DEVELOPMENT AND COMPARISON OF DIFFERENT METHODS OF EVALUATING FREE-RESPONSE ROC SYSTEMS

by

Tao Song

B.S., University of Science and Technology of China, 2003

M.S., University of Toledo, 2005

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Tao Song

It was defended on

November 4, 2008

and approved by:

Andriy Bandos, PhD, Research Assistant Professor
Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

David Gur, PhD, Professor
Department of Radiology, School of Medicine
University of Pittsburgh

Sati Mazumdar, PhD, Professor
Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Dissertation Advisor: Howard E. Rockette, PhD, Professor and Chair
Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

DEVELOPMENT AND COMPARISON OF DIFFERENT METHODS OF EVALUATING FREE-RESPONSE ROC SYSTEMS

Tao Song, PhD

University of Pittsburgh, 2008

Receiver Operating Characteristic (ROC) analysis has been widely used to evaluate diagnostic systems since the 1970s. In diagnostic imaging the decision task often needs the radiologist to locate the specific region on a subject that actually contains the abnormality. A Free-Response ROC experiment has been more and more accepted for evaluating this type of a diagnostic task. It entails detecting and marking the locations of all suspected abnormalities, as well as indicating a level of suspicion regarding the specific abnormality at each marked location. Several existing approaches of analyzing FROC data used the maximum rating to represent the multiple responses of a subject and then applied an analysis in an ROC concept to summarize the diagnostic system's discriminative ability in a randomly selected pair of actually negative and actually positive subjects. This dissertation proposes and evaluates new methods of subject-based discriminative ability by considering approaches based on the average of multiple ratings and approaches based on the stochastic order. Indices are also formulated by improving the modified *JAFROC* indices, in order to summarize the diagnostic performance with correct location information. We also propose new indices that can penalize and reward for the number of correct and incorrect marks on the subjects by modifying the Wilcoxon statistic. Asymptotic procedures are developed to compare the discriminative ability between two FROC systems. These asymptotic approaches are then extended to the multi-reader setting, taking into consideration the correlation and heterogeneity between readers. We also apply three different approaches to fit a smooth FROC curve, namely Box-Cox transformation approach, kernel smoothing approach and kernel regression approach. The public health significance of the work lies in our efforts to improve the statistical tools for evaluating medical diagnostic devices, which can help in the development of more specific and affordable diagnostic methods. Our contribution to early diagnosis could improve the timely recognition of reportable diseases.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

I would like to express my first and most earnest acknowledge to my advisor, Dr Howard E. Rockette, whose kindness, enthusiasm and insightful statistical thinking made all the difference in my academic career. His mentorship in my entire PhD study provides me with the foundation for becoming a biostatistician and researcher. I would also like to express my gratitude to Dr Andriy Bandos and Dr David Gur, who I worked with for the last two years. The insightful guidance this group (Dr Rockette, Dr Bandos and Dr Gur) offered me has been as general as statistical methodologies, but also as detailed as scientific writing and presentation skills. I am also greatly thankful to Dr Sati Mazumdar for her help and support in preparation of the dissertation. Finally I want to thank my family, friends and our faculty members for their encouragement and support.

This research was supported in part by Grants EB006388, EB002106, EB001694 (to the University of Pittsburgh) from the National Institute for Biomedical Imaging and Bioengineering (NIBIB), National Institute of Health.

1.0 INTRODUCTION

1.1 ROC METHODOLOGY

An ROC experiment has been widely used in the evaluation and comparison of diagnostic technologies, practices or systems (often termed as *modalities*) [1,2,3,4]. In such an experiment, the true disease status of every subject is assumed to be known with certainty, either by an existing gold standard for indication of presence of such a disease/abnormality or by an independent assessment. Subjects with a known abnormality of interest are usually termed as "actually positive subjects" and those without a known abnormality are termed as "actually negative subjects". In an ROC experiment each subject is assigned a rating and by convention a higher rating indicates greater evidence of the presence of an abnormality. A diagnostic test is considered positive if the rating exceeds a certain threshold c representing the level of aggressiveness. We denote X and Y as the test results for actually negative and actually positive subjects respectively. The agreement between a diagnostic test and the true disease status can be summarized using two quantities: *True Positive Fraction* (*TPF*) and *False Positive Fraction* (*FPF*), and we define $TPF(c) = P(Y > c)$ and $FPF(c) = P(X > c)$.

An ROC curve is the plot of *TPF* versus *FPF* where the points on the graph are determined as the level of aggressiveness, c, is varied (Figure 1.1). Thus the ROC curve summarizes the agreement between ratings and the presence of an abnormality for all thresholds simultaneously.

One advantageous feature of ROC analysis is that it allows a more general comparison of diagnostic systems when they operate at different thresholds and a single pair of sensitivity-specificity estimates under this scenario is insufficient to describe the full range of diagnostic performance [4,5,6]. As discussed by Metz [4], there are many scenarios in the practice of diagnostic imaging where knowledge of the full ROC curve for different diagnostic systems is necessary in order to have adequate information to compare competing systems.

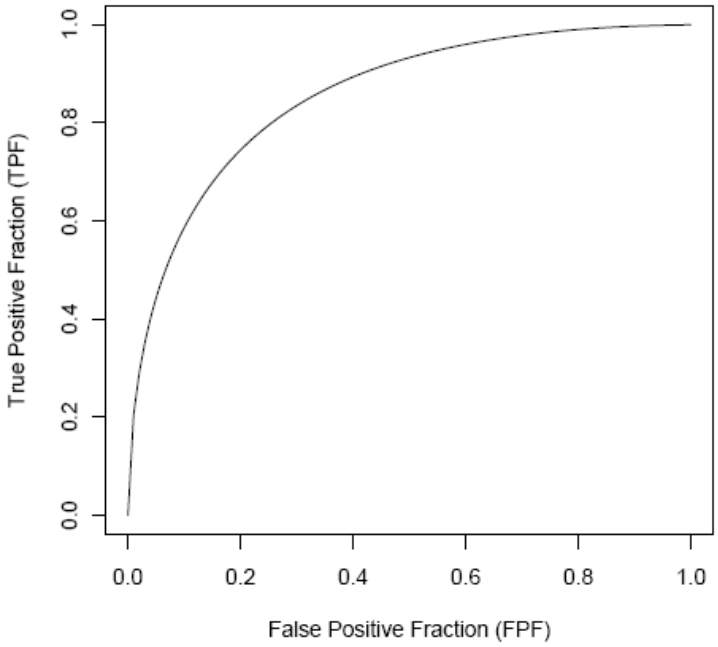

Figure 1.1 An ROC curve

In diagnostic imaging as well as in many other fields one of the most commonly used summary statistics derived from the ROC curve is the area under the ROC curve (AUC) [1,2]. The AUC reflects the inherent discriminative ability of a diagnostic system and can also be interpreted as the probability of correct discrimination between a randomly chosen pair consisting of an actually negative and an actually positive subject [1,7,8]. If ratings of two subjects do differ, then the subject with greater rating is declared as abnormal; otherwise, when both have equal ratings, the actually positive subject is selected randomly. An AUC of 0.5

corresponds to a poor diagnostic test, while an AUC equal to 1 implies a perfect diagnostic test. The area under the ROC curve can be expressed in the following way:

$$AUC = P(X < Y) + \frac{1}{2}P(X = Y)$$

The AUC can be estimated using parametric, nonparametric, and semi-parametric approaches [1,4,7-18]. For the parametric approaches, researchers usually assume a binormal [1,4] or transformable to a binormal [9] distribution for *FP* and *TP* ratings and use maximum likelihood estimation (MLE) to fit the so-called "binormal" ROC curves and to draw inference on AUC under the binormal assumption. Nonparametric approaches [7,8,10] utilize empirical ROC points by connecting them with straight lines. The resulting AUC is equivalent to the Wilcoxon statistic [7,8]. The most commonly used nonparametric inferential procedure is the approach based on the work by Delong *et al* [10], which is equivalent to the two-sample jackknife approach [19]. Several papers [11,12,13] discussed the estimation of AUC when data are correlated or clustered. In one paper, Obuchowski [11] proposed an asymptotic approach to account for the possible correlation within a cluster.

Several authors [9,14-17] investigated the use of kernel smoothing techniques, which lie between parametric and nonparametric approaches and are usually categorized as a semi-parametric approach. This approach applies kernel smoothing techniques to estimate the density function for *FP* and *TP* ratings respectively. By choosing the kernel bandwidths, h_x for *FP* ratings and h_y for *TP* ratings, the overall smoothness of the fit for an ROC curve can be varied. AUC can then be calculated from the estimated ROC curve and statistical inference can be drawn. A detailed comparison of several of the above methods is in Faraggi and Reiser [18].

1.2 TASK OF LOCALIZATION AND FROC METHODOLOGY

Since the 1970s, the ROC curve has been a valuable tool for describing and comparing the discriminative ability of a diagnostic system to separate actually negative subjects from actually positive subjects for the purpose of medical decision making [1]. Despite its wide use, experiments conducted in the ROC framework typically limit their measurements to an observer's overall rating of the subject for abnormality. In clinical situations where treatment is administered on the basis of the location of the abnormality, it may be important to summarize the observer's ability both to detect and correctly locate the abnormality [20-23].

To compare the diagnostic accuracy of different systems for such a task, three conceptual approaches have been considered. The Location ROC (LROC) approach defined by Starr *et al* [20] and further studied by Swensson [21] focuses on subjects with a single abnormality. The LROC approach suggests examining the locations of the reported "positive" diagnostic for each subject. A correct diagnostic (true positive) in an LROC concept is determined if the location of the reported "positive" contains the actual abnormality. An LROC curve is plotted as the fraction of true positive classifications with correct localization versus the conventional false positive fraction by varying all aggressiveness thresholds. For subjects with multiple abnormalities and hence with multiple ratings, Swensson [21] suggested the use of the maximum rating to represent the observer's overall opinion for the subject. An ROC curve and an LROC curve can be constructed using the maximum ratings of all subjects. With the assumption that the highest rating of those with correct localization is independent of the highest rating of those with incorrect localization, the area under the LROC curve is related to the area under the ROC curve in a simple equation. Swensson [21] concluded that the diagnostic performance with correct localization agrees with the performance that ignores correct localization.

An alternative approach of evaluating the diagnostic performance with correct localization is the region-of-interest (ROI) approach, proposed by Obuchowski *et al* [22]. The original ROI approach suggests partitioning the image into subimages based on clinical considerations and requests a response for each region separately. Each region becomes the unit of interest and can be analyzed in an ROC concept. However, due to the possible correlation of the multiple ratings on the same subject, standard errors cannot be correctly obtained using conventional ROC analysis, which assumes independence of the ratings and thus is invalid. Obuchowski *et al* [22] therefore presented a clustered analysis to account for the possible correlation of the ratings within each subject. A simple alternative for analyzing ROI data has also been proposed by Rutter [24], who suggests using a bootstrap method, with patients as the re-sampling unit to obtain standard errors.

As early as 1961, Egan *et al* [26] suggested the use of the "free response" method to address the task of detecting and locating multiple signals from noises in acoustics. Bunch *et al* [27] restudied this method, applied it to radiographic signal detection, and defined the Free-Response Operating Characteristic (FROC) approach. The FROC experiment involves presenting multiple abnormalities per image to the observer, and allowing the observer to make multiple responses on the image. Such an experiment entails detecting and marking the locations of all suspected abnormalities, as well as indicating a level of suspicion regarding the specific abnormality at each marked location. The number of marked locations is a variable that depends upon the image, the task and the observer's experience. A "mark" refers to a reported location on an examination of the image; a "rating" is a number representing the observer's degree of suspicion with regard to the abnormality associated with that mark. By analogy with conventional ROC analysis, a marked location is classified as a "positive" diagnostic if the associated rating is

above the aggressiveness threshold. A mark with a rating above the threshold is classified as a True Positive (*TP*) if the mark falls within a specified acceptable distance from the abnormality. The acceptable distance parameter is usually determined by the investigator in consultation with the clinicians. The commonly used rules for the acceptable distance parameter have been studied in [28-30]. It is important to keep the acceptance distance parameter constant when comparing competing FROC diagnostic systems.

Unlike the previous methods, in an FROC experiment the number of marks on a subject is completely determined by the observer, and the number of marks and the frequency of correct localization are considered to be important characteristics of the Free-Response system. The data obtained from the experiment is summarized in an FROC curve [27] which plots the true positive fraction (*TPF*) versus the average number of false positive marks (False Positive Rate or *FPR*) generated per image as the threshold varies (Figure 1.2).

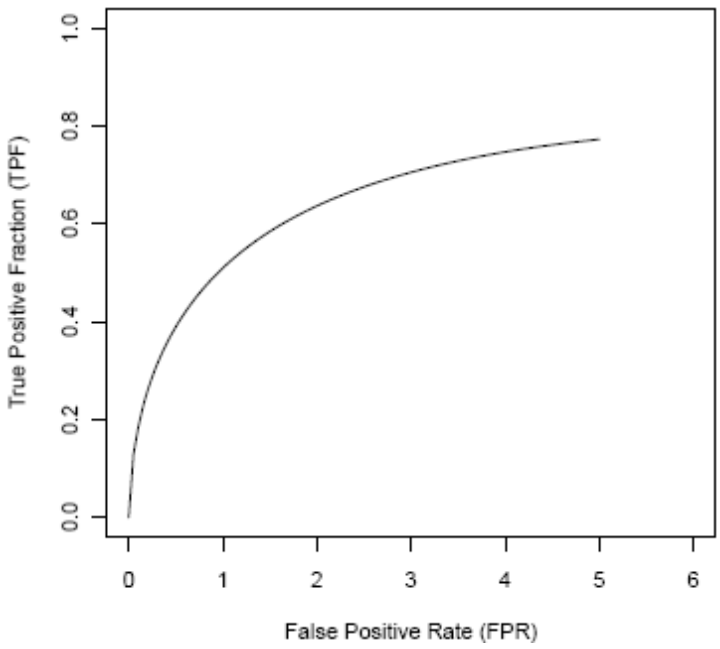

Figure 1.2 An FROC curve

1.3 METHODS OF FROC ANALYSIS

Edwards *et al* [31] attempted to characterize the process for an FROC experiment using an initial-detection-and-candidate-analysis (IDCA) model in which initial "candidate detections" (marks) are first located in the image by the observer who then produces a decision variable (rating) for each individual mark. Similar to the IDCA model [31], Chakraborty [32] presented a search model defined by a "pre-attentive" stage in which the observer uses mainly peripheral vision to identify likely lesion candidates, followed by a second stage in which the observer attentively decides whether to report an individual lesion candidate. In addition, Chakraborty [32] used three parameters to characterize the performance of an FROC diagnostic system: namely, the degree of separation between *FP* and *TP* ratings, the average number of *FP* marks per subject, and the proportion of lesions that are detected by the observer.

To summarize free-response diagnostic performance, two parametric approaches [23,31] have been proposed to fit an FROC curve using the MLE method. Chakraborty [23] modeled the likelihood function assuming a Poisson distribution for the number of false positive marks and a binormal distribution for the ratings. Edwards *et al* [31] suggested a formally different approach using the IDCA model. The detailed modeling process is described as follows. For FROC data, at the most aggressive threshold (the smallest threshold), the average number of *FP* marks per subject is denoted as FPR_0 and the *TP* fraction, obtained by dividing the total number of *TP* marks by the total number of lesions, is denoted as TPF_0. Regardless of the subject, conditioned on the total number of all marked locations, each rating can be treated as a unit of analysis and can be categorized as actually negative or actually positive for each threshold, *c*. An ROC curve ($FPF'(c)$, $TPF'(c)$) based on all of the ratings can be summarized for all thresholds simultaneously. The FROC curve (($FPR(c)$, $TPF(c)$) can be obtained by stretching the

corresponding ROC curve ($FPR_0 \times FPF'(c)$, $TPF_0 \times TPF'(c)$) to the last experimental point (TPF_0, FPR_0). Edwards *et al* [31] then modeled the likelihood function of an FROC curve, under the assumption of a binormal distribution for the ratings, a Poisson distribution for FPR_0 and a binomial distribution for TPF_0, and applied the MLE method to obtain the parameters in the distributions. This approach did not relax any of the distributional assumptions of Chakraborty [23], and both methods assumed independence among the observations within each subject.

Several summary indices characterizing the performance of a free-response system have been proposed. One method uses the estimate of the area under the empirical FROC curve. Although the area under the FROC curve summarizes the performance of the FROC system for all thresholds simultaneously, it may have inadequate information on the diagnostic system. By rewarding for *TPF* and penalizing for *FPR*, Bandos *et al* [33] proposed a summary index related to the area under the FROC curve and demonstrated that the index has well behaved clinical and statistical properties. Two threshold-dependent summary indices of an FROC curve also have been proposed. These include the *TPF* at a specific *FPR* [23] and the area under the FROC curve up to a specific *FPR* [34]. Similar to the corresponding indices for an ROC analysis, these threshold-dependent indices suffer from the subjectivity of selecting the *FPR* range, entail analytical complications associated with the uncertainty of the *FPR* related threshold and are potentially less precise than indices that consider all thresholds simultaneously [1,33].

An alternative approach summarizes the discriminative ability of an FROC diagnostic system by conducting the analysis in the format of an ROC approach. This approach assesses the ability to discriminate between “actually negative” and “actually positive” subjects by assuming a specific method of deducing summary opinion on a subject as a whole from the collection of all the ratings within this subject. Swensson [21] assumed that the highest rating on each individual

subject could be used to represent an observer's "first-choice" opinion. Based on the maximum ratings of all subjects, Swensson [21] constructed an ROC curve and used an ROC approach to analyze FROC data. Chakraborty [32] incorporated a parametric framework for the calculation of the figure of merit, θ, which was defined as the probability that the highest rating on an actually positive subject exceeds the highest rating on an actually negative subject. A companion paper [35] showed that the figure of merit, θ, is equivalent to the area of the ROC curve constructed with the maximum ratings of all subjects.

A limitation of the ROC approach of analyzing FROC data is that it does not incorporate correct location information of the ratings and the *FP* marks on an actually positive subject are compared to the *FP* marks on an actually negative subject. Thus, an actually positive subject could be correctly identified as more likely to be abnormal than an actually negative subject for the wrong reason. To evaluate the diagnostic performance with correct location information, Chakraborty and Berbaum [36] considered two indices incorporating the correct location of all ratings. They suggested treating the *FP* marks for an actually positive subject the same way as the *FP* marks for an actually negative subject. For an actually positive subject, the authors suggested that non-marked lesions should be assigned to the lowest rating by default and be treated the same way as the *TP* marks. Non-marked lesions and *TP* marks were then termed as lesion ratings. The first index denoted as *JAFROC1*, was defined by the average probability of a lesion rating exceeding the highest *FP* rating where the highest *FP* rating is measured on each subject. The second index ignored the *FP* ratings for an actually positive subject and *JAFROC2* was defined as the average probability of a lesion rating on an actually positive subject exceeding the highest *FP* rating on an actually negative subject. Chakraborty [36,38] also suggested incorporating different weights for each lesion within a subject when formulating

JAFROC indices. The sum of the lesion weights for each actually positive subject is equal to one and the values can be determined by the clinical significance in collaboration with clinicians [36].

Once a summary index is developed, it is often of interest to construct a valid statistical test to compare two FROC systems. To evaluate the validity of the statistical test, Chakraborty and Berbaum [36] suggested conducting a hypothesis test for the type I error rate. A 95% confidence interval was constructed on the estimate of the type I error rate, by assuming a binomial distribution with trial size equal to the number of simulations (2,000) and success rate equal to a nominal value of 0.05. Under the null hypothesis if the rejection rate of the statistical test based on the index fell outside of this confidence interval, they suggested that the underlying statistical test is not validated [36]. Furthermore, to compare validated statistical tests based on different indices, it is important to assess their statistical power for a pre-specified difference between two diagnostic systems. Chakraborty [36-38] considered the difference with regard to the degree of separation between *FP* and *TP* ratings.

Chakraborty and Berbaum [36] investigated the validity of the statistical tests based on the three indices discussed above (θ, *JAFROC1* and *JAFROC2* with equal weights) and assessed their statistical power in a multi-reader design setting. In such an FROC experiment, every subject is evaluated (marked and rated) by each reader under both modalities. The number of marks and ratings of the same subject evaluated by different readers are both likely to be correlated. In addition, the number of marks and ratings for two modalities obtained for the same reader may also be correlated. In the simulation, Chakraborty and Berbaum [36] evaluated all three indices using 100 actually negative and 100 actually positive subjects. It was assumed that there was only one marked lesion on each actually positive subject and the number of *FP* marks

on each subject was a constant value of T. They first simulated *FP* and *TP* ratings from a multi-reader ROC experiment with the use of a mixed-effect model presented by Roe and Metz [39], where modality was treated as a fixed effect and subject and reader were treated as random effects. To obtain the data for an FROC experiment, they assigned a *TP* rating and T *FP* ratings on an actually positive subject, and assigned T *FP* ratings on an actually negative subject. The random effects in the mixed-effect model allow the existence of correlation among ratings for the same subject evaluated by different modalities and/or different readers. To construct a statistical test for each of the indices, they applied the Dorfman-Berbaum-Metz (DBM) method [49], which is one of the most commonly used approaches for analyzing multi-reader ROC data. Jackknife pseudo-values of each index were generated by removing each subject separately. An ANOVA-based procedure was used to test the hypothesis of whether the modality-specific indices are equivalent when there is a system difference with regard to the degree of separation between *FP* and *TP* ratings. Then they evaluated the statistical tests based on θ, *JAFROC1* and *JAFROC2* respectively. The simulation results showed that the statistical test based on *JAFROC1* did not pass the above validation test. Chakraborty and Berbaum [36] also suggested ignoring the *FP* ratings for an actually positive subject (*JAFROC2*). Using a similar analysis, *JAFROC2* was found to have passed the validation test and to have more statistical power than θ.

1.4 OBJECTIVES

1.4.1 To characterize the discriminative ability of an FROC diagnostic system.

We propose to develop different groups of indices to reflect the discriminative ability of the FROC system. Several existing approaches of analyzing FROC data [21,32,36] use the maximum function to combine the multiple ratings on a subject and summarize the diagnostic

system's subject-based discriminative ability. We propose and evaluate new methods by considering combination functions based on the average function and on the Wilcoxon statistic. Indices are also formulated in order to summarize the diagnostic performance with correct location information by improving modified *JAFROC* indices. Indices are also proposed to modify the Wilcoxon statistic to summarize the diagnostic performance that incorporates the number of "correct" and "incorrect" marks on each subject. We will evaluate properties of the proposed indices in a simulation study. We will compare the statistical tests based on the different indices by estimating the type I error rate and the statistical power for selected differences. The complex structure of FROC data results in multiple ways in which two diagnostic systems may differ. In this dissertation, we will focus on two types of differences: namely, the degree of separation between *FP* and *TP* marks and the average number of *FP* marks per subject.

(a). The first group of indices is based on the marked ratings within a subject, ignoring whether the mark contains an abnormality. This can be viewed as an analog of the assessment that results when the system uses of all the ratings to evaluate a subject as a whole for abnormalities. We will consider an existing index of this type based on the maximum function and propose two new indices based on a comparison of the average ratings and a comparison based on stochastic order. The two combination functions can incorporate more information on a subject and they are potentially more stable than the maximum function. We will derive closed-form expressions for the variances of these indices, and compare the power of the statistical tests based on the three indices in a simulation study.

(b). The indices in (a) ignores the correct location information of the marks within a subject. *JAFROC* indices in literature [36] were constructed by incorporating location information of the

ratings. They considered comparing the maximum *FP* ratings to the lesion ratings. The second group of indices that we propose tend to improve modified *JAFROC* indices. As will be discussed in Chapter 3, we will propose different types of indices based on three different handling methods and three different comparison functions that might outperform existing methods for the task of assessing the ability of a diagnostic system to separate the *FP* and *TP* ratings. Closed form expressions of the two-sample jackknife variances for these indices will be derived and used to develop asymptotic procedures. Clustered ROC index [11] will be applied to FROC setting and studied in this Chapter. A simulation study will be conducted to evaluate the statistical tests based on these indices.

(c). Reader evaluations in an FROC system are partially reflected by the average number of *FP* marks per subject [32,35,36]. Thus it is important to propose an index that can successfully penalize for an increased number of generated "incorrect" marks. In (b) when we compared *FP* ratings to lesion ratings, our proposed indices emphasized whether the lesion ratings were in some sense larger than the *FP* ratings and treated the number of *FP* ratings on each subject as a nuisance parameter. Although the indices using the maximum indirectly penalized for an increased number of *FP* marks and rewarded for the number of *TP* marks by allowing the comparison between *FP* marks and non-marked lesions, there was no attempt in part (b) to explicitly adjust for a difference in the number of *FP* marks for each of the two systems. Thus, the third group of indices we propose to explore in this dissertation attempts to incorporate additional information on the number of marks when comparing *FP* and *TP* ratings. Specifically, when comparing two groups of ratings, we modify the Wilcoxon statistic so that it can successfully penalize for an increased number of generated "incorrect" (*FP*) marks and reward for an increased number of generated "correct" (*TP*) marks. For the proposed indices we will

derive asymptotic inferential procedures, investigate their performance and compare them to modified *JAFROC* indices in a simulation study.

1.4.2 To develop inferential procedures in the multi-reader setting

The statistical procedures in 1.4.1 are not specifically constructed for problems where the subjects are evaluated by multiple readers. In the 1990s, it became widely recognized that several sources of variability within and between readers should be considered in the assessment of medical imaging [40,42,44-48]. In the free-response multi-reader setting, both the number of marks and the ratings of the same subject may be correlated. We will develop a multi-reader model that allows one to impose the correlations between the number of marks and their ratings, as well as incorporating reader heterogeneity for multi-reader FROC data. We will evaluate our proposed indices for the single reader setting and develop inferential procedures for the FROC indices in a multi-reader design setting. Specifically, the two-sample jackknife approach will be evaluated for the reader-averaged FROC indices. We will also apply the traditional ANOVA-based approach (DBM method [49]) for the reader-averaged FROC indices. The proposed inferential procedures will be investigated in a simulation study.

1.4.3 To apply three different approaches of fitting an FROC curve

In this section, we propose to extend to the FROC setting two approaches [9,17] that were originally developed to fit a smooth ROC curve. The first approach applied the Box-Cox power transformation to the ROC ratings, assumed a binormal distribution for the transformed ratings and used the MLE method to construct a smooth binormal ROC curve [9]. Under an ROC setting, Lloyd and Yong [17] used a two-stage plug-in method to find kernel bandwidths,

estimated the kernel density function for *FP* and *TP* ratings, and fit a smooth ROC curve. For the smooth ROC curves estimated by the above two approaches, the smooth FROC curves can be obtained by stretching the corresponding ROC curves to the last experimental point [31]. We will also propose to use a kernel regression approach to regress *TPF* on *FPR* using the empirical points in the FROC plot and construct a smooth FROC curve. This kernel regression approach allows us to estimate a smooth FROC curve without the independence assumption between the number of marks and the ratings of the subjects. For the three considered approaches, we will develop explicit formulations for the smooth FROC curves estimated using Box-Cox power transformation, kernel smoothing and kernel regression approaches. The areas under the estimated FROC curves by three different approaches will also be formulated and evaluated in a simulation study.

2.0 SUBJECT-BASED APPROACH TO EVALUATE FROC SYSTEMS

2.1 BACKGROUND

The consequences of misspecifying the location of an abnormality on a subject and the consequences of failing to identify a subject that has an abnormality are fundamentally different. To evaluate the diagnostic performance at the subject's level may be of interest in applications where the diagnostic test under study attempts to identify actually positive subjects for additional testing [42]. Several researchers [21,32,35,36] have already investigated the aspect of an FROC diagnostic system that relates to correctly discriminating between an actually negative and an actually positive subject by using concepts inherent to the ROC paradigm. Swensson [21] assumed the highest-rated rating on each individual subject as an observer's "first-choice" report. A figure of merit θ can then be defined as the probability that the highest rating on an actually positive subject exceeds the highest rating on an actually negative subject, and this approach has been considered under both nonparametric [36] and parametric [32,35] frameworks.

Given a set of FROC data a subject-based assessment can be achieved by combining the information on all suspicious locations within a subject. However, such reduction of the FROC data is not unique and can be achieved by using a wide range of combination functions. Although the highest (maximum) rating on the subject has long been assumed to represent the

observer's opinion regarding the subject's abnormality status, no comprehensive investigation of this assumption has been performed. It seems reasonable to expect that an effective decision scheme could "form an opinion" about the entire subject based not only on the single most suspicious location but also considering all suspicious locations and their associated ratings. It is possible that depending on the pattern of location-based ratings in actually negative and actually positive subjects, different combination functions can lead to different conclusions about the subject-based discriminative ability of a system. At the same time, the use of the maximum rating within a subject as a combination function may have inferior statistical properties as compared to the mean, which is the most commonly used combination function to compare two sets of multiple ratings.

In this Chapter we present a general framework that encompasses the maximum rating approach (θ) as well as other indices. As an alternative to θ, we consider two natural indices, develop simple nonparametric procedures for statistical inferences, and compare the three indices in a simulation study. All three indices belong to a general family that includes indices for estimating the probability of a correct discrimination in a corresponding subject-based two-alternative-forced-choice (2-AFC) experiment. The considered indices quantify the ability of the system to discriminate between actually negative and actually positive subjects. Each index in the considered family is determined by a specific function for comparing the collection of ratings on two different subjects. The maximum rating index, θ, corresponds to the comparison of maximum ratings on two subjects, while the two newly proposed indices correspond to the comparison based on the average ratings (mean), A_1, and the comparison based on the stochastic order of the sets of ratings on two different subjects (the Wilcoxon statistic), A_2.

2.2 METHODS

The data from an FROC experiment for S_0 actually negative and S_t actually positive subjects with a fixed number of abnormalities *t* can be summarized as follows:

$$\begin{array}{ll} \{x_{s'c}^0\}_{c=1}^{n_{s'}^0}, & s'=1,...S_0 \quad \leftrightarrow \text{ "actually negative"} \\ \left(\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}\}_{c=1}^{m_s}\right), & s=1,...S_t \quad \leftrightarrow \text{ "actually positive"} \end{array} \tag{2.1}$$

where s' indexes an actually negative subject and s indexes an actually positive subject. We use $n_{s'}^0$ and n_s^t to represent the number of *FP* marks on an actually negative and an actually positive subject respectively; m_s to represent the number of *TP* marks on an actually positive subject and $\vec{x}^0$ and $\vec{x}^t$ to represent the collection of ratings for the *FP* marks for an actually negative subject and an actually positive subject respectively. $\vec{y}$ is the collection of ratings for *TP* marks and *c* is used to index the collection of ratings for each individual subject. We use $x_{s'c}^0$ to represent the *c*-th *FP* rating on the s'-th actually negative subject, x_{sc}^t to represent the *c*-th *FP* rating on the s-th actually positive subject, and y_{sc} to represent the value for the *c*-th *TP* rating on the actually positive subject *s.* We treat the observed data as a realization of a collection of random variables and distinguish between the random quantities from their realizations with capital letters. Thus, $\vec{X}^0$, $\vec{X}^t$ and $\vec{Y}$ are the random vectors of the ratings on a subject with length N^0, N^t and M respectively, namely: $\vec{X}^0 = \begin{pmatrix} X_1^0 \\ \vdots \\ X_{N^0}^0 \end{pmatrix}$ $\vec{X}^t = \begin{pmatrix} X_1^t \\ \vdots \\ X_{N^t}^t \end{pmatrix}$ $\vec{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_M \end{pmatrix}$. The data structure introduced here will be used throughout the entire dissertation.

A natural and commonly used index for summarizing the discriminative ability of a diagnostic system is the percent of correct discrimination in all possible pairs of actually

negative-actually positive subjects. The task of discrimination in such a pair of subjects is called a 2-Alternative Forced Choice (2AFC) task. If the decision in 2AFC is guided by the comparison of a certain subject-specific scalar (ordinal rating), the percent correct in 2AFC is equivalent to the area under the ROC curve constructed on these ratings [7,8]. However, not every 2AFC task can be described by an underlying ordinal rating and hence the percent correct in 2AFC is a more general index than the area under the ROC curve. In order to compute the percent correct in a 2AFC experiment we need to know the value of a corresponding comparison function ψ for every pair of actually negative–actually positive subjects. For the 2AFC task that is guided by comparing subjects-specific ratings the ψ function can be defined as a result of comparison of two numbers *b* and *c*, namely:

$$\psi(b,c) = \begin{cases} 1 & b < c \\ ½ & b = c \\ 0 & b > c \end{cases}$$

With FROC data the 2-AFC task is complicated by the need to compare vectors of observations $\vec{b}$ and $\vec{c}$. For this purpose we define a generalization of the ψ function, i.e.:

$$\tilde{\psi}\left(\vec{b},\vec{c}\right) = \begin{cases} 1 & \vec{b} \lhd \vec{c} \\ ½ & \vec{b} \lhd\rhd \vec{c} \\ 0 & \vec{b} \rhd \vec{c} \end{cases} \tag{2.2}$$

where $\vec{b} \lhd \vec{c}$ indicates that the collection of ratings in $\vec{c}$ dominates the collection of ratings in $\vec{b}$ based on a pre-selected rule ($\lhd$); and $\vec{b} \lhd\rhd \vec{c}$ corresponds to the subject where the ratings in $\vec{b}$ and $\vec{c}$ are equivalent according to a pre-selected rule ($\lhd\rhd$).

For the problem considered in this Chapter we use $\vec{b} = \vec{x}^0$ and $\vec{c} = \{\vec{x}^t, \vec{y}\}$, i.e., we compare the ratings on an actually negative subject to the ratings on an actually positive subject. In an

FROC experiment there might be subjects with no marks, hence subjects without any ratings. This leads to the possibility of having empty vectors $\vec{b}$ and $\vec{c}$. Therefore we augment the definition of $\tilde{\psi}$ in (2.2) by adopting the approach proposed by Chakraborty [32]. For a pair of actually negative and actually positive subjects, if only the actually negative ($\vec{x}^0 = \varnothing \; and \; \{\vec{x}^t, \vec{y}\} \neq \varnothing$) or only the actually positive subject ($\vec{x}^0 \neq \varnothing \; and \; \{\vec{x}^t, \vec{y}\} = \varnothing$) is not marked, we assign $\tilde{\psi}$ to be 1 or 0 respectively; if both subjects are not marked ($\vec{x}^0 = \varnothing \; and \; \{\vec{x}^t, \vec{y}\} = \varnothing$), we assign $\tilde{\psi}$ to be 0.5. Specifically, we define $\tilde{\psi}$ as follows:

$$\tilde{\psi}\left(\vec{b}, \vec{c}\right) = \begin{cases} 1 & \vec{b} \lhd \vec{c}, \quad or \; \vec{b} = \varnothing \; and \; \vec{c} \neq \varnothing \\ \frac{1}{2} & \vec{b} \lhd\rhd \vec{c}, \; or \; \vec{b} = \vec{c} = \varnothing \\ 0 & \vec{b} \rhd \vec{c}, \quad or \; \vec{b} \neq \varnothing \; and \; \vec{c} = \varnothing \end{cases}$$

In this Chapter we consider three specific indices from a family that quantifies the probability of correct discrimination in a 2-AFC task. Each of the indices in this family can be written as

$$A = E\left[\tilde{\psi}\left(\left\{\vec{X}^0\right\}, \left\{\vec{X}^t, \vec{Y}\right\}\right)\right] \tag{2.3}$$

Let's first focus on a pair of actually negative and actually positive subjects on which at least one subject has no marks. If only the actually negative subject is not marked, it contributes a value of $1 \times P(\vec{X}^0 = \varnothing)\left[1 - P(\{\vec{X}^t, \vec{Y}\} = \varnothing)\right]$ to the expectation; if only the actually positive subject is not marked, it contributes a value of zero to the expectation; if neither of them is marked, it contributes a value of $\frac{1}{2} \times P(\vec{X}^0 = \varnothing)\, P(\{\vec{X}^t, \vec{Y}\} = \varnothing)$ to the expectation. All pairs of subjects with at least one subject without any marks contribute $P\left(\vec{X}^0 = \varnothing\right) \times \left[1 - \frac{1}{2} P\left(\vec{X}^t = \varnothing, \vec{Y} = \varnothing\right)\right]$ to the expectation in (2.3) for each of the indices in the family.

Using the maximum combination function is a special case of (2.3), and we can represent the figure of merit θ [32] in the following manner:

$$\theta = A_0 = E\left[\psi\left(\max\left\{\vec{X}^0\right\}, \max\left\{\vec{X}^t, \vec{Y}\right\}\right) \times I\left(\left\{\vec{X}^0\right\} \neq \varnothing, \left\{\vec{X}^t, \vec{Y}\right\} \neq \varnothing\right)\right] + P\left(\vec{X}^0 = \varnothing\right) \times \left[1 - \frac{1}{2} P\left(\vec{X}^t = \varnothing, \vec{Y} = \varnothing\right)\right] \quad (2.4)$$

where $I(\bullet)$ is the indicator function.

The maximum rating is only one of multiple possible functions that can be used with this approach. One commonly used combination function is the mean and our first index (A_1) is based on the comparison of averages ratings in a pair of subjects, namely:

$$A_1 = E\left[\psi\left(\frac{\sum_{c_1=1}^{N^0} X_{c_1}^0}{N^0}, \frac{\sum_{c_2=1}^{N^t} X_{c_2}^t + \sum_{c_3=1}^{M} Y_{c_3}}{N^t + M}\right) \times I\left(\left\{\vec{X}^0\right\} \neq \varnothing, \left\{\vec{X}^t, \vec{Y}\right\} \neq \varnothing\right)\right] + P\left(\vec{X}^0 = \varnothing\right) \times \left[1 - \frac{1}{2} P\left(\vec{X}^t = \varnothing, \vec{Y} = \varnothing\right)\right] \quad (2.5)$$

Another commonly using index is the Wilcoxon statistic and our second index (A_2) is based on the comparison of stochastic order of the sets of ratings in a pair of subjects, namely:

$$A_2 = E\left\{\psi\left[0.5, w\left(\vec{X}^0, \left\{\vec{X}^t, \vec{Y}\right\}\right)\right] \times I\left(\left\{\vec{X}^0\right\} \neq \varnothing, \left\{\vec{X}^t, \vec{Y}\right\} \neq \varnothing\right)\right\} + P\left(\vec{X}^0 = \varnothing\right) \times \left[1 - \frac{1}{2} P\left(\vec{X}^t = \varnothing, \vec{Y} = \varnothing\right)\right] \quad (2.6)$$

where $w\left(\vec{b}, \vec{c}\right) = \dfrac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \psi(b_i, c_j)}{n_1 \times n_2}$, if $\vec{b} = (b_1, ..., b_{n_1})$ and $\vec{c} = (c_1, ..., c_{n_2})$.

The three indices formulated above have the interpretation of the percent of correct decisions in a 2AFC task where the decisions are determined according to a chosen comparison function (based on maximum, average or stochastic order). The two indices based on the maximum ratings (θ or A_0) and based on the average ratings (A_1) are equivalent to the areas under the ROC curves constructed based on the maximum and average of the within-subject ratings correspondingly (with an artificial lowest rating assigned to the subjects with no marks). This however is not true in general for the index A_2 which is based on the stochastic order of the sets

of ratings on two different subjects, because the stochastic order relation does not in general possess the transitivity property inherent to standard order relations.

2.3 STATISTICAL INFERENCE

The nonparametric estimators of the indices considered in section 2.2 can be applied for the data in (2.1):

$$\hat{\theta} = \hat{A}_0 = \frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t} \tilde{\psi}_{\max}\left(\{x_{s'c}^0\}_{c=1}^{n_{s'}^0}, \left\{\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}\}_{c=1}^{m_s}\right\}\right)}{S_0 \times S_t} \text{ or}$$

$$\hat{\theta} = \hat{A}_0 = \frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t} \psi\left(\max(\{x_{s'c}^0\}_{c=1}^{n_{s'}^0}), \max(\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}\}_{c=1}^{m_s})\right) \times I\left(n_{s'}^0 n_s^t m_s \neq 0\right)}{S_0 \times S_t} + \left[\frac{\sum_{s'=1}^{S_0} I\left(n_{s'}^0 = 0\right)}{S_0}\right] \times \left[1 - \frac{1}{2} \times \frac{\sum_{s=1}^{S_t} I\left(n_s^t + m_s = 0\right)}{S_t}\right] \quad (2.7)$$

$$\hat{A}_1 = \frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t} \tilde{\psi}_{mean}\left(\{x_{s'c}^0\}_{c=1}^{n_{s'}^0}, \left\{\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}\}_{c=1}^{m_s}\right\}\right)}{S_0 \times S_t} \text{ or}$$

$$\hat{A}_1 = \frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t} \psi\left(\frac{\sum_{c_1=1}^{n_{s'}^0} x_{s'c_1}^0}{n_{s'}^0}, \frac{\sum_{c_2=1}^{n_s^t} x_{sc_2}^t + \sum_{c_3=1}^{m_s} y_{sc_3}}{n_s^t + m_s}\right) \times I\left(n_{s'}^0 n_s^t m_s \neq 0\right)}{S_0 \times S_t} + \left[\frac{\sum_{s'=1}^{S_0} I\left(n_{s'}^0 = 0\right)}{S_0}\right] \times \left[1 - \frac{1}{2} \times \frac{\sum_{s=1}^{S_t} I\left(n_s^t + m_s = 0\right)}{S_t}\right] \quad (2.8)$$

$$\hat{A}_2 = \frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t} \tilde{\psi}_{wilcoxon}\left(\{x_{s'c}^0\}_{c=1}^{n_{s'}^0}, \left\{\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}\}_{c=1}^{m_s}\right\}\right)}{S_0 \times S_t} \text{ or}$$

$$\hat{A}_2 = \frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t} \psi\left\{0.5, \frac{\sum_{c_1=1}^{n_{s'}^0}\sum_{c_2=1}^{n_s^t} \psi(x_{s'c_1}^0, x_{sc_2}^t) + \sum_{c_1=1}^{n_{s'}^0}\sum_{c_3=1}^{m_s} \psi(x_{s'c_1}^0, y_{sc_3})}{n_{s'}^0 \times (n_s^t + m_s)}\right\} \times I\left(n_{s'}^0 n_s^t m_s \neq 0\right)}{S_0 \times S_t} + \left[\frac{\sum_{s'=1}^{S_0} I\left(n_{s'}^0 = 0\right)}{S_0}\right] \times \left[1 - \frac{1}{2} \times \frac{\sum_{s=1}^{S_t} I\left(n_s^t + m_s = 0\right)}{S_t}\right] \quad (2.9)$$

To illustrate the estimation of all three indices, we provide an example using 4 actually negative and 4 actually positive subjects in Appendix.

