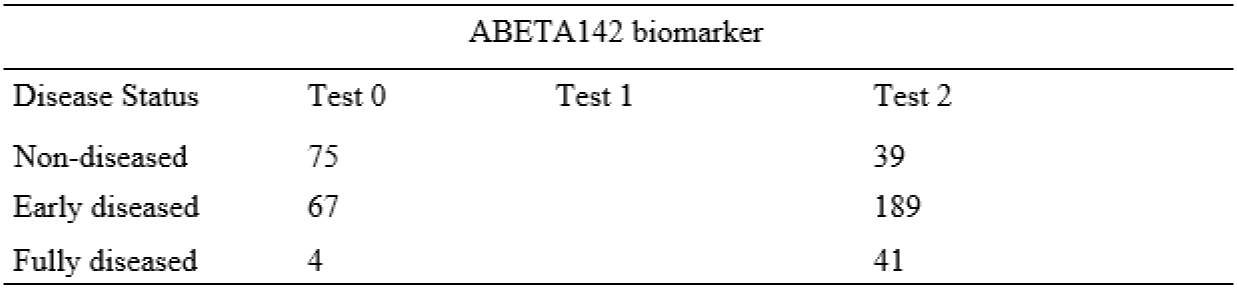
Hippocampus *biomarker*			
Disease Status	Test 0	Test 1	Test 2
Non-diseased	74	21	19
Early diseased	63	70	123
Fully diseased	2	7	36

Table 5.13. *Diagnostic Yield of Whole-Brain biomarker for Alzheimer's disease*



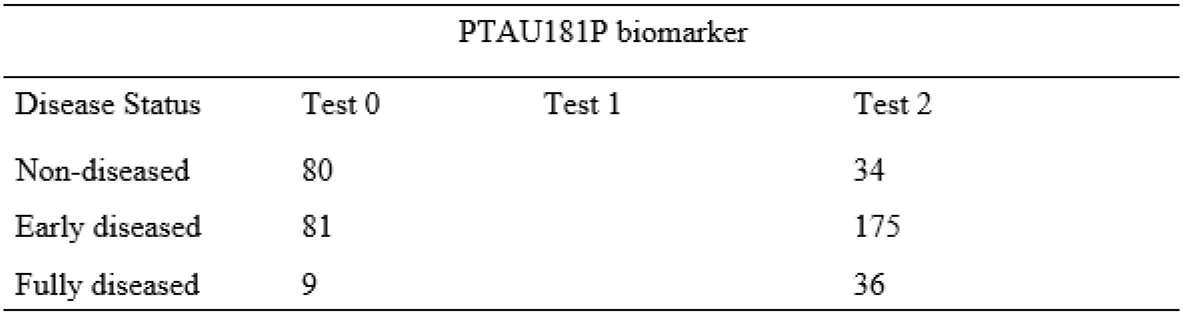
Whole-Brain biomarker			
Disease Status	Test 0	Test 1	Test 2
Non-diseased	54	27	33
Early diseased	98	80	78
Fully diseased	12	8	25

Table 5.14. *Diagnostic Yield of ABETA142 biomarker for Alzheimer's disease*



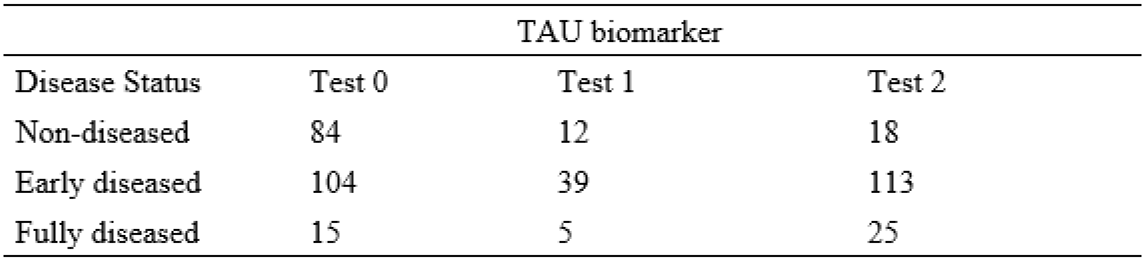
ABETA142 biomarker			
Disease Status	Test 0	Test 1	Test 2
Non-diseased	75		39
Early diseased	67		189
Fully diseased	4		41

