SIMULTANEOUS AND SEQUENTIAL ROC ANALYSES FOR DIAGNOSTIC TESTS

By

RAYMOND RUI FANG

B.Sc., Beijing Broadcasting University, 1983

M.Sc., Beijing Institute of Remote Sensing, 1989

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Department of Statistics)

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September 1991

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Oct. 10, 1991

Abstract

Relative or receiver operating characteristic (ROC) analysis is a simple procedure which can be used to measure the accuracy of diagnostic tests. Diagnostic tests are often used to classify an individual as belonging to one of two populations. Based on statistical decision theory, ROC was first developed to evaluate the performance of electronic signal detection, and has been used to evaluate the accuracy of diagnostic tests. The ROC theory for evaluating one single test, or comparing individual tests is reasonably well understood. The question arises in cases where multiple tests are available as to whether some combination of the tests are better than any single one. In this paper, two ROC procedures of evaluating the aggregate performance of multiple diagnostic tests were presented, one is for evaluating simultaneous multiple diagnostic tests, and the other is for sequential diagnostic tests. These procedures are illustrated by using a breast cancer data set.

Contents

List of Tables

List of Figures

Acknowledgement

I would like to thank my advisor, Dr. Andy Coldman, whose valuable support, help, and numerous constructive suggestions are crucial to this research. I also like to give my thanks to Professor Michael Schulzer for critical reading of the manuscript, Mr. Alec Lai for assistance in part of programming work, and the Department of Statistics for the financial assistance during the period of my graduate study at UBC.

Chapter 1

Introduction

Diagnosis is one of the central problems of medicine and is frequently achieved using physical, bio-chemical, or functional tests on tissue samples. Diagnostic tests are often used to classify individuals as belonging to one of two populations. For convenience, we will refer to one population as $\mathbf{D}_+$ (suggesting the diseased, positive, or case population) and the other population as $\mathbf{D}_-$ (suggesting the non-diseased, negative, or control population).

Here, we will be concerned with test whose outcome may be measured on a continuous numerical scale and that the larger values of the diagnostic test are associated with diseased population, and the smaller values of the test are associated with the non-diseased population. After the numerical scores are obtained, a cutpoint $\mathbf{c}$ is chosen by some criterion such that any individual whose test value, T, is larger than $\mathbf{c}$ is classified as from the diseased population, and any individual whose test value is smaller than $\mathbf{c}$ is classified as non-diseased. *Sensitivity* and *specificity* are measures of the accuracy of a binary diagnostic test. The probability that a randomly chosen individual is classified as diseased when one is, in fact, from diseased population is called the *sensitivity*, or probability of true positive of the test and the probability that an individual is classified as non-diseased when one is from the non-diseased population is called the *specificity*, or probability of true negative.

Two questions arise regarding diagnostic test. The first question is whether there exists a test which is free of error. Secondly, If a perfect test exists, then why do we need a new test. No test may be 100 % accurate, but many are very good. For example, by surgical operation, we can know with high probability whether a patient has breast cancer. In addition, by pathognomonic tests, if the test result is positive, the patient will be diagnosed as diseased with 100 % accuracy. The existence of a gold standard is assumed. In other words, there exists accurate tests for diagnosis. But the problem arises if we operate on a patient before we confirm that one has in fact breast cancer. In many cases, the best tests, which are considered as gold standard, have disadvantages, they may be dangerous, expensive, slow, etc.. Consequently, simple diagnostic tests, which are safe, cheap, and quick, are required. Unfortunately, such tests are usually not 100 percent accurate.

As described previously, a cutpoint is necessary for a diagnostic test to classify objects into one of the two populations. The problem is how to choose an optimal cutpoint. When a cutpoint is chosen, a *sensitivity/specificity* pair is determined. Intuitively, we hope to choose such a cutpoint that the corresponding *sensitivity* and *specificity* would be maximized at the same time (e.g. 1.0), but this is seldom possible. When the cutpoint is chosen to be minus infinity, the corresponding *sensitivity* will be unity, and the *specificity* will be zero. In contrast to this, when the cutpoint is taken to be plus infinity, the *sensitivity* will be zero and the *specificity* will be one.

Sensitivity and *specificity* are important measures of the performance of the diagnostic test. There are direct relations between *sensitivity* and *specificity* and statistical *type I* and *type II* errors, which are defined as the error that an observation is classified as positive when in fact it comes from $\mathbf{D}_-$, and the error that an observation is classified as negative when otherwise. Under the null hypothesis that an observation is from $\mathbf{D}_-$ population, the probabilities of the *type I* and the *type II* errors are

$$\begin{aligned} P(I) &= P(T > \mathbf{c}|\mathbf{D}_-) = P(FP) = 1 - specificity, \\ P(II) &= P(T < \mathbf{c}|\mathbf{D}_+) = 1 - P(T > \mathbf{c}|\mathbf{D}_+) = 1 - P(TP) = 1 - sensitivity. \end{aligned}$$

We cannot, in general, simply compare the *sensitivities* and *specificities* of two tests. For example[1], suppose that there are two potential breast tumor markers: A and B. Each marker gives its test result as a score, and for each test we select a cutpoint to classify samples as diseased or non-diseased. The reason that we cannot simply compare them is that the choice of a different cutpoint may result in a different classification result. At one cutpoint, test A may give a higher correct classification rate, but at another cutpoint, this conclusion may be reversed.

Often in such cases, a single measure of the performance of the diagnostic test which is independent of cutpoint choice is desired. One such measure is based on the ROC curve. ROC analysis removes the arbitrary nature of selecting the cutpoint, and can be used to evaluate and compare the performances of the diagnostic tests.

Based on statistical decision theory, ROC was first developed to evaluate the performance of electronic signal detection methods[2]. In electronic signal detection, an experimenter obtains observations which may include a signal (with background noise), SN, or none (noise only), N. In other words, an observation may come either from SN population or from N population. The experimenter does not know which of the two populations the observation is drawn from and wishes to classify the signal. In this case, the experimenter works as a decision maker, and the action space is composed of two actions: α_1, claim that a signal is present, and α_2, claim that it is not. It is assumed that there are two elements in the state space, $\mathbf{\Theta}$. θ_1 parameterizes the distribution of the SN population, and θ_2 parameterizes the distribution of the N population.

The ROC curve, in this case, is a plot of the *hit rate* versus the the $false\ alarm\ rate$, where the *hit rate* is the probability of correct identification of the signal, and the $false\ alarm\ rate$ is the probability of affirmative response when no signal is present.

Just as an observation can be classified as from SN population or N population in electronic detection, an individual may be classified as from $\mathbf{D}_+$ or $\mathbf{D}_-$ in diagnostic test, where SN population and N population correspond to $\mathbf{D}_+$ and $\mathbf{D}_-$, respectively, and the *hit rate* and $false\ alarm\ rate$ are replaced by *sensitivity* and 1-$specificity$. Drawing from this concept, the ROC analysis may be used in medical applications to evaluate the accuracy of diagnostic tests.

ROC curve is a trace of (1-*specificity*, *sensitivity*) pairs in the unit square that are formed as the cutpoint varies in its range of possible values. In other words, corresponding to each particular value of cutpoint, we obtain a single point on the plot of 1-*specificity* versus *sensitivity*. By varying the cutpoint, we obtain an ROC curve, a locus of points from a plot of empirical 1-*specificity* versus *sensitivity* at each cutpoint. If the sample size is infinite, varying the cutpoint continuously gives the whole ROC curve. If the sample size is finite, we can only obtain finite ROC points on the unit grid, but by some parametric or nonparametric methods which will be discussed in the next chapter, we can still fit a continuous ROC curve. A test which perfectly discriminates would pass through the point (0,1) on the unit grid. In contrast, a test with no discriminating ability will have equal expected values for the *sensitivity* and 1 - *specificity*. Consequently, the ROC curve associated with non-informative test would follow the diagonal of the grid. In practice, most diagnostic tests would have ROC curves which fall between the above two extremes. The better test would have its ROC curve reaching more sharply upward to ideal point (0,1) and being farther away from the diagonal on the grid. One typical ROC curve is shown in Figure 1.1. In Figure 1.2 ROC curves of two different tests are given. Test A with (*sensitivity*, 1-*specificity*) values lying on the upper ROC curve is more discriminating than test B lying on the lower ROC curve.

After the ROC curve is obtained, we can derive some indices as the measures of the performance of the corresponding test from the ROC curve. Based on such indices, diagnostic tests can be evaluated and compared reasonably and effectively. The most commonly used index is the area under the ROC curve of the test, with greater area indicating better tests. The area index as well as some other indices from the ROC curve is described in detail in the next chapter.

In diagnosis of a particular disease, there are often several tests available for testing the same disease. When the result of a single test is subject to substantial error, we can carry out several different diagnostic tests on the same patient, whether simultaneously or sequentially. The ROC theory for evaluating one single test, or comparing several single tests is reasonably well understood. The question arises in cases where multiple tests are available as to whether some combination of the tests are better than any single one. In other words, the question is how we can combine several test results and provide as much information as possible.

In this paper, we will connect multivariate discrimination theory with the ROC methodology, and present an ROC method to evaluate the accuracy and compare the performance of the combinations of several tests, which will be referred as to the ROC analysis for multivariate-measurement test. Also, in many clinical situations, sequential discrimination procedures are increasingly used to classify patients into $\mathbf{D}_+$ or $\mathbf{D}_-$ populations. How to evaluate the performance of sequential discrimination procedures by ROC techniques will also be discussed in this paper.

Chapter 2

ROC Analysis Methodology

ROC principles are based on the notion of a "*decision variable*". The concept is needed because very few clinical diagnostic tests produce results which fall into two obviously defined categories with unequivocal boundaries. With some tests, an "*explicit*" variable exists; for example, a medical test may yield a numerical test result. In such situations, one can choose among an infinity of decision thresholds or cutpoints along the continuum of this variable to serve as the boundary above which one would declare the test *positive*: each different choice will yield a different true-positive fraction and a different false-positive fraction, with a decrease in one being accompanied by an increase in the other. The resulting 2 by 2 table (see, for example, Table 2.1) corresponds to one particular point on the ROC curve.

Performance:

In $\mathbf{D}_{-}$ group

False positive fraction $\quad$ $\mathrm{P}(FP) = 0.05$ (25/500)

Table 2.1: Illustrative Data Table and Measures of Performance from a Binary Diagnostic Test

Test result

	−	+	Total
$\mathbf{D}_-$	475	25	500
$\mathbf{D}_+$	150	350	500

True negative fraction(*specificity*) P(TN) = 1 − P(FP) = 0.95

In $\mathbf{D}_+$ group

True positive fraction(*sensitivity*) P(TP) = 0.70 (350/500)

False negative fraction P(FN) = 1 − P(TP) = 0.30

2.1 Constructing Empirical ROC Points

The ROC curve shows the trade-off between P(TP) (or *sensitivity*) successes and P(FP) (or 1-*specificity*) errors as one employs different decision boundaries.

Consider the hypothetical data of Table 2.2, which represent numbers of rating responses made in each of five categories of a rating scale both for cases truly diseased and for cases truly non-diseased. In each cell the raw frequencies of responses are given at the left. The cumulative response proportions given in parentheses at the right of each cell (moving downward) are the estimates of conditional true-positive TP probabilities (right column). The TP and FP probability (or *sensitivity* and 1-*specificity*) estimates are simply obtained by successively considering each rating-category boundary as if it were a binary-decision criterion.

For convenience, Table 2.2 is usually arranged in the form of Table 2.3, where ratings of 1 corresponds to category ++, ratings of 2 corresponds to category +, and so on, finally, ratings of 5 corresponds to category − −.

Considering only ratings of 1 as diseased, and ratings of 2 to 5 as non-diseased, we obtain an estimate of the FP probability of 0.05, and an estimate of the TP probability of 0.70. This ROC point of (0.05, 0.70) is the equivalent of one generated by a binary-decision criterion that is relatively stringent. Move on now to regard ratings of both 1 and 2 as indicating diseased, and ratings of 3 to 5 as indicating non-diseased–a somewhat more lenient criterion. Then P(FP) equals 0.15 and P(TP) equals 0.80, and so on, through the fifth rating category, yielding the points (0.30, 0.88), (0.60, 0.95), and (1.0, 1.0). The corresponding empirical ROC points are shown in Figure 2.1. We see, incidentally, because the last category always yields P(FP) = P(TP) = 1.0, that rating judgments yield a number of ROC points which equals the number of judgment categories.

Besides the construction of the ROC points described above, some other aspects must also be considered. One important aspect is ROC curve fitting, and another important one is to develop indices to summarize ROC curves for the purpose of performance evaluation and comparison. ROC curve fitting will be described in Section 2.2. In Section 2.3 we will discuss some frequently used ROC indices, as well as some other

Table 2.2: Illustrative Rating-Scale Data [3]

	Stimulus	
Response	Diseased	Non-diseased
Very likely diseased, 1	350 (0.70)	25 (0.05)
Probably diseased, 2	50 (0.80)	50 (0.15)
Possibly diseased, 3	40 (0.88)	75 (0.30)
Probably nondiseased, 4	35 (0.95)	150 (0.60)
Very likely nondiseased, 5	25 (1.00)	200 (1.00)

aspects of the ROC analysis.

2.2 Fitting a Smooth ROC Curve

After a number of points on the ROC curve space are generated from a diagnostic test, it is frequently desirable to obtain a smoothed estimate of the ROC curve. If the sole purpose is to give a very rough, almost qualitative statement and picture of whether the ROC points are close to the upper left corner, close to the diagonal, or halfway in-between, we can use a simple eye-fit. Otherwise, if a more quantitative description is required, as in a formal comparison of two diagnostic tests, some form of objective curve fitting and inferential procedure is necessary.

Basically, ROC curve fitting methods can be divided into two categories, one is

Table 2.3: Illustrative 2 by 5 Table of Rating Data

	Test result (rating category)					
	- -	-	+/-	+	++	Total
$\mathbf{D}_-$	200(1.00)	150(0.60)	75(0.30)	50(0.15)	25(0.05)	500
$\mathbf{D}_+$	25(1.00)	35(0.95)	40(0.88)	50(0.80)	350(0.70)	500

parametric, and the other is nonparametric. The objective is to fit a function $Y = f(X)$, where $Y = \mathrm{P}(TP)$ (or *sensitivity*), and $X = \mathrm{P}(FP)$ (or 1-*specificity*).

2.2.1 Parametria Fitting Methods

The ROC curve can be estimated by a parametric method if the distributions of the test measurements from disease (or signal-plus-noise) and non-disease (or noise) are assumed known. For example, let us simply assume that the two underlying populations are normal with cumulative probability functions $F_{D_-}(x) \sim N(\mu, \sigma^2)$ and $F_{D_+}(x) \sim N(\mu + \Delta, \sigma^2)$. Then given a value of cutpoint, $\mathbf{c} = c_k$, we can calculate the coordinates of the point $t_{c_k} = (P(TP), P(FP))_{c=c_k}$ on the ROC grid as

$$\begin{aligned} P(TP)_{c=c_k} &= P(X > c_k \mid X \in \mathbf{D}_+) \\ &= 1 - \Phi(\frac{c_k - (\mu + \Delta)}{\sigma}) \qquad (2.1) \\ P(FP)_{c=c_k} &= P(X > c_k \mid X \in \mathbf{D}_-) \\ &= 1 - \Phi(\frac{c_k - \mu}{\sigma}) \qquad (2.2) \end{aligned}$$

where Φ is the cumulative probability function of standard normal distribution. If the parameters, such as μ, Δ, and σ in the above normal case, are known, we can decide each point on ROC grid corresponding to each cutpoint, and obtain an accurate continuous ROC curve. Unfortunately, in practice, the parameters of the underlying distributions are usually unknown and are to be estimated from random samples. Morgan [4], Dorfman and Alf [5, 6], and Ogilvie and Creelman [7] have proposed estimation procedures for the cases when noise and signal-plus-noise are assumed to have uniform, normal, and logistic distributions, respectively. Grey and Morgan [8] dealt with both normal and logistic distribution cases. In these papers, maximum likelihood estimates of the parameters for any family of distributions are used.

2.2.2 Nonparametric Fitting Methods

If one is reluctant to impose any prior assumptions or, in the absence of sufficient information, unable to justify a certain distribution, it is preferable to have a distribution-free method, i.e. nonparametric method, of finding the relationship between P(TP) and P(FP). In other words, it is necessary to determine the equation $Y = f(X)$ of an ROC curve without any prior distribution considerations.

There are quite few references related to nonparametric procedures of ROC curve fitting in the literature. One algebraic method of fitting $Y = f(X)$ available was proposed by Birdsall [9] in the context of pure signal detection. According to Birdsall, the ROC curve can be viewed as part of a conic section in general. Jaraiedi and Herrin [10] devised a method which can be applied to any available data set to yield the equation of an ROC curve. They assumed that the ROC locus had a function form of $Y = f(X)$, where X and Y correspond to P(FT) and P(TP) respectively. In their paper, two alternative approaches are developed for estimation of its parameters. Suppose that J is the number of data points available and m is the number of parameters in deciding the function $Y = f(X)$. In the first approach the ROC curve is passed through the first m points, and then the sum of deviations for the remaining $J - m$ points, including (1,1), is minimized. The second approach differs from the first only in the way the objective function is defined: all points are assigned equal weight (equal to 1) and the sum of squared deviations around all points, including (1,1), is minimized.