The estimators in (2.7)-(2.9) have a structure similar to a two-sample U-statistic which permits the development of a closed form expression for the jackknife or bootstrap variance when the re-sampling techniques consider subject as a sampling unit. We propose to use a two-sample jackknife variance [19] for the indices in this Chapter. The two-sample jackknife variance when applied to the area under the empirical ROC curve is known to be equivalent to the variance proposed by Delong *et al* [10]. The two-sample jackknife approach involves stratifying the subjects (in our case, actually negative subjects and actually positive subjects) and calculating the variance of pseudo-values in each stratum respectively. When applied to the AUC (area under the ROC curve) variance formula under the ROC analysis, it is formally different than the commonly used one-sample jackknife approach and the two-sample variance is uniformly smaller [43] than or equal to the one-sample variance. Note that for our indices when we remove an actually negative or an actually positive subject, we calculate the pseudo-values from the $\{\tilde{\psi}_{s's}\}$ matrix the same way as we do for the ROC analysis. We will use a computation algorithm in a manner similar to Delong's approach. Specifically, we:

1. Find the simple averages of each row and column in the matrix $\{\tilde{\psi}_{s's}\}$ where $\tilde{\psi}_{s's} = \tilde{\psi}\left(\{\vec{x}^0_{s'}\}, \{\vec{x}^t_s, \vec{y}_s\}\right)$. Namely, for the s' th row (corresponding to the s' th actually negative subject): $\overline{\tilde{\psi}_{s'\bullet}} = \frac{\sum_{s=1}^{S_t} \tilde{\psi}_{s's}}{S_t}$ and for the sth column (corresponding to the sth actually positive subject): $\overline{\tilde{\psi}_{\bullet s}} = \frac{\sum_{s'=1}^{S_0} \tilde{\psi}_{s's}}{S_0}$.

2. Compute the unbiased estimates of the variance of the averages of the rows and columns or the variance elements that result from actually negative and actually positive subjects correspondingly.

Namely, for the rows (due to actually negative subjects): $\hat{V}_N = \frac{\sum_{s'=1}^{S_0} (\overline{\tilde{\psi}_{s'\bullet}} - \overline{\tilde{\psi}_{\bullet\bullet}})^2}{S_0 - 1}$

and for the columns (due to actually positive subjects): $\hat{V}_P = \frac{\sum_{s=1}^{S_t} \left(\overline{\tilde{\psi}_{\bullet s}} - \overline{\tilde{\psi}_{\bullet\bullet}}\right)^2}{S_t - 1}$

3. Compute the variance using: $\hat{V}_{\text{2s-jackk}}(\hat{A}) = \frac{\hat{V}_N}{S_0} + \frac{\hat{V}_P}{S_t}$ or

$$\hat{V}_{\text{2s-jackk}}(\hat{A}) = \frac{\sum_{s'=1}^{S_0} \left(\overline{\tilde{\psi}_{s'\bullet}} - \overline{\tilde{\psi}_{\bullet\bullet}}\right)^2}{S_0 \times (S_0 - 1)} + \frac{\sum_{s=1}^{S_t} \left(\overline{\tilde{\psi}_{\bullet s}} - \overline{\tilde{\psi}_{\bullet\bullet}}\right)^2}{S_t \times (S_t - 1)} \tag{2.10}$$

When comparing two diagnostic systems evaluated under the FROC paradigm with the proposed indices we use the difference between the modality-specific indices. When assessing statistical significance of the differences in the indices observed for the two modalities, we propose an asymptotic procedure with the test statistics: $Z_i = \frac{\hat{A}_i^2 - \hat{A}_i^1}{\sqrt{\hat{V}_{\text{2s-jackk}}\left(\hat{A}_i^2 - \hat{A}_i^1\right)}}$, i=0,1,2.

To conduct a statistical test we compare the Z-statistics described above with the pre-specified percentile of the standard normal distribution [10]. In an unpaired design where different subjects are evaluated using different systems the estimator of the variance of the difference is simply the sum of the corresponding variance estimators. In a paired design, where the same set of subjects is evaluated under both modalities the estimator of the variance of the difference can be computed using equation (2.10) where $\tilde{\psi}$ is replaced with $\tilde{\psi}^1 - \tilde{\psi}^2$.

2.4 SIMULATION RESULTS

We evaluated all three indices under different scenarios in each of which we generate 10,000 independent datasets based on a set of pre-determined parameters. Each dataset consisted of 20 actually negative and 20 actually positive subjects. The notations for the observations that are typically obtained in an FROC experiment are summarized in expression (2.1). In this simulation study we consider the scenario where the sample consists of a group of actually negative subjects with zero known abnormalities and a group of actually positive subjects with *t* lesions. Originally proposed by Bunch *et al* [27] and further employed by other researchers [23,31,32], the number of *FP* marks *N* on a subject is usually treated as a Poisson variable with parameter λ. The parameter λ can be viewed as the mean number of *FP* marks on a subject and a smaller value is expected for an experienced observer [32,35,36]. The number of *TP* marks *M* on a subject is typically modeled by a binomial distribution. The trial size is equal to the total number of lesions *t* on a subject. The success rate ν regulates the proportion of lesions that are actually detected namely, marked at the right locations.

For all subjects, regardless of their actual ratings, the number of *FP* marks, *n*, was generated from a Poisson distribution with expectation λ of 0.5, 1.0, and 2.0. For every actually positive subject the number of *TP* marks, *m*, was generated from a binomial distribution with number of trials *t* of 1 and 3 and probability of success ν of 0.5, 0.7 and 0.9. The ratings for *FP*, $\vec{x}$, and *TP*, $\vec{y}$, marks were generated independently from normal distributions with means and variances chosen to achieve a pre-specified degree of separation between *FP* and *TP* ratings corresponding to *AUC* of 0.5, 0.7 and 0.9 and we allow $\frac{1}{b}=\frac{\sigma_Y}{\sigma_X}$ to be either 1 or 2, the latter representing the

case where the ratings for the abnormalities are more variable than the ratings for normal regions as it is often the case in ROC studies [1]. In the evaluation of all three indices we used a range of parameters that include values that we have observed in breast cancer imaging studies. The parameters we choose are a reasonable expansion of the dataset in Bandos *et al* [33].

We also evaluated the three indices when ratings follow a pair of skewed distributions with a non-zero mass at the extremes. These distributions were created by grouping the normal distributions. The ratings below the 40^{th} percentile of the distribution of the *FP* ratings were assigned to the 40^{th} percentile and the ratings above the 60^{th} percentile of the distribution of the *TP* ratings were assigned the value of the 60^{th} percentile. As a result, the distribution of the *FP* ratings becomes right-skewed (e.g., the skewness is 0.832 for AUC=0.8 and b=1) and the distribution of the *TP* ratings becomes left-skewed (e.g., the skewness is -0.832 for AUC=0.8 and b=1). The histograms of the skewed ratings are listed after the text of this Chapter (Figure 2.1).

The simulation model described above is slightly different from the search model presented by Chakraborty [32]. In our model we allow for more flexibility by permitting the variance of the ratings to be different for *FP* and *TP* marks. We also evaluate the different methods under a non-normal distribution. Because of the large number of simulation scenarios we considered, only a fraction of these are included in the tables which are also listed after the text of this Chapter.

Table 2.1 summarizes the estimated expectations and standard errors of the three estimators when data are generated from the simulation model assuming a binormal distribution and λ=1. The standard errors were estimated both with a sample variance over the simulated realizations and by using the two-sample jackknife variance (2.10). For each of the three indices and for all scenarios that we considered the empirical standard error is closed to the estimated two-sample jackknife standard error. Specifically it is covered by the inter-quartile interval of the empirical

distribution of the two-sample jackknife estimate of the standard error. The standard error of $\hat{\theta}$ tends to be lower than the standard error of $\hat{A}_1$ or $\hat{A}_2$.

In the adopted simulation model, when the distributions of the ratings for the actually negative and actually positive subjects (the latter have a mixture of distribution of ratings for *FP* and *TP* marks) have the same location (corresponding to *AUC*=0.5), the estimated expectations range from about 0.59 to 0.87, with $\hat{A}_1$ and $\hat{A}_2$ being below 0.7 for most of the simulated scenarios. In fact, under the simulation model in which the ratings of the *FP* and *TP* marks follow the same normal distributions (*AUC*=0.5, *b*=1), the expectations for the indices A_1 and A_2 can be computed directly using formulas (2.5) and (2.6). We list their expectations and expected frequencies of subjects without any marks for all indices in Table 2.2.

The phenomenon that the expectation of A_1 and A_2 are substantially greater than 0.5 when *AUC*=0.5 can be partially attributed to an imbalance in frequencies of actually negative and actually positive subjects without any marks. Specifically, the substantially larger frequency of actually positive subjects without any marks as compared with the actually negative subjects (Table 2.2) poses an imbalanced frequency of $\tilde{\psi}$=1 (high) and $\tilde{\psi}$=0 (low), and hence, shifts the expectation of the comparison function $\tilde{\psi}$ towards higher values. Thus, unlike the index of the average performance in a conventional ROC experiment, the expectations of the considered indices are not necessarily 0.5 when the ratings of actually negative and actually positive subjects have the same location (*AUC*=0.5).

A comparison of the estimated expectations of the three indices suggests that $E\left(\hat{\theta}\right)$ is always higher than $E\left(\hat{A}_1\right)$ and $E\left(\hat{A}_2\right)$. In fact, under our simulation scenario in which *FP* and *TP* ratings follow the same normal distribution (*AUC*=0.5, *b*=1) the expectation of $\hat{\theta}$, unlike that of

$\hat{A}_1$ and $\hat{A}_2$, would be higher than 0.5 even in a subpopulation of subjects with at least one mark. The reason for this phenomenon is that the "maximum of rating", hence, $\hat{\theta}$, is substantially affected not only by the actual value of ratings but also by the number of marks, and in our simulation model the actually positive subjects have on average a higher total number of marks (*FP* and *TP*) than that on actually negative subjects (*FP* only). This property also partially contributes to the fact that the expectation of $\hat{\theta}$ is greater than that of $\hat{A}_1$ and $\hat{A}_2$ when distributions of the *FP* and *TP* ratings are separated ($AUC(FP,TP)>0.5$).

From the formulation of the indices in (2.4)-(2.6) and the results in Table 2.2 one can also see that a pair of actually negative and actually positive subjects in which at least one subject has no marks affects all three considered indices in the same manner. Hence, an increasing frequency of subjects without any marks can be expected to make the difference between the indices less profound and thus to attenuate the difference in the power of the corresponding statistical tests.

For both normal and skewed distributions with nonzero mass at the extremes (Table 2.3), the estimated type I error rate is close to the nominal value. However, when both the number of abnormalities on the actually positive subjects and the degree of separation between *FP* and *TP* ratings (as measured by *AUC*) is extremely large the estimated type I error rate is low for all three indices, and the procedure based on the maximum rating index, θ, demonstrates the greatest degree of conservativeness.

As we discussed in Chapter 1, the complex structure of FROC data results in multiple ways in which two diagnostic systems may differ. Here we focus on two types of such difference. First, we consider two diagnostic systems which are equal with respect to all the parameters except in regard to the discrimination between *FP* and *TP* ratings (degree of separation between

FP and *TP* ratings, parameter *AUC*). Second, we consider two diagnostic systems that differ only with respect to the average number of the *FP* marks on a subject (λ).

Table 2.4 shows the estimates of the statistical power for the scenario where the two diagnostic systems differ in regard to the degree of separation between *FP* and *TP* ratings. From this table one can observe that for samples of subjects in which all actually positive subjects have a single abnormality (t=1), and where *FP* and *TP* ratings follow normal distributions with equal variance (b=1), in most of the scenarios the statistical test based on θ has the greatest statistical power to detect the difference between the two systems. However, in all other instances where *FP* and *TP* ratings follow normal distributions, the mean rating index, A_1, results in a more powerful statistical test. In the scenarios for skewed distributions with non-zero mass at extremes the statistical power of the test based on θ is greater than that for the other two indices in most instances.

Tables 2.5 and 2.6 demonstrate the estimates of the statistical power for the scenario when the two diagnostic systems differ only with respect to the average number of *FP* marks on a subject. In this scenario the index based on stochastic order of the ratings, A_2, results in higher statistical power for all considered scenarios. For the considered combinations of parameters and a sample size of 20 actually negative and 20 actually positive subjects the estimates of statistical power are quite low. To verify whether the patterns remain the same for larger samples we additionally considered a sample size of 100 actually negative and 100 actually positive subjects, (Table 2.6). They demonstrate the same trend as we observed for 20 actually negative and 20 actually positive subjects.

2.5 SUMMARY

In this Chapter we have investigated several specific indices from a family of proposed indices quantifying subject-based discriminative ability of an FROC system. The indices we proposed were developed in part by applying concepts commonly used in ROC analysis. Specifically, all indices are defined in the format of correctly discriminating in every possible actually negative - actually positive pair. Although in ROC analysis there is a well known relationship of the proportion of correct discriminations and the area under a corresponding ROC curve, our indices permit a more general discrimination for which there may be no corresponding ROC curve. Furthermore, even when there is a corresponding ROC curve (e.g. our procedures based on the maximum or average), when applied to an FROC process, a system which has no ability to discriminate between an individual abnormality and non-abnormality on a location level may not result in the area under the subject-based ROC curve of 0.5.

We proposed a nonparametric method for statistical analysis of this type of indices. The proposed statistical approach can be used to compare two indices and we evaluate our proposed indices in the simulation. The simulation model that we used was based on an approach previously used in FROC analysis [32]. Our simulation model is simplistic since it uses simple parametric distributions, fixed number of lesions within an actually positive subject, and does not describe the correlations that are likely to exist in real FROC data. However, the proposed statistical procedure is based on re-sampling subjects as a unit, and hence it may not be affected by the within subject correlations.

Our simulations demonstrate a reasonable type I error rate for the statistical test based on our indices for sample sizes as small as 20 actually negative and 20 actually positive subjects. Different indices of the considered type, despite their apparent similarity, characterize slightly

different features of the FROC data. In the analysis of the statistical power we demonstrated that even using simple models for FROC data it is possible to construct scenarios where the use of different indices leads to different conclusions. Thus, there is no statistically superior index for comparing subject-based discriminative ability of two arbitrarily different FROC systems. The choice of an index should be based on clinical considerations or by utilizing information from other studies that provide information on how correct localization is related to subject-level discrimination for the particular task being addressed.

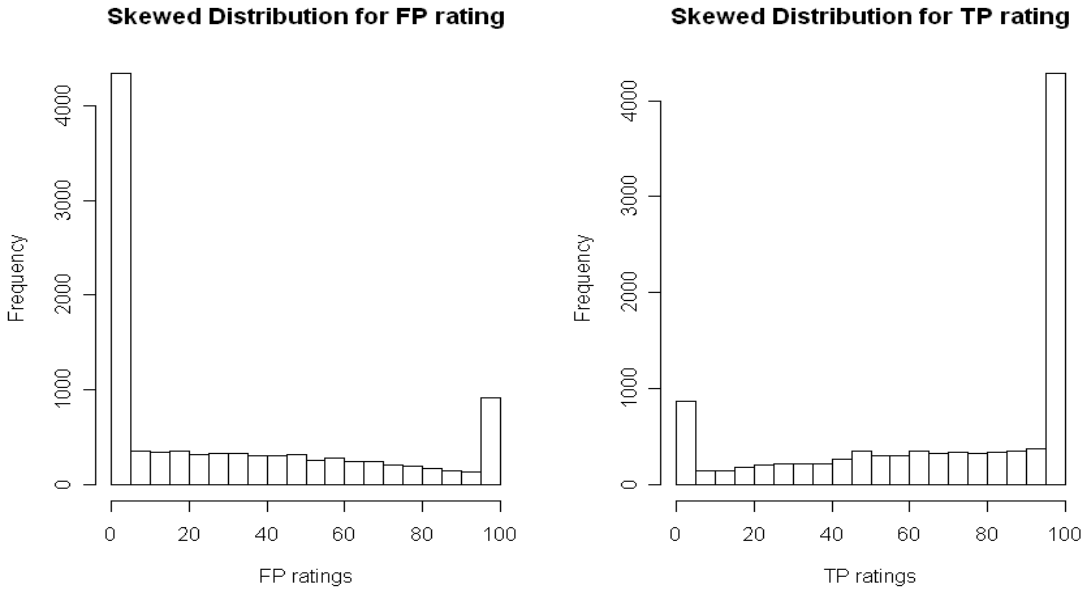

Figure 2.1 Histograms of skewed distributions.

The ratings below the 40th percentile of the distribution of the FP ratings were assigned to the 40th percentile and the ratings above the 60th percentile of the distribution of the TP ratings were assigned the value of the 60th percentile. The skewness for FP ratings is 0.832, and the skewness for TP ratings is -0.832.

Estimated skewness for other scenarios:

1/b	*AUC*	*Skewness for FP ratings*	*Skewness for TP ratings*
1	0.5	0.001	0.000
	0.7	0.555	-0.553
	0.9	1.097	-1.098
2	0.5	0.102	-0.198
	0.7	0.354	-0.951
	0.9	0.893	-1.247

Table 2.1. Estimated expectations and standard errors of the summary indices.

			AUC=0.5								
			$\widehat{E(\hat{\theta})}$	Empirical $se(\hat{\theta})$	2s-jackk. $se(\hat{\theta})$	$\widehat{E(\widehat{A}_1)}$	Empirical $se(\widehat{A}_1)$	2s-jackk. $se(\widehat{A}_1)$	$\widehat{E(\widehat{A}_2)}$	Empirical $se(\widehat{A}_2)$	2s-jackk. $se(\widehat{A}_2)$
1/b	*t*	*v*									
1	1	0.5	0.609	0.0894	0.0900	0.593	0.0909	0.0910	0.593	0.0881	0.0880
		0.9	0.695	0.0842	0.0850	0.666	0.0870	0.0880	0.666	0.0832	0.0840
	3	0.5	0.728	0.0794	0.0810	0.660	0.0874	0.0890	0.661	0.0837	0.0850
		0.9	0.810	0.0707	0.0700	0.683	0.0901	0.0900	0.683	0.0864	0.0860
2	1	0.5	0.621	0.0873	0.0890	0.592	0.0888	0.0900	0.592	0.0863	0.0880
		0.9	0.718	0.0794	0.0810	0.666	0.0854	0.0870	0.666	0.0819	0.0830
	3	0.5	0.764	0.0746	0.0750	0.660	0.0864	0.0870	0.660	0.0827	0.0830
		0.9	0.871	0.0541	0.0540	0.684	0.0853	0.0860	0.684	0.0825	0.0820

			AUC=0.9								
			$\widehat{E(\hat{\theta})}$	Empirical $se(\hat{\theta})$	2s-jackk. $se(\hat{\theta})$	$\widehat{E(\widehat{A}_1)}$	Empirical $se(\widehat{A}_1)$	2s-jackk. $se(\widehat{A}_1)$	$\widehat{E(\widehat{A}_2)}$	Empirical $se(\widehat{A}_2)$	2s-jackk. $se(\widehat{A}_2)$
1/b	*t*	*v*									
1	1	0.5	0.711	0.0811	0.0820	0.691	0.0824	0.0840	0.678	0.0804	0.0820
		0.9	0.883	0.0545	0.0530	0.845	0.0628	0.0620	0.822	0.0613	0.0610
	3	0.5	0.896	0.0524	0.0500	0.859	0.0606	0.0590	0.844	0.0597	0.0580
		0.9	0.981	0.0188	0.0150	0.940	0.0383	0.0350	0.932	0.0376	0.0360
2	1	0.5	0.719	0.0810	0.0810	0.702	0.0826	0.0830	0.678	0.0817	0.0810
		0.9	0.894	0.0519	0.0500	0.864	0.0578	0.0570	0.821	0.0608	0.0600
	3	0.5	0.908	0.0492	0.0470	0.880	0.0548	0.0540	0.845	0.0578	0.0570
		0.9	0.992	0.0110	0.0080	0.964	0.0262	0.0240	0.936	0.0336	0.0320

Estimates are obtained under a simulation model based on normal distributions for λ=1 with a sample size of 20 subjects and 10,000 simulations.

Table 2.2. Expected frequencies of subjects without any marks.

		λ=0.5		λ=1			λ=2	
		actually positive subject	actually negative subject	actually positive subject	actually negative subject	$E(\widehat{A}_1)$ and $E(\widehat{A}_2)$ under *AUC*=0.5, *b*=1	actually positive subject	actually negative subject
t	*v*							
1	0.5	0.30	0.61	0.18	0.37	0.595	0.07	0.14
	0.9	0.06	0.61	0.04	0.37	0.665	0.01	0.14
3	0.5	0.08	0.61	0.05	0.37	0.660	0.02	0.14
	0.9	0.00	0.61	0.00	0.37	0.685	0.00	0.14

Expected frequencies are calculated based on binomial and Poisson distributions. The frequency of no mark on an actually positive subject is

$$P(no\ TP\ mark)\times P(no\ FP\ marks)=P\left(M=0\middle|M\sim Bin\left(t,\upsilon\right)\right)\times P\left(N^t=0\middle|N^t\sim Poisson(\lambda^t)\right)$$

The frequency of no marks on an actually negative subject is

$$P(no\ FP\ marks)=P\left(N^0=0\middle|N^0\sim Poisson(\lambda^0)\right)$$

When the distribution of the FP and TP marks are identically distributed (AUC=0.5, b=1), $E(\widehat{A}_1)$ *and* $E(\widehat{A}_2)$ *can be calculated based on the parametric assumptions using (2.5) and (2.6). For a pair of actually negative and actually positive subjects that both are marked:* $P(mean(\{Y,X^t\})>mean(X^0))=0.5$ *and* $P(w(X^0,\{Y,X^t\})>0.5)=0.5$.

Table 2.3. Estimated type I error rates for normal and skewed distributions.

			AUC=0.5			*AUC*=0.7			*AUC*=0.9		
			λ=1			λ=1			λ=1		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
1/b	*t*	*v*									
1	1	0.5	0.055	0.057	0.055	0.056	0.052	0.053	0.051	0.050	0.051
			(0.056)	(0.058)	(0.058)	(0.057)	(0.053)	(0.054)	(0.051)	(0.051)	(0.050)
		0.9	0.048	0.052	0.053	0.051	0.053	0.052	0.041	0.046	0.050
			(0.048)	(0.053)	(0.052)	(0.052)	(0.055)	(0.055)	(0.042)	(0.047)	(0.048)
	3	0.5	0.055	0.056	0.054	0.053	0.053	0.051	0.042	0.046	0.049
			(0.057)	(0.056)	(0.057)	(0.054)	(0.051)	(0.052)	(0.044)	(0.047)	(0.048)
		0.9	0.046	0.054	0.055	0.037	0.051	0.050	0.002	0.025	0.031
			(0.053)	(0.055)	(0.055)	(0.048)	(0.054)	(0.053)	(0.004)	(0.027)	(0.030)
2	1	0.5	0.054	0.055	0.058	0.054	0.053	0.054	0.053	0.054	0.053
			(0.057)	(0.058)	(0.059)	(0.054)	(0.054)	(0.054)	(0.054)	(0.051)	(0.053)
		0.9	0.049	0.048	0.050	0.046	0.051	0.050	0.038	0.045	0.047
			(0.050)	(0.047	(0.050)	(0.047)	(0.049)	(0.051)	(0.039)	(0.042)	(0.046)
	3	0.5	0.049	0.054	0.053	0.046	0.050	0.049	0.039	0.044	0.047
			(0.051)	(0.056)	(0.056)	(0.048)	(0.052)	(0.050)	(0.039)	(0.045)	(0.046)
		0.9	0.035	0.045	0.046	0.019	0.049	0.050	0.000	0.011	0.027
			(0.049)	(0.049)	(0.050)	(0.031)	(0.052)	(0.050)	(0.000)	(0.009)	(0.025)

			AUC=0.8								
			λ=0.5			λ=1			λ=2		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
1/b	*t*	*v*									
1	1	0.5	0.056	0.055	0.054	0.052	0.053	0.054	0.051	0.055	0.054
			(0.056)	(0.055)	(0.056)	(0.052)	(0.054)	(0.055)	(0.052)	(0.055)	(0.056)
		0.9	0.046	0.046	0.045	0.046	0.049	0.053	0.049	0.046	0.048
			(0.046)	(0.047)	(0.047)	(0.048)	(0.052)	(0.050)	(0.050)	(0.052)	(0.049)
	3	0.5	0.047	0.049	0.051	0.045	0.050	0.048	0.051	0.056	0.057
			(0.049)	(0.052)	(0.051)	(0.048)	(0.051)	(0.048)	(0.054)	(0.056)	(0.056)
		0.9	0.005	0.028	0.033	0.018	0.042	0.045	0.035	0.047	0.047
			(0.012)	(0.033)	(0.032)	(0.036)	(0.045)	(0.048)	(0.051)	(0.048)	(0.049)
2	1	0.5	0.054	0.053	0.053	0.054	0.055	0.056	0.052	0.054	0.055
			(0.061)	(0.061)	(0.061)	(0.051)	(0.051)	(0.051)	(0.052)	(0.055)	(0.054)
		0.9	0.041	0.043	0.046	0.041	0.047	0.048	0.046	0.051	0.052
			(0.040)	(0.043)	(0.045)	(0.043)	(0.045)	(0.047)	(0.047)	(0.052)	(0.052)
	3	0.5	0.041	0.046	0.046	0.047	0.045	0.047	0.047	0.052	0.054
			(0.039)	(0.041)	(0.040)	(0.046)	(0.048)	(0.048)	(0.049)	(0.052)	(0.050)
		0.9	0.001	0.021	0.033	0.003	0.032	0.036	0.014	0.042	0.048
			(0.001)	(0.016)	(0.027)	(0.006)	(0.036)	(0.043)	(0.017)	(0.046)	(0.044)

Estimated type I error rates are obtained when testing the equality of the indices under a simulation model based on normal and skewed distributions (in parenthesis) with a sample size of 20 subjects and 10,000 simulations.

Table 2.4. Estimated power for detecting system difference with regard to *AUC*.

under normal distributions

			AUC=0.7 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.7		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
1/b	*t*	*v*									
1	1	0.5	0.071	0.072	0.071	0.140	0.133	0.120	0.083	0.075	0.069
		0.9	0.115	0.114	0.109	0.460	0.374	0.315	0.200	0.148	0.121
	3	0.5	0.124	0.139	0.132	0.416	0.450	0.420	0.162	0.168	0.156
		0.9	0.185	0.217	0.212	0.698	0.762	0.762	0.195	0.308	0.324
2	1	0.5	0.068	0.076	0.072	0.135	0.157	0.121	0.078	0.080	0.071
		0.9	0.113	0.132	0.109	0.456	0.485	0.334	0.183	0.184	0.118
	3	0.5	0.124	0.179	0.145	0.353	0.553	0.426	0.128	0.198	0.158
		0.9	0.188	0.320	0.249	0.631	0.922	0.843	0.088	0.433	0.370

under skewed distributions

			AUC=0.7 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.7		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
1/b	*t*	*v*									
1	1	0.5	0.071	0.072	0.072	0.150	0.143	0.122	0.091	0.081	0.074
		0.9	0.121	0.117	0.113	0.506	0.410	0.339	0.219	0.172	0.134
	3	0.5	0.136	0.138	0.130	0.490	0.464	0.420	0.191	0.184	0.160
		0.9	0.227	0.187	0.192	0.818	0.720	0.726	0.305	0.305	0.312
2	1	0.5	0.076	0.075	0.074	0.156	0.158	0.125	0.081	0.082	0.071
		0.9	0.145	0.131	0.113	0.538	0.496	0.343	0.194	0.192	0.123
	3	0.5	0.188	0.170	0.147	0.489	0.530	0.406	0.144	0.195	0.144
		0.9	0.398	0.268	0.230	0.890	0.858	0.771	0.140	0.356	0.310

Estimated statistical powers are obtained when detecting system difference with regard to the degree of separation between FP and TP ratings (AUC) for λ=1 under a simulation model based on normal and skewed distribution with a sample size of 20 subjects and 10,000 simulations.

Table 2.5. Estimated power for detecting system difference with regard to λ.

under normal distributions

			λ=1 versus λ=0.5			λ=2 versus λ=0.5			λ=2 versus λ=1		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
1/b	*t*	*v*									
1	1	0.5	0.056	0.063	0.069	0.081	0.112	0.127	0.061	0.069	0.074
		0.9	0.081	0.109	0.130	0.222	0.334	0.415	0.095	0.129	0.146
	3	0.5	0.067	0.092	0.103	0.152	0.248	0.291	0.081	0.098	0.112
		0.9	0.028	0.093	0.113	0.147	0.264	0.309	0.065	0.104	0.112
2	1	0.5	0.055	0.064	0.069	0.071	0.097	0.132	0.056	0.068	0.077
		0.9	0.069	0.093	0.133	0.131	0.264	0.414	0.066	0.103	0.144
	3	0.5	0.055	0.079	0.107	0.085	0.188	0.291	0.054	0.084	0.112
		0.9	0.004	0.065	0.111	0.020	0.203	0.349	0.014	0.079	0.116

under skewed distributions

			λ=1 versus λ=0.5			λ=2 versus λ=0.5			λ=2 versus λ=1		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
1/b	*t*	*v*									
1	1	0.5	0.059	0.062	0.067	0.084	0.106	0.118	0.062	0.068	0.073
		0.9	0.086	0.103	0.116	0.237	0.308	0.372	0.098	0.118	0.136
	3	0.5	0.073	0.088	0.098	0.177	0.228	0.268	0.093	0.098	0.107
		0.9	0.053	0.093	0.107	0.237	0.236	0.291	0.104	0.093	0.109
2	1	0.5	0.057	0.063	0.068	0.073	0.093	0.123	0.057	0.064	0.073
		0.9	0.069	0.092	0.122	0.135	0.240	0.365	0.069	0.100	0.132
	3	0.5	0.056	0.077	0.096	0.090	0.172	0.259	0.056	0.081	0.103
		0.9	0.005	0.062	0.102	0.027	0.182	0.322	0.021	0.077	0.107

Estimated statistical powers are obtained when detecting system difference with regard to the average numbers of FP marks (λ) for AUC=0.8 under a simulation model based on normal and skewed distribution with a sample size of 20 subjects and 10,000 simulations.

Table 2.6. Estimated type I error rates and power for detecting system difference system difference with regard to λ (n=100).

Type I error rate

			λ=0.5			λ=1			λ=2		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
1/b	*t*	*v*									
1	1	0.5	0.049	0.048	0.046	0.066	0.066	0.065	0.056	0.053	0.056
		0.9	0.056	0.054	0.056	0.046	0.044	0.037	0.057	0.057	0.052
	3	0.5	0.049	0.055	0.061	0.062	0.051	0.052	0.048	0.039	0.038
		0.9	0.029	0.051	0.052	0.045	0.040	0.034	0.052	0.056	0.051
2	1	0.5	0.060	0.053	0.054	0.061	0.055	0.056	0.055	0.046	0.051
		0.9	0.048	0.053	0.046	0.058	0.062	0.066	0.049	0.053	0.054
	3	0.5	0.058	0.040	0.044	0.057	0.053	0.049	0.054	0.064	0.055
		0.9	0.039	0.045	0.049	0.055	0.065	0.063	0.051	0.047	0.041

Statistical power

			λ=1 versus λ=0.5			λ=2 versus λ=0.5			λ=2 versus λ=1		
			θ	A_1	A_2	θ	A_1	A_2	θ	A_1	A_2
1/b	*t*	*v*									
1	1	0.5	0.090	0.114	0.132	0.207	0.376	0.431	0.101	0.136	0.153
		0.9	0.268	0.411	0.512	0.805	0.945	0.973	0.312	0.457	0.547
	3	0.5	0.160	0.255	0.314	0.582	0.818	0.878	0.234	0.348	0.378
		0.9	0.234	0.333	0.401	0.783	0.859	0.914	0.312	0.310	0.368
2	1	0.5	0.071	0.087	0.129	0.125	0.302	0.462	0.076	0.121	0.162
		0.9	0.178	0.315	0.488	0.569	0.880	0.976	0.166	0.367	0.541
	3	0.5	0.093	0.219	0.345	0.295	0.725	0.916	0.132	0.282	0.407
		0.9	0.128	0.328	0.486	0.462	0.811	0.961	0.141	0.265	0.428

Estimated statistical powers are obtained when detecting system difference with regard to the average numbers of FP marks (λ) for AUC=0.8 under a simulation model based on normal distributions with a sample size of 100 subjects and 1,000 simulations.

3.0 FROC-TYPE INDICES INVOLVING LOCATION

3.1 BACKGROUND

The indices in Chapter 2 ignored the correct location information of the marks within a subject and used all marked ratings to compare a pair of actually negative and actually positive subjects. In this case, an actually positive subject may be identified as abnormal because of the existence of one or more higher *FP* ratings on the actually positive subject. In addition, there are many cases in the clinical practice of radiology [20-23], that it is not only important to identify an actually positive subject, but also important to offer further guidance regarding the specific location of one or more abnormalities on the subject.

Chakraborty and Berbaum [36] suggested treating the *FP* marks for an actually positive subject the same way as the *FP* marks for an actually negative subject. For simplicity, we denote the *FP* marks for each subject as an *FP* population. For an actually positive subject the authors suggested that non-marked lesions should be assigned to the lowest rating by default and be treated the same way as the *TP* marks. Non-marked lesions and *TP* marks were then termed as lesion ratings. Here we denote the *TP* marks and non-marked lesions on an actually positive subject as an *LR* (Lesion Rating) population. The diagnostic performance with correct location information was evaluated by comparing the *FP* population to the *LR* population. Specifically, to compare two populations with multiple ratings, Chakraborty and Berbaum [36] used the

maximum *FP* rating to summarize the *FP* population and defined an index *JAFROC1* as the weighted average probability that a lesion rating on the actually positive subject exceeds the highest *FP* rating on each subject. As discussed in Chapter 1, they found that the statistical test based on *JAFROC1* (with equal weights for each lesion) may not have a valid type I error rate in a multi-reader simulation study. They then defined an index *JAFROC2*, which ignored the *FP* marks (population) on actually positive subjects. In the same simulation study, the statistical test based on *JAFROC2* was found to have a reasonable type I error rate and have greater power than the statistical test based on θ (index studied in Chapter 2, which uses maximum rating to summarize subject-based discrimination) in detecting the system difference with regard to the degree of separation between *FP* and *TP* ratings (parameter *AUC*).

In this Chapter, we propose new indices to address the two following issues, which might improve *JAFROC* indices in terms of power to detect the system difference with regard to *AUC*:

1. We consider different methods of handling the *FP* marks (*FP* populations) in a pair of actually negative and actually positive subjects. *JAFROC1* suggests treating the *FP* population on each subject as an individual unit. It can be viewed as "splitting" the actually positive subjects into an *FP* population and an *LR* population. We term this method as SPLIT. *JAFROC2* ignores the *FP* population on the actually positive subject and similarly we use such an approach in constituting our IGNORE method. There are potentially other methods of combining the two *FP* populations. We propose as a third approach by switching the *FP* population for the actually positive subject to the actually negative subject (the SWITCH method).

2. When comparing the *FP* population to the *LR* population, *JAFROC1* and *JAFROC2* use the maximum *FP* rating to compare all lesion ratings. As presented in Chapter 2, we compared actually negative subjects to actually positive subjects using three considered comparison

functions, $\tilde{\psi}$, defined by three combination functions (maximum, mean and the Wilcoxon statistic). Each comparison function has been shown to have improved statistical power in certain simulated scenarios. Note: *JAFROC1* and *JAFROC2* do not use the same comparison function defined above. We hereby propose to apply all three comparison functions to compare the *FP* and *LR* populations.

Using the three handling methods (SPLIT, IGNORE, SWITCH) and the three comparison functions ($\tilde{\psi}$ based on max, mean and the Wilcoxon statistic), we have nine indices to evaluate in this Chapter. We compare them to the modified *JAFROC* indices and consider equal weights for each lesion. For statistical inference, we propose to use a two-sample jackknife approach and construct an asymptotic test for each of the indices. We then evaluate and compare the properties of different indices in a simulation study, such as the type I error rates and the statistical power to detect the system difference with regard to *AUC*.

In addition, the clustered ROC approach proposed by Obuchowski [11] can be applied to the FROC paradigm. Each subject can be treated as an independent cluster and the number of marks on a subject is considered to be a nuisance parameter. The index, θ_c, defined in [11] estimates the probability that an *LR* rating exceeds an *FP* rating. Following Obuchowski [11], we construct the variance formula that accounts for the possible correlation within a subject and develop a valid asymptotic test for θ_c. The index θ_c will also be evaluated in the simulation.

3.2 METHODS

Using the notation from Chapter 2, the *FP* population for each subject is described by a random vector $\vec{X}$ of length *N* and the *TP* population for an actually positive subject is described by a random vector $\vec{Y}$ of length *M*. The number of non-marked lesions are therefore *t*-*M*. We assign

the non-marked lesions the lowest rating by default (*L*) and denote it as $\overrightarrow{L_{t-M}}$. An *LR* population is thus equivalent to $\{\vec{Y}, \overrightarrow{L_{t-M}}\}$. Throughout this Chapter, we use $\vec{X}$ to denote the *FP* population on each subject; use $\vec{X}^0$ to denote the *FP* population on an actually negative subject; and use $\vec{X}^t$ to denote the *FP* population on an actually positive subject. We consider different indices that estimate the probability of a correct discrimination between the *FP* population and the *LR* population in an FROC study, for all of which the following formulation applies:

$$A = E\left[\tilde{\psi}\left(\left\{\vec{X}\right\}, \{\vec{Y}, \overrightarrow{L_{t-M}}\}\right)\right]$$

JAFROC1 [36] can be viewed as "splitting" an actually positive subject into an *FP* population and an *LR* population. We defined a similar index *J1* (modification of *JAFROC*1 [36]), which estimates the average probability that a lesion rating on the actually positive subject exceeds the highest *FP* rating on each subject.

$$J1 = E\left[w\left(\max\{\vec{X}\}, \{\vec{Y}, \overrightarrow{L_{t-M}}\}\right) \times I\left(\left\{\vec{X}\right\} \neq \varnothing, \left\{\vec{Y}\right\} \neq \varnothing\right)\right] + \left[P\left(\vec{X} = \varnothing\right) - \frac{1}{2}P\left(\vec{X} = \varnothing, \vec{Y} = \varnothing\right)\right] \quad (3.1)$$

The *w* function is the Wilcoxon statistic defined in (2.6). Note the second term of (3.1) is a natural modification of the second terms in (2.4-2.6). It is the expectation of the scenarios in which there are no *FP* or *TP* marks on an *FP* or an *LR* population. For the comparison of the *FP* population without any *FP* marks and the *LR* population with at least one *TP* mark, we assign 1; for the comparison of the *FP* population with at least one *FP* mark and the *LR* population without any *TP* marks we assign 0; and for the comparison of the *FP* population without any *FP* marks and the *LR* population without any *TP* marks we assign 0.5. *J1* is slightly different from *JAFROC1*. For the comparison of the *FP* population without any *FP* marks $(\vec{X} = \varnothing)$ and the *LR* population with at least one *TP* mark $\left(\{\vec{Y}, \overrightarrow{L_{t-M}}\} \neq \varnothing\right)$, *JAFROC1* [36] considered assigning the

lowest rating L to $\vec{X}$ and had a value of $\frac{M+0.5\times(t-M)}{t}$ which is smaller than or equal to 1 as we defined in (3.1).

The first group of indices applies the SPLIT method with the three comparison functions as we presented in Chapter 2. Following the notation in Chapter 2, we use subscript 0 to denote an index formulated by using the maximum; use subscript 1 to denote an index formulated by using the mean and use subscript 2 to denote an index formulated by using the Wilcoxon statistic. The second component of (3.1) is incorporated into all three of the following SPLIT indices:

$$SPLIT_0 = E\left[\psi\left(\max\left\{\vec{X}\right\},\max\{\vec{Y},\overrightarrow{L_{t-M}}\}\right)\times I\left(\left\{\vec{X}\right\}\neq\varnothing,\left\{\vec{Y}\right\}\neq\varnothing\right)\right]+\left[P\left(\vec{X}=\varnothing\right)-\frac{1}{2}P\left(\vec{X}=\varnothing,\vec{Y}=\varnothing\right)\right]$$

$$SPLIT_1 = E\left[\psi\left(\frac{\sum_{c_1=1}^{N}X_{c_1}}{N},\frac{\sum_{c_2=1}^{M}Y_{c_2}+(t-M)L}{t}\right)\times I\left(\left\{\vec{X}\right\}\neq\varnothing,\left\{\vec{Y}\right\}\neq\varnothing\right)\right]+\left[P\left(\vec{X}=\varnothing\right)-\frac{1}{2}P\left(\vec{X}=\varnothing,\vec{Y}=\varnothing\right)\right] \quad (3.2)$$

$$SPLIT_2 = E\left[\psi\left[0.5,w\left(\vec{X},\{\vec{Y},\overrightarrow{L_{t-M}}\}\right)\right]\times I\left(\left\{\vec{X}\right\}\neq\varnothing,\left\{\vec{Y}\right\}\neq\varnothing\right)\right]+\left[P\left(\vec{X}=\varnothing\right)-\frac{1}{2}P\left(\vec{X}=\varnothing,\vec{Y}=\varnothing\right)\right]$$

The IGNORE method was suggested by Chakraborty and Berbaum [36] as they found that *JAFROC1* may not provide a statistical test with a valid type I error rate. As an alternative to the SPLIT method, we hereby formulate *J2* (modification of *JAFROC2* [36]) and the three indices using different comparison functions (the second group of indices):

$$J2 = E\left[w\left(\max\{\vec{X}^0\},\{\vec{Y},\overrightarrow{L_{t-M}}\}\right)\times I\left(\left\{\vec{X}^0\right\}\neq\varnothing,\left\{\vec{Y}\right\}\neq\varnothing\right)\right]+P\left(\vec{X}^0=\varnothing\right)\times\left[1-\frac{1}{2}P\left(\vec{Y}=\varnothing\right)\right]$$

$$IGNORE_0 = E\left[\psi\left(\max\left\{\vec{X}^0\right\},\max\{\vec{Y},\overrightarrow{L_{t-M}}\}\right)\times I\left(\left\{\vec{X}^0\right\}\neq\varnothing,\left\{\vec{Y}\right\}\neq\varnothing\right)\right]+P\left(\vec{X}^0=\varnothing\right)\times\left[1-\frac{1}{2}P\left(\vec{Y}=\varnothing\right)\right]$$

$$IGNORE_1 = E\left[\psi\left(\frac{\sum_{c_1=1}^{N^0} X_{c_1}^0, \sum_{c_2=1}^{M} Y_{c_2} + (t-M)L}{N^0, t}\right) \times I\left(\left\{\vec{X}^0\right\} \neq \varnothing, \left\{\vec{Y}\right\} \neq \varnothing\right)\right] + P\left(\vec{X}^0 = \varnothing\right) \times \left[1 - \frac{1}{2}P\left(\vec{Y} = \varnothing\right)\right] \quad (3.3)$$

$$IGNORE_2 = E\left[\psi\left[0.5, w\left(\vec{X}^0, \{\vec{Y}, \overrightarrow{L_{t-M}}\}\right)\right] \times I\left(\left\{\vec{X}^0\right\} \neq \varnothing, \left\{\vec{Y}\right\} \neq \varnothing\right)\right] + P\left(\vec{X}^0 = \varnothing\right) \times \left[1 - \frac{1}{2}P\left(\vec{Y} = \varnothing\right)\right]$$

The third group of indices applies the SWITCH method. When we have a pair of actually negative and actually positive subjects, we "switch" the *FP* population of the actually positive subject to the actually negative subject.

$$SWITCH_0 = E\left\{\psi\left(\max\left\{\vec{X}^0, \vec{X}^t\right\}, \max\{\vec{Y}, \overrightarrow{L_{t-M}}\}\right) \times I\left(\left\{\vec{X}^0, \vec{X}^t\right\} \neq \varnothing, \left\{\vec{Y}\right\} \neq \varnothing\right)\right\}$$

$$+\left[P\left(\vec{X}^0 = \varnothing, \vec{X}^t = \varnothing\right) - \frac{1}{2}P\left(\vec{X}^0 = \varnothing, \vec{X}^t = \varnothing, \vec{Y} = \varnothing\right)\right]$$

$$SWITCH_1 = E\left\{\psi\left(\frac{\sum_{c_1=1}^{N^0} X_{c_1}^0 + \sum_{c_2=1}^{N^t} X_{c_2}^t, \sum_{c_3=1}^{M} Y_{c_3} + (t-M)L}{N^0 + N^t, t}\right) \times I\left(\left\{\vec{X}^0, \vec{X}^t\right\} \neq \varnothing, \left\{\vec{Y}\right\} \neq \varnothing\right)\right\}$$

$$+\left[P\left(\vec{X}^0 = \varnothing, \vec{X}^t = \varnothing\right) - \frac{1}{2}P\left(\vec{X}^0 = \varnothing, \vec{X}^t = \varnothing, \vec{Y} = \varnothing\right)\right] \quad (3.4)$$

$$SWITCH_2 = E\left\{\psi\left[0.5, w\left(\left\{\vec{X}^0, \vec{X}^t\right\}, \{\vec{Y}, \overrightarrow{L_{t-M}}\}\right)\right] \times I\left(\left\{\vec{X}^0, \vec{X}^t\right\} \neq \varnothing, \left\{\vec{Y}\right\} \neq \varnothing\right)\right\}$$

$$+\left[P\left(\vec{X}^0 = \varnothing, \vec{X}^t = \varnothing\right) - \frac{1}{2}P\left(\vec{X}^0 = \varnothing, \vec{X}^t = \varnothing, \vec{Y} = \varnothing\right)\right]$$

Note $\left\{\vec{X}^0, \vec{X}^t\right\}$ is different from $\vec{X}$. We use $\left\{\vec{X}^0, \vec{X}^t\right\}$ to denote the combined *FP* population in a pair of actually negative and actually positive subjects and use $\vec{X}$ to denote the *FP* population on each subject.

