

# **Receiver Operating Characteristic analysis for Intelligent Medical Systems – a new approach for finding non-parametric confidence intervals.**

**Julian Tilbury<sup>1</sup>, Peter Van Eetvelt<sup>2</sup>, Jonathan Garibaldi<sup>1</sup>, John Curnow<sup>1,3</sup>, Emmanuel Ifeachor<sup>1</sup>**

<sup>1</sup> SMART Medical Systems Research Group

School of Electronic, Communication and Electrical Engineering,

University of Plymouth,

Drake Circus

Plymouth

PL1 8AA

<sup>2</sup> School of Electronic, Communication and Electrical Engineering,

University of Plymouth,

Drake Circus

Plymouth

PL1 8AA

<sup>3</sup> Department of Medical Physics

Derriford Hospital

Derriford

Plymouth

PL8 6DH

## **Receiver Operating Characteristic analysis for Intelligent Medical Systems – a new approach for finding non-parametric confidence intervals.**

**Intelligent systems are increasing being deployed in medicine and healthcare, but there is a need for a robust and objective methodology for evaluating such systems. Potentially, Receiver Operating Characteristic (ROC) analysis could form a basis for objective evaluation of intelligent medical systems. However, it has several weaknesses when applied to the types of data used to evaluate intelligent medical systems. Firstly, small data sets are often used, which are unsatisfactory with existing methods. Secondly, many existing ROC methods use parametric assumptions which may not always be valid for the test cases selected. Thirdly, system evaluations are often more concerned with particular, clinically meaningful, points on the curve, than on global indexes such as the area under the curve which is commonly used.**

**A novel, robust and accurate method is proposed, derived from first principles, that calculates the probability distribution for each point on a ROC curve for any given sample size. Confidence intervals are produced as contours on the probability distribution. The theoretical work has been validated by Monte Carlo simulations. It has also been applied to two real-world examples of ROC analysis taken from the literature (classification of mammograms and differential diagnosis of pancreatic diseases), to investigate the confidence surfaces produced for real cases, and to illustrate how analysis of system performance can be enhanced. We illustrate the impact of sample size on system performance from analysis of ROC probability distributions and 95% confidence boundaries. This work establishes an important new method for generating probability distributions, and provides an accurate and robust method of producing confidence intervals for ROC curves for small sample sizes typical of intelligent medical systems. It is conjectured that, potentially, the method could be extended to determine risks associated with the deployment of intelligent medical systems in clinical practice.**

# 1.Introduction

Intelligent systems are increasingly being deployed in medicine and healthcare, to practically aid the busy clinician and to improve the quality of patient care [1][2][3][4][5][6][7]. The need for an objective methodology for evaluating such systems is widely recognised [2][4][8][9][10][11][12]. In medicine and healthcare where safety is critical, this is important if techniques such as medical expert systems and neural systems are to be widely accepted in clinical practice.

The work described here arose from a critical investigation into the potential role of ROC analysis as a basis for objective evaluation of intelligent medical systems. The work forms part of an initiative to develop a theoretical framework for an objective methodology for evaluating intelligent systems in this safety critical area.

ROC analysis is now common in medicine and healthcare [4][13][14], particularly in radiology [8][10][15] where it is used to quantify the accuracy of diagnostic tests [10][16][17]. The performance of an ‘expert’, human or machine, can be represented objectively by ROC curves [10][18]. Such curves show, for example, the trade off between a diagnostic test correctly identifying diseased patients as diseased, rather than healthy, versus correctly identifying healthy patients as healthy, rather than diseased. Many intelligent medical systems carry out this type of task, whether actually classified as ‘diagnostic’ [19], or not, e.g. ‘prognostic’ [20].

To serve as a basis for objective evaluation of intelligent medical systems, ROC analysis will need to be extended to address a number of limitations. In practice, ROC analysis can be either parametric or non-parametric. In parametric analysis, the underlying population distributions of the diseased and healthy patients are often assumed to be normal, although other types of distributions, such as gamma and negative exponential, are sometimes used. In contrast, non-parametric analysis does not make any assumptions about the form of the underlying population distributions. When an intelligent medical system is tested against human experts a selection of a small number of cases is picked, often by an independent expert. If the underlying population distributions of diseased and healthy patients gave a binormal ROC curve, a random sample of patients from the population would preserve these distributions. However, if the cases are picked by an independent expert, particularly one who is deliberately biased in favour of picking difficult cases, the distribution may not be preserved. Thus, non-parametric methods may be more appropriate for intelligent medical system testing.