2.2.3 Linear and Transformed ROC Curve

It should be mentioned that some authors prefer to plot the empirical and the formally fitted ROC curves on transformed (normal, logistic, ...) axes rather than on the conventional linear (0, 1) ones [3]. Such a transformation is most useful when it corresponds to the inverse probability transform projecting ROC curve from probability space to parameter space. The main advantages for doing this are that it is easier for the human eye to compare straight lines rather than curves, and that the lack of fit of a parametric model is more readily apparent on suitably transformed axes. Also, these axes make it easier to plot several curves, since the scales expand as one approaches the extremes, i.e., the crowding at the upper left corner of the unit square is avoided.

2.3 How to Summarize ROC Curves

Finally, some indices to summarize ROC curves must be derived as the measures of accuracy.

2.3.1 Some Typical Indices

As an index of accuracy, one can use the TP fraction corresponding to a particular FP fraction, which one might refer to as $\mathrm{P}(TP)_{FP}$. A second popular index is the area under the ROC curve. Also, one can use $\mathrm{P}(TP)_{FP_{range}}$, which is the average $\mathrm{P}(TP)$ over a restricted high level of FPs.

The $\mathrm{P}(TP)_{FP}$ index was originally discussed by Greenhouse and Mantel [11] and, more recently, by Linnett[12]. The main advantage of the $\mathrm{P}(TP)_{FP}$ index is that it is readily understood. However, it is unlikely that different authors will standardize their $\mathrm{P}(TP)$s to the same value of $\mathrm{P}(FP)$, and one cannot always gather from a published report whether a $\mathrm{P}(FP)$ value was chosen in advance of a study or after inspection of the curve(s). Therefore, Swets and Pickett [3] recommended the area, A, under the ROC curve plotted on ordinary axes as the index of choice.

2.3.2 Simple Trapezoidal-Area Index

The trapezoidal area, A_t, under the "curve" that has been formed simply by joining the empirical ROC points is one of the choices, which is nonparametric and easy to calculate. However, it is likely to be affected by the location or spread (or small number) of the

ROC points and generally yields a smaller area than one derived from a smooth curve, because the smoothed ROC curve is usually convex.

2.3.3 Area Index Based on Statistical Models

In order to distinguish the trapezoidal area from those which are model-based areas of the ROC curve fitted using one probability model, we refer to A_z as the area that is based on a model indexed by the parameter z, e.g., normal model.

Generally, if the distributions of the two underlying populations are known, then we can exactly calculate the area under the ROC curve. Suppose that X and Y are independent continuous variables representing test scores from $\mathbf{D}_-$ and $\mathbf{D}_+$, respectively. Then the area under the ROC curve will be

$$\begin{aligned} A(X,Y) &= \int_0^1 P(Y \geq c) dP(X \geq c) \\ &= -\int_{\infty}^{-\infty} P(Y \geq c) f_X(c) dc \\ &= \int_{-\infty}^{\infty} P(Y \geq c) f_X(c) dc = P(Y \geq X) \qquad (2.3) \end{aligned}$$

which provides an intuitive meaning for the area under an ROC curve. In other words, the area under the ROC curve of a diagnostic test equals to the probability that the random measurement X from $\mathbf{D}_-$ is stochastically dominated by Y from $\mathbf{D}_+$. The above formula can also be used to calculate the area after the model is decided. Replacing the quantities in (2.3) by normal distribution ones, $A(X,Y)$ becomes A_z.

2.3.4 Estimation of Area Index Using U Statistic

The area under an ROC curve can be estimated by non-parametric method, in which the estimation is only based on samples and no statistical model is specified. Suppose that x_1, x_2,...,$x_{n_{D_-}}$, and y_1, y_2,..., $y_{n_{D_+}}$ are samples from $\mathbf{D}_-$ and $\mathbf{D}_+$, respectively, and the x's and y's are independent. Let c_k, k = 1, 2, ..., K, be cutpoints such that c_k > c_{k-1}, where K is the number of cutpoints chosen, and c_k's can be decided from the order statistics from the combined samples. Usually we choose K to be 5, 10, or 20. The maximum K is $n_{D_+} + n_{D_-} + 1$ when there are no tied observations. From Figure 2.2, the k^{th} sub-area, $A(X,Y)_k$ can be calculated by corresponding trapezoidal area as

$$\begin{aligned} A(X,Y)_k &= \frac{1}{2}\left[P(x \geq c_{k-1}) - P(x \geq c_k)\right]\left[P(y \geq c_k) + P(y \geq c_{k-1})\right] \\ &= \frac{1}{2}P(x = c_{k-1})\left[P(y \geq c_k) + P(y \geq c_{k-1})\right] \qquad (2.4) \end{aligned}$$

and the estimated total area is

$$\begin{aligned} A(X,Y) &= \sum_k A(X,Y)_k = \frac{1}{2}\sum_k P(x = c_{k-1})P(y = c_{k-1}) + \sum_k P(x = c_{k-1})P(y \geq c_k) \\ &= \frac{1}{2}\left[Proportion\ of\ all\ x,\ y\ pairs\ in\ data\ with\ x\ =\ y\right] \\ &\quad + \left[Proportion\ of\ x,\ y\ pairs\ with\ x < y\right] \\ &= \frac{\mathbf{U}}{n_{D_+}n_{D_-}} \qquad (2.5) \end{aligned}$$

where

$$\mathbf{U} = (Number\ of\ pairs\ with\ X < Y) + \frac{1}{2}(Number\ of\ pairs\ with\ X = Y)$$

is the Mann-Whitney U statistic when sample size is large. If we assume that x's and y's are continuous variables, which is usually the case in practice, $Prop(x = y) \approx 0$, then,

$$A \quad = \quad [Proportion\ of\ x,\ y\ pairs\ with\ x < y] \tag{2.6}$$

Since the area can be estimated by **U**-statistic, many properties for **U** statistic can be applied to area estimation, such as the property that A is an asymptotically unbiased estimator of the true area. Bamber [13] showed the relationship between the area falling under the points comprising an empirical ROC curve and the Mann-Whitney **U**-statistic. For the general theory of **U**-statistic, see Lehmann [26].

2.3.5 Comments on Area Index

The area index has also been studied by various investigators, including Bamber[13], Hanley and McNeil [14, 15], Goddard [16], Dorfman and Alf [6], and Delong et al [17]. Under certain circumstances, the area index may not be an adequate measure for comparing diagnostic tests. For example, the areas under the two ROC curves may be equal although, in fact, one test dominates the other over an interval of clinically relevant FPs. A solution to this problem is to use $\mathrm{P}(TP)_{FP_{range}}$ index. Part of the statistical theory for this type of index has been developed by Wieand et al [18], in which the statistical procedures are established in the context of a class of indices and illustrated using continuous rather than rating data; however, they can be adapted to rating method data.

2.3.6 Complements

Finally, it should be pointed out that besides the construction, fitting, and summarizing of the ROC curves, some other important aspects should also be taken into consideration. For example, in order to score the correctness of each diagnosis ,the true state of each observation (i.e., the population, $\mathbf{D}_-$ or the population, $\mathbf{D}_+$) must be known. Unfortunately, obtaining such truth data for ROC studies in dealing with real cases is often difficult as described in the first chapter, and investigators sometimes resort to using surrogate truth data. Revesz et al. [19] investigated various methods of approximating the truth on the conclusions of a study that compared three radiographic techniques. They found that any of the three techniques could be shown to be more accurate than the others, depending on which method was used to define the truth.

Henkelman et al. [20] have proposed a method of ROC analysis that does not require truth data, for use when several very accurate tests are being compared, but their suggestion has not been followed up. The main drawback seems to be that by essentially relying on the other tests to "define" the truth, it is difficult for the new modality to appear better. For the sake of our emphasis on the construction of ROC curve for the combination of several tests, we just simply assume that the data in our research is truth data.

Other aspects involved in ROC analysis can be found in Hanley's paper [21], which is an useful survey paper about ROC methodology. How to use ROC methodology to evaluate combined diagnostic tests as well as sequential clinical discrimination tests will be developed in the rest of this paper.

Chapter 3

ROC Analysis For The Test With Multivariate Measurements

As mentioned in the first chapter a problem in ROC curve plotting arises if we have more than one diagnostic test, and we would like to plot a single curve to represent the aggregate performance of those multiple test results. In fact, a combination of several tests can be considered as a single test with multivariate measurements. Here, for our convenience, we refer the test with univariate measurements as to univariate measurement test, or simply univariate test, and a combination of several tests as to multivariate measurement test, or simply multivariate test. One purpose of this paper is to develop an ROC methodology to evaluate the performance of such multivariate diagnostic tests.

There are two obvious strategies of carrying out ROC analysis for evaluating multivariate tests. First, instead of using cutpoint in constructing ROC curve as we did for univariate test, we can use cutline for bivariate tests, cutface for trivariate tests, or even super-cutface for more than three variate tests to calculate TP proportions and FP proportions. If the cutface is 'controlled' by a single parameter, we may then vary this to obtain an ROC curve. Secondly, we can project multivariate measurements on to a straight line by which the resulting univariate measurements from the two populations are separated as much as possible, and then use simple cutpoint to obtain TP and FP proportions. It is clear that the second strategy is a special case of the first. The second approach may be implemented by using multivariate discriminant functions, which is well illustrated by Figure 3.1 in a simple two dimensional example.

For convenience, we begin by defining the notation which will be used in this Chapter. We use upper case $\vec{\mathbf{X}}$ to denote a random vector, and lower case $\vec{\mathbf{x}}$ to a realization. Further, let Ψ be the sample space, Ψ_1 be the sub-set of Ψ for which an observation would be classified as being $\mathbf{D}_-$ and Ψ_2 be the sub-set of $\vec{\mathbf{x}}$ values for which the objects are classified as being $\mathbf{D}_+$, where $\Psi = \Psi_1 \cup \Psi_2$, and $\Psi_1 \cap \Psi_2 = \phi$. In addition, let $f_{D_-}(\vec{\mathbf{x}})$ and $f_{D_+}(\vec{\mathbf{x}})$ be the probability density functions associated with the $p \times 1$ random vector $\vec{\mathbf{X}}$ for the two populations $\mathbf{D}_-$ and $\mathbf{D}_+$, respectively.

In this chapter, we first introduce several typical multivariate discriminant functions, and then describe how to use those functions as well as ROC methodology to plot a single ROC curve for multivariate measurement tests.

Table 3.1: The Form of Data for a Discrimination Analysis

Population	*Individual*	X_1	X_2	...	X_p
1	1	x_{111}	x_{112}	...	x_{11p}
1	2	x_{211}	x_{212}	...	x_{21p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	m	x_{m11}	x_{m12}	...	x_{m1p}
2	1	x_{121}	x_{122}	...	x_{12p}
2	2	x_{221}	x_{222}	...	x_{22p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	n	x_{n21}	x_{n22}	...	x_{n2p}

3.1 Survey of Multivariate Discrimination Procedures

Multivariate discriminant analysis is used to separate two or more groups of individuals, given measurements for these individuals on several variables. Discriminant function is a univariate score based on $\vec{\mathbf{X}}$. In the two population case there are random samples from the two populations, of sizes m and n, respectively, and values are available for p variables X_1, X_2, ..., X_p for each sample member. Thus the data for a discriminant analysis takes the form shown in Table 3.1.

In order to do multivariate discriminant analysis, many procedures have been suggested in the literature. Fisher's linear discriminant function(LDF) [22], [23], and [24] is most widely used. Fisher's idea was to transform the multivariate observations $\vec{\mathbf{x}}$ to univariate observations y such that the y's in the two groups were separated as much as possible. With the assumption that the two populations are normally distributed with the equal covariance matrices, LDF arises as a likelihood ratio. Fisher suggested taking linear combinations of $\vec{\mathbf{x}}$ to create the y's because they are simple functions of $\vec{\mathbf{x}}$ and are easily handled mathematically. If the assumption that the two populations have the same covariance matrices is not true, then it is appropriate to use quadratic discriminant functions (QDF) instead of linear ones[25].

Logistic regression[27] is another useful way for multivariate discrimination with assumption that logit q is linear in the X's, where q devotes the probability of a subject being diseased. One advantage is that we only need to estimate the $p+1$ parameters α_0 and β from the sample data, without having to specify the the underlying distributions, whereas in LDF, for example, there are $2p + (\frac{p!}{2!(p-2)!})$ parameters to estimate. The logistic model is particularly useful for handling diagnostic data[27]. For example, as indicated in [27], if the sampling distributions are multivariate normal with identical dispersion matrices, then the diagnostic variable will be linear-logistically distributed.

Given a discriminant function $D(\vec{\mathbf{x}})$, we can also determine an assignment rule based only on ranked values of D. Some ranking procedures are considered by Randles et al.[28] and [29] and Beckman and Johnson[30], respectively. Alternatively, instead of forming the discriminant function and then using ranks, Conover and Iman[31] proposed ranking the data first and then using the LDF or QDF, called RLDF and RQDF. The ranked data are no longer normally distributed, but RLDF and RQDF work well. They compared RLDF and RQDF with several other nonparametric techniques, and concluded that if the data are multivariate normal, very little is lost by using the RLDF and RQDF methods instead of the LDF and QDF methods. When the data were non-normal, the ranking methods were superior to LDF and QDF and they compared favorably with the other nonparametric methods, such as nearest neighbor and kernel estimator. LDF, QDF, RLDF, and RQDF will be discussed and used in our ROC analysis of this and next Chapters.

In addition, there are several other multivariate discrimination procedures available. For example, sequential discrimination method [32], in which the dimensions of $\vec{\mathbf{X}}$ are assumed to be sequentially gathered, is completely distribution-free. After each step the sequential discrimination method decides either to classify the current $\vec{\mathbf{x}}$ to one of the population, $\mathbf{D}_-$ and $\mathbf{D}_+$ or to introduce another variable into $\vec{\mathbf{x}}$. The sequential discrimination method will be discussed in further detail in Chapter 4. The other discrimination procedures include nearest neighbor techniques [33], partitioning methods [34], and Kernel Method [35].

3.2 Linear and Quadratic Discriminant Functions

In this section, our discussion will follow up from the general discrimination problem to the simple LDF and QDF procedures.

3.2.1 The General Discrimination Problem

It is clear that discrimination rules will not usually provide an error–free method of assignment. This is because there may not be a clear distinction between the measured characteristics of the populations; that is, the groups may overlap. It is then possible to incorrectly classify a $\mathbf{D}_+$ object as belonging to $\mathbf{D}_-$ or a $\mathbf{D}_-$ object as belonging to $\mathbf{D}_+$. A good discrimination procedure should result in few misclassifications. In addition, an optimal discrimination rule should take some prior information into account, such as prior probability of occurrence of each population. Finally, an optimal discrimination procedure should account for the costs associated with misclassification. For example, failing to diagnose an illness is generally more costly than concluding that the disease is present when it is not.

Let P$(+|\mathbf{D}_-)$ denote the conditional probability of classifying an object as $\mathbf{D}_+$ when, in fact, it is from $\mathbf{D}_-$, and P$(-|\mathbf{D}_+)$ denote the conditional probability of classifying an object as $\mathbf{D}_-$ when, in fact, it is from $\mathbf{D}_+$. Similarly, we define P$(+|\mathbf{D}_+)$ and P$(-|\mathbf{D}_-)$. Then we have

$$\begin{aligned} P(+|\mathbf{D}_-) &= P(\vec{\mathbf{x}} \in \Psi_2|\mathbf{D}_-) = \int_{\Psi_2} f_{D_-}(\vec{\mathbf{x}})d\vec{\mathbf{x}} \\ P(-|\mathbf{D}_+) &= P(\vec{\mathbf{x}} \in \Psi_1|\mathbf{D}_+) = \int_{\Psi_1} f_{D_+}(\vec{\mathbf{x}})d\vec{\mathbf{x}} \end{aligned} \tag{3.1}$$

Let π_1 and π_2 be the prior probabilities that an observation comes from $\mathbf{D}_-$ and $\mathbf{D}_+$, respectively, where $\pi_1 + \pi_2 = 1$, then the overall probabilities of correctly or incorrectly classifying objects will be

$$P(correct) = \pi_2 P(+|\mathbf{D}_+) + \pi_1 P(-|\mathbf{D}_-)$$

$$P(incorrect) \;=\; \pi_1 P(+|\mathbf{D}_-) + \pi_2 P(-|\mathbf{D}_+) \tag{3.2}$$

If we take misclassification cost into consideration, then an optimal classification rule could be found in the sense of minimum expected cost of misclassification ($MECM$) by

$$E_{CM} = \rho(+|\mathbf{D}_-)P(+|\mathbf{D}_-)\pi_1 + \rho(-|\mathbf{D}_+)P(-|\mathbf{D}_+)\pi_2 \tag{3.3}$$

where $\rho(+|\mathbf{D}_-)$ denotes the cost of classifying a patient as diseased when he or she is from $\mathbf{D}_-$ population. $\rho(-|\mathbf{D}_+)$, $\rho(+|\mathbf{D}_+)$, and $\rho(-|\mathbf{D}_-)$ can be defined similarly. We assume that $\rho(-|\mathbf{D}_-) = \rho(+|\mathbf{D}_+) = 0$.