3.3 STATISTICAL INFERENCE

In contrast to (2.1) where s' indexes for an actually negative subject and s indexes for an actually positive subject, we use s_1 to index all subjects. The nonparametric estimators of the indices considered in section 3.2 can be written as follows.

Apply FROC data in (2.1) to estimate *J1* and SPLIT indices:

$$\widehat{J1} = \frac{1}{(S_0 + S_t) \times S_t} \sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} \tilde{\psi}_{J1} \left(\{x_{s_1 c_1}\}_{c_1=1}^{n_{s1}}, \left\{ \{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}} \right\} \right) \text{ or}$$

$$\widehat{J1} = \frac{1}{(S_0 + S_t) \times S_t} \sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} \frac{\sum_{c_2=1}^{m_s} \psi \left(\max(\{x_{s_1 c_1}\}_{c_1=1}^{n_{s1}}), y_{sc_2} \right) + (t - m_s) \times 0}{t} \times I\left(n_{s_1} m_s \neq 0\right)$$

$$+ \frac{\sum_{s_1=1}^{S_0+S_t} I\left(n_{s_1} = 0\right)}{S_0 + S_t} - \frac{1}{2} \times \left[\frac{\sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} I\left(n_{s_1} + m_s = 0\right)}{(S_0 + S_t) \times S_t} \right] \quad (3.5)$$

$$\widehat{SPLIT}_0 = \frac{1}{(S_0 + S_t) \times S_t} \sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} \tilde{\psi}_{\max} \left(\{x_{s_1 c_1}\}_{c_1=1}^{n_{s1}}, \left\{ \{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}} \right\} \right) \text{ or}$$

$$\widehat{SPLIT}_0 = \frac{1}{(S_0 + S_t) \times S_t} \sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} \psi \left\{ \max \left\{ x_{s_1 c_1} \right\}_{c_1=1}^{n_{s1}}, \max \left\{ y_{sc_2} \right\}_{c_2=1}^{m_s} \right\} \times I\left(n_{s_1} m_s \neq 0\right)$$

$$+ \frac{\sum_{s_1=1}^{S_0+S_t} I\left(n_{s_1} = 0\right)}{S_0 + S_t} - \frac{1}{2} \times \left[\frac{\sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} I\left(n_{s_1} + m_s = 0\right)}{(S_0 + S_t) \times S_t} \right] \quad (3.6)$$

$$\widehat{SPLIT}_1 = \frac{1}{(S_0 + S_t) \times S_t} \sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} \tilde{\psi}_{mean} \left(\{x_{s_1 c_1}\}_{c_1=1}^{n_{s1}}, \left\{ \{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}} \right\} \right) \text{ or}$$

$$\widehat{SPLIT}_1 = \frac{1}{(S_0 + S_t) \times S_t} \sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} \psi \left\{ \frac{\sum_{c_1=1}^{n_{s1}} x_{s_1 c_1}}{n_{s_1}}, \frac{\sum_{c_2=1}^{m_s} y_{sc_2} + (t - m_s) \times L}{t} \right\} \times I\left(n_{s_1} m_s \neq 0\right)$$

$$+\frac{\sum_{s_1=1}^{S_0+S_t} I\left(n_{s_1}=0\right)}{S_0+S_t}-\frac{1}{2}\times\left[\frac{\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t} I\left(n_{s_1}+m_s=0\right)}{\left(S_0+S_t\right)\times S_t}\right] \quad (3.7)$$

$$\widehat{SPLIT}_2=\frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\tilde{\psi}_{wilcoxon}\left(\{x_{s_1c_1}\}_{c_1=1}^{n_{s_1}},\left\{\{y_{sc_2}\}_{c_2=1}^{m_s},\overrightarrow{L_{t-m_s}}\right\}\right) \text{ or}$$

$$\widehat{SPLIT}_2=\frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\psi\left\{0.5,\frac{\sum_{c_1=1}^{n_{s_1}}\sum_{c_2=1}^{m_s}\psi\left(x_{s_1c_1},y_{sc_2}\right)+n_{s_1}\times(t-m_s)\times 0}{n_{s_1}\times t}\right\}\times I\left(n_{s_1}m_s\neq 0\right)$$

$$+\frac{\sum_{s_1=1}^{S_0+S_t} I\left(n_{s_1}=0\right)}{S_0+S_t}-\frac{1}{2}\times\left[\frac{\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t} I\left(n_{s_1}+m_s=0\right)}{\left(S_0+S_t\right)\times S_t}\right] \quad (3.8)$$

Apply FROC data (2.1) to estimate *J2* and IGNORE indices:

$$\widehat{J2}=\frac{1}{S_0\times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{J2}\left(\{x^0_{s'c_1}\}_{c_1=1}^{n^0_{s'}},\left\{\{y_{sc_2}\}_{c_2=1}^{m_s},\overrightarrow{L_{t-m_s}}\right\}\right) \text{ or}$$

$$\widehat{J2}=\frac{1}{S_0\times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\frac{\sum_{c_2=1}^{m_s}\psi\left(\max(\{x^0_{s'c_1}\}_{c_1=1}^{n^0_{s'}}),y_{sc_2}\right)+(t-m_s)\times 0}{t}\times I\left(n^0_{s'}m_s\neq 0\right)$$

$$+\left[\frac{\sum_{s'=1}^{S_0} I\left(n^0_{s'}=0\right)}{S_0}\right]\times\left[1-\frac{1}{2}\times\frac{\sum_{s=1}^{S_t} I\left(m_s=0\right)}{S_t}\right] \quad (3.9)$$

$$\widehat{IGNORE}_0=\frac{1}{S_0\times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{\max}\left(\{x^0_{s'c_1}\}_{c_1=1}^{n^0_{s'}},\left\{\{y_{sc_2}\}_{c_2=1}^{m_s},\overrightarrow{L_{t-m_s}}\right\}\right) \text{ or}$$

$$\widehat{IGNORE}_0=\frac{1}{S_0\times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\psi\left\{\max\left\{x^0_{s'c_1}\right\}_{c_1=1}^{n^0_{s'}},\max\left\{y_{sc_2}\right\}_{c_2=1}^{m_s}\right\}\times I\left(n^0_{s'}m_s\neq 0\right)$$

$$+\left[\frac{\sum_{s'=1}^{S_0} I\left(n^0_{s'}=0\right)}{S_0}\right]\times\left[1-\frac{1}{2}\times\frac{\sum_{s=1}^{S_t} I\left(m_s=0\right)}{S_t}\right] \quad (3.10)$$

$$\widehat{IGNORE}_1 = \frac{1}{S_0 \times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{mean}\left(\{x^0_{s'c_1}\}_{c_1=1}^{n^0_{s'}}, \left\{\{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}}\right\}\right) \text{ or}$$

$$\widehat{IGNORE}_1 = \frac{1}{S_0 \times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\psi\left\{\frac{\sum_{c_1=1}^{n^0_{s'}} x^0_{s'c_1}}{n^0_{s'}}, \frac{\sum_{c_2=1}^{m_s} y_{sc_2} + (t-m_s)\times L}{t}\right\} \times I\left(n^0_{s'} m_s \neq 0\right)$$

$$+\left[\frac{\sum_{s'=1}^{S_0} I\left(n^0_{s'}=0\right)}{S_0}\right]\times\left[1-\frac{1}{2}\times\frac{\sum_{s=1}^{S_t} I\left(m_s=0\right)}{S_t}\right] \quad (3.11)$$

$$\widehat{IGNORE}_2 = \frac{1}{S_0 \times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{wilcoxon}\left(\{x^0_{s'c_1}\}_{c_1=1}^{n^0_{s'}}, \left\{\{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}}\right\}\right) \text{ or}$$

$$\widehat{IGNORE}_2 = \frac{1}{S_0 \times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\psi\left\{0.5, \frac{\sum_{c_1=1}^{n^0_{s'}}\sum_{c_2=1}^{m_s}\psi\left(x^0_{s'c_1}, y_{sc_2}\right) + n^0_{s'}\times(t-m_s)\times 0}{n^0_{s'}\times t}\right\} \times I\left(n^0_{s'} m_s \neq 0\right)$$

$$+\left[\frac{\sum_{s'=1}^{S_0} I\left(n^0_{s'}=0\right)}{S_0}\right]\times\left[1-\frac{1}{2}\times\frac{\sum_{s=1}^{S_t} I\left(m_s=0\right)}{S_t}\right] \quad (3.12)$$

Apply FROC data (2.1) to estimate SWITCH indices:

$$\widehat{SWITCH}_0 = \frac{1}{S_0 \times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{\max}\left(\left\{\left\{x^0_{s'c_1}\right\}_{c_1=1}^{n^0_{s'}}, \left\{x^t_{sc_2}\right\}_{c_2=1}^{n^t_s}\right\}, \left\{\{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}}\right\}\right) \text{ or}$$

$$\widehat{SWITCH}_0 = \frac{1}{S_0 \times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\psi\left\{\max\left\{\left\{x^0_{s'c_1}\right\}_{c_1=1}^{n^0_{s1}}, \left\{x^t_{sc_2}\right\}_{c_2=1}^{n^t_s}\right\}, \max\left\{y_{sc_3}\right\}_{c_3=1}^{m_s}\right\} \times I\left((n^0_{s'}+n^t_s)m_s \neq 0\right)$$

$$+\frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t} I\left(n^0_{s'}+n^t_s=0\right) - \frac{1}{2}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t} I\left(n^0_{s'}+n^t_s+m_s=0\right)}{S_0 \times S_t} \quad (3.13)$$

$$\widehat{SWITCH}_1 = \frac{1}{S_0 \times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{mean}\left(\left\{\left\{x^0_{s'c_1}\right\}_{c_1=1}^{n^0_{s'}}, \left\{x^t_{sc_2}\right\}_{c_2=1}^{n^t_s}\right\}, \left\{\{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}}\right\}\right) \text{ or}$$

$$\widehat{SWITCH}_1 = \frac{1}{S_0 \times S_t} \sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \psi \left\{ \frac{\sum_{c_1=1}^{n^0_{s'}} x^0_{s'c_1} + \sum_{c_2=1}^{n^t_s} x^t_{sc_2}}{n^0_{s'} + n^t_s}, \frac{\sum_{c_3=1}^{m_s} y_{sc_3} + (t - m_s) \times L}{t} \right\} \times I\left((n^0_{s'} + n^t_s) m_s \neq 0\right)$$

$$+ \frac{\sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} I\left(n^0_{s'} + n^t_s = 0\right) - \frac{1}{2} \sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} I\left(n^0_{s'} + n^t_s + m_s = 0\right)}{S_0 \times S_t} \quad (3.14)$$

$$\widehat{SWITCH}_2 = \frac{1}{S_0 \times S_t} \sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \tilde{\psi}_{wilcoxon} \left(\left\{ \left\{ x^0_{s'c_1} \right\}_{c_1=1}^{n^0_{s'}}, \left\{ x^t_{sc_2} \right\}_{c_2=1}^{n^t_s} \right\}, \left\{ \left\{ y_{sc_2} \right\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}} \right\} \right) \text{ or}$$

$$\widehat{SWITCH}_2 = \frac{1}{S_0 \times S_t} \sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \psi \left\{ 0.5, \frac{\sum_{c_1=1}^{n^0_{s'}} \sum_{c_3=1}^{m_s} \psi(x^0_{s'c_1}, y_{sc_3}) + \sum_{c_2=1}^{n^t_s} \sum_{c_3=1}^{m_s} \psi(x^t_{sc_2}, y_{sc_3}) + (n^0_{s'} + n^t_s) \times (t - m_s) \times 0}{(n^0_{s'} + n^t_s) \times t} \right\}$$

$$\times I\left((n^0_{s'} + n^t_s) m_s \neq 0\right) + \frac{\sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} I\left(n^0_{s'} + n^t_s = 0\right) - \frac{1}{2} \sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} I\left(n^0_{s'} + n^t_s + m_s = 0\right)}{S_0 \times S_t} \quad (3.15)$$

To illustrate the estimation of all the indices in (3.5)-(3.15), we provide an example using 4 actually negative and 4 actually positive subjects in Appendix.

For the generalized U-statistics in (3.5)-(3.15) we can construct the closed form two-sample jackknife variance when the re-sampling techniques consider subject as the sampling unit. When we do two-sample jackknifing for the indices in Chapter 2, removing an actually positive subject only influences a column of the $\{\tilde{\psi}_{s's}\}$ matrix that represents the actually positive subject. Similarly we can construct the two-sample variance for indices using the IGNORE and SWITCH methods (3.9)-(3.15) using the formula (2.10) presented in Chapter 2:

$$\hat{V}_{\text{2s-jackk}}(\hat{A}) = \frac{\sum_{s'=1}^{S_0} \left(\overline{\tilde{\psi}_{s'\bullet}} - \overline{\tilde{\psi}_{\bullet\bullet}} \right)^2}{S_0 \times (S_0 - 1)} + \frac{\sum_{s=1}^{S_t} \left(\overline{\tilde{\psi}_{\bullet s}} - \overline{\tilde{\psi}_{\bullet\bullet}} \right)^2}{S_t \times (S_t - 1)}$$

For *J1* (3.5) and the three indices using the SPLIT method (3.6)-(3.8), the two-sample jackknife approach is formally different than (2.10). This is because removing an actually positive subject influences both a row and a column of the $\{\tilde{\psi}_{s_1 s}\}$ matrix that represents the actually positive subject. To develop the two-sample jackknife variance formula for SPLIT indices, we denote the $\tilde{\psi}$ functions for comparing actually negative and actually positive subjects as $\{\tilde{\psi}^0_{s_1 s}\}$, where $s_1 = 1,...,S_0; s = 1,...,S_t$ and for comparing actually positive and actually positive subjects as $\{\tilde{\psi}^t_{s_1 s}\}$ where $s_1 = S_0 + 1,...,S_0 + S_t; s = 1,...,S_t$. $\tilde{\psi}_{s_1 \bullet}$ denotes the summation of the s_1 th row (*FP* population). $\tilde{\psi}_{\bullet s}$ denotes the summation of the s th column (*LR* population). To estimate the index, we use $\hat{A} = \frac{\tilde{\psi}^0_{\bullet\bullet} + \tilde{\psi}^t_{\bullet\bullet}}{(S_0 + S_t)S_t}$. Now we construct the two-sample jackknife variance for the SPLIT indices:

1. Find the estimates after removing an actually negative subject:

$$\hat{A}^{s_1}_{S_0+S_t-1,S_t} = \frac{\tilde{\psi}^0_{\bullet\bullet} - \tilde{\psi}^0_{s_1\bullet} + \tilde{\psi}^t_{\bullet\bullet}}{(S_0 - 1)S_t + S_t \times S_t}$$

The pseudo-values are:

$$\hat{A}^{s_1}_{negative;J2} = S_0 \times \widehat{A} - (S_0 - 1) \times \hat{A}^{s_1}_{S_0+S_t-1,S_t} = S_0 \times \frac{\tilde{\psi}^0_{\bullet\bullet} + \tilde{\psi}^t_{\bullet\bullet}}{(S_0 + S_t)S_t} - (S_0 - 1) \times \frac{\tilde{\psi}^0_{\bullet\bullet} - \tilde{\psi}^0_{s_1\bullet} + \tilde{\psi}^t_{\bullet\bullet}}{(S_0 - 1)S_t + S_t \times S_t}$$
$$= \frac{\tilde{\psi}^0_{\bullet\bullet} + \tilde{\psi}^t_{\bullet\bullet}}{(S_0 + S_t)(S_0 + S_t - 1)} + \frac{(S_0 - 1)\tilde{\psi}^0_{s_1\bullet}}{(S_0 + S_t - 1)S_t}.$$

The average pseudo-values due to actually negative subjects are

$$\hat{A}^1 = \frac{\sum_{s_1=1}^{S_0} \hat{A}^{s_1}_{negative;J2}}{S_0} = \frac{\tilde{\psi}^0_{\bullet\bullet} + \tilde{\psi}^t_{\bullet\bullet}}{(S_0 + S_t)(S_0 + S_t - 1)} + \frac{(S_0 - 1)\tilde{\psi}^0_{\bullet\bullet}}{(S_0 + S_t - 1)S_0 S_t}$$

The variance component due to actually negative subjects is $\hat{V}_N = \dfrac{\sum_{s_1=1}^{S_0} (\hat{A}^{s_1}_{negative;J2} - \hat{A}^1)^2}{S_0(S_0-1)}$

2. Find the estimates after removing an actually positive subject:

$$\hat{A}^s_{S_0+S_t-1,S_t-1} = \frac{(\tilde{\psi}^0_{\bullet\bullet} - \tilde{\psi}^0_{\bullet s}) + (\tilde{\psi}^t_{\bullet\bullet} - \tilde{\psi}^t_{s\bullet} - \tilde{\psi}^t_{\bullet s} + \tilde{\psi}^t_{ss})}{S_0(S_t-1) + (S_t-1)^2}$$

The pseudo-values are:

$$\hat{A}^s_{positive;J2} = S_t \times \widehat{A} - (S_t-1) \times \hat{A}^s_{S_0+S_t-1,S_t-1} = S_t \times \frac{\tilde{\psi}^0_{\bullet\bullet} + \tilde{\psi}^t_{\bullet\bullet}}{(S_0+S_t)S_t} - (S_t-1) \times \frac{(\tilde{\psi}^0_{\bullet\bullet} - \tilde{\psi}^0_{\bullet s}) + (\tilde{\psi}^t_{\bullet\bullet} - \tilde{\psi}^t_{s\bullet} - \tilde{\psi}^t_{\bullet s} + \tilde{\psi}^t_{ss})}{S_0(S_t-1) + (S_t-1)^2}$$

$$= \frac{-(\tilde{\psi}^0_{\bullet\bullet} + \tilde{\psi}^t_{\bullet\bullet})}{(S_0+S_t)(S_0+S_t-1)} + \frac{\tilde{\psi}^0_{\bullet s} + \tilde{\psi}^t_{s\bullet} + \tilde{\psi}^t_{\bullet s} - \tilde{\psi}^t_{ss}}{S_0+S_t-1}.$$

The average pseudo-values due to actually positive subjects is given by

$$\hat{A}^2 = \frac{\sum_{s=1}^{S_t} \hat{A}^s_{positive;J2}}{S_t} = \frac{-(\tilde{\psi}^0_{\bullet\bullet} + \tilde{\psi}^t_{\bullet\bullet})}{(S_0+S_t)(S_0+S_t-1)} + \frac{\tilde{\psi}^0_{\bullet\bullet} + 2\tilde{\psi}^t_{\bullet\bullet} - \sum_{s=1}^{S_t} \tilde{\psi}^t_{ss}}{(S_0+S_t-1)S_t}$$

and the variance component due to actually positive subjects is given by $\hat{V}_P = \dfrac{\sum_{s=1}^{S_t} (\hat{A}^s_{positive;J2} - \hat{A}^2)^2}{S_t(S_t-1)}$

3. Compute the variance using: $\hat{V}_{\text{2s-jackk}}(\hat{A}_{J2}) = \hat{V}_N + \hat{V}_P$ or

$$\hat{V}_{\text{2s-jackk}}(\hat{A}) = \frac{\sum_{s_1=1}^{S_0} (\hat{A}^{s_1}_{negative;J2} - \hat{A}^1)^2}{S_0(S_0-1)} + \frac{\sum_{s=1}^{S_t} (\hat{A}^s_{positive;J2} - \hat{A}^2)^2}{S_t(S_t-1)} \qquad (3.16)$$

We propose an asymptotic procedure based on (3.16) with the following test statistics for comparing two FROC diagnostic systems: $Z_i = \dfrac{\hat{A}^2_i - \hat{A}^1_i}{\sqrt{\hat{V}_{\text{2s-jackk}}\left(\hat{A}^2_i - \hat{A}^1_i\right)}}$, i=1,…,11.

Now we apply the clustered ROC approach presented by Obuchowski [11] to the FROC paradigm. Denote lesion ratings to be $\{z_{sc}\}_{c=1}^{t} = \left\{\{y_{sc}\}_{c=1}^{m_s}, \overrightarrow{L_{t-m_s}}\right\}$ for an actually positive subject. Using our notation, there are $S_0 + S_t$ total clusters (subjects). The total number of *FP* marks is given by $N' = \sum_{s_1=1}^{S_0+S_t} n_{s_1}$ and the total number of lesions is given by $M' = tS_t$. The total number of subjects with at least one *FP* mark is denoted by I_{01}, and the total number of subjects with at least one lesion rating is denoted by I_{10} ($I_{10} = S_t$).

This approach considers the number of marks on a subject as a nuisance parameter and it estimates the degree of separation between *FP* and *LR* ratings

$$\hat{\theta}_c = \frac{1}{M'N'} \sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} \sum_{c_1=1}^{n_{s_1}} \sum_{c_2=1}^{t} \psi(x_{s_1c_1}, z_{sc_2}) .$$

This estimate is different from our proposed index $\widehat{SPLIT}_2$ of (3.8) in that $\hat{\theta}_c$ ignores the subject effect. Specifically, $\widehat{SPLIT}_2$ considered the comparison between the *FP* and *LR* population without any marks. In addition, $\widehat{SPLIT}_2$ calculated the statistic $\frac{\sum_{c_1=1}^{n_{s_1}} \sum_{c_2=1}^{t} \psi\left(x_{s_1c_1}, z_{sc_2}\right)}{n_{s_1} \times t}$ in a pair of *FP* and *LR* population and compared it to 0.5 before taking a simple average over all possible pairs.

Following Obuschowski [11], the X- and Y- components for clustered data are

$$V_{01}(x_{s_1c_1}) = \frac{1}{M'} \sum_{s=1}^{I_{10}} \sum_{c_2=1}^{t} \psi(x_{s_1c_1}, z_{sc_2}) \text{ and } V_{10}(z_{sc_2}) = \frac{1}{N'} \sum_{s_1=1}^{I_{01}} \sum_{c_1=1}^{n_{s_1}} \psi(x_{s_1c_1}, z_{sc_2}) .$$

Let $V_{01}(x_{s_1\bullet})$ and $V_{10}(z_{s\bullet})$ be the sum of the X- and Z- components respectively for the s_1 th subject and for the s th actually positive subject. When there are no *FP* marks in subject s', $n_{s'}$

and $V_{01}(x_{s_1\bullet})$ equal zero. Similarly, when there are no *TP* marks in actually positive subject s, $V_{10}(z_{s\bullet})$ equal zero. The sum of squares of the X-and Z- components are

$$S_{01} = \frac{I_{01}}{(I_{01}-1)N'}\sum_{s_1=1}^{I_{01}}\left[V_{01}(x_{s_1\bullet}) - n_{s_1}\hat{\theta}_c\right]^2 \text{ and } S_{10} = \frac{I_{10}}{(I_{10}-1)M'}\sum_{s=1}^{I_{10}}\left[V_{10}(z_{s\bullet}) - t\hat{\theta}_c\right]^2 .$$

The cross-product, which takes into consideration the possible correlation between the *LR* and *FP* marks within the same subject, is formulated as

$$S_{11} = \frac{S_0 + S_t}{(S_0 + S_t - 1)}\sum_{s_1=1}^{S_t}\left(\left[V_{01}(x_{s_1\bullet}) - n_{s_1}\hat{\theta}_c\right]\left[V_{10}(z_{s_1\bullet}) - t\hat{\theta}_c\right]\right)$$

The estimator of the variance of $\hat{\theta}_c$ is thus given by $\hat{V}(\hat{\theta}_c) = \frac{1}{N'}S_{01} + \frac{1}{M'}S_{10} + \frac{2}{N'M'}S_{11}$.

Following Obuschowski [11], $\left(\hat{\theta}_c - \theta_c\right)/\sqrt{\hat{V}(\hat{\theta}_c)}$ is asymptotically $N(0,1)$ if $\lim_{S_0+S_t\to\infty}\frac{I_{10}}{I_{01}}$ is bounded and nonzero.

3.4 SIMULATION RESULTS

We evaluated all the indices in section 3.3 under the simulation model presented in Chapter 2 for both normal and skewed distributions with 10,000 simulation runs. In the simulation, the estimated two-sample jackknife variances for all the indices are close to the variances of sample realization in most considered scenarios. In Table 3.1 we list estimated expectations and standard errors for *J1*, $SPLIT_2$ and $SWITCH_2$. The standard errors were estimated both with a sample variance over the simulated realizations and by using the corresponding two-sample jackknife variances using (3.16) and (2.10).

When the distributions of the ratings for *FP* and *TP* marks have the same location (corresponding to *AUC*=0.5), the estimated expectations range from 0.26 to 0.64, with $SWITCH_2$

being the lowest for most of the simulated scenarios. The frequencies of the *LR* and *FP* populations with no marks for three different handling methods are listed in Table 3.2. The phenomenon that the expectation of $SPLIT_2$ and $SWITCH_2$ are different from 0.5 when AUC=0.5 can still be partially attributed to an imbalance in frequencies of no *TP* marks or no *FP* marks on an *LR* or *FP* population. Specifically, when the frequency of no *TP* marks is higher, it shifts the expectation to be smaller than 0.5; and when the frequency of no *FP* marks is higher, it shifts the expectation to be larger than 0.5. From Table 3.2, one can also observe that the frequency of no *FP* marks on an *FP* population for SWITCH indices is smaller than that for *J1* and SPLIT indices. This partially explains the phenomenon that $SWITCH_2$ has a smaller expectation than that of *J1* and $SPLIT_2$.

For both normal and skewed distributions, the estimated type I error rates for the proposed indices are close to the nominal value of 0.05. We list the estimated type I error rates for $SPLIT_2$ and $SWITCH_2$ indices, as well as the index θ_c, using the clustered ROC approach [11], in Table 3.3. Similar to what was observed in Chapter 2, when both the number of abnormalities on the actually positive subjects and the degree of separation between *FP* and *TP* ratings (AUC) is large, the estimated type I error rate is low for $SPLIT_2$ and $SWITCH_2$. However, we can observe that the estimated type I error rate for θ_c, using the cluster ROC approach [11], is close to 0.05 in all considered scenarios.

We compare the estimates of the statistical power for the various scenarios where the two diagnostic systems differ with regard to the degree of separation between *FP* and *TP* ratings (AUC). The simulation results are shown in Table 3.4-3.8 for both normal and skewed distributions. From Table 3.4-3.6, we list the indices using three different handling methods (SPLIT, SWITCH, IGNORE). We highlight the greatest statistical power under each scenario in

the tables. It can be seen that similar to the indices in Chapter 2, for each handling method (SPLIT, SWITCH or IGNORE) that is used, none of these three comparison functions (based on max, mean and the Wilcoxon statistic) perform the best in all simulated scenarios. Each comparison function has a power advantage in different scenarios.

It can be also observed that, compared to the SPLIT method, the IGNORE method has a smaller power at each of the simulated scenarios. *J2* (not shown in table) is also found to have a smaller power than *J1*. Due to the fact that the IGNORE method only uses half of the *FP* populations as compared to the SPLIT method, it may not be surprising to see that it results in such a power loss. The loss of power for the IGNORE method (including *J2*) compared to the SPLIT method (including *J1*) ranges from 5.5% to 52.6%. The average loss of power for all scenarios is 20.3%.

Table 3.7 exhibits the simulation results for *J1* and $SWITCH_2$, as well as those for the clustered ROC index θ_c [11]. For both normal and skewed distributions, the statistical test based on $SWITCH_2$ tends to have greater power than the statistical test based on based on *J1*. It can be also observed that the clustered ROC index θ_c has greatest power in detecting the system difference with regard to *AUC* at each of the simulated scenarios.

In Table 3.8, we use a paired *t* test to test the power equivalence of the statistical tests based on *J1* and all our proposed indices for normal and skewed distributions respectively (each with 36 simulated scenarios). The Bonferroni correction method is used to adjust for multiple comparisons. In the first hypothesis, we test whether the statistical tests based on the SPLIT method are equivalent to the statistical tests based on the IGNORE method using three different comparison functions. The SPLIT method is found to be significantly better than the IGNORE method with each of the three comparison functions. The average power improvement is 0.097

(26.2%). In the second hypothesis, we test whether the statistical tests based on the SWITCH method are equivalent to the statistical tests based on the SPLIT method. The SWITCH method is found to be significantly better than the SPLIT method. The average power improvement is 0.036 (15.6%). For the third hypothesis, we allow the comparison between the statistical test based on *J1* and the statistical test based on each of all our proposed indices. The statistical test based on *J1* shows no significant difference with the statistical test based on other SPLIT indices, except for the maximum function for skewed distributions. It can also be observed that the statistical tests based on all SWITCH indices are significantly better than the statistical test based on *J1*. The average power improvement is 0.084 (25.6%).

3.5 SUMMARY

In this Chapter, we propose new indices that are natural extensions of *JAFROC1* (SPLIT method) and *JAFROC2* (IGNORE method) with different comparison functions. We also propose a new family of indices with a different handling method for *FP* populations and term it as SWITCH method. All indices estimate the probability of correct discrimination between the *FP* and *LR* populations. For comparisons of the discriminative ability of two FROC diagnostic systems, the statistical tests based on all indices seem to have reasonable type I error rates in most simulated scenarios. As might be expected, because the *FP* marks on actually positive subjects are ignored, the statistical tests based on the IGNORE method (including *J2*, modified *JAFROC2*) lose an average of 20.3% statistical power to detect the system difference with regard to the degree of separation between *FP* and *TP* ratings (*AUC*), as compared to the statistical tests based on the SPLIT method (including *J1*, modified *JAFROC1*). From the simulations, there is no significant power advantage of applying the three comparison functions (based on max, mean

and the Wilcoxon statistic) over the comparison function of *JAFROC*-type. The proposed SWITCH indices show some power improvement under certain scenarios than *J1* or SPLIT indices.

Compared to our proposed groups of indices, the clustered ROC index θ_c [11] characterizes the FROC diagnostic performance by estimating the degree of separation between *FP* and *LR* ratings. Despite that it has greatest power advantage to detect the system difference with regard to *AUC*, this approach ignores the subject effect, which is an important feature of the FROC system. In fact, if we furthermore ignore non-marked lesions, the refined clustered index, $\widehat{AUC}$, directly estimates the degree of separation between *FP* and *TP* ratings and the statistical test based on it has greater power in the simulation (not shown).

Table 3.1. Estimated expectations and standard errors of the summary indices.

			AUC=0.5								
			$J1$	Empirical SE	Jackk. SE	$SPLIT_2$	Empirical SE	Jackk. SE	$SWITCH_2$	Empirical SE	Jackk. SE
1/b	*t*	*v*									
1	1	0.5	0.408	0.0742	0.0730	0.435	0.0769	0.0750	0.319	0.0866	0.0850
		0.9	0.587	0.0736	0.0730	0.633	0.0713	0.0710	0.516	0.0913	0.0910
	3	0.5	0.477	0.0664	0.0660	0.453	0.0732	0.0720	0.264	0.0778	0.0780
		0.9	0.605	0.0626	0.0610	0.643	0.0684	0.0670	0.504	0.0907	0.0910
2	1	0.5	0.418	0.0777	0.0770	0.434	0.0793	0.0780	0.318	0.0890	0.0880
		0.9	0.604	0.0769	0.0760	0.633	0.0757	0.0750	0.517	0.0954	0.0960
	3	0.5	0.486	0.0663	0.0650	0.446	0.0757	0.0750	0.262	0.0799	0.0800
		0.9	0.624	0.0594	0.0580	0.639	0.0712	0.0710	0.504	0.0924	0.0940

			AUC=0.9								
			$J1$	Empirical SE	Jackk. SE	$SPLIT_2$	Empirical SE	Jackk. SE	$SWITCH_2$	Empirical SE	Jackk. SE
1/b	*t*	*v*									
1	1	0.5	0.548	0.0875	0.0880	0.561	0.0888	0.0890	0.496	0.1013	0.1020
		0.9	0.842	0.0604	0.0590	0.865	0.0581	0.0560	0.840	0.0716	0.0690
	3	0.5	0.617	0.0620	0.0610	0.615	0.0810	0.0800	0.498	0.0991	0.1000
		0.9	0.860	0.0364	0.0360	0.933	0.0335	0.0320	0.917	0.0488	0.0460
2	1	0.5	0.552	0.0891	0.0890	0.559	0.0898	0.0900	0.491	0.1022	0.1030
		0.9	0.850	0.0617	0.0600	0.863	0.0601	0.0580	0.834	0.0744	0.0730
	3	0.5	0.622	0.0617	0.0610	0.614	0.0820	0.0820	0.497	0.1004	0.1010
		0.9	0.867	0.0344	0.0340	0.937	0.0354	0.0330	0.919	0.0512	0.0470

Estimates are obtained under a simulation model based on normal distributions for $\lambda=1$ with a sample size of 20 subjects and 10,000 simulations.

Table 3.2. Expected frequencies of *LR* and *FP* populations with no marks.

J1 & SPLIT indices

		λ=0.5		λ=1		λ=2	
		LR without *TP* marks	*FP* without *FP* marks	*LR* without *TP* marks	*FP* without *FP* marks	*LR* without *TP* marks	*FP* without *FP* marks
t	*v*						
1	0.5	0.50	0.61	0.50	0.37	0.50	0.14
	0.9	0.10	0.61	0.10	0.37	0.10	0.14
3	0.5	0.13	0.61	0.13	0.37	0.13	0.14
	0.9	0.00	0.61	0.00	0.37	0.00	0.14

J2 & IGNORE indices

		λ=0.5		λ=1		λ=2	
		LR without *TP* marks	*FP* without *FP* marks	*LR* without *TP* marks	*FP* without *FP* marks	*LR* without *TP* marks	*FP* without *FP* marks
t	*v*						
1	0.5	0.50	0.61	0.50	0.37	0.50	0.14
	0.9	0.10	0.61	0.10	0.37	0.10	0.14
3	0.5	0.13	0.61	0.13	0.37	0.13	0.14
	0.9	0.00	0.61	0.00	0.37	0.00	0.14

SWITCH indices

		λ=0.5		λ=1		λ=2	
		LR without *TP* marks	*FP* without *FP* marks	*LR* without *TP* marks	*FP* without *FP* marks	*LR* without *TP* marks	*FP* without *FP* marks
t	*v*						
1	0.5	0.50	0.37	0.50	0.14	0.50	0.02
	0.9	0.10	0.37	0.10	0.14	0.10	0.02
3	0.5	0.13	0.37	0.13	0.14	0.13	0.02
	0.9	0.00	0.37	0.00	0.14	0.00	0.02

Expected frequencies are calculated based on binomial and Poisson distributions. The frequency of no marks on an FP population is

For SPLIT: $P(\text{no FP marks}) = P\left(N^0 \text{ or } N^t = 0 \middle| N^0, N^t \sim Poisson(\lambda)\right)$

For IGNORE: $P(\text{no FP marks}) = P\left(N^0 = 0 \middle| N^0 \sim Poisson(\lambda)\right)$

For SWITCH: $P(\text{no FP marks}) = P\left(N^0 = 0 \text{ and } N^t = 0 \middle| N^0, N^t \sim Poisson(\lambda)\right)$

The frequency of no marks on a TP population is $P(\text{no TP mark}) = P\left(M = 0 \middle| M \sim Bin(t, \upsilon)\right)$

Table 3.3. Estimated type I error rates for normal and skewed distributions.

under normal distributions

			AUC=0.5				*AUC*=0.7				*AUC*=0.9			
			J1	$SPLIT_2$	$SWITCH_2$	θ_c	*J1*	$SPLIT_2$	$SWITCH_2$	θ_c	*J1*	$SPLIT_2$	$SWITCH_2$	θ_c
b	*t*	*v*												
1	1	0.5	0.056	0.057	0.053	0.055	0.057	0.058	0.054	0.057	0.059	0.058	0.058	0.056
		0.7	0.058	0.060	0.056	0.057	0.058	0.060	0.059	0.057	0.052	0.053	0.054	0.054
		0.9	0.058	0.055	0.054	0.054	0.058	0.055	0.053	0.054	0.055	0.053	0.055	0.053
	3	0.5	0.060	0.062	0.050	0.050	0.059	0.056	0.057	0.056	0.056	0.056	0.058	0.053
		0.7	0.061	0.055	0.052	0.049	0.059	0.059	0.056	0.051	0.055	0.051	0.054	0.050
		0.9	0.056	0.056	0.055	0.051	0.059	0.057	0.057	0.054	0.056	0.041	0.035	0.052
2	1	0.5	0.059	0.059	0.057	0.057	0.061	0.059	0.060	0.058	0.058	0.058	0.059	0.057
		0.7	0.059	0.060	0.057	0.059	0.057	0.057	0.057	0.057	0.054	0.056	0.057	0.057
		0.9	0.056	0.057	0.054	0.057	0.049	0.050	0.052	0.052	0.054	0.054	0.055	0.056
	3	0.5	0.056	0.057	0.049	0.054	0.059	0.057	0.055	0.052	0.060	0.058	0.060	0.052
		0.7	0.060	0.057	0.055	0.050	0.057	0.054	0.057	0.052	0.054	0.051	0.057	0.051
		0.9	0.053	0.053	0.053	0.051	0.055	0.055	0.056	0.057	0.049	0.038	0.035	0.055

under skewed distributions

			AUC=0.5				*AUC*=0.7				*AUC*=0.9			
			J1	$SPLIT_2$	$SWITCH_2$	θ_c	*J1*	$SPLIT_2$	$SWITCH_2$	θ_c	*J1*	$SPLIT_2$	$SWITCH_2$	θ_c
b	*t*	*v*												
1	1	0.5	0.057	0.058	0.054	0.051	0.057	0.058	0.055	0.055	0.060	0.059	0.058	0.057
		0.7	0.059	0.061	0.060	0.060	0.059	0.059	0.058	0.056	0.052	0.052	0.053	0.054
		0.9	0.057	0.054	0.055	0.053	0.056	0.055	0.053	0.054	0.055	0.052	0.056	0.052
	3	0.5	0.060	0.062	0.050	0.053	0.059	0.057	0.055	0.054	0.056	0.056	0.058	0.053
		0.7	0.061	0.057	0.052	0.049	0.059	0.057	0.056	0.050	0.055	0.051	0.055	0.049
		0.9	0.055	0.055	0.054	0.050	0.060	0.062	0.057	0.051	0.054	0.040	0.036	0.050
2	1	0.5	0.058	0.058	0.059	0.057	0.060	0.059	0.060	0.057	0.057	0.058	0.057	0.057
		0.7	0.059	0.059	0.057	0.059	0.059	0.058	0.056	0.057	0.054	0.055	0.055	0.057
		0.9	0.060	0.059	0.055	0.056	0.049	0.052	0.051	0.054	0.055	0.054	0.055	0.056
	3	0.5	0.057	0.056	0.051	0.051	0.059	0.057	0.056	0.052	0.061	0.059	0.059	0.052
		0.7	0.059	0.058	0.056	0.050	0.057	0.056	0.056	0.054	0.054	0.051	0.057	0.050
		0.9	0.053	0.055	0.053	0.052	0.058	0.055	0.055	0.056	0.048	0.037	0.036	0.055

Estimated type I error rates are obtained when testing the equality of the indices under a simulation model based on normal and skewed distributions for λ=1 with a sample size of 20 subjects and 10,000 simulations.

