EVALUATION OF DIAGNOSTIC PERFORMANCE USING PARTIAL AREA UNDER THE ROC CURVE

by

Hua Ma

B.S. Sichuan Normal University, Chengdu, China, 2007

M.S. Xiamen University, Xiamen, China, 2010

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

2014

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Hua Ma

It was defended on

May 8, 2014

and approved by

Howard E. Rockette, PhD, Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Jong-Hyeon Jeong, PhD, Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

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

Chung-Chou Chang, PhD, Associate Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Dissertation Director: Andriy Bandos, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Andriy Bandos, PhD

EVALUATION OF DIAGNOSTIC PERFORMANCE USING PARTIAL AREA UNDER THE ROC CURVE

Hua Ma, PhD

University of Pittsburgh, 2014

ABSTRACT

Evaluation of diagnostic performance is critical in many fields including but not limited to diagnostic medicine. The Receiver Operating Characteristic (ROC) curve is the most widely used methodology for describing the intrinsic performance of diagnostic tests, with the area under the curve (AUC) being the most commonly used summary index of overall performance. The partial area under the ROC curve (pAUC), when focused on the range of practical/clinical relevance, is considered a more relevant summary index than the full AUC. However, several conceptual and analytical difficulties frequently prevent the pAUC from being used. First, in many diagnostic setting the relevant range is difficult to determine objectively. Second, in theory, due to potential use of less information, analysis based on the pAUC could lead to the loss of statistical precision and therefore would require larger sample sizes. Through mathematical derivation, extensive simulation studies and practical examples, this work investigates statistical properties when using the pAUC. First, this work demonstrates that in many practical scenarios inferences based on pAUC could be more powerful than inferences based on the full AUC. Thus, the use of the pAUC may lead to not only more clinically relevant but also more conclusive results in analyses of experimental data. Second, this investigation demonstrates that the advantages of pAUC-based inferences depend on the shape of ROC curves. The conventional binormal model does not always adequately describe scenarios where the pAUC is more

statistically efficient. The bi-gamma family of concave ROC curves is shown to describe practically reasonable scenarios where either pAUC or full AUC could be advantageous. Programs for sample size estimation based on bi-gamma model are then developed. Finally, this work investigates the properties of pAUC-based inferences in scenarios where diagnostic results have substantial ties (or a “mass”) at the lowest diagnostic results. For certain type of the ROC curves the existence of ties could lead to an increase in statistical efficiency. Forcing a diagnostic system to resolve ties could detrimentally affect reliability and conclusiveness of statistical inferences. In conclusion, this work provides investigators with insights into and tools for generating practically relevant conclusions using pAUC. The public health importance of this work stems from the relevance of the ROC analysis at different stages of development and regulatory approval of diagnostic systems in medicine. Enhanced methodology for evaluation of diagnostic accuracy helps in the development of improved diagnostic systems and could accelerate the delivery and clinical adoption of truly beneficial diagnostic technologies and/or clinical practices.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 BACKGROUND

A basic problem in evaluation of diagnostic performance involves assessment of the accuracy of a diagnostic test in identifying a patient with a specific, predefined condition (abnormal subject) and a patient without the condition (normal subject). The true status (presence or absence of the abnormality in question) of a subject is assumed to be known for all subjects used for accuracy evaluation. The diagnostic test results can be measured using a binary scale indicating that the subject is assessed as “positive” or “negative”, or an ordinal multi-category (e.g. 7) scale typically with larger values representing higher probability of the abnormality being present, or a continuous scale indicating the likelihood of a pre-specified abnormality being present. For a multi-category diagnostic test, a subject can be classified into a “positive” or “negative” class according to whether the test result is greater than or less than a pre-specified threshold. The Receiver Operating Characteristic (ROC) analysis is the most widely used methodology to investigate this type of research objectives.

ROC analysis originated from signal detection theory (Green and Swets, 1966) (Egan, 1975) and has been well developed over the past 50 years in particular as related to diagnostic imaging and decision making (Metz, 1989) (Hanley, 1989) (McNeil *et al.*, 1975) (Zhou *et al.*, 2002). However, many issues remain and new methods are constantly being developed.

The ROC curve is the plot of sensitivity versus 1-specificity for all possible decision threshold values of c (Figure 1.1). Let X and Y denote the ratings for normal and abnormal subjects respectively. Sensitivity, or true positive fraction (TPF), is the probability of test results being positive for abnormal subjects, and can be defined as follows:

$$sensitivity(c) = TPF(c) = \Pr(Y > c)$$

Specificity, or true negative fraction (TNF), is the probability of test results being negative for normal subjects, and can be defined as follows:

$$specificity(c) = TNF(c) = \Pr(X \le c)$$

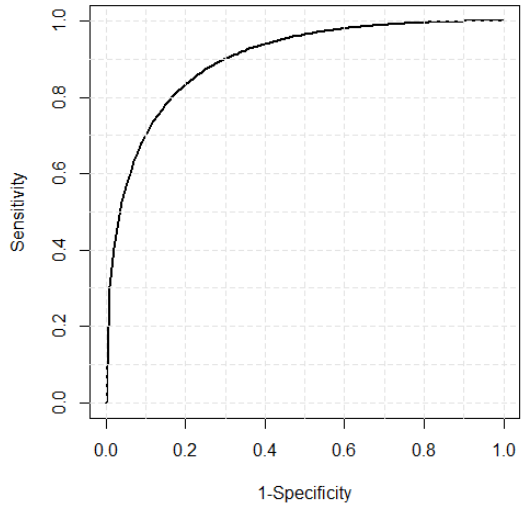


Figure 1.1 ROC curve

The most commonly used summary index associated with the ROC curve is the area under the ROC curve (AUC). It is defined as

$$A = \int_0^1 ROC(f)\,df$$

The AUC has several interpretations. First, it can be interpreted as the weighted average value of sensitivity of all possible values of specificity (Zhou *et al.*, 2002), or the weighted average value of specificity of all possible values of sensitivity (Metz, 1989). Second, it can be

also interpreted as the probability that a test result for a randomly selected abnormal patient indicates a greater suspicion of disease than the test result for a randomly selected normal patient (Hanley and McNeil, 1982) (Bamber, 1975). If X and Y are continuous (i.e., no ties in results are possible), then the AUC can be defined as $\Pr(Y > X)$. The value of AUC of 1 indicates a perfect system whereas the non-informative (i.e., guessing) diagnostic system would have AUC of 0.5. An unbiased non-parametric AUC estimate is the area under the empirical ROC curve which is the same as the Mann-Whitney form of the two-sample Wilcoxon rank-sum statistic (Shapiro, 1999). A number of parametric and non-parametric methods based on AUC have been developed to make statistical inferences (Zhou *et al.*, 2002) (Pepe, 2003).

AUC offers a single value to indicate the accuracy of diagnostic performance by considering both sensitivity and specificity across all possible threshold values; its major limitation is that it summarizes the entire ROC curve including the region which may not be of interest or practical relevance, for example, the region with very low specificity levels.

To remedy this limitation, partial area under the ROC curve (pAUC) can be used to describe the intrinsic accuracy of diagnostic tests in the range of practical (clinical) interest. The pAUC is frequently defined as

$$A_{e_1,e_2} = \int_{e_1}^{e_2} ROC(f)\,df$$

In practice, a range of $(0,e)$ is often used due to the importance of high-specificity range in practice (McClish, 1989) (Jiang *et al.*, 1996). Since the pAUC over an arbitrary interval (e_1,e_2) is equivalent to the difference in $\text{pAUC}_{(0,e2)}$ and $\text{pAUC}_{(0,e1)}$, $A_e = \int_0^e ROC(f)\,df$ will be in focus of my work.

A number of statistical methods and inferences based on the pAUC using both parametric and non-parametric approaches have been developed. These include the parametric estimator of the pAUC and its variance using the bi-normal model (McClish, 1989) (Jiang *et al.*, 1996). Wieand *et al.* (1989) proposed a non-parametric method for estimating pAUC and its variance. Based on Delong's approach (DeLong *et al.*, 1988), Zhang *et al.* (2002) proposed a simpler method to compute the variance of pAUC which was subsequently improved by He and Escobar (2008). An alternative nonparametric variance estimator of the pAUC using its expected value was proposed by Liu *et al*. (2005). Other non-parametric methods have been developed such as empirical likelihood methods, for comparing two pAUCs (Huang *et al.*, 2012) (Qin and Zhou, 2006) (Chen and Wong, 2009), and semi-parametric regression approaches on pAUC by Dodd and Pepe (2003) and Cai and Dodd (2008). However, several conceptual and analytical difficulties prevent pAUC from being widely used.

In general, parametric approaches offer improved efficiency of statistical inferences, but could introduce substantial bias if the needed parametric assumptions are not satisfied. Under the correctly specified model the relative efficiency of nonparametric estimates of partial AUC can be as low as 50% for short ranges of interest (e.g., 0-0.05), but increase beyond 80% efficiency for ranges wider than 0-0.2 (Dodd and Pepe, 2003). For the full AUC, results of parametric and nonparametric inferences are very similar (Hajian-Tilaki *et al.*, 1997) (Hajian-Tilaki and Hanley, 2002). However, it is not always easy to verify appropriateness of the parametric assumptions, and for mis-specified models parametric estimates of pAUC could easily have bias as high as 40% (e.g., Dodd and Pepe, 2003). For this reason it is often recommended to use non-parametric approaches for inferences about partial AUC (Dodd and Pepe, 2003) (Zhang *et al.*, 2002) (He and Escobar, 2008). Non-parametric analysis of pAUC is in primary focus of this work as well.

One of these difficulties is that the scale of values of pAUC increases with increasing range of interest. To partially overcome this limitation, several partial area indices have been proposed (Zhou *et al.*, 2002) (McClish, 1989). A natural transformation of the partial area aimed to "standardize" the range of its values can be written as follows (McClish, 1989):

$$\tilde{A}_e = \frac{1}{2}\left(1+\frac{A_e - e^2/2}{e - e^2/2}\right) = \frac{1}{2}\left(1+\frac{\int_0^e ROC(f)df - e^2/2}{e - e^2/2}\right) \quad (1.1)$$

Here, we term this index as the "standardized partial AUC" (spAUC). For ROC curves describing better-than-chance performance, $\tilde{A}_e$ varies from 0.5 to 1 regardless of *e*, and for *e*=1 it reduces to the conventional AUC.

Second, the relevant range should be pre-specified during study design but it is often difficult to determine a priori. In addition, it is often assumed that because of the use of less information, analysis based on the pAUC may result in the loss of statistical precision as compared with statistical inferences based on the full AUC, and thus its use may require larger sample sizes (Zhou *et al.*, 2002) (Obuchowski and McClish, 1997) (Wieand *et al.*, 1989). Conjectures about the relative stability of the spAUC with respect to the range of interest and the decrease in variance with increasing range are intuitively appealing and could affect the way statistical analysis is planned and interpreted. In analyzing experimentally ascertained datasets from observer performance studies we frequently encountered scenarios that contradicted the two conjectures. The work presented here primarily focuses on the investigation of properties of statistical inferences based on the pAUC.

In diagnostic radiology, it is natural to observe multiple subjects having the same diagnostic test results (a tie), in particular at the lowest range, even when the original scale is

continuous or pseudo-continuous (e.g. 0-100 confidence rating scale). A tie at the lower rating could reflect an important characteristic such as the prevalence of the obviously "normal" subjects (e.g., chest images) in a sample, or frequency of the natural absence of a tested substance (Schisterman *et al.*, 2006), or assigning default value to subjects with biomarker levels below a certain limit of concentration and/or a limit of detection (Perkins *et al.*, 2007). When these multiple ties occur at the lowest rating level, the ROC curve includes a straight line segment joining the point corresponding to the lowest threshold and the corner point (1, 1). Since this type of test results has a spike (mass) at the lower threshold, for brevity we term such a curve as an "ROC curve with mass". For the ROC curve with mass, a parametric mixed model combined with Box-Cox transformation and a non-parametric approach based on the Mann-Whitney statistic for the estimation of AUC has been proposed and discussed (Schisterman *et al.*, 2006). The parametric mixed model approach can be further used to estimate Youden's Index and determine the optimal threshold for test results with mass (Schisterman *et al.*, 2008). However, issues related to the evaluation of a single pAUC and the comparisons of two correlated pAUCs associated with ROC curves with mass remain unsolved to date.

1.2 OBJECTIVES

The emphasis of this dissertation will be on investigations of statistical properties when evaluating diagnostic performance using pAUC. We believe that in many practical scenarios inferences based on pAUC could be no less statistically advantageous than inferences based on the full AUC. Thus the use of pAUC may actually lead to not only more relevant but also more conclusive results in analyses of experimental data and/or require smaller sample sizes in planed

studies. This should encourage researchers and practitioners to more frequently apply this highly relevant, but currently underused summary index. The results of our investigation could also provide foundation for decisions about optimal thresholds to achieve greatest statistical power and therefore smaller sample sizes when using pAUC.

This dissertation includes the following three objectives.

Objective 1:

As related to evaluation of a single diagnostic system, we investigate the effect, if any, of the range of interest $(0, e)$ on statistical inferences when the $pAUC_{(0,e)}$ is used as a summary measure of performance. We analyze the properties of nonparametric and parametric estimates of standardized pAUCs and their variances. Using extensive simulation studies, we investigate the statistical power for different families of ROC curves such as binormal ROC curves, bi-gamma ROC curves and straight-line ROC curves. Based on the results of this research, we develop a program for estimating sample size in the evaluation of a single pAUC in a range of practically relevant scenarios.

Objective 2:

We extend the developments from objective 1 for the task of comparison of accuracy levels of two diagnostic systems on the basis of pAUC computed from the paired data collected with each case rated under every modality. First, we analytically investigate conditions for the increasing difference in the standardized pAUC with increasing size of the range of interest. Based on extensive simulation studies, we investigate the statistical power for comparisons of pAUCs over different ranges of interest under the ROC scenarios (such as binormal ROC curves, bi-gamma ROC curves and straight-line ROC curves) which lead to different patterns of changes in pAUC with increasing range. Based on the result of this research, we develop a program for

estimating sample size for comparison of two correlated pAUCs for a variety of practical scenarios.

Objective 3:

The task of evaluation of diagnostic modalities is often complicated by presence of substantial ties in the data. Using mathematical considerations and extensive simulations, we investigate the properties of the differences in the pAUCs and statistical power over different ranges of interest. The expectation is that the trends of increasing/decreasing variance with increasing range of interest would become less pronounced for data with ties at the lowest rating value (corresponding to the ROC curve with mass) as compared with data without ties. This could affect the expected patterns in statistical power. The results of this investigation will help plan the analyses of diagnostic accuracy using data with ties at the lowest rating levels and make more informative decisions about the data collection protocols.

2.0 FACTS RELATED TO THE PRESEARCH

2.1 FAMILIES OF ROC CURVE, THEIR AUCS AND PAUCS

2.1.1 BINORMAL ROC CURVES

Bi-normal ROC curve is the most widely used model in ROC analysis (Zhou *et al.*, 2002). The name "binormal" reflects the shape of ROC curves and stems from the fact that "binormal" ROC curve can result from the two (independent) normally distributed random variables. However, the use of the binormal ROC curve does not necessarily imply that the test results are assumed to follow normal distributions in the subpopulation of normal and abnormal patients. Rather, the use of a binormal ROC curve implies that the observed diagnostic result is related (according to a certain monotonically increasing transformation, with possible grouping for discrete case) to normally distributed latent scores.

For a pair of latent scores for normal and abnormal patients which follow two normal distributions, i.e. $X \sim N\left(\mu_x, \sigma_x^2\right)$ and $Y \sim N\left(\mu_y, \sigma_y^2\right)$ respectively, the ROC curve can be expressed as:

$$ROC\left(e\right) = \Phi\left(a + b\Phi^{-1}\left(x\right)\right)$$

where $a = \frac{\left(\mu_y - \mu_x\right)}{\sigma_y}$ $b = \frac{\sigma_x}{\sigma_y}$ and Φ is the cumulative normal distribution function. This relationship between (a,b) and the parameters of the distribution of the latent scores is rarely used in practice. One of the exceptions is to fit the ROC curve for continuous data using Box-Cox transformation (Zou and Hall, 2000); however, this relationship is very useful in simulation studies.

The AUC for the binormal ROC can be expressed as:

$$A = \int_0^1 \Phi\left(a + b\Phi^{-1}\left(x\right)\right)dx = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$$

and the pAUC as:

$$A_e = \int_0^e \Phi\left(a + b\Phi^{-1}\left(x\right)\right)dx$$

Hillis and Metz provided an analytic expression for pAUC in the case of binormal ROC curves (Hillis and Metz, 2012),

$$A_e = F_{BVN}\left(\frac{a}{\sqrt{1+b^2}}, \Phi^{-1}\left(e\right); -\frac{b}{\sqrt{1+b^2}}\right)$$

where $F_{BVN}\left(z, x; \rho\right)$ is the standardized bivariate normal distribution function with correlation ρ.

2.1.2 POWER-LAW ROC CURVES

Another well-known, but simpler and less flexible (due to a single-parameter type) family of ROC curves is described by the "power-law" curves (Egan, 1975) (Hanley, 1988), or Lehman family of the ROC curves (Gonen and Glenn, 2010). One of the reasons to consider this model was for investigating the consequences of deviation from the binormal assumption (Hanley,

1988) and enabling simple inferences using built-in software (Gonen and Heller, 2010). Power-law ROC curve can result from two exponentially distributed variables. However, the use of the power-law ROC curve does not necessarily imply that the test results are assumed to follow exponential distributions in the subpopulation of normal and abnormal patients. Rather, similar to other parametric ROC curves, the use of a power-low ROC curve implies that the observed diagnostic result is related (according to a certain monotonically increasing transformation, with possible grouping for discrete case) to exponentially distributed latent scores.

For a pair of latent scores for normal and abnormal patients which follow two exponential distributions, i.e. $X \sim Exp(\theta_x)$ and $Y \sim Exp(\theta_y)$ respectively, the ROC curve (power-law) can be expressed as:

$$ROC(e) = \exp\left(\frac{\theta_x}{\theta_y} e\right).$$

with the AUC of:

$$A = \int_0^1 \exp\left(\frac{\theta_x}{\theta_y} f\right) df = \frac{\theta_y}{\theta_x}\left(\exp\left(\frac{\theta_x}{\theta_y}\right) - 1\right),$$

and the pAUC of:

$$A_e = \int_0^1 \exp\left(\frac{\theta_x}{\theta_y} x\right) dx = \frac{\theta_y}{\theta_x}\left(\exp\left(\frac{\theta_x}{\theta_y} e\right) - 1\right).$$

2.1.3 BI-GAMMA ROC CURVES

Bi-gamma family is another of the well-known families of the ROC curves (Egan, 1975) (Dorfman *et al*., 1996) (Faraggi *et al*., 2003) (Huang and Pepe, 2009). In general bi-gamma ROC

curves constitute a three-parameter family, however, in practice a subfamily of concave curves represented by “constant-shape bi-gamma ROC curves” is used (Dorfman *et al.*, 1996). Similar to the binormal ROC curves, the constant shape bi-gamma ROC curves constitute a two-parameter family, however, it offers more flexible shapes of the practically reasonable concave ROC curves (a subfamily of concave binormal ROC curve is a one-parameter family).

The primary disadvantage of bi-gamma ROC curve lies in the relative complexity of computations. However, the computational complexity is alleviated with the development of software packages and theoretical investigations of the properties of bi-gamma ROC curves (Constantine *et al.*, 1986). A bi-gamma ROC curve can be parameterized with parameters of the gamma distribution of the latent (as opposed to actual) ratings for normal and abnormal subjects. We note that similar to other ROC models, the underlying assumption of a bi-gamma-type shape of the ROC curve does not imply an assumption of a bi-gamma distribution of the actual ratings (due to the invariance of the ROC curve with respect to monotonically increasing transformation of the ratings). In other words, the distributions of latent ratings are simply intermediate steps for parameterization of the ROC curve. The probability density function of the underlying rating model of the bi-gamma ROC curve has the following form:

$$f(x;k,\theta)=\frac{1}{\theta^{k}}\frac{1}{\tau(k)}x^{k-1}e^{-\frac{x}{\theta}},$$

In general parameters θ and κ could be different for the latent normal and abnormal ratings. The constant-shape bi-gamma ROC curves are obtained by constraining the shape parameter κ to be the same for two distributions. When κ approaches 0, the bi-gamma ROC curve approaches the shape of a straight-line and when κ>1 the shape of the bi-gamma ROC curve resembles a binormal ROC curve due to the fact that gamma distribution approaches to

normal distribution when shape parameter κ is large (We note however, that this does not guarantee convergence of the ROC curves). When κ=1 the bi-gamma ROC curve is equivalent to the power-law ROC curve (Egan, 1975) (Hanley, 1989).

For a pair of latent scores for normal and abnormal patients which follow two gamma distributions, i.e. $X \sim Gamma\left(\theta_x, \kappa_x\right)$ and $Y \sim Gamma\left(\theta_y, \kappa_y\right)$ respectively, the ROC curve can be expressed as:

$$ROC\left(e\right) = S_y\left(S_x^{-1}\left(e\right)\right).$$

The density of the Gamma distribution is given by $f(x;k,\theta) = \frac{1}{\theta^k}\frac{1}{\tau\left(k\right)}x^{k-1}e^{-\frac{x}{\theta}}$ and S denotes the survival function of Gamma distribution.

Due to the relationship between Gamma and Beta distribution the AUC of the bi-gamma ROC curve can then be expressed (Constantine *et al*., 1986) (Hussain, 2012) as:

$$A = \frac{1}{B\left(\kappa_y, \kappa_x\right)}\int_0^{\frac{\theta_y}{\theta_x+\theta_y}} x^{\kappa_x-1}\left(1-x\right)^{\kappa_y-1}dx = 1 - F_{\mathrm{F}}\left(\left(\kappa_y/\kappa_x\right)\left(\theta_x/\theta_y\right); 2\kappa_y, 2\kappa_x\right) = F_{beta}\left(\frac{\theta_y}{\theta_x+\theta_y}; \kappa_x, \kappa_y\right)$$

where $F_{\mathrm{F}}\left(*; 2\kappa_y, 2\kappa_x\right)$ is the cumulative distribution function (CDF) of an F random variable with parameters $2\kappa_y$ and $2\kappa_x$, and $F_{beta}\left(*; \kappa_x, \kappa_y\right)$ is the CDF of a beta random variable with parameters κ_x and κ_y.

As of now there are no simplified expressions for the pAUC, and it is usually computed using numerical integration according to the original definition:

$$A_e = \int_0^e S_y\left(S_x^{-1}\left(e\right)\right)dx.$$

2.1.4 STRAIGHT-LINE ROC CURVES

We define a "straight-line" ROC curve as the curve consisting of two line segments the vertical segment connecting the point (0, 0) and the point (0, 1/a), where a>1, and a line segment connecting the point (0, 1/a) and the point (1, 1). Namely:

$$ROC(e) = \frac{1}{a}e + 1 - \frac{1}{a} \tag{2.1}$$

Such a curve describes a theoretically important scenario where diagnostic result perfectly separates the most obvious "abnormal" patients, while being non-informative for discriminating between normal and abnormal patients in the remaining population. Indeed, using a flip of a coin it is possible to create a diagnostic test with operating characteristics anywhere on the straight line extending to (1, 1) (Wagner *et al.*, 2010) (Bandos *et al.*, 2010). Theoretical importance of this type of a ROC curve for the current work stems from the ancillary nature of the operating points with non-zero FPF. In practice the pure straight-line ROC curves (i.e., with empirical points aligned around the straight line) could occur when a diagnostic system is forced to produce continuous (untied) results in situations when there is little or no information for distinguishing between subjects (Gur *et al.*, 2006).

Straight-line ROC curve has a constant value of standardized partial AUC regardless of the range of interest (Ma *et al*., 2013); this offers an important scenario for investigating pAUC-based inferences.

The name "straight-line" simply reflects the shape of the curve. The ROC curve with a straight-line shape would result from the two (independent) random variables with uniform distributions. However, due to the ROC invariance property the use of the straight-line ROC curve does not necessarily imply that the test results are assumed to follow uniform distributions

in the subpopulation of normal and abnormal patients. In general it can be viewed as a curve corresponding to a diagnostic result that perfectly separates the most obvious "abnormal" patients, while is non-informative for discriminating between normal and abnormal patients in the remaining population.