Table 5.15. *Diagnostic Yield of PTAU181P biomarker for Alzheimer's disease*



PTAU181P biomarker			
Disease Status	Test 0	Test 1	Test 2
Non-diseased	80		34
Early diseased	81		175
Fully diseased	9		36

Table 5.16. *Diagnostic Yield of TAU biomarker for Alzheimer's disease*



TAU biomarker			
Disease Status	Test 0	Test 1	Test 2
Non-diseased	84	12	18
Early diseased	104	39	113
Fully diseased	15	5	25

5.3. Discussion

Based on the optimal statistics in Table 5.2, Hippocampus has the highest statistics compared to other biomarkers. 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, and the correct classification rate at the early disease stage is zero for both biomarkers. The results imply that Abeta and PTAU are not suitable biomarkers for detecting between subjects among the three stages in this study. However, the optimal statistics of these two biomarkers are close to the optimal statistics of TAU and slightly higher than the optimal statistics of WholeBrain. The optimal statistic of TAU lies between the values of Abeta and PTAU, and it is somewhat higher than the optimal statistics for the whole-Brain. Hippocampus has much higher optimal statistics than other biomarkers. Thus, Hippocampus is the best biomarker to use to discriminate between subjects among the three stages of Alzheimer's disease. It is important to properly diagnose subjects in the early stage of Alzheimer's disease since it is an irreversible condition that progresses over time, and brain changes caused by Alzheimer's disease may begin 20 years or more before any signs and symptoms appear (Gaugler, James, Johnson, Marin, & Weuve, 2019).

Early diagnosis, intervention, prevention, and treatment by application of new diagnostic methods at the earliest possible stage of AD are essential in slowing down the progression of the disease. Gaugler et al. (2019) mention that seniors believe the early diagnosis is important because of early intervention for the disease, allowing them time to understand what is happening with the disease and all concerned. This helps them to adjust and offers access to advice, financial support, and non-pharmacological and pharmacological treatments, allowing the family to plan for the future. Compared to existing measures, our proposed measure has high correct classification rates of stage 1 and stage 2, the specificity of stages 1 and 2, along with the sensitivity of stages 1 and 2, respectively, for the Hippocampus. The specificity of stages 1 and 2 emphasizes rule-in information and provides essential information for the early diagnosis of Alzheimer's disease. The sensitivity for stages 1 and 2 highlights rule-out information and provides crucial information to avoid unnecessary additional workup of Alzheimer's disease early on.

In conclusion, the net benefit approach is the most effective approach to evaluating biomarkers of Alzheimer's disease 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. In summary, among the five biomarkers, Hippocampus has presented the best performance results in the scenario of a three-stage setting, while Abeta has presented an acceptable performance in the two-stage setting.

CHAPTER 6

FINAL REMARKS, CONCLUSIONS, LIMITATIONS, AND FUTURE RESEARCH

6.1 Final Remarks and Conclusions

Comparing tests based on benefit-risk plays a crucial role in convincing clinicians because it involves the accuracy of the test and the clinical consequences of diagnostic errors. Evaluating diagnostic tests is essential in placing patients on appropriate treatment plans; thus, measuring the benefit-risk of a test has significant clinical implications. In my study, benefit-risk approaches are used for binary tests, where benefit and harm are put on the same scale to determine whether a diagnostic test has better, worse, or the same outcomes when assessing a test's clinical consequences. No studies have investigated the accuracy of measures and the clinical consequences of the medical diagnostic test errors in multi-stage disease settings. However, the clinical implications of treating or not treating patients at a different stage of the disease have different benefit-risk consequences. In practice, it is vital to detect the early stage of illness for timely medical interventions to reduce the cost of the treatment and improve the quality of life for patients. Diagnostic tests that can identify multiple stages are precious, desirable, and in need. As I have studied, I have realized that the benefit-risk approach for multi-stage diseases requires more research than two-stage diseases.

Motivated by Pennello's approach, this dissertation proposes a descriptive, diagnostic yield table for multi-stage clinical conditions in practice. I extend the net benefit approach of evaluating diagnostic tests to multi-stage clinical conditions. Consequently, I extend the diagnostic yield table to multi-stage clinical conditions. I develop a decision theory based on net benefit for evaluating diagnostic tests. It provides additional interpretation for rule-in or rule-out clinical conditions and their adverse consequences from unnecessary work-up in multi-stage diseases. Numerical examples are conducted for the three-stage disease to illustrate the proposed measures, assuming different prevalence settings with varying parameters for the diseased underlying distributions. Numerical examples show that the higher correct classification rates test has a better relative net benefit. Consequently, it will reduce adverse events by treating the non-diseased and helping to avoid not treating the right stage of the disease. My results indicate that using our proposed research methods will yield the advantage of most effectively deciding which biomarker should be used based on the diagnostic purpose to identify rule-in or rule-out patients.