ROC curves are a complex representation of performance and for convenience, the area under the ROC curve (AUC) is used as a single index of accuracy. Intuitively, the area under the curve can be viewed as the probability of the test being able to correctly identify a healthy patient from a pair of one healthy and one diseased patient drawn at random from each respective population [21]. As human experts are available for a limited time, the number of cases used for evaluation is often small. Existing methods of ROC analysis are often unsatisfactory with small numbers

of cases, especially if the area under the curve is high. Under these circumstances their confidence intervals can be erroneous, or the algorithms can even fail to produce any result. Obuchowski and Lieber [22] compared 11 methods of parametric and non-parametric analysis and could not find a single best alternative for constructing the confidence interval when the sample size was small.

Further, the intuitive definition of the area under a ROC curve is not directly related to the task an intelligent medical system may be required to perform. For example, an intelligent medical system may be required to make a decision about the disease status of a single patient of unknown health. The area under the ROC curve gives the probability of correctly identifying a healthy patient from a pair where one is known to be diseased and the other one is known to be health, which is of limited practical value in the clinical situation.

The aim of the work reported here was therefore to find a non-parametric method of ROC analysis that would be robust and accurate over any number in the sample, particularly small numbers, and could be applied to particular points on the ROC curve of clinical interest. To achieve this, the underlying probability theory was re-examined, and a novel method of producing a probability distribution over the whole ROC graph, for each point on the curve, was derived. The theoretical work has been validated by Monte Carlo simulations and it has also been applied to two real-world examples taken from the literature. While many Monte Carlo simulations assume a fixed population ROC curve and examine the distribution of samples generated randomly from the fixed population, our simulation generated random samples from random population ROC curves.

## 2.Receiver Operating Characteristic (ROC) Curves

Take the situation where there are two types of events, signal (diseased), and noise (healthy), and it is hoped to distinguish between them by measuring a characteristic property of these events, on an ordinal, interval or ratio scale. Fig 1. gives a hypothetical example of the relative frequency with which two types of events give different values of the measured property. To distinguish between the types of event a threshold is chosen such that events with a measurement lower than the threshold are labelled as noise, and events with a measurement greater than the threshold are labelled as signal. Since the two distributions overlap, no threshold value will completely separate them. Table 1 shows the 2 by 2 contingency table of the actual type of a event, against its test classification according to the threshold. This table assumes there is a standard by which the actual type of the event is known.

The test can then be characterised by two ratios:

$$\text{Hit Rate} = \frac{\text{True} + \text{ve}}{\text{True} + \text{ve} + \text{False} - \text{ve}} = \frac{b_0}{b_0 + b_1}$$

$$\text{False Alarm Rate} = \frac{\text{False} + \text{ve}}{\text{False} + \text{ve} + \text{True} - \text{ve}} = \frac{a_0}{a_0 + a_1}$$

If multiple thresholds are used, for example, to categorise events into ‘definitely signal’ ‘possibly signal’, ‘possibly noise’, and ‘definitely noise’, the contingency table can be expanded to a 2 by  $n+1$  table, where  $n$  is the number of thresholds; for example, for a 2 by 4 table, as shown in table 2.

From the table,  $n$  pairs of Hit Rate and False Alarm Rate can be calculated. The example, table 2 gives the following three pairs:

$$\begin{aligned} \text{Hit Rate}_0 &= \frac{b_0}{b_0 + b_1 + b_2 + b_3} & \text{False Alarm Rate}_0 &= \frac{a_0}{a_0 + a_1 + a_2 + a_3} \\ \text{Hit Rate}_1 &= \frac{b_0 + b_1}{b_0 + b_1 + b_2 + b_3} & \text{False Alarm Rate}_1 &= \frac{a_0 + a_1}{a_0 + a_1 + a_2 + a_3} \\ \text{Hit Rate}_2 &= \frac{b_0 + b_1 + b_2}{b_0 + b_1 + b_2 + b_3} & \text{False Alarm Rate}_2 &= \frac{a_0 + a_1 + a_2}{a_0 + a_1 + a_2 + a_3} \end{aligned}$$

With a sufficiently large number of thresholds changing in small discrete steps, a plot of Hit Rate (along the y axis) against False Alarm Rate (along the x axis) for each threshold gives a receiver operating characteristic (ROC) curve. A typically shaped curve for a multi threshold plot is given in Fig. 2.

The curve thus shows the trade off between correctly detecting a signal and mistaking noise for a signal. If the two underlying population distributions are well separated the curve will immediately rise to the top left corner (0.0, 1.0), and then proceed horizontally, while if the distributions tend to overlap, such that they cannot be distinguished by the measurement, the curve will approach the diagonal (0.0, 0.0 to 1.0, 1.0).