For a pair of latent scores for normal and abnormal patients which follow two uniform distributions, i.e. $X \sim U(0,1)$ and $Y \sim U(0,a)$ respectively, the ROC curve can be expressed as:

$$ROC(e) = \frac{1}{a}e + 1 - \frac{1}{a}.$$

and the AUC can be expressed as

$$A = \int_0^1 \left(\frac{1}{a}e + 1 - \frac{1}{a} \right) dx = 1 - \frac{1}{2a},$$

while the pAUC can be expressed as:

$$A_e = \int_0^e \left(\frac{1}{a}e + 1 - \frac{1}{a} \right) dx = \left(1 - \frac{1}{a} \right) e + \frac{1}{2a}e^2.$$

2.2 ESTIMATION OF ROC CURVES

2.2.1 PARAMETRIC ESTIMATES OF ROC CURVES

A number of approaches exist for parametric estimation of the ROC curve. The two general classes of parametric estimation methods are "distribution-free" and "distribution-based" approaches.

Distribution-free approaches may place parametric assumption on the shape of the ROC curve, e.g., binormal ROC curve, $ROC(e) = \Phi\left(a + b\Phi^{-1}(e)\right)$, but not on the distributions of scores for diseased and non-diseased subjects. For continuous test results, Pepe (2003) proposed an estimation method involving the methods of generalized estimating equations and generalized linear models which can incorporate covariate information. Zou and Hall (2000) performed MLE rank-based estimation of binormal ROC curves. For discrete test results, a maximum likelihood approach was introduced by Dorman and Alf (1969).

Distribution-based approaches, on the other hand, estimate conditional distribution of the test results (given subjects' true status); the ROC curve is then estimated indirectly as the composition quantile and distribution function. For example, a naïve distribution-based approach for estimating the binormal ROC curve (which is rarely used in practice), assumes a normal distribution of the test results. If X and Y are the test results for the random samples of m normal and n abnormal subjects, based on the invariance principle, the maximum likelihood estimate (MLE) of the binormal ROC curve can be expressed as follows (Zhou *et al.*, 2002):

$$\hat{ROC}(e) = \Phi\left(\hat{a} + \hat{b}\Phi^{-1}(x)\right)$$

where $\hat{a} = \dfrac{\left(\hat{\mu}_y - \hat{\mu}_x\right)}{\hat{\sigma}_y}$, $\hat{b} = \dfrac{\hat{\sigma}_x}{\hat{\sigma}_y}$, $\hat{\mu}_x$, $\hat{\mu}_y$, $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the ML estimates of the means and standard deviations, and Φ is the cumulative normal distribution function. Given that the binormal distribution assumption is restrictive and based on the invariance property of monotonic transformation of ROC curves, Faraggi *et al.* (2003) applied a Box-Cox type power transformation to the data, and after obtaining the appropriate transformation used binormal model.

2.2.2 EMPIRICAL ROC CURVE

The empirical ROC curve is a collection of the empirical operating points ($F\hat{P}F$ and $T\hat{P}F$) where The empirical true and false positive fractions are computed as follows:

$$T\hat{P}F(c) = \frac{\sum_{j=1}^{n} I\left[Y_j > c\right]}{n},$$

$$F\hat{P}F(c) = \frac{\sum_{i=1}^{m} I\left[X_i > c\right]}{m}.$$

where $I(x) = 1$ if x is true and 0 otherwise. However, frequently the empirical ROC curve is plotted by connecting the empirical points with straight line segments. Some analytical methods however, do not use the points on the straight-lines (Greenhouse and Mantel, 1950) (Wieand *et al.*, 1989) (Zhang *et al.*, 2002) (He and Escobar, 2008). The points on the straight-line segments between the empirical points describe operating characteristics which might not be attainable by applying specific thresholds to the observable test results. However, these could be attained by random guessing between the decisions at the adjacent operating points (Fawcett, 2006) (Wagner *et al.*, 2010) (Bandos *et al.*, 2010).

We will use the term "linearly-interpolated" empirical ROC curve to distinguish it from the set of discrete (fpf, tpf) points.

2.3 ESTIMATION OF AUC AND PAUC

2.3.1 PARAMETRIC ESTIMATES OF AUC AND PAUC

Parametric analyses based on AUC and partial AUC are reasonably straightforward. The previously mentioned methods can be used to estimate smooth ROC curves. Once a smooth curve is fitted, the partial area can be estimated for any specified range of interest; its variance can be evaluated using the "delta" method based on the variance of the model parameters (Zhou *et al.*, 2003). For naïve binormal model, the estimated AUC or partial AUC can be computed as follows:

$$\hat{A} = \int_0^1 \Phi\left(\hat{a} + \hat{b}\Phi^{-1}(x)\right)dx = \Phi\left(\frac{\hat{a}}{\sqrt{1+\hat{b}^2}}\right)$$

$$\hat{A}_e = \int_0^e \Phi\left(\hat{a} + \hat{b}\Phi^{-1}(x)\right)dx$$

where $\hat{a} = \frac{\left(\hat{\mu}_y - \hat{\mu}_x\right)}{\hat{\sigma}_y}$, $\hat{b} = \frac{\hat{\sigma}_x}{\hat{\sigma}_y}$, $\hat{\mu}_x$, $\hat{\mu}_y$, $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the MLE of the mean and standard deviations of the latent scores (e.g., MLE estimates for a and b can be obtained from the probit regression model of the discrete test results), and Φ is the cumulative normal distribution function. Or by using analytic expression for pAUC in the case of binormal ROC curves (Hillis and Metz, 2012),

$$\hat{A}_e = F_{BVN}\left(\frac{\hat{a}}{\sqrt{1+\hat{b}^2}}, \Phi^{-1}(e); -\frac{\hat{b}}{\sqrt{1+\hat{b}^2}}\right)$$

where $F_{BVN}(z, x; \rho)$ is the standardized bivariate normal distribution function with correlation ρ.

2.3.2 EMPIRICAL ESTIMATES OF AUC AND PAUC

If $\{X_i\}_{i=1}^{m}$ and $\{Y_j\}_{j=1}^{n}$ are the test results for random samples of *m* normal and *n* abnormal subjects then the estimate of the AUC can be expressed as follows:

$$\hat{A} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\psi(X_i, Y_j)}{nm} \text{ where } \psi(X,Y) = \begin{cases} 1 & X < Y \\ ½ & X = Y \\ 0 & X > Y \end{cases}$$

This non-parametric AUC estimator is equal to the area under the empirical ROC curve where the points on the plot are connected by straight lines.

The partial area can be estimated by (He and Escobar, 2008):

$$\hat{A}_e = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi(X_i, Y_j)$$

where

$$\phi(X_i, Y_j) = \begin{cases} 1, & Y_j > X_i \quad and \quad X_i \in [r_0, \infty) \\ \frac{1}{2}, & Y_j = X_i \quad and \quad X_i \in [r_0, \infty) \\ 0, & Y_j < X_i \quad and \quad X_i \in [r_0, \infty) \end{cases}$$

$$r_0 = \tilde{F}_x^{-1}(1-e)$$

and $\tilde{F}_x$ is the empirical distribution of X_i.

For any consecutive ratings r_1 and r_2 where $r_2 < r_1$, $e_1 = FPF(r_1)$ and $e_2 = FPF(r_2)$, if $e \in (e_1, e_2)$, one can a use linear interpolation to compute the pAUC which can be expressed as follows:

$$\hat{A}_e = \hat{A}_{e_1} + \left\{ TPF(r_1) + \frac{(e-e_1)(TPF(r_2) - TPF(r_1))}{2(e_2 - e_1)} \right\}(e - e_1).$$

2.4 ESTIMATION OF VARIANCE OF AUC AND PAUC

In general variance of the parametric AUC and pAUC estimators can be obtained from a variance matrix of the estimated parameters (corresponding to the ROC fitting approach) using delta method (Zhou *et al.*, 2002). For the naïve fitting of the binormal ROC curve (assuming normally distributed test result) the variance estimator attains the following closed-form expression in terms of *a* and *b* parameters of the binormal model (Obuchowski and McClish, 1997):

$$\hat{V}\left(\hat{A}_e\right) = f^2 V\left(\hat{a}\right) + g^2 V\left(\hat{b}\right) + 2\,fgC\left(\hat{a},\hat{b}\right) \tag{2.1}$$

where:

$$\hat{V}\left(\hat{a}\right) = \frac{m\left(a^2+2\right)+2nb^2}{2mn}$$

$$\hat{V}\left(\hat{b}\right) = \frac{\left(n+m\right)b^2}{2mn}$$

$$f = \frac{\exp\left\{-\dfrac{a^2}{2\left(1+b^2\right)}\right\}}{\sqrt{2\pi\left(1+b^2\right)}}\left\{\Phi\left(h\right)\right\}$$

$$\hat{C}\left(\hat{a},\hat{b}\right) = \frac{ab}{2n}$$

and

$$h = \left\{\Phi^{-1}\left(e\right) + \frac{ab}{1+b^2}\right\}\sqrt{1+b^2}$$

When $e=1$, this formula will reduce to the variance estimator for full AUC.

Due to the close relationship to the Mann-Whitney test statistics (Bamber, 1975), variance of the AUC estimator for an empirical ROC curve can be derived from the formula for the Wilcoxon statistics proposed by Noether (1967):

$$Var(\hat{A}) = \frac{m-1}{mn}\xi_{10} + \frac{n-1}{mn}\xi_{01} + \frac{1}{mn}\xi_{11} \quad \forall i,k = 1,...,m \;\; j,l = 1,...,n$$

where

$$\xi_{10} = Cov\{\psi(X_i, Y_j), \psi(X_i, Y_l)\} = E\{\psi(X_i, Y_j)\psi(X_i, Y_l)\} - A^2, \quad j \neq l$$

$$\xi_{01} = Cov\{\psi(X_i, Y_j), \psi(X_k, Y_j)\} = E\{\psi(X_i, Y_j), \psi(X_k, Y_j)\} - A^2, \quad i \neq k$$

$$\xi_{11} = Var\{\psi(X_i, Y_j)\} = E\{\psi(X_i, Y_j)^2\} - A^2$$

$$A = E\{\psi(X_i, Y_j)\} = E\{\hat{A}\}$$

For continuous test results which are often encountered in many scenarios such as genetic research, He and Escobar (2008) proposed a non-parametric variance estimator for the partial area. Alternatively, the variance of empirical estimators of AUC and pAUC can also be estimated using a nonparametric bootstrap approaches (Efron and Tibshirani, 1993). The variance can be estimated by resampling the normal and abnormal subjects and linearly interpolating the empirical ROC curves. Another variance estimator of the pAUC using a non-parametric approach was proposed by Liu *et al.* (2005). If $\{X_i\}_{i=1}^{m}$ and $\{Y_j\}_{j=1}^{n}$ are the test results for random samples of *m* normal and *n* abnormal subjects with distribution functions F_x and F_y, and empirical distribution functions $\tilde{F}_x$ and $\tilde{F}_y$ respectively, then the asymptotically unbiased estimator $\hat{A}_e$ of pAUC can be expressed as:

$$\hat{A}_e = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i\in P} I\left(Y_j > X_i\right) = \frac{1}{m}\sum_{i\in P}\tilde{S}_y\left(X_i\right)$$

where $\tilde{S}_y(z) = 1 - \tilde{F}_y(z)$, R_i is the rank of X_i among the X's, that is, $R_i = \sum_{k=1}^{m} I(X_k \le X_i)$, and $\mathrm{P} = \{i : m(1-e) \le R_i \le m\}$. They also showed that:

$$\hat{A}_e \xrightarrow{d} N\left(A_e, \frac{\sigma_e^2}{m+n}\right)$$

Where:

$$\sigma_e^2 = \sigma_H^2 / \lambda' + \sigma_W^2 / \lambda$$

$$\sigma_W^2 = \int_{1-e}^{1}\int_{1-e}^{1} S_y\left(F_x^{-1}(s \vee t)\right) ds dt - A_e^2,$$

$$s \vee t = \max(s,t),$$

$$\sigma_H^2 = \int_{1-e}^{1} S_y^2\left(F_x^{-1}(p)\right) dp - A_e^2,$$

$$\lambda' = 1 - \lambda = m/(m+n),$$

$$S_x(z) = 1 - F_x(z) \text{ and } S_y(z) = 1 - F_y(z).$$

Moreover, the consistent estimators of σ_H^2 and σ_W^2, respectively can be obtained:

$$\hat{\sigma}_H^2 = \frac{1}{m}\sum_{i \in \mathrm{P}} \left\{\tilde{S}_y(X_i)\right\}^2 - \hat{A}_e^2, \quad \hat{\sigma}_W^2 = \frac{1}{m(m-1)} \sum_{i \ne k \in \mathrm{P}} \tilde{S}_y(X_i \vee X_k) - \hat{A}_e^2.$$

3.0 EVALUATION OF A SINGLE PAUC

The statistical inference regarding diagnostic accuracy of a single modality (diagnostic system, classification tool, etc.) is often made on the basis of summary indices such as pAUC and AUC. For example, diagnostic accuracy for classifying images as depicting or not-depicting lung nodules can be assessed using both point estimation and interval estimation of pAUC and AUC. In the evaluation of a single pAUC, we investigated the effect of the size of the range of interest (0, e) on statistical inferences regarding the pAUC. We analyzed the properties of the nonparametric and parametric estimates of spAUCs and their variances. We derived two important properties of the relationship between the spAUC and a defined range of interest, which could facilitate a wider and more appropriate use of this important summary index. First, we mathematically proved that the spAUC increases with increasing range of interest for common ROC curves. Second, using a comprehensive numerical investigation we demonstrated that, contrary to common belief, the uncertainty about the estimated spAUC can either decrease or increase with an increasing range of interest.

Our results indicated that the pAUC could offer advantages in some scenarios in terms of statistical uncertainty of the estimation. In addition, selection of a wider range would likely lead to an increased estimate even in the case of spAUC. We demonstrated that the bi-gamma family of the concave ROC curves adequately describes a wide range of scenarios including cases where pAUC is statistically advantageous. This family was used to develop sample size

estimation software offering a better insight in relative merits of analyzing part of the curve. This portion of the research is published in Statistics in Medicine (Appendix A).

3.1 METHOD

3.1.1 STANDARDIZED PARTIAL AUC AND ITS PROPERTIES

Based on the definition of standardized pAUC (1.1), it can be shown that the standardized pAUC and the variance of its estimate are always larger than conventional pAUC and the variance of its estimate. Indeed since 1/e/(2e-1), is less than 1 for all e≤1:

$$\tilde{A}_e \geq \frac{1}{2}\left\{1+2\left(A_e-\frac{e^2}{2}\right)\right\} = A_e+\frac{1}{2}-\frac{e^2}{2} \geq A_e$$

and

$$V\left(\hat{\tilde{A}}_e\right) = V\left(\hat{A}_e\right) \Big/ 4\left(e-\frac{e^2}{2}\right)^2 \geq V\left(\hat{A}_e\right).$$

Unfortunately, “standardization” of the partial area in (1.1) is not ideal. Indeed, although the range of $\tilde{A}_e$ is independent of *e*, the actual value of $\tilde{A}_e$ for a given ROC curve could depend on *e*. Moreover, as we demonstrate in Proposition 3.1 below, theoretically it can either increase or decrease with increasing range while remaining constant only in the case of a “straight-line ROC curve” (Chapter 2.1.4) composed of two straight-line segments – one vertical and the other passing through the point (1,1). Using equation (2.1) it is easy to see that partial AUC for the straight-line ROC curve passing through the point (f,t) is:

$$A_{e,straight,(f,t)} = e^2(1-t)/2(1-f) + e\{1-(1-t)/(1-f)\}$$

and the standardized partial AUC does not depend on the range of interest (independent of e):

$$\tilde{A}_{straight,(f,t)} = 1-(1-t)/2(1-f) \tag{3.1}$$

Proposition 3.1

For *any* $e \in (0,1)$,

i. $\frac{\partial \tilde{A}_e}{\partial e} > 0 \Leftrightarrow ROC(e) > 2(1-\tilde{A}_e)e + (2\tilde{A}_e - 1)$

ii. $\frac{\partial \tilde{A}_e}{\partial e} = 0 \Leftrightarrow ROC(e) = 2(1-\tilde{A}_e)e + (2\tilde{A}_e - 1)$

iii. $\frac{\partial \tilde{A}_e}{\partial e} < 0 \Leftrightarrow ROC(e) < 2(1-\tilde{A}_e)e + (2\tilde{A}_e - 1)$

Proof:

By straightforward differentiation of (1.1) we obtain:

$$\frac{\partial \tilde{A}_e}{\partial e} = \frac{1}{2}\left(e - \frac{e^2}{2}\right)^{-2}\left\{\left(ROC(e) - e\right)\left(e - \frac{e^2}{2}\right) - \left(A_e - \frac{e^2}{2}\right)(1-e)\right\}$$

Since $A_e - \frac{e^2}{2} = (2\tilde{A}_e - 1)\left(e - \frac{e^2}{2}\right)$, the derivative of standardized partial AUC can be written as follows:

$$\frac{\partial \tilde{A}_e}{\partial e} = \frac{1}{2}\left(e - \frac{e^2}{2}\right)^{-1}\left\{\left(ROC(e) - e\right) - (2\tilde{A}_e - 1)(1-e)\right\}$$

The three claims of this proposition immediately follow.

Proposition 3.1 implies that given the area over the range (0,*e*), we can determine whether a small increase in the range would lead to an increase in the standardized pAUC by comparing whether the point on the ROC curve ROC(*e*) is actually above or below the fixed

straight line, that passes through the point (1,1) and has a slope of $2\left(1-\tilde{A}_e\right)$. Alternatively, this comparison can be conducted by comparing the negative diagnostic likelihood ratio (1-ROC(*e*))/(1-e) with $2\left(1-\tilde{A}_e\right)$.

The negative diagnostic likelihood ratio, DLR-(e), is an important characteristic of binary diagnostic test (Zhou *et al.*, 2002) (Norman, 1964) (Biggerstaff, 2000) (Bandos *et al.*, 2010). For a point on the ROC curve (e,ROC(e)) it is defined as (1-ROC(*e*))/(1-e). The ROC curve with a decreasing negative diagnostic likelihood ratio is practically important. Such an ROC curve ensures that starting at any given operating point, a threshold-driven improvement in sensitivity will be better than an improvement achieved by randomly selected subjects that were tested "negative" at the given operating point (Norman, 1964) (Bandos *et al.*, 2010). Thus, a decreasing negative diagnostic likelihood ratio in the region where experimental operating points are observed is a natural property for many practical diagnostic tests.

While results of proposition 3.1 are important for judging the dependence of spAUC on small changes in the range of interest, they provide little insight into the more global behavior of the spAUC, or the general form of curves with always increasing/decreasing $\tilde{A}_e$. These questions are addressed by the following proposition and its corollaries.

Proposition 3.2

If the ROC curve has a decreasing negative diagnostic likelihood ratio in (0, e_0), namely, $\frac{\partial}{\partial e}\left\{\frac{1-ROC(e)}{1-e}\right\}<0$, then $\frac{\partial \tilde{A}_e}{\partial e}>0$ in the same range.

Proof:

Let us consider e from $(0, e_0)$. Since for any e`$\in$(0,e) $\frac{\partial}{\partial f}\left\{\frac{1-ROC(f)}{1-f}\right\}\Big|_{f=e`} < 0$, we can obtain the following inequality :

$$\frac{1-ROC(e)}{1-e} < \frac{1-ROC(e`)}{1-e`} \quad or \quad ROC(e`) < 1-(1-e`)\times\frac{ROC(e)-1}{e-1}$$

Hence over the range (0, e], the partial area (A_e) and the standardized partial area under the ROC curve $(\tilde{A}_e)$ are smaller than the corresponding areas under the straight line ROC curve passing though (e, ROC (e)). Indeed:

$$A_e = \int_0^e ROC(f)\,df < \int_0^e \left\{1+(f-1)\times\frac{ROC(e)-1}{e-1}\right\} df = A_{e,straight,(e,ROC(e))} \Rightarrow \tilde{A}_e < \tilde{A}_{straight,(e,ROC(e))}.$$

On the other hand, from (2) we have:

$$ROC(e) = 2\left(1-\tilde{A}_{straight,(e,ROC(e))}\right)e + \left(2\tilde{A}_{straight,(e,ROC(e))}-1\right) = 2(1-e)\,\tilde{A}_{straight,(e,ROC(e))}+1.$$

Also, since $\tilde{A}_e < \tilde{A}_{straight,(e,ROC(e))}$, from above we obtain:

$$ROC(e) > 2(1-e)\,\tilde{A}_e + 1 = 2\left(1-\tilde{A}_e\right)e + \left(2\tilde{A}_e - 1\right)$$

Finally, applying the result (i) of proposition 3.1 we obtain $\frac{\partial \tilde{A}_e}{\partial e} > 0$.

As we discussed previously in this section, a decreasing negative diagnostic likelihood ratio is a natural property for many practical diagnostic tests. We also note that the result of proposition 3.2 is directly applicable to concave ROC curves, as it can be demonstrated that concavity immediately implies a decreasing diagnostic likelihood ratios. Figure 3.1 illustrates the increase of the standardized partial AUC with increasing range for five concave binormal ROC curves.

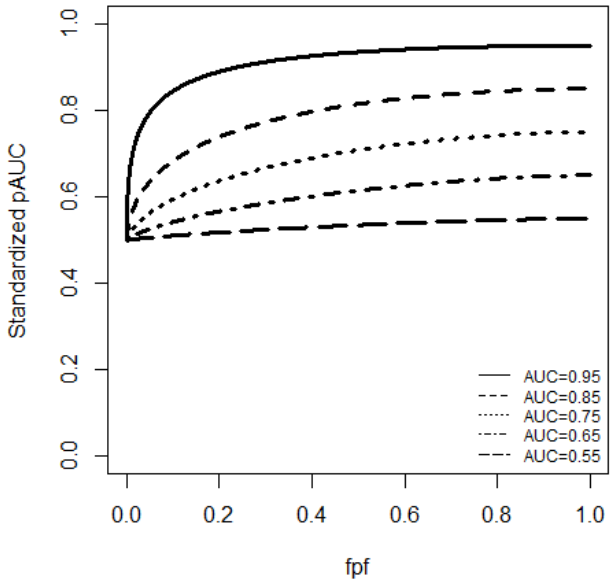


Figure 3.1 Values of the standardized partial AUC for concave binormal ROC curves.

We note that proposition 3.2 is directly extendable to the partial area index (McClish, 1989) (Jiang *et al.*, 1996) as well as to the "non-standardized" partial area. Results summarized in this section indicate that in practical scenarios, current approaches to standardization of the partial AUC do not necessarily eliminate the effect of the range of interest on values of the standardized pAUC. Moreover, increasing range of can frequently increase the apparent level of diagnostic performance. In the next two sections we examine the statistical uncertainty in the estimated standardized partial AUC.

3.1.2 VARIANCE OF THE PARAMETRIC ESTIMATE OF SPAUC

The partial AUC and other ROC related characteristics are typically estimated from a sample of m normal and n abnormal subjects with observed diagnostic test results of $\{x_i\}_{i=1}^{m}$ and $\{y_j\}_{j=1}^{n}$

correspondingly. We focus here on the relationship between the variance of the spAUC and the size of the range of interest. In particular, we examine the common conjecture that variance would decrease with increasing range, since a larger range incorporates more available information in regards to the operating characteristics.

We begin by considering a simple variance estimate for the partial area under the binormal ROC curve (McClish, 1989). Then, in section 5 we present simulation results that demonstrate the generality of the derived conclusions.

We can compute the variance of the estimated standardized partial AUC as:

$$V\left(\hat{\tilde{A}}_e\right) = \frac{V\left(\hat{A}_e\right)}{4\left(e - \frac{e^2}{2}\right)^2}, \text{ where } V\left(\hat{A}_e\right) \text{ is computed according to (2.1).}$$

Figure 3.2 demonstrates the variance of the estimated standardized pAUC as a function of the length of the range *e*, for two different binormal as well as straight-line ROC scenarios. These scenarios are based on a sample size of 100, (m=n=50) and describe different shapes of ROC curves, including concave curves (*b=1*) and typical improper curves (*b=0.5*) (Obuchowski and McClish, 1997). Each figure shows variance functions for five ROC curves with AUCs of 0.55, 0.65, 0.75, 0.85, and 0.95. We note that here, as well as in the investigations that follow, we consider binormal ROC curves with b≤1 since the corresponding shapes of these ROC curves are more common in practical applications including, but not limited to, medical imaging. Indeed, a binormal ROC curve with b>1 implies a worse-than-chance performance in evaluations of highly suspicious subjects (i.e., in the range of high specificity) – which rarely happens in practice.