This study also provides an example of applying the proposed measure for multi-stage diseases, using a dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Results concur with the numerical study and highlight the strengths of the proposed measure that benefits and risks could be different from stage to stage. This study provides clues for the early diagnosis of Alzheimer's disease. For example, the net benefit approach of evaluating diagnostic tests can detect subjects in the first and the second stages with the highest correct classification rates in these two stages of the biomarkers. This study provides exciting exploratory outcomes in improving both diagnostic accuracy tests and the clinical consequences of diagnostic error for multi-stage diseases, especially for those who want to discover biomarkers for early diagnosis.

6.2 Limitations and Future Work

In this study, the net benefit approach has shown some advantages in detecting subjects in the last stage of a multi-stage disease and has been beneficial for diagnosis; however, the correct classification rate has been found to be slightly lower in the last stage of the disease. Additionally, when comparing the benefit-risk among all disease stages, I have found that there can be no existing standardized methods to compare the performance of the measures since they have different properties. For two-stage diseases, high correct classification rates in both stages are desired. For multi-stage disease cases, the demand for a high correct classification rate of a specific stage relies on clinical needs, clinical consequences, and treatment cost. For example, when a clinical test or biomarkers are considered to identify subjects in the early stages, the correct classification rates are expected to be as high as needed. In contrast, when a clinical test or biomarkers are considered to identify patients in late stages, the correct classification rate of the last stage would be more critical than the others. In reality, it is hard to achieve a balanced diagnostic test or biomarker with high correct classification rates among all stages. Hence, a method that can evaluate the performance of diagnostic test accuracy and clinical consequences of diagnostic errors, besides the benefit-risk, is desired in future studies.

Mainly the focus of this study is on the ordinal stages of diseases. But many cases of disease deal with multiple nominal classes, including genomic studies. Additional research is needed for nominal cases, including a general outline of specific benefit-risk comparisons of tests that permit multiple nominal test results.

Furthermore, the findings of this study's estimations are restricted to using the kernel approach. The simulation of estimation using other methods is highly encouraged to compare the performance of diagnostic test accuracy and clinical consequences of the diagnostic errors of multi-stage disease. Additional research is needed into the properties and strengths of different measures under various distributions.

Diagnosis of multi-stage disease at an early stage provides enough time for health care practitioners to make a plan, fight severe diseases, and minimize the cost, specifically for conditions without a cure. However, I have found a lack of practical applications of net benefit analysis of diagnostic tests for multi-stage diseases to research data. Therefore, there is a top priority to develop reliable and reasonable measures to compare diagnostic test accuracy and clinical consequences of diagnostic errors, which will further improve diagnosis and assist in designing clinical treatments and guidelines.

Lastly, a single biomarker is less than satisfactory for confirmation of the clinical diagnosis of a disease. For example, genetic and epigenetic biomarkers are not enough to identify the subtypes of cancer or confirm the staging of cancer patients and treatment evaluation. As a result, the generalization of single biomarker measures to multiple biomarkers is encouraged for future study of the diagnosis of multi-stage diseases.

Although the net benefit approach is relatively novel, the net benefit of comparing diagnostic tests has recently gained increasing attention. Unlike traditional measures such as sensitivity, specificity, or area under the curve, studying net benefit provides information about the clinical judgment of the relative value of benefit and harm associated with diagnostic errors (Vickers, Van Calster, & Steyerberg, 2016). This approach helps doctors make better decisions and make wider use of net benefits. It also quantifies the good and harm of a clinical decision based on a biomarker, diagnostic test, or statistical model and better matches the clinical aim of much medical research. As a result, the broader use of the net benefit approach should be an exciting focus in the future study of the diagnosis of 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.

Abe, O., Abe, R., Enomoto, K., Kikuchi, K., Koyama, H., Masuda, H., . . . Caffier, H. (2005). EBCTCGEarly Breast Cancer Trialist's Collaborative Group (EBCTCG)@ Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet 365: 1687-1717. *The Lancet, 365*, 1687-1717.

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 Association. (2020). 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.

Alzheimer’s Disease Neuroimaging Initiative (ADNI). (2020). 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)*.