It is well known that the regions Ψ_1 and Ψ_2 that minimize the ECM are determined by the values $\vec{\mathbf{x}}$ satisfying the following.

$$\begin{aligned} \Psi_2 : \quad & \frac{f_{D_+}(\vec{\mathbf{x}})}{f_{D_-}(\vec{\mathbf{x}})} \geq \left(\frac{\rho(+|\mathbf{D}_-)}{\rho(-|\mathbf{D}_+)}\right)\left(\frac{\pi_1}{\pi_2}\right) \\ \Psi_1 : \quad & \frac{f_{D_+}(\vec{\mathbf{x}})}{f_{D_-}(\vec{\mathbf{x}})} < \left(\frac{\rho(+|\mathbf{D}_-)}{\rho(-|\mathbf{D}_+)}\right)\left(\frac{\pi_1}{\pi_2}\right) \end{aligned} \tag{3.4}$$

In practice, when both the prior probabilities and misclassification cost ratios are unity or one ratio is the reciprocal of the other, the optimal classification regions are determined simply by comparing the values of the density functions, i.e.,

$$\Psi_2 : \frac{f_{D_+}(\vec{\mathbf{x}})}{f_{D_-}(\vec{\mathbf{x}})} \geq 1; \qquad \Psi_1 : \frac{f_{D_+}(\vec{\mathbf{x}})}{f_{D_-}(\vec{\mathbf{x}})} < 1 \tag{3.5}$$

3.2.2 Classification of Two multivariate Normal Populations

Here we first suppose that the two populations have the same covariance matrix $\mathbf{\Sigma}$. Let $f_{D_-}(\vec{\mathbf{x}})$ and $f_{D_+}(\vec{\mathbf{x}})$ be the joint densities for population D_- and D_+, respectively. Then,

$$f_{D_-}(\vec{\mathbf{x}}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(\vec{\mathbf{x}} - \vec{\mu}_1)'\Sigma^{-1}(\vec{\mathbf{x}} - \vec{\mu}_1)\right] \tag{3.6}$$

$$f_{D_+}(\vec{\mathbf{x}}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(\vec{\mathbf{x}}-\vec{\mu}_2)'\Sigma^{-1}(\vec{\mathbf{x}}-\vec{\mu}_2)\right]$$

where $\vec{\mu}_1$ and $\vec{\mu}_2$ are the mean vectors of $\mathbf{D}_-$ and $\mathbf{D}_+$, and $\vec{\mu}_1$, $\vec{\mu}_2$ and Σ are supposed to be known here. Then

$$\begin{aligned}\frac{f_{D_+}(\vec{\mathbf{x}})}{f_{D_-}(\vec{\mathbf{x}})} &= exp\left[-\frac{1}{2}(\vec{\mathbf{x}}-\vec{\mu}_2)'\Sigma^{-1}(\vec{\mathbf{x}}-\vec{\mu}_2)+\frac{1}{2}(\vec{\mathbf{x}}-\vec{\mu}_1)'\Sigma^{-1}(\vec{\mathbf{x}}-\vec{\mu}_1)\right] \\ &= exp\left[(\vec{\mu}_2-\vec{\mu}_1)'\Sigma^{-1}\vec{\mathbf{x}}-\frac{1}{2}(\vec{\mu}_2-\vec{\mu}_1)'\Sigma^{-1}(\vec{\mu}_1+\vec{\mu}_2)\right] \qquad (3.7)\end{aligned}$$

Consequently, the MECM regions Ψ_1 and Ψ_2 in (3.4) become

$$\begin{aligned}\Psi_2: &\quad (\vec{\mu}_2-\vec{\mu}_1)'\Sigma^{-1}\vec{\mathbf{x}}-\frac{1}{2}(\vec{\mu}_2-\vec{\mu}_1)'\Sigma^{-1}(\vec{\mu}_1+\vec{\mu}_2) \geq ln\left[(\frac{\rho(+|\mathbf{D}_-)}{\rho(-|\mathbf{D}_+)})(\frac{\pi_1}{\pi_2})\right] \\ \Psi_1: &\quad (\vec{\mu}_2-\vec{\mu}_1)'\Sigma^{-1}\vec{\mathbf{x}}-\frac{1}{2}(\vec{\mu}_2-\vec{\mu}_1)'\Sigma^{-1}(\vec{\mu}_1+\vec{\mu}_2) < ln\left[(\frac{\rho(+|\mathbf{D}_-)}{\rho(-|\mathbf{D}_+)})(\frac{\pi_1}{\pi_2})\right] \qquad (3.8)\end{aligned}$$

If we ignore prior probabilities and misclassification cost and simply take both the prior probability and misclassification cost ratios unity, then (3.8) becomes

$$\begin{aligned}\Psi_2: &\quad (\vec{\mu}_2-\vec{\mu}_1)'\Sigma^{-1}\vec{\mathbf{x}} \geq \frac{1}{2}(\vec{\mu}_2-\vec{\mu}_1)'\Sigma^{-1}(\vec{\mu}_1+\vec{\mu}_2) \\ \Psi_1: &\quad (\vec{\mu}_2-\vec{\mu}_1)'\Sigma^{-1}\vec{\mathbf{x}} < \frac{1}{2}(\vec{\mu}_2-\vec{\mu}_1)'\Sigma^{-1}(\vec{\mu}_1+\vec{\mu}_2) \qquad (3.9)\end{aligned}$$

which is exactly the same as Fisher's linear discrimination rule.

Fisher's linear discriminant function converts the $\mathbf{D}_-$ and $\mathbf{D}_+$ multivariate populations into univariate populations such that the corresponding univariate population means are separated as much as possible. In fact, (3.9) can be used as a classification device, which gives one score to each observation vector. Then for a new observation $\vec{\mathbf{x}}_0$, we obtain the linear discriminnt function (LDF) as

$$y = (\vec{\mu}_2-\vec{\mu}_1)'\Sigma^{-1}\vec{\mathbf{x}}_0 \qquad (3.10)$$

where y and $\vec{\mathbf{x}}$ have a linear relation, and the optimal cutpoint between the two univariate population means as

$$cp = \frac{1}{2}(\vec{\mu}_2 - \vec{\mu}_1)'\Sigma^{-1}(\vec{\mu}_1 + \vec{\mu}_2) \tag{3.11}$$

In practice, the population parameters $\vec{\mu}_1$, $\vec{\mu}_2$, and Σ are usually unknown. They may be estimated from observations that have already been correctly classified. Then the estimated LDF is

$$\hat{y} = (\vec{\bar{\mathbf{x}}}_2 - \vec{\bar{\mathbf{x}}}_1)'\mathbf{S}_{\text{pooled}}^{-1}\vec{\mathbf{x}}_0 \tag{3.12}$$

and the optimal cutpoint is

$$\hat{cp} = \frac{1}{2}(\vec{\bar{\mathbf{x}}}_2 - \vec{\bar{\mathbf{x}}}_1)'\mathbf{S}_{\text{pooled}}^{-1}(\vec{\bar{\mathbf{x}}}_1 + \vec{\bar{\mathbf{x}}}_2) \tag{3.13}$$

where $\vec{\bar{\mathbf{x}}}_1$ and $\vec{\bar{\mathbf{x}}}_2$ are the sample mean vectors, and $\mathbf{S}_{\text{pooled}}$ is the pooled sample covariance matrix.

If the assumption that the two underlying populations have equal covariance matrices is not satisfied, the classification rule becomes more complicated. Substituting densities with unequal covariance matrices into (3.5) gives the quadratic classification rule that classifies $\mathbf{x}_0$ to

$$\begin{array}{ll} \mathbf{D}_+ \ if & \frac{1}{2}\vec{\mathbf{x}}_0'(\mathbf{\Sigma}_2^{-1} - \mathbf{\Sigma}_1^{-1})\vec{\mathbf{x}}_0 - (\vec{\mu}_2'\mathbf{\Sigma}_2^{-1} - \vec{\mu}_1'\mathbf{\Sigma}_1^{-1})\vec{\mathbf{x}}_0 - b \geq ln\left[(\frac{\rho(+|\mathbf{D}_-)}{\rho(-|\mathbf{D}_+)})(\frac{\pi_1}{\pi_2})\right] \\ \mathbf{D}_- \ if & \frac{1}{2}\vec{\mathbf{x}}_0'(\mathbf{\Sigma}_2^{-1} - \mathbf{\Sigma}_1^{-1})\vec{\mathbf{x}}_0 - (\vec{\mu}_2'\mathbf{\Sigma}_2^{-1} - \vec{\mu}_1'\mathbf{\Sigma}_1^{-1})\vec{\mathbf{x}}_0 - b < ln\left[(\frac{\rho(+|\mathbf{D}_-)}{\rho(-|\mathbf{D}_+)})(\frac{\pi_1}{\pi_2})\right] \end{array} \tag{3.14}$$

where

$$b = \frac{1}{2} ln(\frac{|\Sigma_2|}{|\Sigma_1|}) - \frac{1}{2}(\vec{\mu}_2' \Sigma_2^{-1} \vec{\mu}_2 - \vec{\mu}_1' \Sigma_1^{-1} \vec{\mu}_1) \qquad (3.15)$$

we obtain the quadratic discriminant function (QDF) and the optimal cutpoint when we ignore the prior probabilities and misclassification costs

$$y = \frac{1}{2} \vec{\mathbf{x}}_0' (\Sigma_2^{-1} - \Sigma_1^{-1}) \vec{\mathbf{x}}_0 - (\vec{\mu}_2' \Sigma_2^{-1} - \vec{\mu}_1' \Sigma_1^{-1}) \vec{\mathbf{x}}_0 \qquad (3.16)$$

$$cp = \frac{1}{2} ln(\frac{|\Sigma_2|}{|\Sigma_1|}) - \frac{1}{2}(\vec{\mu}_2' \Sigma_2^{-1} \vec{\mu}_2 - \vec{\mu}_1' \Sigma_1^{-1} \vec{\mu}_1) \qquad (3.17)$$

and similarly the estimated QDF and cutpoint

$$\hat{y} = \frac{1}{2} \vec{\mathbf{x}}_0' (\mathbf{S}_2^{-1} - \mathbf{S}_1^{-1}) \vec{\mathbf{x}}_0 - (\bar{\vec{\mathbf{x}}}_2' \mathbf{S}_2^{-1} - \bar{\vec{\mathbf{x}}}_1' \mathbf{S}_1^{-1}) \vec{\mathbf{x}}_0 \qquad (3.18)$$

$$\hat{cp} = \frac{1}{2} ln(\frac{|\mathbf{S}_2|}{|\mathbf{S}_1|}) - \frac{1}{2}(\vec{\mathbf{x}}_2' \mathbf{S}_2^{-1} \vec{\mathbf{x}}_2 - \vec{\mathbf{x}}_1' \mathbf{S}_1^{-1} \vec{\mathbf{x}}_1) \qquad (3.19)$$

If either outliers exist or the two underlying populations are not normally distributed, RLDF(or RQDF), or logistic regression methods could be used for the same purpose.

3.3 Evaluating Multi-measurement Tests By ROC

In the above sections we discussed several useful multivariate discrimination functions. By using such functions, multivariate measurements can be transformed into univariate measurements, and much of the information for separating populations is included in the univariate measurements. Meanwhile, each corresponding optimal cutpoint has been derived, upon which the samples can be classified into one of the two underlying populations.

In order to perform an ROC analysis two requirements must be met. First, the measurements must be univariate. Secondly, cutpoints must be set up, varied across the whole range. Based on multivariate discrimination functions, such univariate measurements are automatically produced. Then the traditional ROC analysis described in Chapter 2 can be applied.

In this section, we will emphasize the utility of rank transformation in ROC analysis. On the one hand, the area under the ROC curve only depends on the relative placement of the observed measurements. Since rank transformation of a univariate measurement does not change such relative placement, the area under the ROC curve remains the same whether rank transformation is taken or not. On the other hand, outliers are influential upon the multivariate discriminant functions described previously, because those outliers will change the estimated location and dispersion parameters of the two sample populations. After rank transformation, those outliers, such as the very high observed values for the diseased individuals, or very low values for non-diseased individuals, will agree in magnitude with the other measurements in the same population. In other words, discrimination using rank transformations is not influenced greatly by outliers. Finally, it will be seen in the next chapter, rank transformation can also play an important part in the ROC analysis for the sequential diagnostic tests.

The rank transformation for the test with multivariate measurements involves ranking the k^{th} components of all observations from smallest, with rank 1, to largest, with rank $N=m+n$. Each component is ranked separately for $k=1$, through $k=p$. Then the sample means $\overline{\mathbf{x}}_i(\mathbf{R})$ and sample covariance matrices $\mathbf{S}_i(\mathbf{R})$ are computed on the ranks of the observations from each population separately.

In applying the ROC analysis described in this Chapter to multivariate data, we first check outliers by Q-Q plot. If obvious outliers exist, RLDF or RQDF can be tried. Meanwhile, if the univariate normality for each component is not acceptable, even if there is no obvious outlier, RLDF or RQDF should be used because RLDF and RQDF have no model assumptions. Otherwise, LDF or QDF can be applied. In other words, the advisability of the rank transformation will depend on the data set. The only difference between LDF and RLDF (or QDF and RQDF) is that in RLDF the observations are replaced by their ranks.

Finally, we summarize the major steps to carry out ROC analysis for the test with multivariate measurements as follow.

Step 1:

Consider p diagnostic tests as a single test with p-variate measurements.

Step 2:

Check data set for outliers and normality and choose suitable discriminant function. Then do discriminant analysis, and obtain univariate scores.

Step 3:

Use ROC methodology to analyze those univariate scores from discriminant functions, and obtain the final ROC curve, which represents the performance of the corresponding multiple diagnostic test results.

In Chapter 5, we will apply the ROC techniques described in this chapter to a breast cancer data set, which contains three diagnostic test measurements obtained from 381 individuals by three different tumor markers.

3.4 Summary

In this Chapter we have discussed the ROC technique to evaluate the aggregate performance of several diagnostic tests. We consider several tests as a single test with multivariate measurements. Then we use discrimination functions, such as LDF and QDF or RLDF and RQDF, to project the multivariate measurements on a straight line by which the measurements from the two populations are separated as much as possible. Finally, the traditional ROC analysis is applied to the univariate scores obtained during the projection, and the aggregate performance of the original multiple tests can be evaluated.

As we mentioned before, in diagnosis of a particular disease, there are often several tests available for testing the same disease. If those tests are applied sequentially, we carry out one test and then decide if we need a second test based on the test result of the first test. This process continues until the status of the patient has been decided or we run out of tests. In the next Chapter, we will describe sequential discrimination procedures, and then examine ROC technique for evaluating sequential tests.

Chapter 4

ROC Analysis For Evaluating Sequential Diagnostic Procedures

Sequential discrimination procedure can be of significant interest in clinical diagnosis. For the sake of risk, expense, time, and etc., we hope to use fewer tests to determine whether an individual is really diseased or not. For example, suppose that we have three diagnostic tests for diagnosing breast cancer, if we are quite sure that an individual is of cancer free by the result from the first diagnostic test, we do not need carry out a second on that individual. Otherwise, we should apply successive tests to that individual until we identify their status.

In the following we will describe a sequential discrimination procedure. Then, by simulation, we compare the performance of sequential discrimination procedure with that of simultaneous multivariate discrimination procedure in terms of total error rate and number of tests needed. Finally, we will introduce our new method to apply ROC to evaluate sequential diagnostic procedures.

4.1 Sequential Discrimination Procedures

A sequential discrimination procedure is as follows. For each individual, after each step we decide to either allocate the current $\vec{\mathbf{x}}$ to one of the two populations, $\mathbf{D}_-$ and $\mathbf{D}_+$, or introduce another variable into $\vec{\mathbf{x}}$. In other words, for each patient, after each diagnostic test we decide to either classify him/her to diseased or non-diseased population, or delay judgement and carry out a further test.

One simple sequential discrimination procedure is suggested by Kendall and Stuart [36] and Kendall [37]. They first order the x_1 values of all the $m + n$ sample points and choose a_1 and b_1 such that all observations with $x_1 < a_1$ belong to $\mathbf{D}_-$, and all observations with $x_1 > b_1$ belong to $\mathbf{D}_+$. For those observations with $a_1 < x_1 < b_1$ we go through the same procedure with x_2, choosing a_2 and b_2 such that the observations with $x_2 < a_2$ belong to $\mathbf{D}_-$ and those with $x_2 > b_2$ belong to $\mathbf{D}_+$. For those observations with $a_1 < x_1 < b_1$ and $a_2 < x_2 < b_2$ we proceed to x_3 and continue the process until all $m + n$ observations are classified to their correct groups, or we run out of variables. In this way, there are usually some unclassified observations left.

How to choose the first test to start the sequential diagnosis depends on circumstances. If our emphasis is put on accuracy of the sequential tests, x_1 should be the most accurate test. If we are concerned with economic factor, we can choose the cheapest test as x_1 even if it is not very accurate.

Assuming that potentially $\vec{x}$ has infinite dimensions, most sequential discrimination procedures are based on log-likelihood ratio and can be expressed as follow. Suppose that $f_{D_-}(\vec{\mathbf{x}})$ and $f_{D_+}(\vec{\mathbf{x}})$ are density functions for $\vec{\mathbf{x}}$ from $\mathbf{D}_-$ and $\mathbf{D}_+$, respectively. Let $LR_k = log(f_{D_+}(x_1, x_2, ..., x_k)/f_{D_-}(x_1, x_2, ..., x_k))$, then the classification rule is

$$\begin{aligned} &If\ LR_k < B_k,\ classify\ to\ \mathbf{D}_-; \\ &If\ LR_k > A_k,\ classify\ to\ \mathbf{D}_+; \\ &If\ B_k \leq LR_k \leq A_k,\ introduce\ x_{k+1}\ and\ calculate\ LR_{k+1}. \end{aligned} \tag{4.1}$$

A general theory of sequential discrimination is given by Hora[32].