Table 3.4. Estimated power for detecting system difference with regard to *AUC* --- the SPLIT method (based on max, mean, Wilcoxon).

under normal distributions

			AUC=0.7 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.7		
			$SPLIT_0$	$SPLIT_1$	$SPLIT_2$	$SPLIT_0$	$SPLIT_1$	$SPLIT_2$	$SPLIT_0$	$SPLIT_1$	$SPLIT_2$
1/b	*t*	*v*									
1	1	0.5	0.097	**0.098**	0.095	**0.245**	0.210	0.205	**0.106**	0.092	0.091
		0.7	**0.148**	0.146	0.140	**0.430**	0.371	0.362	**0.170**	0.138	0.140
		0.9	0.226	**0.242**	0.236	**0.747**	0.701	0.693	**0.327**	0.275	0.278
	3	0.5	**0.214**	0.121	0.121	**0.627**	0.425	0.333	**0.225**	0.181	0.135
		0.7	**0.316**	0.217	0.235	**0.861**	0.743	0.722	**0.380**	0.310	0.273
		0.9	0.403	0.396	**0.418**	**0.968**	0.961	0.969	0.507	0.520	**0.573**
2	1	0.5	0.089	**0.089**	0.087	**0.222**	0.200	0.199	**0.098**	0.089	0.089
		0.7	0.137	**0.139**	0.138	**0.383**	0.345	0.345	**0.148**	0.128	0.128
		0.9	0.202	**0.202**	0.202	**0.696**	0.659	0.661	**0.268**	0.237	0.243
	3	0.5	**0.191**	0.166	0.119	0.497	**0.574**	0.329	0.166	**0.244**	0.135
		0.7	0.281	**0.292**	0.230	0.729	**0.855**	0.703	0.265	**0.388**	0.271
		0.9	0.370	**0.491**	0.426	0.897	**0.987**	0.968	0.260	**0.600**	0.536

under skewed distributions

			AUC=0.7 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.7		
			$SPLIT_0$	$SPLIT_1$	$SPLIT_2$	$SPLIT_0$	$SPLIT_1$	$SPLIT_2$	$SPLIT_0$	$SPLIT_1$	$SPLIT_2$
1/b	*t*	*v*									
1	1	0.5	0.090	0.089	**0.091**	**0.232**	0.203	0.202	**0.111**	0.096	0.095
		0.7	**0.131**	0.126	0.128	**0.417**	0.362	0.360	**0.180**	0.146	0.144
		0.9	0.198	0.207	**0.211**	**0.743**	0.686	0.685	**0.351**	0.293	0.290
	3	0.5	**0.203**	0.101	0.128	**0.646**	0.410	0.363	**0.254**	0.214	0.151
		0.7	**0.313**	0.173	0.237	**0.895**	0.757	0.751	**0.448**	0.389	0.302
		0.9	**0.415**	0.316	0.391	**0.986**	0.952	0.967	**0.623**	0.569	0.592
2	1	0.5	**0.093**	0.086	0.086	**0.230**	0.195	0.197	**0.097**	0.091	0.090
		0.7	**0.148**	0.133	0.132	**0.403**	0.341	0.340	**0.149**	0.131	0.129
		0.9	**0.231**	0.197	0.202	**0.737**	0.657	0.660	**0.277**	0.241	0.242
	3	0.5	**0.259**	0.181	0.123	0.618	**0.651**	0.344	0.181	**0.291**	0.133
		0.7	**0.439**	0.338	0.242	0.878	**0.907**	0.713	0.324	**0.448**	0.267
		0.9	**0.629**	0.517	0.447	**0.989**	0.987	0.966	0.411	**0.616**	0.520

Estimated statistical powers are obtained when detecting system difference with regard to the degree of separation between FP and TP ratings (AUC) for λ=1 under a simulation model based on normal and skewed distribution with a sample size of 20 subjects and 10,000 simulations. The greatest statistical power for each scenario is in bold font.

Table 3.5. Estimated power for detecting system difference with regard to *AUC* --- the SWITCH method (based on max, mean, Wilcoxon).

under normal distributions

			AUC=0.7 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.7		
			$SWITCH_0$	$SWITCH_1$	$SWITCH_2$	$SWITCH_0$	$SWITCH_1$	$SWITCH_2$	$SWITCH_0$	$SWITCH_1$	$SWITCH_2$
1/b	*t*	*v*									
1	1	0.5	0.121	**0.121**	0.115	**0.371**	0.285	0.277	**0.144**	0.105	0.106
		0.7	0.175	**0.182**	0.175	**0.575**	0.475	0.460	**0.231**	0.160	0.164
		0.9	0.239	**0.279**	0.271	**0.818**	0.780	0.772	**0.381**	0.295	0.300
	3	0.5	**0.233**	0.150	0.149	**0.746**	0.556	0.458	**0.291**	0.215	0.164
		0.7	**0.303**	0.268	0.284	**0.877**	0.811	0.793	**0.384**	0.323	0.292
		0.9	0.355	0.441	**0.453**	0.922	0.962	**0.971**	0.421	0.477	**0.519**
2	1	0.5	**0.116**	0.112	0.110	**0.322**	0.268	0.266	**0.126**	0.105	0.103
		0.7	**0.162**	0.160	0.160	**0.509**	0.439	0.439	**0.187**	0.152	0.154
		0.9	0.221	**0.234**	0.228	**0.770**	0.723	0.724	**0.313**	0.257	0.261
	3	0.5	**0.222**	0.210	0.151	0.604	**0.692**	0.440	0.204	**0.283**	0.156
		0.7	0.283	**0.332**	0.279	0.763	**0.888**	0.767	0.263	**0.389**	0.286
		0.9	0.339	**0.505**	0.446	0.848	**0.980**	0.963	0.194	**0.526**	0.486

under skewed distributions

			AUC=0.7 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.7		
			$SWITCH_0$	$SWITCH_1$	$SWITCH_2$	$SWITCH_0$	$SWITCH_1$	$SWITCH_2$	$SWITCH_0$	$SWITCH_1$	$SWITCH_2$
1/b	*t*	*v*									
1	1	0.5	0.103	0.106	**0.109**	**0.346**	0.271	0.275	**0.153**	0.114	0.110
		0.7	0.150	0.151	**0.160**	**0.568**	0.456	0.457	**0.249**	0.175	0.169
		0.9	0.201	0.224	**0.240**	**0.828**	0.756	0.767	**0.411**	0.320	0.314
	3	0.5	**0.221**	0.105	0.152	**0.781**	0.562	0.500	**0.338**	0.295	0.183
		0.7	**0.309**	0.188	0.281	**0.925**	0.827	0.821	**0.486**	0.442	0.324
		0.9	0.387	0.345	**0.435**	**0.975**	0.953	0.973	**0.590**	0.534	0.545
2	1	0.5	**0.125**	0.103	0.106	**0.366**	0.259	0.262	**0.131**	0.108	0.105
		0.7	**0.187**	0.150	0.156	**0.579**	0.429	0.434	**0.196**	0.159	0.157
		0.9	**0.276**	0.213	0.224	**0.844**	0.719	0.728	**0.334**	0.270	0.267
	3	0.5	**0.325**	0.236	0.162	0.777	**0.784**	0.460	0.240	**0.346**	0.161
		0.7	**0.486**	0.382	0.288	0.935	**0.941**	0.778	0.343	**0.484**	0.287
		0.9	**0.622**	0.521	0.461	**0.989**	0.984	0.961	0.387	**0.563**	0.472

Estimated statistical powers are obtained when detecting system difference with regard to the degree of separation between FP and TP ratings (AUC) for λ=1 under a simulation model based on normal and skewed distribution with a sample size of 20 subjects and 10,000 simulations. The greatest statistical power for each scenario is in bold font.

Table 3.6. Estimated power for detecting system difference with regard to *AUC* --- the IGNORE method (based on max, mean, Wilcoxon).

under normal distributions

			AUC=0.7 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.7		
			IGNORE$_0$	*IGNORE*$_1$	*IGNORE*$_2$	*IGNORE*$_0$	*IGNORE*$_1$	*IGNORE*$_2$	*IGNORE*$_0$	*IGNORE*$_1$	*IGNORE*$_2$
1/b	*t*	*v*									
1	1	0.5	0.084	**0.085**	0.084	**0.205**	0.184	0.178	**0.097**	0.084	0.086
		0.7	0.120	**0.126**	0.119	**0.351**	0.314	0.307	**0.146**	0.126	0.127
		0.9	0.162	**0.179**	0.172	**0.607**	0.588	0.579	**0.245**	0.216	0.220
	3	0.5	**0.159**	0.094	0.095	**0.513**	0.285	0.242	**0.184**	0.134	0.110
		0.7	**0.215**	0.143	0.161	**0.714**	0.537	0.552	**0.263**	0.205	0.204
		0.9	0.243	0.250	**0.269**	0.805	0.811	**0.851**	0.253	0.332	**0.385**
2	1	0.5	**0.079**	0.079	0.078	**0.195**	0.181	0.180	**0.090**	0.083	0.084
		0.7	**0.121**	0.120	0.119	**0.330**	0.303	0.301	**0.132**	0.119	0.120
		0.9	0.155	**0.164**	0.163	**0.599**	0.575	0.574	**0.226**	0.203	0.208
	3	0.5	**0.162**	0.128	0.101	**0.423**	0.417	0.245	0.145	**0.180**	0.108
		0.7	**0.210**	0.193	0.159	0.637	**0.698**	0.558	0.205	**0.287**	0.213
		0.9	0.265	**0.343**	0.304	0.773	**0.932**	0.900	0.129	**0.440**	0.422

under skewed distributions

			AUC=0.7 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.7		
			IGNORE$_0$	*IGNORE*$_1$	*IGNORE*$_2$	*IGNORE*$_0$	*IGNORE*$_1$	*IGNORE*$_2$	*IGNORE*$_0$	*IGNORE*$_1$	*IGNORE*$_2$
1/b	*t*	*v*									
1	1	0.5	0.081	0.081	**0.082**	**0.200**	0.177	0.177	**0.101**	0.089	0.088
		0.7	0.107	0.110	**0.112**	**0.344**	0.311	0.307	**0.154**	0.135	0.132
		0.9	0.145	0.154	**0.159**	**0.599**	0.579	0.576	**0.261**	0.234	0.230
	3	0.5	**0.154**	0.083	0.097	**0.536**	0.268	0.263	**0.203**	0.160	0.120
		0.7	**0.209**	0.120	0.158	**0.759**	0.558	0.578	**0.316**	0.278	0.227
		0.9	**0.255**	0.204	0.247	**0.860**	0.807	0.842	0.346	0.394	**0.396**
2	1	0.5	**0.083**	0.076	0.076	**0.204**	0.176	0.176	**0.090**	0.083	0.083
		0.7	**0.125**	0.114	0.115	**0.351**	0.297	0.297	**0.133**	0.121	0.119
		0.9	**0.173**	0.158	0.161	**0.627**	0.564	0.564	**0.230**	0.210	0.208
	3	0.5	**0.205**	0.135	0.103	**0.525**	0.467	0.252	0.158	**0.208**	0.107
		0.7	**0.313**	0.222	0.164	**0.775**	0.767	0.554	0.237	**0.335**	0.212
		0.9	**0.440**	0.342	0.297	**0.931**	0.923	0.872	0.195	**0.440**	0.393

Estimated statistical powers are obtained when detecting system difference with regard to the degree of separation between FP and TP ratings (AUC) for λ=1 under a simulation model based on normal and skewed distribution with a sample size of 20 subjects and 10,000 simulations. The greatest statistical power for each scenario is in bold font.

Table 3.7. Estimated power for detecting system difference with regard to *AUC* --- the clustered ROC index θ_c.

under normal distributions

			AUC=0.7 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.7		
			J1	$SWITCH_2$	θ_c	*J1*	$SWITCH_2$	θ_c	*J1*	$SWITCH_2$	θ_c
1/b	*t*	*v*									
1	1	0.5	0.097	0.115	**0.133**	0.245	0.277	**0.340**	0.106	0.106	**0.121**
		0.7	0.148	0.175	**0.217**	0.430	0.460	**0.576**	0.170	0.164	**0.198**
		0.9	0.226	0.271	**0.341**	0.747	0.772	**0.880**	0.327	0.300	**0.388**
	3	0.5	0.124	0.149	**0.273**	0.360	0.458	**0.724**	0.148	0.164	**0.233**
		0.7	0.200	0.284	**0.413**	0.684	0.793	**0.932**	0.285	0.292	**0.394**
		0.9	0.322	0.453	**0.595**	0.942	0.971	**0.997**	0.563	0.519	**0.692**
2	1	0.5	0.089	0.110	**0.128**	0.222	0.266	**0.325**	0.098	0.103	**0.114**
		0.7	0.137	0.160	**0.195**	0.383	0.439	**0.532**	0.148	0.154	**0.183**
		0.9	0.202	0.228	**0.287**	0.696	0.724	**0.822**	0.268	0.261	**0.329**
	3	0.5	0.120	0.151	**0.270**	0.334	0.440	**0.704**	0.135	0.156	**0.227**
		0.7	0.201	0.279	**0.411**	0.654	0.767	**0.922**	0.266	0.286	**0.384**
		0.9	0.353	0.446	**0.611**	0.951	0.963	**0.997**	0.568	0.486	**0.701**

under skewed distributions

			AUC=0.7 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.5			*AUC*=0.9 versus *AUC*=0.7		
			J1	$SWITCH_2$	θ_c	*J1*	$SWITCH_2$	θ_c	*J1*	$SWITCH_2$	θ_c
1/b	*t*	*v*									
1	1	0.5	0.090	0.109	**0.125**	0.232	0.275	**0.340**	0.111	0.110	**0.130**
		0.7	0.131	0.160	**0.201**	0.417	0.457	**0.586**	0.180	0.169	**0.213**
		0.9	0.198	0.240	**0.322**	0.743	0.767	**0.894**	0.351	0.314	**0.418**
	3	0.5	0.114	0.152	**0.254**	0.345	0.500	**0.735**	0.154	0.183	**0.251**
		0.7	0.177	0.281	**0.389**	0.669	0.821	**0.939**	0.299	0.324	**0.428**
		0.9	0.280	0.435	**0.574**	0.937	0.973	**0.998**	0.585	0.545	**0.725**
2	1	0.5	0.093	0.106	**0.129**	0.230	0.262	**0.326**	0.097	0.105	**0.112**
		0.7	0.148	0.156	**0.203**	0.403	0.434	**0.551**	0.149	0.157	**0.178**
		0.9	0.231	0.224	**0.319**	0.737	0.728	**0.852**	0.277	0.267	**0.327**
	3	0.5	0.125	0.162	**0.273**	0.347	0.460	**0.702**	0.134	0.161	**0.218**
		0.7	0.211	0.288	**0.421**	0.669	0.778	**0.923**	0.267	0.287	**0.375**
		0.9	0.372	0.461	**0.626**	0.954	0.961	**0.997**	0.563	0.472	**0.689**

Estimated statistical powers are obtained when detecting system difference with regard to the degree of separation between FP and TP ratings (AUC) for λ=1 under a simulation model based on normal and skewed distribution with a sample size of 20 subjects and 10,000 simulations. The greatest statistical power for each scenario is in bold font.

Table 3.8. Comparison of statistical power for different types of indices using *t* test.

under normal distributions

Index1	*Index2*	*difference*	*Lower CL*	*Upper CL*	*t statistic*	*critical*	*reject H_0*
$SPLIT_0$	$IGNORE_0$	0.0737	0.0504	0.0970	7.92	2.51	yes
$SPLIT_1$	$IGNORE_1$	0.0748	0.0502	0.0995	7.61	2.51	yes
$SPLIT_2$	$IGNORE_2$	0.0625	0.0416	0.0834	7.49	2.51	yes
$SWITCH_0$	$SPLIT_0$	0.0316	0.0085	0.0547	3.43	2.51	yes
$SWITCH_1$	$SPLIT_1$	0.0363	0.0196	0.0531	5.43	2.51	yes
$SWITCH_2$	$SPLIT_2$	0.0365	0.0207	0.0523	5.79	2.51	yes
J1	*J2*	0.0799	0.0494	0.1104	7.50	2.86	yes
J1	$SPLIT_0$	-0.0260	-0.0680	0.0154	1.81	2.86	no
J1	$SPLIT_1$	-0.0250	-0.0580	0.0086	2.12	2.86	no
J1	$SPLIT_2$	0.0034	-0.0130	0.0196	0.59	2.86	no
J1	$SWITCH_0$	-0.0580	-0.1140	-0.0020	2.95	2.86	yes
J1	$SWITCH_1$	-0.0610	-0.1010	-0.0210	4.37	2.86	yes
J1	$SWITCH_2$	-0.0330	-0.0550	-0.0120	4.41	2.86	yes

under skewed distributions

Index1	*Index2*	*difference*	*Lower CL*	*Upper CL*	*t statistic*	*critical*	*reject H_0*
$SPLIT_0$	$IGNORE_0$	0.0807	0.0541	0.1072	7.62	2.51	yes
$SPLIT_1$	$IGNORE_1$	0.0755	0.0503	0.1007	7.51	2.51	yes
$SPLIT_2$	$IGNORE_2$	0.0660	0.0437	0.0884	7.41	2.51	yes
$SWITCH_0$	$SPLIT_0$	0.0509	0.0284	0.0735	5.67	2.51	yes
$SWITCH_1$	$SPLIT_1$	0.0383	0.0213	0.0553	5.65	2.51	yes
$SWITCH_2$	$SPLIT_2$	0.0379	0.0219	0.0538	5.94	2.51	yes
J1	*J2*	0.0811	0.0497	0.1124	7.41	2.86	yes
J1	$SPLIT_0$	-0.0640	-0.1120	-0.0160	3.81	2.86	yes
J1	$SPLIT_1$	-0.0300	-0.0720	0.0121	2.04	2.86	no
J1	$SPLIT_2$	-0.0170	0.0028	0.0227	0.39	2.86	no
J1	$SWITCH_0$	-0.1150	-0.1720	-0.0580	5.79	2.86	yes
J1	$SWITCH_1$	-0.0680	-0.1180	-0.0180	3.92	2.86	yes
J1	$SWITCH_2$	-0.0350	-0.0610	-0.0090	3.87	2.86	yes

Comparison of statistical power in detecting system difference resulting from the degree of separation between FP and TP ratings (AUCs) for λ=1 for different types of indices using t test with adjustment for multiple comparison.

4.0 INDICES THAT INCORPORATE THE NUMBER OF MARKS

4.1 BACKGROUND

An FROC experiment offers the observer the freedom to detect and mark all suspicious locations within a subject. Differing from other methodologies that address the localization task such as LROC [20,21] and ROI [22,24], the number of marks on a subject is not pre-determined by the experiment and is considered to be an important characteristic of the FROC system. When multiple readers use a diagnostic system in an FROC experiment, researchers have noticed that the average number of *FP* marks per subject partially reflect reader experiences [32,35,36]. It would be reasonable to expect an expert to find fewer *FP* marks (a smaller λ) and identify more *TP* marks (a larger ν) on each subject.

Several existing indices have been derived to reflect these important characteristics of an FROC system. One index based on the area under the FROC curve [33] penalized for the number of *FP* marks, rewarded for the fraction of *TP* marks, and adjusted for the effect of the target size. *JAFROC* indices [36] used the maximum *FP* rating to summarize an *FP* population and thus penalized for an increased number of *FP* marks. In addition, *JAFROC* indices, as well as our proposed FROC indices in Chapter 3, rewarded for the number of *TP* marks by combining non-marked lesions and *TP* marks into an *LR* population, and comparing it to an *FP* population.

Specifically, when the number of *TP* marks in the *LR* population increases, there is fewer number of non-marked lesions and the values of these indices increase.

However, in Chapter 3, when we compared the *FP* population to the *LR* population, our proposed indices, based on the comparison of the mean and the Wilcoxon statistic, focused on estimating the stochastic order of the underlying two populations and treated the number of *FP* marks in the *FP* population as a nuisance parameter. When we use the mean or the Wilcoxon statistic to compare the *FP* and *LR* populations, the indices values do not change when the *FP* population increases in size without changing the shape of the distribution. Thus the mean or the Wilcoxon statistic do not penalize for an increased number of *FP* marks. In addition, although the indices in Chapter 3 rewarded for the number of *TP* marks by comparing the *FP* marks to the non-marked lesions, this "rewarding" approach is indirect and might be improved.

In this Chapter we propose new indices to compare the *FP* and *TP* populations with the use of a family of comparison functions based on the modified Wilcoxon statistic. The comparison functions incorporate additional information on the number of marks of the two populations. Thus, they can successfully penalize the indices when there is an increased number of generated "wrong" (*FP*) marks and reward the indices when there is an increased number of generated "correct" (*TP*) marks. For the proposed indices we will derive asymptotic inferential procedures, investigate their performance and compare them to *J1* index (modified *JAFROC1*) in a simulation study.

4.2 METHODS

Using the notation of data (2.1), when comparing *N FP* ratings ("wrong" marks), $\vec{X}$, to *M TP* ratings, $\vec{Y}$, ("correct" marks) in a pair of *FP* and *TP* populations with both of them being marked

(N, M>0), the Wilcoxon statistic is formulated as $\frac{\sum_{c_1=1}^{N}\sum_{c_2=1}^{M} I(X_{c_1} < Y_{c_2})}{N \times M}$. In this Chapter, we propose new comparison functions that are based on a modification of the Wilcoxon statistic. We add a second function to the Wilcoxon statistic so that it incorporates the number of marks (N, M) in a pair of *FP* and *TP* populations:

$$\frac{\sum_{c_1=1}^{N}\sum_{c_2=1}^{M} I(X_{c_1} < Y_{c_2})}{N \times M} + f(N,M)$$

The underlying concept tells us that $f(N,M)$ should be a decreasing function with N and an increasing function with M. With the second function $f(N,M)$, indices can penalize for an increased number of *FP* marks (N) and reward for an increased number of *TP* marks (M), as well as, assess the stochastic order of the two underlying populations. We hereby propose one family of the modified Wilcoxon statistic, W_k:

$$f(N,M) = \frac{1}{k}\left(\frac{M}{N} - \frac{1}{C}\right), \text{ and } W_k\left(\vec{X},\vec{Y}\right) = \frac{\sum_{c_1=1}^{N}\sum_{c_2=1}^{M} I(X_{c_1} < Y_{c_2})}{N \times M} + \frac{1}{k}\left(\frac{M}{N} - \frac{1}{C}\right), \quad k > 0, C > 0$$

We add a function $\frac{1}{k}\left(\frac{M}{N} - \frac{1}{C}\right)$ to the Wilcoxon statistic, so the indices based on the modified Wilcoxon statistic, W_k, can incorporate the number of marks of the two underlying populations, as well as penalize for an increasing ratio $\frac{N}{M}$. Additionally, $\frac{1}{k}\left(\frac{M}{N} - \frac{1}{C}\right)$ is smaller than zero when $\frac{N}{M} > C$ and hence, it penalizes the indices. Similarly, $\frac{1}{k}\left(\frac{M}{N} - \frac{1}{C}\right)$ is larger than zero when $\frac{N}{M} < C$ and hence, it rewards the indices. This may reflect the diagnostic practice

when the medical investigator wants to set a "cut point" or "tolerance level" for the ratio $\frac{N}{M}$ that can be accepted for the underlying FROC study. Here, without additional information, we set C=1 and W_k is identical to the original Wilcoxon statistic when N=M. For the parameter k, it can be considered as the weight to the second function that incorporates the numbers. We hereby evaluate W_k when k changes from 2, 4, to 8.

Now we apply the modified Wilcoxon statistic W_k to the SPLIT and the SWITCH method, respectively. For comparing a pair of *FP* and *TP* populations with at least one of them being unmarked, we follow the same approaches used for the SPLIT method of (3.2) and SWITCH method of (3.4) presented in Chapter 3. Applying W_k to the SPLIT method of (3.2), we have:

$$SPLIT_{(k)} = E\left[\left\{\frac{\sum_{c_1=1}^{N}\sum_{c_2=1}^{M}\psi\left(X_{c_1},Y_{c_2}\right)}{N\times M}+\frac{1}{k}\left(\frac{M}{N}-1\right)\right\}\times I\left(\left\{\vec{X}\right\}\neq\varnothing,\left\{\vec{Y}\right\}\neq\varnothing\right)\right]$$

$$+\left[P\left(\vec{X}=\varnothing\right)-\frac{1}{2}P\left(\vec{X}=\varnothing,\vec{Y}=\varnothing\right)\right];\quad k=2,4,\text{ and }8$$

Applying W_k to the SWITCH method of (3.4), we have:

$$SWITCH_{(k)} = E\left\{\left\{\frac{\sum_{c_1=1}^{N^0}\sum_{c_3=1}^{M}\psi\left(X_{c_1}^{0},Y_{c_3}\right)+\sum_{c_2=1}^{N^t}\sum_{c_3=1}^{M}\psi\left(X_{c_2}^{t},Y_{c_3}\right)}{(N^0+N^t)\times M}+\frac{1}{k}\left(\frac{M}{N^0+N^t}-1\right)\right\}\times I\left(\left\{\vec{X}^0,\vec{X}^t\right\}\neq\varnothing,\left\{\vec{Y}\right\}\neq\varnothing\right)\right\}$$

$$+\left[P\left(\vec{X}^0=\varnothing,\vec{X}^t=\varnothing\right)-\frac{1}{2}P\left(\vec{X}^0=\varnothing,\vec{X}^t=\varnothing,\vec{Y}=\varnothing\right)\right];\quad k=2,4,\text{ and }8$$

4.3 STATISTICAL INFERENCE

Using the notation in (2.1), we have the following estimates for $\widehat{SPLIT}_{(k)}$ and $\widehat{SWITCH}_{(k)}$:

$$\widehat{SPLIT}_{(k)} = \frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\tilde{\psi}_{SPLIT_{(k)}}\left(\{x_{s_1c_1}\}_{c_1=1}^{n_{s_1}},\{y_{sc_2}\}_{c_2=1}^{m_s}\right) \text{ or}$$

$$\widehat{SPLIT}_{(k)} = \frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\left\{\frac{\sum_{c_1=1}^{n_{s_1}}\sum_{c_2=1}^{m_s}\psi\left(x_{s_1c_1},y_{sc_2}\right)}{n_{s_1}\times m_s}+\frac{1}{k}\left(\frac{m_s}{n_{s_1}}-1\right)\right\}\times I\left(m_s\neq 0, n_{s_1}\neq\varnothing\right)$$

$$+\frac{\sum_{s_1=1}^{S_0+S_t}I\left(n_{s_1}=0\right)}{S_0+S_t}-\frac{1}{2}\times\left[\frac{\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}I\left(n_{s_1}+m_s=0\right)}{\left(S_0+S_t\right)\times S_t}\right];\quad k=2,4,\text{ and }8 \quad (4.1)$$

$$\widehat{SWITCH}_{(k)} = \frac{1}{S_0\times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{SWITCH_{(k)}}\left(\left\{\left\{x^0_{s'c_1}\right\}_{c_1=1}^{n^0_{s'}},\left\{x^t_{sc_2}\right\}_{c_2=1}^{n^t_s}\right\},\{y_{sc_2}\}_{c_2=1}^{m_s}\right) \text{ or}$$

$$\widehat{SWITCH}_{(k)} = \frac{1}{S_0\times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\left\{\frac{\sum_{c_1=1}^{n^0_{s'}}\sum_{c_3=1}^{m_s}\psi(x^0_{s'c_1},y_{sc_3})+\sum_{c_2=1}^{n^t_s}\sum_{c_3=1}^{m_s}\psi(x^t_{sc_2},y_{sc_3})}{(n^0_{s'}+n^t_s)\times m_s}+\frac{1}{k}\left(\frac{m_s}{n^0_{s'}+n^t_s}-1\right)\right\}$$

$$\times I\left((n^0_{s'}+n^t_s)m_s\neq 0\right)+\frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}I\left(n^0_{s'}+n^t_s=0\right)-\frac{1}{2}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}I\left(n^0_{s'}+n^t_s+m_s=0\right)}{S_0\times S_t};\ k=2,4,\text{ and }8 \quad (4.2)$$

To illustrate the estimation of all the indices in (4.1)-(4.2), we provide an example using 4 actually negative and 4 actually positive subjects in Appendix.

For (4.1)-(4.2), we can still construct a closed form two-sample jackknife variance when the re-sampling techniques consider subject as a sampling unit. When we do two-sample jackknifing

for the indices in Chapter 2, removing an actually positive subject only influences a column of the $\tilde{\psi}$ matrix that represents the actually positive subject. Similarly we can construct the two-sample variance for SWITCH indices of (4.2) using the approach in Chapter 2 (2.10):

$$\hat{V}_{\text{2s-jackk}}(\widehat{SWITCH}_{(k)}) = \frac{\sum_{s'=1}^{S_0}\left(\overline{\tilde{\psi}_{s'\bullet}} - \overline{\tilde{\psi}_{\bullet\bullet}}\right)^2}{S_0 \times (S_0 - 1)} + \frac{\sum_{s=1}^{S_t}\left(\overline{\tilde{\psi}_{\bullet s}} - \overline{\tilde{\psi}_{\bullet\bullet}}\right)^2}{S_t \times (S_t - 1)}.$$

For the SPLIT indices of (4.1), removing an actually positive subject influences both a row and a column of the $\tilde{\psi}$ matrix that represent the actually positive subject. We derived their closed form two-sample jackknife variance for Chapter 3 in (3.16):

$$\hat{V}_{\text{2s-jackk}}(\widehat{SPLIT}_{(k)}) = \frac{\sum_{s_1=1}^{S_0}(\hat{A}^{s_1}_{negative;J2} - \hat{A}^2)^2}{S_0(S_0 - 1)} + \frac{\sum_{s=1}^{S_t}(\hat{A}^{s}_{positive;J2} - \hat{A}^1)^2}{S_t(S_t - 1)}$$

When comparing two diagnostic systems evaluated under the FROC paradigm with the proposed indices we use the difference between the modality-specific indices. When assessing statistical significance of the differences in the indices observed for the two modalities, we propose an asymptotic procedure with the test statistics: $Z_i = \frac{\hat{A}_i^2 - \hat{A}_i^1}{\sqrt{\hat{V}_{\text{2s-jackk}}\left(\hat{A}_i^2 - \hat{A}_i^1\right)}}$, i=1,…, 6.

4.4 SIMULATION RESULTS

In Chapter 3, the pattern of the observations for all indices is similar for both normal and skewed distributions. Thus we evaluate all the proposed indices in this Chapter under the simulation model in Chapter 2 for normal distributions with 10,000 simulation runs. The estimated type I error rates for all our proposed indices as well as *J1* index (modified *JAFROC1*) are shown in Table 4.1. It can be observed that in all simulated scenarios, the estimated type I error rates of all

the constructed asymptotic tests are close to the nominal value of 0.05 and they range from 0.044 to 0.063.

Table 4.2 exhibits the statistical power estimates in detecting the system difference with regard to *AUC*. As one can observe, compared to the statistical test based on *J1*, the statistical test based on $SPLIT_{(k)}$ tends to have smaller power in most considered scenarios for t=1 and tends to have greater power in most considered scenarios for t=2 and t=3. The statistical test based on $SWITCH_{(k)}$ tends to have greater power than the statistical test based on *J1* in most considered scenarios for k=4, 8. The statistical test based on $SWITCH_{(2)}$ tends to have greater power than the statistical test based on *J1* in most considered scenarios except for scenarios of (t=2, 3; ν=0.9). When k increases from 2 to 8, the statistical power tends to increase for both SPLIT and SWITCH indices except for scenarios of (t=1; ν=0.5, 0.7).

Table 4.3 exhibits the power estimates for detecting the system difference with regard to λ. From the table one can observe that for the task of identifying a difference in λ, compared to the statistical test based on *J1*, the statistical tests based on $SPLIT_{(2)}$, $SPLIT_{(4)}$ and $SWITCH_{(4)}$ have greater power in most considered scenarios for t=1 and have smaller power in most considered scenarios for t=2 and t=3. The statistical tests based on $SPLIT_{(8)}$ and $SWITCH_{(8)}$ have smaller power than the statistical test based on *J1* in all considered scenarios. The statistical test based on $SWITCH_{(2)}$ tends to have greater power than the statistical test based on *J1* in most considered scenarios except for scenarios when (t=2, 3; ν=0.5). When k increases from 2 to 8, the statistical power tends to decrease for both SPLIT and SWITCH indices except for SPLIT indices at scenarios of (t=3; ν=0.7, 0.9).

Table 4.4 and 4.5 exhibit the summary results for all indices. The mean statistical power averaged over three ν's (ν=0.5,0.7,0.9) and three statistical tests (three statistical test for *AUC*:

AUC=0.7 versus *AUC*=0.5, *AUC*=0.9 versus *AUC*=0.5, *AUC*=0.9 versus *AUC*=0.7 or three statistical tests for λ: λ=1 versus λ=0.5, λ=2 versus λ=0.5, λ=2 versus λ=1) are listed. From the tables one can observe that for detecting the system difference with regard to *AUC*, the statistical test based on $SWITCH_{(8)}$ tends to have the greatest average power among all indices. For detecting the system difference with regard to λ, the statistical test based on $SWITCH_{(2)}$ tends to have the greatest average power in all scenarios. For both tasks, the statistical test based on $SWITCH_{(2)}$ has greater average power than the statistical test based on *J1* in almost all considered scenarios.

4.5 SUMMARY

In this Chapter, we incorporated the number of marks when comparing *FP* and *TP* populations to evaluate the FROC diagnostic systems. We proposed indices with the use of a family of comparison functions based on the modified Wilcoxon statistic W_k. In the simulation study, the statistical tests based on the proposed indices have been shown to achieve a nominal type I error rate and have different power advantages for detecting a system difference with regard to *AUC* and λ, respectively. The statistical test based on $SWITCH_{(2)}$ has greater average power than the statistical test based on *J1* (modified *JAFROC1*) in almost all considered scenarios.

We propose to add a function, $\frac{1}{k}\left(\frac{M}{N}-\frac{1}{C}\right)$, to the Wilcoxon statistic. In addition to assessing the stochastic order between the *FP* and *TP* populations, indices based on the modified Wilcoxon statistic ,W_k, penalize for an increased number of *FP* marks (*N*) and reward for an increased number of *TP* marks (*M*). The parameter *k* can be viewed as the weight for the numbers to be penalized or rewarded. When applying W_k to the SWITCH method, the statistical tests still show

power advantages for detecting the system difference with regard to *AUC* than the statistical test based on *J1*. For detecting the system difference with regard to λ, the statistical test based on $SWITCH_{(2)}$ has greater average power than the statistical test based on *J1*, although the statistical tests do not behave that well for $SWITCH_{(4)}$ and $SWITCH_{(8)}$.

In this Chapter, we focused on two types of differences in FROC systems (with regard to *AUC* and λ). As we observed from the tables, an index may perform better in detecting one system difference but not as well in detecting others. The benefit of our proposed indices is that they are a family of indices, where it may not be hard to find an index that has greater average power than *J1* index. Despite its improvement, it may not be reasonable to expect our indices to have greater power than the naïve parameter estimator, in detecting the system difference with regard to that parameter. As we discussed in Chapter 3, when ignoring the other features of the FROC system, the refined clustered index, $\widehat{AUC}$, only estimates the parameter *AUC*. Likewise, although we do not study it in this Chapter, an index based on the simple count of the average number of *FP* marks per subject ($\hat{\lambda}$) can be formulated by ignoring the other features of the FROC system and only estimating the parameter λ.

It should be noted that the two types of differences we evaluated are not the only types of differences between two FROC systems. Specifically, we do not consider system difference due to the fraction of *TP* marks on each subject (ν), even though our proposed method, as well as *J1*, can reward for a higher ν, as we introduced at the beginning of this Chapter.

Table 4.1 Estimated type I error rates for normal distributions.

			AUC=0.5						
1/b	*t*	ν	*J1*	$SPLIT_{(2)}$	$SPLIT_{(4)}$	$SPLIT_{(8)}$	$SWITCH_{(2)}$	$SWITCH_{(4)}$	$SWITCH_{(8)}$
1	1	0.5	0.056	0.057	0.057	0.057	0.053	0.052	0.054
		0.7	0.058	0.058	0.059	0.058	0.052	0.055	0.055
		0.9	0.058	0.057	0.058	0.058	0.056	0.057	0.056
	2	0.5	0.059	0.057	0.057	0.057	0.053	0.053	0.054
		0.7	0.055	0.057	0.057	0.057	0.056	0.056	0.054
		0.9	0.057	0.055	0.055	0.057	0.054	0.053	0.054
	3	0.5	0.054	0.055	0.057	0.057	0.053	0.054	0.056
		0.7	0.055	0.051	0.052	0.052	0.051	0.052	0.052
		0.9	0.055	0.059	0.056	0.054	0.052	0.052	0.052
2	1	0.5	0.059	0.057	0.059	0.059	0.055	0.056	0.056
		0.7	0.057	0.058	0.059	0.059	0.055	0.056	0.056
		0.9	0.057	0.057	0.058	0.057	0.052	0.054	0.053
	2	0.5	0.060	0.058	0.058	0.056	0.051	0.053	0.054
		0.7	0.061	0.056	0.055	0.054	0.055	0.054	0.055
		0.9	0.054	0.053	0.053	0.052	0.053	0.054	0.054
	3	0.5	0.063	0.055	0.058	0.060	0.057	0.061	0.060
		0.7	0.059	0.052	0.053	0.056	0.054	0.055	0.058
		0.9	0.055	0.057	0.056	0.055	0.053	0.053	0.054

			AUC=0.9						
1/b	*t*	ν	*J1*	$SPLIT_{(2)}$	$SPLIT_{(4)}$	$SPLIT_{(8)}$	$SWITCH_{(2)}$	$SWITCH_{(4)}$	$SWITCH_{(8)}$
1	1	0.5	0.055	0.056	0.056	0.056	0.053	0.056	0.056
		0.7	0.054	0.052	0.054	0.053	0.053	0.052	0.053
		0.9	0.054	0.058	0.055	0.054	0.057	0.056	0.056
	2	0.5	0.059	0.059	0.061	0.061	0.058	0.058	0.059
		0.7	0.057	0.056	0.056	0.057	0.055	0.057	0.059
		0.9	0.055	0.054	0.051	0.044	0.053	0.053	0.052
	3	0.5	0.056	0.058	0.062	0.059	0.056	0.059	0.059
		0.7	0.056	0.056	0.053	0.047	0.051	0.053	0.053
		0.9	0.054	0.059	0.058	0.053	0.052	0.051	0.052
2	1	0.5	0.056	0.055	0.057	0.056	0.053	0.056	0.056
		0.7	0.058	0.058	0.058	0.059	0.057	0.058	0.058
		0.9	0.053	0.058	0.055	0.053	0.056	0.056	0.056
	2	0.5	0.058	0.055	0.057	0.058	0.057	0.059	0.059
		0.7	0.058	0.059	0.057	0.054	0.055	0.055	0.057
		0.9	0.053	0.054	0.051	0.046	0.053	0.052	0.052
	3	0.5	0.055	0.055	0.055	0.054	0.056	0.054	0.054
		0.7	0.054	0.055	0.053	0.048	0.055	0.053	0.051
		0.9	0.049	0.059	0.057	0.052	0.047	0.047	0.049

Estimated type I error rates are obtained when testing the equality of the indices under a simulation model based on normal distributions for λ=1 with a sample size of 20 subjects and 10,000 simulations.

Table 4.2 Estimated power for detecting system difference with regard to *AUC*.

			AUC=0.7 versus *AUC*=0.5						
1/b	*t*	ν	*J1*	$SPLIT_{(2)}$	$SPLIT_{(4)}$	$SPLIT_{(8)}$	$SWITCH_{(2)}$	$SWITCH_{(4)}$	$SWITCH_{(8)}$
1	1	0.5	0.097	0.097	0.096	0.095	0.126	0.120	0.117
		0.7	0.148	0.142	0.141	0.140	0.172	0.176	0.174
		0.9	0.226	0.211	0.225	0.229	0.229	0.254	0.262
	2	0.5	0.116	0.145	0.159	0.162	0.172	0.193	0.196
		0.7	0.172	0.201	0.235	0.245	0.219	0.269	0.290
		0.9	0.291	0.313	0.384	0.396	0.273	0.362	0.403
	3	0.5	0.123	0.169	0.213	0.229	0.193	0.246	0.266
		0.7	0.202	0.222	0.332	0.366	0.227	0.337	0.388
		0.9	0.319	0.260	0.446	0.492	0.249	0.392	0.473
2	1	0.5	0.095	0.095	0.095	0.094	0.111	0.111	0.110
		0.7	0.139	0.139	0.138	0.136	0.164	0.166	0.164
		0.9	0.210	0.195	0.205	0.209	0.212	0.228	0.233
	2	0.5	0.112	0.137	0.144	0.146	0.164	0.175	0.181
		0.7	0.183	0.203	0.230	0.236	0.216	0.261	0.274
		0.9	0.303	0.301	0.369	0.376	0.276	0.354	0.393
	3	0.5	0.123	0.166	0.203	0.216	0.189	0.241	0.260
		0.7	0.198	0.224	0.326	0.357	0.223	0.319	0.371
		0.9	0.345	0.270	0.462	0.508	0.255	0.401	0.483

			AUC=0.9 versus *AUC*=0.5						
1/b	*t*	ν	*J1*	$SPLIT_{(2)}$	$SPLIT_{(4)}$	$SPLIT_{(8)}$	$SWITCH_{(2)}$	$SWITCH_{(4)}$	$SWITCH_{(8)}$
1	1	0.5	0.240	0.215	0.208	0.201	0.321	0.299	0.285
		0.7	0.432	0.383	0.382	0.374	0.514	0.503	0.489
		0.9	0.748	0.670	0.687	0.691	0.739	0.775	0.779
	2	0.5	0.316	0.382	0.415	0.426	0.504	0.547	0.552
		0.7	0.588	0.629	0.720	0.746	0.692	0.795	0.825
		0.9	0.909	0.879	0.958	0.968	0.841	0.946	0.971
	3	0.5	0.350	0.487	0.605	0.643	0.585	0.711	0.752
		0.7	0.673	0.683	0.887	0.924	0.719	0.901	0.942
		0.9	0.947	0.755	0.980	0.995	0.787	0.964	0.989
2	1	0.5	0.218	0.211	0.203	0.199	0.304	0.286	0.272
		0.7	0.384	0.361	0.357	0.354	0.481	0.471	0.458
		0.9	0.689	0.631	0.650	0.655	0.690	0.726	0.733
	2	0.5	0.294	0.374	0.406	0.417	0.486	0.523	0.532
		0.7	0.561	0.608	0.690	0.715	0.674	0.768	0.794
		0.9	0.896	0.867	0.949	0.957	0.829	0.939	0.962
	3	0.5	0.339	0.484	0.598	0.635	0.574	0.701	0.744
		0.7	0.659	0.680	0.880	0.918	0.715	0.898	0.941
		0.9	0.943	0.756	0.981	0.995	0.786	0.966	0.990

Estimated statistical powers are obtained when detecting system difference with regard to the degree of separation between FP and TP ratings (AUC) for λ=1 under a simulation model based on normal distribution with a sample size of 20 subjects and 10,000 simulations.

Table 4.2 Continued.