Association, A. s. (2018). 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia, 14*(3), 367-429.

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

Baker, S. G., Cook, N. R., Vickers, A., & Kramer, B. S. J. J. o. t. R. S. S. S. A. (2009). Using relative utility curves to evaluate risk prediction. *172*(4), 729-748.

Baker, S. G., & Kramer, B. S. J. D. m. (2012). Evaluating a new marker for risk prediction: decision analysis to the rescue. *14*(76), 181-188.

Baker, S. G., Van Calster, B., & Steyerberg, E. W. J. T. i. j. o. b. (2012). Evaluating a new marker for risk prediction using the test tradeoff: an update. *8*(1), 1-37.

Baker, S. G. J. J. J. o. t. N. C. I. (2009). Putting risk prediction in perspective: relative utility curves. *101*(22), 1538-1542.

Beason-Held, L. L., Goh, J. O., An, Y., Kraut, M. A., O'Brien, R. J., & Ferrucci, L. (2013). Changes in Brain Function Occur Years before the Onset of Cognitive Impairment. *The Journal of Neuroscience, 33*(46), 18008 –18014.

Biggerstaff, B. J. J. S. i. m. (2000). Comparing diagnostic tests: a simple graphic using likelihood ratios. *19*(5), 649-663.

Bjerke, M., & Engelborghs, S. (2018). Cerebrospinal fluid biomarkers for early and differential Alzheimer's disease diagnosis. *Journal of Alzheimer's Disease, 62*(3), 1199-1209.

Blennow, K., Hampel, H., Weiner, M., & Zetterberg, H. (2010). Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. *Nat Rev Neurol, 6*(3), 131-144. doi:10.1038/nrneurol.2010.4

Bossuyt, P. M., Lijmer, J. G., & Mol, B. W. (2000). Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet, 356*(9244), 1844-1847. doi:10.1016/s0140-6736(00)03246-3

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.

Chudecka-Głaz, A. M. (2015). ROMA, an algorithm for ovarian cancer. *Clin Chim Acta, 440*, 143-151. doi:10.1016/j.cca.2014.11.015

Cramer, D. W., Bast, R. C., Jr., Berg, C. D., Diamandis, E. P., Godwin, A. K., Hartge, P., . . . Urban, N. (2011). Ovarian cancer biomarker performance in prostate, lung, colorectal, and ovarian cancer screening trial specimens. *Cancer Prev Res (Phila), 4*(3), 365-374. doi:10.1158/1940-6207.Capr-10-0195

DAFFNEr, K. R., & Scinto, L. F. (2000). Early diagnosis of Alzheimer's disease. In *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.

Duthey, B. (2013). Background paper 6.11: Alzheimer disease and other dementias. *A public health approach to innovation, 6*, 1-74.

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

Evans, S. R., Pennello, G., Pantoja-Galicia, N., Jiang, H., Hujer, A. M., Hujer, K. M., . . . for the Antibacterial Resistance Leadership, G. (2016). Benefit-risk Evaluation for Diagnostics: A Framework (BED-FRAME). *Clinical Infectious Diseases, 63*(6), 812-817. doi:10.1093/cid/ciw329

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.

Gail, M. H., & Pfeiffer, R. M. J. B. (2005). On criteria for evaluating models of absolute risk. *6*(2), 227-239.

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. doi:10.1186/1750-1326-4-19

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.

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.

Hoering, A., Leblanc, M., & Crowley, J. J. (2008). Randomized phase III clinical trial designs for targeted agents. *Clin Cancer Res, 14*(14), 4358-4367. doi:10.1158/1078-0432.Ccr-08-0288

Holmboe, E. S., & Durning, S. J. J. D. (2014). Assessing clinical reasoning: moving from in vitro to in vivo. *1*(1), 111-117.

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.

Johns Hopkins Medicine. (2019a). Alzheimer's Disease: Stages of Alzheimer's. Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/alzheimers-disease/stages-of-alzheimer-disease.

Johns Hopkins Medicine. (2019b). Stages of Alzheimer's Disease. Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/alzheimers-disease/stages-of-alzheimer-disease.

Jutel, A. J. S. o. h., & illness. (2009). Sociology of diagnosis: a preliminary review. *31*(2), 278-299.

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.

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.

MAYO Clinic. (2020). Alzheimer's stages: How the disease progresses. Retrieved from https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers-stages/art-20048448.