It is a difficult problem to decide optimal bounds at each stage of the sequential tests, especially when joint distribution functions are involved. Since the main purpose of this Chapter is to develop ROC technique to evaluate sequential diagnostic test, not to design the test itself, we simply pick up the method described in [33] to determine the discrimination bounds. In 4.4, we will show some initial research results in determining optimal bounds in the sense that the P(TP) is maximized at any given P(FT).

In this paper, assuming $\vec{\mathbf{x}}$ to be normally distributed, we will use discrimination rule (4.1) for the first $p-1$ stages, which utilizes all the test measurements available at each stage, and forced discrimination procedure at the p^{th} stage, where p is the total number of tests available.

By the method in [33], we observe x_1 and calculate

$$A = log(\frac{1-\beta}{\alpha}) \tag{4.2}$$

$$B = log(\frac{\beta}{1-\alpha}) \tag{4.3}$$

$$LR_1 = log\left[\frac{f_{D+}(x_1)}{f_{D-}(x_1)}\right] \tag{4.4}$$

where α and β are *type I* and *type II* errors. If x_1 follows either $N(\mu_1, \sigma^2)$ when it comes from $\mathbf{D}_-$ or $N(\mu_2, \sigma^2)$ when it comes from $\mathbf{D}_+$, then (4.4) becomes

$$LR_1 = \frac{1}{\sigma}\left[x_1 - \frac{1}{2}(\mu_1 + \mu_2)\right](\mu_2 - \mu_1) \equiv -D(x_1) \tag{4.5}$$

Then, we will assign x_1 either to $\mathbf{D}_-$ if $LR_1 \leq B$ or to $\mathbf{D}_+$ if $LR_1 \geq A$; otherwise, take a second observation and compute the log-likelihood ratio for the two bivariate distributions,

$$LR_2 = log\left[\frac{f_{D+}(x_1, x_2)}{f_{D-}(x_1, x_2)}\right] \tag{4.6}$$

When $\vec{\mathbf{x}}$=(x_1, x_2) follows either $N(\vec{\mu}_1, \mathbf{\Sigma})$ or $N(\vec{\mu}_2, \mathbf{\Sigma})$, then, similarly to (4.6), we get

$$LR_2 = \left[\vec{\mathbf{x}} - \frac{1}{2}(\vec{\mu}_1 + \vec{\mu}_2)\right]' \mathbf{\Sigma}^{-1}(\vec{\mu}_2 - \vec{\mu}_1) \equiv -\mathbf{D}(\vec{\mathbf{x}}) \tag{4.7}$$

Then, we will assign $\vec{\mathbf{x}}$ either to $\mathbf{D}_-$ if $LR_2 \leq B$ or to $\mathbf{D}_+$ if $LR_2 \geq A$; otherwise, take a third observation and compute the log-likelihood ratio for the two trivariate distributions, LR_3 and so on. Generally, at the first p-1 stages, the discrimination rule is to assign to $\mathbf{D}_-$ if

$$LR_k = log\left[\frac{f_{D+}(x_1, .., x_k)}{f_{D-}(x_1, .., x_k)}\right] < B \tag{4.8}$$

to $\mathbf{D}_+$ if

$$LR_k = log\left[\frac{f_{D+}(x_1, .., x_k)}{f_{D-}(x_1, .., x_k)}\right] > A \tag{4.9}$$

and observe x_{k+1} otherwise. LR_k can be written in the same form as (4.6) in multinormal case but with $\vec{\mathbf{x}}$ having k components.

At the p^{th} stage, in which x_p is the last test measurement available, we simply apply forced discriminant procedure discussed in Chapter 3 to those observations which were not classified at the previous stages, that is, assign $\vec{\mathbf{x}}$ to $\mathbf{D}_-$ if $LR_p < 0$ to $\mathbf{D}_+$ if $LR_p > 0$.

4.2 Performance Evaluation

The performance of sequential diagnostic procedures can be compared with that of simultaneous multivariate discrimination procedures in terms of total error rate and number of tests needed, which will be defined later. The performance of sequential diagnostic procedures can also be evaluated by ROC. In this chapter we will develop a new method based on rank transformation and ROC methodology to evaluate sequential diagnostic procedures.

The main advantage of sequential discrimination procedure is that the patients could be classified to either $\mathbf{D}_-$ or $\mathbf{D}_+$ populations by fewer diagnostic tests compared with simultaneous procedure. Of course, benefit of test saving will result in loss of accuracy. By limited simulation below, these two procedures were compared in terms of total error rate and number of tests needed. Though we did not take account all the different situations, the limited comparisons still gave us some rough idea that in large sample case the sequential procedure is asymptotically as good as simultaneous procedure in accuracy, but, on the other hand, results in a significant reduction in the number of tests needed.

4.2.1 Comparison under Different Distribution Overlap

Samples were generated from each of the two populations, $\mathbf{D}_-$ and $\mathbf{D}_+$, in which the parameters were chosen roughly based on one, two, and three times of standard deviations of separation. The trivariate normal distributions sampled were as follows:

a. Weak overlap:

$$\vec{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{\mu}_2 = \begin{bmatrix} 3 \\ 2.5 \\ 2.25 \end{bmatrix}, \quad \textbf{and} \quad \Sigma = \begin{bmatrix} 1.00 & 0.25 & 0.50 \\ 0.25 & 1.00 & 0.75 \\ 0.50 & 0.75 & 1.00 \end{bmatrix}$$

b. Medium overlap:

$$\vec{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{\mu}_2 = \begin{bmatrix} 2 \\ 1.5 \\ 1.25 \end{bmatrix}, \quad \textbf{and} \quad \Sigma = \begin{bmatrix} 1.00 & 0.25 & 0.50 \\ 0.25 & 1.00 & 0.75 \\ 0.50 & 0.75 & 1.00 \end{bmatrix}$$

c. Strong overlap:

$$\vec{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{\mu}_2 = \begin{bmatrix} 1 \\ 0.5 \\ 0.25 \end{bmatrix}, \quad \textbf{and} \quad \Sigma = \begin{bmatrix} 1.00 & 0.25 & 0.50 \\ 0.25 & 1.00 & 0.75 \\ 0.50 & 0.75 & 1.00 \end{bmatrix}$$

We generated samples of size 500 from each of these distributions and classified them by the two procedures, in which $\alpha = 0.05$ and $\beta = 0.20$ were used in the sequential

Table 4.1: Comparison under Different Overlap (1,000 observations)

Method	Overlap	Error	Tests Needed :	RE
		Rate	Correctly Classified:	
			Incorrectly Classified	
Simultaneous	Weak	0.045	3,000:955:45	0.00
Sequential	Weak	0.052	1,233:948:52	0.8835
			(1,000:813:30/157:78:3/76:57:19)	
Simultaneous	Medium	0.121	3,000:879:121	0.00
Sequential	Medium	0.138	1,892:862:138	0.554
			(1,000:458:22/520:133:15/372:271:101)	
Simultaneous	Strong	0.274	3,000:726:274	0.00
Sequential	Strong	0.276	2,908:724:276	0.046
			(1,000:7:0/993:64:14/915:653:262)	

procedure (Equation 4.2 and 4.3). Repeated this process by five times and calculated the average proportions of the observations which were misclassified, and the average number of tests needed to classify all the observations. The results are given in Table 4.1. From the Table 4.1, the error rates, which are defined as

$$Error\ Rate = \frac{Number\ of\ False\ Positive\ +\ Number\ of\ False\ Negative}{Total\ number\ of\ the\ Observations}$$

of sequential discrimination procedure are larger than, but quite close to the corresponding error rates of simultaneous procedure. Further, let p be the number of different tests available, $N(Sequential)$ be the number of tests actually done by the sequential test, and M be the number of subjects, then we define the relative efficiency (RE) of sequential test as

$$\begin{aligned} RE &= \frac{Number\ of\ tests\ saved}{M \times (p-1)} \\ &= \frac{N(Simultaneous) - N(Sequential)}{M \times (p-1)} \\ &= \frac{M \times p - N(Sequential)}{M \times (p-1)} \end{aligned}$$

When no test can be saved, the $RE = 0$, and when all the second tests and the later tests can be saved, the $RE = 1$. From Table 4.1, the number of the test saved by the sequential procedure is significant, especially when the overlap between the two populations is not great. Meanwhile, as the distribution overlap increases, there were fewer observations which could be classified by the first two sequential discrimination rules. Therefore, more and more observations could not be classified until last stage, i.e., until the forced discrimination stage.

4.2.2 Comparison under Different Correlation Coefficients

To investigate the effect of correlation coefficients among different diagnostic test measurements on the relative efficiency of sequential procedure, we generated 500 samples from each of the two populations in the three situations.

a. Weak correlated:

$$\vec{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{\mu}_2 = \begin{bmatrix} 3 \\ 2.5 \\ 2.25 \end{bmatrix}, \quad \textbf{and} \quad \Sigma = \begin{bmatrix} 1.00 & 0.10 & 0.10 \\ 0.10 & 1.00 & 0.10 \\ 0.10 & 0.10 & 1.00 \end{bmatrix}$$

b. Medium correlated:

$$\vec{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{\mu}_2 = \begin{bmatrix} 3 \\ 2.5 \\ 2.25 \end{bmatrix}, \quad \textbf{and} \quad \Sigma = \begin{bmatrix} 1.00 & 0.50 & 0.50 \\ 0.50 & 1.00 & 0.50 \\ 0.50 & 0.50 & 1.00 \end{bmatrix}$$

c. Strong correlated:

$$\vec{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \vec{\mu}_2 = \begin{bmatrix} 3 \\ 2.5 \\ 2.25 \end{bmatrix}, \quad \textbf{and} \quad \Sigma = \begin{bmatrix} 1.00 & 0.90 & 0.90 \\ 0.90 & 1.00 & 0.90 \\ 0.90 & 0.90 & 1.00 \end{bmatrix}$$

The proportions of misclassification and the relative efficiencies are given in Table 4.2, where $\alpha = 0.05$, and $\beta = 0.20$ for the sequential procedure. As the correlations increased, the information contained in three tests approached that of a single test. If the three tests were identical, then three tests would be as good as a single one. In Table 4.2 the error rates increased and relative efficiencies decreased as the correlations increased.

4.2.3 Comparison under Different Discrimination Bounds

The two discrimination bounds, A and B, which depend upon α and β, are crucial to the relative efficiency of sequential procedure. In the following situations we compared

Table 4.2: Comparison under Different Correlation (1,000 observations)

Method	Correlation	Error Rate	Tests Needed : Correctly Classified: Incorrectly Classified	RE
Simultaneous	Weak	0.015	3,000:985:15	0.00
Sequential	Weak	0.037	1,205:963:37	0.8975
			(1,000:813:30/157:106:3/48:44:4)	
Simultaneous	Medium	0.047	3,000:953:47	0.00
Sequential	Medium	0.057	1,274:943:57	0.863
			(1,000:813:30/157:39:1/117:91:26)	
Simultaneous	Strong	0.067	3,000:933:67	0.00
Sequential	Strong	0.085	1,253:922:78	0.843
			(1,000:813:30/157:0:0/157:102:55)	

sequential and simultaneous procedures at

$$a: \quad \alpha = 0.10 \qquad and \quad \beta = 0.20$$

$$b: \quad \alpha = 0.05 \qquad and \quad \beta = 0.20$$

$$c: \quad \alpha = 0.01 \qquad and \quad \beta = 0.01$$

Using samples of size 500 from each of the two populations with medium correlation, the results are given in Table 4.3. It is obvious that as α and β decrease, the upper (lower) bound would become high (low). At the first two stages, more and more observations fell into the intervals between the two bounds, and could not be classified. Thus, the total tests needed increased and the error rates decreased as the intervals enlarged.

As a whole, sequential discrimination procedure is nearly as good as simultaneous procedure in accuracy, and the number of tests needed for classification reduces significantly compared to simultaneous procedure. In addition, it should be pointed out that the performance of sequential test depends heavily on the separation of the two populations on x_1, i.e., the best test among all the tests. If the best test is good enough, the second and the third tests will be unnecessary. In contrast, if all the three tests are quite poor, the sequential test will not yield sufficient savings.

4.3 Evaluating Sequential Procedures by ROC

In the above sections we discussed sequential discrimination procedures, and compared them with the simultaneous procedures. Although the sequential procedure is suboptimal in the sense of minimum misclassification rate, the number of tests saved is

Table 4.3: Comparison under Different Discrimination Bounds

Method	Bounds	Error Rate	Tests Needed : Correctly Classified: Incorrectly Classified	RE
Simultaneous		0.047	3,000:953:47	0.00
Sequential	α=0.10; β=0.20 log(B)=-1.5 log(A)=2.08	0.069	1,195:931:69 (1,000:828:47/125:54:1/70:49:21)	0.9025
Sequential	α =0.05; β=0.20 log(B)=-1.56 log(A)=2.77	0.057	1,274:943:57 (1,000:813:30/157:39:1/117:91:26)	0.863
Sequential	α=0.01; β=0.01 log(B)=-4.6 log(A)=4.6	0.049	1,682:951:49 (1,000:647:7/346:20:0/326:284:42)	0.659

significant relative to simultaneous ones. In this section, we will develop a procedure to evaluate the performance of sequential diagnostic procedures.

The major problem in using ROC here is how we set high dimensional cutpoints in order to obtain an ROC curve. At the first stage we need a cutpoint to separate the univariate measures in x_1 space, at the second stage we need a surface to separate bivariate measures in (x_1, x_2) space, and so on. Following the idea of projection in Chapter 3, we can use multivariate discrimination functions to transform multivariate measurements to an univariate one, and use cutpoints at each stage. The problem arises that how we should combine those univariate measurements obtained at each stage.

The purpose here is the same as that in Chapter 3, i.e., combine multivariate sequential measurements into univariate scores. The univariate scores will be expressed in terms of ranks. The diseased individuals should have high ranks, and the non-diseased ones should have low ranks. Further, the individuals who are classified by the first test as diseased (or non-diseased) are considered to have highest ranks (or lowest ranks). The individuals who can not be classified until the last test should have moderate ranks. Such idea will be described in detail in the following.

In sequential discrimination, only those individuals with either very high or very low values among all the individuals would be classified by the first test. We consider these people as very diseased and very non-diseased respectively. Therefore, the ranks of those classified individuals would be either highest ones or lowest ones. Then, at the second stage, discriminant function based on the first two best tests on those unclassified at the first stage gives univariate measurements, among which those with either relatively high or low values would be classified at this stage. Those people are considered as diseases and non-diseased respectively. Thus, the ranks of the individuals with relatively very high observations classified at the second stage would follow that of the individuals with very high observations classified at the first stage, while the ranks of the individuals with relatively very low observations classified at the second stage would be followed by that of the individuals with very low observations at the first stage, and etc..

For example, suppose that there are a total of N observations to be classified. At the first stage, n_1 and m_1 observations are classified as *positive* and *negative*, respectively. Thus, those n_1 observations would rank from N down to $N - n_1 + 1$, and those m_1 observations would rank from 1 up to m_1. Then, at the second stage, n_2 and m_2 observations are classified as *positive* and *negative*, respectively. Therefore, those n_2 would rank from $N - n_1$ down to $N - n_1 - n_2 + 1$, and those m_2 would rank from $m_1 + 1$ up to $m_1 + m_2$. Finally, the remained $N - n_1 - n_2 - m_1 - m_2$ observations would be forced to be classified at the third stage, and their ranks would be in the middle. The relative ranks of the observations classified at each stage depend on the absolute values of the univariate measurements.

After the above rank transformation based on sequential discrimination procedure, an univariate rank "measurements" are obtained. Sequently, traditional ROC technique can be applied to evaluate the performance of the corresponding sequential discrimination procedures. We summarize the above ROC procedure as follows, where we suppose that the best test is in the sense of most accurate.

Step 1:

Apply ROC procedure to each diagnostic test, respectively, and find out the best in terms of the largest area under the ROC curve. In the same way decide the second best discriminator and so on.

Step 2:

The sequential discrimination procedure will start from the test which is the best discriminator with given α and β. Using (4.6) and comparing with A and B, partial observations will be classified and the corresponding ranks are decided.

Step 3:

For those which have not been classified at the first stage, add the second test measurements, which is best conditional on that the first best test is used and construct consequent bivariate measurements. Using (4.8) and (4.9), and comparing with discrimination bounds, those observations which fall outside the interval will be classified and their ranks, as described earlier, are determined.

Step 4:

Suppose that there are only three diagnostic tests available, then at the third stage all the observations unclassified before would be classified.

Step 5:

Combining all the observations by their ranks which are derived at different stages, we obtain an univariate rank measurements. Finally, ROC analysis is applied to those rank measurements, and the performance of the sequential diagnostic procedure could be evaluated.

4.4 Some Initial Work On Determining Bounds

In the previous discussion of sequential discrimination procedures, discrimination bounds were determined by the preassumed *type I* and *type II* errors. In other words, when the two errors α and β are chosen, the discrimination bounds A_k and B_k are decided. Usually, α and β are chosen according to the required power ($=$ 1- β) and the significance level ($= \alpha$) of the test.