			AUC=0.9 versus *AUC*=0.7						
1/b	*t*	ν	*J1*	$SPLIT_{(2)}$	$SPLIT_{(4)}$	$SPLIT_{(8)}$	$SWITCH_{(2)}$	$SWITCH_{(4)}$	$SWITCH_{(8)}$
1	1	0.5	0.101	0.090	0.088	0.087	0.117	0.112	0.108
		0.7	0.168	0.144	0.142	0.138	0.178	0.175	0.172
		0.9	0.330	0.259	0.274	0.278	0.291	0.314	0.321
	2	0.5	0.131	0.141	0.150	0.155	0.176	0.190	0.194
		0.7	0.244	0.240	0.292	0.312	0.262	0.333	0.361
		0.9	0.489	0.394	0.566	0.623	0.361	0.517	0.592
	3	0.5	0.152	0.177	0.234	0.258	0.215	0.280	0.304
		0.7	0.282	0.246	0.423	0.513	0.263	0.430	0.519
		0.9	0.552	0.265	0.568	0.741	0.299	0.536	0.681
2	1	0.5	0.099	0.094	0.092	0.091	0.116	0.111	0.107
		0.7	0.145	0.134	0.131	0.130	0.162	0.157	0.155
		0.9	0.280	0.243	0.250	0.253	0.263	0.282	0.284
	2	0.5	0.122	0.135	0.149	0.154	0.171	0.183	0.186
		0.7	0.210	0.216	0.265	0.287	0.242	0.305	0.326
		0.9	0.461	0.368	0.533	0.592	0.344	0.492	0.558
	3	0.5	0.139	0.167	0.217	0.238	0.201	0.261	0.282
		0.7	0.266	0.240	0.415	0.507	0.262	0.424	0.510
		0.9	0.565	0.268	0.583	0.770	0.301	0.561	0.708

Estimated powers are obtained when detecting system difference with regard to AUC for λ=1 under a simulation model based on normal distribution with 20 subjects and 10,000 simulations.

Table 4.3 Estimated power for detecting system difference with regard to *λ*.

			λ=1 versus λ=0.5						
1/b	*t*	ν	*J1*	$SPLIT_{(2)}$	$SPLIT_{(4)}$	$SPLIT_{(8)}$	$SWITCH_{(2)}$	$SWITCH_{(4)}$	$SWITCH_{(8)}$
1	1	0.5	0.144	0.173	0.145	0.130	0.204	0.149	0.123
		0.7	0.132	0.175	0.133	0.117	0.225	0.143	0.114
		0.9	0.154	0.232	0.162	0.132	0.282	0.170	0.121
	2	0.5	0.320	0.241	0.159	0.123	0.280	0.133	0.087
		0.7	0.340	0.320	0.211	0.153	0.437	0.215	0.120
		0.9	0.348	0.451	0.334	0.229	0.646	0.367	0.190
	3	0.5	0.931	0.481	0.402	0.333	0.827	0.540	0.322
		0.7	0.931	0.425	0.486	0.472	0.949	0.782	0.523
		0.9	0.914	0.266	0.463	0.580	0.990	0.939	0.731
2	1	0.5	0.134	0.169	0.144	0.130	0.195	0.141	0.121
		0.7	0.120	0.171	0.135	0.116	0.215	0.139	0.110
		0.9	0.126	0.219	0.152	0.122	0.273	0.162	0.114
	2	0.5	0.267	0.229	0.149	0.117	0.259	0.131	0.088
		0.7	0.273	0.307	0.197	0.144	0.423	0.203	0.110
		0.9	0.248	0.408	0.288	0.197	0.616	0.334	0.167
	3	0.5	0.903	0.459	0.372	0.308	0.815	0.511	0.298
		0.7	0.889	0.407	0.464	0.439	0.947	0.773	0.507
		0.9	0.843	0.259	0.446	0.545	0.991	0.937	0.729

Estimated powers are obtained when detecting system difference with regard to λ for AUC=0.8 under a simulation model based on normal distribution with 20 subjects and 10,000 simulations.

Table 4.3 Continued.

			λ=2 versus λ=0.5						
1/b	*t*	ν	*J1*	$SPLIT_{(2)}$	$SPLIT_{(4)}$	$SPLIT_{(8)}$	$SWITCH_{(2)}$	$SWITCH_{(4)}$	$SWITCH_{(8)}$
1	1	0.5	0.146	0.176	0.144	0.132	0.205	0.147	0.120
		0.7	0.132	0.172	0.134	0.118	0.223	0.142	0.109
		0.9	0.161	0.235	0.170	0.140	0.288	0.174	0.128
	2	0.5	0.327	0.250	0.162	0.126	0.286	0.134	0.091
		0.7	0.332	0.319	0.206	0.145	0.436	0.211	0.116
		0.9	0.347	0.441	0.333	0.230	0.649	0.364	0.187
	3	0.5	0.931	0.472	0.398	0.331	0.832	0.538	0.320
		0.7	0.937	0.422	0.490	0.473	0.949	0.788	0.524
		0.9	0.910	0.261	0.461	0.570	0.989	0.938	0.724
2	1	0.5	0.131	0.174	0.142	0.129	0.196	0.142	0.118
		0.7	0.118	0.169	0.130	0.116	0.207	0.137	0.105
		0.9	0.130	0.223	0.159	0.130	0.272	0.166	0.121
	2	0.5	0.276	0.231	0.153	0.118	0.261	0.134	0.092
		0.7	0.261	0.298	0.188	0.134	0.408	0.192	0.113
		0.9	0.249	0.413	0.292	0.196	0.610	0.330	0.163
	3	0.5	0.903	0.460	0.375	0.302	0.817	0.522	0.298
		0.7	0.888	0.402	0.453	0.431	0.946	0.770	0.502
		0.9	0.850	0.266	0.452	0.553	0.990	0.938	0.733

			λ=2 versus λ=1						
1/b	*t*	ν	*J1*	$SPLIT_{(2)}$	$SPLIT_{(4)}$	$SPLIT_{(8)}$	$SWITCH_{(2)}$	$SWITCH_{(4)}$	$SWITCH_{(8)}$
1	1	0.5	0.142	0.172	0.142	0.129	0.203	0.145	0.120
		0.7	0.135	0.177	0.138	0.118	0.223	0.142	0.112
		0.9	0.151	0.234	0.165	0.132	0.284	0.172	0.122
	2	0.5	0.328	0.245	0.159	0.122	0.280	0.137	0.091
		0.7	0.340	0.326	0.217	0.155	0.438	0.222	0.123
		0.9	0.356	0.449	0.340	0.235	0.652	0.370	0.189
	3	0.5	0.930	0.477	0.399	0.323	0.831	0.537	0.318
		0.7	0.932	0.423	0.487	0.475	0.948	0.785	0.527
		0.9	0.914	0.275	0.472	0.582	0.991	0.938	0.734
2	1	0.5	0.130	0.167	0.141	0.129	0.193	0.136	0.117
		0.7	0.119	0.176	0.134	0.116	0.217	0.144	0.111
		0.9	0.124	0.220	0.155	0.126	0.267	0.160	0.118
	2	0.5	0.271	0.227	0.147	0.115	0.260	0.130	0.088
		0.7	0.259	0.294	0.185	0.133	0.401	0.195	0.111
		0.9	0.258	0.413	0.294	0.199	0.614	0.340	0.168
	3	0.5	0.897	0.463	0.378	0.307	0.817	0.515	0.306
		0.7	0.895	0.413	0.465	0.442	0.948	0.784	0.509
		0.9	0.851	0.265	0.455	0.560	0.990	0.944	0.736

Estimated statistical powers are obtained when detecting system difference with regard to the average numbers of FP marks (λ) for AUC=0.8 under a simulation model based on normal distribution with a sample size of 20 subjects and 10,000 simulations.

Table 4.4 Summary of all indices in detecting system difference in *AUC*.

1/b	*t*	Averaged over 3 tests in *AUC*s and 3 v's	Indices Estimator	Mean power
1	1	9	*J1*	0.277
			$SPLIT_{(2)}$	0.246
			$SPLIT_{(4)}$	0.249
			$SPLIT_{(8)}$	0.248
			$SWITCH_{(2)}$	0.299
			$SWITCH_{(4)}$	**0.303**
			$SWITCH_{(8)}$	0.301
	2	9	*J1*	0.362
			$SPLIT_{(2)}$	0.369
			$SPLIT_{(4)}$	0.431
			$SPLIT_{(8)}$	0.448
			$SWITCH_{(2)}$	0.389
			$SWITCH_{(4)}$	0.461
			$SWITCH_{(8)}$	**0.487**
	3	9	*J1*	0.400
			$SPLIT_{(2)}$	0.363
			$SPLIT_{(4)}$	0.521
			$SPLIT_{(8)}$	0.573
			$SWITCH_{(2)}$	0.393
			$SWITCH_{(4)}$	0.533
			$SWITCH_{(8)}$	**0.590**
2	1	9	*J1*	0.251
			$SPLIT_{(2)}$	0.234
			$SPLIT_{(4)}$	0.236
			$SPLIT_{(8)}$	0.236
			$SWITCH_{(2)}$	0.278
			$SWITCH_{(4)}$	**0.282**
			$SWITCH_{(8)}$	0.280
	2	9	*J1*	0.349
			$SPLIT_{(2)}$	0.357
			$SPLIT_{(4)}$	0.415
			$SPLIT_{(8)}$	0.431
			$SWITCH_{(2)}$	0.378
			$SWITCH_{(4)}$	0.444
			$SWITCH_{(8)}$	**0.467**
	3	9	*J1*	0.397
			$SPLIT_{(2)}$	0.362
			$SPLIT_{(4)}$	0.518
			$SPLIT_{(8)}$	0.572
			$SWITCH_{(2)}$	0.390
			$SWITCH_{(4)}$	0.530
			$SWITCH_{(8)}$	**0.588**

Table 4.5 Summary of all indices in detecting system difference in λ.

1/b	*t*	Averaged over 3 tests in λ s and 3 v's	Indices Estimator	Mean power
1	1	9	*J1*	0.144
			$SPLIT_{(2)}$	0.194
			$SPLIT_{(4)}$	0.148
			$SPLIT_{(8)}$	0.128
			$SWITCH_{(2)}$	**0.237**
			$SWITCH_{(4)}$	0.154
			$SWITCH_{(8)}$	0.119
	2	9	*J1*	0.338
			$SPLIT_{(2)}$	0.338
			$SPLIT_{(4)}$	0.236
			$SPLIT_{(8)}$	0.169
			$SWITCH_{(2)}$	**0.456**
			$SWITCH_{(4)}$	0.239
			$SWITCH_{(8)}$	0.133
	3	9	*J1*	**0.926**
			$SPLIT_{(2)}$	0.389
			$SPLIT_{(4)}$	0.451
			$SPLIT_{(8)}$	0.460
			$SWITCH_{(2)}$	0.923
			$SWITCH_{(4)}$	0.754
			$SWITCH_{(8)}$	0.525
2	1	9	*J1*	0.126
			$SPLIT_{(2)}$	0.188
			$SPLIT_{(4)}$	0.144
			$SPLIT_{(8)}$	0.124
			$SWITCH_{(2)}$	**0.226**
			$SWITCH_{(4)}$	0.147
			$SWITCH_{(8)}$	0.115
	2	9	*J1*	0.262
			$SPLIT_{(2)}$	0.313
			$SPLIT_{(4)}$	0.210
			$SPLIT_{(8)}$	0.150
			$SWITCH_{(2)}$	**0.428**
			$SWITCH_{(4)}$	0.221
			$SWITCH_{(8)}$	0.122
	3	9	*J1*	0.880
			$SPLIT_{(2)}$	0.377
			$SPLIT_{(4)}$	0.429
			$SPLIT_{(8)}$	0.432
			$SWITCH_{(2)}$	**0.918**
			$SWITCH_{(4)}$	0.744
			$SWITCH_{(8)}$	0.513

5.0 MULTI-READER STUDY OF FROC SYSTEMS

5.1 BACKGROUND

For many diagnostic tests in radiology, the test results usually depend on the subjective interpretation of a reader. Since the 1990s, it has been widely recognized that several sources of variability within and between readers should be considered in the assessment of medical imaging [40,42,44-48]. Therefore, most studies to compare competing diagnostic modalities are designed to include multiple readers. However, such a multi-reader study design may raise two additional statistical issues. First, the ratings assigned to different modalities by different readers are obtained for the same set of subjects and hence, they may be correlated due to the replicates from multiple modalities and due to the replicates from multiple readers. Some phenomena may lead to a positive correlation between the number and ratings, while others, such as "satisfaction of search" [41], may lead to a negative correlation. Secondly, there are several sources of variability to consider when obtaining the estimates of diagnostic accuracy. These include estimating subject variability and reader variability. Under the ROC paradigm, a variety of methods are available to compare AUCs (area under the ROC curve) estimated using the same set of subjects evaluated by multiple readers [13, 47, 49-55].

The two most frequently used approaches for analyzing multi-reader ROC data are the Dorfman–Berbaum–Metz (DBM) approach [49] and Obuchowski and Rockette (OR) approach

[50]. The DBM approach applies an ANOVA-based procedure to the pseudo-values obtained from a jackknife estimate of the AUC. The pseudo-values of AUC are computed by removing each subject separately for each modality and reader combination. A mixed-effect model ANOVA procedure is then performed on the pseudo-values for testing whether the reader-averaged AUC is equivalent between two modalities. Instead of modeling the jackknife pseudo-values, Obuchowski and Rockette (OR) [50] model all of the original AUCs computed for each modality and reader combination with a two-way ANOVA. The OR approach [50] allows the random errors in the ANOVA model to be correlated and the usual ANOVA F tests are modified to correct for these correlations. Although they are different in their original formulations, Hillis *et al* [54] show that the DBM and OR approaches yield the same test statistic when based on the same ROC index, such as AUC, and the same covariance estimation method. However, the statistical inferences depend on which denominator degrees of freedom method (ddf), the DBM or OR approach, is used. Hillis [55] proposed a new ddf method that can be used with either the DBM or OR approach and it has been shown to have better properties than the ddf methods originally used in DBM and OR.

To evaluate the FROC indices in a multi-reader design setting, we need to use a reasonable simulation model that allows for the consideration of the multiple sources of correlations and reader heterogeneity in a multi-reader FROC experiment. To develop a multi-reader FROC model, Chakraborty [36-38] considered a fixed number of *FP* and *TP* marks on each subject. Then they simulated *FP* and *TP* ratings from a multi-reader ROC experiment with the use of a mixed-effect model presented by Roe and Metz [39], where modality was treated as a fixed effect and subject and reader were treated as random effects. The random effects in the mixed-

effect model allow for the existence of a correlation among ratings for the same subject evaluated by different modalities and/or different readers.

In this Chapter, we will develop a reasonable simulation model by considering different sources of correlations and reader heterogeneity in a multi-reader FROC study. In the free-response diagnostic setting, both the number of marks and the ratings of the same subject may be correlated. In the proposed model, we will treat the number of *FP* and *TP* marks on each subject as random variables and allow them to be correlated for different modalities/readers, in addition to modeling different sources of correlations for the ratings. We consider reader heterogeneity in terms of two aspects of the FROC system, namely *AUC* and λ. We will evaluate our proposed $SWITCH_{(2)}$ index, as well as the *J1* index (modified *JAFROC1*), where the statistical test based on $SWITCH_{(2)}$ was shown to have improved power under the single reader setting. We will develop inferential procedures for the FROC indices in a multi-reader design setting. Specifically, the two-sample jackknife approach will be evaluated for reader-averaged indices. We will also apply the ANOVA-based approach (DBM approach [49]) for the reader-averaged FROC indices. The proposed inferential procedures will be investigated in a simulation study.

5.2 SOURCES OF CORRELATION

Similar to the notions defined in Chapter 2, we denote the random variables on each subject indexed by s_1 evaluated by two modalities indexed by *l* and five readers indexed by *r* as follows: denote *FP* ratings for modality *l* and reader *r* as $\left\{X^{lr}_{s_1 1},...,X^{lr}_{s_1 n^{lr}_{s_1}}\right\}$, where $n^{lr}_{s_1}$ is a random variable indicating the number of *FP* marks observed on subject s_1 for modality *l* and reader *r*, where *l*=1,2; *r*=1,…5; and denote *TP* ratings for modality *l* and reader *r* as $\left\{Y^{lr}_{s_1 1},...,Y^{lr}_{m^{lr}_{s_1}}\right\}$, where $m^{lr}_{s_1}$ is a

random variable indicating the number of *TP* marks observed on subject s_1 for modality *l* and reader *r*, where *l*=1,2; *r*=1,…5. Note s_1 indexes each subject and when it refers to an actually negative subject, $m_{s_1}^{lr}$ is set to 0.

We assume that the random variables for subject s_1 follow multivariate normal distributions:

$$\begin{pmatrix} X_{s_1 1}^{11},...,X_{s_1 n_{s_1}^{11}}^{11},Y_{s_1 1}^{11},...,Y_{s_1 m_{s_1}^{11}}^{11},X_{s_1 1}^{21},...,X_{s_1 n_{s_1}^{21}}^{21},Y_{s_1 1}^{21},...,Y_{s_1 m_{s_1}^{21}}^{21}, \\ X_{s_1 1}^{12},...,X_{s_1 n_{s_1}^{12}}^{12},Y_{s_1 1}^{12},...,Y_{s_1 m_{s_1}^{12}}^{12},X_{s_1 1}^{22},...,X_{s_1 n_{s_1}^{22}}^{22},Y_{s_1 1}^{22},...,Y_{s_1 m_{s_1}^{22}}^{22}, \\ X_{s_1 1}^{13},...,X_{s_1 n_{s_1}^{13}}^{13},Y_{s_1 1}^{13},...,Y_{s_1 m_{s_1}^{13}}^{13},X_{s_1 1}^{23},...,X_{s_1 n_{s_1}^{23}}^{23},Y_{s_1 1}^{23},...,Y_{s_1 m_{s_1}^{23}}^{23}, \\ X_{s_1 1}^{14},...,X_{s_1 n_{s_1}^{14}}^{14},Y_{s_1 1}^{14},...,Y_{s_1 m_{s_1}^{14}}^{14},X_{s_1 1}^{24},...,X_{s_1 n_{s_1}^{24}}^{24},Y_{s_1 1}^{24},...,Y_{s_1 m_{s_1}^{24}}^{24}, \\ X_{s_1 1}^{15},...,X_{s_1 n_{s_1}^{15}}^{15},Y_{s_1 1}^{15},...,Y_{s_1 m_{s_1}^{15}}^{15},X_{s_1 1}^{25},...,X_{s_1 n_{s_1}^{25}}^{25},Y_{s_1 1}^{25},...,Y_{s_1 m_{s_1}^{25}}^{25} \end{pmatrix} \sim iid\ N\left(\begin{pmatrix} \mu_X^{11},...,\mu_X^{11},\mu_Y^{11},...,\mu_Y^{11},\mu_X^{21},...,\mu_X^{21},\mu_Y^{21},...,\mu_Y^{21}, \\ \mu_X^{12},...,\mu_X^{12},\mu_Y^{12},...,\mu_Y^{12},\mu_X^{22},...,\mu_X^{22},\mu_Y^{22},...,\mu_Y^{22}, \\ \mu_X^{13},...,\mu_X^{13},\mu_Y^{13},...,\mu_Y^{13},\mu_X^{23},...,\mu_X^{23},\mu_Y^{23},...,\mu_Y^{23}, \\ \mu_X^{14},...,\mu_X^{14},\mu_Y^{14},...,\mu_Y^{14},\mu_X^{24},...,\mu_X^{24},\mu_Y^{24},...,\mu_Y^{24}, \\ \mu_X^{15},...,\mu_X^{15},\mu_Y^{15},...,\mu_Y^{15},\mu_X^{25},...,\mu_X^{25},\mu_Y^{25},...,\mu_Y^{25} \end{pmatrix},\Sigma \right)$$

We assume independence for the multiple responses of different subjects. Now we consider the formulation of the variance-covariance matrix Σ for each subject and categorize the possible sources of correlations among these random variables as within- and between-subject correlations. For within-subject correlations, we consider equal correlation within each subject on replicate evaluation, regardless of by modality and/or reader.

Within-subject correlations for modality *l* and reader *r*

We consider the following within-subject correlations for subject s_1:

(w1). Correlation between *FP* ratings ($X_{s_1 i}^{lr}$) and *FP* ratings ($X_{s_1 j}^{lr}$) on each subject but on different locations for modality *l* and reader *r*. *l*=1,2; *r*=1,…,5; $i, j = 1,...,n^{lr};\ i \neq j$.

(w2). Correlation between *FP* ratings ($X_{s_1 i}^{lr}$) and *TP* ratings ($Y_{s_1 j}^{lr}$) on an actually positive subject for modality *l* and reader *r*. *l*=1,2; *r*=1,…,5; $i = 1,...,n^{lr};\ j = 1,...,m^{lr}$.

(w3). Correlation between *TP* ratings ($Y_{s_1 j}^{lr}$) and *TP* ratings ($Y_{s_1 j}^{lr}$) on an actually positive subject but on different locations for modality *l* and reader *r*. *l*=1,2; *r*=1,…,5; $i, j = 1,...,m^{lr}$; $i \neq j$.

(w4). Correlation between the number of the number of *FP* marks ($n_{s_1}^{lr}$) and *TP* marks ($m_{s_1}^{lr}$) on an actually positive subject for modality *l* and reader *r*. *l*=1,2; *r*=1,…,5.

Between-subject correlations between modality l_1 reader r_1 and modality l_2 reader r_2

We consider the following between-subject correlations for subject s_1:

(b1). Correlation between *FP* ratings ($X_{s_1 i}^{l_1 r_1}$) and *FP* ratings ($X_{s_1 j}^{l_2 r_2}$) on subject s_1 evaluated by two different modalities/readers. l_1, l_2=1,2; r_1, r_2=1,…,5; $l_1 \neq l_2 / r_1 \neq r_2$; $i = 1,...,n^{l_1 r_1}$; $j = 1,...,n^{l_2 r_2}$.

(b2). Correlation between *FP* ratings ($X_{s_1 i}^{l_1 r_1}$) and *TP* ratings ($Y_{s_1 j}^{l_2 r_2}$) on subject s_1 evaluated by two different modalities/readers. l_1, l_2=1,2; r_1, r_2=1,…,5; $l_1 \neq l_2 / r_1 \neq r_2$; $i = 1,...,n^{l_1 r_1}$; $j = 1,...,m^{l_2 r_2}$.

(b3). Correlation between *TP* ratings ($Y_{s_1 i}^{l_1 r_1}$) and *TP* ratings ($Y_{s_1 j}^{l_2 r_2}$) on subject s_1 evaluated by two different modalities/readers. l_1, l_2=1,2; r_1, r_2=1,…5; $l_1 \neq l_2 / r_1 \neq r_2$; $i = 1,...,m^{l_1 r_1}$; $j = 1,...,m^{l_2 r_2}$.

(b4). Correlation between the number of *FP* marks ($n_{s_1}^{l_1 r_1}$ and $n_{s_1}^{l_2 r_2}$) on subject s_1 evaluated by two different modalities/readers. l_1, l_2=1,2; r_1, r_2=1,…,5; $l_1 \neq l_2 / r_1 \neq r_2$.

(b5). Correlation between the number of *FP* marks ($n_{s_1}^{l_1 r_1}$) and the number of *TP* marks ($m_{s_1}^{l_2 r_2}$) on subject s_1 evaluated by two different modalities/readers. l_1, l_2=1,2; r_1, r_2=1,…,5; $l_1 \neq l_2 / r_1 \neq r_2$.

(b6). Correlation between the number of *TP* marks ($m_{s_1}^{l_1 r_1}$ and $m_{s_1}^{l_2 r_2}$) on subject s_1 evaluated by two different modalities/readers. l_1, l_2=1,2; r_1, r_2=1,…,5; $l_1 \neq l_2 / r_1 \neq r_2$.

Parameters to be considered in the simulation

We use $\rho_{0\bullet}$ to denote the within-subject correlations, $\rho_{1\bullet}$ to denote the correlations for the same subject for different modalities, $\rho_{2\bullet}$ to denote the correlations for the same subject for different readers, and $\rho_{3\bullet}$ to denote the correlations for the same subject for different modalities and different readers. Furthermore, we use a second subscript, 1 to denote the correlations between *FP* and *FP* ratings, 2 to denote the correlations between *FP* and *TP* ratings, and 3 to denote the correlations between *TP* and *TP* ratings. Therefore, we have the following:

$$\text{within-subject:}\quad \text{(w1) } \mathrm{cor}(X^{lr}_{s_1 i}, X^{lr}_{s_1 j}) = \rho_{01},\ \text{(w2) } \mathrm{cor}(X^{lr}_{s_1 i}, Y^{lr}_{s_1 j}) = \rho_{02},\ \text{(w3) } \mathrm{cor}(Y^{lr}_{s_1 i}, Y^{lr}_{s_1 j}) = \rho_{03}.$$

$$\text{between-subject:}\begin{cases} \text{(b1) } \mathrm{cor}(X^{1r}_{s_1 i}, X^{2r}_{s_1 j}) = \rho_{11},\ \text{(b1) } \mathrm{cor}(X^{1r}_{s_1 i}, Y^{2r}_{s_1 j}) = \rho_{12},\ \text{(b1) } \mathrm{cor}(Y^{1r}_{s_1 i}, Y^{2r}_{s_1 j}) = \rho_{13}. \\ \text{(b2) } \mathrm{cor}(X^{lr_1}_{s_1 i}, X^{lr_2}_{s_1 j}) = \rho_{21},\ \text{(b2) } \mathrm{cor}(X^{lr_1}_{s_1 i}, Y^{lr_2}_{s_1 j}) = \rho_{22},\ \text{(b2) } \mathrm{cor}(Y^{lr_1}_{s_1 i}, Y^{lr_2}_{s_1 j}) = \rho_{23}. \\ \text{(b3) } \mathrm{cor}(X^{1r_1}_{s_1 i}, X^{2r_2}_{s_1 j}) = \rho_{31},\ \text{(b3) } \mathrm{cor}(X^{1r_1}_{s_1 i}, Y^{2r_2}_{s_1 j}) = \rho_{32},\ \text{(b3) } \mathrm{cor}(Y^{1r_1}_{s_1 i}, Y^{2r_2}_{s_1 j}) = \rho_{33}. \end{cases}$$

For the simple scenario of $\sigma_X = \sigma_Y = \sigma$ and t=1 (one lesion on each actually positive subject), the variance-covariance matrix, Σ, is then specified as follows:

$$\Sigma = \sigma^2 \times \begin{pmatrix} \vec{\rho}_0 & \vec{\rho}_1 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 \\ \vec{\rho}_1 & \vec{\rho}_0 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 \\ \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_0 & \vec{\rho}_1 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 \\ \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_1 & \vec{\rho}_0 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 \\ \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_0 & \vec{\rho}_1 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 \\ \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_1 & \vec{\rho}_0 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 \\ \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_0 & \vec{\rho}_1 & \vec{\rho}_2 & \vec{\rho}_3 \\ \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_1 & \vec{\rho}_0 & \vec{\rho}_3 & \vec{\rho}_2 \\ \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_0 & \vec{\rho}_1 \\ \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_3 & \vec{\rho}_2 & \vec{\rho}_1 & \vec{\rho}_0 \end{pmatrix}, \textit{where } \vec{\rho}_0 = \begin{pmatrix} 1 & \rho_{01} & . & . & . & \rho_{01} & \rho_{02} \\ \rho_{01} & 1 & . & . & . & \rho_{01} & \rho_{02} \\ . & . & . & . & . & \rho_{01} & \rho_{02} \\ . & . & . & . & . & \rho_{01} & \rho_{02} \\ . & . & . & . & . & \rho_{01} & \rho_{02} \\ \rho_{01} & \rho_{01} & . & . & . & 1 & \rho_{02} \\ \rho_{02} & \rho_{02} & . & . & . & \rho_{02} & 1 \end{pmatrix}$$

$$\vec{\rho}_1=\begin{pmatrix}\rho_{11} & \rho_{11} & \cdot & \cdot & \rho_{11} & \rho_{12}\\ \cdot & \cdot & \cdot & \cdot & \rho_{11} & \rho_{12}\\ \cdot & \cdot & \cdot & \cdot & \rho_{11} & \rho_{12}\\ \cdot & \cdot & \cdot & \cdot & \rho_{11} & \rho_{12}\\ \rho_{11} & \rho_{11} & \cdot & \cdot & \rho_{11} & \rho_{12}\\ \rho_{12} & \rho_{12} & \cdot & \cdot & \rho_{12} & \rho_{13}\end{pmatrix}, \vec{\rho}_2=\begin{pmatrix}\rho_{21} & \rho_{21} & \cdot & \cdot & \rho_{21} & \rho_{22}\\ \cdot & \cdot & \cdot & \cdot & \rho_{21} & \rho_{22}\\ \cdot & \cdot & \cdot & \cdot & \rho_{21} & \rho_{22}\\ \cdot & \cdot & \cdot & \cdot & \rho_{21} & \rho_{22}\\ \rho_{21} & \rho_{21} & \cdot & \cdot & \rho_{21} & \rho_{22}\\ \rho_{22} & \rho_{22} & \cdot & \cdot & \rho_{22} & \rho_{23}\end{pmatrix}, \vec{\rho}_3=\begin{pmatrix}\rho_{31} & \rho_{31} & \cdot & \cdot & \rho_{31} & \rho_{32}\\ \cdot & \cdot & \cdot & \cdot & \rho_{31} & \rho_{32}\\ \cdot & \cdot & \cdot & \cdot & \rho_{31} & \rho_{32}\\ \cdot & \cdot & \cdot & \cdot & \rho_{31} & \rho_{32}\\ \rho_{31} & \rho_{31} & \cdot & \cdot & \rho_{31} & \rho_{32}\\ \rho_{32} & \rho_{32} & \cdot & \cdot & \rho_{32} & \rho_{33}\end{pmatrix}.$$

$\vec{\rho}_0$ represents the within-subject correlation matrix; $\vec{\rho}_1$ represents the correlation matrix between different modalities; $\vec{\rho}_2$ represents the matrix between different readers; and $\vec{\rho}_3$ represents the correlation matrix between different modalities and different readers. In the simulation, we consider $\rho_{01}=0.7, \rho_{02}=0.1$; $\rho_{11}=0.6, \rho_{12}=0.1, \rho_{13}=0.6$; $\rho_{21}=\rho_{31}=0.4$, $\rho_{22}=\rho_{32}=0.1$, $\rho_{23}=\rho_{33}=0.4$.

To incorporate the discussed correlation structure to the number of marks, we assume that the numbers of marks for the two modalities and five readers have the following distributions:

$$n_{s_1}^{lr} \sim Poisson\left(\lambda^{lr}\right), \quad m_{s_1}^{lr} \sim Binomial\left(t,\upsilon\right)$$

(w4) $\text{cor}(n_{s_1}^{lr}, m_{s_1}^{lr}) = \rho_{n0}$,

(b4) $\text{cor}(n_{s_1}^{1r}, n_{s_1}^{2r}) = \text{cor}(n_{s_1}^{lr_1}, n_{s_1}^{lr_2}) = \text{cor}(n_{s_1}^{1r_1}, n_{s_1}^{2r_2}) = \rho_{n1}$,

(b5) $\text{cor}(n_{s_1}^{1r}, m_{s_1}^{2r}) = \text{cor}(n_{s_1}^{lr_1}, m_{s_1}^{lr_2}) = \text{cor}(n_{s_1}^{1r_1}, m_{s_1}^{2r_2}) = \rho_{n2}$,

(b6) $\text{cor}(m_{s_1}^{1r}, m_{s_1}^{2r}) = \text{cor}(m_{s_1}^{lr_1}, m_{s_1}^{lr_2}) = \text{cor}(m_{s_1}^{1r_2}, m_{s_1}^{2r_2}) = \rho_{n3}$.

where ρ_{n0} represents the within-subject correlation; ρ_{n1} represents the correlation between two numbers of *FP* marks; ρ_{n2} represents the correlation between the number of *FP* marks and the number of *TP* marks; and ρ_{n3} represents the correlation between two different numbers of *TP* marks. The between-subject correlations between the numbers are assumed to be equal for different modalities and/or readers in (b4), (b5) and (b6).

To impose the correlation structure for multivariate Poisson and/or binomial distributions, we first simulate scores from the following multivariate normal distribution:

$$\begin{pmatrix} n_{score}^{11}, & m_{score}^{11}, & n_{score}^{21}, & m_{score}^{21}, \\ n_{score}^{12}, & m_{score}^{12}, & n_{score}^{22}, & m_{score}^{22}, \\ n_{score}^{13}, & m_{score}^{13}, & n_{score}^{23}, & m_{score}^{23}, \\ n_{score}^{14}, & m_{score}^{14}, & n_{score}^{24}, & m_{score}^{24}, \\ n_{score}^{15}, & m_{score}^{15}, & n_{score}^{25}, & m_{score}^{25}, \end{pmatrix} \sim N\left(0, \begin{pmatrix} \vec{\rho}_0' & \vec{\rho}_1' & \vec{\rho}_1' & \vec{\rho}_1' & \vec{\rho}_1' \\ \vec{\rho}_1' & \vec{\rho}_0' & \vec{\rho}_1' & \vec{\rho}_1' & \vec{\rho}_1' \\ \vec{\rho}_1' & \vec{\rho}_1' & \vec{\rho}_0' & \vec{\rho}_1' & \vec{\rho}_1' \\ \vec{\rho}_1' & \vec{\rho}_1' & \vec{\rho}_1' & \vec{\rho}_0' & \vec{\rho}_1' \\ \vec{\rho}_1' & \vec{\rho}_1' & \vec{\rho}_1' & \vec{\rho}_1' & \vec{\rho}_0' \end{pmatrix}\right)$$

where $\vec{\rho}_0' = \begin{pmatrix} 1 & \rho_{n0}' & \rho_{n1}' & \rho_{n2}' \\ \rho_{n0}' & 1 & \rho_{n2}' & \rho_{n3}' \\ \rho_{n1}' & \rho_{n2}' & 1 & \rho_{n0}' \\ \rho_{n2}' & \rho_{n3}' & \rho_{n0}' & 1 \end{pmatrix}$ can be viewed as the within-modality correlation matrix and

$\vec{\rho}_1' = \begin{pmatrix} \rho_{n1}' & \rho_{n2}' & \rho_{n1}' & \rho_{n2}' \\ \rho_{n2}' & \rho_{n3}' & \rho_{n2}' & \rho_{n3}' \\ \rho_{n1}' & \rho_{n2}' & \rho_{n1}' & \rho_{n2}' \\ \rho_{n2}' & \rho_{n3}' & \rho_{n2}' & \rho_{n3}' \end{pmatrix}$ can be viewed as the between-reader correlation matrix.

The simulated scores are then truncated into the desired Poisson or binomial distributions based on the following transformation of $\left(n_{score}^{lr},\ m_{score}^{lr}\right) \to \left(n^{lr},\ m^{lr}\right)$:

$$if\ \ \Phi^{-1}(poisson(\lambda^{lr}, i-1)) < n_{score}^{lr} \le \Phi^{-1}(poisson(\lambda^{lr}, i)),\ \ n^{lr} = i$$
$$if\ \ \Phi^{-1}(binomial(t, \upsilon, i-1)) < m_{score}^{lr} \le \Phi^{-1}(binomial(t, \upsilon, i)),\ \ m^{lr} = i$$

However, the correlation structure is not identically preserved after transformation. In the simulation, we consider $\rho_{n0}' = \rho_{n1}' = \rho_{n2}' = \rho_{n3}' = 0.65$. A separate simulation study shows that the corresponding correlations for the transformed numbers of marks are $\rho_{n0} = \rho_{n2} = 0.40$, $\rho_{n1} = 0.55$, and $\rho_{n3} = 0.44$.

In the simulation model, we assume independence between the ratings and the number of marks of each subject.

5.3 READER HETEROGENEITY

In this section, we consider reader heterogeneity in terms of the degree of separation between *FP* and *TP* ratings (*AUC*) and the average number of *FP* marks (λ) identified per subject for an FROC system (modality).

Reader heterogeneity in term of *AUC*

We denote the degree of separation between *FP* and *TP* ratings for two modalities and five readers as AUC^{lr} , l=1,2; r=1,…,5; the reader-averaged *AUC* for two modalities as $\overline{AUC}^{l\bullet} = \frac{\sum_{r=1}^{5} AUC^{lr}}{5}$, l=1,2; and the reader-specific *AUC* differences between two modalities as $d_{AUC}^{r} = AUC^{1r} - AUC^{2r}$, r=1,…,5. We define the difference between two modalities (FROC systems) with regard to reader-averaged *AUC* as $\overline{d_{AUC}^{\bullet}} = \overline{AUC}^{1\bullet} - \overline{AUC}^{2\bullet}$ (specifically, $\overline{d_{AUC}^{\bullet}} = 0$ indicates two identical systems and $\overline{d_{AUC}^{\bullet}} \neq 0$ indicates two different systems). Equivalently, it is also the reader-averaged *AUC* difference between two modalities, $\overline{d_{AUC}^{\bullet}} = \frac{\sum_{r=1}^{5}(AUC^{1r} - AUC^{2r})}{5} = \frac{\sum_{r=1}^{5} d_{AUC}^{r}}{5}$. We further denote the reader heterogeneity of the reader-specific *AUC* differences as $\gamma_{AUC} = \frac{\sum_{r=1}^{5}(d_{AUC}^{r} - \overline{d_{AUC}^{\bullet}})^2}{5}$ and denote reader-specific deviates as $\left\{\Delta_{AUC}^{r}\right\}$, where $\Delta_{AUC}^{r} = d_{AUC}^{r} - \overline{d_{AUC}^{\bullet}}$. In the simulation, we consider $\sum_{r=1}^{5} \Delta_{AUC}^{r} = 0$ and $\left\{\Delta_{AUC}^{r}\right\}$ as constant values.

Parameters to be considered in simulations for reader-specific *AUC*s

In the simulation, we will consider reader-averaged $\overline{AUC}^{l\bullet} = 0.6, 0.7, 0.8; l = 1, 2$ for two identical or two different systems. We consider reader heterogeneity due to AUC as $\gamma_{AUC} = 0.004$. We first obtain 1/6th to 5/6th percentiles of a standard normal distribution (-.967, -.431, 0, .431, .967) which all sum to 0. Then we scale their heterogeneity to γ_{AUC} and add the scaled numbers to $\overline{AUC}^{l\bullet}$ to generate modality- and reader-specific AUC^{lr} s. For example, for two FROC systems with reader-averaged $\overline{AUC}^{1\bullet} = \overline{AUC}^{2\bullet} = 0.7$ and two FROC systems with reader-averaged $\overline{AUC}^{1\bullet} = 0.8$ and $\overline{AUC}^{2\bullet} = 0.7$, for all AUC^{lr} s, l=1, 2 and r=1,…5, we have

$\overline{AUC}^{1\bullet} = \overline{AUC}^{2\bullet} = 0.7;\ \overline{d^{\bullet}_{AUC}} = 0$	$\overline{AUC}^{1\bullet} = 0.8,\ \overline{AUC}^{2\bullet} = 0.7;\ \overline{d^{\bullet}_{AUC}} = 0.2$
$AUC^{11} = 0.659\ \ AUC^{21} = 0.741$	$AUC^{11} = 0.759\ \ AUC^{21} = 0.741$
$AUC^{12} = 0.682\ \ AUC^{22} = 0.718$	$AUC^{12} = 0.782\ \ AUC^{22} = 0.718$
$AUC^{13} = 0.700\ \ AUC^{23} = 0.700$	$AUC^{13} = 0.800\ \ AUC^{23} = 0.700$
$AUC^{14} = 0.718\ \ AUC^{24} = 0.682$	$AUC^{14} = 0.818\ \ AUC^{24} = 0.682$
$AUC^{15} = 0.741\ \ AUC^{25} = 0.659$	$AUC^{15} = 0.841\ \ AUC^{25} = 0.659$

Reader heterogeneity in term of λ

Similarly, we denote the average number of *FP* marks for two modalities and five readers as λ^{lr}, l=1,2; r=1,…,5; the reader-averaged λ for two modalities as $\overline{\lambda}^{l\bullet} = \frac{\sum_{r=1}^{5} \lambda^{lr}}{5}$, l=1,2; and the reader-specific λ differences between two modalities as $d^{r}_{\lambda} = \lambda^{1r} - \lambda^{2r}$, r=1,…,5. We define the difference between two modalities (FROC systems) with regard to the reader-averaged λ as $\overline{d^{\bullet}_{\lambda}} = \overline{\lambda}^{1\bullet} - \overline{\lambda}^{2\bullet}$ and equivalently, it is the reader-averaged λ difference between two modalities,

$\overline{d_\lambda^\bullet} = \frac{\sum_{r=1}^{5}(\lambda^{1r} - \lambda^{2r})}{5} = \frac{\sum_{r=1}^{5} d_\lambda^r}{5}$. We further denote the reader heterogeneity of the reader-specific λ differences as $\gamma_\lambda = \frac{\sum_{r=1}^{5}(d_\lambda^r - \overline{d_\lambda^\bullet})^2}{5}$ and denote reader-specific deviates as $\left\{\Delta_\lambda^r\right\}$, where $\Delta_\lambda^r = d_\lambda^r - \overline{d_\lambda^\bullet}$.

In the simulation, we consider $\sum_{r=1}^{5}\Delta_\lambda^r = 0$ and $\left\{\Delta_\lambda^r\right\}$ as constant values.