For concave ROC curves (Figure 3.2a) the variance of the full AUC can exhibit both patterns, namely, it can be either smaller or larger than variance of the standardized partial AUCs on (0, *e*). The decrease in variance with increasing range is observed only for ROC curves with

AUC values greater than 0.75. In the straight-line ROC scenarios for which all standardized partial AUCs are exactly the same as the full AUC, the variance of the standardized partial AUC increases. As shown in Figure 3.2b, for an improper binormal ROC curve, the variance frequently increases with increasing range, in particular the variance of the full AUC ($e=1$) tends to be larger than the variances for standardized partial AUCs over most ranges considered. The anticipated decrease in the variance when switching to full AUC is evident only for the ROC curve with the largest AUC (0.95) considered herein.

a) b=1 b) b=0.5

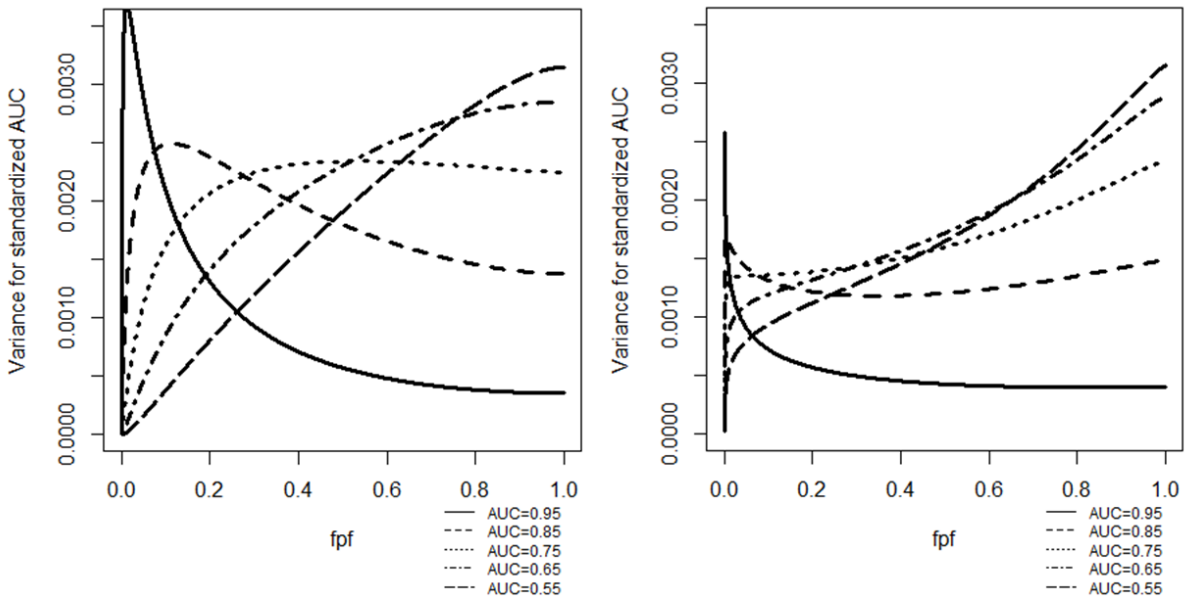


Figure 3.2 Variance of standardized pAUC(0,e) estimates for binormal ROC curves as a function of the size of the range of interest *e*.

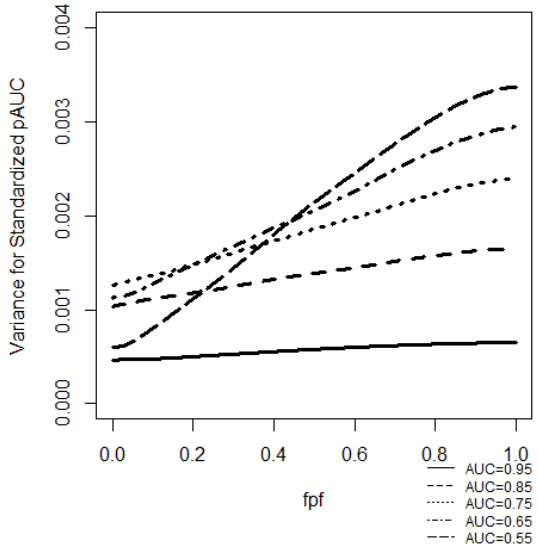


Figure 3.3 Variance of standardized pAUC estimates for straight-line ROC curves over (0,*e*) as a function of the size of the range of interest *e*.

These results provide an important indication that there are a number of practical scenarios in which the estimated partial AUC may be no less precise than the estimated variance for the full AUC. Variance is an important characteristic of the statistical uncertainty, however, its usefulness for non-symmetric distributions is limited (e.g., sampling distribution of estimates of high pAUC). Furthermore, the trends shown in Figure 3.2 are based on the assumption of normality of the test result, and hence might not be generalizable. In order to verify these trends we conducted a simulation study as described in the following section.

3.2 NUMERICAL STUDY

In this section we considered several families of ROC curves including binormal, bi-gamma and straight-line ROC curves. For each scenario, we computed the standardized pAUC by numerical integration. We conducted a simulation study to assess the length of the equal-tail 95% range (97.5^{th} -2.5^{th} percentile) and variance of the sampling distribution of the standardized pAUC. The statistical power was estimated from 1000 results of the bootstrap tests and the sample size was computed by established results for Wald-type tests (Flahault, 2005). In the simulation study the test results for normal and abnormal subjects were generated from normal distributions with parameters selected to generate binormal ROC curves with specific values of AUC (ranging from 0.55 to 0.95) and for three values for the shape parameter b (1, 0.5 and 0.33). Values for the parameters of binormal ROC curves were selected to reflect shapes typically encountered in experimental performance assessment studies in diagnostic medicine. For the bi-gamma ROC curves the test results were generated from gamma distributions with the same shape parameter (Chapter 2.1.3). For the straight-line ROC curves the test results were generated from the

uniform distributions of different width. For each scenario we generated 10,000 datasets of with m=50 and n=50 subjects.

For each generated dataset we estimated the empirical ROC curve and, using numerical integration, computed the standardized partial AUC over different ranges starting from 0 and ending at 0.2, 0.4, 0.6, 0.8, and 1. The difference between the 9750^{th} (largest) and 250^{th} (smallest) estimate of the AUC for a given scenario was used to estimate the length of the equal-tail 95% range of the sampling distribution. We note that transformations (e.g., logit) are often used to improve on Wald-type confidence intervals. In the simulation study, however, we have the ability to assess the width of distribution more precisely by using percentiles of the simulated distribution.

Scenario 1:

We first investigated the properties of standardized pAUC for binormal ROC curves. Table 3.1 showed that the standardized pAUC increased with increasing range for concave binormal ROC curves and improper ROC curves with high AUC, i.e. b=0.5 AUC≥0.85, and for b=0.33, AUC=0.95.

Table 3.1 Theoretical spAUC for binormal ROC curves with different *b*'s and full AUCs

Parameters of the ROC curves	Ranges of False Positive Fractions				
	0-0.2	**0-0.4**	**0-0.6**	**0-0.8**	**0-1**
b=0.33					
auc=0.55	0.651	0.642	0.621	0.589	0.550
auc=0.65	0.710	0.709	0.697	0.676	0.650
auc=0.75	0.776	0.781	0.777	0.765	0.750
auc=0.85	0.852	0.860	0.861	0.857	0.850
auc=0.95	0.942	0.948	0.951	0.951	0.950
b=0.50					
auc=0.55	0.607	0.607	0.595	0.574	0.550
auc=0.65	0.665	0.676	0.674	0.664	0.650
auc=0.75	0.734	0.753	0.758	0.756	0.750
auc=0.85	0.818	0.840	0.848	0.851	0.850
auc=0.95	0.926	0.940	0.946	0.949	0.950
b=1.0					
auc=0.55	0.518	0.530	0.539	0.547	0.550
auc=0.65	0.567	0.602	0.626	0.643	0.650
auc=0.75	0.637	0.690	0.723	0.743	0.750
auc=0.85	0.739	0.797	0.828	0.845	0.850
auc=0.95	0.890	0.926	0.941	0.948	0.950

The results for the empirical estimator of the standardized pAUC are summarized in Tables 3.2 and 3.3. These results closely agree with results from the previous section (Figure 4.1). In particular, the variances and lengths of the equal-tail 95% ranges of the sampling distributions of the estimated standardized pAUCs increase with increasing ranges for the ROC curves with lower AUCs (e.g., AUC for concave ROC curves is less than 0.75). With increasing "improperness" of the ROC curves (i.e., decreasing b) decreasing trends, even for ROC curve with large AUCs, are gradually diminishing. For example, for a binormal ROC curve with b=0.33, the variance and length of the equal-tail 95% interval of sampling distribution of standardized pAUC increases with increasing range of interest for all considered ROC curves.

Table 3.2 Variance of sampling distributions of standardized pAUC for binormal ROC curves ($\times 10^{-3}$)

Parameters of the ROC curves	**Ranges of False Positive Fractions**				
	0-0.2	**0-0.4**	**0-0.6**	**0-0.8**	**0-1**
b=0.33					
auc=0.55	1.487	1.822	2.300	2.974	3.880
auc=0.65	1.599	1.845	2.216	2.767	3.486
auc=0.75	1.494	1.636	1.909	2.309	2.812
auc=0.85	1.212	1.237	1.370	1.578	1.846
auc=0.95	0.539	0.496	0.509	0.552	0.613
b=0.50					
auc=0.55	1.467	1.827	2.297	2.915	3.615
auc=0.65	1.723	1.966	2.307	2.771	3.268
auc=0.75	1.776	1.833	2.012	2.276	2.563
auc=0.85	1.539	1.427	1.449	1.534	1.645
auc=0.95	0.718	0.563	0.512	0.500	0.507
b=1.0					
auc=0.55	1.058	1.852	2.561	3.125	3.407
auc=0.65	1.763	2.388	2.753	2.954	3.010
auc=0.75	2.447	2.510	2.449	2.349	2.289
auc=0.85	2.811	2.162	1.775	1.548	1.462
auc=0.95	1.537	0.784	0.528	0.423	0.392

*data consisted of ratings for 50 normal and 50 abnormal subjects; 10,000 datasets were simulated.

Table 3.3 Differences of 2.5% and 97.5% estimated percentiles of sampling distributions of standardized pAUC for binormal ROC curves

Parameters of the ROC curves	Ranges of False Positive Fractions				
	0-0.2	**0-0.4**	**0-0.6**	**0-0.8**	**0-1**
b=0.33					
auc=0.55	0.1500	0.1681	0.1890	0.2150	0.2440
auc=0.65	0.1567	0.1669	0.1838	0.2050	0.2312
auc=0.75	0.1511	0.1575	0.1714	0.1896	0.2088
auc=0.85	0.1356	0.1369	0.1448	0.1554	0.1676
auc=0.95	0.0900	0.0862	0.0890	0.0925	0.0972
b=0.50					
auc=0.55	0.1489	0.1669	0.1867	0.2117	0.2340
auc=0.65	0.1622	0.1744	0.1886	0.2067	0.2244
auc=0.75	0.1644	0.1669	0.1748	0.1858	0.1976
auc=0.85	0.1533	0.1481	0.1495	0.1542	0.1592
auc=0.95	0.1033	0.0919	0.0867	0.0854	0.0864
b=1.0					
auc=0.55	0.1256	0.1681	0.1986	0.2200	0.2272
auc=0.65	0.1644	0.1919	0.2057	0.2138	0.2156
auc=0.75	0.1922	0.1956	0.1943	0.1892	0.1864
auc=0.85	0.2067	0.1813	0.1652	0.1538	0.1492
auc=0.95	0.1522	0.1094	0.0886	0.0792	0.0764

*data consisted of ratings for 50 normal and 50 abnormal subjects; 10,000 datasets were simulated.

Table 3.4 shows that the statistical power for evaluation of a single pAUC for binormal ROC curves increases with increasing range for concave binormal ROC curves with AUC less than 0.75 and decreases with increasing range for improper ROC curves with AUC less than 0.65.

Table 3.4 Statistical power for testing spAUC=0.5 for binormal ROC curves

Parameters of the ROC curves	**Ranges of False Positive Fractions**				
	0-0.2	**0-0.4**	**0-0.6**	**0-0.8**	**0-1**
b=0.33					
auc=0.55	1.000	0.980	0.832	0.455	0.157
auc=0.65	1.000	0.999	0.998	0.957	0.771
auc=0.75	1.000	1.000	1.000	1.000	0.999
auc=0.85	1.000	1.000	1.000	1.000	1.000
auc=0.95	1.000	1.000	1.000	1.000	1.000
b=0.50					
auc=0.55	0.961	0.850	0.615	0.341	0.157
auc=0.65	0.999	0.997	0.978	0.917	0.771
auc=0.75	1.000	1.000	1.000	1.000	1.000
auc=0.85	1.000	1.000	1.000	1.000	1.000
auc=0.95	1.000	1.000	1.000	1.000	1.000
b=1.0					
auc=0.55	0.133	0.145	0.155	0.157	0.152
auc=0.65	0.559	0.676	0.736	0.754	0.760
auc=0.75	0.950	0.990	0.998	0.998	0.998
auc=0.85	1.000	1.000	1.000	1.000	1.000
auc=0.95	1.000	1.000	1.000	1.000	1.000

*data consisted of ratings for 50 normal and 50 abnormal subjects; 1000 datasets were simulated and 1000 results of the bootstrap tests for testing the null hypothesis spAUC=0.5 were performed.

Table 3.5 shows the sample size requirements for a two-sided 95% confidence interval for a standardized pAUC to be narrower than 0.1 with probability corresponding to 1-β=0.8. Sample size was estimated using the method proposed by Flahault *et al.* (2005). There exists an increasing trend in sample size for pAUC-based statistical inference for concave binormal ROC curves with AUC ≤ 0.65 and improper ROC curves with AUC ≤ 0.85.

Table 3.5 Sample size requirements for two-sided 95% confidence interval for a standardized pAUC to be narrower than 0.1 when the ROC curve has a binormal shape

Parameters of the ROC curves	**Ranges of False Positive Fractions**				
	0-0.2	**0-0.4**	**0-0.6**	**0-0.8**	**0-1**
b=0.33					
auc=0.55	234	287	362	467	610
auc=0.65	252	290	348	435	548
auc=0.75	235	257	300	363	442
auc=0.85	191	195	215	248	290
auc=0.95	85	78	80	87	97
b=0.50					
auc=0.55	231	287	361	458	568
auc=0.65	271	309	363	435	514
auc=0.75	279	288	316	358	403
auc=0.85	242	224	228	241	259
auc=0.95	113	89	81	79	80
b=1.0					
auc=0.55	167	291	403	491	535
auc=0.65	277	375	433	464	473
auc=0.75	385	394	385	369	360
auc=0.85	442	340	279	243	230
auc=0.95	242	124	83	67	62

*data consisted of ratings for 50 normal and 50 abnormal subjects; 10,000 datasets were simulated for evaluating the variance of empirical spAUCs.

Table 3.6 shows that the estimated sample size for testing the null hypothesis of spAUC=0.5 (with power of 80% to detect the simulated difference) frequently decreases with increasing range. The increasing trend can be observed only for improper binormal ROC curves with AUC less than 0.75. The discrepancy between trends in Tables 3.5 and 3.6 stems from the tendency of the spAUC to change with increasing size of the range.

Table 3.6 Sample size requirements for testing spAUC=0.5 when the ROC curve has a binormal shape

Parameters of the ROC curves	**Ranges of False Positive Fractions**				
	0-0.2	**0-0.4**	**0-0.6**	**0-0.8**	**0-1**
b=0.33					
auc=0.55	26	35	62	147	609
auc=0.65	14	17	22	35	61
auc=0.75	8	8	10	13	18
auc=0.85	4	4	4	5	6
auc=0.95	1	1	1	1	1
b=0.50					
auc=0.55	50	63	100	209	567
auc=0.65	25	25	30	40	57
auc=0.75	13	11	12	14	16
auc=0.85	6	5	5	5	5
auc=0.95	2	1	1	1	1
b=1.0					
auc=0.55	1282	808	661	555	535
auc=0.65	154	90	68	57	53
auc=0.75	51	27	19	16	14
auc=0.85	19	10	6	5	5
auc=0.95	4	2	1	1	1

*data consisted of ratings for 50 normal and 50 abnormal subjects; 10,000 datasets were simulated for evaluating the variance of empirical spAUCs.

Scenario 2:

In this scenario, we investigated the properties of standardized pAUC for straight-line ROC curves. As discussed previously, straight-line ROC curves guarantee constancy of the standardized pAUC regardless of range of interest. Thus the trend for statistical properties is driven purely by sampling variability.

In contrast to the binormal scenario, Tables 3.7, 3.8, 3.9, 3.10 and 3.11, show that with increasing range for straight-line ROC curves, the variance and width of the sampling distribution always increases, statistical power decreases, and therefore, sample size requirement increases.

Table 3.7 Variance of sampling distributions of standardized pAUC for straight-line ROC curves ($\times 10^{-3}$)

Parameters of the ROC curves	Ranges of False Positive Fractions				
	0-0.2	**0-0.4**	**0-0.6**	**0-0.8**	**0-1**
auc=0.55	1.085	1.708	2.339	2.937	3.269
auc=0.65	1.478	1.923	2.348	2.748	2.987
auc=0.75	1.491	1.782	2.087	2.370	2.529
auc=0.85	1.173	1.316	1.468	1.616	1.708
auc=0.95	0.483	0.522	0.566	0.606	0.631

*data consisted of ratings for 50 normal and 50 abnormal subjects; 10,000 datasets were simulated.

Table 3.8 Differences of 2.5% and 97.5% estimated percentiles of sampling distributions of standardized pAUC for straight-line ROC curves

Parameters of the ROC curves	Ranges of False Positive Fractions				
	0-0.2	**0-0.4**	**0-0.6**	**0-0.8**	**0-1**
auc=0.55	0.128	0.161	0.190	0.214	0.224
auc=0.65	0.151	0.173	0.190	0.206	0.215
auc=0.75	0.150	0.164	0.178	0.190	0.196
auc=0.85	0.133	0.143	0.150	0.156	0.160
auc=0.95	0.088	0.088	0.093	0.096	0.098

*data consisted of ratings for 50 normal and 50 abnormal subjects; 10,000 datasets were simulated.

Table 3.9 Statistical power for testing spAUC=0.5 when the ROC curve has a straight-line shape

Parameters of the ROC curves	Ranges of False Positive Fractions				
	0-0.2	**0-0.4**	**0-0.6**	**0-0.8**	**0-1**
auc=0.55	0.388	0.226	0.181	0.147	0.133
auc=0.65	0.988	0.954	0.869	0.776	0.712
auc=0.75	1.000	1.000	1.000	0.997	0.992
auc=0.85	1.000	1.000	1.000	1.000	1.000
auc=0.95	1.000	1.000	1.000	1.000	1.000

*data consisted of ratings for 50 normal and 50 abnormal subjects; 1000 datasets were simulated and 1000 results of the bootstrap tests for testing the null hypothesis spAUC=0.5 were performed.

Table 3.10 Sample size requirements for two-sided 95% confidence interval for a standardized pAUC to be narrower than 0.1 when the ROC curve has a straight-line shape

Parameters of the ROC curves	Ranges of False Positive Fractions				
	0-0.2	**0-0.4**	**0-0.6**	**0-0.8**	**0-1**
auc=0.55	170	268	367	461	513
auc=0.65	232	302	369	431	469
auc=0.75	234	280	328	372	397
auc=0.85	184	207	230	254	268
auc=0.95	76	82	89	95	99

*data consisted of ratings for 50 normal and 50 abnormal subjects; 10,000 datasets were simulated for evaluating the variance of empirical spAUCs.

Table 3.11 Sample size requirements for testing spAUC=0.5 when the ROC curve has a straight-line shape

Parameters of the ROC curves	Ranges of False Positive Fractions				
	0-0.2	**0-0.4**	**0-0.6**	**0-0.8**	**0-1**
auc=0.55	170	268	367	461	513
auc=0.65	26	34	41	48	52
auc=0.75	9	11	13	15	16
auc=0.85	4	4	5	5	5
auc=0.95	1	1	1	1	1

*data consisted of ratings for 50 normal and 50 abnormal subjects; 10,000 datasets were simulated for evaluating the variance of empirical spAUCs.

Scenario 3:

In this scenario, we investigate the properties of standardized pAUC for bi-gamma ROC curves. Table 3.12 shows that the standardized pAUC for bi-gamma ROC curves with the same shape parameter κ in the distribution of ratings for diseased and non-diseased subjects always increases with increasing range. This increase is expected since the constant shape bi-gamma ROC curves are concave (Dorfman *et al.*, 1996), and therefore, according to the proposition 3.2 the spAUC is always increasing.

Table 3.12 Theoretical value of spAUCs for bi-gamma ROC curves with different k's and full AUCs

Parameters for lower ROC curve	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
k=3					
AUC=0.55	0.523	0.534	0.542	0.548	0.550
AUC=0.65	0.583	0.614	0.633	0.645	0.650
AUC=0.75	0.666	0.708	0.731	0.745	0.750
AUC=0.85	0.775	0.814	0.835	0.846	0.850
AUC=0.95	0.914	0.934	0.944	0.948	0.950
k=2					
AUC=0.55	0.524	0.535	0.543	0.548	0.550
AUC=0.65	0.587	0.617	0.635	0.646	0.650
AUC=0.75	0.673	0.712	0.733	0.746	0.750
AUC=0.85	0.783	0.818	0.837	0.847	0.850
AUC=0.95	0.919	0.936	0.944	0.949	0.950
k=1					
AUC=0.55	0.526	0.537	0.544	0.549	0.550
AUC=0.65	0.596	0.623	0.638	0.647	0.650
AUC=0.75	0.688	0.720	0.738	0.747	0.750
AUC=0.85	0.800	0.827	0.841	0.848	0.850
AUC=0.95	0.929	0.941	0.946	0.949	0.950
k=1/2					
AUC=0.55	0.530	0.540	0.546	0.549	0.550
AUC=0.65	0.609	0.631	0.642	0.648	0.650
AUC=0.75	0.708	0.731	0.743	0.748	0.750
AUC=0.85	0.820	0.837	0.845	0.849	0.850
AUC=0.95	0.939	0.945	0.948	0.950	0.950
k=1/3					
AUC=0.55	0.533	0.543	0.547	0.549	0.550
AUC=0.65	0.618	0.636	0.645	0.649	0.650
AUC=0.75	0.720	0.737	0.745	0.749	0.750
AUC=0.85	0.830	0.842	0.847	0.849	0.850
AUC=0.95	0.943	0.947	0.949	0.950	0.950

The results for the empirical estimator of the standardized pAUC for bi-gamma ROC curves are summarized in Tables 3.13 and 3.14. For $\kappa \geq 1$, the variance as well as the length of 95% confidence interval decrease with increasing range for higher AUC (AUC greater than or equal to 0.85). These results are similar to those obtained for concave bi-normal ROC curves. For $\kappa=1$, the bi-gamma ROC curves degenerate to power-law ROC curves. For $\kappa<1$, the variance and the length of 95% confidence interval always increase with increasing range. These results are similar to those obtained for straight-line ROC curves.