The constant bounds given by (4.2) and (4.3) are simple, but unlikely optimal. There may be some other ways to choose optimal discrimination bounds in the sense that for a given value of P(FP), say P_{fp}, we find bounds such that P(TP) is maximized. For example, in a simple two-stage diagnostic test, we have three bounds, A, B, and C, to be decided. Suppose that object x has two sequential measurements, x_1 and x_2, and our classification rule is

At the first stage:

$$\begin{aligned} &If \quad x_1 \geq A, \ classify \ x \ to \ \mathbf{D}_+; \\ &If \quad x_1 \leq B, \ classify \ x \ to \ \mathbf{D}_-; \qquad (4.10) \end{aligned}$$

$$If \quad x_1 \in (B, A), \; go \; to \; the \; second \; measurement \; x_2.$$

and at the second stage:

$$\begin{aligned} &If \quad x_2 \geq C, \quad classify \; x \; to \; \mathbf{D}_+; \\ &If \quad x_2 < C, \quad classify \; x \; to \; \mathbf{D}_-. \end{aligned} \tag{4.11}$$

Assume that the two populations are normally distributed with known parameters

$$\vec{\mu}_1 = \begin{bmatrix} \mu_{11} \\ \mu_{12} \end{bmatrix} \quad \vec{\mu}_2 = \begin{bmatrix} \mu_{21} \\ \mu_{22} \end{bmatrix} \quad \mathbf{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

then we can calculate P(TP) and P(FP) as

$$\begin{aligned} P(TP) &= P\left\{x_1 \geq A | f_{D_+}\right\} + P\left\{x_1 \in (B, A), x_2 \geq C | f_{D_+}\right\} \\ &= 1 - \Phi(A - \mu_{11}) + \int_C^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_2 - \mu_{12})^2} \\ &\quad \left[\Phi(\frac{A - \mu_{11} - \rho(x_2 - \mu_{12})}{\sqrt{1-\rho^2}}) - \Phi(\frac{B - \mu_{11} - \rho(x_2 - \mu_{12})}{\sqrt{1-\rho^2}})\right] dx_2 \quad (4.12) \\ P(FP) &= P\left\{x_1 \geq A | f_{D_-}\right\} + P\left\{x_1 \in (B, A), x_2 \geq C | f_{D_-}\right\} \\ &= 1 - \Phi(A - \mu_{21}) + \int_C^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_2 - \mu_{22})^2} \\ &\quad \left[\Phi(\frac{A - \mu_{21} - \rho(x_2 - \mu_{22})}{\sqrt{1-\rho^2}}) - \Phi(\frac{B - \mu_{21} - \rho(x_2 - \mu_{22})}{\sqrt{1-\rho^2}})\right] dx_2 \quad (4.13) \end{aligned}$$

Let ψ(A,B,C) = P(TP), and ϕ(A,B,C) = P_{fp} - P(FP), then the problem becomes that we find A, B, and C so as to maximize ψ(A,B,C) subject to ϕ(A,B,C) = 0. In other words, we need maximize

$$\begin{aligned} \Gamma &= \psi(A, B, C) + \lambda\phi(A, B, C) \\ &= P(TP) + \lambda(P_{fp} - P(FP)) \end{aligned} \tag{4.14}$$

where λ is the Lagrange multiplier.

Direct solution for the above optimization problem is difficult. One suggested way to solve the problem iteratively is to use the Newton-Raphson method. Unfortunately, we can not give a general criterion of deciding initial values of A, B, and C. In the following we will suggest a possible procedure of choosing optimal bounds in the sense that P(TP) is maximized when P(FP) is given.

The idea is that for the given values of A, and B, choose C such that P(FP) $= P_{fp}$, then calculate the corresponding P$(TP)_{(A,B,C)}$. Different values of A and B will result in different C and P$(TP)_{(A,B,C)}$. The optimal values of (A, B, C) will be the values at which P(TP) reaches its maximum.

Then the problem is how to choose limited reasonable pair of (A, B) among the infinite combinations. By (4.13) we have

$$P(FP) \geq 1 - \Phi(A - \mu_{11}) \tag{4.15}$$

$$P(FP) \leq 1 - \Phi(B - \mu_{11}) \tag{4.16}$$

When P(FP) $= P_{fp}$, there exist constants a and b such that (4.15) and (4.16) are equivalent to the following equalities.

$$1 - \Phi(A - \mu_{11}) = aP_{fp} \tag{4.17}$$

$$\Phi(B - \mu_{11}) = b(1 - P_{fp}) \tag{4.18}$$

where $0 \leq a \leq 1$, and $0 \leq b \leq 1$. Selecting a and b, the bounds A and B can be solved from (4.17) and (4.18), and C can be computed from (4.13). For example, let a as well as

b to be 0.1, 0.3, 0.5, 0.7, and 0.9, respectively. We have 25 combinations of corresponding (A, B). Given P$(FP) = P_{fp}$, the bound C can be calculated, assuming that the two distributions are known. Thus, the resulting 25 values of P(TP) are obtained. Choose the approximate optimal bounds A, B, and C such that the corresponding value of P(TP) is the largest among the 25 values. Of course, the more values a and b take, the more accurate the values of A, B, and C will be, and the more amount of calculations will be needed. We can also write a program to do automatic searching.

For example, given the distribution parameters as $\mu_{11} = \mu_{12} = 0$, $\mu_{21} = 1.5$, $\mu_{22} = 1.0$, and $\rho = 0$, we find the maximum value of P(TP) and corresponding A, B, and C with a given value of P_{fp}. Choosing P_{fp} to be a batch of values, we summarize the results in Table 4.4.

From Table 4.4, as the value of P_{fp} increased, the corresponding P(TP) increased as well. The bounds A and B shifted to the left, and the bound C changed accordingly.

Finally, we compare the P(TP) value at each P_{fp} point of our nearly optimal procedure (OP) with that of the procedure with discrimination bounds calculated by Kendall and Stuart's method (KS), and give the results in Table 4.5.

It is clear that when P_{fp} is fixed, the larger the P(TP) is, the stronger the discrimination ability will be. From Table 4.5, all the P(TP) values by OP are higher than those by KS, especially when P_{fp} is small. Thus the improvement by OP method is obvious.

Finally, it should be pointed out that the results from the optimal algorithm in the

Table 4.4: The Maximum P(TP) and the Corresponding A, B, and C

P_{fp}	Maximum P(TP)	A	B	C
0.01	0.2599	2.6525	1.2320	1.5788
0.02	0.3622	2.4093	1.1851	1.2309
0.03	0.4309	2.2576	1.1408	1.0084
0.04	0.4828	2.1449	1.0985	0.8411
0.05	0.5243	2.0542	1.0581	0.7059
0.06	0.5588	1.9778	1.0194	0.5917
0.08	0.6172	1.8526	0.6307	0.8173
0.10	0.6606	1.4758	0.3319	1.23
0.20	0.8271	1.2817	-1.3514	1.1600
0.30	0.9442	1.0364	-1.4132	0.8630
0.50	0.9561	0.3853	- 0.6745	0.32
0.90	0.9986	- 0.8779	-0.18808	- 0.15

Table 4.5: Comparison of P(TP) Values for Two Methods

	False Positive Proportion (P_{fp})								
Method	0.01	0.02	0.04	0.06	0.10	0.20	0.30	0.50	0.90
OP	0.2599	0.3622	0.4828	0.5588	0.6643	0.8271	0.9442	0.9561	0.9987
KS	0.2278	0.3118	0.4181	0.4908	0.5922	0.7407	0.8272	0.9242	0.9958

Table 4.4 and 4.5 are theoretical values with the assumption that the two populations are normally distributed with known parameters and common variance. If these If these assumptions are not satisfied, the performance will decrease. Meanwhile, the optimal bounds A, B, and C, calculated froom a fixed value of P(FP) do not necessarily result in the same value of false positive proportion for the data set because the estimated parameters are not exactly equal to the true values.

Chapter 5

Analyses Of Breast Cancer Data Using ROC Procedures

In the previous chapters we described the traditional ROC methodology, and developed two modified ROC procedures, which can be used to evaluate the aggregate performances of several diagnostic tests and of sequential diagnostic methods, respectively. In this chapter we will apply those two modified ROC procedures to breast cancer data obtained from three tumor markers. The data set is used with permission from British Columbia Cancer Agency. The data set was first studied by Silver et al. [1], and considered of three potential breast tumor markers, CEA, CA15.3, and MCA.

5.1 Tumor Markers and Breast Cancer Data

Data were obtained using three tumor markers, CEA, CA15.3, and MCA. We rearrange the data by three groups: normal(N), breast cancer(BC), and high risk patients(HR).

Tumor markers are important methods in clinical diagnosis, in which they work as discriminators between breast-cancer patients and non-breast-cancer patients. CEA was first used and well accepted as a tumor marker for diagnosing cancer[38]. In recent years, more and more potentially specific tumor markers have been developed [39], among which are CA15.3 and MCA.

Here, we will not go into too much detail of tumor makers, but only give some general descriptions of the three tumor markers, as described in [1]. Carcinoembryonic antigen(CEA) is a large complex molecular weight glycoprotein associated with the cellular glycocalyx. The measurement of CEA employs an enzyme immunoassay method using heat extraction and polyclonal anti-CEA antibody. The CA15.3 test uses a double determinant radioimmunoassay utilizing two different monoclonal antibodies, 115D8 and DF3. 115D8 is immobilized on polystyrene beads to complete the double antibody sandwich with 125I-DF3. Mucinous-like carcinoma-associated antigen(MCA) is a 350 Kd glycoprotein produced by mammary carcinomas and some normal tissues. The monoclonal antibody used to detect MCA is known as b12.

The data to be used here are from three groups. The normal group(N) contained samples from 100 female donors, aged from 16 to 91, collected and donated by the Canadian Red Cross. Donors did not suffer from known neoplastic, inflammatory or liver disease; In the breast cancer(BC) group, samples were from 158 female patients with histologically confirmed breast carcinoma. The third group (HR) consisted of samples from 33 women who were believed to be at high risk of developing breast carcinoma on the basis of physical examination or mammogram. Each of the 33 women had undergone diagnostic cytology which did not reveal evidence of malignancy.

5.2 Exploratory Data Analysis

Before we start to evaluate the three tumor markers by modified ROC procedures, we first undertook exploratory data analyses on the 291 original observations from the three groups. In Table the basic statistical descriptions for the measurements from three tumor markers in each group are summarized.

In Table 5.1 we notice that the dispersions of the observations from both Group N and Group HR are not severe. The mean and median values in each group are quite close, and the standard deviations are relatively small. These imply that the data in these two groups agree with one another well in magnitude. These were proven to be nearly true by the histograms and the Q-Q plots shown in Figure 5.1 to Figure 5.12 The Q-Q plots show that the normalities for the observations in the two groups are satisfied.

Also in Table 5.1 the distributions for the three measurements in Group BC are not symmetric at all. From Figure 5.13, 5.15, and 5.17 the normalities for the three measurements in this group even worse. The reason for these is that there are several extreme large measurements, which appear obviously in the histograms in Figure 5.14,

Table 5.1: Statistical Descriptions for Original Data in Three Groups

Group	Test	Mean	SD	Max	Q3	Median	Q1	Min
Normal	CEA	0.95	0.67	2.8	1.45	0.80	0.40	0.1
	CA15.3	15.02	4.93	28.2	17.95	15.10	11.00	5.5
	MCA	6.01	3.36	13.5	8.15	5.95	3.15	0.6
High Risk	CEA	1.55	1.11	4.1	2.3	1.0	0.8	0.4
	CA15.3	20.10	8.60	51.0	24.3	19.7	14.1	8.0
	MCA	7.95	4.64	18.0	10.6	9.2	3.4	0.6
Cancer	CEA	125.58	1431.96	18000	4.2	1.60	0.7	0.1
	CA15.3	274.87	1977.19	24600	48.0	22.00	16.0	6.2
	MCA	46.97	190.80	2100	15.8	7.45	3.7	0.5

Table 5.2: Statistical Descriptions for Rank-transformed Data

Group	Test	Mean	SD	Max	Q3	Median	Q1	Min
Normal	CEA	108.7	69.2	237.0	170.5	105.0	41.5	11.5
	CA15.3	99.0	62.5	228.0	142.3	99.0	41.5	1.5
	MCA	122.3	69.0	242.0	174.8	129.8	61.0	3.0
High Risk	CEA	150.4	66.6	250.5	221.5	131.5	105.0	41.5
	CA15.3	147.5	77.7	254.5	212.0	167.0	76.5	13.0
	MCA	151.0	81.0	255.0	212.0	188.0	69.0	3.0
Cancer	CEA	168.7	87.9	291.0	252.0	181.5	93.0	11.5
	CA15.3	175.4	84.3	291.0	251.0	193.5	107.0	3.5
	MCA	160.0	90.4	291.0	251.5	161.3	78.6	1.0

5.16, and 5.18. Unfortunately, these observations can not be simply treated as outliers and deleted, because they agree with the other observations in the sense that large values correspond to disease. But the existence of such extreme values will influence the estimations of the population parameters, which is crucial to the implement of the multivariate discriminant functions in ROC analyses for evaluating the aggregate performance of multiple tests. Under such circumstance, as described in Chapter 3, rank transformation could be used so that the observations in each group will agree with one the other in magnitude. The resulting statistical descriptions for the rank-transformed data are shown in Table 5.2.

5.3 The Aggregate Performance of CEA, CA15.3, and MCA

Based on above rank-transformed data, modified ROC analysis is carried out in this section to evaluate the aggregate performance of the three tumor markers.

As indicated in Chapter 2, the area under an ROC curve represents the probability that the random sample X from $\mathbf{D}_-$ is stochastically dominated by Y from $\mathbf{D}_+$. The more separated the samples from two populations, the larger the area under the corresponding ROC curve. In other words, a good diagnostic test will result in large separation between the samples from $\mathbf{D}_-$ and $\mathbf{D}_+$. The separation between the two populatons can be illustrated using a boxplot. In order to evaluate the performance of multiple tests quantitatively, we also present the performance of each single test. For comparing Group N and Group BC and comparing Group HR and Group BC, Figure 5.19, 5.20, and 5.21 show the boxplots of CEA, CA15.3, and MCA, respectively.

It is appeared in these figures that CA15.3 separates the two populations more than that either CEA or MCA does. CEA appears a little better than MCA. These impressions from boxplots are confirmed by the traditional ROC analysis. The results of ROC analysis are given in the first three rows in Table 5.3 and Table 5.4.

Then we consider those three tests as a single test with three measurements, and use RLDF or RQDF, described in Chapter 3 which does not need normality assumption, to obtain the univariate scores. The corresponding boxplots for pairwise comparison are

Table 5.3: ROC Areas of Individual Test on the Patients in Group N and Group BC

Test	A_z	SD_{A_z}	A_t	SD_A
CEA	0.7141	0.0311	0.7022	0.0321
CA15.3	0.7722	0.0283	0.7573	0.0294
MCA	0.6462	0.0332	0.6239	0.0349
RLDF	0.8105	0.0262	0.8153	0.0258
RQDF	0.8208	0.0254	0.8159	0.0258

Note:

A_z and A_t are the area under the fitted curve and trapezoidal (Wilcoxon) area, and the SD_{A_z} and SD_A are the corresponding estimated standard deviations. RLDF and RQDF are combined tests based on linear or quadratic discrimination function, respectively.

Table 5.4: ROC Areas of Individual Test on the Patients in Group HR and Group BC

Test	A_z	SD_{A_z}	A_t	SD_A
CEA	0.5714	0.0446	0.5744	0.0528
CA15.3	0.6259	0.0488	0.6116	0.0510
MCA	0.5625	0.0494	0.5485	0.0539
RLDF	0.6657	0.0471	0.6799	0.0466
RQDF	0.7121	0.0425	0.7096	0.0443

given in Figure 5.22, 5.23, 5.24, and 5.25. Finally, we follow the steps listed in the end of Chapter 3 and calculate the ROC index values. The values for RLDF and RQDF are given in the last two rows in Table 5.3 and Table 5.4, which evaluate the aggregate performance of the three tests.

From the results of ROC analysis in Table 5.3 when we compared the normal (N) people and breast cancer (BC) patients, it is immediately apparent that CA15.3 was the best discriminator of the three tumor markers in this case, and CEA was the second best one. Further, the combined test from the three markers based on either linear discrimination function or quadratic discrimination function was better than any single tumor marker, which means that we can get more information from the combined test than from any single one of the three markers. Similarly, when we tested how well tumor markers discriminated between the high risk (HR) and breast cancer (BC) groups, CA15.3 was still the best discriminator among the three single ones, as indicated in Table

5.4. In both tables the combined tests based on RQDF, according to the areas under the fitted ROC curves, were more or less better than the ones based on RLDF. Since we could not obtain valid random samples from underlying two populations, we did not carry out hypothetical test of equal covariance matrix, and simply gave all those results in Table 5.3 and 5.4. The empirical ROC curves are shown in Figure 5.26 and 5.27.

5.4 Sequential Diagnosis Using CEA, CA15.3 and MCA

Sequential diagnostic procedures are usually important and necessary in clinical diagnosis. For the sake of risk, expenses, time, and etc., we hope to diagnose a cancer patient based on minimal number of tests. Here we can start from the diagnostic test in the sense of minimum misclassification rate, and check with the measurement from that test. If its value is greater than one preset value, A, we may come to the conclusion that the individual is diseased. In contrast, if the value is below another preset value, B, we diagnose that individual to be non-diseased. In such cases, we do not need carry out any more diagnostic tests on that individual. Otherwise, we need use more tests to classify that individual. The above diagnostic procedure is called sequential discrimination procedure, and was formally introduced in Chapter 4.

Following the steps of the modified ROC analysis summarized in Chapter 4, we evaluated the performance of such sequential diagnostic procedure based on the three tumor markers.