If a ROC curve is plotted for a sample of cases the curve will only be an estimate of the actual ROC curve of the population. Confidence intervals therefore need to be given. Many methods of producing confidence intervals for the area under the curve, both parametrically [23] and non-parametrically [17][18][21][22][24][25], and for each individual point on a non-parametric curve [26][27][28] have been given. However, for small samples sizes, typical in intelligent medical system testing, none of these methods is ideal [22]. In the case of binormal parametric models the methods can fail to produce any results at all when the diseased and healthy samples do not overlap. This can happen particularly with small samples when the population AUC approaches 1.0 [29].

### 3.A New Approach to ROC Analysis

The method proposed here assumes a non-parametric model that is robust and accurate over all data sets. It returns to the underlying probability theory to construct a probability distribution over the entire ROC graph for each point of the curve. The method is based on asking the following question for every possible point on the surface of the ROC graph:

*'If this point represents the true Hit Rate and False Alarm Rate of the population, what would be the probability of getting the sample actually obtained?'*

If that question can be answered for every point on the graph, the relative probability of every point can then be calculated, and normalised such that the total relative probability sums to 1. This generates the normalised relative probability surface for the true Hit Rate and False Alarm Rate. By dividing the surface into a fine grid, and integrating the expression for the relative probability of every point over each square of the grid, the surface can then be presented as a 3D mesh, or contour lines can be drawn to enclose an arbitrary percentage of the probability, e.g. 95% of the probability, which gives the 95% confidence interval for the location of the true Hit Rate and False Alarm Rate.

First consider the situation where  $x$  is the False Alarm Rate of the population, given as a probability between 0 and 1. If  $a_0$  is the number of false positives, and  $a_1$  the number of true negatives in the sample, then  $P$ , the probability of obtaining  $a_0$  False Alarms in  $a_0 + a_1$  healthy cases when the probability of a False Alarm is  $x$ , is given by:

$$P_{xa_0a_1} = \binom{a_0 + a_1}{a_0} x^{a_0} (1 - x)^{a_1}$$

The ratio between the probability at two arbitrary points  $x_i$  and  $x_j$  is given by:

$$\frac{P_{x_i a_0 a_1}}{P_{x_j a_0 a_1}} = \frac{\binom{a_0 + a_1}{a_0} x_i^{a_0} (1 - x_i)^{a_1}}{\binom{a_0 + a_1}{a_0} x_j^{a_0} (1 - x_j)^{a_1}} = \frac{x_i^{a_0} (1 - x_i)^{a_1}}{x_j^{a_0} (1 - x_j)^{a_1}}$$

The relative probability for the Hit Rate can be derived in the same way.

Now consider the full situation where  $y$  is the Hit Rate of the population, given as a probability between 0 and 1,  $x$  is the False Alarm Rate of the population, given as a probability between 0 and 1, and  $f$  is the frequency of disease events in the population, again given as a probability between 0 and 1. If  $b_0$  is the number of true positives,  $a_0$  the number of false positives,  $b_1$  the number of false negatives and  $a_1$  the number of true negatives in the sample, then  $P$ , the probability of a ROC point being at the location  $x, y$ , is given by the probability of obtaining  $b_0 + b_1$  diseased cases in  $a_0 + a_1 + b_0 + b_1$  cases when the probability of disease is  $f$ , multiplied by the probability of obtaining  $a_0$  False Alarms in  $a_0 + a_1$  healthy cases when the probability of a False Alarm is  $x$ , multiplied by the probability of obtaining  $b_0$  Hits in  $b_0 + b_1$  diseased cases when the probability of a Hit is  $y$ :

$$P_{xyf a_0 b_1 a_1} = \binom{a_0 + a_1 + b_0 + b_1}{b_0 + b_1} f^{b_0 + b_1} (1 - f)^{a_0 + a_1} \binom{a_0 + a_1}{a_0} x^{a_0} (1 - x)^{a_1} \binom{b_0 + b_1}{b_0} y^{b_0} (1 - y)^{b_1}$$