Table 3.13 Variance of sampling distributions of standardized pAUC for bi-gamma ROC curves ($\times 10^{-3}$)

Parameters for lower ROC curve	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
k=3					
AUC=0.55	1.117	1.863	2.550	3.147	3.466
AUC=0.65	1.757	2.242	2.614	2.876	3.004
AUC=0.75	2.284	2.292	2.305	2.316	2.333
AUC=0.85	2.090	1.725	1.561	1.487	1.462
AUC=0.95	0.870	0.606	0.501	0.451	0.434
k=2					
AUC=0.55	1.076	1.786	2.448	2.999	3.307
AUC=0.65	1.759	2.228	2.595	2.865	3.009
AUC=0.75	2.221	2.238	2.285	2.333	2.368
AUC=0.85	2.006	1.710	1.585	1.531	1.520
AUC=0.95	0.819	0.612	0.529	0.492	0.480
k=1					
AUC=0.55	1.105	1.829	2.490	3.067	3.391
AUC=0.65	1.744	2.163	2.518	2.831	3.008
AUC=0.75	2.045	2.086	2.174	2.268	2.336
AUC=0.85	1.741	1.571	1.530	1.535	1.554
AUC=0.95	0.658	0.555	0.521	0.511	0.514
k=1/2					
AUC=0.55	1.129	1.774	2.388	2.968	3.304
AUC=0.65	1.704	2.118	2.484	2.803	2.998
AUC=0.75	1.909	2.008	2.166	2.331	2.440
AUC=0.85	1.443	1.403	1.454	1.530	1.587
AUC=0.95	0.542	0.518	0.525	0.542	0.558
k=1/3					
AUC=0.55	1.121	1.787	2.402	2.951	3.266
AUC=0.65	1.704	2.090	2.492	2.863	3.074
AUC=0.75	1.739	1.893	2.104	2.318	2.454
AUC=0.85	1.354	1.398	1.496	1.608	1.676
AUC=0.95	0.521	0.523	0.544	0.571	0.592

*data consisted of ratings for 50 normal and 50 abnormal subjects; 10,000 datasets were simulated.

Table 3.14 Differences of 2.5% and 97.5% estimated percentiles of sampling distributions of standardized pAUC for bi-gamma ROC curves

Parameters for lower ROC curve	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
k=3					
AUC=0.55	0.129	0.169	0.198	0.220	0.230
AUC=0.65	0.164	0.186	0.200	0.211	0.215
AUC=0.75	0.186	0.188	0.187	0.188	0.188
AUC=0.85	0.178	0.163	0.155	0.151	0.149
AUC=0.95	0.114	0.096	0.087	0.083	0.081
k=2					
AUC=0.55	0.128	0.166	0.196	0.214	0.226
AUC=0.65	0.163	0.184	0.200	0.210	0.215
AUC=0.75	0.183	0.186	0.186	0.189	0.190
AUC=0.85	0.174	0.161	0.154	0.151	0.152
AUC=0.95	0.110	0.096	0.090	0.086	0.085
k=1					
AUC=0.55	0.128	0.166	0.195	0.217	0.229
AUC=0.65	0.163	0.183	0.197	0.207	0.214
AUC=0.75	0.177	0.178	0.180	0.185	0.187
AUC=0.85	0.162	0.154	0.150	0.151	0.152
AUC=0.95	0.099	0.091	0.088	0.087	0.088
k=1/2					
AUC=0.55	0.130	0.164	0.190	0.210	0.225
AUC=0.65	0.161	0.181	0.195	0.208	0.215
AUC=0.75	0.171	0.174	0.181	0.189	0.192
AUC=0.85	0.148	0.148	0.150	0.153	0.155
AUC=0.95	0.090	0.088	0.089	0.090	0.092
k=1/3					
AUC=0.55	0.130	0.164	0.192	0.211	0.223
AUC=0.65	0.161	0.179	0.194	0.209	0.217
AUC=0.75	0.164	0.171	0.179	0.190	0.196
AUC=0.85	0.142	0.145	0.151	0.157	0.161
AUC=0.95	0.089	0.088	0.090	0.093	0.094

*data consisted of ratings for 50 normal and 50 abnormal subjects; 10,000 datasets were simulated.

Table 3.15 shows that the statistical power for evaluation of a single pAUC for bi-gamma frequently decreased with increasing range for AUC<0.85. The increasing trend can only be observed for $\kappa>1$.

Table 3.15 Statistical power for testing spAUC=0.5 when the ROC curve has a bi-gamma shape

Parameters for	Ranges of false positive fractions				
lower ROC curve	0-0.2	0-0.4	0-0.6	0-0.8	0-1
k=3					
AUC=0.55	0.151	0.170	0.170	0.158	0.144
AUC=0.65	0.652	0.720	0.733	0.717	0.708
AUC=0.75	0.986	0.995	0.999	0.999	0.997
AUC=0.85	1.000	1.000	1.000	1.000	1.000
AUC=0.95	1.000	1.000	1.000	1.000	1.000
k=2					
AUC=0.55	0.151	0.165	0.153	0.148	0.140
AUC=0.65	0.704	0.762	0.773	0.773	0.748
AUC=0.75	0.995	1.000	1.000	0.997	0.997
AUC=0.85	1.000	1.000	1.000	1.000	1.000
AUC=0.95	1.000	1.000	1.000	1.000	1.000
k=1					
AUC=0.55	0.133	0.137	0.141	0.143	0.133
AUC=0.65	0.782	0.825	0.812	0.772	0.749
AUC=0.75	0.999	0.998	0.998	0.994	0.989
AUC=0.85	1.000	1.000	1.000	1.000	1.000
AUC=0.95	1.000	1.000	1.000	1.000	1.000
k=1/2					
AUC=0.55	0.185	0.193	0.165	0.138	0.132
AUC=0.65	0.878	0.862	0.828	0.796	0.760
AUC=0.75	1.000	1.000	0.999	0.998	0.995
AUC=0.85	1.000	1.000	1.000	1.000	1.000
AUC=0.95	1.000	1.000	1.000	1.000	1.000
k=1/3					
AUC=0.55	0.213	0.208	0.190	0.163	0.152
AUC=0.65	0.919	0.902	0.864	0.799	0.754
AUC=0.75	1.000	1.000	1.000	1.000	0.999
AUC=0.85	1.000	1.000	1.000	1.000	1.000
AUC=0.95	1.000	1.000	1.000	1.000	1.000

*data consisted of ratings for 50 normal and 50 abnormal subjects; 1000 datasets were simulated and 1000 results of the bootstrap tests for testing the null hypothesis spAUC=0.5 were performed.

We developed a program (Appendix C) for estimating sample size for evaluation of a single pAUC under the bi-gamma assumption for the ROC curves. Table 3.16 shows the sample size requirements for a two-sided 95% confidence interval for a standardized pAUC to be

narrower than 0.1 with probability corresponding to 1-β=0.8. Sample size was estimated using the method proposed by Flahault *et al.* (2005). For κ≥1, the decreasing trend in sample size can only be observed for bi-gamma ROC curves with AUC ≥ 0.85, which was similar to bi-normal ROC curves (Table 3.4). For κ<1, sample size requirements always increase with increasing range, which was similar to straight-line ROC curves.

Table 3.16 Sample size requirements for two-sided 95% confidence interval for a standardized pAUC to be narrower than 0.1 when the ROC curve has a bi-gamma shape

Parameters for lower ROC curve	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
k=3					
AUC=0.55	157	263	362	450	492
AUC=0.65	265	330	382	422	438
AUC=0.75	345	337	339	343	343
AUC=0.85	317	252	229	216	212
AUC=0.95	136	91	76	70	67
k=2					
AUC=0.55	167	280	384	475	520
AUC=0.65	275	343	399	444	462
AUC=0.75	335	340	349	358	361
AUC=0.85	296	253	234	226	224
AUC=0.95	126	93	80	75	74
k=1					
AUC=0.55	167	276	367	454	502
AUC=0.65	277	336	387	435	460
AUC=0.75	326	332	347	363	373
AUC=0.85	282	252	241	242	243
AUC=0.95	107	91	85	83	83
k=1/2					
AUC=0.55	179	291	400	489	538
AUC=0.65	279	344	405	456	486
AUC=0.75	301	317	343	370	387
AUC=0.85	229	225	234	245	254
AUC=0.95	90	86	87	91	92
k=1/3					
AUC=0.55	167	277	380	467	518
AUC=0.65	257	313	377	430	466
AUC=0.75	265	287	320	355	380
AUC=0.85	200	203	223	245	259
AUC=0.95	90	91	95	102	105

*data consisted of ratings for 50 normal and 50 abnormal subjects; 10,000 datasets were simulated for evaluating the variance of empirical spAUCs.

Table 3.17 Sample size requirements for testing spAUC=0.5 when the ROC curve has a bi-gamma shape

Parameters for lower ROC curve	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
k=3					
AUC=0.55	829	632	567	536	544
AUC=0.65	100	68	58	54	52
AUC=0.75	33	21	17	15	15
AUC=0.85	11	7	5	5	5
AUC=0.95	2	1	1	1	1
k=2					
AUC=0.55	733	572	520	511	519
AUC=0.65	91	64	56	53	52
AUC=0.75	29	20	17	15	15
AUC=0.85	10	7	5	5	5
AUC=0.95	2	1	1	1	1
k=1					
AUC=0.55	642	524	505	501	532
AUC=0.65	74	56	52	51	52
AUC=0.75	23	17	15	15	15
AUC=0.85	8	6	5	5	5
AUC=0.95	1	1	1	1	1
k=1/2					
AUC=0.55	492	435	443	485	519
AUC=0.65	56	48	48	50	52
AUC=0.75	17	15	14	15	15
AUC=0.85	6	5	5	5	5
AUC=0.95	1	1	1	1	1
k=1/3					
AUC=0.55	404	379	427	482	513
AUC=0.65	48	44	47	51	54
AUC=0.75	14	13	14	15	15
AUC=0.85	5	5	5	5	5
AUC=0.95	1	1	1	1	1

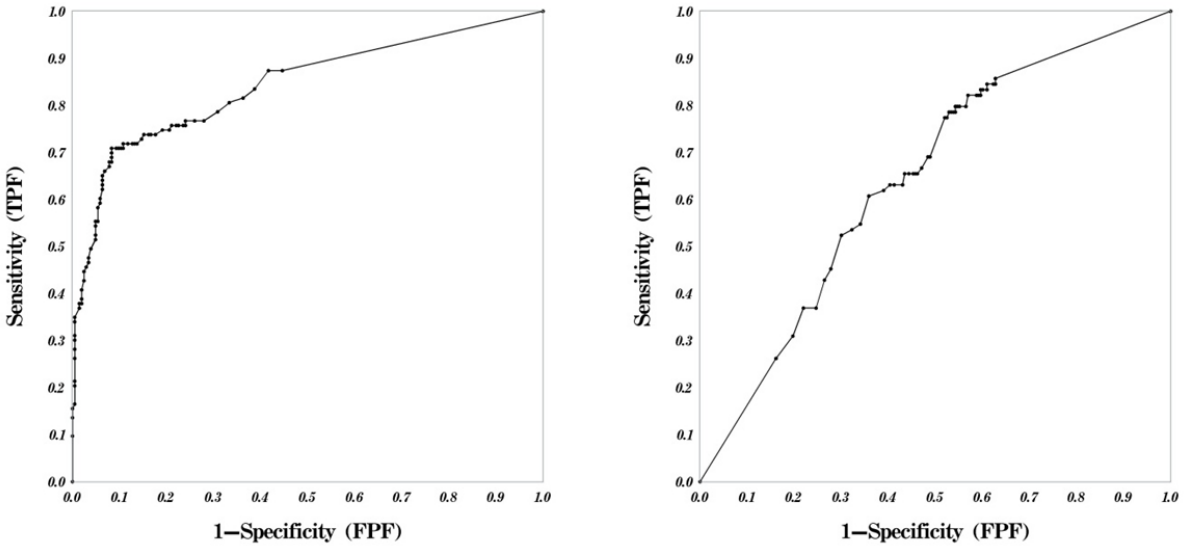
*data consisted of ratings for 50 normal and 50 abnormal subjects; 10,000 datasets were simulated for evaluating the variance of empirical spAUCs.

3.3 EXAMPLES

In this section we illustrate the patterns described in the previous sections with an example obtained from two datasets from observer performance studies we previously conducted (Gur *et al.*, 2009). One dataset (307 cases, 103 abnormal and 204 normal) includes observer's ratings for classification of images as depicting/non-depicting lung nodules. The second dataset (307 cases, 84 abnormal and 223 normal) includes observer's ratings for classification of images in regard to

presence/absence of subtle interstitial disease. For both datasets the diagnostic ratings were provided by a group of radiologists using a pseudo-continuous scale from 0 to 100.

For each dataset we estimated empirical ROC curves by connecting empirical points with straight lines (Zhou *et al.*, 2002) (Pepe, 2003). The estimates of the standardized partial AUCs were computed by integration for ranges starting at 0 and ending at 0.2, 0.4, 0.6, 0.8, and 1. Variance of the empirical estimator of the standardized partial AUC was estimated using a nonparametric bootstrap approach (Efron and Tibshirani, 1993). The bootstrap percentile confidence intervals were computed using 10,000 random bootstrap samples.



a) Chest nodules (AUC=0.843) b) Interstitial lung disease (AUC=0.644)

Figure 3.4 Empirical ROC curves for the two datasets

Figure 3.4 illustrates the empirical ROC curves for the two datasets. Table 3.18 summarizes the standardized partial area, its bootstrap variance, and the length of the 95% bootstrap confidence interval. In agreement with our findings in Chapter 3.1 for both empirical ROC curves the standardized partial areas were increasing with increasing range. In agreement

with our findings in Chapter 3.2, for the ROC curve with AUC=0.84, the variance estimator of the standardized partial area first decreased and then remained virtually unchanged. Since data for interstitial disease included very subtle cases, the ROC curve had a relatively low AUC of 0.64 and the bootstrap variance of standardized partial area for the ROC curve increases over the considered ranges. The same trend was observed for the length of the 95% bootstrap confidence interval.

Table 3.18 Example: Empirical standardized partial areas and their variance for sample data from studies of detection of lung nodules and interstitial disease

	0-0.2	0-0.4	0-0.6	0-1
Nodule				
Stand pAUC	0.796	0.819	0.835	0.843
Standard deviation	0.0270	0.0261	0.0257*	0.0257*
Length of 95% bootstrap CI	0.1058	0.1020	0.0993	0.0998
Interstitial				
Stand pAUC	0.534	0.579	0.613	0.644
Standard deviation	0.0206	0.0298	0.0329	0.0334
Length of 95% bootstrap CI	0.0799	0.1160	0.1271	0.1304

*Further increase of the range does not increase the number of included empirical operating points.

3.4 SUMMARY

In practice inferences based on the partial AUC could be both more clinically relevant and more statistically conclusive than inference based on full AUC.

In many practical problems increasing the range of interest for partial area would lead to an increase in the estimated level of diagnostic accuracy, even after application of existing standardizations.

Effect of the increasing range on the sampling variability depends on the shape of the ROC curve.

There exists ROC curves for which inference based on shorter ranges for partial AUC are always preferable. At the same time evaluation of binormal ROC curves can often be more efficiently performed using partial AUC over the full range (full AUC).

The approaches for sample size estimation based on binormal ROCs often mask statistical advantages of the partial AUC that may be real in practice.

We demonstrated that family of constant shape bi-gamma ROC curves allows more realistic reflection of properties of pAUC analysis. Bi-gamma family of ROC curves covers many practically reasonable and plausible shapes and includes ROC curves that are close to the straight line, as well as concave ROC curves that are similar in shape to binormal ROC curves.

4.0 COMPARISON OF TWO CORRELATED PAUCS

In comparison of two diagnostic systems, the primary interest is often in comparing two modalities on the basis of pAUC and AUC. Data for this problem is often collected under the paired design where each case is rated under every modality. Analysis of data collected under the paired design requires addressing the possible correlation between the ratings assigned to the same case.

We analyzed properties of the difference in partial AUC as a function of the size of the range of interest and conducted extensive simulation studies of statistical power for comparisons of correlated pAUCs in families of binormal, straight-line, and bi-gamma ROC curves.

We demonstrated that, in contrast to the single standardized partial AUC, the difference in two pAUC does not always increases even for concave non-crossing ROC curves. The approximate graphical approach was described for determining whether the difference would increase with increasing range. In simulation studies we demonstrated that the use of pAUC was statistically advantageous in several types of performance curves. For binormal ROC curves with low AUC, an increase in range often leads to an increase in spAUCs differences, thereby contributing to increasing statistical power. However, when ROC curves approached the shape of a specific straight-line shape, the difference in standardized pAUCs became more stable, and the statistical power decreased with increasing range. Thus, the relative statistical power for pAUC-based comparisons is affected not only by the height, but also by the shape of ROC curves. For

adequately planning studies based on the pAUC, we propose to use the bi-gamma ROC model which includes curves with nearly binormal shape as well as curves with nearly straight-line shape. For many practical ROC curves, studies focusing on clinically relevant pAUCs would actually require smaller sample sizes than studies based on the full AUC. This portion of the research has been submitted in 2014 (Appendix B).

4.1 METHOD

The range of clinical interest (relevance) has a natural effect on the magnitude of the partial area under the ROC curve (pAUC). Several approaches to standardization of the pAUC (McClish, 1989) (Jiang *et al*., 1996) alleviate the problem, but do not address it completely (Ma *et al*., 2013). Since the magnitude of differences in standardized pAUCs could directly affect the statistical power of comparisons of partial AUCs, to investigate the statistical properties of comparisons, it is important to understand the patterns of these differences in standardized pAUCs. In addition, knowledge of the conditions when the differences between standardized pAUCs increase or decrease helps one to better interpret reported results of analyses based on the pAUC.

The absolute difference in pAUCs always increases for non-crossing ROC curves. Indeed, the derivative of the difference is the difference in the ROC points corresponding to the end of the range of interest, which does not change signs for non-crossing curves, i.e.:

$$\frac{\partial}{\partial e}\left(\int_0^e ROC_2(f)\,df - \int_0^e ROC_1(f)\,df\right) = ROC_2(e) - ROC_1(e) > 0\,.$$

This relationship however, offers little insight into the ability to declare statistically significant differences, since the variability of the pAUC also increases with increasing range (Ma *et al*., 2013). Since the standardized pAUC is a linear function of the pAUC, the test for comparison of partial AUCs could be viewed as a test for equality of the standardized pAUC. Since the standardized pAUC is more stable with increasing range, its properties are also more relevant for investigating statistical power. In contrast with the difference between pAUCs, the difference between standardized pAUCs could either increase or decrease. Indeed, based on the definition of the spAUC (1.1), the difference between standardized pAUCs can be written as:

$$\tilde{A}_e^2 - \tilde{A}_e^1 = \frac{1}{2}\frac{A_e^2 - A_e^1}{e - \frac{e^2}{2}} = \frac{A_e^2 - A_e^1}{2e - e^2} \tag{4.1}$$

If the increase in the value of $2e\text{-}e^2$ with increasing range cannot compensate for the increase in pAUC, the difference in the standardized pAUC will increase. Otherwise, the difference in the standardized pAUC will either remain unchanged or decrease.

The following proposition establishes the fact that the spAUC difference increases as long as it is smaller than half of the difference between the negative diagnostic likelihood ratios (DLR-) at the end of the range of interest. For a given point (e, ROC(e)) on the ROC curve the negative diagnostic likelihood ratio is defined as follows:

$$DLR^{-}(e) = \frac{1 - ROC\ (e)}{1 - e} \tag{4.2}$$

We note that the negative diagnostic likelihood ratio for a given point on the ROC curve is different from the “likelihood ratio” (which is equal to the slope of the ROC curve at any given point).

Proposition 4.1:

For any $e \in (0,1)$,

$$sgn\left\{\frac{\partial \Delta \tilde{A}_e}{\partial e}\right\} = sgn\left\{\Delta DLR^{-}(e) - 2\Delta \tilde{A}_e\right\}$$

where $\Delta \tilde{A}_e = \tilde{A}_e^2 - \tilde{A}_e^1 \quad \Delta DLR^{-}(e) = DLR_1^{-}(e) - DLR_2^{-}(e)$

and $\text{sgn}(x) = \begin{cases} -1 & if\ x < 0 \\ 0 & if\ x = 0 \\ 1 & if\ x > 0 \end{cases}$

Proof:

Based on the definition of the spAUC (1.1), the difference of the standardized partial areas can be written as:

$$\tilde{A}_e^2 - \tilde{A}_e^1 = \frac{1}{2}\frac{\int_0^e ROC_2(f)df - \int_0^e ROC_1(f)df}{e - \frac{e^2}{2}}$$

By differentiation of the difference in the spAUCs $\tilde{A}_e^2 - \tilde{A}_e^1$ we obtain:

$$\frac{\partial\left(\tilde{A}_e^2 - \tilde{A}_e^1\right)}{\partial e} = \frac{1}{2}\left(e - \frac{e^2}{2}\right)^{-1}\left\{ROC_2(e) - ROC_1(e) - (1-e)\frac{\left(A_e^2 - A_e^1\right)}{e - \frac{e^2}{2}}\right\}$$

Since $\tilde{A}_e^2 - \tilde{A}_e^1 = \frac{1}{2}\frac{A_e^2 - A_e^1}{e - \frac{e^2}{2}}$, the derivative of the difference of the spAUCs can then be written as:

$$\frac{\partial\left(\tilde{A}_e^2 - \tilde{A}_e^1\right)}{\partial e} = \frac{1}{2}\left(e - \frac{e^2}{2}\right)^{-1}\left\{ROC_2(e) - ROC_1(e) - 2(1-e)\left(\tilde{A}_e^2 - \tilde{A}_e^1\right)\right\}$$

The conclusion of this proposition follows immediately from the above equation, the definition of DLR^- and the fact that $\left(2e-e^2\right)\Big/\left(1-e\right)$ is positive for any e from (0, 1).

Negative diagnostic likelihood ratio, $DLR^-(e)$, is easy to visualize as the slope of the line extending from a given point on the ROC curve to (1, 1). It is known to decrease for any concave curve. However, the difference in DLR^-'s of points between two concave ROC curves may either increase or decrease. Figure 4.1 illustrates the difference in the DLR^-'s and the difference in the spAUCs for two concave binormal ROC curves with AUCs of 0.80 and 0.85 respectively. At the FPF point where the $\Delta DLR^-/2$ and ΔspAUC curves cross, the difference in the spAUCs reaches its maximum value for the ROC curves being compared.

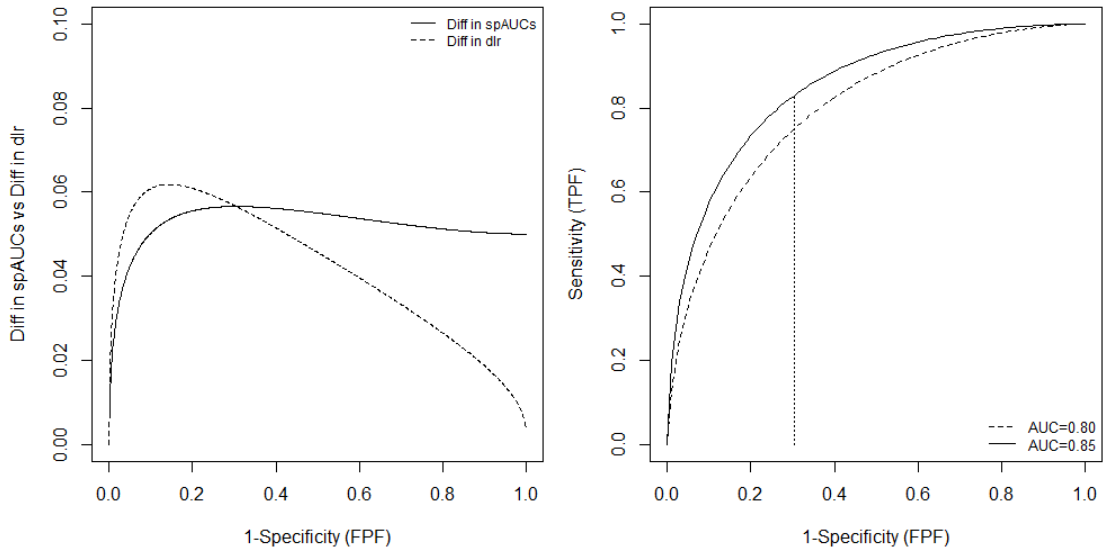


Figure 4.1 b=1 and lower AUC=0.8

Furthermore, it is possible that two ROC curves are concave with continually increasing differences in DLR^-'s, and it is also possible for two concave ROC curves to have a constant difference in DLR^-'s. One simple type of a curve that has a constant difference in standardized partial AUC is the straight-line ROC curve. From Chapter 2.1.4, we define the straight-line ROC

curve as linear ROC curve passing through the point (1, 1) (Ma *et al*., 2013); the straight-line ROC curve with full AUC of *A* has the following functional form:

$$ROC(e) = (2A - 1) + 2(1 - A)e \tag{4.3}$$

The straight-line ROC curve has a constant DLR^- of 2(1-*A*) and the standardized pAUC for any range of interest is constant and equal to *A* (Ma *et al*., 2013). Furthermore, from the reformulation of the straight-line ROC curve in terms of its DLR^- it can be seen that the difference in standardized pAUCs of the two straight-line ROC curves equals to the half of the difference in their DLR^-'s. Combined with the fact that at a fixed point an ROC curve has the same DLR^- as the straight-line ROC curve passing through this point, proposition 1 can be reformulated as follows: "The difference in standardized pAUCs increases/decreases if it is smaller/larger than the difference in the standardized pAUCs of straight-line ROC curves passing through the same points at the end of the range of interest". Since these standardized pAUCs are considered over the same range, the proposition can be equivalently formulated in terms of the difference in pAUCs between the ROC curves of interest and the corresponding straight-line ROC curves. This enables an approximate visual inspection of changes in the standardized pAUCs difference with increasing range by visually comparing the area between the ROC curves over the range of interest to the corresponding area between the two straight line-ROC curves passing through the same points at the end of the range of interest. In particular, the difference in standardized pAUCs reaches its maximum when it is the same as the difference in pAUCs of the corresponding straight-line ROC curves. This leads to the Corollary 4.2 and Figure 4.2. The shaded area in the left plot of Figure 4.2 shows the difference in pAUCs for two ROC curves over the range of interest and the shaded area in the right plot of Figure 4.2 shows the difference in pAUCs for corresponding straight-line ROC curves.