Table 5.5: ROC Areas When a Second Test is Added (N and BC)

Test	A_z	SD_{A_z}	A_t	SD_A
CA15.3	0.7722	0.0283	0.7573	0.0294
CA15.3 & CEA	0.7791	0.0279	0.7714	0.0286
CA15.3 & MCA	0.8093	0.0264	0.8089	0.0263

By the results of ROC analysis on the three tumor markers in Table 5.3, globally CA15.3 is the best discriminator in terms of largest area under the ROC curve. MCA is the worst discriminator, and CEA is in the middle. Therefore, we use CA15.3 as the first test. But which test should be used as the second test? We know that this second test is not the globally second best test, but the best test conditional on the first best test. Thus, under the condition that the CA15.3 is used, we introduce CEA and MCA, respectively, and check the performance increase by adding a second test.

Using the RLDF and ROC technique described in Chapter 3, the aggregate performances in terms of ROC areas for adding a second test to CA15.3 are given in Table 5.5.

Although the MCA test is the worst globally among the three tests, from Table 5.5, adding MCA to CA15.3 results in more increase in discrimination performance (A_z increases 6.81%) than that of adding CEA (A_z increases 1.86%). Therefore, CA15.3 is the best test, and MCA is the second best test under the condition that CA15.3 is used. In our sequential diagnosis, we will use CA15.3 measurements first, then introduce MCA

Table 5.6: ROC Areas of Sequential Tests on the Patients in Group N and Group BC

α	β	A_z	SD_{A_z}	A_t	SD_A	Relative Efficiency
0.01	0.01	0.8105	0.0262	0.8153	0.0258	0.00%
0.10	0.20	0.8045	0.0266	0.8094	0.0262	9.69%
0.10	0.30	0.7999	0.0269	0.8008	0.0268	19.51%

to CA15.3 for those unclassified by CA15.3 only. For those which are not classified by the CA15.3 and MCA, CAE will be included at last.

As mentioned in the Chapter 4, the discrimination bounds, A and B, which depend upon type I and type II errors, are crucial to the performance of the sequential discrimination procedure. Choosing different values of type I and type II errors, we first applied sequential diagnostic procedure to the patients in the Group N and Group BC, and used modified ROC procedure to evaluate the resulting performances. The results are shown in Table 5.6.

As we increased the type I and type II errors, the interval between the two bounds was getting smaller, and the area under the ROC curve was getting smaller. Thus, accuracy of the sequential diagnostic procedure decreased. Fortunately, such performance decreases were very small, and were not sensible to the discrimination bounds. On the other hand, as the interval between the two bounds was getting smaller, the number of tests needed to classify those patients reduced. For example, when $\alpha = 0.10$ and $\beta = 0.30$, the trapezoidal ROC area reduced only about 2% compared with the simultaneous

Table 5.7: ROC Areas of Sequential Tests on the Patients in Group HR and Group BC

α	β	A_z	SD_{A_z}	A_t	SD_A	Relative Efficiency
0.05	0.10	0.6649	0.0471	0.6799	0.0466	0%
0.35	0.35	0.6646	0.0501	0.6799	0.0466	4.19%
0.40	0.40	0.6569	0.0474	0.6391	0.0494	31.24%

procedure, but about 20% tests were saved in that case.

Then, we did the same ROC analysis on the patients in the Group HR and Group BC, similar conclusions could be derived and results of ROC analysis are given in Table 5.7.

From the above analyses, it is immediately apparent that the sequential diagnostic procedure is a potentially efficient and economic diagnostic procedure in clinical diagnosis. Compared with the simultaneous diagnostic procedure, a small decrease in accuracy results in significant test savings.

Chapter 6

Conclusions

In this paper, two ROC-based techniques have been developed to evaluate the aggregate performance of several diagnostic tests, one is for evaluating simultaneous multiple diagnostic tests, and the other is for sequential tests.

Quite often, there are several tests available in diagnosis of a particular disease. When these univariate tests are applied to the same patients in the group, the tests can be treated as one test with multivariate measurements. In this case, Fisher's linear discrimination functions or their quadratic forms could be used to combine such multivariate measurements as univariate measurements, upon which the traditional ROC analysis can be applied. The indices derived from the ROC analysis will represent the aggregate performance of the multiple diagnostic tests involved.

On the other hand, in consideration of the risk, expenses, and time consuming of the diagnostic tests, we prefer to use as few tests as possible to classify an individual as belonging to diseased or non-diseased population. In such situation, sequential diagnostic procedures are strongly recommended. From both the simulation study of comparing the sequential discrimination procedures and the simultaneous ones, and the data analyses on the breast cancer data, the sequential procedures are nearly as good as the corresponding simultaneous ones in accuracy, but the number of tests needed for classifying individuals reduces significantly. By the method of taking ranks described in Chapter 4, sequential diagnostic test scores can be transformed into corresponding univariate rank scores, upon which the traditional ROC analysis can be applied. The indices derived from the ROC analysis will represent the aggregate performance of such sequential diagnostic procedures.

In addition, the importance of rank transformation in ROC analysis was emphasized in this paper. First, the indices under the ROC curve remain the same no matter rank transformation is applied to the original data or not. Second, after the rank transformation, outliers (extremely large values in the diseased population, or extremely small values in the non-diseased population) will agree with the other observations in the same population. Finally, when the data are not normally distributed, multivariate discrimination methods based on rank transformation are superior to the ones without rank transformation.

It should be mentioned that the use of ROC techniques for evaluating the performance of clinical diagnosis has grown considerably in the past years. Methods for statistical inference have been derived for most situations. However, a number of questions remain to be answered. For example, what types of questions can or cannot be answered by the ROC studies? Can ROC analysis be improved to the conduct of multicenter imaging trials? What is the best way to evaluate the quantitative tests, and what is the best way to compare quantitative tests with qualitative ones? Therefore, ROC-based techniques need to be developed by our further studies.

Bibliography

[1] Silver, H. K. B., Archibald, B. L., Ragaz, J., and Coldman, A. J., Relative Operating Characteristic Analysis and Group Modelling for Tumor Markers; Comparison of CA 15.3, Carcinoembryonic Antigen, and Mucin-like Carcinoma-associated Antigen in Breast Carcinoma, Cancer Research, 51, p1904-1909, 1991.

[2] Green, D. M. and Swets, J. A., Signal Detection Theory and Psychophysics. John Wiley & Sons, New York, 1966.

[3] Swets, J. A. and Pickett, R. M., Evaluation of Diagnostic Systems: Methods from Signal Detection Theory, Academic Press, New York, 1982.

[4] Morgan, B. J., The Uniform Distribution in Signal Detection Theory, British Journal of Statistical Psychology, 29, p81-88, 1976.

[5] Dorfman, D. D. and Alf, E., Maximum Likelihood Estimation of Parameters of Signal Detection Theory–A Direct Solution, Psychometrika, 33, p117-124, 1968.

[6] Dorfman, D. D. and Alf, E., Maximum Likelihood Estimation of Parameters of Signal Detection Theory and Determination of Confidence Intervals: Rating Method Data, Journal of Mathematical Psychology, 6, p487-496, 1969.

[7] Ogilvie, J. C. and Creelman, C. D., Maximum Likelihood Estimation of ROC Curve Parameters, Journal of Mathematical Psychology, 5, p377-391, 1968.

[8] Grey, D. R. and Morgan, B. J., Some Aspects of ROC Curve Fitting: Normal and Logistic Models, Journal of Mathematical Psychology, 9, p128-139, 1972.

[9] Birdsall, T. G., The Theory of Signal Detectability: ROC Curves and Their Characters, unpublished dissertation, Department of Electrical and Computer Engineering, The University of Michigan, Ann Arbor, 1973.

[10] Jaraiedi, M. and Herrin, G. D., Effect of Human Inspector Error on Sample Plan Design, Proceedings, 1985 Fall Industrial Engineering Conference, Chicago 1985, p436-439.

[11] Greenhouse, S. W. and Mantel, N., The Evaluation of Diagnostic Tests, Biometrics, 6, p399-412, 1950.

[12] Linnett, K., Comparison of Quantitative Diagnostic Tests: Type I Error, Power, and Sample Size, Statistics in Medicine, 6, p147-158, 1987.

[13] Bamber, D., The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph, Journal of Mathematical Psychology, 12, p387-415, 1975.

[14] Hanley, J. A. and McNeil, B. J., The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve, Radiology, 143, p29-36, 1982.

[15] Hanley, J. A. and McNeil, B. J., A Method of Comparing the Areas Under Receiver Operating Characteristic Curves Derived from the Same Cases, Radiology, 148, p839-843, 1983.

[16] Goddard, M. J. and Hinbery, I., ROC Curves for Non-normal Data, Presented paper, joint statistical meetings, Chicago, 1986.

[17] Delong, E. R., Delong, D. M., and Clarke-Oearson, D. L., Comparing the Area Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, Biometrics, 44, p837-845, 1988.

[18] Wieand, S. et al, A Family of Nonparametric Statistics for Comparing Diagnostic Tests with Paired or Unpaired Data, Biometrika, 76, p585-592, 1989.

[19] Revesz, G., Kundel, H. L., and Bonitatibus,M., The Effect of Verification on the Assessment of Imaging Techniques, Investigative Radiology, 18, p194, 1983.

[20] Henkelman, R. M., Kay, I. B., and Bronskill, M. L., Receiver Operator Characteristic (ROC) Analysis without Truth, unpublished document, Department of Medical Biopgysics, University of Toronto, Toronto, 1986.

[21] Hanley, J. A., Receiver Operating Characteristic (ROC) Methodology: The State of Art, Critical Reviews in Diagnostic Imaging, 29, Issue 3, p307-335, 1989.

[22] Fisher, R. A., The Use of Multiple Measurements in Taxonomic Problems, Annals of Eugenics, 7, p179-188, 1936.

[23] Fisher, R. A., The Statistical Utilization of Multiple Measurements, Annals of Eugenics, 8, p376-386, 1938.

[24] Johnson, R. A. and Wichern, D. W., Applied Multivariate Statistical Analysis, 2nd edition, Prentice Hall, 1988.

[25] Seber, G. A. F., Multivariate Observations, John Wiley & Sons, 1984.

[26] Lehmann, E. L., Nonparametrics : statistical methods based on ranks, Holden-Day, 1975.

[27] Dawid, A. P., Properties of Diagnostic Data Distributions, Biometrics, 32, p647-658, 1976.

[28] Randles, R. H., Broffitt, J. D., and Hogg,R.V., Discriminant Analysis Based on Ranks, Journal of Am. Stat. Assoc., 73, p379-384, 1978.

[29] Randles, R. H., Broffitt, J. D., and Hogg,R.V., Generalized Linear and Quadratic Discriminant Functions Using Robust Estimates, Journal of Am. Stat. Assoc., 73, p564-568, 1978.

[30] Beckman, R. J. and Johnson, M. E., A Ranking Procedure for Partial Discriminant Analysis, J. of Am. Stat. Assoc., 76, p671-675, 1981.

[31] Conover, W. J. and Iman, R. L., The Rank Transformation as a Method of Discrimination with Some Examples, Communications in Statistics, A9(5), p465-487, 1980.

[32] Hora, S. C., Sequential Discrimination, Commun. Statist. Theor. Meth., A9(9), p905-916, 1980.

[33] Lachenbruch, P. A., Discriminant Analysis, Hafner: New York, 1975.

[34] Gordon, L. and Olshen, R. A., Asymptotically Efficient Solution to the Classification Problem, Ann. Stat., 6, p515-533, 1978.

[35] Glick, N., Sample-based Classification Procedures Derived from Density Estimators, J. Am. Stat. Assoc., 67, p116-122, 1972.

[36] Kendall, M. G. and Stuart, A.,The Advanced Theory of Statistics, Vol.III. Griffin: London, 1966.

[37] Kendall, M. G., Multivariate Analysis. Griffin: London, 1975.

[38] Beard, D. B. and Haskell, C. M., Carcinoembryonic Antigen in Breast Cancer, Am. J. Med., 80, p241-245, 1986.

[39] Tondini, C., Hayes, D. F., and Kufe, D. W., Circulating Tumor Markers in Breast Cancer, Hematol. Oncol. Clin. North Am., 3, p653-674, 1989.