The ratio between the probability at the point  $x_i, y_i$ , and an arbitrary fixed point  $x_j, y_j$  is given by:

$$\frac{P_{x_i y_j f b_0 a_0 b_1 a_1}}{P_{x_j y_j f b_0 a_0 b_1 a_1}} = \frac{\binom{a_0 + a_1 + b_0 + b_1}{b_0 + b_1} f^{b_0 + b_1} (1 - f)^{a_0 + a_1} \binom{a_0 + a_1}{a_0} x_i^{a_0} (1 - x_i)^{a_1} \binom{b_0 + b_1}{b_0} y_i^{b_0} (1 - y_i)^{b_1}}{\binom{a_0 + a_1 + b_0 + b_1}{b_0 + b_1} f^{b_0 + b_1} (1 - f)^{a_0 + a_1} \binom{a_0 + a_1}{a_0} x_j^{a_0} (1 - x_j)^{a_1} \binom{b_0 + b_1}{b_0} y_j^{b_0} (1 - y_j)^{b_1}}$$

$$\therefore \text{PointRatio}_{xy} = \frac{P_{x_i y_j f b_0 a_0 b_1 a_1}}{P_{x_j y_j f b_0 a_0 b_1 a_1}} = \frac{x_i^{a_0} (1 - x_i)^{a_1}}{x_j^{a_0} (1 - x_j)^{a_1}} \frac{y_i^{b_0} (1 - y_i)^{b_1}}{y_j^{b_0} (1 - y_j)^{b_1}}$$

This shows the relative probability is independent of  $f$ , the relative frequency of noise events in the population, which would be expected for a ROC graph [10].

In order to scale the relative probability at each point  $x_i, y_i$ , to sum to 1, when integrated across the whole surface, the probability ratio is divided by the integral over the surface:

$$\text{NormalisedPointRatio}_{xy} = \frac{\frac{x_i^{a_0} (1 - x_i)^{a_1}}{x_j^{a_0} (1 - x_j)^{a_1}}}{\int_0^1 x_i^{a_0} (1 - x_i)^{a_1} dx_i} \frac{\frac{y_i^{b_0} (1 - y_i)^{b_1}}{y_j^{b_0} (1 - y_j)^{b_1}}}{\int_0^1 y_i^{b_0} (1 - y_i)^{b_1} dy_i}$$

$$\therefore \text{NormalisedPointRatio}_{xy} = \frac{x_i^{a_0} (1 - x_i)^{a_1}}{\int_0^1 x_i^{a_0} (1 - x_i)^{a_1} dx_i} \frac{y_i^{b_0} (1 - y_i)^{b_1}}{\int_0^1 y_i^{b_0} (1 - y_i)^{b_1} dy_i} \quad (1)$$

The Beta function, by definition, and given that  $m$  and  $n$  are integers, is:

$$\beta(m, n) = \int_0^1 x^{m-1} (1 - x)^{n-1} dx = \frac{(m-1)!(n-1)!}{(m+n-1)!}$$

If  $m = a_0 + 1$  and  $n = a_1 + 1$

$$\int_0^1 x^{a_0+1-1} (1 - x)^{a_1+1-1} dx = \frac{(a_0 + 1 - 1)!(a_1 + 1 - 1)!}{(a_0 + 1 + a_1 + 1 - 1)!}$$

$$\therefore \int_0^1 x^{a_0} (1 - x)^{a_1} dx = \frac{a_0! a_1!}{(a_0 + a_1 + 1)!} \quad (2)$$

Substitute (2) into (1)

$$\therefore \text{NormalisedPointRatio}_{xy} = \frac{x_i^{a_0} (1 - x_i)^{a_1}}{\frac{a_0!a_1!}{(a_0 + a_1 + 1)!}} \quad \frac{y_i^{b_0} (1 - y_i)^{b_1}}{\frac{b_0!b_1!}{(b_0 + b_1 + 1)!}}$$

To represent the surface, it is divided into a fine grid and the probability of each quantized grid square is calculated by intergating the probability at a point, over the area of each grid square. The integral over the area is equal to the product of two, one diminsional intergals along the Hit Rate and False Alarm Rate axes. Therefore two vectors,  $\mathbf{X}$  and  $\mathbf{Y}$ , each with  $i$  elements, are defined to hold the one dimensional integrals:

$$\mathbf{X}_i = \frac{\int_{\frac{i-1}{n}}^{\frac{i}{n}} x^{a_0} (1 - x)^{a_1} dx}{\frac{a_0!a_1!}{(a_0 + a_1 + 1)!}} \quad \text{and} \quad (3)$$

$$\mathbf{Y}_i = \frac{\int_{\frac{i-1}{n}}^{\frac{i}{n}} y^{a_0} (1 - y)^{a_1} dy}{\frac{b_0!b_1!}{(b_0 + b_1 + 1)!}} \quad \text{For all } i \text{ from } i=1 \text{ to } i=n \quad (4)$$