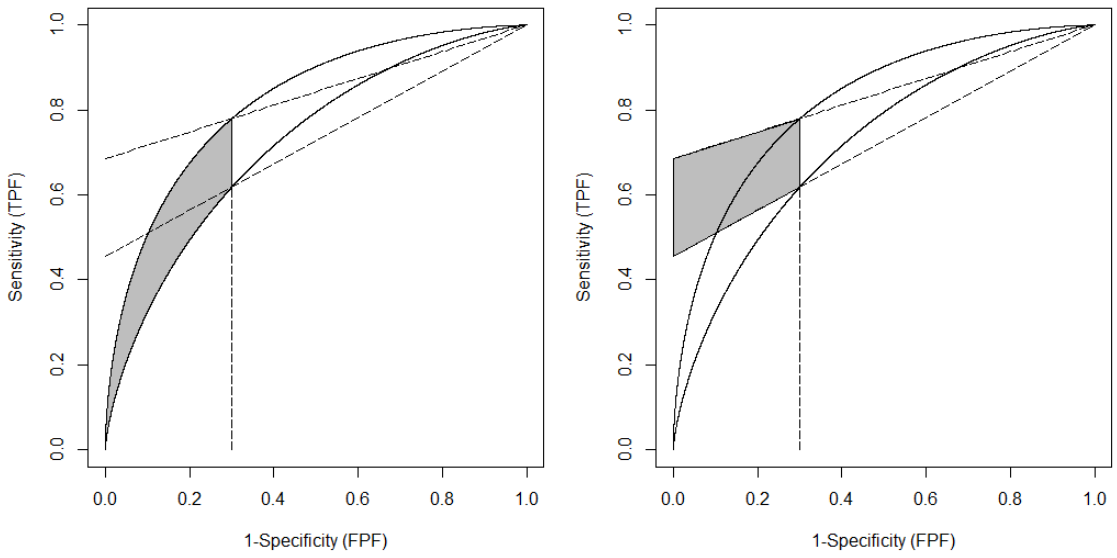


Figure 4.2 Difference in pAUCs for ROC curves of interest (left) vs. Difference in pAUCs for straight-line ROC curves (right)

Corollary 4.2:

For any $e \in (0,1)$ and two ROC curves ROC^1(e) and ROC^2(e), let $\Delta A_e = A_e^2 - A_e^1$ represents the difference in pAUCs over the range (0,e) and let $\Delta A_e^{straight} = A_e^{2,straight} - A_e^{1,straight}$ represents the corresponding difference in pAUCs for straight-line ROC curves passing through (e,ROC^2(e)) and (e,ROC^1(e)) correspondingly. If $\Delta \tilde{A}_e^{straight}$ is the difference in the standardized pAUCs for ROC^2 and ROC^1, then

$$sgn\left\{\frac{\partial \Delta \tilde{A}_e}{\partial e}\right\} = sgn\left\{\Delta A_e^{straight} - \Delta A_e\right\}$$

$$\text{sgn}(x) = \begin{cases} -1 & if\ x < 0 \\ 0 & if\ x = 0 \\ 1 & if\ x > 0 \end{cases}$$

Proof:

From definition of the straight-line ROC curves (4.3) and DLR- it follows that

$$\Delta\tilde{A}_e^{straight} = \Delta DLR^{-}(e)/2 .$$

Using this fact the Proposition 1 can be rewritten as follows:

$$sgn\left\{\frac{\partial\Delta\tilde{A}_e}{\partial e}\right\} = sgn\left\{2\Delta\tilde{A}_e^{straight} - 2\Delta\tilde{A}_e\right\} = sgn\left\{\Delta\tilde{A}_e^{straight} - \Delta\tilde{A}_e\right\}$$

The conclusion of this corollary immediately follows from the definition of the standardized pAUC.

Since for two almost-linear ROC curves the difference in the standardized pAUCs is approximately constant, to better understand a direct effect of increasing range of interest on the statistical power when comparing two partial AUCs we can consider binormal and "almost" linear ROC curves. Both binormal and almost linear ROC curves can be encountered in practice (Hanley, 1988) (Gur *et al.*, 2007). Thus, it is important to understand the properties of the statistical comparison of pAUCs for both types of ROC curves. Statistical power can be affected by both the magnitude of the differences attempted to be detected and the sampling variability of the estimates.

The difference between spAUCs will be approximately constant for two ROC curves, each with "almost" a linear shape. For example, the standardized pAUC (and hence the difference) remains virtually constant for the bi-gamma ROC curve with a small value of the shape parameter κ (e.g., ranges from 0.25 to 1.00 for κ=1/3), which we discuss in detail later in this chapter. This type of curve, however, cannot be well approximated by a binormal ROC curve unless it has an improper shape. Yet, both binormal and "almost" linear types of ROC curves could approximate reasonably well some empirical data (Hanley, 1988), and, as we

demonstrate later, different types of curves could have substantially different properties during statistical comparisons of pAUCs. In the next section we perform a comprehensive numerical investigation of the properties of statistical comparisons for several types of ROC curves.

4.2 NUMERICAL STUDY

In this section we consider several families of ROC curves. A pair of ROC curves from the same family is used to represent the performance of two diagnostic tests being compared. Computations of true parameters of these ROC curves, including pAUCs, were conducted using numerical integration (Piessens *et al*., 1983).

In the simulation studies parameters for each ROC curve were determined, as well as the distribution of ratings (diagnostic scores) for 150 normal and 150 abnormal subjects. Pairs of ratings for the same subjects (representing results of the two diagnostic tests being compared) were correlated by sharing a subject-specific random effect adjusted to generate correlation of the targeted magnitude. For each generated dataset, pAUCs were estimated using area under the empirical (linearly interpolated) ROC curves over the given range of interest. The statistical test for equality of two pAUCs was performed using non-parametric bootstrap approach based on 1000 resamples of normal and abnormal subjects, separately. Statistical power was estimated from 1000 results of the bootstrap tests.

Scenario 1:

We first investigated properties of comparisons of pAUCs, A_e^1 and A_e^2, for two binormal ROC curves with the same shape parameter *b* and a constant difference between the full AUCs, A^1 and A^2. Since the two binormal ROC curves have the same parameter *b*, they do not cross

each other. Thus, as noted in the previous section, the difference in the pAUCs ($A_e^2 - A_e^1$) increases with increasing range. $A_e^2 - A_e^1$ reaches a maximum value (equal to the difference in full AUCs) at e=1. The standardized difference also increases in most scenarios.

For this scenario, with increasing range of interest the difference in standardized pAUCs $\tilde{A}_e^2 - \tilde{A}_e^1$ does not always increase, but does increase rather frequently. Table 4.1 shows the differences in the standardized pAUCs when the difference in the full AUCs of the two ROC curves is 0.05. The difference in the standardized pAUCs decreases with increasing range of interest for ROC curves with high AUCs (e.g., AUC of 0.8 and 0.9 for concave ROC curve with b=1, and AUC of 0.9 for improper ROC curve with b=0.5). This agrees with proposition 4.1, since in proximity to the point (1, 1), binormal ROC curves with large AUC tend to have a small slope, thereby leading to a small difference in DLR^-'s, and eventually to a decreasing difference in the standardized pAUCs.

Table 4.1 Theoretical $\tilde{A}_e^2 - \tilde{A}_e^1$ for binormal ROC curves with same *b* and a constant difference between full AUCs

The lower AUC	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
b=0.33					
AUC=0.60	0.030	0.034	0.038	0.044	0.050
AUC=0.70	0.034	0.037	0.040	0.045	0.050
AUC=0.80	0.039	0.040	0.043	0.046	0.050
AUC=0.90	0.047	0.046	0.046	0.048	0.050
b=0.50					
AUC=0.60	0.030	0.035	0.040	0.045	0.050
AUC=0.70	0.036	0.040	0.043	0.046	0.050
AUC=0.80	0.044	0.045	0.046	0.048	0.050
AUC=0.90	0.058	0.052	0.050	0.050	0.050
b=1.0					
AUC=0.60	0.027	0.038	0.045	0.049	0.050
AUC=0.70	0.038	0.046	0.049	0.050	0.050
AUC=0.80	0.056	0.056	0.054	0.051	0.050
AUC=0.90	0.083	0.067	0.057	0.052	0.050

To investigate properties of the variance of the difference in spAUCs and statistical power, we conducted a simulation study. Each generated dataset consisted of ratings for 150 normal (X_{i1}, X_{i2}) and 150 abnormal subjects (Y_{j1}, Y_{j2}) where i, j=1,2…,150. Ratings were generated from bivariate normal distributions with a correlation of 0.5. Exploiting the invariance property of the ROC curve to monotonically increasing transformation of the ratings, the distributions of ratings of normal subjects were set to bivariate normal distribution with mean $\mu=(0,0)^T$, and variance covariance matrix $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

Parameters for the distributions of ratings of abnormal subjects were selected to reflect the pre-specified shape of ROC curves and areas under these curves while preserving the correlation of 0.5 between ratings corresponding to the same subjects.

Table 4.2 summarizes the estimated variance of difference in spAUCs between two binormal ROC curves with a difference of 0.05 in full AUCs. The results show that the variance

frequently increases with increasing range. The decreasing trend can only be observed for ROC curves with high AUC (e.g., AUC of 0.8 and 0.9 for concave ROC curve with *b*=1, and AUC of 0.9 for improper ROC curve with *b*=0.5). This trend looks as similar to the difference in spAUCs.

Table 4.2 Variance of empirical $\tilde{A}_e^2 - \tilde{A}_e^1$ for binormal ROC curves with same *b* and a constant difference between full AUCs ($\times 10^{-4}$)

The lower AUC	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
b=0.33					
AUC=0.60	6.654	7.724	9.301	11.377	13.883
AUC=0.70	7.209	7.644	8.804	10.533	12.596
AUC=0.80	5.475	5.581	6.218	7.333	8.642
AUC=0.90	3.549	3.497	3.746	4.116	4.601
b=0.50					
AUC=0.60	6.617	7.358	8.823	10.670	12.493
AUC=0.70	7.032	7.035	7.640	8.666	9.695
AUC=0.80	6.908	6.665	6.750	7.202	7.653
AUC=0.90	4.429	3.519	3.264	3.260	3.383
b=1.0					
AUC=0.60	7.001	8.646	9.499	9.911	9.955
AUC=0.70	9.826	10.569	10.763	10.470	10.178
AUC=0.80	12.026	9.078	7.524	6.601	6.271
AUC=0.90	8.406	4.805	3.391	2.762	2.568

*Data consisted of pairs of ratings for 150 normal and 150 abnormal subjects, with between-modality correlation of 0.5; 1000 datasets were simulated.

Table 4.3 summarizes the estimated statistical power for comparisons of pAUCs of two binormal ROC curves with a difference of 0.05 in full AUCs. The results show that the statistical power frequently increases with increasing range. The decreasing trend in statistical power can only be observed for improper ROC curves (*b*<1) with relatively high AUCs (e.g., AUC=0.9).

The observed increase of statistical power or decrease of sample size requirements with increasing range could be affected by the concurrent tendency of the difference in spAUCs to

increase. To circumvent this difficulty we investigated a family of straight-line ROC curves (4.3) in which the difference in true spAUC remains constant regardless of the range.

Table 4.3 Statistical power for comparisons of two partial AUCs of bi-normal ROC curves with differences in full AUCs of 0.05

Parameters for the lower ROC curve	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
b=0.33					
AUC=0.60	0.196	0.215	0.235	0.246	0.256
AUC=0.70	0.281	0.309	0.327	0.323	0.332
AUC=0.80	0.382	0.395	0.404	0.412	0.423
AUC=0.90	0.722	0.711	0.698	0.688	0.686
b=0.50					
AUC=0.60	0.211	0.244	0.261	0.265	0.277
AUC=0.70	0.245	0.317	0.337	0.333	0.340
AUC=0.80	0.379	0.407	0.425	0.431	0.432
AUC=0.90	0.775	0.798	0.791	0.782	0.781
b=1.0					
AUC=0.60	0.175	0.236	0.286	0.304	0.327
AUC=0.70	0.245	0.308	0.336	0.367	0.377
AUC=0.80	0.333	0.435	0.474	0.504	0.513
AUC=0.90	0.794	0.860	0.885	0.887	0.888

*Data consisted of pairs of ratings for 150 normal and 150 abnormal subjects, with between-modality correlation of 0.5; 1000 datasets were simulated and 1000 results of the bootstrap tests for testing the null hypothesis $\tilde{A}_e^2 - \tilde{A}_e^1 = 0$ were performed.

Table 4.4 summarizes the sample size requirements for comparisons of pAUCs of two binormal ROC curves with a difference of 0.05 in full AUCs. The sample size was computed using code provided in Appendix B based on the original sample of 150 diseased and 150 non-diseased subjects. The results show that the sample size requirements frequently decrease with increasing range. The increasing trend in sample size requirements can only be observed for improper ROC curves (b<1) with relatively high AUCs (e.g., AUC=0.9).

Table 4.4 Sample size requirements for comparisons of two partial AUCs of bi-normal ROC curves with differences in full AUCs of 0.05 (between-modality correlation of 0.5)

Parameters for the lower ROC curve	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
b=0.33					
AUC=0.60	870	787	758	692	654
AUC=0.70	734	657	648	612	593
AUC=0.80	424	411	396	408	407
AUC=0.90	189	195	208	210	217
b=0.50					
AUC=0.60	866	707	649	620	588
AUC=0.70	639	518	487	482	457
AUC=0.80	420	388	376	368	360
AUC=0.90	155	153	154	154	159
b=1.0					
AUC=0.60	1131	705	552	486	469
AUC=0.70	801	588	528	493	479
AUC=0.80	452	341	304	299	295
AUC=0.90	144	126	123	120	121

*Data consisted of pairs of ratings for 150 normal and 150 abnormal subjects, with between-modality correlation of 0.5; 1000 datasets were simulated for evaluating the variance of difference in empirical spAUCs.

Scenario 2:

In this section we investigate the properties of comparisons of pAUCs (A_e^1 and A_e^2), in the case of two straight-line ROC curves (4.3) with constant differences of 0.05 between the full AUCs. As discussed previously, for these ROC curves the difference in the spAUCs was also 0.05 regardless of the range of interest.

Ratings with bivariate uniform distribution and AUCs of A_i (i=1, 2) were generated by probability integral transformation of bivariate normal random variables with adjusted correlations (Rachev, 2003) (Hotelling and Pabst, 1936). The marginal distributions were X_1~Uniform(0,1) X_2~Uniform(0,1) for normal subjects and Y_1~Uniform(0,1/(2-2A_1)) and Y_2~Uniform(0,1/(2-2A_2)) for abnormal subjects, respectively. The variance covariance matrix

used was $\sum = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ for both normal and abnormal subjects. Generated ratings were then used to conduct a bootstrap test for equality of two pAUCs (as described previously). The estimated variance and statistical power are summarized in Table 4.5 and 4.6.

Results in Table 4.5 demonstrate that the variance of the difference in spAUCs in the case of two straight-line ROC curves always increases with increasing range of interest.

Table 4.5 Variance of difference between spAUCs of two straight-line ROC curves with differences in full AUCs of 0.05 ($\times 10^{-4}$)

The AUC for lower ROC curve	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
AUC=0.60	5.741	7.445	9.081	10.390	11.136
AUC=0.70	6.465	7.563	8.629	9.280	9.531
AUC=0.80	6.017	6.758	7.338	7.799	7.952
AUC=0.90	3.385	3.733	4.076	4.362	4.487

*Data consisted of pairs of ratings for 150 normal and 150 abnormal subjects, with between-modality correlation of 0.5; 1000 datasets were simulated.

Results in Table 4.6 demonstrate that statistical power for comparisons of pAUCs in the case of two straight-line ROC curves always decreases with increasing range of interest. Results in Table 4.7 demonstrated that the sample size requirements increased with increasing range.

This should also hold quite well for “almost” or nearly straight-line ROC curves. In the next section, we verified our findings using a flexible, bi-gamma, family of ROC curves that cover both nearly-linear ROC curves as well as binormal-looking ROC curves.

Table 4.6 Statistical power of comparisons of two partial AUCs of straight-line ROC curves with differences in full AUCs of 0.05

The AUC for lower ROC curve	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
AUC=0.60	0.525	0.438	0.384	0.345	0.335
AUC=0.70	0.489	0.433	0.383	0.363	0.355
AUC=0.80	0.561	0.517	0.492	0.448	0.442
AUC=0.90	0.783	0.754	0.710	0.696	0.691

*Data consisted of pairs of ratings for 150 normal and 150 abnormal subjects, with between-modality correlation of 0.5; 1000 datasets were simulated and 1000 results of the bootstrap tests for testing the null hypothesis $\tilde{A}_e^2 - \tilde{A}_e^1 = 0$ were performed.

Table 4.7 Sample size requirements of comparisons of two partial AUCs of straight-line ROC curves with differences in full AUCs of 0.05 (data consisted of pairs of ratings for 150 normal and 150 abnormal subjects, with between-modality correlation of 0.5)

The AUC for lower ROC curve	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
AUC=0.60	270	351	428	489	524
AUC=0.70	304	356	406	437	449
AUC=0.80	283	318	346	367	374
AUC=0.90	159	176	192	205	211

*Data consisted of pairs of ratings for 150 normal and 150 abnormal subjects, with between-modality correlation of 0.5; 1000 datasets were simulated for evaluating the variance of difference in empirical spAUCs.

Scenario 3:

In this section we investigated the properties of comparisons of pAUCs of two correlated bi-gamma ROC curves with a fixed difference in full AUCs. We introduced bi-gamma ROC curves and demonstrated the merits in Chapter 2.1.3.

Figure 4.3 illustrates the three types of bi-gamma ROC curves each with AUC equal to 0.8 and κ=3, 2, 1, 1/2 and 1/3, which are the values used in the simulation study.

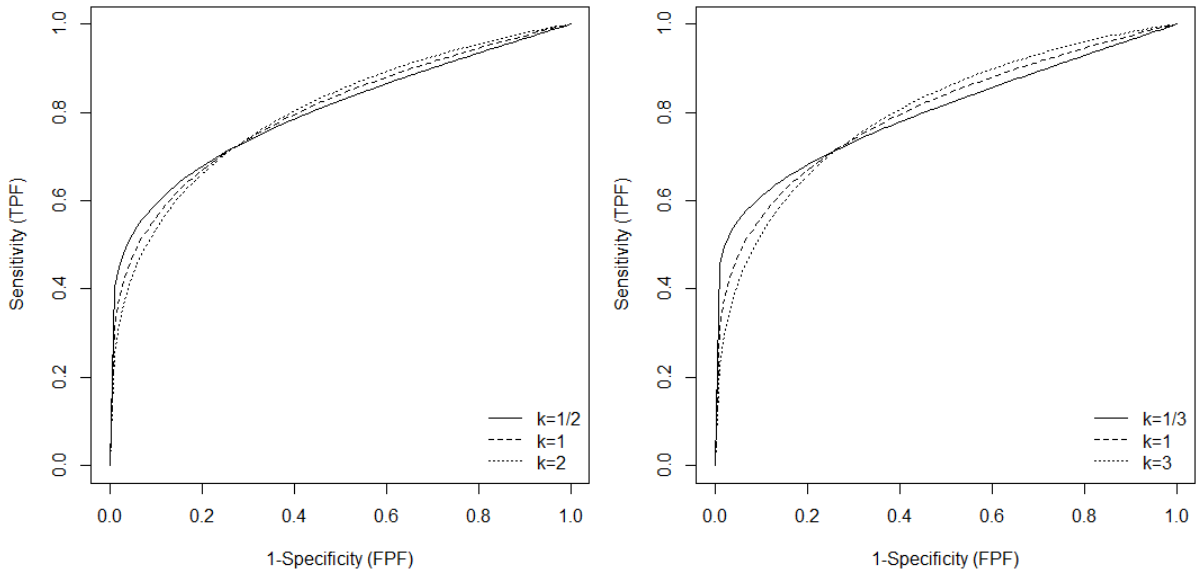


Figure 4.3 Bi-gamma ROC curves with AUC=0.8

Since the two bi-gamma ROC curves have the same shape parameter κ, they do not cross each other. Thus the difference in the pAUCs ($A_e^2 - A_e^1$) increases with increasing range. $A_e^2 - A_e^1$ reaches a maximum value (equal to the difference in full AUCs) at e=1.

For this scenario, with increasing range of interest the difference in standardized pAUCs $\tilde{A}_e^2 - \tilde{A}_e^1$ decreases for ROC curves with high AUC. Table 4.8 shows the differences in the spAUCs where the difference in the full AUCs between the two ROC curves is remains 0.05. The difference in the spAUCs decreases with increasing range of interest for ROC curves with high AUCs (e.g., AUC of 0.8 and 0.9 for bi-gamma ROC curve with $\kappa \geq 1$, and AUC of 0.7, 0.8 and 0.9 for bi-gamma ROC curve with $\kappa < 1$). This agrees with proposition 4.1, since in the proximity of the point (1,1) concave bi-gamma ROC curves with large AUC tend to have a small slope, thereby leading to a small difference in DLR$^-$'s, and eventually to a decreasing difference in the spAUCs.