Figure 1.1 One Typical ROC Curve

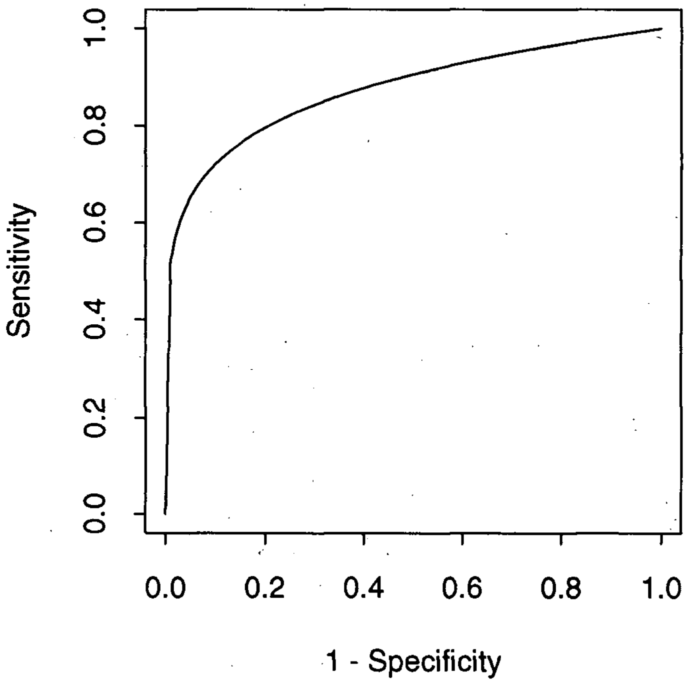


Figure 1.2 Comparing Two ROC Curves

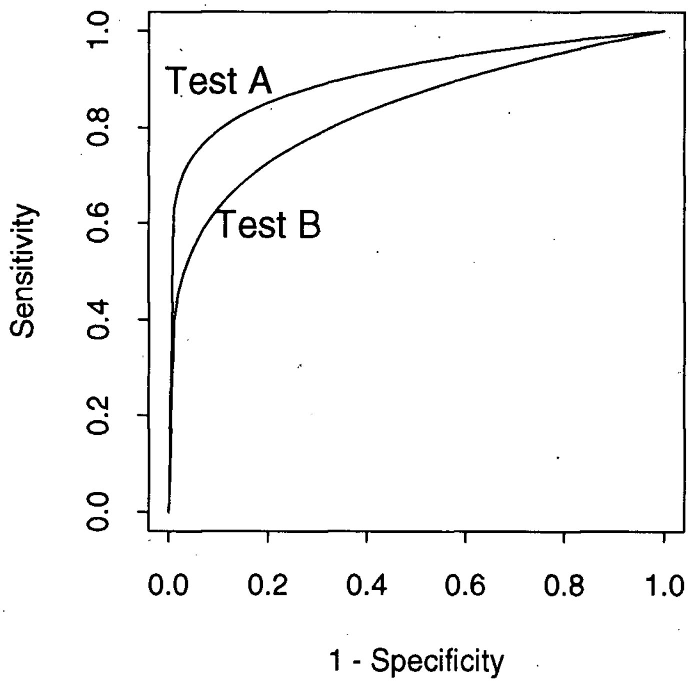


Figure 2.1 Empirical ROC Points

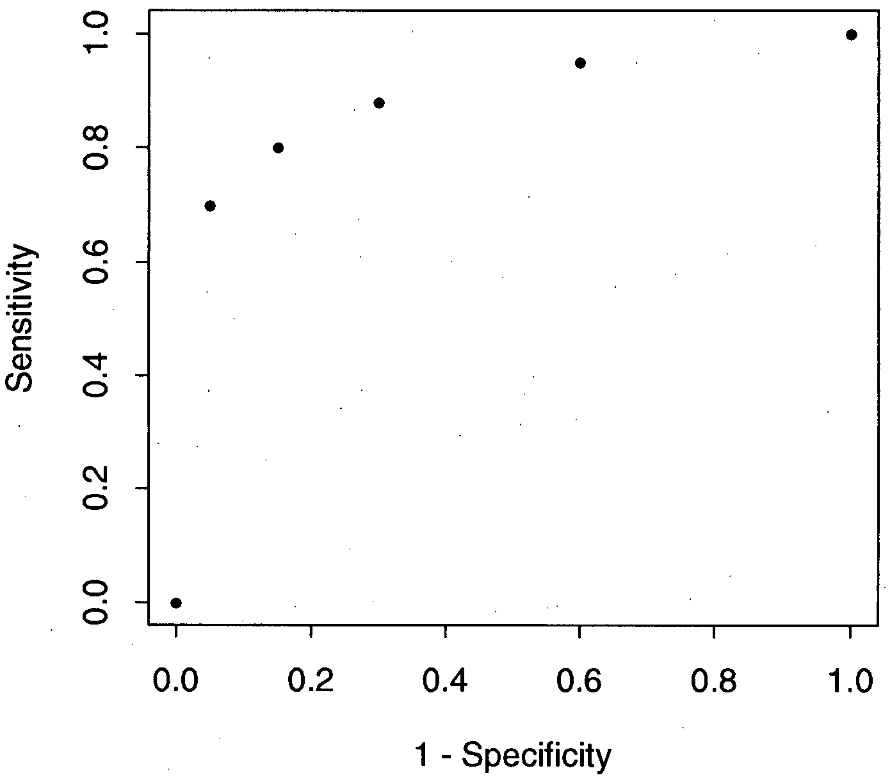


Figure 2.2 Sub-area under ROC Curve

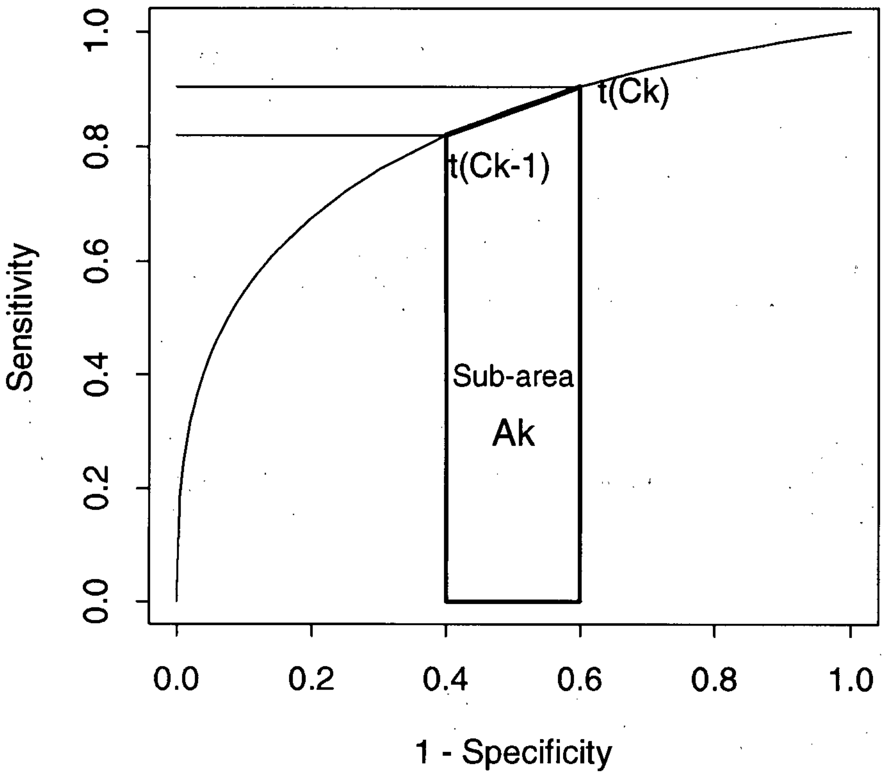


Figure 3.1. Cutlines in Bivariate Case

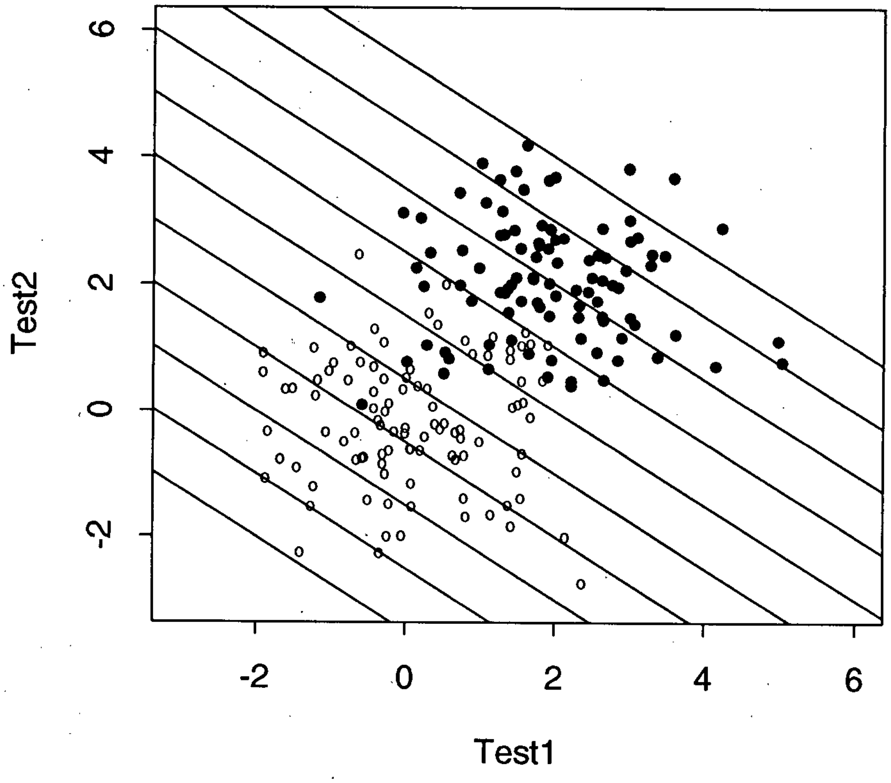


Fig.5.1 Q-Q Plot For Group N

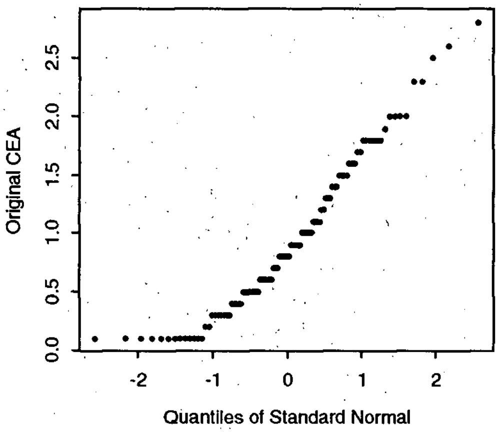


Fig.5.2 Histogram For Group N

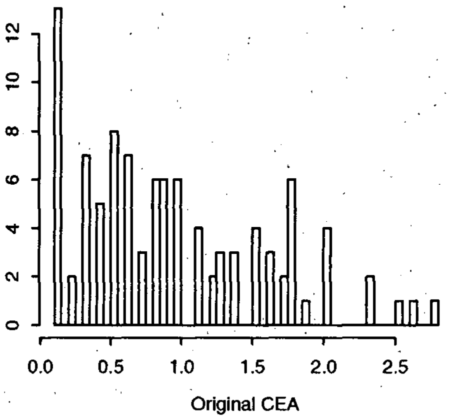


Fig.5.3 Q-Q Plot For Group N

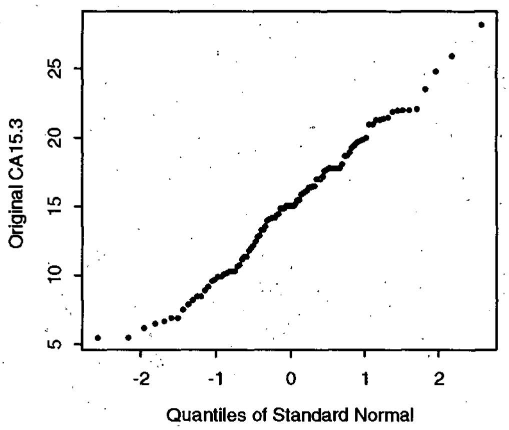


Fig.5.4 Histogram For Group N

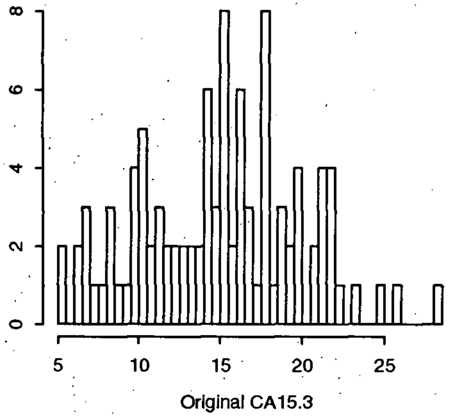


Fig.5.5 Q-Q Plot For Group N

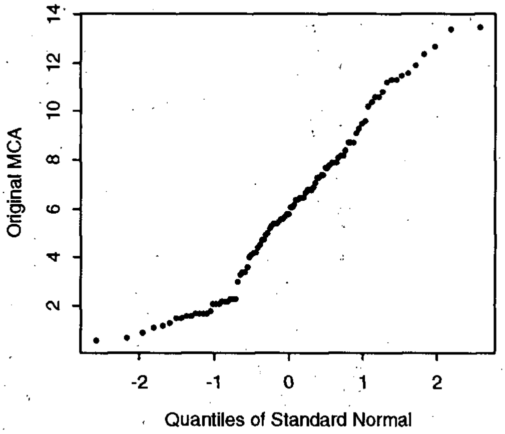


Fig.5.6 Histogram For Group N

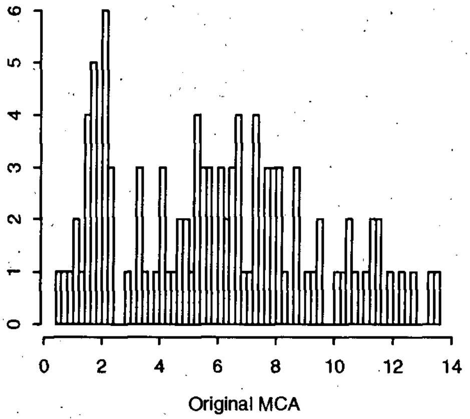


Fig.5.7 Q-Q Plot For Group HR

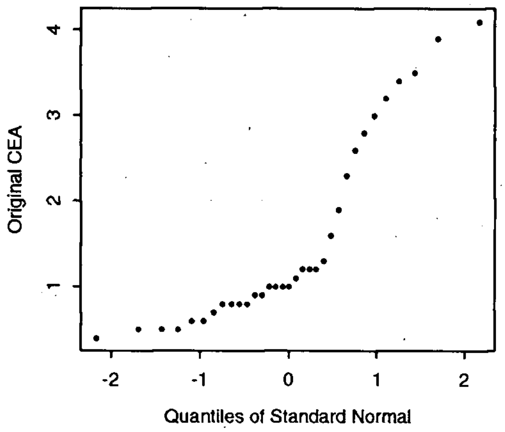


Fig.5.8 Histogram For Group HR

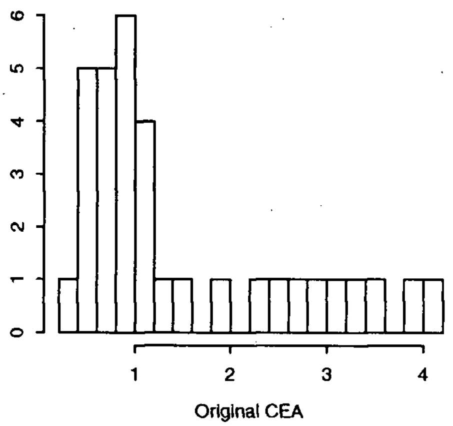


Fig.5.9 Q-Q Plot For Group HR

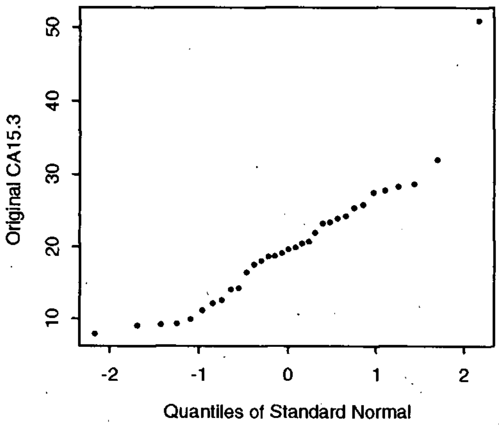


Fig.5.10 Histogram For Group HR

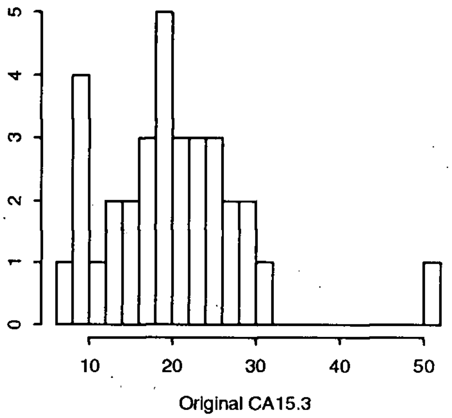


Fig.5.11 Q-Q Plot For Group HR

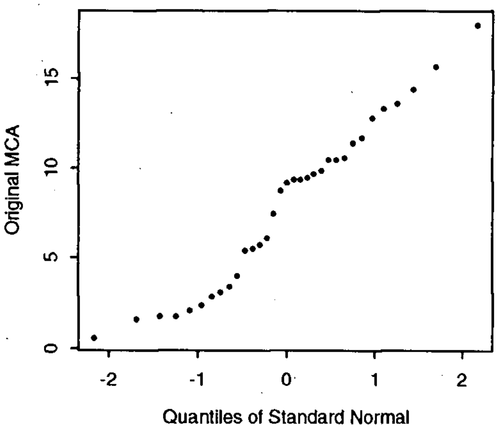


Fig.5.12 Histogram For Group HR

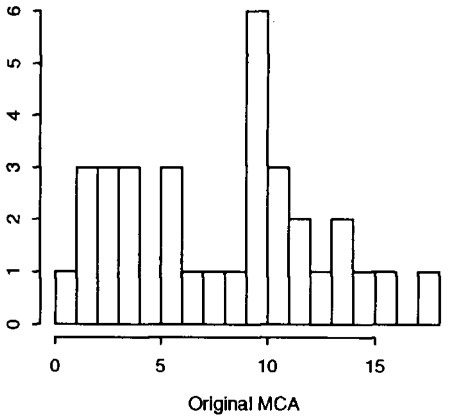