Parameters to be considered in simulation for reader-specific λ's

In the simulation, we will consider reader-averaged $\overline{\lambda}^{l\bullet} = 0.75, 1.00, 1.25; l = 1, 2$ for two identical or two different systems. We consider reader heterogeneity due to λ as $\gamma_\lambda = \gamma_{AUC} = 0.004$. Similarly, we add the scaled numbers to $\overline{\lambda}^{l\bullet}$ to generate modality- and reader-specific λ^{lr}s. For example, for two FROC systems with reader-averaged $\overline{\lambda}^{1\bullet} = \overline{\lambda}^{2\bullet} = 1.00$ and two FROC systems with reader-averaged $\overline{\lambda}^{1\bullet} = 1.25$ and $\overline{\lambda}^{2\bullet} = 1.00$, for all λ^{lr}s, *l*=1,2 and *r*=1,…,5, we have

$\overline{\lambda}^{1\bullet} = \overline{\lambda}^{2\bullet} = 1.00;\ \overline{d^\bullet} = 0$	$\overline{\lambda}^{1\bullet} = 1.25,\ \overline{\lambda}^{2\bullet} = 1.00;\ \overline{d^\bullet} = 1.0$
$\lambda^{11} = 0.959\ \ \lambda^{21} = 1.041$	$\lambda^{11} = 1.209\ \ \lambda^{21} = 1.041$
$\lambda^{12} = 0.982\ \ \lambda^{22} = 1.018$	$\lambda^{12} = 1.232\ \ \lambda^{22} = 1.018$
$\lambda^{13} = 1.000\ \ \lambda^{23} = 1.000$	$\lambda^{13} = 1.250\ \ \lambda^{23} = 1.000$
$\lambda^{14} = 1.018\ \ \lambda^{24} = 0.982$	$\lambda^{14} = 1.268\ \ \lambda^{24} = 0.982$
$\lambda^{15} = 1.041\ \ \lambda^{25} = 0.959$	$\lambda^{15} = 1.291\ \ \lambda^{25} = 0.959$

5.4 STATISTICAL INFERENCE

For comparing the overall diagnostic performance of two FROC systems in a multi-reader study, we propose to use the reader-averaged difference between the modality-specific FROC indices. For the single reader setting in (3.5), the estimate of *J1* (modified *JAFROC1*) can be written as

$$\widehat{J1} = \frac{1}{(S_0 + S_t) \times S_t} \sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} \tilde{\psi}_{J1} \left(\{x_{s_1 c_1}\}_{c_1=1}^{n_{s_1}}, \left\{ \{y_{s c_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}} \right\} \right)$$

We denote the number of readers as *R*. For the reader-averaged difference between the modality-specific *J1* indices, we have

$$\overline{\hat{D}^{\bullet}_{J1}} = \frac{\sum_{r=1}^{R} \sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} (\tilde{\psi}^{1r}_{s_1 s} - \tilde{\psi}^{2r}_{s_1 s})}{R \times (S_0 + S_t) \times S_t} = \frac{\sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} \tilde{\omega}_{s_1 s}}{S_0 \times S_t}, \ where\ \tilde{w}_{s_1 s} = \frac{\sum_{r=1}^{R} (\tilde{\psi}^{1r}_{s_1 s} - \tilde{\psi}^{2r}_{s_1 s})}{R}, and\ \ \tilde{\psi} = \tilde{\psi}_{J1}.$$

For the single reader setting in (4.2), the estimate of $SWITCH_{(2)}$ can be written as

$$\widehat{SWITCH}_{(2)} = \frac{1}{S_0 \times S_t} \sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \tilde{\psi}_{SWITCH_{(2)}} \left(\left\{ \{x^0_{s' c_1}\}_{c_1=1}^{n^0_{s'}}, \{x^t_{s c_2}\}_{c_2=1}^{n^t_s} \right\}, \{y_{s c_2}\}_{c_2=1}^{m_s} \right)$$

The reader-averaged difference between the modality-specific $SWITCH_{(2)}$ indices is

$$\overline{\hat{D}^{\bullet}_{SWITCH_{(2)}}} = \frac{\sum_{r=1}^{R} \sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} (\tilde{\psi}^{1r}_{s's} - \tilde{\psi}^{2r}_{s's})}{R \times S_0 \times S_t} = \frac{\sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \tilde{\omega}_{s's}}{S_0 \times S_t}, \ where\ \tilde{w}_{s's} = \frac{\sum_{r=1}^{R} (\tilde{\psi}^{1r}_{s's} - \tilde{\psi}^{2r}_{s's})}{R}, and\ \ \tilde{\psi} = \tilde{\psi}_{SWITCH_{(2)}}.$$

The estimators $\overline{\hat{D}^{\bullet}_{J1}}$ and $\overline{\hat{D}^{\bullet}_{SWITCH_{(2)}}}$ can still be treated as a two-sample U-statistic, which permits the development of a closed form expression for the two-sample jackknife variance when the re-sampling techniques consider subject as a sampling unit. For the reader-averaged *J1* index $\overline{\hat{D}^{\bullet}_{J1}}$, removing an actually positive subject influences both a row and a column of the reader-averaged $\left\{\tilde{\psi}_{s_1 s},\ \tilde{\psi} = \tilde{\psi}_{J1}\right\}$ matrix, namely the $\left\{\tilde{\omega}_{s_1 s}\right\}$ matrix that represents the actually positive subject. Similar to the approach of deriving the closed form two-sample jackknife

variance in (3.16), we replace the $\tilde{\psi}_{J1}$ functions with the reader-averaged $\tilde{\psi}_{J1}$ functions, namely $\tilde{\omega}$ functions. We also denote the $\tilde{\omega}$ functions for comparing actually negative and actually positive subjects as $\left\{\tilde{\omega}^0_{s_1 s}\right\}$, where $s_1 = 1,...,S_0; s = 1,...,S_t$ and for comparing actually positive and actually positive subjects as $\left\{\tilde{\omega}^t_{s_1 s}\right\}$ where $s_1 = S_0 + 1,...,S_0 + S_t; s = 1,...,S_t$.

$$\hat{V}_{\text{2s-jackk}}\left(\overline{\hat{D}^{\bullet}_{J1}}\right) = \frac{\sum_{s_1=1}^{S_0}(\hat{A}^{s_1}_{negative;J2} - \hat{A}^1)^2}{S_0(S_0-1)} + \frac{\sum_{s=1}^{S_t}(\hat{A}^{s}_{positive;J2} - \hat{A}^2)^2}{S_t(S_t-1)} \tag{5.1}$$

where $\hat{A}^{s_1}_{negative;J2} = \frac{\tilde{\omega}^0_{\bullet\bullet} + \tilde{\omega}^t_{\bullet\bullet}}{(S_0+S_t)(S_0+S_t-1)} + \frac{(S_0-1)\tilde{\omega}^0_{s_1\bullet}}{(S_0+S_t-1)S_t}$, $\hat{A}^{s}_{positive;J2} = \frac{-(\tilde{\omega}^0_{\bullet\bullet} + \tilde{\omega}^t_{\bullet\bullet})}{(S_0+S_t)(S_0+S_t-1)} + \frac{\tilde{\omega}^0_{\bullet s} + \tilde{\omega}^t_{s\bullet} + \tilde{\omega}^t_{\bullet s} - \tilde{\omega}^t_{ss}}{S_0+S_t-1}$,

$\hat{A}^1 = \frac{\tilde{\omega}^0_{\bullet\bullet} + \tilde{\omega}^t_{\bullet\bullet}}{(S_0+S_t)(S_0+S_t-1)} + \frac{(S_0-1)\tilde{\omega}^0_{\bullet\bullet}}{(S_0+S_t-1)S_0S_t}$, and $\hat{A}^2 = \frac{-(\tilde{\omega}^0_{\bullet\bullet} + \tilde{\omega}^t_{\bullet\bullet})}{(S_0+S_t)(S_0+S_t-1)} + \frac{\tilde{\omega}^0_{\bullet\bullet} + 2\tilde{\omega}^t_{\bullet\bullet} - \sum_{s=1}^{S_t}\tilde{\omega}^t_{ss}}{(S_0+S_t-1)S_t}$.

For the reader-averaged *SWITCH(2)* index $\overline{\hat{D}^{\bullet}_{SWITCH_{(2)}}}$, removing an actually positive subject only influences a column of the reader-averaged $\left\{\tilde{\psi}_{s's},\ \tilde{\psi} = \tilde{\psi}_{SWITCH_{(2)}}\right\}$ matrix, namely $\{\tilde{\omega}_{s's}\}$ matrix that represents the actually positive subject. Similarly we can construct the two-sample jackknife variance for *SWITCH(2)* using (2.10) by replacing the $\tilde{\psi}_{SWITCH_{(2)}}$ functions with the reader-averaged $\tilde{\psi}_{SWITCH_{(2)}}$ functions, namely $\tilde{\omega}$ functions:

$$\hat{V}_{\text{2s-jackk}}\left(\overline{\hat{D}^{\bullet}_{SWITCH_{(2)}}}\right) = \frac{\sum_{s'=1}^{S_0}\left(\overline{\tilde{\omega}_{s'\bullet}} - \overline{\tilde{\omega}_{\bullet\bullet}}\right)^2}{S_0 \times (S_0-1)} + \frac{\sum_{s=1}^{S_t}\left(\overline{\tilde{\omega}_{\bullet s}} - \overline{\tilde{\omega}_{\bullet\bullet}}\right)^2}{S_t \times (S_t-1)} \tag{5.2}$$

For the analysis of multi-reader setting, the most common approach to test such hypotheses is to apply an ANOVA-based approach in a mixed-effect model for the pseudo-values of the

indices $\overline{\hat{D}^{\bullet}_{J1}}$ and $\overline{\hat{D}^{\bullet}_{SWITCH_{(2)}}}$ (the DBM approach [49]) and the closed form solution for the variance of the average pseudo-values for fixed reader analysis is shown in Bandos *et a*l [58] as:

$$\widehat{V_{DBM}\left(\overline{\hat{D}^{\bullet}_{\bullet;pv}}\right)} = \hat{V}_{1\text{s-jackk}}(\overline{\hat{D}^{\bullet}})$$

where $\overline{\hat{D}^{\bullet}_{\bullet;pv}}$ is the reader-averaged difference of the pseudo-values of the FROC indices and $\hat{V}_{1\text{s-jackk}}(\overline{\hat{D}^{\bullet}})$ is the one-sample jackknife variance estimator for the difference between the reader-averaged index.

Using the closed form expressions of the two-sample or ANOVA-based variance formula of (5.1) and (5.2), we propose an asymptotic procedure with the following test statistics for comparing two FROC diagnostic systems: $\dfrac{\overline{\hat{D}^{\bullet}}}{\sqrt{\hat{V}_{2\text{s-jackk}}\left(\overline{\hat{D}^{\bullet}}\right)}}$ or $\dfrac{\overline{\hat{D}^{\bullet}_{\bullet;pv}}}{\sqrt{\widehat{V_{DBM}\left(\overline{\hat{D}^{\bullet}_{\bullet;pv}}\right)}}} \sim N(0,1)$.

5.5 SIMULATION RESULTS

In the proposed model, we evaluate reader-averaged *J1* ($\overline{\hat{D}^{\bullet}_{J1}}$) and *SWITCH*$_{(2)}$ ($\overline{\hat{D}^{\bullet}_{SWITCH_{(2)}}}$), where *SWITCH*$_{(2)}$ on average results in a more powerful statistical test than *J1* for detecting system differences with regard to AUC and λ, respectively, for the single reader setting. We assess their statistical power in detecting the single difference between the two systems, namely, with regard to reader-averaged AUC or $\overline{AUC}^{l\bullet}$ (scenario 1) and with regard to reader-averaged λ or $\overline{\lambda}^{l\bullet}$ (scenario 2). In addition, we construct two additional scenarios that allow the two systems to be different for the two parameters simultaneously. Specifically, for scenario 3 we consider two different systems where one system has a larger $\overline{AUC}^{1\bullet}$ and a larger $\overline{\lambda}^{1\bullet}$, while the other has a

smaller $\overline{AUC}^{2\bullet}$ and a smaller $\overline{\lambda}^{2\bullet}$. For scenario 4, we consider two different systems where one system has a larger $\overline{AUC}^{1\bullet}$ and a smaller $\overline{\lambda}^{1\bullet}$, while the other has a smaller $\overline{AUC}^{2\bullet}$ and a larger $\overline{\lambda}^{2\bullet}$. We exhibit the simulation results in Table 5.1-5.8.

For the simulated results, we list the standard errors and the type I error rates for testing two identical systems, as well as the standard errors and the statistical powers for testing two different systems. For the two inferential procedures, one can observe that the estimated two-sample jackknife variances and the ANOVA-based variances are close to the variances of sample realization for testing both identical and separated hypotheses in all considered four scenarios. The two-sample jackknife variances are uniformly smaller than or equal to the ANOVA-based variances [49] for both indices. In all simulated scenarios, the type I error rates for two indices using the two-sample jackknife approach range from 0.053 to 0.063. The type I error rates for two indices the ANOVA-based approach range from 0.050 to 0.060.

For detecting the system difference with regard to $\overline{AUC}^{l\bullet}$ (scenario 1) and with regard to $\overline{\lambda}^{l\bullet}$ (scenario 2), one can observe from Table 5.2 and Table 5.4 that the statistical test based on $SWITCH_{(2)}$ tends to have greater power, similar to what was also observed for the single reader-setting. For detecting the system difference in scenario 3 (Table 5.6), the statistical test based on *J1* tends to have greater power to detect such reader-averaged system differences. For the simulation results in scenario 4 (Table 5.8), the statistical test based on $SWITCH_{(2)}$ has greater power to detect such reader-averaged system differences in all simulated scenarios.

5.6 SUMMARY

In this Chapter, we proposed a reasonable simulation model for the multi-reader FROC data. It incorporates different sources of correlations among the multiple responses on the same subject evaluated by different modalities and different readers. Similar to the multi-reader model in Chakraborty [36-38], we allowed the extra consideration of the numbers on the same subject to be correlated for different modalities/readers. In the proposed model, we considered reader heterogeneity in term of two aspects of the FROC system, namely $\overline{AUC}^{l\bullet}$ and $\bar{\lambda}^{l\bullet}$.

Inferential procedures were developed for reader-averaged *J1* index (modified *JAFROC1*) and $SWITCH_{(2)}$ index. The two-sample jackknife approach is a natural extension of the previous proposed method for the single reader setting. We also applied the ANOVA-based approach [49] for the reader-averaged FROC indices. We evaluated the properties of the statistical tests in the proposed model. In the simulation, the type I error rates for the two reader-averaged indices are both close to the nominal value of 0.05, for two inferential approaches.

For the power of the statistical tests based on the reader-averaged indices, similar patterns are observed for both the two-sample jackknife approach and the ANOVA-based approach. Similar to the simulation results in Chapter 4, for detecting system difference with regard to $\overline{AUC}^{l\bullet}$ (scenario 1) or with regard to $\bar{\lambda}^{l\bullet}$ (scenario 2), the statistical test based on $SWITCH_{(2)}$ tends to have greater power. We further consider two FROC systems differing in two parameters simultaneously. For scenario 3, the statistical test based on *J1* tends to have greater power to detect such difference in two systems. For scenario 4, the statistical test based on $SWITCH_{(2)}$ has greater power to detect such differences in two systems in all simulated scenarios.

Table 5.1 Estimated type I error rates for normal distributions for Scenario 1.

			$\overline{AUC}^{1\bullet} = \overline{AUC}^{2\bullet} = 0.6$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Type I of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Type I of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0013	0.0014 (0.0014)	0.061(0.057)	0.0017	0.0017 (0.0018)	0.059(0.056)
		0.7	0.0013	0.0013 (0.0014)	0.060(0.057)	0.0020	0.0020 (0.0021)	0.059(0.057)
		0.9	0.0011	0.0011 (0.0011)	0.058(0.055)	0.0020	0.0021 (0.0021)	0.053(0.051)
			$\overline{AUC}^{1\bullet} = \overline{AUC}^{2\bullet} = 0.7$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Type I of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Type I of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0014	0.0015 (0.0015)	0.060(0.056)	0.0017	0.0018 (0.0018)	0.058(0.054)
		0.7	0.0014	0.0014 (0.0014)	0.061(0.057)	0.0020	0.0020 (0.0020)	0.061(0.056)
		0.9	0.0010	0.0010 (0.0011)	0.060(0.056)	0.0019	0.0019 (0.0019)	0.057(0.054)
			$\overline{AUC}^{1\bullet} = \overline{AUC}^{2\bullet} = 0.8$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Type I of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Type I of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0016	0.0016 (0.0017)	0.063(0.059)	0.0018	0.0018 (0.0019)	0.063(0.060)
		0.7	0.0015	0.0015 (0.0015)	0.060(0.056)	0.0019	0.0019 (0.0020)	0.060(0.057)
		0.9	0.0009	0.0009 (0.0010)	0.057(0.053)	0.0016	0.0016 (0.0017)	0.059(0.055)

Estimated type I error rates are obtained for Scenario 1 for $\overline{\lambda}^{1\bullet} = \overline{\lambda}^{2\bullet} = 1.0$.

Table 5.2 Estimated power for detecting system difference with regard to Scenario 1.

			$\overline{AUC}^{1\bullet} = 0.7$ versus $\overline{AUC}^{2\bullet} = 0.6$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Power of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Power of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0015	0.0015 (0.0015)	0.154(0.144)	0.0018	0.0018 (0.0019)	0.218(0.212)
		0.7	0.0014	0.0014 (0.0015)	0.252(0.239)	0.0021	0.0021 (0.0021)	0.311(0.303)
		0.9	0.0011	0.0011 (0.0012)	0.445(0.431)	0.0020	0.0020 (0.0021)	0.444(0.434)
			$\overline{AUC}^{1\bullet} = 0.8$ versus $\overline{AUC}^{2\bullet} = 0.6$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Power of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Power of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0017	0.0017 (0.0018)	0.380(0.364)	0.0021	0.0021 (0.0022)	0.561(0.552)
		0.7	0.0017	0.0017 (0.0017)	0.639(0.623)	0.0023	0.0023 (0.0024)	0.774(0.765)
		0.9	0.0013	0.0013 (0.0013)	0.906(0.898)	0.0021	0.0022 (0.0022)	0.935(0.932)
			$\overline{AUC}^{1\bullet} = 0.8$ versus $\overline{AUC}^{2\bullet} = 0.7$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Power of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Power of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0016	0.0016 (0.0016)	0.156(0.146)	0.0019	0.0019 (0.0020)	0.216(0.209)
		0.7	0.0015	0.0015 (0.0016)	0.256(0.244)	0.0020	0.0021 (0.0021)	0.313(0.304)
		0.9	0.0011	0.0011 (0.0011)	0.489(0.475)	0.0018	0.0019 (0.0019)	0.467(0.459)

Estimated powers are obtained for Scenario 1 for $\overline{\lambda}^{1\bullet} = \overline{\lambda}^{2\bullet} = 1.0$.

Table 5.3 Estimated type I error rates for normal distributions for Scenario 2.

			$\overline{\lambda}^{1\bullet} = \overline{\lambda}^{2\bullet} = 0.75$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Type I of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Type I of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0015	0.0015 (0.0015)	0.063(0.058)	0.0021	0.0021 (0.0021)	0.062(0.058)
		0.7	0.0013	0.0013 (0.0014)	0.063(0.060)	0.0021	0.0020 (0.0021)	0.061(0.058)
		0.9	0.0008	0.0008 (0.0009)	0.060(0.056)	0.0016	0.0017 (0.0017)	0.057(0.054)
			$\overline{\lambda}^{1\bullet} = \overline{\lambda}^{2\bullet} = 1.00$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Type I of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Type I of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0016	0.0016 (0.0016)	0.060(0.057)	0.0018	0.0018 (0.0019)	0.062(0.059)
		0.7	0.0015	0.0015 (0.0015)	0.062(0.059)	0.0019	0.0019 (0.0020)	0.060(0.057)
		0.9	0.0010	0.0010 (0.0010)	0.060(0.057)	0.0016	0.0016 (0.0017)	0.058(0.055)
			$\overline{\lambda}^{1\bullet} = \overline{\lambda}^{2\bullet} = 1.25$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Type I of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Type I of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0017	0.0017 (0.0018)	0.060(0.057)	0.0016	0.0016 (0.0017)	0.063(0.060)
		0.7	0.0016	0.0016 (0.0016)	0.059(0.055)	0.0017	0.0018 (0.0018)	0.061(0.058)
		0.9	0.0010	0.0010 (0.0011)	0.054(0.051)	0.0015	0.0015 (0.0016)	0.055(0.052)

Estimated type I error rates are obtained for Scenario 2 for $\overline{AUC}^{1\bullet} = \overline{AUC}^{2\bullet} = 0.8$.

Table 5.4 Estimated power for detecting system difference with regard to Scenario 2.

			$\overline{\lambda}^{1\bullet} = 1.00$ versus $\overline{\lambda}^{2\bullet} = 0.75$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Power of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Power of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0016	0.0016 (0.0016)	0.196(0.188)	0.0020	0.0020 (0.0020)	0.321(0.313)
		0.7	0.0014	0.0014 (0.0015)	0.180(0.173)	0.0020	0.0020 (0.0021)	0.355(0.348)
		0.9	0.0009	0.0009 (0.0009)	0.190(0.181)	0.0016	0.0017 (0.0017)	0.425(0.416)
			$\overline{\lambda}^{1\bullet} = 1.25$ versus $\overline{\lambda}^{2\bullet} = 0.75$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Power of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Power of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0017	0.0017 (0.0017)	0.466(0.451)	0.0019	0.0019 (0.0020)	0.721(0.711)
		0.7	0.0015	0.0015 (0.0016)	0.401(0.390)	0.0020	0.0020 (0.0020)	0.774(0.765)
		0.9	0.0010	0.0010 (0.0010)	0.465(0.454)	0.0016	0.0017 (0.0017)	0.874(0.869)
			$\overline{\lambda}^{1\bullet} = 1.25$ versus $\overline{\lambda}^{2\bullet} = 1.00$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Power of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Power of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0016	0.0017 (0.0017)	0.140(0.133)	0.0017	0.0017 (0.0018)	0.201(0.196)
		0.7	0.0015	0.0015 (0.0016)	0.138(0.132)	0.0018	0.0019 (0.0019)	0.247(0.240)
		0.9	0.0010	0.0010 (0.0010)	0.149(0.143)	0.0016	0.0016 (0.0016)	0.317(0.309)

Estimated powers are obtained for Scenario 2 for $\overline{AUC}^{1\bullet} = \overline{AUC}^{2\bullet} = 0.8$.

Table 5.5 Estimated type I error rates for normal distributions for Scenario 3.

			$\overline{AUC}^{l\bullet} = 0.6$ & $\bar{\lambda}^{l\bullet} = 0.75$, l=1,2.					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Type I of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Type I of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0013	0.0013 (0.0013)	0.061(0.056)	0.0020	0.0020 (0.0020)	0.057(0.053)
		0.7	0.0013	0.0013 (0.0013)	0.059(0.056)	0.0022	0.0023 (0.0023)	0.060(0.056)
		0.9	0.0010	0.0010 (0.0010)	0.058(0.054)	0.0021	0.0022 (0.0023)	0.053(0.050)
			$\overline{AUC}^{l\bullet} = 0.7$ & $\bar{\lambda}^{l\bullet} = 1.00$, l=1,2.					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Type I of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Type I of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0014	0.0015 (0.0017)	0.060(0.056)	0.0017	0.0018 (0.0018)	0.058(0.054)
		0.7	0.0014	0.0014 (0.0017)	0.061(0.057)	0.0020	0.0020 (0.0020)	0.061(0.056)
		0.9	0.0010	0.0010 (0.0013)	0.060(0.056)	0.0019	0.0019 (0.0019)	0.057(0.054)
			$\overline{AUC}^{l\bullet} = 0.8$ & $\bar{\lambda}^{l\bullet} = 1.25$, l=1,2.					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Type I of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Type I of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0017	0.0017 (0.0018)	0.060(0.057)	0.0016	0.0016 (0.0017)	0.063(0.060)
		0.7	0.0016	0.0016 (0.0016)	0.059(0.055)	0.0017	0.0018 (0.0018)	0.061(0.058)
		0.9	0.0010	0.0010 (0.0011)	0.054(0.051)	0.0015	0.0015 (0.0016)	0.055(0.052)

Table 5.6 Estimated power for detecting system difference with regard to Scenario 3.

			$\overline{AUC}^{1\bullet} = 0.7$ & $\bar{\lambda}^{1\bullet} = 1.00$ versus $\overline{AUC}^{2\bullet} = 0.6$ & $\bar{\lambda}^{2\bullet} = 0.75$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Power of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Power of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0015	0.0015 (0.0015)	0.090(0.086)	0.0020	0.0020 (0.0021)	0.091(0.087)
		0.7	0.0014	0.0014 (0.0015)	0.066(0.062)	0.0022	0.0022 (0.0023)	0.073(0.069)
		0.9	0.0010	0.0011 (0.0011)	0.069(0.064)	0.0021	0.0022 (0.0022)	0.055(0.052)
			$\overline{AUC}^{1\bullet} = 0.8$ & $\bar{\lambda}^{1\bullet} = 1.25$ versus $\overline{AUC}^{2\bullet} = 0.6$ & $\bar{\lambda}^{2\bullet} = 0.75$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Power of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Power of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0018	0.0018 (0.0019)	0.083(0.078)	0.0023	0.0023 (0.0024)	0.086(0.082)
		0.7	0.0018	0.0017 (0.0018)	0.080(0.074)	0.0024	0.0025 (0.0025)	0.057(0.054)
		0.9	0.0012	0.0013 (0.0013)	0.257(0.244)	0.0022	0.0023 (0.0023)	0.078(0.074)
			$\overline{AUC}^{1\bullet} = 0.8$ & $\bar{\lambda}^{1\bullet} = 1.25$ versus $\overline{AUC}^{2\bullet} = 0.7$ & $\bar{\lambda}^{2\bullet} = 1.00$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Power of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Power of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0017	0.0017 (0.0017)	0.057(0.053)	0.0018	0.0018 (0.0019)	0.058(0.055)
		0.7	0.0016	0.0016 (0.0016)	0.091(0.086)	0.0020	0.0020 (0.0021)	0.069(0.066)
		0.9	0.0011	0.0011 (0.0011)	0.212(0.203)	0.0018	0.0018 (0.0019)	0.095(0.091)

Table 5.7 Estimated type I error rates for normal distributions for Scenario 4.

			$\overline{AUC}^{l\bullet} = 0.6$ & $\overline{\lambda}^{l\bullet} = 1.25$, l=1,2.					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Type I of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Type I of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0014	0.0014 (0.0014)	0.059(0.054)	0.0014	0.0015 (0.0015)	0.058(0.055)
		0.7	0.0014	0.0014 (0.0014)	0.059(0.055)	0.0018	0.0018 (0.0018)	0.060(0.058)
		0.9	0.0011	0.0012 (0.0012)	0.055(0.051)	0.0018	0.0019 (0.0019)	0.054(0.051)
			$\overline{AUC}^{l\bullet} = 0.7$ & $\overline{\lambda}^{l\bullet} = 1.00$, l=1,2.					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Type I of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Type I of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0014	0.0015 (0.0015)	0.060(0.056)	0.0017	0.0018 (0.0018)	0.058(0.054)
		0.7	0.0014	0.0014 (0.0014)	0.061(0.057)	0.0020	0.0020 (0.0020)	0.061(0.056)
		0.9	0.0010	0.0010 (0.0011)	0.060(0.056)	0.0019	0.0019 (0.0019)	0.057(0.054)
			$\overline{AUC}^{l\bullet} = 0.8$ & $\overline{\lambda}^{l\bullet} = 1.25$, l=1,2.					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Type I of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Type I of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0015	0.0015 (0.0015)	0.063(0.058)	0.0021	0.0021 (0.0021)	0.062(0.058)
		0.7	0.0013	0.0013 (0.0014)	0.063(0.060)	0.0021	0.0020 (0.0021)	0.061(0.058)
		0.9	0.0008	0.0008 (0.0009)	0.060(0.056)	0.0016	0.0017 (0.0017)	0.057(0.054)

Table 5.8 Estimated power for detecting system difference with regard to Scenario 4.

			$\overline{AUC}^{1\bullet} = 0.7$ & $\overline{\lambda}^{1\bullet} = 1.00$ versus $\overline{AUC}^{2\bullet} = 0.8$ & $\overline{\lambda}^{2\bullet} = 0.75$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Power of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Power of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0015	0.0015 (0.0016)	0.492(0.474)	0.0020	0.0020 (0.0021)	0.705(0.697)
		0.7	0.0015	0.0015 (0.0015)	0.607(0.593)	0.0022	0.0022 (0.0022)	0.826(0.819)
		0.9	0.0010	0.0010 (0.0011)	0.811(0.801)	0.0019	0.0019 (0.0020)	0.932(0.928)
			$\overline{AUC}^{1\bullet} = 0.6$ & $\overline{\lambda}^{1\bullet} = 1.25$ versus $\overline{AUC}^{2\bullet} = 0.8$ & $\overline{\lambda}^{2\bullet} = 0.75$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Power of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Power of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0017	0.0017 (0.0017)	0.958(0.953)	0.0021	0.0021 (0.0022)	0.995(0.995)
		0.7	0.0017	0.0017 (0.0018)	0.983(0.981)	0.0024	0.0024 (0.0025)	0.999(0.999)
		0.9	0.0014	0.0014 (0.0015)	0.998(0.998)	0.0022	0.0022 (0.0023)	1.000(1.000)
			$\overline{AUC}^{1\bullet} = 0.6$ & $\overline{\lambda}^{1\bullet} = 1.25$ versus $\overline{AUC}^{2\bullet} = 0.7$ & $\overline{\lambda}^{2\bullet} = 1.00$					
			Em se. *J1*	Jackk.(DBM) se. *J1*	Power of *J1*	Em se. $SWITCH_{(2)}$	Jackk.(DBM) se. $SWITCH_{(2)}$	Power of $SWITCH_{(2)}$
b	*t*	ν						
1	1	0.5	0.0015	0.0015 (0.0015)	0.511(0.492)	0.0017	0.0017 (0.0018)	0.657(0.649)
		0.7	0.0015	0.0015 (0.0015)	0.654(0.640)	0.0020	0.0020 (0.0020)	0.791(0.783)
		0.9	0.0012	0.0012 (0.0012)	0.844(0.832)	0.0020	0.0020 (0.0020)	0.912(0.908)

6.0 SMOOTH FROC CURVES

6.1 BACKGROUND

As introduced in Chapter 1, Edwards *et al* [31] characterized the process for an FROC experiment using an initial-detection-and-candidate-analysis (IDCA) model and with an independence assumption between the number of marks and the ratings of the subjects. This related the technique of fitting an FROC curve to the well developed technique of fitting an ROC curve. Specifically, at the most aggressive threshold of the FROC data, the average number of *FP* marks per subject is denoted as FPR_0 and the *TP* fraction obtained by dividing the total number of *TP* marks by the total number of lesions is denoted as TPF_0. The corresponding point (TPF_0, FPR_0) in the FROC plot is usually termed as the last experimental point. Conditioned on the total number of all marked locations, the *FP* and *TP* ratings can be summarized in an ROC curve ($FPF'(c)$, $TPF'(c)$) for all thresholds, *c*. The FROC curve ($FPR(c)$, $TPF(c)$) can be obtained by stretching the corresponding ROC curve ($FPR_0 \times FPF'(c)$, $TPF_0 \times TPF'(c)$) to the last experimental point. In their paper, Edwards *et al* [31] modeled the likelihood function of the FROC curve under the assumption of a binormal distribution for the ratings, a Poisson distribution for FPR_0 and a binomial distribution for TPF_0, and used the MLE method to fit a smooth FROC curve.

In this Chapter, we propose to extend to the FROC setting two approaches [9,17] that were originally developed to fit a smooth ROC curve. The first approach applied the Box-Cox power transformation to the ROC ratings, assumed the transformed ratings to be binormally distributed and used the MLE method to fit a smooth binormal ROC curve [9]. Several authors [9,14-17] investigated the use of a kernel smoothing technique and applied a density smoothing technique separately for the *FP* ratings and *TP* ratings of the ROC data. By choosing the corresponding kernel bandwidths, h_x and h_y, the overall smoothness of the fit for an ROC curve can be varied. The second approach presented in this Chapter uses a two-stage plug-in method to find kernel bandwidths, estimates kernel density function for the *FP* and *TP* ratings respectively, and fits the smooth ROC curve [17].

We will also apply the kernel regression approach as presented by Wand and Jones [57] to fit a smooth FROC curve. This kernel regression approach will regress the *TPF* on *FPR* using the empirical points in the FROC plot. The third approach allows us to estimate a smooth FROC curve without the independence assumption between the number of marks and the ratings of the subjects. For the three considered approaches, we will develop explicit formulations for the estimated smooth FROC curves. The areas under the estimated FROC curves by the three different approaches are also formulated and evaluated in a simulation study.

6.2 METHODS

Box-Cox transformation approach

In the ROC analysis, the ROC curve is invariant to any monotonically increasing transformation of the underlying data [1]. This nice property allows the researchers to relax the assumption on the binormally distribution of the underlying data, and to make the more general assumption that

the data can be transformable to a binormal distribution. The Box-Cox power transformation [56] is one such approach. When the ROC data are transformed with the same power function, the resulting empirical ROC curve remains the same, but the transformed ratings can fit a better binormal ROC curve, when the binormal assumption is untenable for the original ratings [9,18].

For the *FP* ratings X and *TP* ratings Y of the ROC data, Zou *et al* [9] applied the Box-Cox power transformation to X and Y, namely $X^{(\eta)} = \begin{cases} \dfrac{X^{\eta}-1}{\eta}, & \eta \neq 0 \\ \log(X), & \eta = 0 \end{cases}$ and $Y^{(\eta)} = \begin{cases} \dfrac{Y^{\eta}-1}{\eta}, & \eta \neq 0 \\ \log(Y), & \eta = 0 \end{cases}$.

It then can be assumed that $X^{(\eta)} \sim N(\mu_X, \sigma_X^2)$ and $Y^{(\eta)} \sim N(\mu_Y, \sigma_Y^2)$. The area under the ROC curve can be obtained as $AUC_{BC} = P(Y > X) = P(Y^{(\eta)} > X^{(\eta)}) = \Phi\left(\dfrac{\mu_Y - \mu_X}{\sqrt{\sigma_X^2 + \sigma_Y^2}}\right)$.

Following the notation for the FROC data of (2.1), all the ratings for S_0 actually negative and S_t actually positive subjects can be pooled together and summarized as follows:

$$\{x_i\}_{i=1}^{n} \leftrightarrow \text{FP marks}; \quad \{y_j\}_{j=1}^{m} \leftrightarrow \text{TP marks} \tag{6.1}$$

A profile log-likelihood function can be formulated and used to estimate η [9],

$$l = -\frac{n}{2}\log\left\{\frac{\sum_{i=1}^{n}\left(x_i^{(\eta)} - \dfrac{\sum_{i=1}^{n} x_i^{(\eta)}}{n}\right)^2}{n}\right\} - \frac{m}{2}\log\left\{\frac{\sum_{i=1}^{m}\left(y_j^{(\eta)} - \dfrac{\sum_{j=1}^{m} y_j^{(\eta)}}{m}\right)^2}{m}\right\} + (\eta - 1)\left(\sum_{i=1}^{n}\log x_i + \sum_{j=1}^{m}\log y_j\right) + c$$

The MLE of η is the solution to equation $\frac{\partial l}{\partial \eta} \triangleq f = 0$. We apply the Newton-Raphson algorithm to find the iterative solution for $\hat{\eta}$:

$$\eta^{k+1} = \eta^k - \frac{f'}{f''}, \text{ and } \hat{\eta} = \eta^{\infty}. \tag{6.2}$$

For practical applications, the functional forms of f' and f'' are too complex and one can use numerical derivatives instead of deriving the exact functional forms. We then apply $\hat{\eta}$ to obtain the transformed ratings of $\left\{x_i^{(\hat{\eta})}\right\}_{i=1}^{n}$ and $\left\{y_j^{(\hat{\eta})}\right\}_{j=1}^{m}$, and use them to estimate the parameters for the binormal distribution:

$$\hat{\mu}_X = \frac{\sum_{i=1}^{n} x_i^{(\hat{\eta})}}{n}, \quad \hat{\mu}_Y = \frac{\sum_{j=1}^{m} y_j^{(\hat{\eta})}}{m}, \quad \hat{\sigma}_X^2 = \frac{\sum_{i=1}^{n}\left(x_i^{(\hat{\eta})} - \hat{\mu}_X\right)^2}{n}, \quad \hat{\sigma}_Y^2 = \frac{\sum_{j=1}^{m}\left(y_j^{(\hat{\eta})} - \hat{\mu}_Y\right)^2}{m}.$$

Based on the transformed FROC ratings, the smooth ROC curve based on the Box-Cox transformation approach (BC) is formulated as $ROC_{BC}(p) = \Phi\left(\frac{\widehat{\mu_Y} - \widehat{\mu_X}}{\widehat{\sigma_Y}} + \frac{\widehat{\sigma_X}}{\widehat{\sigma_Y}} \times \Phi^{-1}(p)\right),\ p \subset (0,1).$

and the area under the ROC curve is estimated as $\widehat{AUC_{BC}} = \Phi\left(\frac{\widehat{\mu_Y} - \widehat{\mu_X}}{\sqrt{\widehat{\sigma_X^2} + \widehat{\sigma_Y^2}}}\right)$. Following the approach presented by Edwards *et al* [31], the smooth FROC curve from (0,0) to the last experimental point $\left(\widehat{FPR}_0, \widehat{TPF}_0\right)$ where $\widehat{FPR}_0 = \frac{n}{S_0 + S_t}$ and $\widehat{TPF}_0 = \frac{m}{tS_t}$ is formulated as

$FROC_{BC}(p') = \widehat{TPF_0} \times \Phi\left(\frac{\widehat{\mu_Y} - \widehat{\mu_X}}{\widehat{\sigma_Y}} + \frac{\widehat{\sigma_X}}{\widehat{\sigma_Y}} \times \Phi^{-1}\left(\frac{p'}{\widehat{FPR_0}}\right)\right)$, $p' \subset (0, \widehat{FPR_0})$ and the area under FROC curve can be estimated as $\widehat{FAUC_{BC}} = \widehat{AUC_{BC}} \times \widehat{FPR_0} \times \widehat{TPF_0}$.

Kernel smoothing approach

Several authors [9,14-17] applied well developed kernel smoothing techniques to fit a smooth ROC curve. The kernel smoothing approach estimates the density function for the *FP* ratings and *TP* ratings by choosing the corresponding kernel bandwidths h_x and h_y. A detailed comparison of the two kernel smoothing approaches presented by [9,17] is described in Faraggi and Reiser [18]. To develop the kernel smoothing approach for FROC data (6.1), following [9,17,18], a Gaussian kernel is chosen to estimate the probability density function of *FP* ratings *X* and *TP* ratings *Y* respectively: $\hat{f}_X(t) = \frac{1}{nh_X}\sum_{i=1}^{n}\phi(\frac{t-x_i}{h_X})$ and $\hat{f}_Y(t) = \frac{1}{mh_Y}\sum_{j=1}^{m}\phi(\frac{t-y_j}{h_Y})$, where ϕ is the density of the standard normal distribution.

As discussed by Wand and Jones [57], the kernel density estimator $\hat{f}_X(t) = \frac{1}{nh_X}\sum_{i=1}^{n}\phi(\frac{t-x_i}{h_X})$ can be interpreted as $\hat{f}_X(t) = \frac{1}{n}\sum_{i=1}^{n}\phi(t-x_i, 0, h_X)$ where $\phi(t-x_i, 0, h_X)$ is the density function of $N(0, h_X^2)$. At each value of *t*, the distances from *t* ($\{t-x_i\}_{i=1}^{n}$) are considered to be distributed as $N(0, h_X^2)$. The value of the kernel density estimator $\hat{f}_X(t)$ at *t* is simply the average of the *n* kernel densities at that point. In this way, the observations closer to *t* have more influence on the estimate than those farther away. The amount of relative influence is controlled by the bandwidth h_x. If h_x is small, then the estimate at each *t* depends heavily on those observations that are closest to *t*, and vice versa.

To estimate the optimal kernel bandwidths h_x and h_y from the FROC data in (6.1), Lloyd and Yong [17] used a two-stage plug-in bandwidth selection method presented by Wand and Jones [57]. The selection criteria for the optimal kernel bandwidth is to minimize the asymptotic mean integrated square error (AMISE) of the kernel density function.

With the optimal kernel bandwidths, the smooth ROC curve based on the kernel smoothing approach (KS) is formulated as $ROC_{KS}(p) = 1 - \hat{F}_Y(\hat{F}_X^{-1}(1-p)),\ p \subset (0,1)$ where $\hat{F}_X(t) = \frac{1}{n}\sum_{i=1}^{n}\Phi(\frac{t-x_i}{h_X})$ and $\hat{F}_Y(t) = \frac{1}{m}\sum_{j=1}^{m}\Phi(\frac{t-y_j}{h_Y})$. Lloyd [16] has shown that the resulting estimate of the area under the ROC curve can be estimated as $\widehat{AUC_{KS}} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\Phi(\frac{y_j - x_i}{\sqrt{h_X^2 + h_Y^2}})$. Following the approach presented by Edwards *et al* [31], the smooth FROC curve from (0,0) to the last experimental point $\left(\widehat{FPR}_0, \widehat{TPF}_0\right)$ is formulated as $FROC_{KS}(p') = \widehat{TPF_0} \times (1 - \hat{F}_Y(\hat{F}_X^{-1}(1 - \frac{p'}{\widehat{FPR_0}})))$, $p' \subset (0, \widehat{FPR_0})$ and the area under FROC curve can be estimated as $\widehat{FAUC_{KS}} = \widehat{AUC_{KS}} \times \widehat{FPR_0} \times \widehat{TPF_0}$.

As was noted in the paper [17], when applying a two-stage plug-in method for optimal bandwidth for FP ratings and TP ratings separately to fit a smooth ROC curve, the kernel bandwidths will probably not be optimal for estimating the entire ROC curve or any index related to the curve.

Kernel regression approach

For the third approach, we propose to estimate a smooth FROC curve from the empirical points in the FROC plot using a kernel regression method, and this approach allows us to relax the independence assumption between the number of marks and the ratings of the subjects. We apply a local linear kernel estimator presented by Wand and Jones [57]. The empirical points ($\{fpr_c\}_{c=1}^{C}$, $\{tpf_c\}_{c=1}^{C}$) can be obtained from our FROC data $\{x_i\}_{j=1}^{n}$ and $\{y_j\}_{j=1}^{m}$ in (6.1), where C is the total number of thresholds.

Considering the *tpf* and *fpr* as random variables and $\{fpr_c\}_{c=1}^{C}$ and $\{tpf_c\}_{c=1}^{C}$) as sample realizations, we regress the *tpf* on *fpr* as $tpf = m(fpr) + \varepsilon$, where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. At each point of *fpr*, *tpf* is estimated by fitting a straight line to the data $\{tpf_c\}_{c=1}^{C}$ using a weighted least squares (WLS) method where the weights are determined by the values of the kernel functions $\{\phi(fpr - fpr_c, 0, h_r)\}_{c=1}^{C}$ at that point. We use h_r to distinguish from h_X and h_Y that are formulated for the regression approach and $\phi(fpr - fpr_c, 0, h_r)$ is the standard normal density function scaled by h_r. Similarly, this means that those observations $\{fpr_c, tpf_c\}_{c=1}^{C}$ closer to *fpr* have more influence on the regression estimate at *fpr* than those farther away. Specifically, at each point of *fpr* the straight line $\beta_0 + \beta_1(t - fpr)$ is obtained by choosing the (β_0, β_1) that minimizes $SS = \sum_{c=1}^{C}\left\{\left[tpf_c - \beta_0 - \beta_1(fpr_c - fpr)\right]^2 \phi(fpr - fpr_c, 0, h_r)\right\}$.