Table 4.8 Theoretical $\tilde{A}_e^2 - \tilde{A}_e^1$ of two bi-gamma ROC curves with differences in full AUCs of 0.05

Parameters for	Ranges of false positive fractions				
lower ROC curve	0-0.2	0-0.4	0-0.6	0-0.8	0-1
k=3					
AUC=0.60	0.033	0.042	0.046	0.049	0.050
AUC=0.70	0.045	0.049	0.050	0.050	0.050
AUC=0.80	0.058	0.055	0.053	0.051	0.050
AUC=0.90	0.074	0.062	0.055	0.051	0.050
k=2					
AUC=0.60	0.034	0.043	0.047	0.049	0.050
AUC=0.70	0.046	0.049	0.050	0.050	0.050
AUC=0.80	0.058	0.055	0.052	0.051	0.050
AUC=0.90	0.072	0.060	0.054	0.051	0.050
k=1					
AUC=0.60	0.038	0.044	0.048	0.049	0.050
AUC=0.70	0.049	0.050	0.050	0.050	0.050
AUC=0.80	0.058	0.054	0.052	0.050	0.050
AUC=0.90	0.067	0.058	0.053	0.051	0.050
k=1/2					
AUC=0.60	0.042	0.047	0.049	0.050	0.050
AUC=0.70	0.051	0.051	0.050	0.050	0.050
AUC=0.80	0.057	0.053	0.051	0.050	0.050
AUC=0.90	0.060	0.055	0.052	0.050	0.050
k=1/3					
AUC=0.60	0.045	0.048	0.049	0.050	0.050
AUC=0.70	0.052	0.051	0.050	0.050	0.050
AUC=0.80	0.056	0.052	0.051	0.050	0.050
AUC=0.90	0.057	0.053	0.051	0.050	0.050

Table 4.9 Variance of empirical spAUC difference for two non-crossing concave bi-gamma ROC curves with differences in full AUCs of 0.05

Parameters for lower ROC curve	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
k=3					
AUC=0.60	6.906	8.680	9.833	10.398	10.696
AUC=0.70	8.712	8.891	8.989	9.010	8.980
AUC=0.80	**9.072**	**7.690**	**7.057**	**6.821**	**6.717**
AUC=0.90	5.623	4.160	3.509	3.198	3.081
k=2					
AUC=0.60	6.805	8.332	9.319	10.076	10.371
AUC=0.70	8.758	9.005	9.279	9.305	9.291
AUC=0.80	**9.288**	**7.759**	**7.323**	**7.078**	**7.007**
AUC=0.90	4.995	3.870	3.487	3.274	3.214
k=1					
AUC=0.60	6.826	8.297	9.476	10.389	10.765
AUC=0.70	8.297	8.670	9.017	9.345	9.451
AUC=0.80	**7.922**	**7.159**	**6.880**	**6.732**	**6.704**
AUC=0.90	4.362	3.665	3.540	3.550	3.587
k=1/2					
AUC=0.60	6.360	7.579	8.844	9.936	10.525
AUC=0.70	7.963	8.207	8.680	9.269	9.568
AUC=0.80	**6.736**	**6.616**	**6.662**	**6.857**	**7.057**
AUC=0.90	3.831	3.654	3.720	3.870	3.934
k=1/3					
AUC=0.60	6.441	7.641	8.912	10.237	10.799
AUC=0.70	6.829	7.422	8.219	8.874	9.156
AUC=0.80	**6.167**	**6.491**	**6.957**	**7.467**	**7.685**
AUC=0.90	3.414	3.443	3.620	3.852	3.952

*Data consisted of pairs of ratings for 150 normal and 150 abnormal subjects, with between-modality correlation of 0.5; 1000 datasets were simulated for evaluating the variance of difference in empirical spAUCs; bold-faced scenarios correspond to curves in Figure 4.3.

Each simulated dataset consisted of correlated pairs of ratings generated from a gamma distribution. Due to the invariance property of the ROC curves, without any loss of generality, we set $\theta=1$ for latent ratings of abnormal subjects. We then selected θ for the latent normal ratings to reflect the targeted area under the ROC curve (given κ of 2, 1, or ½). The between-modality correlation of 0.5 was established using a Gaussian copula model (Nelsen, 1999).

Table 4.10 summarizes the statistical power for comparisons of pAUCs of two bi-gamma ROC curves with a difference in full AUCs of 0.05. The results show that statistical power frequently increases with increasing range for $\kappa \geq 1$, but always decreases with increasing range for $\kappa \leq ½$. Even for $\kappa \geq 1$ the decreasing trend in statistical power can be observed for ROC curves with high AUCs, but with increasing κ (i.e., higher curvature) the use of the full AUC becomes increasing more beneficial (statistically more powerful). For example, for scenarios with $\kappa=1$ the statistical power increases with increasing range of interest when AUC<0.8, whereas, for $\kappa=2$ (or higher curvature) the increasing pattern is observed for most scenarios, except for AUC of 0.9.

Table 4.10 Statistical power for comparisons of two partial AUCs of concave non-crossing bi-gamma ROC type curves with differences in full AUCs of 0.05

Parameters for lower ROC curve	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
k=3					
AUC=0.60	0.242	0.300	0.309	0.328	0.340
AUC=0.70	0.302	0.364	0.380	0.386	0.382
AUC=0.80	**0.482**	**0.515**	**0.526**	**0.528**	**0.537**
AUC=0.90	0.882	0.884	0.873	0.875	0.865
k=2					
AUC=0.60	0.267	0.306	0.309	0.323	0.339
AUC=0.70	0.357	0.400	0.412	0.418	0.415
AUC=0.80	**0.488**	**0.525**	**0.528**	**0.535**	**0.519**
AUC=0.90	0.903	0.887	0.859	0.830	0.822
k=1					
AUC=0.60	0.295	0.327	0.333	0.324	0.323
AUC=0.70	0.370	0.418	0.418	0.408	0.401
AUC=0.80	**0.561**	**0.582**	**0.555**	**0.521**	**0.526**
AUC=0.90	0.900	0.867	0.829	0.810	0.806
k=1/2					
AUC=0.60	0.359	0.374	0.356	0.345	0.332
AUC=0.70	0.463	0.454	0.413	0.391	0.384
AUC=0.80	**0.616**	**0.562**	**0.525**	**0.496**	**0.491**
AUC=0.90	0.887	0.837	0.792	0.768	0.745
k=1/3					
AUC=0.60	0.419	0.415	0.368	0.335	0.322
AUC=0.70	0.494	0.460	0.433	0.407	0.395
AUC=0.80	**0.596**	**0.553**	**0.525**	**0.485**	**0.461**
AUC=0.90	0.849	0.808	0.761	0.741	0.736

*Data consisted of pairs of ratings for 150 normal and 150 abnormal subjects, with between-modality correlation of 0.5; 1000 datasets were simulated and 1000 results of the bootstrap tests for testing the null hypothesis $\tilde{A}_e^2 - \tilde{A}_e^1 = 0$ were performed; bold-faced scenarios correspond to curves in Figure 4.3.

It is interesting to note that despite the substantial discrepancy in observed trends among bi-gamma curves with different κ's, visually they may not look very different. Figure 4.3 illustrates three bi-gamma curves with AUC of 0.8 and κ=1/3, 1, and 3, respectively; the corresponding trends in statistical power are presented in bold-face in Table 5.10.

We developed a program (Appendix D) to compute sample size for comparisons of two partial AUCs of bi-gamma ROC curves, the result were shown in Table 4.11. For κ<1, sample

size requirements increased with increasing range; for κ=1, sample size requirements increased with increasing range only for AUC≥0.8; for κ>1, sample size requirements increased with increasing range only for AUC≥0.9.

Table 4.11 Sample size requirements for comparisons of two partial AUCs of bi-gamma ROC type curves with differences in full AUCs of 0.05

Parameters for	Ranges of false positive fractions				
lower ROC curve	0-0.2	0-0.4	0-0.6	0-0.8	0-1
k=3					
AUC=0.60	730	581	538	514	505
AUC=0.70	549	450	429	419	419
AUC=0.80	324	290	283	289	293
AUC=0.90	119	122	132	140	142
k=2					
AUC=0.60	669	553	525	509	505
AUC=0.70	508	433	421	418	420
AUC=0.80	303	282	283	292	297
AUC=0.90	115	123	135	145	149
k=1					
AUC=0.60	558	501	500	504	508
AUC=0.70	432	398	407	417	425
AUC=0.80	265	270	284	300	309
AUC=0.90	113	129	146	159	164
k=1/2					
AUC=0.60	446	443	472	498	511
AUC=0.70	357	364	394	419	433
AUC=0.80	237	263	290	313	326
AUC=0.90	120	141	161	176	183
k=1/3					
AUC=0.60	389	414	460	496	515
AUC=0.70	325	351	391	423	440
AUC=0.80	232	267	298	324	337
AUC=0.90	127	150	170	185	192

*Data consisted of pairs of ratings for 150 normal and 150 abnormal subjects, with between-modality correlation of 0.5; 1000 datasets were simulated for evaluating the variance of difference in empirical spAUCs.

Full versus partial AUC

The three families of ROC curves we investigated lead to different trends in the "usefulness" of the partial AUC as compared with the inferences based on the full AUC. In

particular, for concave binormal ROC curves, comparisons of full AUCs leads to a higher statistical power than comparisons of partial AUCs over any range. Conversely, within the family of straight-line ROC curves comparisons of full AUCs always have smaller statistical power than comparisons of partial AUCs. The family of concave bi-gamma ROC curves could favor either the full or the partial AUC (in terms of statistical power) depending on the shape parameter κ (nearly straight-line ROC curves for $\kappa<1$, and binormal-looking ROC curve for $\kappa>1$).

In practice, bi-gamma and binormal ROC curves may look similar; however, the sample size requirement for the AUC and the pAUC could be quite different. Figure 4.4 illustrates binormal (b=1), bi-gamma (κ=1), and straight-line ROC curves with a full AUC of 0.8. For comparisons of these curves to the curves of the same shape but with a true AUC of 0.85, in order to achieve the same power as that computed for $pAUC_{(0, 0.2)}$ (for 150 diseased and 150 non-diseased subjects as shown in table 3), using the full AUC we would need 88 diseased subjects for the concave binormal ROC curve, 163 diseased subjects for the bi-gamma ROC curve and 204 diseased subjects for the straight-line ROC curve.

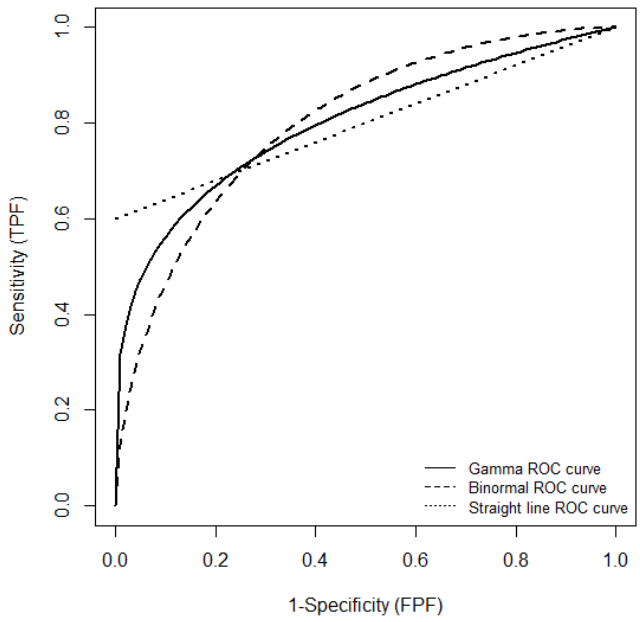


Figure 4.4 Binormal ROC curve (b=1), Bi-gamma ROC curve (κ=1) and a straight-line ROC curve with AUC=0.8

Table 4.12 summarizes sample size requirements for comparisons of full AUCs to achieve the same power as comparisons of pAUCs $_{(0,\ 0.2)}$ for the same ROC curves estimated based on 150 diseased and 150 non-diseased subjects. We observe that in agreement with our findings in scenarios 1-3, for concave binormal ROC curves, improper binormal ROC curves with an AUC less than or equal to 0.8, and bi-gamma ROC curve with $\kappa>1$ and an AUC less than or equal to 0.8, using the full AUC leads to smaller sample size requirements as compared with the requirements for using the pAUC over (0, 0.2). In contrast, for other scenarios, using the pAUC requires a smaller sample size.

Table 4.12 Sample size requirements for inferences based on full AUC to achieve the same power as comparison of pAUC (0, 0.2) shown in tables 2-4

family	Shape parameter of ROC curves	AUC for lower ROC curve			
		0.6	0.7	0.8	0.9
Binormal	b=0.33	107	123	133	163
	b=0.50	107	101	128	148
	b=1.00	69	89	**88**	115
Straight-line		261	222	**204**	187
Bi-gamma	κ=3	99	113	131	158
	κ=2	113	125	139	192
	κ=1	135	136	**163**	198
	κ=1/2	165	189	203	220
	κ=1/3	206	198	210	200

*Based on 150 diseased and 150 non-diseased subjects; shaded cells indicate scenarios where use of partial AUC is preferable over full AUC; bold-faced results correspond to scenarios with ROC curves of shape shown in Figure 2

4.3 EXAMPLES

In this example we provide analysis of a small dataset for comparing accuracy of two diagnostic modalities evaluated using 50 diseased and 50 non-diseased subjects. We simulated diagnostic ratings from bi-gamma distributions with a correlation of 0.5 for diseased subjects and 50 non-diseased subjects, respectively. We estimated empirical ROC curves by connecting empirical points with straight lines. The estimates of the standardized pAUC were computed by integration over the ranges starting at 0 and ending at 0.2, 0.4, 0.6, 0.8 and 1. Variances of the differences in estimated standardized pAUCs were estimated using non-parametric bootstrap approach (Efron and Tibshirani, 1993). The bootstrap percentile confidence intervals, and corresponding p-values were computed using 10,000 random bootstrap samples.

Table 4.13 summarizes the differences in standardized pAUCs, their bootstrap variances, and the 95% bootstrap confidence intervals. In this example, the differences in standardized pAUCs decreased with increasing range, while the variances remained relatively stable across all

ranges of interest. As illustrated in Figure 4.5, the two ROC curves do not cross. The difference in the full AUCs was not statistically significant (p=0.118) while the difference in partial AUCs over the range (0, 0.2) was statistically significant (p=0.041).

Table 4.13 Results for comparisons of correlated ROC curves presented in example #1.

	Ranges of false positive fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
Difference in spAUCs	0.1122	0.0988	0.0867	0.0788	0.0784
Bootstrap CI					
2.5% percentiles	0.0056	-0.0056	-0.0152	-0.0200	-0.0196
97.5% percentiles	0.2100	0.2031	0.1876	0.1800	0.1792
Bootstrap variance	0.0027	0.0028	0.0027	0.0026	0.0026
p-value	0.041	0.062	0.100	0.117	0.118

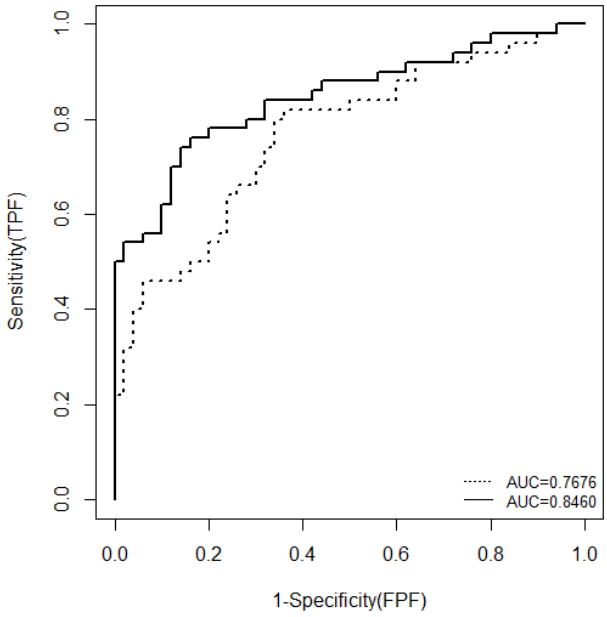


Figure 4.5 Empirical estimates of correlated ROC curve from example #1

It is important to highlight that the sample sizes estimated for the bi-gamma family of the ROC curves are different than estimates obtained from a standard approach assuming a binormal ROC model (Obuchowski and McClish, 1997). In particular, under the binormal model, for 80% statistical power in comparisons of concave ROC curves with areas 0.8 and 0.85 one would need sample sizes of 452 for the pAUC over (0, 0.2) and 295 for the full AUC, while in the case of the improper ROC curve with b=1/3 the required sample sizes would be 424 and 407. In fact, we were not able to find the scenarios under which the sample size estimation for concave binormal ROC curves would favor inferences based on the pAUC. In contrast, assuming bi-gamma model with $\kappa=1/3$, the sample sizes for the partial and the full AUC would be 232 and 337 respectively. This suggests that there may be advantages to using the pAUC in practical scenarios where the underlying distributions of ratings are reasonable but not necessarily binormal.

4.4 SUMMARY

In some practical scenarios comparison of two ROC curves based on the partial AUC could be both more clinically relevant and more statistically conclusive than inference based on full AUC.

Increase of the range of interest could lead to either an increase or decrease in difference between two partial areas. And effect of the increasing range on the sampling variability depends on the shape of the performance curve.

For ROC curves with nearly straight-line shape comparisons based on shorter range of partial AUC are always preferable. At the same time, the comparison of two correlated binormal ROC curves can be more efficiently performed using full AUC. Thus, approaches for sample

size estimation based on binormal ROC models often mask statistical the possible advantages of using partial AUCs.

We demonstrated that family of constant shape bi-gamma ROC curves allows more realistic and flexible reflection of properties of pAUC analysis. Bi-gamma family of ROC curves provides better coverage of practically reasonable and plausible shapes. It can accommodate ROC curves that are close to the straight line, as well as ROC curves that are similar to the binormal ROC curves. The developed R program allows estimating sample size for comparison based on pAUCs for the bi-gamma ROC curves with different values of the shape parameter.

5.0 PARTIAL AREA UNDER THE ROC CURVE WITH MASS

Diagnostic test results (ratings) often have ties in particular in the region of low rating levels. These ties could results from various phenomena including absence of apparent signs of disease in a subsample of subjects (including some actually diseased), natural absence of a tested substance, or artificial assignment of a default value to subjects with biomarker levels below a predetermined threshold or below the limit of detection. The corresponding ROC curves have straight-line shape (with no deterministic operating points) in the regions with low specificity, and sometimes called ROC curve with mass (at '0'). ROC curves with mass can be constructed from any given ROC curve by replacing the right-most part with a straight-line segment (or equivalently by grouping data below a certain threshold).

In this chapter we investigate statistical properties of evaluation of a single diagnostic test as well as a comparison of performance levels of two diagnostic modalities using pAUC over different ranges for ROC curves with mass. We demonstrate that due to virtual absence of empirical points in the ranges with low specificity, the selection of wider range leads to increasing power for ROC curves with mass obtained from originally concave binormal ROC curves and decreasing power for ROC curves with mass obtained from the originally straight-line ROC curves. However, the increasing or decreasing trend tends to gradually disappear after the point where mass occurs, and thus the statistical power becomes stable. For comparison of

two full AUC of the ROC curves with nearly straight-line shape, statistical power is higher for ROC curves with mass than that for curves without mass.

Thus, as similar as the regular ROC curve, the statistical power for ROC curve with mass, and thereby sample size requirement for inferences based on pAUC are affected by the shape of the performance curves. The presence of "mass" (i.e., grouping diagnostic results below certain level) can alleviate the decrease in variability, but it can disturb the estimated accuracy levels if the grouped results are informative. However, if the diagnostic results below a certain threshold have little information, grouping could be beneficial.

5.1 METHOD

In evaluation of a single pAUC, as a direct application of proposition 2 (Ma *et al.*, 2013), since ROC curves having a mass does not change concavity, the standardized pAUC increases with increasing range for concave binormal ROC curves, whether these have mass, or not.

In comparison of two correlated pAUCs, we presented previously two conditions that determine whether the difference in standardized pAUCs increases or decreases in the proximity of the FPF of interest. Here we can demonstrate that, for all types of ROC curves with mass, the increasing or decreasing trend of the difference in standardized pAUCs beyond the point where mass occurs will remain the same as the difference up to the point at which mass occurs. In other words, if the difference in standardized pAUCs increases or decreases in the proximity of the FPF where mass occurs, then the difference in standardized pAUCs keeps increasing or decreasing beyond that point.

5.2 NUMERICAL STUDY

In this section we consider ROC family of curves under the distribution assumptions of normality and uniformity for underlying continuous test results. For each scenario, i.e. binormal ROC curves and straight-line ROC curves, we investigate statistical inferences based on pAUC and AUC for conventional ROC curves, partial ROC curve with mass at FPF equal 0.5, and partial ROC curve with mass at FPF equal 0.2. The partial ROC curves with mass have exactly the same shapes as the conventional curves in the range before the mass occurs. For example, Figure 5.1 to 5.3 show the concave binormal ROC curves, partially concave binormal ROC curves with mass at FPF equal 0.5, and partially concave binormal ROC curves with mass at FPF equal 0.2 respectively, where the full range curves have AUC ranging from 0.65 to 0.95.

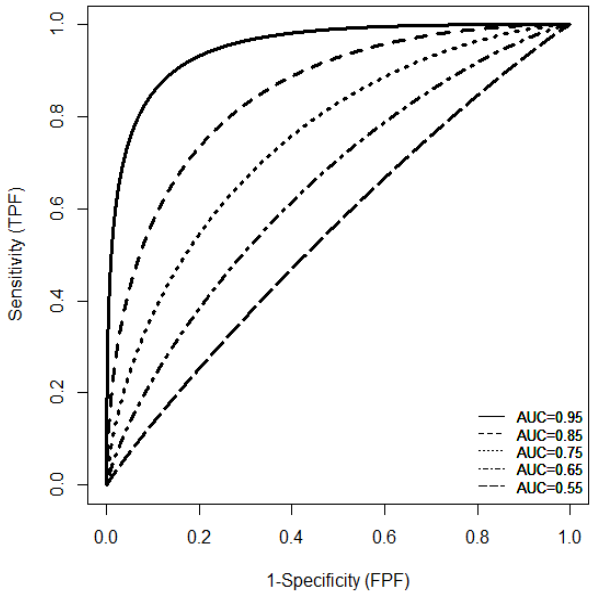


Figure 5.1 Concave binormal ROC curves

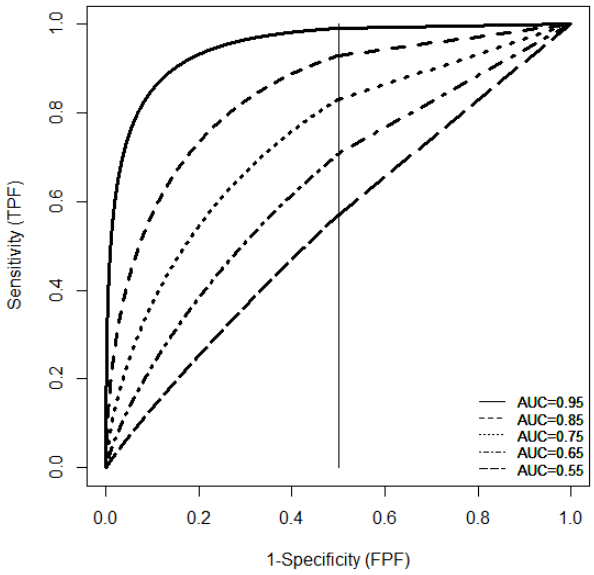


Figure 5.2 Partial concave binormal ROC curves with mass at FPF equal 0.5

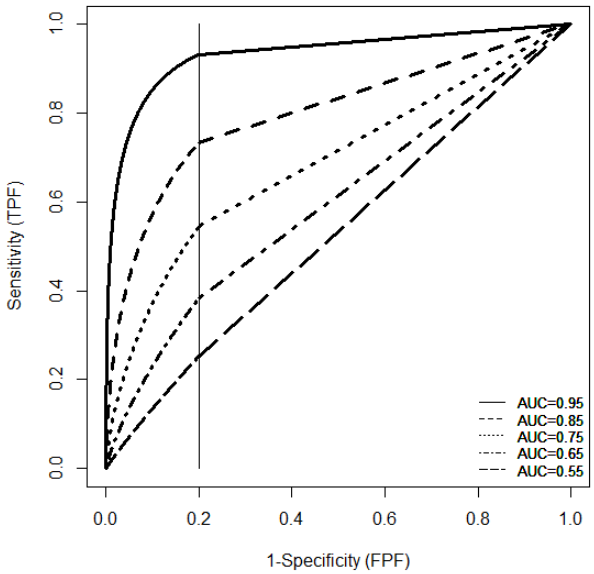


Figure 5.3 Partial concave binormal ROC curves with mass at FPF equal 0.2

5.2.1 EVALUATION OF A SINGLE PAUC

1. Standardized pAUC

We previously showed that the standardized pAUC increases with increasing range for concave binormal ROC curves and partially concave binormal ROC curves with mass, and it remains constant for straight-line ROC curves and partially straight-line ROC curves with mass.

Table 5.1 shows that standardized pAUC increases with increasing range. However, the linear segment on ROC curves with mass results in a smaller increasing trend for the standardized pAUC, namely, the standardized pAUC of partially concave binormal ROC curves with mass tends to be smaller than the concave binormal ROC curves beyond the point where mass occurs. This trend contributes to the almost constant standardized pAUC for partially concave binormal ROC curves.