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

Pauker, S. G., & Kassirer, J. P. J. N. E. J. o. M. (1975). Therapeutic decision making: a cost-benefit analysis. *293*(5), 229-234.

Pennello, G., Pantoja-Galicia, N., & Evans, S. (2016a). Comparing diagnostic tests on benefit-risk. *Journal of biopharmaceutical statistics, 26*(6), 1083-1097. doi:10.1080/10543406.2016.1226335

Pennello, G., Pantoja-Galicia, N., & Evans, S. J. J. o. b. s. (2016b). Comparing diagnostic tests on benefit-risk. *26*(6), 1083-1097.

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

Pepe, M. S., Janes, H., Li, C. I., Bossuyt, P. M., Feng, Z., & Hilden, J. J. C. c. (2016). Early-phase studies of biomarkers: what target sensitivity and specificity values might confer clinical utility? *, 62*(5), 737-742.

Rapsomaniki, E., White, I. R., Wood, A. M., Thompson, S. G., & Emerging Risk Factors, C. (2012). A framework for quantifying net benefits of alternative prognostic models. *Statistics in medicine, 31*(2), 114-130. doi:10.1002/sim.4362

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.

Rosenberg, C. E. J. T. M. Q. (2002). The tyranny of diagnosis: specific entities and individual experience. *80*(2), 237-260.

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., Chen, D.-G., Ahmed, F., & Kersey, J. (2021). Medical diagnostics accuracy measures and cut-point selection: An innovative approach based on relative net benefit. *Communications in Statistics - Theory and Methods,* 1-16. doi:10.1080/03610926.2021.2001016

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.

Shaw, L. M., & Trojanowski, J. Q. (2017). Methods and Tools. *Alzheimer's Disease Neuroimaging Initiative (ADNI).* Retrieved from https://www.alz.org/media/Documents/ww-adni-may-2017-biomarker-core-shaw.pdf.

Simon, R. (2010). Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Per Med, 7*(1), 33-47. doi:10.2217/pme.09.49

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

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

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

Sperlinga, R. A., Aisenb, P. S., Beckettc, L. A., Bennettd, D. A., Crafte, S., Faganf, A. M., . . . Kayei, J. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelinesfor Alzheimer's disease. *Alzheimers Dement., 7*(3), 280–292.

Study design. (2020). *Alzheimer's Disease Neuroimaging Initiative*. Retrieved from http://adni.loni.usc.edu/study-design/#background-container.

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

Tsalik, E. L., Li, Y., Hudson, L. L., Chu, V. H., Himmel, T., Limkakeng, A. T., . . . Welty-Wolf, K. E. J. A. o. t. A. T. S. (2016). Potential cost-effectiveness of early identification of hospital-acquired infection in critically ill patients. *13*(3), 401-413.

Vickers, A. J., & Elkin, E. B. J. M. D. M. (2006). Decision curve analysis: a novel method for evaluating prediction models. *26*(6), 565-574.

Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *bmj, 352*.

Vijayakumar, A., & Vijayakumar, A. (2013). Comparison of hippocampal volume in dementia subtypes. *International Scholarly Research Notices, 2013*.

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

World Health Organization [WHO]. (2012). International classification of diseases (ICD). Geneva: World Health Organization.

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

Youden, W. J. J. C. (1950). Index for rating diagnostic tests. *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.

2022 Alzheimer's disease facts and figures. (2022). *Alzheimers Dement, 18*(4), 700-789. doi:10.1002/alz.12638

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

Abe, O., Abe, R., Enomoto, K., Kikuchi, K., Koyama, H., Masuda, H., . . . Caffier, H. (2005). EBCTCGEarly Breast Cancer Trialist's Collaborative Group (EBCTCG)@ Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet 365: 1687-1717. *The Lancet, 365*, 1687-1717.

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 Association. (2020). 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.

Alzheimer’s Disease Neuroimaging Initiative (ADNI). (2020). 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)*.

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

Baker, S. G., Cook, N. R., Vickers, A., & Kramer, B. S. J. J. o. t. R. S. S. S. A. (2009). Using relative utility curves to evaluate risk prediction. *172*(4), 729-748.

Baker, S. G., & Kramer, B. S. J. D. m. (2012). Evaluating a new marker for risk prediction: decision analysis to the rescue. *14*(76), 181-188.

Baker, S. G., Van Calster, B., & Steyerberg, E. W. J. T. i. j. o. b. (2012). Evaluating a new marker for risk prediction using the test tradeoff: an update. *8*(1), 1-37.