The solution of (β_0, β_1) determines the straight line $\widehat{\beta_0} + \widehat{\beta_1}(t - fpr)$ and the estimator of *tpf*, namely $\hat{m}(fpr, h_r)$ at *fpr* is obtained at $\widehat{\beta_0} + \widehat{\beta_1}(fpr - fpr) = \widehat{\beta_0}$, and

$$\hat{m}(fpr, h_r) = \widehat{\beta_0} = \frac{s_2 \sum_{c=1}^{C}\left[tpf_c \phi(fpr - fpr_c, 0, h_r)\right] - s_1 \sum_{c=1}^{C}\left[tpf_c (fpr_c - fpr)\phi(fpr - fpr_c, 0, h_r)\right]}{s_2 s_0 - s_1^2} \text{ where}$$

$$s_0 = \sum_{c=1}^{C}\phi(fpr - fpr_c, 0, h_r), s_1 = \sum_{c=1}^{C}\left[\phi(fpr - fpr_c, 0, h_r)(fpr - fpr_c)\right], s_2 = \sum_{c=1}^{C}\left[\phi(fpr - fpr_c, 0, h_r)(fpr - fpr_c)^2\right].$$

To find the bandwidth h_r, we use a two-stage plug-in bandwidth selection method for *fpr* from the data $\{fpr_c\}_{c=1}^{C}$. The details follow the same approach as we derived h_x and h_y. It should be noted that, similar to h_x and h_y, h_r is chosen to better estimate the true distribution of *fpr*. It is probably not optimal for estimating the entire FROC curve or any index related to the curve.

The resulting FROC curve from (0,0) to the last experimental point $\left(\widehat{FPR}_0, \widehat{TPF}_0\right)$ using kernel regression approach (KR) is estimated as $FROC_{KR}(p') = \hat{m}(p', h_r)$, $p' \subset (0, \widehat{FPR}_0)$ where $\hat{m}(p', h_r)$ is formulated in (6.8). The area under the FROC curve is estimated as

$$\widehat{FAUC}_{KR} = \int_0^{\widehat{FPR}_0} \hat{m}(fpr, h_r)\, dfpr = \int_0^{\frac{n}{S_0+S_t}} \frac{s_2 \sum_{c=1}^{C} \left[tpf_c \phi(fpr - fpr_c, 0, h_r) \right] - s_1 \sum_{c=1}^{C} \left[tpf_c (fpr_c - fpr) \phi(fpr - fpr_c, 0, h_r) \right]}{s_2 s_0 - s_1^2} dfpr$$

6.3 SIMULATION RESULTS

We evaluate the estimators of the area under the FROC curve using all three proposed approaches in terms of the bias and root mean square error. We use BC to denote the Box-Cox transformation approach, KS to denote the kernel smoothing approach, and KR to denote the kernel regression approach. We simulate different scenarios in each of which we generate 10,000 independent datasets and each dataset consists of 20 actually negative and 20 actually positive subjects. The true expectation of the area under the FROC curve (FAUC) is estimated as the average of 10,000 empirical values based on 100 actually negative and 100 actually positive subjects. We consider the scenarios where the sample consists of a group of actually negative subjects and a group of actually positive subjects with *t*=1 and *t*=2 lesions. The number of *FP* marks on each subject n_{s_1} is simulated from a Poisson distribution (λ) where λ=1. The number of *TP* marks m_s on each actually positive subject is simulated from a binomial (*t*,ν) where ν=0.7, 0.9.

The ratings for *FP* $\vec{x}_{s_1}'$ on each subject and *TP* marks $\vec{y}_s'$ on each actually positive subject are first generated independently from normal distributions with means and variances chosen to

achieve a pre-specified degree of separation between *FP* and *TP* ratings (*AUC*) of 0.7 and 0.9. Then we apply $\vec{x}_{s_1} = \vec{x}_{s_1}' + \xi_x(n_{s_1} - 1)$, $\vec{y}_s = \vec{y}_s' + \xi_y(m_s - t\upsilon)$ to incorporate the possible correlation between the number of marks and the ratings of the subjects. This transformation keeps the same expectation of the ratings. A naïve algebra shows that ξ_x and ξ_y can be related to the desired correlation $r = cor(\vec{x}_{s_1}, n_{s_1}) = cor(\vec{y}_s, m_s)$ by $\xi_x = sign(r)sqrt(\frac{100r^2}{1-r^2})$ and $\xi_y = sign(r)sqrt(\frac{100r^2}{(1-r^2)t\upsilon(1-\upsilon)})$. We consider scenarios of small correlation (r=0.1), moderate correlations ($r = \pm 0.5$) and strong correlation (r=0.8) between the number of marks and the ratings of the subjects. Since conditional on the number of marks, $\vec{x}_{s_1}$ and $\vec{y}_s$ follow normal distributions. We denote the distribution of $\vec{x}_{s_1}$ and $\vec{y}_s$ as conditional normal distributions.

In our proposed simulation model we evaluate the three approaches when ratings follow conditional normal distributions and skewed distributions with a non-zero mass at the extremes. The skewed distributions were created by grouping the conditional normal distributions. The ratings below the 40th percentile of the distribution of the *FP* ratings were assigned to the 40th percentile value and the ratings above the 60th percentile of the distribution of the *TP* ratings were assigned the 60th percentile value. The simulation results for conditional normal distributions are shown in Table 6.1 and those for skewed distributions are shown in Table 6.2. The average bias and average root mean square errors of the area estimators over all selected parameter combinations are shown in Table 6.3.

For the simulation results of conditional normal distributions (Table 6.1), the bias of the BC approach tends to be larger for scenarios with strong correlation (r=0.8, bias ranges from -0.0103 to 0.0430 and the average bias is 0.0086). The bias tends to be small for scenarios with small to moderate correlations (bias ranges from -0.0080 to 0.0040 and the average bias is -0.0001). For

the simulation results of the skewed distributions (Table 6.2), the BC approach still tends to have larger bias for strong correlation (the average bias is 0.0181) than that for small to moderate correlations (the average bias is 0.0079). In addition, the BC approach tends to have smaller bias for conditional normal distributions (the average bias is 0.0021) than for skewed distributions (the average bias is 0.0105).

The two kernel approaches (KS and KR) underestimate the true FAUC in almost all simulated scenarios. Differing from the Box-Cox approach, the two kernel approaches tend to have smaller bias in skewed distributions (the average bias is -0.0080 for KS and -0.0069 for KR) than those for conditional normal distributions (the average bias is -0.0187 for KS and -0.0159 for KR) with the KR approach being less biased.

The root mean square errors for the two kernel approaches are similar and they both tend to be smaller than BC approach. For conditional normal distributions, the average RMSE is 0.1227 for BC, 0.1209 for KS and 0.1217 for KR. For skewed distributions, the average RMSE is 0.1229 for BC, 0.1180 for KS and 0.1183 for KR. For the two kernel approaches, the RMSE tends to be smaller for skewed distributions.

6.4 SUMMARY

In this Chapter we investigated three different approaches to estimating a smooth FROC curve for the evaluation of an FROC diagnostic system. The Box-Cox transformation [9] and kernel smoothing [17] approaches have already been widely investigated under the ROC setting and we naturally extend them for the analysis of FROC curves. We also present a third approach, which utilizes a kernel regression technique with the allowance of a dependence assumption between the number of marks and the ratings of the subjects. The area under the estimated FROC curves

are formulated and studied in the simulation under a variety of scenarios. We choose parameters that are commonly observed in an FROC study and consider a wide range of correlations between the number of marks and the ratings of the subjects.

As we observed in the simulation, the Box-Cox approach has the smallest bias for small to moderate correlations, although this approach has the largest bias when the correlation is strong (r=0.8). This phenomenon suggests that the Box-Cox approach may not be appropriate when there is strong correlation between the number of marks and the ratings of the subjects. In recognition of the construction of the conditional normal distributions of $\vec{x}_{s_1} = \vec{x}_{s_1}' + \xi_x(n_{s_1} - 1)$, $\vec{y}_s = \vec{y}_s' + \xi_y(m_s - t\upsilon)$, when the correlation is strong, ξ_x, ξ_y ($\xi_x, \xi_y \propto r$) serves as a scaling parameter that spreads the conditional normal distributions ($\vec{x}_{s_1}, \vec{y}_s$) from the original normal distributions ($\vec{x}_{s_1}', \vec{y}_s'$). In the case of strong correlation, applying Box-Cox transformation may not work well to "transform" the conditional normal distributions to the normal distributions.

The two kernel approaches both tend to have smaller RMSE for the estimate of the FAUC as compared to the Box-Cox approach. This might not be surprising since the formulation of the FROC curve, as well as the resulting FAUC depends highly on the bandwidth selection of h_x, h_y or h_r. As we described in the methods section, we selected the bandwidth that minimizes the asymptotic mean integrated square error (AMISE) of the density function for X and Y or the density function for *fpr*, although it may be hard to choose them by directly minimizing the AMISE of FAUC. Such minimization of the AMISE might not lead to the estimator that is least biased and in fact, as we observed in the simulation, selecting bandwidth to minimize the AMISE leads to an underestimate of the true FAUC for most of the simulated scenarios. If we select the bandwidth according to other criteria, such as bias, the results may differ.

It should be noted that our proposed kernel regression approach to fit a smooth FROC curve has not been shown to have much superiority over the traditional approach presented by Edwards *et al* [31]. It has smaller bias but similar RMSE for the estimate of the FAUC as compared to the traditional kernel approach. We also note that although the estimated areas under the FROC curve for KR and KS approaches studied in the simulation have similar results, the estimated FROC curve for the two approaches may be different in shape. Further improvement on the regression approach might be made by choosing the bandwidth h_r. As we have already discussed, h_r is not chosen to minimize the FROC curve estimate or FAUC. The derivation of such an h_r is not trivial but may be shown to have better results.

Table 6.1 Bias and root mean square error for conditional normal distributions.

correlation=-0.5

			FAUC	Bias(BC)	RMSE(BC)	Bias(KS)	RMSE(KS)	Bias(KR)	RMSE(KR)
AUC	ν	*t*							
0.7	0.7	1	0.4365	-0.0039	0.1211	-0.0094	0.1168	-0.0054	0.1191
		2	0.4932	-0.0050	0.1056	-0.0153	0.1048	-0.0110	0.1058
	0.9	1	0.6020	-0.0048	0.1394	-0.0167	0.1350	-0.0108	0.1371
		2	0.6311	-0.0006	0.1234	-0.0145	0.1235	-0.0107	0.1253
0.9	0.7	1	0.5933	0.0025	0.1409	-0.0237	0.1389	-0.0194	0.1383
		2	0.6210	0.0000	0.1232	-0.0214	0.1228	-0.0203	0.1234
	0.9	1	0.7875	0.0010	0.1536	-0.0309	0.1535	-0.0290	0.1529
		2	0.7991	-0.0035	0.1430	-0.0233	0.1434	-0.0205	0.1439

correlation=0.1

			FAUC	Bias(BC)	RMSE(BC)	Bias(KS)	RMSE(KS)	Bias(KR)	RMSE(KR)
AUC	ν	*t*							
0.7	0.7	1	0.4968	0.0026	0.1217	-0.0181	0.1178	-0.0113	0.1189
		2	0.4846	0.0025	0.1023	-0.0122	0.0996	-0.0076	0.1006
	0.9	1	0.6319	0.0017	0.1267	-0.0216	0.1230	-0.0143	0.1240
		2	0.6192	0.0001	0.1155	-0.0167	0.1130	-0.0113	0.1140
0.9	0.7	1	0.6326	0.0010	0.1386	-0.0250	0.1378	-0.0260	0.1374
		2	0.6261	0.0031	0.1207	-0.0175	0.1197	-0.0178	0.1202
	0.9	1	0.8102	0.0013	0.1459	-0.0293	0.1458	-0.0315	0.1467
		2	0.8042	0.0014	0.1388	-0.0229	0.1384	-0.0218	0.1390

correlation=0.5

			FAUC	Bias(BC)	RMSE(BC)	Bias(KS)	RMSE(KS)	Bias(KR)	RMSE(KR)
AUC	ν	*t*							
0.7	0.7	1	0.5211	0.0015	0.1155	-0.0207	0.1133	-0.0144	0.1137
		2	0.4488	-0.0022	0.0972	-0.0114	0.0946	-0.0077	0.0957
	0.9	1	0.6338	0.0020	0.1164	-0.0209	0.1138	-0.0141	0.1146
		2	0.5756	-0.0080	0.1069	-0.0134	0.1046	-0.0071	0.1052
0.9	0.7	1	0.6371	0.0025	0.1344	-0.0220	0.1331	-0.0252	0.1339
		2	0.5879	0.0031	0.1163	-0.0178	0.1152	-0.0155	0.1154
	0.9	1	0.8012	0.0040	0.1385	-0.0264	0.1376	-0.0276	0.1393
		2	0.7387	-0.0051	0.1327	-0.0187	0.1307	-0.0173	0.1322

correlation=0.8

			FAUC	Bias(BC)	RMSE(BC)	Bias(KS)	RMSE(KS)	Bias(KR)	RMSE(KR)
AUC	ν	*t*							
0.7	0.7	1	0.5323	0.0032	0.1120	-0.0181	0.1101	-0.0150	0.1107
		2	0.4080	-0.0103	0.0952	-0.0053	0.0924	-0.0031	0.0941
	0.9	1	0.6245	0.0017	0.1067	-0.0185	0.1041	-0.0133	0.1048
		2	0.5610	0.0430	0.1063	-0.0141	0.1004	-0.0054	0.1001
0.9	0.7	1	0.6199	0.0062	0.1284	-0.0177	0.1259	-0.0228	0.1276
		2	0.5099	-0.0011	0.1110	-0.0125	0.1093	-0.0121	0.1096
	0.9	1	0.7597	0.0066	0.1256	-0.0227	0.1241	-0.0234	0.1261
		2	0.6850	0.0196	0.1225	-0.0207	0.1236	-0.0145	0.1235

FAUC stands for average of 10,000 empirical areas under FROC curve. BC is for Box-Cox transformation, KS is for kernel smoothing and KR is for kernel regression.

Table 6.2 Bias and root mean square error for skewed distributions.

correlation=-0.5

			FAUC	Bias(BC)	RMSE(BC)	Bias(KS)	RMSE(KS)	Bias(KR)	RMSE(KR)
AUC	ν	*t*							
0.7	0.7	1	0.4018	0.0106	0.1197	0.0010	0.1118	0.0036	0.1130
		2	0.4555	0.0082	0.1029	-0.0036	0.0976	-0.0052	0.0975
	0.9	1	0.5587	0.0096	0.1344	-0.0064	0.1264	-0.0010	0.1291
		2	0.5814	0.0130	0.1204	-0.0046	0.1148	-0.0072	0.1164
0.9	0.7	1	0.5759	0.0043	0.1409	-0.0227	0.1373	-0.0040	0.1401
		2	0.6068	0.0073	0.1229	-0.0111	0.1205	-0.0083	0.1224
	0.9	1	0.7710	0.0056	0.1540	-0.0260	0.1521	-0.0071	0.1542
		2	0.7829	0.0048	0.1412	-0.0139	0.1406	-0.0076	0.1434

correlation=0.1

			FAUC	Bias(BC)	RMSE(BC)	Bias(KS)	RMSE(KS)	Bias(KR)	RMSE(KR)
AUC	ν	*t*							
0.7	0.7	1	0.4840	0.0050	0.1222	-0.0104	0.1181	-0.0028	0.1185
		2	0.4733	0.0042	0.1020	-0.0064	0.0994	-0.0006	0.1002
	0.9	1	0.6164	0.0046	0.1278	-0.0128	0.1236	-0.0036	0.1240
		2	0.6049	0.0016	0.1152	-0.0096	0.1127	-0.0022	0.1136
0.9	0.7	1	0.6292	0.0102	0.1413	-0.0114	0.1386	-0.0133	0.1368
		2	0.6229	0.0141	0.1237	-0.0061	0.1203	-0.0080	0.1194
	0.9	1	0.8060	0.0142	0.1491	-0.0129	0.1462	-0.0154	0.1450
		2	0.7999	0.0160	0.1424	-0.0090	0.1386	-0.0112	0.1378

correlation=0.5

			FAUC	Bias(BC)	RMSE(BC)	Bias(KS)	RMSE(KS)	Bias(KR)	RMSE(KR)
AUC	ν	*t*							
0.7	0.7	1	0.4841	0.0072	0.1151	-0.0065	0.1096	-0.0072	0.1092
		2	0.4333	-0.0020	0.0946	-0.0041	0.0922	-0.0050	0.0924
	0.9	1	0.5984	0.0115	0.1197	-0.0049	0.1118	-0.0037	0.1125
		2	0.5679	-0.0147	0.1070	-0.0024	0.1022	0.0018	0.1022
0.9	0.7	1	0.6168	0.0152	0.1337	-0.0082	0.1287	-0.0187	0.1279
		2	0.5779	0.0100	0.1166	-0.0074	0.1130	-0.0123	0.1119
	0.9	1	0.7781	0.0215	0.1390	-0.0094	0.1332	-0.0199	0.1328
		2	0.7335	0.0075	0.1317	-0.0073	0.1271	-0.0132	0.1265

correlation=0.8

			FAUC	Bias(BC)	RMSE(BC)	Bias(KS)	RMSE(KS)	Bias(KR)	RMSE(KR)
AUC	ν	*t*							
0.7	0.7	1	0.4710	0.0117	0.1129	-0.0055	0.1068	-0.0027	0.1071
		2	0.3910	-0.0075	0.0908	0.0080	0.0895	0.0093	0.0902
	0.9	1	0.5904	0.0098	0.1159	-0.0052	0.1086	-0.0060	0.1089
		2	0.5566	0.0347	0.1080	-0.0004	0.1010	-0.0003	0.0995
0.9	0.7	1	0.5720	0.0234	0.1258	-0.0089	0.1172	-0.0140	0.1173
		2	0.5045	-0.0001	0.1083	-0.0074	0.1038	-0.0081	0.1043
	0.9	1	0.7127	0.0313	0.1284	-0.0096	0.1166	-0.0160	0.1173
		2	0.6697	0.0418	0.1246	-0.0099	0.1145	-0.0111	0.1140

FAUC stands for average of 10,000 empirical areas under FROC curve. BC is for Box-Cox transformation, KS is for kernel smoothing and KR is for kernel regression.

Table 6.3 Average bias and average root mean square error.

under conditional normal distributions

	Bias(BC)	RMSE(BC)	Bias(KS)	RMSE(KS)	Bias(KR)	RMSE(KR)
Correlation						
-0.5	**-0.0018**	0.1313	-0.0194	**0.1298**	-0.0159	0.1307
0.1	**0.0017**	0.1263	-0.0204	**0.1244**	-0.0177	0.1251
0.5	**-0.0003**	0.1197	-0.0189	**0.1179**	-0.0161	0.1188
0.8	**0.0086**	0.1135	-0.0162	**0.1112**	-0.0137	0.1121
average	**0.0021**	0.1227	-0.0187	**0.1208**	-0.0159	0.1217

under skewed distributions

	Bias(BC)	RMSE(BC)	Bias(KS)	RMSE(KS)	Bias(KR)	RMSE(KR)
Correlation						
-0.5	0.0079	0.1296	-0.0109	**0.1251**	**-0.0046**	0.1270
0.1	0.0087	0.1280	-0.0098	0.1247	**-0.0071**	**0.1244**
0.5	0.0070	0.1197	**-0.0063**	0.1147	-0.0098	**0.1144**
0.8	0.0181	0.1143	**-0.0049**	**0.1073**	-0.0061	0.1073
average	0.0105	0.1229	-0.0080	0.1180	**-0.0069**	0.1183

Average bias and average root mean square error over all eight scenarios for both conditional normal and skewed distributions. BC is for Box-Cox transformation, KS is for kernel smoothing and KR is for kernel regression.

7.0 DISCUSSION AND RECOMMENDATIONS FOR FURTHER RESEARCH

In this Dissertation, we studied different groups of indices characterizing the performance level of an FROC diagnostic system. In Chapter 2, we applied an ROC approach to investigate the subject-based discriminative ability of the FROC system. A family of comparison functions was defined to summarize the discriminative ability in a random pair of actually negative and actually positive subjects. Each comparison function was formulated by a specific combination function such as maximum, mean and the Wilcoxon statistic to combine the multiple ratings on the subjects. With regard to the statistical power, we demonstrated that there is no statistically superior index for comparing subject-based discriminative ability of two FROC systems.

As was studied by Chakraborty and Berbaum [36], FROC indices can be formulated with the consideration of correct location information. They suggested "splitting" an actually positive subject into an *FP* population and an *LR* population. They found such indices could have greater power to detect the systems difference with regard to *AUC* than the approach based on subject-based evaluation. In Chapter 3, we investigated several groups of indices that used different handling methods for the *FP* population (SPLIT, IGNORE, SWITCH methods) and also used different comparison functions (based on maximum, mean and the Wilcoxon statistic) within each handling method. With regard to the statistical power, we demonstrated that when ignoring

the *FP* population on actually positive subjects, indices would lose substantial power (average 20%) to detect the system difference with regard to *AUC*. The SWITCH indices were found to have slightly greater average power than *J1* (modified *JAFROC1*) or SPLIT indices.

To incorporate the number of marks on the subjects in the evaluation of free-response diagnostic performance, we proposed a family of comparison functions by modifying the Wilcoxon statistic to incorporate the number of *FP* and *TP* marks on the subjects. The modified Wilcoxon statistic can then successfully penalize the indices when there is an increased number of generated "wrong" (*FP*) marks and reward the indices when there is an increased number of generated "correct" (*TP*) marks. When applying these functions to the SWITCH indices, the statistical test based on $SWITCH_{(2)}$ was found to have greater average power than that based on *J1* for detecting the system difference with regard to *AUC* and λ respectively. We further demonstrated that the same pattern can also be observed in a multi-reader FROC study (Chapter 5, scenario 1 for detecting system difference with regard to *AUC* only and scenario 2 for detecting system difference with regard to λ only).

Despite the improvement of power for our proposed indices, the naïve parameter estimators have very good statistical power in detecting the system difference with regard to that parameter. Specifically, for parameter *AUC*, it is the refined cluster ROC index, $\widehat{AUC}$, when ignoring non-marked lesions. For parameter λ, it is the index based on the average number of *FP* marks, say $\hat{\lambda}$. Differing from these single-parameter estimators, $SWITCH_{(2)}$ and *J1* summarize at least three features of the FROC diagnostic performance, namely parameters *AUC*, λ and ν. Thus it may not be reasonable to expect a single index to have the greatest power to detect system difference with regard to each of the three parameters. In addition, it is also difficult to compare different indices of this type. Even if we focus only on two parameters (*AUC* and λ), as further demonstrated in

Chapter 5, $SWITCH_{(2)}$ does not have greater power than $J1$ in all four considered scenarios that the diagnostic systems could differ.

Further research should be directed towards finding the most influential features in the evaluation of the free-response diagnostic system. If the researcher is primarily interested in one parameter, the parameter estimator itself can be expected to have greater power than the general FROC indices. If the researcher's interest is on two parameters simultaneously, such as AUC and λ, an index can be further formulated based on the combination of two parameter estimators. Specifically, an index that combines $\widehat{AUC}$ and $\hat{\lambda}$ with different weights, $w_1\widehat{AUC} + w_2\hat{\lambda}$, may be formulated so that the weights are chosen to maximize the statistical power in the two-dimensional space of AUC and λ.

APPENDIX

COMPUTATION EXAMPLE FOR PROPOSED INDICES

For four actually negative subjects ($S_0 = 4$) indexed by s' :

The *FP* ratings ($x^0_{s'c}$) for four actually negative subjects are

	$x^0_{s'1}$	$x^0_{s'2}$
s' =1	98	NA
s' =2	48	NA
s' =3	56	67
s' =4	NA	NA

For four actually positive subjects ($S_t = 4$, t=2) indexed by s :

The *FP* ratings (x^t_{sc}), the *TP* ratings (y^t_{sc}) and the ratings for non-marked lesions $\overrightarrow{L_{t-M}}$ are

	x^t_{s1}	x^t_{s2}	y^t_{s1}	y^t_{s2}	L_1	L_2
s =1	NA	NA	59	98	NA	NA
s =2	54	NA	79	89	NA	NA
s =3	45	87	74	NA	NA	0
s =4	NA	NA	NA	NA	0	0

Note the rating scale we used is from 0 to 100. So the lowest rating by default is L=0.

The subject-based indices as presented in Chapter 2:

In this Chapter we ignore non-marked lesions:

$$\hat{\theta} = \hat{A}_0 = \frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{\max}\left(\{x_{s'c}^0\}_{c=1}^{n_{s'}^0}, \left\{\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}\}_{c=1}^{m_s}\right\}\right)}{S_0 \times S_t} \tag{2.7}$$

$\tilde{\psi}_{\max}$	1st positive subject	2nd positive subject	3rd positive subject	4th positive subject
1st negative subject	98/(59,98) $\tilde{\psi}$ (98,98)=0.5	98/(54,79,89) $\tilde{\psi}$ (98,89)=0	98/(45,87,74) $\tilde{\psi}$ (98,87)=0	98/NA $\tilde{\psi}$ (98,NA)=0
2nd negative subject	48/(59,98) $\tilde{\psi}$ (48,98)=1	48/(54,79,89) $\tilde{\psi}$ (48,89)=1	48/(45,87,74) $\tilde{\psi}$ (48,87)=1	48/NA $\tilde{\psi}$ (48,NA)=0
3rd negative subject	(56,67)/(59,98) $\tilde{\psi}$ (67,98)=1	(56,67)/(54,79,89) $\tilde{\psi}$ (67,89)=1	(56,67)/(45,87,74) $\tilde{\psi}$ (67,87)=1	(56,67)/NA $\tilde{\psi}$ (67,NA)=0
4th negative subject	NA/(59,98) $\tilde{\psi}$ (NA,98)=1	NA/(54,79,89) $\tilde{\psi}$ (NA,89)=1	NA/(45,87,74) $\tilde{\psi}$ (NA,87)=1	NA/NA $\tilde{\psi}$ (NA,NA)=0.5

$$\hat{\theta} = \hat{A}_0 = \frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{\max}\left(\{x_{s'c}^0\}_{c=1}^{n_{s'}^0}, \left\{\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}\}_{c=1}^{m_s}\right\}\right)}{S_0 \times S_t} = \frac{10}{4\times 4} = 0.625$$

$$\hat{A}_1 = \frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{mean}\left(\{x_{s'c}^0\}_{c=1}^{n_{s'}^0}, \left\{\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}\}_{c=1}^{m_s}\right\}\right)}{S_0 \times S_t} \tag{2.8}$$

$\tilde{\psi}_{mean}$	1st positive subject	2nd positive subject	3rd positive subject	4th positive subject
1st negative subject	98/(59,98) $\tilde{\psi}$ (98,78.5)=0	98/(54,79,89) $\tilde{\psi}$ (98,74)=0	98/(45,87,74) $\tilde{\psi}$ (98,68.7)=0	98/NA $\tilde{\psi}$ (98,NA)=0
2nd negative subject	48/(59,98) $\tilde{\psi}$ (48,78.5)=1	48/(54,79,89) $\tilde{\psi}$ (48,74)=1	48/(45,87,74) $\tilde{\psi}$ (48,68.7)=1	48/NA $\tilde{\psi}$ (48,NA)=0
3rd negative subject	(56,67)/(59,98) $\tilde{\psi}$ (67,78.5)=1	(56,67)/(54,79,89) $\tilde{\psi}$ (67,74)=1	(56,67)/(45,87,74) $\tilde{\psi}$ (67,68.7)=1	(56,67)/NA $\tilde{\psi}$ (67,NA)=0
4th negative subject	NA/(59,98) $\tilde{\psi}$ (NA,78.5)=1	NA/(54,79,89) $\tilde{\psi}$ (NA,74)=1	NA/(45,87,74) $\tilde{\psi}$ (NA,68.7)=1	NA/NA $\tilde{\psi}$ (NA,NA)=0.5

$$\hat{A}_1 = \frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{mean}\left(\{x_{s'c}^0\}_{c=1}^{n_{s'}^0}, \left\{\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}\}_{c=1}^{m_s}\right\}\right)}{S_0 \times S_t} = \frac{9.5}{4\times 4} = 0.594$$

$$\hat{A}_2 = \frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{wilcoxon}\left(\{x_{s'c}^0\}_{c=1}^{n_{s'}^0}, \left\{\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}\}_{c=1}^{m_s}\right\}\right)}{S_0 \times S_t} \tag{2.9}$$

$\tilde{\psi}_{wil}$	1st positive subject	2nd positive subject	3rd positive subject	4th positive subject
1st negative subject	98/(59,98) $\tilde{\psi}$ (0.5,0.25)=0	98/(54,79,89) $\tilde{\psi}$ (0.5,0)=0	98/(45,87,74) $\tilde{\psi}$ (0.5,0)=0	98/NA $\tilde{\psi}$ (0.5,0)=0
2nd negative subject	48/(59,98) $\tilde{\psi}$ (0.5,1)=1	48/(54,79,89) $\tilde{\psi}$ (0.5,1)=1	48/(45,87,74) $\tilde{\psi}$ (0.5,0.67)=1	48/NA $\tilde{\psi}$ (0.5,0)=0
3rd negative subject	(56,67)/(59,98) $\tilde{\psi}$ (0.5,0.75)=1	(56,67)/(54,79,89) $\tilde{\psi}$ (0.5,0.67)=1	(56,67)/(45,87,74) $\tilde{\psi}$ (0.5,0.67)=1	(56,67)/NA $\tilde{\psi}$ (0.5,0)=0
4th negative subject	NA/(59,98) $\tilde{\psi}$ (0.5,1)=1	NA/(54,79,89) $\tilde{\psi}$ (0.5,1)=1	NA/(45,87,74) $\tilde{\psi}$ (0.5,1)=1	NA/NA $\tilde{\psi}$ (0.5,0.5)=0.5

$$\hat{A}_2 = \frac{\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{wilcoxon}\left(\{x_{s'c}^0\}_{c=1}^{n_{s'}^0}, \left\{\{x_{sc}^t\}_{c=1}^{n_s^t}, \{y_{sc}\}_{c=1}^{m_s}\right\}\right)}{S_0 \times S_t} = \frac{9.5}{4\times 4} = 0.594$$

The indices that incorporate the correct location information as presented in Chapter 3:

***J1* and three indices using SPLIT method:**

$$\widehat{J1} = \frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\tilde{\psi}_{J1}\left(\{x_{s_1c_1}\}_{c_1=1}^{n_{s_1}}, \left\{\{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}}\right\}\right) \tag{3.5}$$

$\tilde{\psi}_{JF1}$	*LR* (1st positive)	*LR* (2nd positive)	*LR* (3rd positive)	*LR* (4th positive)
FP (1st negative)	98/(59,98) $\tilde{\psi}_{J1}$=0.25	98/(79,89) $\tilde{\psi}_{J1}$=0	98/(74,0) $\tilde{\psi}_{J1}$=0	98/NA $\tilde{\psi}_{J1}$=0
FP (2nd negative)	48/(59,98) $\tilde{\psi}_{J1}$=1	48/(79,89) $\tilde{\psi}_{J1}$=1	48/(74,0) $\tilde{\psi}_{J1}$=0.5	48/NA $\tilde{\psi}_{J1}$=0
FP (3rd negative)	(56,67)/(59,98) $\tilde{\psi}_{J1}$=0.5	(56,67)/(79,89) $\tilde{\psi}_{J1}$=1	(56,67)/(74,0) $\tilde{\psi}_{J1}$=0.5	(56,67)/NA $\tilde{\psi}_{J1}$=0
FP (4th negative)	NA/(59,98) $\tilde{\psi}_{J1}$=1	NA/(79,89) $\tilde{\psi}_{J1}$=1	NA/(74,0) $\tilde{\psi}_{J1}$=1	NA/NA $\tilde{\psi}_{J1}$=0.5
FP (1st positive)	NA/(59,98) $\tilde{\psi}_{J1}$=1	NA/(79,89) $\tilde{\psi}_{J1}$=1	NA/(74,0) $\tilde{\psi}_{J1}$=1	NA/NA $\tilde{\psi}_{J1}$=0.5
FP (2nd positive)	54/(59,98) $\tilde{\psi}_{J1}$=1	54/(79,89) $\tilde{\psi}_{J1}$=1	54/(74,0) $\tilde{\psi}_{J1}$=0.5	54/NA $\tilde{\psi}_{J1}$=0
FP (3rd positive)	(45,87)/(59,98) $\tilde{\psi}_{J1}$=0.5	(45,87)/(79,89) $\tilde{\psi}_{J1}$=0.5	(45,87)/(74,0) $\tilde{\psi}_{J1}$=0	(45,87)/NA $\tilde{\psi}_{J1}$=0
FP (4th positive)	NA/(59,98) $\tilde{\psi}_{J1}$=1	NA/(79,89) $\tilde{\psi}_{J1}$=1	NA/(74,0) $\tilde{\psi}_{J1}$=1	NA/NA $\tilde{\psi}_{J1}$=0.5

$$\widehat{J1} = \frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\tilde{\psi}_{J1}\left(\{x_{s_1c_1}\}_{c_1=1}^{n_{s_1}}, \left\{\{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}}\right\}\right) = \frac{18.75}{8\times 4} = 0.586$$

$$\widehat{SPLIT}_0 = \frac{1}{(S_0 + S_t) \times S_t} \sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} \tilde{\psi}_{\max} \left(\{x_{s_1 c_1}\}_{c_1=1}^{n_{s_1}}, \left\{ \{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}} \right\} \right) \tag{3.6}$$

$\tilde{\psi}_{\max}$	*LR* (1st positive)	*LR* (2nd positive)	*LR* (3rd positive)	*LR* (4th positive)
FP (1st negative)	98/(59,98) $\tilde{\psi}$ (98,98)=0.5	98/(79,89) $\tilde{\psi}$ (98,89)=0	98/(74,0) $\tilde{\psi}$ (98,74)=0	98/NA $\tilde{\psi}$ (98,NA)=0
FP (2nd negative)	48/(59,98) $\tilde{\psi}$ (48,98)=1	48/(79,89) $\tilde{\psi}$ (48,89)=1	48/(74,0) $\tilde{\psi}$ (48,74)=1	48/NA $\tilde{\psi}$ (48,NA)=0
FP (3rd negative)	(56,67)/(59,98) $\tilde{\psi}$ (67,98)=1	(56,67)/(79,89) $\tilde{\psi}$ (67,89)=1	(56,67)/(74,0) $\tilde{\psi}$ (67,74)=1	(56,67)/NA $\tilde{\psi}$ (67,NA)=0
FP (4th negative)	NA/(59,98) $\tilde{\psi}$ (NA,98)=1	NA/(79,89) $\tilde{\psi}$ (NA,89)=1	NA/(74,0) $\tilde{\psi}$ (NA,74)=1	NA/NA $\tilde{\psi}$ (NA,NA)= 0.5
FP (1st positive)	NA/(59,98) $\tilde{\psi}$ (NA,98)=1	NA/(79,89) $\tilde{\psi}$ (NA,89)=1	NA/(74,0) $\tilde{\psi}$ (NA,74)=1	NA/NA $\tilde{\psi}$ (NA,NA)= 0.5
FP (2nd positive)	54/(59,98) $\tilde{\psi}$ (54,98)=1	54/(79,89) $\tilde{\psi}$ (54,89)=1	54/(74,0) $\tilde{\psi}$ (54,74)=1	54/NA $\tilde{\psi}$ (54,NA)=0
FP (3rd positive)	(45,87)/(59,98) $\tilde{\psi}$ (87,98)=1	(45,87)/(79,89) $\tilde{\psi}$ (87,89)=1	(45,87)/(74,0) $\tilde{\psi}$ (87,74)=0	(45,87)/NA $\tilde{\psi}$ (87,NA)=0
FP (4th positive)	NA/(59,98) $\tilde{\psi}$ (NA,98)=1	NA/(79,89) $\tilde{\psi}$ (NA,89)=1	NA/(74,0) $\tilde{\psi}$ (NA,74)=1	NA/NA $\tilde{\psi}$ (NA,NA)= 0.5

$$\widehat{SPLIT}_0 = \frac{1}{(S_0 + S_t) \times S_t} \sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} \tilde{\psi}_{\max} \left(\{x_{s_1 c_1}\}_{c_1=1}^{n_{s_1}}, \left\{ \{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}} \right\} \right) = \frac{22}{8 \times 4} = 0.688$$

$$\widehat{SPLIT}_1 = \frac{1}{(S_0 + S_t) \times S_t} \sum_{s_1=1}^{S_0+S_t} \sum_{s=1}^{S_t} \tilde{\psi}_{mean} \left(\{x_{s_1 c_1}\}_{c_1=1}^{n_{s_1}}, \left\{ \{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}} \right\} \right) \tag{3.7}$$

$\tilde{\psi}_{mean}$	*LR* (1st positive)	*LR* (2nd positive)	*LR* (3rd positive)	*LR* (4th positive)
FP (1st negative)	98/(59,98) $\tilde{\psi}$ (98,78.5)=0	98/(79,89) $\tilde{\psi}$ (98,84)=0	98/(74,0) $\tilde{\psi}$ (98,37)=0	98/NA $\tilde{\psi}$ (98,NA)=0
FP (2nd negative)	48/(59,98) $\tilde{\psi}$ (48,78.5)=1	48/(79,89) $\tilde{\psi}$ (48,84)=1	48/(74,0) $\tilde{\psi}$ (48,37)=0	48/NA $\tilde{\psi}$ (48,NA)=0
FP (3rd negative)	(56,67)/(59,98) $\tilde{\psi}$ (61.5,78.5)=1	(56,67)/(79,89) $\tilde{\psi}$ (61.5,84)=1	(56,67)/(74,0) $\tilde{\psi}$ (61.5,37)=0	(56,67)/NA $\tilde{\psi}$ (61.5,NA)=0
FP (4th negative)	NA/(59,98) $\tilde{\psi}$ (NA,78.5)=1	NA/(79,89) $\tilde{\psi}$ (NA,84)=1	NA/(74,0) $\tilde{\psi}$ (NA,37)=1	NA/NA $\tilde{\psi}$ (NA,NA)=0.5
FP (1st positive)	NA/(59,98) $\tilde{\psi}$ (NA,78.5)=1	NA/(79,89) $\tilde{\psi}$ (NA,84)=1	NA/(74,0) $\tilde{\psi}$ (NA,37)=1	NA/NA $\tilde{\psi}$ (NA,NA)=0.5
FP (2nd positive)	54/(59,98) $\tilde{\psi}$ (54,78.5)=1	54/(79,89) $\tilde{\psi}$ (54,84)=1	54/(74,0) $\tilde{\psi}$ (54,37)=0	54/NA $\tilde{\psi}$ (54,NA)=0
FP (3rd positive)	(45,87)/(59,98) $\tilde{\psi}$ (66,78.5)=1	(45,87)/(79,89) $\tilde{\psi}$ (66,84)=1	(45,87)/(74,0) $\tilde{\psi}$ (66,37)=0	(45,87)/NA $\tilde{\psi}$ (66,NA)=0

FP (4th positive)	NA/(59,98) $\tilde{\psi}$ (NA,78.5)=1	NA/(79,89) $\tilde{\psi}$ (NA,84)=1	NA/(74,0) $\tilde{\psi}$ (NA,37)=1	NA/NA $\tilde{\psi}$ (NA,NA)=0.5

$$\widehat{SPLIT}_1 = \frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\tilde{\psi}_{mean}\left(\{x_{s_1c_1}\}_{c_1=1}^{n_{s_1}}, \left\{\{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}}\right\}\right) = \frac{18.5}{8\times 4} = 0.578$$

$$\widehat{SPLIT}_2 = \frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\tilde{\psi}_{wilcoxon}\left(\{x_{s_1c_1}\}_{c_1=1}^{n_{s_1}}, \left\{\{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}}\right\}\right) \quad (3.8)$$

$\tilde{\psi}_{wil}$	*LR* (1st positive)	*LR* (2nd positive)	*LR* (3rd positive)	*LR* (4th positive)
FP (1st negative)	98/(59,98) $\tilde{\psi}$ (0.5,0.25)=0	98/(79,89) $\tilde{\psi}$ (0.5,0)=0	98/(74,0) $\tilde{\psi}$ (0.5,0)=0	98/NA $\tilde{\psi}$ (0.5,0)=0
FP (2nd negative)	48/(59,98) $\tilde{\psi}$ (0.5,1)=1	48/(79,89) $\tilde{\psi}$ (0.5,1)=1	48/(74,0) $\tilde{\psi}$ (0.5,0.5)=0.5	48/NA $\tilde{\psi}$ (0.5,0)=0
FP (3rd negative)	(56,67)/(59,98) $\tilde{\psi}$ (0.5,0.75)=1	(56,67)/(79,89) $\tilde{\psi}$ (0.5,1)=1	(56,67)/(74,0) $\tilde{\psi}$ (0.5,0.5)=0.5	(56,67)/NA $\tilde{\psi}$ (0.5,0)=0
FP (4th negative)	NA/(59,98) $\tilde{\psi}$ (0.5,1)=1	NA/(79,89) $\tilde{\psi}$ (0.5,1)=1	NA/(74,0) $\tilde{\psi}$ (0.5,1)=1	NA/NA $\tilde{\psi}$ (0.5,0.5)=0.5
FP (1st positive)	NA/(59,98) $\tilde{\psi}$ (0.5,1)=1	NA/(79,89) $\tilde{\psi}$ (0.5,1)=1	NA/(74,0) $\tilde{\psi}$ (0.5,1)=1	NA/NA $\tilde{\psi}$ (0.5,0.5)=0.5
FP (2nd positive)	54/(59,98) $\tilde{\psi}$ (0.5,1)=1	54/(79,89) $\tilde{\psi}$ (0.5,1)=1	54/(74,0) $\tilde{\psi}$ (0.5,0.5)=0.5	54/NA $\tilde{\psi}$ (0.5,0)=0
FP (3rd positive)	(45,87)/(59,98) $\tilde{\psi}$ (0.5,0.75)=1	(45,87)/(79,89) $\tilde{\psi}$ (0.5,0.75)=1	(45,87)/(74,0) $\tilde{\psi}$ (0.5,0.25)=0	(45,87)/NA $\tilde{\psi}$ (0.5,0)=0
FP (4th positive)	NA/(59,98) $\tilde{\psi}$ (0.5,1)=1	NA/(79,89) $\tilde{\psi}$ (0.5,1)=1	NA/(74,0) $\tilde{\psi}$ (0.5,1)=1	NA/NA $\tilde{\psi}$ (0.5,0.5)=0.5

$$\widehat{SPLIT}_2 = \frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\tilde{\psi}_{wilcoxon}\left(\{x_{s_1c_1}\}_{c_1=1}^{n_{s_1}}, \left\{\{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}}\right\}\right) = \frac{20}{8\times 4} = 0.625$$