Table 5.1 Theoretical standardized pAUC for concave binormal ROC curves and corresponding partial binormal ROC curves with mass

Concave binormal ROC curves	Ranges of False Positive Fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
Regular ROC curves					
auc=0.55	0.5180	0.5300	0.5395	0.5466	0.5500
auc=0.65	0.5667	0.6016	0.6263	0.6430	0.6500
auc=0.75	0.6373	0.6902	0.7228	0.7426	0.7500
auc=0.85	0.7390	0.7974	0.8281	0.8445	0.8500
auc=0.95	0.8899	0.9260	0.9411	0.9480	0.9500
Mass at FPF=0.5					
auc=0.55	0.5180	0.5300	0.5388	0.5428	0.5439
auc=0.65	0.5667	0.6016	0.6249	0.6351	0.6380
auc=0.75	0.6373	0.6902	0.7214	0.7349	0.7387
auc=0.85	0.7390	0.7974	0.8272	0.8399	0.8434
auc=0.95	0.8899	0.9260	0.9409	0.9470	0.9487
Mass at FPF=0.2					
auc=0.55	0.5180	0.5247	0.5268	0.5276	0.5278
auc=0.65	0.5667	0.5877	0.5941	0.5966	0.5974
auc=0.75	0.6373	0.6715	0.6819	0.6861	0.6873
auc=0.85	0.7390	0.7804	0.7930	0.7981	0.7995
auc=0.95	0.8899	0.9192	0.9282	0.9318	0.9328

2. Variance of standardized pAUC

We conducted a simulation study to assess variance of standardized pAUC for binormal and straight-line ROC curves and the corresponding ROC curves with mass. In the simulation study for the binormal model data were generated from normal distributions with equal variance and parameters selected to generate binormal ROC curves with AUC ranging from 0.55 to 0.95. For the straight-line ROC curve the test results for normal and abnormal subjects were generated from uniform distributions. To generate ROC curves with mass, we replaced all ratings below the predetermined threshold corresponding to the FPF where mass occurs by the ratings at that threshold. For each scenario, we generated 1000 datasets with ratings for 150 normal and 150

abnormal subjects. pAUCs were estimated using area under the linearly-interpolated empirical ROC curve over the range of interest. The methods for constructing ROC curves with mass and the estimation method for pAUC were the same throughout this section.

For concave ROC curves as well as concave ROC curves with mass, the variance trend can exhibit different patterns, namely, it can either decrease or increase with increasing range. The decrease in variance with increasing range can be observed for ROC curves with AUC values greater than 0.75. This is a similar trend for ROC curves without mass. However, the decreasing/increasing trend tends to be smaller beyond the point where mass occurs. Thus, for the ROC curve originally having increasing variance, the variance of full AUC tends to be smaller for partial concave ROC curves with mass than the corresponding concave ROC curves without mass, and vice versa.

For straight-line ROC curves as well as straight-line ROC curves with mass, the variance of standardized pAUC increases with increasing range. This is a similar trend for straight-line ROC curves without mass. However, the increasing trend in variance, diminishes beyond the point where mass occurs. This leads to a smaller variance of full AUC for straight-line ROC curves with mass as compared with the straight-line ROC curves without mass. (Table 5.2)

Table 5.2 Variance of standardized pAUC for concave binormal and straight-line ROC curves and corresponding partial ROC curves with mass ($\times 10^{-4}$)

	Ranges of False Positive Fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
Regular binormal curves					
auc=0.55	3.312	5.592	7.836	9.706	10.607
auc=0.65	5.937	8.235	9.498	10.104	10.202
auc=0.75	8.254	8.548	8.402	7.995	7.750
auc=0.85	10.072	7.290	5.877	5.099	4.801
auc=0.95	5.161	2.631	1.748	1.385	1.281
Partial binormal curves with mass at FPF=0.5					
auc=0.55	3.312	5.592	7.649	9.022	9.476
auc=0.65	5.937	8.234	9.171	9.633	9.808
auc=0.75	8.254	8.546	8.178	8.009	7.995
auc=0.85	10.072	7.288	5.804	5.222	5.083
auc=0.95	5.161	2.631	1.758	1.445	1.366
Partial binormal curves with mass at FPF=0.2					
auc=0.55	3.228	4.700	5.376	5.677	5.764
auc=0.65	5.699	6.451	6.912	7.133	7.199
auc=0.75	7.894	6.867	6.790	6.806	6.816
auc=0.85	9.763	6.614	6.042	5.880	5.842
auc=0.95	5.087	2.792	2.348	2.211	2.176
Regular straight-line ROC curves					
auc=0.55	3.718	5.492	7.423	9.373	10.391
auc=0.65	5.070	6.542	8.128	9.591	10.206
auc=0.75	5.180	6.141	7.080	8.046	8.595
auc=0.85	3.946	4.369	4.943	5.386	5.607
auc=0.95	1.576	1.694	1.844	1.964	2.028
Partial straight-line curves with mass at FPF=0.5					
auc=0.55	3.718	5.492	7.371	8.768	9.227
auc=0.65	5.070	6.542	8.043	9.063	9.395
auc=0.75	5.180	6.141	7.039	7.716	7.938
auc=0.85	3.946	4.369	4.918	5.282	5.401
auc=0.95	1.576	1.694	1.847	1.952	1.986
Partial straight-line curves with mass at FPF=0.2					
auc=0.55	3.713	5.104	5.738	6.020	6.101
auc=0.65	5.064	6.217	6.729	6.955	7.020

Table 5.2 (continued)

auc=0.75	5.194	5.860	6.174	6.314	6.354
auc=0.85	3.949	4.297	4.459	4.531	4.552
auc=0.95	1.579	1.672	1.715	1.734	1.739

* Data consisted of ratings for 150 normal and 150 abnormal subjects; 1000 datasets were simulated for evaluating the variance of empirical spAUCs.

3. Statistical power in a single modality

We investigated the statistical power for tests based on pAUC and AUC in a one-sample problem for binormal and straight-line ROC curves and the corresponding ROC curves with mass. The statistical test of the null hypothesis for standardized pAUC equal 0.5 versus the alternative hypothesis for standardized pAUC greater than 0.5 was performed using a nonparametric bootstrap approach based on 1000 resamples of 50 normal and 50 abnormal subjects. Statistical power was estimated from 1000 results of the bootstrap results.

Table 5.3 shows that in the case of concave binormal ROC curves as well as partially concave binormal ROC curves with mass, the statistical power in a one-sample problem always increases with increasing range. The statistical power is similar for concave full binormal ROC curves and the corresponding partially concave binormal ROC curves with mass at 0.5. The statistical power for partially concave binormal ROC curves with mass at 0.2 remains nearly a constant after FPF=0.4, and is smaller than the statistical power for the corresponding concave full binormal ROC curves. The smaller statistical power for ROC curves with mass results primarily from the smaller standardized pAUC as we had demonstrated when evaluating a single pAUC.

For straight-line ROC curves as well as for straight-line ROC curves with mass, the statistical power decreases with increasing range. The decreasing trend diminished after the point where mass occurs. This results in a higher statistical power for testing pAUC over wider ranges for straight-line ROC curves with mass as compared with full range straight-line ROC curves.

Table 5.3 Statistical power for concave binormal and straight-line ROC curves and corresponding partial ROC curves with mass

	Ranges of False Positive Fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
Regular binormal curves					
auc=0.55	0.089	0.096	0.106	0.120	0.129
auc=0.65	0.370	0.494	0.594	0.641	0.686
auc=0.75	0.789	0.926	0.969	0.982	0.992
auc=0.85	0.989	0.999	1.000	1.000	1.000
auc=0.95	1.000	1.000	1.000	1.000	1.000
Partial binormal curves with mass at FPF=0.5					
auc=0.55	0.093	0.101	0.123	0.123	0.134
auc=0.65	0.396	0.497	0.600	0.646	0.694
auc=0.75	0.815	0.926	0.969	0.983	0.991
auc=0.85	0.982	0.997	0.999	1.000	1.000
auc=0.95	1.000	1.000	1.000	1.000	1.000
Partial binormal curves with mass at FPF=0.2					
auc=0.55	0.091	0.091	0.093	0.091	0.092
auc=0.65	0.387	0.484	0.540	0.563	0.568
auc=0.75	0.780	0.914	0.964	0.978	0.979
auc=0.85	0.984	0.999	1.000	1.000	1.000
auc=0.95	1.000	1.000	1.000	1.000	1.000
Regular straight-line ROC curves					
auc=0.55	0.409	0.252	0.182	0.158	0.128
auc=0.65	0.997	0.970	0.895	0.792	0.727
auc=0.75	1.000	1.000	1.000	0.996	0.989
auc=0.85	1.000	1.000	1.000	1.000	1.000
auc=0.95	1.000	1.000	1.000	1.000	1.000
Partial straight-line curves with mass at FPF=0.5					
auc=0.55	0.409	0.252	0.183	0.154	0.152
auc=0.65	0.997	0.970	0.894	0.819	0.785
auc=0.75	1.000	1.000	1.000	0.997	0.995
auc=0.85	1.000	1.000	1.000	1.000	1.000
auc=0.95	1.000	1.000	1.000	1.000	1.000
Partial straight-line curves with mass at FPF=0.2					
auc=0.55	0.407	0.275	0.237	0.224	0.217
auc=0.65	0.998	0.978	0.936	0.922	0.913

Table 5.3 (continued)

auc=0.75	1.000	1.000	0.999	0.999	0.999
auc=0.85	1.000	1.000	1.000	1.000	1.000
auc=0.95	1.000	1.000	1.000	1.000	1.000

*Data consisted of ratings for 150 normal and 150 abnormal subjects; 1000 datasets were simulated and 1000 results of the bootstrap tests for testing the null hypothesis spAUC=0.5 were performed.

5.2.2 COMPARISON OF CORRELATED PAUC

1. Difference in standardized pAUCs

We first investigated the properties for comparisons of pAUCs (A_e^1 and A_e^2) for two concave binormal ROC curves and the corresponding partially concave binormal ROC curves with mass. We considered pairs of concave binormal ROC curves that are were constrained to have a constant difference of 0.05 between full AUCs. The corresponding partially concave binormal ROC curves with mass have exactly the same shape as the full range curves throughout the range before the mass occurs.

We previously proved that in general, for ROC curves with mass, if the difference in standardized pAUCs increases or decreases in the proximity of the FPF where mass occurs, the difference in standardized pAUCs keeps increasing or decreasing after that point, as well. For straight-line ROC curves and straight-line ROC curves with mass, the difference in standardized pAUCs remains constant across the entire range from 0 to 1.

Table 5.4 shows the differences between the standardized pAUCs when the difference in the full range AUCs of the two concave binormal ROC curves is 0.05. We show that for larger AUCs (namely, average of 0.825 and 0.925) the difference in standardized pAUCs is greater for ROC curves with mass as compared with the corresponding full ROC curves beyond the point where mass occurs. In addition, for lower AUCs (namely, average of 0.725 or 0.625) the

difference in spAUCs tends to be smaller for ROC curve with mass as compared with the corresponding full range ROC curves beyond the point where mass occurs.

Table 5.4 Theoretical difference in standardized pAUCs for comparisons of two concave binormal ROC curves and comparisons of corresponding partial binormal ROC curves with mass

Average AUC	Ranges of False Positive Fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
Regular ROC curves					
auc=0.625	0.0267	0.0378	0.0447	0.0487	0.0500
auc=0.725	0.0385	0.0465	0.0494	0.0502	0.0500
auc=0.825	0.0555	0.0561	0.0537	0.0512	0.0500
auc=0.925	0.0835	0.0672	0.0574	0.0519	0.0500
Mass at FPF=0.5					
auc=0.625	0.0267	0.0378	0.0444	0.0472	0.0480
auc=0.725	0.0385	0.0465	0.0495	0.0507	0.0510
auc=0.825	0.0555	0.0561	0.0540	0.0529	0.0527
auc=0.925	0.0835	0.0672	0.0577	0.0536	0.0525
Mass at FPF=0.2					
auc=0.625	0.0267	0.0338	0.0360	0.0369	0.0372
auc=0.725	0.0385	0.0448	0.0467	0.0474	0.0477
auc=0.825	0.0555	0.0579	0.0586	0.0589	0.0590
auc=0.925	0.0835	0.0735	0.0704	0.0692	0.0688

2. Variance of the difference in standardized pAUCs

We computed the variance of the difference in standardized pAUCs for two concave binormal curves and the corresponding partially concave binormal ROC curves with mass. The paired ROC curves we compared have a constant difference of 0.05 between the full AUCs. In the simulation study, for binormal model the test results for normal and abnormal subjects were generated from bivariate normal distributions with correlation of 0.5. Exploiting the invariance property of the ROC curve to monotonically increasing transformation of the ratings, the distributions of ratings for normal subjects were set to bivariate normal distribution with mean

$\mu=(0,0)^T$, and a covariance matrix $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. Parameters for the distributions of ratings for abnormal subjects were selected to reflect areas under the curves while preserving a correlation of 0.5 between ratings corresponding to the same subjects. Each generated dataset consisted of ratings for 150 normal (X_{i1}, X_{i2}) and 150 abnormal subjects (Y_{j1}, Y_{j2}) where i, j=1,2…,150. To generate ROC curves with mass, we replaced the ratings below the threshold corresponding to the FPF where mass occurs by the values at that threshold. For each scenario, we generated 1000 datasets with ratings for 150 normal and 150 abnormal subjects.

Table 5.5 shows that similar to one-sample problem, concave binormal ROC curves as well as the partially concave binormal ROC curves with mass exhibit variance trends that can either decrease or increase with increasing range. The decrease in variance with increasing range is observed for ROC curves with average AUC values greater than or equal to 0.825. In other words, considering binormal ROC curves with mass will not change the trend in variance of the difference in standardized pAUCs. However, the decreasing/increasing trend tends to diminish after the point where mass occurs. Thus, for a ROC curve that originally has an increasing variance, the variance of the full AUC tends to be smaller for partially concave ROC curves with mass than the corresponding concave ROC curve, and vice versa.

Table 5.5 Variance of difference in standardized pAUCs for concave binormal and corresponding partially concave ROC curves with mass ($\times 10^{-4}$)

Average AUC	Ranges of False Positive Fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
Regular binormal curves					
auc=0.625	6.206	8.826	9.902	10.466	10.622
auc=0.725	9.227	9.819	9.808	9.622	9.399
auc=0.825	12.309	9.341	7.675	6.687	6.309
auc=0.925	8.930	4.989	3.510	2.839	2.635
Partial binormal curves with mass at FPF=0.5					
auc=0.625	6.206	8.819	9.787	10.572	10.878
auc=0.725	9.227	9.814	9.604	9.707	9.801
auc=0.825	12.309	9.339	7.658	7.043	6.915
auc=0.925	8.930	4.987	3.528	3.012	2.888
Partial binormal curves with mass at FPF=0.2					
auc=0.625	5.917	7.203	7.965	8.328	8.436
auc=0.725	8.758	8.325	8.645	8.855	8.922
auc=0.825	11.830	8.704	8.182	8.061	8.036
auc=0.925	8.785	5.287	4.639	4.453	4.409

*Data consisted of pairs of ratings for 150 normal and 150 abnormal subjects, with between-modality correlation of 0.5; 1000 datasets were simulated for evaluating the variance of difference in empirical spAUCs.

3. Statistical power in comparison of two partial AUCs

Using simulations we investigated the statistical power for comparisons of two concave binormal, straight-line, and the corresponding ROC curves with mass. The statistical test of the null hypothesis, or the difference in standardized pAUCs equal 0, was performed using nonparametric bootstrap approach based on 1000 resamples of 150 normal and 150 abnormal subjects with ratings generated for corresponding ROC curves with a difference between the full AUCs of 0.05. Statistical power was estimated from 1000 results of the bootstrap results.

Table 5.6 showed that for concave ROC curves as well as for partially concave ROC curves with mass, the statistical power for two sample comparisons never decreases with

increasing range. The statistical power for partially concave binormal ROC curves with mass at FPF value of 0.2 remains almost a constant after FPF value of 0.6. The statistical power for partially concave binormal ROC curves having average AUCs equal to wither 0.625 or 0.725, is smaller than that for the corresponding concave full binormal ROC curves. In contrast, the statistical power for partially concave binormal ROC curves having average AUCs equal to wither 0.825 or 0.925 is greater than that for concave full binormal ROC curves. However, this could be driven primarily by the difference in AUCs alone.

Thus, we estimated statistical power for comparisons of two partially concave ROC curves with mass with a constant actual difference in AUC of 0.05. Table 5.6 shows that for lower AUCs (namely, average of either 0.625 or 0.725), the statistical power in the family of partially concave binormal ROC curves is greater than the corresponding family of full binormal ROC curves. When AUC is larger (namely either average of 0.825 or 0.925), statistical power for comparisons is greater in the family of full binormal ROC curves.

Table 5.6 also showed that for straight-line ROC curves as well as for straight-line ROC curves with mass, the statistical power decreases with increasing range. The decreasing trend tends to diminish after the point where mass occurs. This leads to a higher statistical power for comparisons of straight-line ROC curves with mass as compared with conventional straight-line ROC curves.

Table 5.6 Statistical power for comparison of pAUCs within classes concave binormal ROC curves, straight-line ROC curves, and corresponding partial ROC curves with mass

Average AUC	Ranges of False Positive Fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
Regular binormal curves					
auc=0.625	0.179	0.248	0.285	0.304	0.313
auc=0.725	0.220	0.275	0.327	0.349	0.360
auc=0.825	0.335	0.445	0.475	0.499	0.513
auc=0.925	0.779	0.861	0.880	0.886	0.890
Partial binormal curves with mass at FPF=0.5					
auc=0.625	0.179	0.248	0.291	0.286	0.286
auc=0.725	0.220	0.275	0.327	0.343	0.344
auc=0.825	0.335	0.445	0.479	0.508	0.511
auc=0.925	0.779	0.861	0.884	0.889	0.889
Partial binormal curves with mass at FPF=0.2					
auc=0.625	0.186	0.243	0.252	0.253	0.258
auc=0.725	0.233	0.305	0.316	0.320	0.320
auc=0.825	0.354	0.506	0.550	0.564	0.564
auc=0.925	0.796	0.899	0.911	0.913	0.913
Regular straight-line ROC curves					
auc=0.625	0.490	0.418	0.363	0.322	0.322
auc=0.725	0.488	0.439	0.393	0.374	0.362
auc=0.825	0.543	0.511	0.478	0.443	0.432
auc=0.925	0.792	0.748	0.713	0.686	0.683
Partial straight-line curves with mass at FPF=0.5					
auc=0.625	0.490	0.418	0.372	0.328	0.320
auc=0.725	0.488	0.439	0.398	0.370	0.362
auc=0.825	0.543	0.510	0.480	0.461	0.448
auc=0.925	0.792	0.748	0.713	0.689	0.682
Partial straight-line curves with mass at FPF=0.2					
auc=0.625	0.497	0.417	0.385	0.379	0.374
auc=0.725	0.488	0.438	0.416	0.404	0.402
auc=0.825	0.546	0.512	0.497	0.492	0.488
auc=0.925	0.789	0.760	0.752	0.738	0.736

Table 5.6 (continued)

*Data consisted of pairs of ratings for 150 normal and 150 abnormal subjects, with between-modality correlation of 0.5; 1000 datasets were simulated and 1000 results of the bootstrap tests for testing the null hypothesis $\tilde{A}_e^2 - \tilde{A}_e^1 = 0$ were performed.

Table 5.7 Statistical power for concave binormal ROC curves and corresponding partial ROC curves with mass

(fixed AUC difference=0.05)

Average AUC	Ranges of False Positive Fractions				
	0-0.2	0-0.4	0-0.6	0-0.8	0-1
Regular binormal curves without mass					
auc=0.625	0.185	0.260	0.318	0.347	0.363
auc=0.725	0.226	0.317	0.364	0.387	0.394
auc=0.825	0.366	0.454	0.491	0.504	0.505
auc=0.925	0.799	0.871	0.882	0.891	0.889
Partial binormal curves with mass at FPF=0.5					
auc=0.625	0.202	0.275	0.339	0.349	0.350
auc=0.725	0.232	0.323	0.369	0.399	0.405
auc=0.825	0.353	0.428	0.473	0.487	0.485
auc=0.925	0.780	0.847	0.862	0.871	0.869
Partial binormal curves with mass at FPF=0.2					
auc=0.625	0.264	0.381	0.401	0.406	0.409
auc=0.725	0.255	0.380	0.404	0.407	0.404
auc=0.825	0.340	0.453	0.479	0.490	0.495
auc=0.925	0.696	0.814	0.831	0.831	0.832

*Data consisted of pairs of ratings for 150 normal and 150 abnormal subjects, with between-modality correlation of 0.5; 1000 datasets were simulated and 1000 results of the bootstrap tests for testing the null hypothesis $\tilde{A}_e^2 - \tilde{A}_e^1 = 0$ were performed.

5.3 SUMMARY

A substantial number of ties at the lowest diagnostic rating could affect the shape of the ROC curves.

Ties corresponding to grouping of informative diagnostic scores (corresponding to a concave part of the ROC curve) lead to a lower ROC curve with a straight line segment in the range of low specificities (ROC curve with mass). However, differences between the two ROC curves with mass could increase or decrease depending on the shape and height of the ROC curves.

The effect of the increasing range on the sampling variability depends on the shape of the ROC curve with a mass. Concave binormal ROC curves as well as the partially concave binormal ROC curves with mass exhibit variance trends that can either decrease or increase with increasing range. Variance trends always increase with increasing range for straight-line ROC curves as well as the partially straight-line ROC curves with mass.

One of the important implications of these results is that in some scenarios having ties at the lowest diagnostic rating can actually be beneficial for the assessment of performance. If the grouped results are barely informative (e.g., unobserved results below detection limit are similar for diseased and non-diseased), presence of ties would actually decrease the sampling variability of the estimated AUC. This concurs with previous findings that some "well-defined" tasks forcing diagnostic system to break the ties could detrimentally affect reliability and conclusiveness of statistical inferences (Gur *et al.*, 2007).

6.0 CONCLUSION AND DISCUSSION

The results of this work provide useful insights primarily for the design of the diagnostic performance studies; however, they also have important implications of the analysis of the studies. In particular, the results indicate that a conjecture about larger sample sizes requirements of pAUC over shorter range is frequently incorrect and should not discourage planning analysis based on pAUC, since the use of pAUC can lead to an actually more efficient analysis. Similarly, grouping of the diagnostic results could also in some cases be beneficial for the future statistical analysis. However, obtained results should not be directly used for selecting type of analysis after collecting the data since ad hoc alterations of the analysis can affect the error rate. Development of methods for controlling error rate in analysis where the range of interest, or range of grouping, should be selected is the topic of the future research. Similarly, the current result provide important basis for future exploration of methods for estimating the statistically optimal range of interest for given ROC curves.

In light of obtained results on statistical efficiency it is important to note that selection of range of interest for pAUC should be driven primarily by clinical/practical considerations, which may override any statistical considerations. For example, if consideration of pAUC over a certain range of specificity is considered to be clinically important and leading to conclusions potentially different from conclusions based on full AUC, pAUC should be used even if it is statistically less efficient under the expected conditions. Selection of the range of interest is driven by the

operating points that can be used for medical decision making. Medical decision making could be influenced by various factors including the cost/benefit of the consequences of performing a diagnostic test and prevalence of the "disease" in the population (Metz, 1978). Thus, although the ROC curve and all results obtained in this work do not depend on prevalence, in practice the expected prevalence of the disease in the target population, as well as cost-benefit structure could have a substantial effect on decision regarding the integration range for pAUC.

6.1 EVALUATION OF A SINGLE PAUC

When evaluating a single pAUC, we investigated two important properties of the pAUC which should facilitate a wider and more appropriate use of this important summary index. First, for ROC curves typically encountered in practical applications the spAUC actually increases with increasing range of interest. For example, the spAUC is always increasing for concave ROC curves and also, when considering short ranges for improper ($b<1$) binormal ROC curves. Second, the statistical uncertainty of the estimated spAUC in general, and its variance in particular, could frequently be smaller than those of the full AUC. In particular, a decrease of the variance with increasing range (as often conjectured) can be observed only in the case of concave binormal ROC curve ($b=1$) with AUC of at least 0.75, or in the case of improper binormal ROC curves with increasingly larger AUCs. This decrease in the width of distribution with increasing range for large AUCs, is likely to be the result of the true value of standardized pAUC approaching its upper boundary. We demonstrated that in the case of straight-line ROC

curves, where standardized pAUC is a constant, the variance increases regardless of how high the ROC curve is.