Baker, S. G. J. J. J. o. t. N. C. I. (2009). Putting risk prediction in perspective: relative utility curves. *101*(22), 1538-1542.

Biggerstaff, B. J. J. S. i. m. (2000). Comparing diagnostic tests: a simple graphic using likelihood ratios. *19*(5), 649-663.

Blennow, K., Hampel, H., Weiner, M., & Zetterberg, H. (2010). Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. *Nat Rev Neurol, 6*(3), 131-144. doi:10.1038/nrneurol.2010.4

Bossuyt, P. M., Lijmer, J. G., & Mol, B. W. (2000). Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet, 356*(9244), 1844-1847. doi:10.1016/s0140-6736(00)03246-3

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.

Chudecka-Głaz, A. M. (2015). ROMA, an algorithm for ovarian cancer. *Clin Chim Acta, 440*, 143-151. doi:10.1016/j.cca.2014.11.015

Cramer, D. W., Bast, R. C., Jr., Berg, C. D., Diamandis, E. P., Godwin, A. K., Hartge, P., . . . Urban, N. (2011). Ovarian cancer biomarker performance in prostate, lung, colorectal, and ovarian cancer screening trial specimens. *Cancer Prev Res (Phila), 4*(3), 365-374. doi:10.1158/1940-6207.Capr-10-0195

DAFFNEr, K. R., & Scinto, L. F. (2000). Early diagnosis of Alzheimer’s disease. In *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.

Duthey, B. (2013). Background paper 6.11: Alzheimer disease and other dementias. *A public health approach to innovation, 6*, 1-74.

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

Evans, S. R., Pennello, G., Pantoja-Galicia, N., Jiang, H., Hujer, A. M., Hujer, K. M., . . . for the Antibacterial Resistance Leadership, G. (2016). Benefit-risk Evaluation for Diagnostics: A Framework (BED-FRAME). *Clinical Infectious Diseases, 63*(6), 812-817. doi:10.1093/cid/ciw329

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.

Gail, M. H., & Pfeiffer, R. M. J. B. (2005). On criteria for evaluating models of absolute risk. *6*(2), 227-239.

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. doi:10.1186/1750-1326-4-19

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.

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.

Hoering, A., Leblanc, M., & Crowley, J. J. (2008). Randomized phase III clinical trial designs for targeted agents. *Clin Cancer Res, 14*(14), 4358-4367. doi:10.1158/1078-0432.Ccr-08-0288

Holmboe, E. S., & Durning, S. J. J. D. (2014). Assessing clinical reasoning: moving from in vitro to in vivo. *1*(1), 111-117.

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.

Johns Hopkins Medicine. (2019a). Alzheimer's Disease: Stages of Alzheimer's. Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/alzheimers-disease/stages-of-alzheimer-disease.

Johns Hopkins Medicine. (2019b). Stages of Alzheimer's Disease. Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/alzheimers-disease/stages-of-alzheimer-disease.

Jutel, A. J. S. o. h., & illness. (2009). Sociology of diagnosis: a preliminary review. *31*(2), 278-299.

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.

2022 Alzheimer's disease facts and figures. (2022). *Alzheimers Dement, 18*(4), 700-789. doi:10.1002/alz.12638

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

Abe, O., Abe, R., Enomoto, K., Kikuchi, K., Koyama, H., Masuda, H., . . . Caffier, H. (2005). EBCTCGEarly Breast Cancer Trialist's Collaborative Group (EBCTCG)@ Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet 365: 1687-1717. *The Lancet, 365*, 1687-1717.

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 Association. (2020). 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.

Alzheimer’s Disease Neuroimaging Initiative (ADNI). (2020). 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)*.

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

Baker, S. G., Cook, N. R., Vickers, A., & Kramer, B. S. J. J. o. t. R. S. S. S. A. (2009). Using relative utility curves to evaluate risk prediction. *172*(4), 729-748.

Baker, S. G., & Kramer, B. S. J. D. m. (2012). Evaluating a new marker for risk prediction: decision analysis to the rescue. *14*(76), 181-188.

Baker, S. G., Van Calster, B., & Steyerberg, E. W. J. T. i. j. o. b. (2012). Evaluating a new marker for risk prediction using the test tradeoff: an update. *8*(1), 1-37.

Baker, S. G. J. J. J. o. t. N. C. I. (2009). Putting risk prediction in perspective: relative utility curves. *101*(22), 1538-1542.