Fig.5.13 Q-Q Plot For Group BC

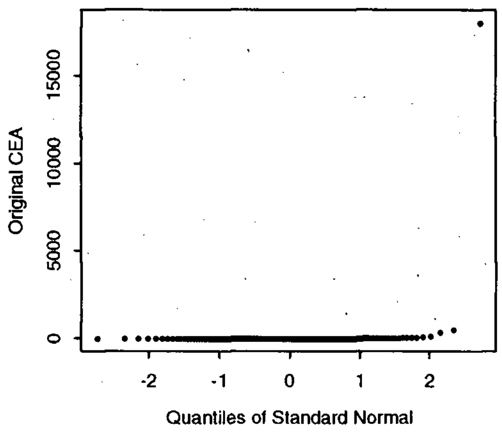


Fig.5.14 Histogram For Group BC

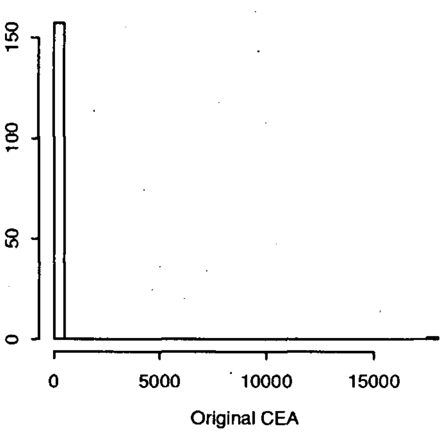


Fig.5.15 Q-Q Plot For Group BC

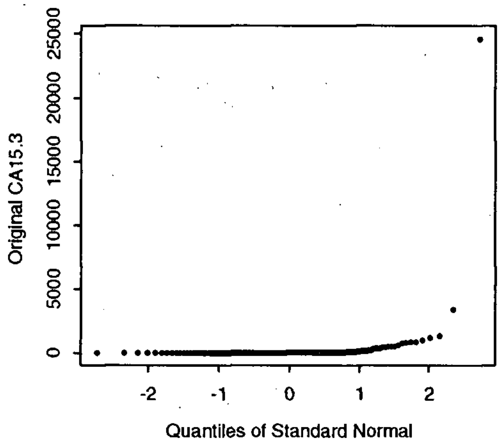


Fig.5.16 Histogram For Group BC

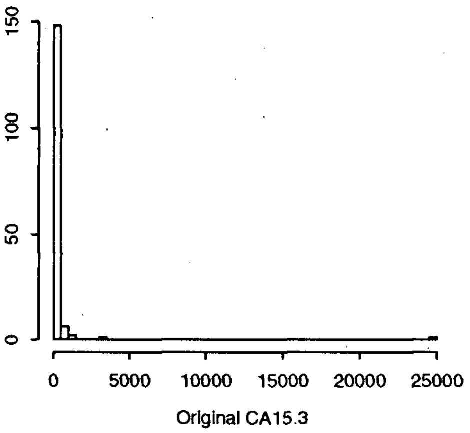


Fig.5.17 Q-Q Plot For Group BC

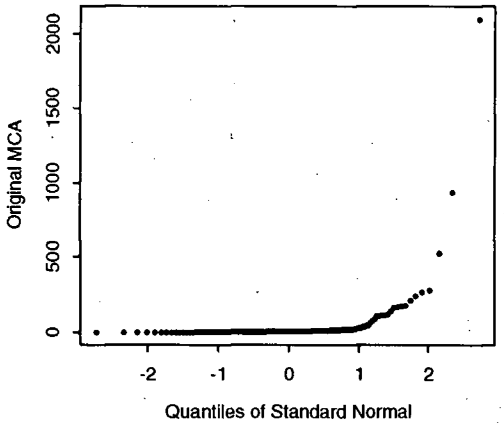


Fig.5.18 Histogram For Group BC

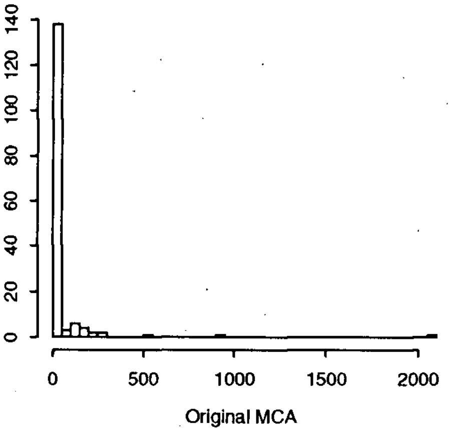


Figure 5.19. Boxplot of the Ranked CEA Measurements

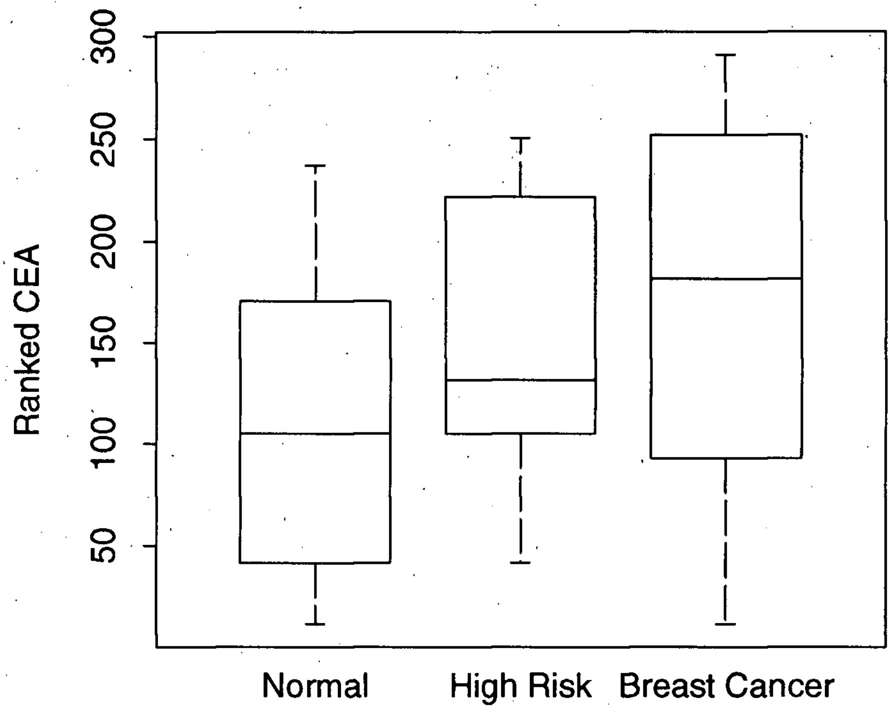


Figure 5.20 Boxplot of the Ranked CA15.3 Measurements

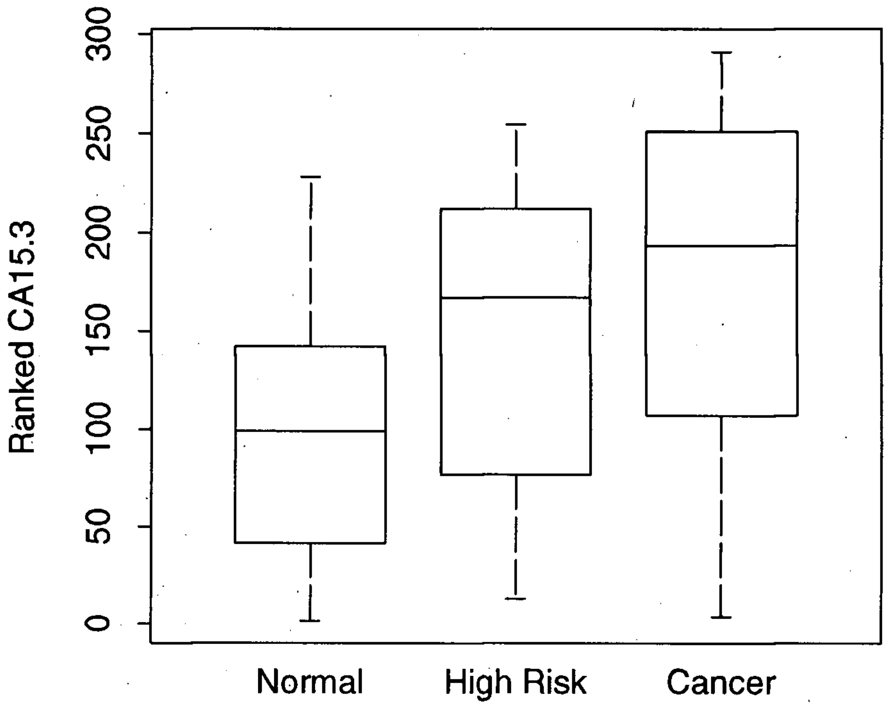


Figure 5.21 Boxplot of the Ranked MCA Measurements

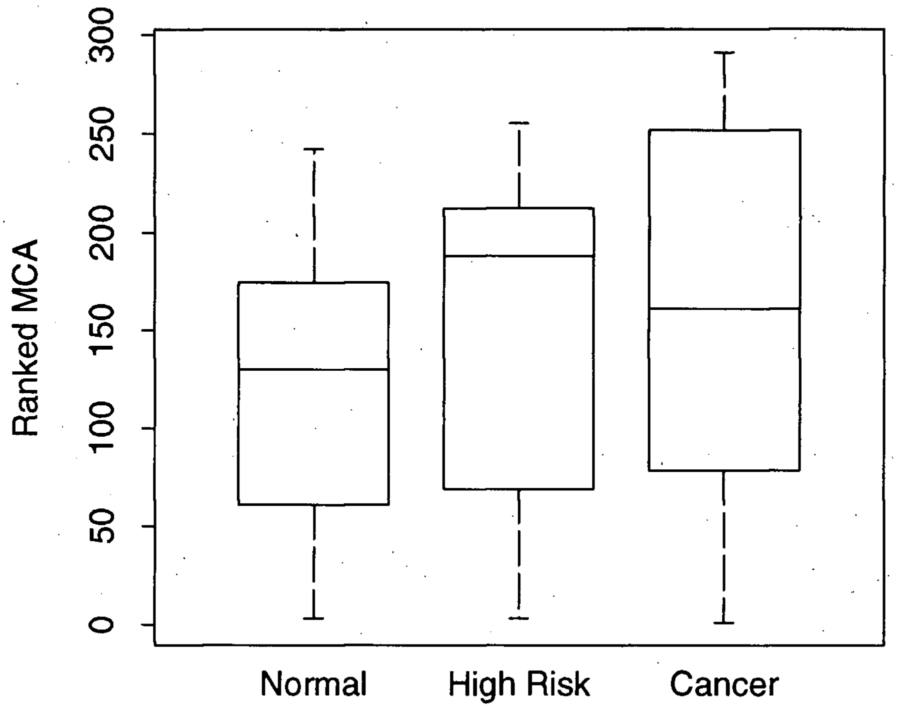


Fig.5.22 Boxplot of RLDF in N and BC

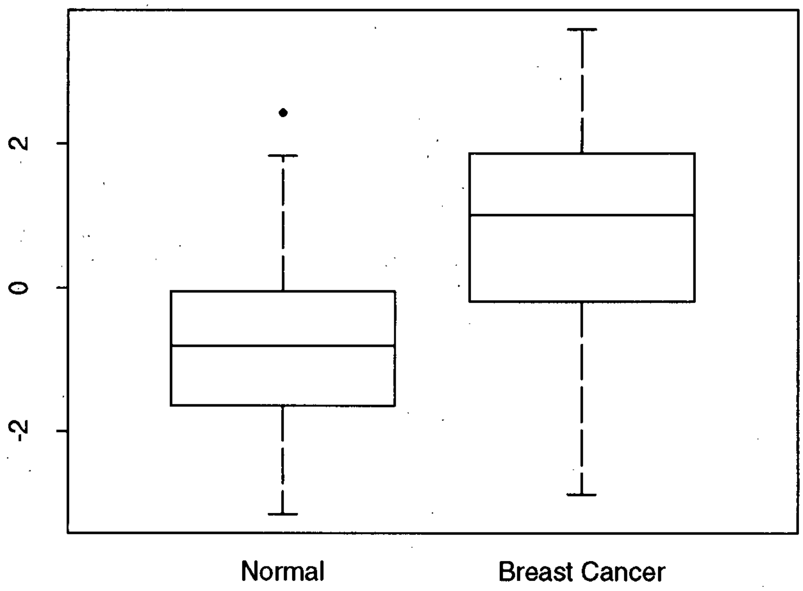


Fig.5.23 Boxplot of RQDF in N and BC

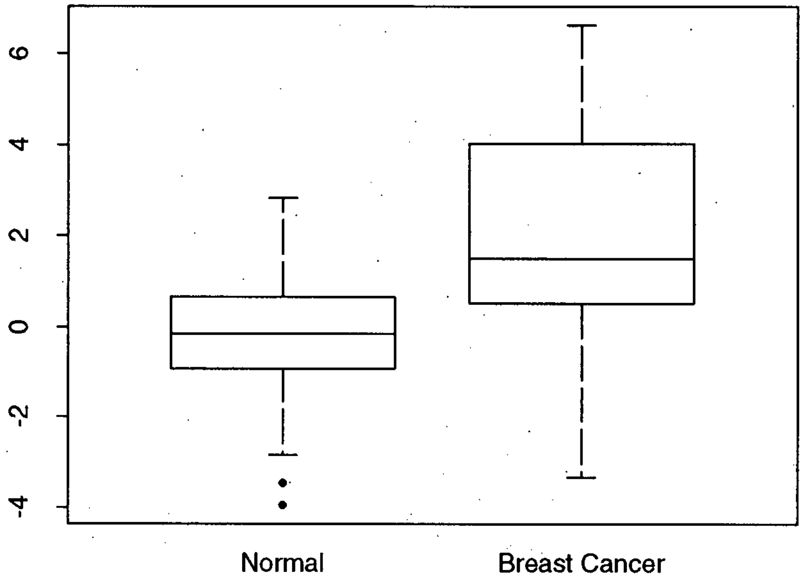


Fig.5.24 Boxplot of RLDF in HR and BC

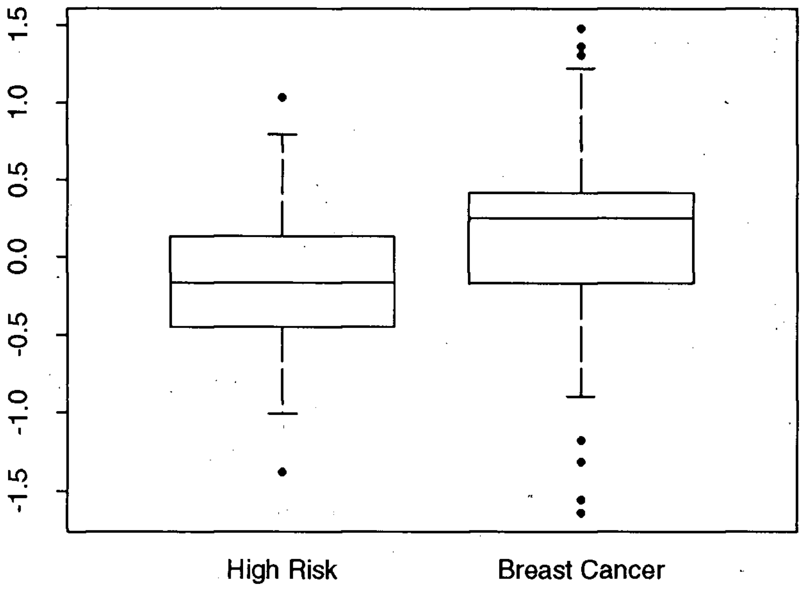


Fig.5.25 Boxplot of RQDF in HR and BC

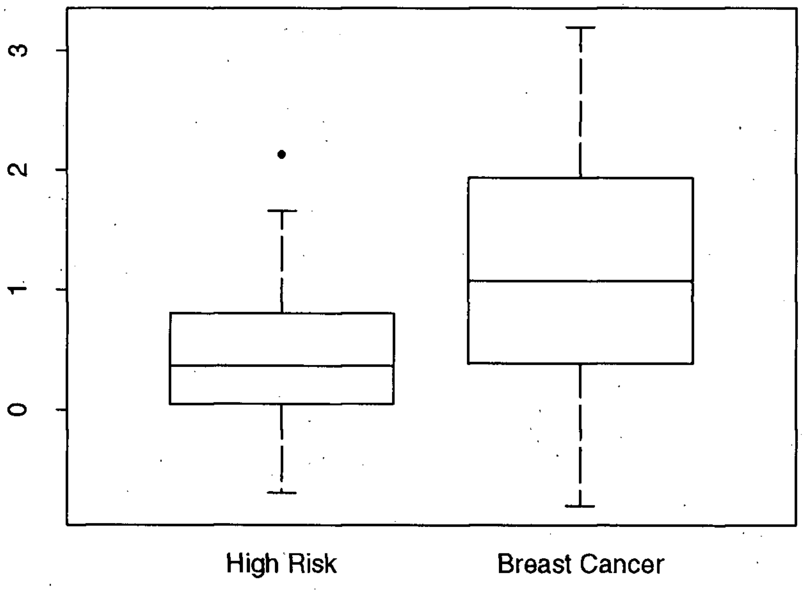


Figure 5.26 ROC Curves (Group N and Group BC)

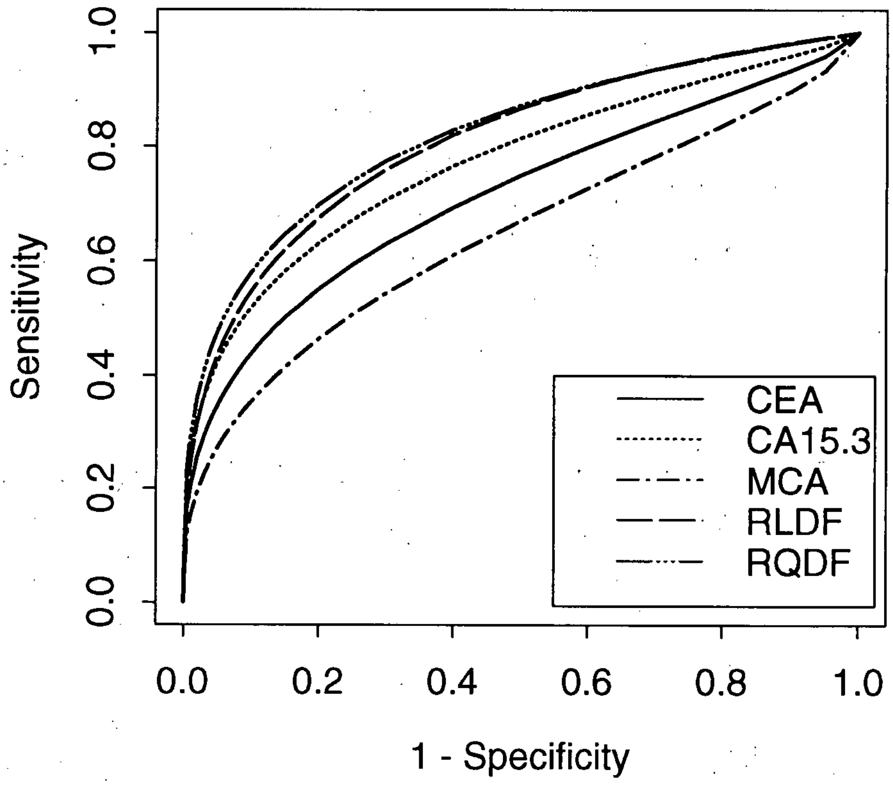
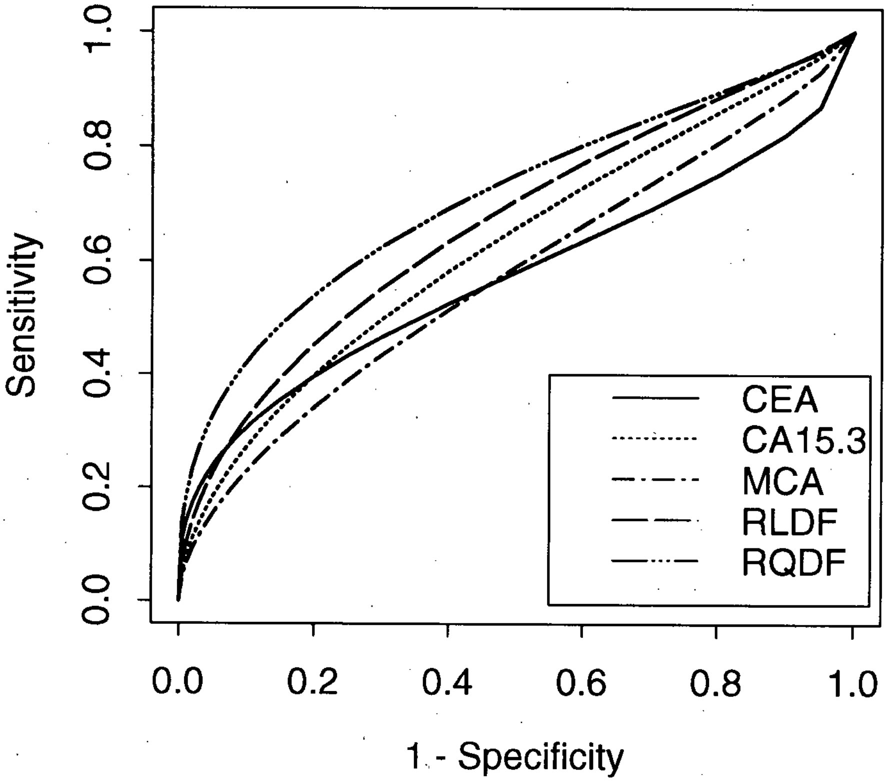


Figure 5.27 ROC Curves (Group HR and Group BC)


Sensitivity
1 - Specificity
CEA
CA15.3
MCA
RLDF
RQDF