Our findings have direct practical implications for the design and analysis of diagnostic performance studies in which it is common to disregard partial area indices in favor of inferences based on full area under the ROC curve. Specifically, our results on the statistical uncertainty of estimation indicate that in a number of practical scenarios inferences based on the pAUC could be no less statistically advantageous than inferences based on the full AUC. A program (Appendix C) was developed to estimate the sample size. As compared with the binormal model, statistical inferences for the bi-gamma model based on partial AUC required smaller sample sizes than full AUC when shape parameter was less than 1. Our results on the values of the standardized pAUC indicate that the estimates should always be interpreted in the context of the range of interest, even if standardization is employed. Using a wider range of interest than that which is of clinical interest clinically could lead to overoptimistic estimates of performance in practically relevant scenarios.

6.2 COMPARISON OF TWO CORRELATED PAUCS

First of all, a number of practically reasonable types of non-crossing ROC curves could have statistically significant differences in partial AUCs, but not in full AUCs. Secondly, depending on the expected shape of the ROC curve, planning for future studies based on pAUC could lead to smaller sample size requirements. Thirdly, we demonstrated that comparisons of pAUCs computed over a wider range of two non-crossing ROC curves could have smaller statistical power. Statistical power for the pAUC over a wider range (and the full AUC in particular) tends

to increase with increasing range for ROC curves that have higher curvature in the range of higher FPFs. In cases of flatter ROC curves, and in particular in cases where the curve is nearly-linear in shape in the range of higher FPFs, the statistical power frequently decreases with increasing range.

Experimentally different shapes of ROC curves can be encountered that are both reasonable and plausible. As we illustrated, very similar curves visually could have drastically different properties in terms of pAUC-based inferences. Binormal ROC curves are reasonably straightforward to fit and these provide good approximation for a large number of different types of ROC curves (Hanley, 1988). Yet, this family does not include curves with nearly straight-line shapes, which could be experimentally observed, thereby leading to a different relationship between inferences based on partial and full AUC. The binormal ROC model offers only one type of concave curves for which it is always more beneficial to use in the analysis the full AUC rather than the pAUC. Although improper binormal ROC curves provide a somewhat different picture, these are not likely to be used in sample size estimations due to the unrealistic hooks associated with improper curves. For planning purposes, it is important to have a tool that is flexible enough to allow for the diversity in the shape of the performance curve in different applications. Hence, the bi-gamma family of ROC curves has been advocated by several investigators (Constantine *et al*., 1986) (Dorfman *et al*., 1996) (Faraggi *et al*., 2003) (Huang and Pepe, 2009). The bi-gamma family represents a flexible family that consists of concave ROC curves that includes both bi-normal looking curves and curves with nearly straight-line shapes. As such, the bi-gamma family of ROC curves offers an important tool for planning for future studies aimed at comparing pAUCs under different ROC curves with varying shapes. As we demonstrated, sample size for bi-gamma ROC curves can be adequately estimated using the code

we provide and, for some parameters, the estimates could be quite different from those obtained under the assumption of binormality. We developed a program (Appendix D) to estimate the sample size. Therefore, in studies where differences in pAUCs are more relevant, there may be no need to resort to comparisons of full AUCs simply because of perceived smaller sample size requirement.

6.3 PARTIAL AREA UNDER THE ROC CURVE WITH MASS

Investigations of the properties of comparisons of pAUCs are complicated by the fact that the rating data are not truly continuous, namely, ties are possible. Ties could result from evaluations of normal images, the properties of the diagnostic tool, or the assignment of a default value to all subjects with biomarker levels below a predetermined threshold or below the limit of detection. Therefore, it is important that we understand the properties of statistical inference for this type of data.

We provide some insight into the properties that may be useful for the design and analysis of comparisons of ROC curves with mass. For partially concave binormal ROC curves with mass, increasing the range of interest leads to increasing power, therefore, the statistical inferences based on the full AUC provide maximum power hence the lowest sample size requirement. For concave binormal ROC curves with expected high AUC, (e.g. average AUC greater than 0.825), the statistical power for partially concave binormal ROC curves with mass can be higher than that for conventional concave binormal ROC curves. For a fixed difference in the AUC in the case of partially concave binormal ROC curves, the opposite result was observed.

In contrast, for partial straight-line ROC curves with mass, increasing the range of interest will not affect statistical power thereby sample size requirement. Thus statistical inferences based on more clinically relevant pAUC could be advantageous due purely to its clinical relevance rather than based on statistical considerations. Furthermore, the statistical power for comparing full AUCs from partial straight-line ROC curves with mass are higher than for conventional straight-line ROC curves.

In practice, ROC curves with mass could be observed for different experimental reasons. If there is no useful information contained in a tie, breaking it up is equivalent to randomly assigning ratings, and thus gives us a partial straight-line looking empirical ROC curve. Our results demonstrate that in this scenario breaking up ties would not result in any statistical advantage. Our results also demonstrate that if there exists useful information in the tie (e.g. assigning default value to subjects with biomarker levels below a limit of detection), breaking up the tie correctly will result in less biased estimates, which agrees with works on the inference for ROC curves with limits of detection (Schisterman *et al.*, 2006).

Increased efficiency of inferences based on the AUC from the grouped data agrees with the consequences of the randomized estimator (Lehmann and Casella, 1998). In particular, according to the Rao-Blackwell Theorem (Lehmann and Casella, 1998), for any randomized estimator that is not a function of a sufficient statistic, there always exists a better estimator depending only on the sufficient statistics. For a straight-line ROC curve, the AUC estimator for continuous data can be considered as a randomized estimators, the sensitivity at FPF=0 can be shown to be a sufficient statistic for the entire ROC curve, and the empirical AUC estimator for the grouped data can be considered as a Rao-Blackwell estimator.

Compared with the results for conventional ROC curves (Appendix B), there are similar trends in statistical power when a wider range selection is taken into consideration, namely, increasing the range leads to increasing power for binormal ROC curves and decreasing power for straight-line ROC curves. However, for ROC curves with mass, this increasing or decreasing trend tends to gradually diminish after the point where mass occurs, and thus the statistical power becomes stable (almost a constant).

Our findings may have direct practical implications for the design and analysis of diagnostic performance assessments when one expects a performance curve with mass. The statistical inference based on full AUC for ROC curve with mass could be advantageous as compared with a regular ROC curve without mass in terms of achievement of greater statistical power and lower sample sizes. However the effect of mass is affected by the shape of ROC curves. In the case of ROC curve with nearly straight-line segments, allowing for ties could provide statistical advantages as compared to breaking them.

APPENDIX A

ON USE OF PARTIAL AREA UNDER THE ROC CURVE FOR EVALUATION OF DIAGNOSTIC PERFORMANCE

Ma H, Bandos A, Rockette H, Gur D. “On use of partial area under the ROC curve for evaluation of diagnostic performance”, *Statistics in Medicine* 2013; **32:** 3449-3458.

The part of results was presented in Chapter 3.

APPENDIX B

ON THE USE OF PARTIAL AREA UNDER THE ROC CURVE FOR COMPARISON OF TWO DIAGNOSTIC TESTS

Ma H, Bandos A, Gur D. “On the use of partial area under the ROC curve for comparison of two diagnostic tests” (submitted to Biometrical Journal, 2014).

The part of results was presented in Chapter 4.

APPENDIX C

R PROGRAM FOR ESTIMATING SAMPLE SIZES FOR BI-GAMMA ROC CURVES IN EVALUATION OF SINGLE PARTIAL AUC

#Input:
#k is shape parameter for gamma distribution
#auc is AUC for the bi-gamma ROC curve
#range is partial area we focus on
#CI is length of pre-specified confidence interval
#pow is the pre-specified statistical power
#alpha is the significance level
#Output: estimated sample sizes

```
rm(list=ls())
sample.size.bigamma<-function(k,auc,range,pow,alpha,CI){

set.seed(19840825)
#Tuning parameters
n.sim=500 #the higher the better, but slower
n.iter=100 #the higher the better, but slower

pAUC.thresholds<-range
theta.x=NULL
theta.y=1
diff.pAUC=CI

eroc<-function(q){
    x=q[,1]
    y=q[,2]
    thresholds<-sort(unique(c(x,y)),decreasing=TRUE)
    matrix.thresholds.x<-matrix(rep(thresholds,length(x)),nrow=length(x),byrow=TRUE)
    matrix.thresholds.y<-matrix(rep(thresholds,length(y)),nrow=length(y),byrow=TRUE)
    matrix.x<-matrix(rep(x,length(thresholds)),nrow=length(x))
    matrix.y<-matrix(rep(y,length(thresholds)),nrow=length(y))
    matrix.comp.x<-(matrix.x>matrix.thresholds.x)
    matrix.comp.y<-(matrix.y>matrix.thresholds.y)
    fpf<-apply(matrix.comp.x,2,mean)
    tpf<-apply(matrix.comp.y,2,mean)
    coordinate.temp<-cbind(fpf,tpf)
    coordinate<-rbind(coordinate.temp,c(1,1))
    return(coordinate)
}

pAUC<-function(r){
    fpf=r[,1]
```

```
    tpf=r[,2]
    p_area<-NULL
    for (i in 1:length(pAUC.thresholds)){
        fpf_l=fpf[max(which(fpf<=pAUC.thresholds[i]))]
        tpf_l=tpf[max(which(fpf==fpf_l))]

        fpf_r=fpf[min(which(fpf>=pAUC.thresholds[i]))]
        tpf_r=tpf[min(which(fpf==fpf_r))]

        if (fpf_l==fpf_r) temp=c(pAUC.thresholds[i],tpf_l) else{
            lambda=(pAUC.thresholds[i]-fpf_l)/(fpf_r-fpf_l)
            tpf.pAUC.thresholds=tpf_l*(1-lambda)+tpf_r*lambda
            temp=c(pAUC.thresholds[i],tpf.pAUC.thresholds)
        }

        temp.eroc<-cbind(fpf,tpf)
        coordinate.pAUC=rbind(temp.eroc[1:max(which(fpf<=pAUC.thresholds[i])),],temp)
        fpf0=coordinate.pAUC[,1]
        tpf0=coordinate.pAUC[,2]
        fpf1=c(0,coordinate.pAUC[,1])
        tpf1=c(0,coordinate.pAUC[,2])
        midline=0.5*(tpf0+tpf1[1:length(tpf0)])
        delta=fpf0-fpf1[1:length(fpf0)]
        p_area[i]=sum(delta*midline)
    }
    return(p_area)
}

for (i in 1:length(k)){
    for (j in 1:length(auc)){
        theta.x[j+length(auc)*(i-1)]=qf(1-auc[j], df1=2*k, df2=2*k)
    }
}

sim.var<-function(t.par.x,t.par.y){
    temp.x=replicate(n.sim, rgamma(n.iter,shape=t.par.x[1],scale=t.par.x[2]))
    temp.y=replicate(n.sim, rgamma(n.iter,shape=t.par.y[1],scale=t.par.y[2]))
    sim.xy<-lapply(1:n.sim,function(i) cbind(temp.x[,i],temp.y[,i]))
    temp.eroc=lapply(sim.xy,eroc)
    temp.pauc=sapply(temp.eroc,pAUC)
    var.temp.spauc=var(temp.pauc)/4/(range-range^2/2)^2
    return(var.temp.spauc)
}

sim.v=sim.var(t.par.x=c(k,theta.x),t.par.y=c(k,theta.y))*n.iter
n=ceiling((qnorm(1-alpha/2)+qnorm(pow))^2*sim.v/(CI/2)^2)
return(n)
}

#example
sample.size.bigamma(k=10,auc=0.8,range=0.2,alpha=0.05,pow=0.8,CI=0.1)
```

APPENDIX D

R PROGRAM FOR ESTIMATING SAMPLE SIZES FOR COMPARISONS OF BI-GAMMA ROC CURVES USING PAUC

#Input:
#k is shape parameter for gamma distribution
#auc1 is the lower AUC for the two bi-gamma ROC curves
#auc2 is the higher AUC for the two bi-gamma ROC curves
#range is partial area we focus on
#pow is the pre-specified statistical power
#alpha is the significance level
#Output: estimated sample sizes

```
sample.size.bigamma<-function(k,auc1,auc2,range,rho,alpha,pow){
set.seed(19840818)
library(copula)

#Tuning parameters
n.sim=500 #the higher the better, but slower
n.iter=100 #the higher the better, but slower

pAUC.thresholds<-range
delta=auc2-auc1
theta.x1=theta.x2=NULL
theta.y1=theta.y2=1

gamma.gen<-function(t.par){
    temp.copula<-
mvdc(normalCopula(rho),c("gamma","gamma"),list(list(shape=t.par[1],scale=t.par[2]),list(shape=t.p
ar[3],scale=t.par[4])))
    return(rMvdc(n.iter,temp.copula))
}

eroc<-function(q){
    x=q[,1]
    y=q[,2]
    thresholds<-sort(unique(c(x,y)),decreasing=TRUE)
    matrix.thresholds.x<-matrix(rep(thresholds,length(x)),nrow=length(x),byrow=TRUE)
    matrix.thresholds.y<-matrix(rep(thresholds,length(y)),nrow=length(y),byrow=TRUE)
    matrix.x<-matrix(rep(x,length(thresholds)),nrow=length(x))
    matrix.y<-matrix(rep(y,length(thresholds)),nrow=length(y))
    matrix.comp.x<-(matrix.x>matrix.thresholds.x)
    matrix.comp.y<-(matrix.y>matrix.thresholds.y)
    fpf<-apply(matrix.comp.x,2,mean)
    tpf<-apply(matrix.comp.y,2,mean)
```

```
    coordinate.temp<-cbind(fpf,tpf)
    coordinate<-rbind(coordinate.temp,c(1,1))
    return(coordinate)
}

pAUC<-function(r){
    fpf=r[,1]
    tpf=r[,2]
    p_area<-NULL
    for (i in 1:length(pAUC.thresholds)){
        fpf_l=fpf[max(which(fpf<=pAUC.thresholds[i]))]
        tpf_l=tpf[max(which(fpf==fpf_l))]

        fpf_r=fpf[min(which(fpf>=pAUC.thresholds[i]))]
        tpf_r=tpf[min(which(fpf==fpf_r))]

        if (fpf_l==fpf_r) temp=c(pAUC.thresholds[i],tpf_l) else{
            lambda=(pAUC.thresholds[i]-fpf_l)/(fpf_r-fpf_l)
            tpf.pAUC.thresholds=tpf_l*(1-lambda)+tpf_r*lambda
            temp=c(pAUC.thresholds[i],tpf.pAUC.thresholds)
        }

        temp.eroc<-cbind(fpf,tpf)
        coordinate.pAUC=rbind(temp.eroc[1:max(which(fpf<=pAUC.thresholds[i])),],temp)
        fpf0=coordinate.pAUC[,1]
        tpf0=coordinate.pAUC[,2]
        fpf1=c(0,coordinate.pAUC[,1])
        tpf1=c(0,coordinate.pAUC[,2])
        midline=0.5*(tpf0+tpf1[1:length(tpf0)])
        delta=fpf0-fpf1[1:length(fpf0)]
        p_area[i]=sum(delta*midline)
    }
    return(p_area)
}

for (i in 1:length(k)){
    for (j in 1:length(auc1)){
        theta.x1[j+length(auc1)*(i-1)]=qf(1-auc1[j], df1=2*k, df2=2*k)
    }
}

for (i in 1:length(k)){
    for (j in 1:length(auc2)){
        theta.x2[j+length(auc2)*(i-1)]=qf(1-auc2[j], df1=2*k, df2=2*k)
    }
}

sim.var<-function(t.par.x,t.par.y){
    temp.x=replicate(n.sim, gamma.gen(t.par.x))
    temp.y=replicate(n.sim, gamma.gen(t.par.y))
    sim.xy1<-lapply(1:n.sim,function(i) cbind(temp.x[,1,i],temp.y[,1,i]))
    sim.xy2<-lapply(1:n.sim,function(i) cbind(temp.x[,2,i],temp.y[,2,i]))
    temp.eroc1=lapply(sim.xy1,eroc)
    temp.eroc2=lapply(sim.xy2,eroc)
    temp.pauc1=sapply(temp.eroc1,pAUC)
    temp.pauc2=sapply(temp.eroc2,pAUC)
    sim.diff=temp.pauc2-temp.pauc1
    return(var(sim.diff))
}

roc1=function(x){

pgamma(qgamma(x,shape=k,scale=theta.x1,lower.tail=FALSE),shape=k,scale=theta.y1,lower.tail=FALSE)
}
roc2=function(x){

pgamma(qgamma(x,shape=k,scale=theta.x2,lower.tail=FALSE),shape=k,scale=theta.y2,lower.tail=FALSE)
}

diff.pAUC=integrate(roc2,lower=0,upper=range)$value-integrate(roc1,lower=0,upper=range)$value
var.alter<-sim.var(t.par.x=c(k,theta.x1,k,theta.x2),t.par.y=c(k,theta.y1,k,theta.y2))
```

```
n=ceiling((qnorm(1-alpha/2)+qnorm(pow))^2*var.alter*n.iter/(diff.pAUC)^2)
return(n)
}

#example
sample.size.bigamma(k=3,auc1=0.8,auc2=0.85,range=0.2,rho=0.5,alpha=0.05,pow=0.8)
```

BIBLIOGRAPHY

Bamber DC. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; 12:387-415.

Biggerstaff BJ. Comparing diagnostic tests: a simple graphic using likelihood ratios. *Statistics in medicine* 2000; 19(5): 649-663.

Bandos AI, Rockette HE, Gur D. Use of likelihood ratios for comparisons of binary diagnostic tests: underlying ROC curves. *Medical Physics* 2010; 37(11):5821–5830.

Cai T, Dodd L. Regression analysis for the partial area under the ROC curve. *Statist. Sinica*. 2008; 18,817–836.

Constantine K, Karson M et al. Estimation of P(Y<X) in the gamma case. *Communication Statistics Simulation* 1986; 15: 365-388.

Chen SX, Wong CM. Smoothed block empirical likelihood for quantiles of weakly dependent processes. *Statist. Sinica*. 2009; 19, 71–82.

Dorfman DD, Alf JrE. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals – rating-method data. *Journal of Mathematical Psychology* 1969; 6:487-496.

DeLong ER, DeLong DM, Clarke-Pearson DL. Comaring the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. *Biometrics* 1988; 44: 1033-1053.

Dodd LE, Pepe MS. Partial AUC estimation and regression. *Biometrics* 2003; 59: 614-623.

Egan JP. *Signal Detection Theory and ROC Analysis*. Academic Press, New York, 1st edition, 1975.

Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 3rd edition, 1993.

Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters* 2006; 27: 861-874.

Flahault A, Cadilhac M, and Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *Journal of Clinical Epidemiology* 2005; 58: 859-862.

Faraggi D, Reiser B, and Schisterman EF. ROC curve analysis for biomarkers based on pooled assessments. *Statistics in Medicine* 2003; 11: 1591-1597.

Gonen M and Heller G. Lehmann family of ROC curves. *Medical Decision Making* 2010; 30: 509.

Gur D, Rockette HE, Bandos AI. "Binary" and "Non-Binary" Detection Tasks: Are Current Performance Measures Optimal? *Academic Radiology* 2007; 14(7): 871-876.

Green DM, Swets JA. *Signal Detection Theory and Psychophsics*. Wiley: New York, 1966.

Hanley JA. The robustness of the 'Binormal' assumption used in fitting ROC curves. *Medical Decision Making* 1988; 8(3): 197-203.

Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. *Critical Reviews in Diagnostic Imaging* 1989; 29:307-335.

Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29-36.

Hajian-Tilaki KO, Hanley JA. Comparison of three methods for estimating the standard error of the area under the curve in ROC analysis of quantitative data. *Academic Radiology* 2002; 9: 1278-1285.

Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP. A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Medical Decision Making* 1997; 17: 94-102.

He Y, Escobar M. Nonparametric statistical inference method for paired areas under receiver operating characteristics curves, with application to genomic studies. *Statistics in Medicine* 2008; 27: 5991-5308.

Hillis SL and Metz CE. An analytic expression for the binormal partial area under the ROC curve. *Academic Radiology* 2012; 19(12): 1491-1498.

Hotelling H and Pabst MR. Rank Correlation and Tests of Significance Involving No Assumption of Normality. *Annals of Mathematical Statistics* 1936; 7: 29-43.

Huang Z, Qin GS, Yan Y, Zhou XH. Confidence intervals for the difference between two partial AUCs. *Australian & New Zealand Journal of Statistics* 2012; 54(1): 63-79.

Hussain E. The Bi-Gamma ROC Curve in a Straightforward Manner. *Journal of Basic & Applied Sciences* 2012; 8: 309-314.

Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology* 1996; 201: 745-750.

Lehmann EL, Casella G. *Theory of Point Estimation*. Springer, New York, 2st edition, 1998.

Liu A, Schisterman EF, Wu C. Nonparametric estimation and hypothesis testing on the partial area under receiver operating characteristic curves. *Communications in Statistics—Theory and Methods* 2005; 34: 2077-2088.

Ma H, Bandos A, Rockette H, and Gur D. On use of partial area under the ROC curve for evaluation of diagnostic performance, *Statistics in Medicine* 2013; **32:** 3449-3458.

Metz CE. Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine* 1978; 8(4): 283-298.

Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigation Radiology* 1989; 24:234-245.

Mantel N, Greenhouse SW. The Evaluation of Diagnostic Tests. *Biometrics* 1950; 6: 399-412.

McClish DK. Analyzing a portion of the ROC curve. *Medical Decision Making* 1989; 9: 190-195.

McNeil BJ, Keeler E, Adelstein SJ. Primer on certain elements of medical decision making. *New England Journal of Medicine* 1975; 293: 211-215.

Nelsen RB. *An Introduction to Copulas*. Springer: New York, 1999.

Noether GE. *Elements if Nonparametric Statistics*. Wiley & Sons Inc.: New York 1967.

Norman DA. A comparison of data obtained with different false-alarm rates. *Psychological Review* 1964; 71(3):243–246.

Obuchowski NA, McClish DK. Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices. Statistics in Medicine 1997; 16: 1529-1542.

Qin GS, Zhou XH. Empirical likelihood inference for the area under the ROC curve. Biometrics. 2006; 62, 613–622.

Pepe MS. *Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, 2003.

Perkins NJ, Schisterman EF, Vexler A. Receiver operating characteristic curve inference from a sample with a limit of detection. *American Journal of Epidemiology* 2007; 165:325–333.

Perkins NJ, Schisterman EF, Vexler A. Generalized ROC curve inference for a biomarker subject to a limit of detection and measurement error. *Statistics in Medicine* 2009; 28:1841-1860.

Rachev ST. *Handbook of Heavy Tailed Distributions in Finance*. Elsevier: Boston, 2003.

Schisterman EF, Reiser B, Faraggi D. ROC analysis for markers with mass at zero. *Statistics in Medicine* 2006; 25:623–638.

Schisterman EF, Faraggi D, Reiser B, Hu J. Youden Index and the optimal threshold for markers with mass at zero. *Statistics in Medicine* 2008; 27:297-315.

Shapiro DE. The interpretation of diagnostic tests. *Statistical Methods in Medical Research* 1999; 8:113-134.

Vexler A, Liu A, Eliseeva E, Schisterman EF. Maximum likelihood ratio tests for comparing the discriminatory ability of biomarkers subject to limit of detection. *Biometrics* 2008; 64:895–903.

Wagner RF, Beiden SV, Metz CE. Continuous versus categorical data for ROC analysis: some quantitative considerations. *Academic Radiology* 2001; 8: 328-334.

Wieand S, Gail MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Bometrika* 1989; 76(3): 585-592.

Zhang DD, Zhou XH, Freeman DH Jr, Freeman JL. A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. *Statistics in Medicine* 2002; 21: 701-715.

Zhou XH, Obuchowski NA, McClish DK. *Statistical methods in diagnostic medicine*. New York: Wiley & Sons Inc, 2002.

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

Zuley ML, Bandos AI, Ganott MA, et al. Digital breast tomosynthesis versus supplemental diagnostic mammographic views for evaluation of noncalcified breast lesions. *Radiology* 2013; 266(1): 89-95.