Biggerstaff, B. J. J. S. i. m. (2000). Comparing diagnostic tests: a simple graphic using likelihood ratios. *19*(5), 649-663.

Blennow, K., Hampel, H., Weiner, M., & Zetterberg, H. (2010). Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. *Nat Rev Neurol, 6*(3), 131-144. doi:10.1038/nrneurol.2010.4

Bossuyt, P. M., Lijmer, J. G., & Mol, B. W. (2000). Randomised comparisons of medical tests: sometimes invalid, not always efficient. *Lancet, 356*(9244), 1844-1847. doi:10.1016/s0140-6736(00)03246-3

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.

Chudecka-Głaz, A. M. (2015). ROMA, an algorithm for ovarian cancer. *Clin Chim Acta, 440*, 143-151. doi:10.1016/j.cca.2014.11.015

Cramer, D. W., Bast, R. C., Jr., Berg, C. D., Diamandis, E. P., Godwin, A. K., Hartge, P., . . . Urban, N. (2011). Ovarian cancer biomarker performance in prostate, lung, colorectal, and ovarian cancer screening trial specimens. *Cancer Prev Res (Phila), 4*(3), 365-374. doi:10.1158/1940-6207.Capr-10-0195

DAFFNEr, K. R., & Scinto, L. F. (2000). Early diagnosis of Alzheimer’s disease. In *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.

Duthey, B. (2013). Background paper 6.11: Alzheimer disease and other dementias. *A public health approach to innovation, 6*, 1-74.

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

Evans, S. R., Pennello, G., Pantoja-Galicia, N., Jiang, H., Hujer, A. M., Hujer, K. M., . . . for the Antibacterial Resistance Leadership, G. (2016). Benefit-risk Evaluation for Diagnostics: A Framework (BED-FRAME). *Clinical Infectious Diseases, 63*(6), 812-817. doi:10.1093/cid/ciw329

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.

Gail, M. H., & Pfeiffer, R. M. J. B. (2005). On criteria for evaluating models of absolute risk. *6*(2), 227-239.

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. doi:10.1186/1750-1326-4-19

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.

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.

Hoering, A., Leblanc, M., & Crowley, J. J. (2008). Randomized phase III clinical trial designs for targeted agents. *Clin Cancer Res, 14*(14), 4358-4367. doi:10.1158/1078-0432.Ccr-08-0288

Holmboe, E. S., & Durning, S. J. J. D. (2014). Assessing clinical reasoning: moving from in vitro to in vivo. *1*(1), 111-117.

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.

Johns Hopkins Medicine. (2019a). Alzheimer's Disease: Stages of Alzheimer's. Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/alzheimers-disease/stages-of-alzheimer-disease.

Johns Hopkins Medicine. (2019b). Stages of Alzheimer's Disease. Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/alzheimers-disease/stages-of-alzheimer-disease.

Jutel, A. J. S. o. h., & illness. (2009). Sociology of diagnosis: a preliminary review. *31*(2), 278-299.

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.

Kersey, J., Samawi, H., Yin, J., Rochani, H., & Zhang, X. (2022). On diagnostic accuracy measure with cut-points criterion for ordinal disease classification based on concordance and discordance. *Journal of Applied Statistics*, 1-18. doi:10.1080/02664763.2022.2041567

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.

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.

MAYO Clinic. (2020). Alzheimer's stages: How the disease progresses. Retrieved from https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers-stages/art-20048448.

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.

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

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

Pauker, S. G., & Kassirer, J. P. J. N. E. J. o. M. (1975). Therapeutic decision making: a cost-benefit analysis. *293*(5), 229-234.

Pennello, G. (2019). *Net Benefit of a Diagnostic Test to Rule-In or Rule-Out a Clinical Condition*. Paper presented at the ENAR.

Pennello, G., Pantoja-Galicia, N., & Evans, S. (2016a). Comparing diagnostic tests on benefit-risk. *Journal of biopharmaceutical statistics, 26*(6), 1083-1097. doi:10.1080/10543406.2016.1226335

Pennello, G., Pantoja-Galicia, N., & Evans, S. J. J. o. b. s. (2016b). Comparing diagnostic tests on benefit-risk. *26*(6), 1083-1097.

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

Pepe, M. S., Janes, H., Li, C. I., Bossuyt, P. M., Feng, Z., & Hilden, J. J. C. c. (2016). Early-phase studies of biomarkers: what target sensitivity and specificity values might confer clinical utility? *, 62*(5), 737-742.