***J2* and three indices using IGNORE method:**

$$\widehat{J2} = \frac{1}{S_0\times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{J2}\left(\{x^0_{s'c_1}\}_{c_1=1}^{n^0_{s'}}, \left\{\{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}}\right\}\right) \quad (3.9)$$

$\tilde{\psi}_{JF2}$	1st positive subject	2nd positive subject	3rd positive subject	4th positive subject
1st negative subject	98/(59,98) $\tilde{\psi}_{J2}$=0.25	98/(79,89) $\tilde{\psi}_{J2}$=0	98/(74,0) $\tilde{\psi}_{J2}$=0	98/NA $\tilde{\psi}_{J2}$=0
2nd negative subject	48/(59,98) $\tilde{\psi}_{J2}$=1	48/(79,89) $\tilde{\psi}_{J2}$=1	48/(74,0) $\tilde{\psi}_{J2}$=0.5	48/NA $\tilde{\psi}_{J2}$=0
3rd negative subject	(56,67)/(59,98) $\tilde{\psi}_{J2}$=0.5	(56,67)/(79,89) $\tilde{\psi}_{J2}$=1	(56,67)/(74,0) $\tilde{\psi}_{J2}$=0.5	(56,67)/NA $\tilde{\psi}_{J2}$=0

4th negative subject	NA/(59,98) $\tilde{\psi}_{J2}$=1	NA/(79,89) $\tilde{\psi}_{J2}$=1	NA/(74,0) $\tilde{\psi}_{J2}$=1	NA/NA $\tilde{\psi}_{J2}$=0.5

$$\widehat{J2}=\frac{1}{S_0\times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{J2}\left(\{x_{s'c_1}^0\}_{c_1=1}^{n_{s'}^0},\left\{\{y_{sc_2}\}_{c_2=1}^{m_s},\overrightarrow{L_{t-m_s}}\right\}\right)=\frac{8.25}{4\times 4}=0.516$$

$$\widehat{IGNORE}_0=\frac{1}{S_0\times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{\max}\left(\{x_{s'c_1}^0\}_{c_1=1}^{n_{s'}^0},\left\{\{y_{sc_2}\}_{c_2=1}^{m_s},\overrightarrow{L_{t-m_s}}\right\}\right) \tag{3.10}$$

$\tilde{\psi}_{\max}$	1st positive subject	2nd positive subject	3rd positive subject	4th positive subject
1st negative subject	98/(59,98) $\tilde{\psi}$(98,98)=0.5	98/(79,89) $\tilde{\psi}$(98,89)=0	98/(74,0) $\tilde{\psi}$(98,74)=0	98/NA $\tilde{\psi}$(98,NA)=0
2nd negative subject	48/(59,98) $\tilde{\psi}$(48,98)=1	48/(79,89) $\tilde{\psi}$(48,89)=1	48/(74,0) $\tilde{\psi}$(48,74)=1	48/NA $\tilde{\psi}$(48,NA)=0
3rd negative subject	(56,67)/(59,98) $\tilde{\psi}$(67,98)=1	(56,67)/(79,89) $\tilde{\psi}$(67,89)=1	(56,67)/(74,0) $\tilde{\psi}$(67,74)=1	(56,67)/NA $\tilde{\psi}$(67,NA)=0
4th negative subject	NA/(59,98) $\tilde{\psi}$(NA,98)=1	NA/(79,89) $\tilde{\psi}$(NA,89)=1	NA/(74,0) $\tilde{\psi}$(NA,74)=1	NA/NA $\tilde{\psi}$(NA,NA)= 0.5

$$\widehat{IGNORE}_0=\frac{1}{S_0\times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{\max}\left(\{x_{s'c_1}^0\}_{c_1=1}^{n_{s'}^0},\left\{\{y_{sc_2}\}_{c_2=1}^{m_s},\overrightarrow{L_{t-m_s}}\right\}\right)=\frac{10}{4\times 4}=0.625$$

$$\widehat{IGNORE}_1=\frac{1}{S_0\times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{mean}\left(\{x_{s'c_1}^0\}_{c_1=1}^{n_{s'}^0},\left\{\{y_{sc_2}\}_{c_2=1}^{m_s},\overrightarrow{L_{t-m_s}}\right\}\right) \tag{3.11}$$

$\tilde{\psi}_{mean}$	1st positive subject	2nd positive subject	3rd positive subject	4th positive subject
1st negative subject	98/(59,98) $\tilde{\psi}$(98,78.5)=0	98/(79,89) $\tilde{\psi}$(98,84)=0	98/(74,0) $\tilde{\psi}$(98,37)=0	98/NA $\tilde{\psi}$(98,NA)=0
2nd negative subject	48/(59,98) $\tilde{\psi}$(48,78.5)=1	48/(79,89) $\tilde{\psi}$(48,84)=1	48/(74,0) $\tilde{\psi}$(48,37)=0	48/NA $\tilde{\psi}$(48,NA)=0
3rd negative subject	(56,67)/(59,98) $\tilde{\psi}$(61.5,78.5)=1	(56,67)/(79,89) $\tilde{\psi}$(61.5,84)=1	(56,67)/(74,0) $\tilde{\psi}$(61.5,37)=0	(56,67)/NA $\tilde{\psi}$(61.5,NA)=0
4th negative subject	NA/(59,98) $\tilde{\psi}$(NA,78.5)=1	NA/(79,89) $\tilde{\psi}$(NA,84)=1	NA/(74,0) $\tilde{\psi}$(NA,37)=1	NA/NA $\tilde{\psi}$(NA,NA)=0.5

$$\widehat{IGNORE}_1=\frac{1}{S_0\times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{mean}\left(\{x_{s'c_1}^0\}_{c_1=1}^{n_{s'}^0},\left\{\{y_{sc_2}\}_{c_2=1}^{m_s},\overrightarrow{L_{t-m_s}}\right\}\right)=\frac{7.5}{4\times 4}=0.469$$

$$\widehat{IGNORE}_2=\frac{1}{S_0\times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{wilcoxon}\left(\{x_{s'c_1}^0\}_{c_1=1}^{n_{s'}^0},\left\{\{y_{sc_2}\}_{c_2=1}^{m_s},\overrightarrow{L_{t-m_s}}\right\}\right) \tag{3.12}$$

$\tilde{\psi}_{wil}$	1st positive subject	2nd positive subject	3rd positive subject	4th positive subject
1st negative subject	98/(59,98) $\tilde{\psi}$ (0.5,0.25)=0	98/(79,89) $\tilde{\psi}$ (0.5,0)=0	98/(74,0) $\tilde{\psi}$ (0.5,0)=0	98/NA $\tilde{\psi}$ (0.5,0)=0
2nd negative subject	48/(59,98) $\tilde{\psi}$ (0.5,1)=1	48/(79,89) $\tilde{\psi}$ (0.5,1)=1	48/(74,0) $\tilde{\psi}$ (0.5,0.5)=0.5	48/NA $\tilde{\psi}$ (0.5,0)=0
3rd negative subject	(56,67)/(59,98) $\tilde{\psi}$ (0.5,0.75)=1	(56,67)/(79,89) $\tilde{\psi}$ (0.5,1)=1	(56,67)/(74,0) $\tilde{\psi}$ (0.5,0.5)=0.5	(56,67)/NA $\tilde{\psi}$ (0.5,0)=0
4th negative subject	NA/(59,98) $\tilde{\psi}$ (0.5,1)=1	NA/(79,89) $\tilde{\psi}$ (0.5,1)=1	NA/(74,0) $\tilde{\psi}$ (0.5,1)=1	NA/NA $\tilde{\psi}$ (0.5,0.5)=0.5

$$\widehat{IGNORE}_2 = \frac{1}{S_0 \times S_t} \sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \tilde{\psi}_{wilcoxon} \left(\{x^0_{s'c_1}\}_{c_1=1}^{n^0_{s'}}, \left\{ \{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}} \right\} \right) = \frac{8.5}{4 \times 4} = 0.531$$

Three indices using SWITCH method:

$$\widehat{SWITCH}_0 = \frac{1}{S_0 \times S_t} \sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \tilde{\psi}_{\max} \left(\left\{ \{x^0_{s'c_1}\}_{c_1=1}^{n^0_{s'}}, \{x^t_{sc_2}\}_{c_2=1}^{n^t_s} \right\}, \left\{ \{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}} \right\} \right) \tag{3.13}$$

$\tilde{\psi}_{\max}$	1st positive subject	2nd positive subject	3rd positive subject	4th positive subject
1st negative subject	98/(59,98) $\tilde{\psi}$ (98,98)=0.5	(98,54)/(79,89) $\tilde{\psi}$ (98,89)=0	(98,45,87)/(74,0) $\tilde{\psi}$ (98,74)=0	98/NA $\tilde{\psi}$ (98,NA)=0
2nd negative subject	48/(59,98) $\tilde{\psi}$ (48,98)=1	(48,54)/(79,89) $\tilde{\psi}$ (54,89)=1	(48,45,87)/(74,0) $\tilde{\psi}$ (87,74)=0	48/NA $\tilde{\psi}$ (48,NA)=0
3rd negative subject	(56,67)/(59,98) $\tilde{\psi}$ (67,98)=1	(56,67,54)/(79,89) $\tilde{\psi}$ (67,89)=1	(56,67,45,87) /(74,0) $\tilde{\psi}$ (87,74)=0	(56,67)/NA $\tilde{\psi}$ (67,NA)=0
4th negative subject	NA/(59,98) $\tilde{\psi}$ (NA,98)=1	54/(79,89) $\tilde{\psi}$ (54,89)=1	(45,87)/(74,0) $\tilde{\psi}$ (87,74)=0	NA/NA $\tilde{\psi}$ (NA,NA)=0.5

$$\widehat{SWITCH}_0 = \frac{1}{S_0 \times S_t} \sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \tilde{\psi}_{\max} \left(\left\{ \{x^0_{s'c_1}\}_{c_1=1}^{n^0_{s'}}, \{x^t_{sc_2}\}_{c_2=1}^{n^t_s} \right\}, \left\{ \{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}} \right\} \right) = \frac{8}{4 \times 4} = 0.500$$

$$\widehat{SWITCH}_1 = \frac{1}{S_0 \times S_t} \sum_{s'=1}^{S_0} \sum_{s=1}^{S_t} \tilde{\psi}_{mean} \left(\left\{ \{x^0_{s'c_1}\}_{c_1=1}^{n^0_{s'}}, \{x^t_{sc_2}\}_{c_2=1}^{n^t_s} \right\}, \left\{ \{y_{sc_2}\}_{c_2=1}^{m_s}, \overrightarrow{L_{t-m_s}} \right\} \right) \tag{3.14}$$

$\tilde{\psi}_{mean}$	1st positive subject	2nd positive subject	3rd positive subject	4th positive subject
1st negative subject	98/(59,98) $\tilde{\psi}$ (98,78.5)=0	(98,54)/(79,89) $\tilde{\psi}$ (76,84)=1	(98,45,87)/(74,0) $\tilde{\psi}$ (76.7,37)=0	98/NA $\tilde{\psi}$ (98,NA)=0
2nd negative subject	48/(59,98) $\tilde{\psi}$ (48,78.5)=1	(48,54)/(79,89) $\tilde{\psi}$ (51,84)=1	(48,45,87)/(74,0) $\tilde{\psi}$ (60,37)=0	48/NA $\tilde{\psi}$ (48,NA)=0
3rd negative subject	(56,67)/(59,98) $\tilde{\psi}$ (61.5,78.5)=1	(56,67,54)/(79,89) $\tilde{\psi}$ (59,84)=1	(56,67,45,87) /(74,0) $\tilde{\psi}$ (63.8,37)=0	(56,67)/NA $\tilde{\psi}$ (61.5,NA)=0
4th negative subject	NA/(59,98) $\tilde{\psi}$ (NA,78.5)=1	54/(79,89) $\tilde{\psi}$ (54,84)=1	(45,87)/(74,0) $\tilde{\psi}$ (66,37)=0	NA/NA $\tilde{\psi}$ (NA,NA)=0.5

$$\widehat{SWITCH}_1 = \frac{1}{S_0 \times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{mean}\left(\left\{\left\{x^0_{s'c_1}\right\}_{c_1=1}^{n^0_{s'}},\left\{x^t_{sc_2}\right\}_{c_2=1}^{n^t_s}\right\},\left\{\{y_{sc_2}\}_{c_2=1}^{m_s},\overrightarrow{L_{t-m_s}}\right\}\right) = \frac{7.5}{4\times 4} = 0.469$$

$$\widehat{SWITCH}_2 = \frac{1}{S_0 \times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{wilcoxon}\left(\left\{\left\{x^0_{s'c_1}\right\}_{c_1=1}^{n^0_{s'}},\left\{x^t_{sc_2}\right\}_{c_2=1}^{n^t_s}\right\},\left\{\{y_{sc_2}\}_{c_2=1}^{m_s},\overrightarrow{L_{t-m_s}}\right\}\right) \qquad (3.15)$$

$\tilde{\psi}_{wil}$	1st positive subject	2nd positive subject	3rd positive subject	4th positive subject
1st negative subject	98/(59,98) $\tilde{\psi}$ (0.5,0.25)=0	(98,54)/(79,89) $\tilde{\psi}$ (0.5,0.5)=0.5	(98,45,87)/(74,0) $\tilde{\psi}$ (0.5,0.17)=0	98/NA $\tilde{\psi}$ (0.5,0)=0
2nd negative subject	48/(59,98) $\tilde{\psi}$ (0.5,1)=1	(48,54)/(79,89) $\tilde{\psi}$ (0.5,1)=1	(48,45,87)/(74,0) $\tilde{\psi}$ (0.5,0.33)=0	48/NA $\tilde{\psi}$ (0.5,0)=0
3rd negative subject	(56,67)/(59,98) $\tilde{\psi}$ (0.5,0.75)=1	(56,67,54)/(79,89) $\tilde{\psi}$ (0.5,1)=1	(56,67,45,87)/(74,0) $\tilde{\psi}$ (0.5,0.38)=0	(56,67)/NA $\tilde{\psi}$ (0.5,0)=0
4th negative subject	NA/(59,98) $\tilde{\psi}$ (0.5,1)=1	54/(79,89) $\tilde{\psi}$ (0.5,1)=1	(45,87)/(74,0) $\tilde{\psi}$ (0.5,0.25)=0	NA/NA $\tilde{\psi}$ (0.5,0.5)=0.5

$$\widehat{SWITCH}_2 = \frac{1}{S_0 \times S_t}\sum_{s'=1}^{S_0}\sum_{s=1}^{S_t}\tilde{\psi}_{wilcoxon}\left(\left\{\left\{x^0_{s'c_1}\right\}_{c_1=1}^{n^0_{s'}},\left\{x^t_{sc_2}\right\}_{c_2=1}^{n^t_s}\right\},\left\{\{y_{sc_2}\}_{c_2=1}^{m_s},\overrightarrow{L_{t-m_s}}\right\}\right) = \frac{7}{4\times 4} = 0.438$$

The indices that incorporate the number of marks presented in Chapter 4:
Three indices using SPLIT method in (4.1):

$$k=2,\ \widehat{SPLIT}_{(2)} = \frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\left\{\frac{\sum_{c_1=1}^{n_{s_1}}\sum_{c_2=1}^{m_s}\tilde{\psi}\left(x_{s_1c_1},y_{sc_2}\right)}{n_{s_1}\times m_s}+\frac{1}{2}\left(\frac{m_s}{n_{s_1}}-1\right)\right\}$$

$\tilde{\psi}_{wil}$	*TP* (1st positive)	*TP* (2nd positive)	*TP* (3rd positive)	*TP* (4th positive)
FP (1st negative)	98/(59,98) $\tilde{\psi}_{wil2}$ =0.25+0.5=0.75	98/(79,89) $\tilde{\psi}_{wil2}$ =0+0.5=0.5	98/74 $\tilde{\psi}_{wil2}$ =0+0=0	98/NA $\tilde{\psi}$ (0.5,0)=0
FP (2nd negative)	48/(59,98) $\tilde{\psi}_{wil2}$ =1+0.5=1.5	48/(79,89) $\tilde{\psi}_{wil2}$ =1+0.5=1.5	48/74 $\tilde{\psi}_{wil2}$ =1+0=1	48/NA $\tilde{\psi}$ (0.5,0)=0
FP (3rd negative)	(56,67)/(59,98) $\tilde{\psi}_{wil2}$ =0.75+0=0.75	(56,67)/(79,89) $\tilde{\psi}_{wil2}$ =1+0=1	(56,67)/74 $\tilde{\psi}_{wil2}$ =1-0.25=0.75	(56,67)/NA $\tilde{\psi}$ (0.5,0)=0
FP (4th negative)	NA/(59,98) $\tilde{\psi}$ (0.5,1)=1	NA/(79,89) $\tilde{\psi}$ (0.5,1)=1	NA/74 $\tilde{\psi}$ (0.5,1)=1	NA/NA $\tilde{\psi}$ (0.5,0.5)=0.5
FP (1st positive)	NA/(59,98) $\tilde{\psi}$ (0.5,1)=1	NA/(79,89) $\tilde{\psi}$ (0.5,1)=1	NA/74 $\tilde{\psi}$ (0.5,1)=1	NA/NA $\tilde{\psi}$ (0.5,0.5)=0.5
FP (2nd positive)	54/(59,98) $\tilde{\psi}_{wil2}$ =1+0.5=1.5	54/(79,89) $\tilde{\psi}_{wil2}$ =1+0.5=1.5	54/74 $\tilde{\psi}_{wil2}$ =1+0=1	54/NA $\tilde{\psi}$ (0.5,0)=0
FP (3rd positive)	(45,87)/(59,98) $\tilde{\psi}_{wil2}$ =0.75+0=0.75	(45,87)/(79,89) $\tilde{\psi}_{wil2}$ =0.75+0=0.75	(45,87)/74 $\tilde{\psi}_{wil2}$ =0.5-0.25=0.25	(45,87)/NA $\tilde{\psi}$ (0.5,0)=0

FP (4th positive)	NA/(59,98) $\tilde{\psi}(0.5,1)=1$	NA/(79,89) $\tilde{\psi}(0.5,1)=1$	NA/74 $\tilde{\psi}(0.5,1)=1$	NA/NA $\tilde{\psi}(0.5,0.5)=0.5$

$$\widehat{SPLIT}_{(2)}=\frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\left\{\frac{\sum_{c_1=1}^{n_{s_1}}\sum_{c_2=1}^{m_s}\tilde{\psi}\left(x_{s_1c_1},y_{sc_2}\right)}{n_{s_1}\times m_s}+\frac{1}{2}\left(\frac{m_s}{n_{s_1}}-1\right)\right\}=\frac{24}{32}=0.750$$

$$k=4,\ \widehat{SPLIT}_{(4)}=\frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\left\{\frac{\sum_{c_1=1}^{n_{s_1}}\sum_{c_2=1}^{m_s}\tilde{\psi}\left(x_{s_1c_1},y_{sc_2}\right)}{n_{s_1}\times m_s}+\frac{1}{4}\left(\frac{m_s}{n_{s_1}}-1\right)\right\}$$

$\tilde{\psi}_{wil}$	*TP* (1st positive)	*TP* (2nd positive)	*TP* (3rd positive)	*TP* (4th positive)
FP (1st negative)	98/(59,98) $\tilde{\psi}_{wil4}$=0.25+0.25=0.5	98/(79,89) $\tilde{\psi}_{wil4}$=0+0.25=0.25	98/74 $\tilde{\psi}_{wil4}$=0+0=0	98/NA $\tilde{\psi}(0.5,0)=0$
FP (2nd negative)	48/(59,98) $\tilde{\psi}_{wil4}$=1+0.25=1.25	48/(79,89) $\tilde{\psi}_{wil4}$=1+0.25=1.25	48/74 $\tilde{\psi}_{wil4}$=1+0=1	48/NA $\tilde{\psi}(0.5,0)=0$
FP (3rd negative)	(56,67)/(59,98) $\tilde{\psi}_{wil4}$=0.75+0=0.75	(56,67)/(79,89) $\tilde{\psi}_{wil4}$=1+0=1	(56,67)/74 $\tilde{\psi}_{wil4}$=1-0.13=0.87	(56,67)/NA $\tilde{\psi}(0.5,0)=0$
FP (4th negative)	NA/(59,98) $\tilde{\psi}(0.5,1)=1$	NA/(79,89) $\tilde{\psi}(0.5,1)=1$	NA/74 $\tilde{\psi}(0.5,1)=1$	NA/NA $\tilde{\psi}(0.5,0.5)=0.5$
FP (1st positive)	NA/(59,98) $\tilde{\psi}(0.5,1)=1$	NA/(79,89) $\tilde{\psi}(0.5,1)=1$	NA/74 $\tilde{\psi}(0.5,1)=1$	NA/NA $\tilde{\psi}(0.5,0.5)=0.5$
FP (2nd positive)	54/(59,98) $\tilde{\psi}_{wil4}$=1+0.25=1.25	54/(79,89) $\tilde{\psi}_{wil4}$=1+0.25=1.25	54/74 $\tilde{\psi}_{wil4}$=1+0=1	54/NA $\tilde{\psi}(0.5,0)=0$
FP (3rd positive)	(45,87)/(59,98) $\tilde{\psi}_{wil4}$=0.75+0=0.75	(45,87)/(79,89) $\tilde{\psi}_{wil4}$=0.75+0=0.75	(45,87)/74 $\tilde{\psi}_{wil4}$=0.5-0.13=0.37	(45,87)/NA $\tilde{\psi}(0.5,0)=0$
FP (4th positive)	NA/(59,98) $\tilde{\psi}(0.5,1)=1$	NA/(79,89) $\tilde{\psi}(0.5,1)=1$	NA/74 $\tilde{\psi}(0.5,1)=1$	NA/NA $\tilde{\psi}(0.5,0.5)=0.5$

$$\widehat{SPLIT}_{(4)}=\frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\left\{\frac{\sum_{c_1=1}^{n_{s_1}}\sum_{c_2=1}^{m_s}\tilde{\psi}\left(x_{s_1c_1},y_{sc_2}\right)}{n_{s_1}\times m_s}+\frac{1}{4}\left(\frac{m_s}{n_{s_1}}-1\right)\right\}=\frac{22.75}{32}=0.711$$

$$k=8,\ \widehat{SPLIT}_{(8)}=\frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\left\{\frac{\sum_{c_1=1}^{n_{s_1}}\sum_{c_2=1}^{m_s}\tilde{\psi}\left(x_{s_1c_1},y_{sc_2}\right)}{n_{s_1}\times m_s}+\frac{1}{8}\left(\frac{m_s}{n_{s_1}}-1\right)\right\}$$

$\tilde{\psi}_{wil}$	*TP* (1st positive)	*TP* (2nd positive)	*TP* (3rd positive)	*TP* (4th positive)

FP (1st negative)	98/(59,98) $\tilde{\psi}_{wil8}$ =.25+0.13=0.38	98/(79,89) $\tilde{\psi}_{wil8}$ =0+0.13=0.13	98/74 $\tilde{\psi}_{wil8}$ =0+0=0	98/NA $\tilde{\psi}$ (0.5,0)=0
FP (2nd negative)	48/(59,98) $\tilde{\psi}_{wil8}$ =1+0.13=1.13	48/(79,89) $\tilde{\psi}_{wil8}$ =1+0.13=1.13	48/74 $\tilde{\psi}_{wil8}$ =1+0=1	48/NA $\tilde{\psi}$ (0.5,0)=0
FP (3rd negative)	(56,67)/(59,98) $\tilde{\psi}_{wil8}$ =0.75+0=0.75	(56,67)/(79,89) $\tilde{\psi}_{wil8}$ =1+0=1	(56,67)/74 $\tilde{\psi}_{wil8}$ =1-0.06=0.94	(56,67)/NA $\tilde{\psi}$ (0.5,0)=0
FP (4th negative)	NA/(59,98) $\tilde{\psi}$ (0.5,1)=1	NA/(79,89) $\tilde{\psi}$ (0.5,1)=1	NA/74 $\tilde{\psi}$ (0.5,1)=1	NA/NA $\tilde{\psi}$ (0.5,0.5)=0.5
FP (1st positive)	NA/(59,98) $\tilde{\psi}$ (0.5,1)=1	NA/(79,89) $\tilde{\psi}$ (0.5,1)=1	NA/74 $\tilde{\psi}$ (0.5,1)=1	NA/NA $\tilde{\psi}$ (0.5,0.5)=0.5
FP (2nd positive)	54/(59,98) $\tilde{\psi}_{wil8}$ =1+0.13=1.13	54/(79,89) $\tilde{\psi}_{wil8}$ =1+0.13=1.13	54/74 $\tilde{\psi}_{wil8}$ =1+0=1	54/NA $\tilde{\psi}$ (0.5,0)=0
FP (3rd positive)	(45,87)/(59,98) $\tilde{\psi}_{wil8}$ =0.75+0=0.75	(45,87)/(79,89) $\tilde{\psi}_{wil8}$ =0.75+0=0.75	(45,87)/74 $\tilde{\psi}_{wil8}$ =0.5-0.06=0.44	(45,87)/NA $\tilde{\psi}$ (0.5,0)=0
FP (4th positive)	NA/(59,98) $\tilde{\psi}$ (0.5,1)=1	NA/(79,89) $\tilde{\psi}$ (0.5,1)=1	NA/74 $\tilde{\psi}$ (0.5,1)=1	NA/NA $\tilde{\psi}$ (0.5,0.5)=0.5

$$\widehat{SPLIT}_{(8)} = \frac{1}{(S_0+S_t)\times S_t}\sum_{s_1=1}^{S_0+S_t}\sum_{s=1}^{S_t}\left\{\frac{\sum_{c_1=1}^{n_{s_1}}\sum_{c_2=1}^{m_s}\tilde{\psi}\left(x_{s_1c_1}, y_{sc_2}\right)}{n_{s_1}\times m_s} + \frac{1}{8}\left(\frac{m_s}{n_{s_1}}-1\right)\right\} = \frac{18.625}{32} = 0.665$$

]Three indices using SWITCH method in (4.2):

$$k=2,\ \widehat{SWITCH}_{(2)} = \frac{1}{S_0\times S_t}\sum_{s_1=1}^{S_0}\sum_{s=1}^{S_t}\left\{\frac{\sum_{c_1=1}^{n_{s_1}^0}\sum_{c_2=1}^{n_s^t}\sum_{c_3=1}^{m_s}\left[\tilde{\psi}(x_{s_1c_1}^0, y_{sc_3}) + \tilde{\psi}(x_{sc_2}^t, y_{sc_3})\right]}{(n_{s_1}^0+n_s^t)\times m_s} + \frac{1}{2}\left(\frac{m_s}{n_{s_1}^0+n_s^t}-1\right)\right\}$$

$\tilde{\psi}_{wil}$	1st positive subject	2nd positive subject	3rd positive subject	4th positive subject
1st negative subject	98/(59,98) $\tilde{\psi}_{wil2}$ =.25+0.5=0.75	(98,54)/(79,89) $\tilde{\psi}_{wil2}$ =0.5+0=0.5	(98,45,87)/74 $\tilde{\psi}_{wil2}$ =0.33-0.33=0	98/NA $\tilde{\psi}$ (0.5,0)=0
2nd negative subject	48/(59,98) $\tilde{\psi}_{wil2}$ =1+0.5=1.5	(48,54)/(79,89) $\tilde{\psi}_{wil2}$ =1+0=1	(48,45,87)/74 $\tilde{\psi}_{wil2}$ =.67-0.33=0.33	48/NA $\tilde{\psi}$ (0.5,0)=0
3rd negative subject	(56,67)/(59,98) $\tilde{\psi}_{wil2}$ =0.75+0=0.75	(56,67,54)/(79,89) $\tilde{\psi}_{wil2}$ =1-0.17=0.83	(56,67,45,87)/74 $\tilde{\psi}_{wil2}$ =.75-0.38=0.38	(56,67)/NA $\tilde{\psi}$ (0.5,0)=0
4th negative subject	NA/(59,98) $\tilde{\psi}$ (0.5,1)=1	54/(79,89) $\tilde{\psi}_{wil2}$ =1+0.5=1.5	(45,87)/74 $\tilde{\psi}_{wil2}$ =0.5-0.25=0.25	NA/NA $\tilde{\psi}$ (0.5,0.5)=0.5

$$\widehat{SWITCH}_{(2)} = \frac{1}{S_0\times S_t}\sum_{s_1=1}^{S_0}\sum_{s=1}^{S_t}\left\{\frac{\sum_{c_1=1}^{n_{s_1}^0}\sum_{c_2=1}^{n_s^t}\sum_{c_3=1}^{m_s}\left[\tilde{\psi}(x_{s_1c_1}^0, y_{sc_3}) + \tilde{\psi}(x_{sc_2}^t, y_{sc_3})\right]}{(n_{s_1}^0+n_s^t)\times m_s} + \frac{1}{2}\left(\frac{m_s}{n_{s_1}^0+n_s^t}-1\right)\right\} = \frac{9.292}{4\times 4} = 0.581$$

k=4, $\widehat{SWITCH}_{(4)} = \frac{1}{S_0 \times S_t} \sum_{s_1=1}^{S_0} \sum_{s=1}^{S_t} \left\{ \frac{\sum_{c_1=1}^{n_{s_1}^0} \sum_{c_2=1}^{n_s^t} \sum_{c_3=1}^{m_s} \left[\tilde{\psi}(x_{s_1 c_1}^0, y_{s c_3}) + \tilde{\psi}(x_{s c_2}^t, y_{s c_3}) \right]}{(n_{s_1}^0 + n_s^t) \times m_s} + \frac{1}{4} \left(\frac{m_s}{n_{s_1}^0 + n_s^t} - 1 \right) \right\}$

$\tilde{\psi}_{wil}$	1st positive subject	2nd positive subject	3rd positive subject	4th positive subject
1st negative subject	98/(59,98) $\tilde{\psi}_{wil4}$=.25+0.25=0.5	(98,54)/(79,89) $\tilde{\psi}_{wil4}$=0.5+0=0.5	(98,45,87)/74 $\tilde{\psi}_{wil4}$=.33-0.17=0.17	98/NA $\tilde{\psi}$(0.5,0)=0
2nd negative subject	48/(59,98) $\tilde{\psi}_{wil4}$=1+0.25=1.25	(48,54)/(79,89) $\tilde{\psi}_{wil4}$=1+0=1	(48,45,87)/74 $\tilde{\psi}_{wil4}$=.67-0.17=0.5	48/NA $\tilde{\psi}$(0.5,0)=0
3rd negative subject	(56,67)/(59,98) $\tilde{\psi}_{wil4}$=0.75+0=0.75	(56,67,54)/(79,89) $\tilde{\psi}_{wil4}$=1-0.08=0.92	(56,67,45,87)/74 $\tilde{\psi}_{wil4}$=.75-0.19=0.56	(56,67)/NA $\tilde{\psi}$(0.5,0)=0
4th negative subject	NA/(59,98) $\tilde{\psi}$(0.5,1)=1	54/(79,89) $\tilde{\psi}_{wil4}$=1+0.25=1.25	(45,87)/74 $\tilde{\psi}_{wil4}$=0.5-0.13=0.37	NA/NA $\tilde{\psi}$(0.5,0.5)=0.5

$$\widehat{SWITCH}_{(4)} = \frac{1}{S_0 \times S_t} \sum_{s_1=1}^{S_0} \sum_{s=1}^{S_t} \left\{ \frac{\sum_{c_1=1}^{n_{s_1}^0} \sum_{c_2=1}^{n_s^t} \sum_{c_3=1}^{m_s} \left[\tilde{\psi}(x_{s_1 c_1}^0, y_{s c_3}) + \tilde{\psi}(x_{s c_2}^t, y_{s c_3}) \right]}{(n_{s_1}^0 + n_s^t) \times m_s} + \frac{1}{4} \left(\frac{m_s}{n_{s_1}^0 + n_s^t} - 1 \right) \right\} = \frac{9.271}{4 \times 4} = 0.579$$

k=8, $\widehat{SWITCH}_{(8)} = \frac{1}{S_0 \times S_t} \sum_{s_1=1}^{S_0} \sum_{s=1}^{S_t} \left\{ \frac{\sum_{c_1=1}^{n_{s_1}^0} \sum_{c_2=1}^{n_s^t} \sum_{c_3=1}^{m_s} \left[\tilde{\psi}(x_{s_1 c_1}^0, y_{s c_3}) + \tilde{\psi}(x_{s c_2}^t, y_{s c_3}) \right]}{(n_{s_1}^0 + n_s^t) \times m_s} + \frac{1}{8} \left(\frac{m_s}{n_{s_1}^0 + n_s^t} - 1 \right) \right\}$

$\tilde{\psi}_{wil}$	1st positive subject	2nd positive subject	3rd positive subject	4th positive subject
1st negative subject	98/(59,98) $\tilde{\psi}_{wil8}$=.25+.13=0.38	(98,54)/(79,89) $\tilde{\psi}_{wil8}$=0.5+0=0.5	(98,45,87)/74 $\tilde{\psi}_{wil8}$=.33-0.08=0.25	98/NA $\tilde{\psi}$(0.5,0)=0
2nd negative subject	48/(59,98) $\tilde{\psi}_{wil8}$=1+0.13=1.13	(48,54)/(79,89) $\tilde{\psi}_{wil8}$=1+0=1	(48,45,87)/74 $\tilde{\psi}_{wil8}$=.67-0.08=0.58	48/NA $\tilde{\psi}$(0.5,0)=0
3rd negative subject	(56,67)/(59,98) $\tilde{\psi}_{wil8}$=0.75+0=0.75	(56,67,54)/(79,89) $\tilde{\psi}_{wil8}$=1-0.04=0.96	(56,67,45,87)/74 $\tilde{\psi}_{wil8}$=.75-0.09=0.66	(56,67)/NA $\tilde{\psi}$(0.5,0)=0
4th negative subject	NA/(59,98) $\tilde{\psi}$(0.5,1)=1	54/(79,89) $\tilde{\psi}_{wil8}$=1+0.13=1.13	(45,87)/74 $\tilde{\psi}_{wil8}$=0.5-0.06=0.44	NA/NA $\tilde{\psi}$(0.5,0.5)=0.5

$$\widehat{SWITCH}_{(8)} = \frac{1}{S_0 \times S_t} \sum_{s_1=1}^{S_0} \sum_{s=1}^{S_t} \left\{ \frac{\sum_{c_1=1}^{n_{s_1}^0} \sum_{c_2=1}^{n_s^t} \sum_{c_3=1}^{m_s} \left[\tilde{\psi}(x_{s_1 c_1}^0, y_{s c_3}) + \tilde{\psi}(x_{s c_2}^t, y_{s c_3}) \right]}{(n_{s_1}^0 + n_s^t) \times m_s} + \frac{1}{8} \left(\frac{m_s}{n_{s_1}^0 + n_s^t} - 1 \right) \right\} = \frac{9.260}{4 \times 4} = 0.579$$

BIBLIOGRAPHY

1. Zhou XH, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. *Wiley & Sons Inc.*: New York, 2002.
2. Metz CE. ROC methodology in radiologic imaging. *Investigative Radiology* 1986; 21(9), 720-733.
3. Campbell G. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine* 1994; 13, 499-508.
4. Metz CE. Basic Principles of ROC analysis. *Seminars in Nuclear Medicine* 1978; 8(4), 283-298.
5. Lusted LB. Signal detectability and medical decision-making. *Science* 1971; 171, 1217–1219.
6. Lusted LB. ROC recollected (editorial). *Medical Decision Making* 1984; 4, 131–135.
7. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; 12, 387-415.
8. Hanley JA, McNeil BJ. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 1982; 143(11), 29-36.
9. Zou KH, Tempany CM, Fielding JR, Silverman SG. Original smooth receiver operating characteristic curves estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of urethral stones. *Academic Radiology* 1998: 5, 680-687.
10. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Area under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 1988; 44(3), 837-845.
11. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics* 1997; 53, 567-578.
12. Rosner B, Grove D. Use of the Mann-Whitney U- test for clustered data. *Statistics in Medicine* 1999; 18, 1387-1400.
13. Lee MT, Rosner B. The Average Area under Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach Based on Generalized Two-Sample Wilcoxon statistic. *Applied Statistics* 2001, 50(3), 337-344.
14. Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* 1997; 16, 2143–2156.

15. Zou KH, Hall WJ. Two transformation models for estimating an ROC curve derived from continuous data *Journal of Applied Statistics* 2000; 27(5), 621-631

16. Lloyd CJ. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association* 1998; 93, 1356–1364.

17. Lloyd CJ, Yong Z. Kernel estimators of the ROC curves are better than empirical. *Statistics and Probability Letters* 1999; 44, 221-228.

18. Faraggi D, Reiser B. Estimation of the area under the ROC curve. *Statistics in Medicine* 2002; 21: 3093-3106.

19. Arvesen JN. Jackknifing U-statistics. *Annals of Mathematical Statistics* 1969; 40, 2076–2100.

20. Starr SJ, Metz CE, Lusted LB, et al. Visual detection and localization of radiographic images. *Radiology* 1975; 116, 533–538.

21. Swensson, RG. Unified measurement of observer performance in detecting and localizing target objects on images. *Medical Physics* 1996; 23, 1709–25.

22. Obuchowski NA, Lieber ML, Powell KA. Data analysis for detection and localization of multiple abnormalities with application to mammography. *Academic Radiology* 2000; 7,516–525.

23. Chakraborty, DP. Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. *Medical Physics* 1989; 16(4), 561-568.

24. Rutter CM. Bootstrap estimation of diagnostic accuracy with patient clustered data. *Academic Radiology* 2000; 7, 413–419.

25. Chakraborty D, Obuchowski NA, Lieber ML, et al. Point-counterpoint. *Academic Radiology* 2000; 7, 553–556.

26. Egan JP, Greenberg GZ, Schulman AI. Operating characteristics, signal delectability, and the methods of free response. *Journal of the Acoustical Society of America* 1961; 33(8), 993-1007.

27. Bunch, PC, Hamilton, JF, Sanderson, GK and Simmons, AH. A free response approach to the measurement and characterization of radiographic-observer performance. *Journal of Applied Photographic Engineering* 1978; 4(4), 165-171.

28. Nishikawa RM, Yarusso LM. Variations in measured performance of CAD schemes due to database composition and scoring protocol. *Proc SPIE* 1998; 3338:840–844.

29. Kallergi M, Carney GM, Gaviria J. Evaluating the performance of detection algorithms in digital mammography. *Medical Physics* 1999; 26, 267–275.

30. Yoon HJ, Zheng B, Sahiner B, Chakraborty DP. Evaluating computer-aided detection algorithms. *Medical Physics* 2007; 34, 2024–2038.

31. Edwards DC, Kupinski MA, Metz CE, Nishikawa RM. Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. *Medical Physics* 2002; 29(12), 2861-2870.

32. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. *Physics in Medicine and Biology* 2006; 51(14), 3449-3462.

33. Bandos A, Rockette H, Song T, Gur D. Area under the FROC curve and a related summary index. *Biometrics*; under revision.

34. Samuelson FW, Petrick N. Comparing image detection algorithms using re-sampling. *Biomedical Imaging: Macro to Nano, 3rd IEEE International Symposium* 2006; 1312-1315.

35. Chakraborty DP. ROC curves predicted by a model of visual search. *Phys Med Biol* 2006; 51, 3463–3482.

36. Chakraborty DP, Berbaum KS. "Observer studies involving detection and localization: Modeling, analysis and validation." *Medical Physics* 2004; 31(8), 2313-2330.

37. Chakraborty DP. Statistical power in observer performance studies: a comparison of the ROC and free-response methods in tasks involving localization. *Academic Radiology* 2002; 9, 147–156.

38. Chakraborty DP. Analysis of Location Specific Observer Performance Data: Validated Extensions of the Jackknife Free-Response (JAFROC) Method. *Academic Radiology* 2006; 13, 1187–1193

39. Roe C and Metz CE, ‘‘Dorfman–Berbaum–Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: Validation with computer simulation.’’ *Academic Radiology* 1997; 4, 298–303.

40. Beam C, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. *Arch Intern Med* 1996; 156:209–213

41. Berbaum KS, Franken EA, Dorfman DD, Rooholamini SA, Kathol MH, Barloon TJ, Behlke FM, Sato Y, Lu CC, El-Khoury GY, Flickinger FW and Montgomery WJ. Satisfaction of search in diagnostic radiology. *Investigative Radiology* 1990; 25, 133–140.

42. Wagner FW, Metz CE and Compbell G. Assessment of Medical Imaging Systems and Computer Aids: A Tutorial Review. *Academic Radiology* 2007; 14, 723-748.

43. Bandos A. Nonparametric methods in comparing two correlated ROC curves. PhD Dissertation 2005, *University of Pittsburgh*.

44. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists’ interpretations of mammograms. *N Engl J Med* 1994; 331,1493–1499.

45. D’Orsi CJ, Swets JA. Variability in the interpretation of mammograms (letter). *N Engl J Med* 1995; 332,1172.

46. Metz CE, Shen JH. Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis. *Medical Decision Making* 1992; 12, 60–75.

47. Obuchowski NA. Multi-reader multi-modality ROC studies: hypothesis testing and sample size estimation using an ANOVA approach with dependent observations with rejoinder. *Academic Radiology* 1995; 2(suppl), S22–S29.

48. Rockette HE, Campbell WL, Britton CA, *et al*. Empiric assessment of parameters that affect the design of multireader receiver operating characteristic studies. *Academic Radiology* 1999; 6, 723–729.

49. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. *Investigative Radiology* 1992; 27, 723-731.

50. Obuchowski NA, Rockette HE. Hypothesis testing of the diagnostic accuracy for multiple diagnostic tests: an ANOVA approach with dependent observations. *Communications in Statistics: Simulations and Computations* 1995; 24, 285-308.

51. Song HH. Analysis of correlated ROC areas in diagnostic testing. *Biometrics* 1997; 53(1), 370-382.

52. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative method for random effects receiver operating characteristic analysis. *Academic Radiology* 2000; 7, 341–349.

53. Gallas B. One-shot estimate of MRMC variance: AUC. *Academic Radiology* 2006; 13, 353-362.

54. Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette Methods for receiver operating characteristic (ROC) data. *Statistics in Medicine* 2005; 24, 1579–1607.

55. Hillis SL. A Comparison of Denominator Degrees of Freedom Methods for Multiple Observer ROC Analysis. *Statistics in Medicine* 2007; 26(3), 596–619.

56. Box GEP, Cox DR. An analysis of transformations. *Journal of the Royal Statistical Society: Series B* 1964; 42, 71-78.

57. Wand MP, Jones MC. Kernel Smoothing. *Chapman & Hall/CRC*, 1995.

58. Bandos A, Rockette HE, Gallas BD, Gur D. Multi-Reader ROC methods: Explicit formulations and relationships between DBM, multi-WMW, BWC and Gallas's methods. *International Biometric Society, Eastern North-American Region Conference (ENAR)*, Atlanta Georgia, 2007.