Rapsomaniki, E., White, I. R., Wood, A. M., Thompson, S. G., & Emerging Risk Factors, C. (2012). A framework for quantifying net benefits of alternative prognostic models. *Statistics in medicine, 31*(2), 114-130. doi:10.1002/sim.4362

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.

Rosenberg, C. E. J. T. M. Q. (2002). The tyranny of diagnosis: specific entities and individual experience. *80*(2), 237-260.

Samawi, H., Chen, D.-G., Ahmed, F., & Kersey, J. (2021). Medical diagnostics accuracy measures and cut-point selection: An innovative approach based on relative net benefit. *Communications in Statistics - Theory and Methods*, 1-16. doi:10.1080/03610926.2021.2001016

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.

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.

Simon, R. (2010). Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Per Med, 7*(1), 33-47. doi:10.2217/pme.09.49

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

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

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

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

Tsalik, E. L., Li, Y., Hudson, L. L., Chu, V. H., Himmel, T., Limkakeng, A. T., . . . Welty-Wolf, K. E. J. A. o. t. A. T. S. (2016). Potential cost-effectiveness of early identification of hospital-acquired infection in critically ill patients. *13*(3), 401-413.

Vickers, A. J., & Elkin, E. B. J. M. D. M. (2006). Decision curve analysis: a novel method for evaluating prediction models. *26*(6), 565-574.

Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *bmj, 352*.

Vijayakumar, A., & Vijayakumar, A. (2013). Comparison of hippocampal volume in dementia subtypes. *International Scholarly Research Notices, 2013*.

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

World Health Organization [WHO]. (2012). International classification of diseases (ICD). Geneva: World Health Organization.

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

Youden, W. J. J. C. (1950). Index for rating diagnostic tests. *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.

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.

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.

MAYO Clinic. (2020). Alzheimer's stages: How the disease progresses. Retrieved from https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers-stages/art-20048448.

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.

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

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

Pauker, S. G., & Kassirer, J. P. J. N. E. J. o. M. (1975). Therapeutic decision making: a cost-benefit analysis. *293*(5), 229-234.

Pennello, G., Pantoja-Galicia, N., & Evans, S. (2016a). Comparing diagnostic tests on benefit-risk. *Journal of biopharmaceutical statistics, 26*(6), 1083-1097. doi:10.1080/10543406.2016.1226335

Pennello, G., Pantoja-Galicia, N., & Evans, S. J. J. o. b. s. (2016b). Comparing diagnostic tests on benefit-risk. *26*(6), 1083-1097.

Pennello, G. (2019). *Net Benefit of a Diagnostic Test to Rule-In or Rule-Out a Clinical Condition*. Paper presented at the ENAR.

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

Pepe, M. S., Janes, H., Li, C. I., Bossuyt, P. M., Feng, Z., & Hilden, J. J. C. c. (2016). Early-phase studies of biomarkers: what target sensitivity and specificity values might confer clinical utility? , *62*(5), 737-742.

Rapsomaniki, E., White, I. R., Wood, A. M., Thompson, S. G., & Emerging Risk Factors, C. (2012). A framework for quantifying net benefits of alternative prognostic models. *Statistics in medicine, 31*(2), 114-130. doi:10.1002/sim.4362

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.

Rosenberg, C. E. J. T. M. Q. (2002). The tyranny of diagnosis: specific entities and individual experience. *80*(2), 237-260.

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.

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.

Simon, R. (2010). Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Per Med, 7*(1), 33-47. doi:10.2217/pme.09.49

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

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

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

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

Tsalik, E. L., Li, Y., Hudson, L. L., Chu, V. H., Himmel, T., Limkakeng, A. T., . . . Welty-Wolf, K. E. J. A. o. t. A. T. S. (2016). Potential cost-effectiveness of early identification of hospital-acquired infection in critically ill patients. *13*(3), 401-413.

Vickers, A. J., & Elkin, E. B. J. M. D. M. (2006). Decision curve analysis: a novel method for evaluating prediction models. *26*(6), 565-574.

Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *bmj, 352*.

Vijayakumar, A., & Vijayakumar, A. (2013). Comparison of hippocampal volume in dementia subtypes. *International Scholarly Research Notices, 2013*.

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

World Health Organization [WHO]. (2012). International classification of diseases (ICD). Geneva: World Health Organization.

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

Youden, W. J. J. C. (1950). Index for rating diagnostic tests. *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.