The probability surface, quantized as a fine grid, is therefore the product of the two vectors:

$$\text{Surface} = \mathbf{X} \cdot \mathbf{Y}^T \quad (5)$$

The numerators of the vectors  $\mathbf{X}$  and  $\mathbf{Y}$  ((3), (4)) are given in terms of an expression of the following form:

$$\text{Numerator} = \int_q^r x^{a_0} (1 - x)^{a_1} dx \quad (6)$$

Where  $q$  is the probability at the lower boundary of the element, and  $r$  is the probability at the upper boundary of the element.

$$\therefore \text{Numerator} = \int_0^r x^{a_0} (1 - x)^{a_1} dx - \int_0^q x^{a_0} (1 - x)^{a_1} dx \quad (7)$$

Dealing with one partial beta function at a time, where either  $q$  or  $r$  can be substituted for  $s$ :

$$\int_0^s x^{a_0} (1 - x)^{a_1} dx = \int_0^s x^{a_0} ( (1 - s) + (s - x) )^{a_1} dx$$

Applying the binomial expansion:

$$\begin{aligned}
&= \int_0^s x^{a_0} \sum_{k=0}^{a_1} \frac{a_1!}{k!(a_1-k)!} (1-s)^k (s-x)^{a_1-k} dx \\
&= \sum_{k=0}^{a_1} \frac{a_1!}{k!(a_1-k)!} (1-s)^k \int_0^s x^{a_0} (s-x)^{a_1-k} dx \tag{8}
\end{aligned}$$

Change the limits on the integral from 0 to  $s$ , to 0 to 1, by letting  $x = s t$ , which implies  $dx = s dt$ , and letting  $a_2 = a_1 - k$ :

$$\begin{aligned}
\int_0^s x^{a_0} (s-x)^{a_2} dx &= \int_0^1 (s t)^{a_0} (s - s t)^{a_2} s dt \\
&= \int_0^1 s^{a_0} t^{a_0} (s(1-t))^{a_2} s dt \\
&= \int_0^1 s^{a_0} s^{a_2} s t^{a_0} (1-t)^{a_2} dt \\
&= s^{a_0+a_2+1} \int_0^1 t^{a_0} (1-t)^{a_2} dt
\end{aligned}$$

Substituting the modified Beta function as given by (2):

$$= s^{(a_0+a_2+1)} \frac{a_0! a_2!}{(a_0 + a_1 + 1)!} \tag{9}$$

Substituting (9) into (8):

$$\begin{aligned}
\int_0^s x^{a_0} (1-x)^{a_1} dx &= \sum_{k=0}^{a_1} \frac{a_1!}{k!(a_1-k)!} (1-s)^k s^{a_0+a_1-k+1} \frac{a_0! (a_1-k)!}{(a_0 + a_1 - k + 1)!} \\
&= a_0! a_1! \sum_{k=0}^{a_1} \frac{(1-s)^k s^{a_0+a_1-k+1}}{k! (a_0 + a_1 - k + 1)!} \tag{10}
\end{aligned}$$

Substituting (10) into (7) and then substituting into (6) and simplifying:

$$\text{Numerator} = a_0! a_1! \sum_{k=0}^{a_1} \frac{r^{a_0+a_1+1-k} (1-r)^k - q^{a_0+a_1+1-k} (1-q)^k}{k!(a_0 + a_1 + 1 - k)!} \tag{11}$$

Substituting (11) into (3) gives the expression for each element of the  $X$  vector:

$$\begin{aligned} X_i &= \frac{a_0!a_1!}{(a_0+a_1+1)!} \sum_{k=0}^{a_1} \frac{\left(\frac{i}{n}\right)^{a_0+a_1+1-k} \left(1-\frac{i}{n}\right)^k - \left(\frac{i-1}{n}\right)^{a_0+a_1+1-k} \left(1-\frac{i-1}{n}\right)^k}{k!(a_0+a_1+1-k)!} \\ &= (a_0+a_1+1)! \sum_{k=0}^{a_1} \frac{\left(\frac{i}{n}\right)^{a_0+a_1+1-k} \left(1-\frac{i}{n}\right)^k - \left(\frac{i-1}{n}\right)^{a_0+a_1+1-k} \left(1-\frac{i-1}{n}\right)^k}{k!(a_0+a_1+1-k)!} \end{aligned} \quad (12)$$

For all  $i$  from  $i=1$  to  $i=n$

Similarly, for  $Y$ , (by substituting (11) into (4)):

$$Y_i = (b_0+b_1+1)! \sum_{k=0}^{b_1} \frac{\left(\frac{i}{n}\right)^{b_0+b_1+1-k} \left(1-\frac{i}{n}\right)^k - \left(\frac{i-1}{n}\right)^{b_0+b_1+1-k} \left(1-\frac{i-1}{n}\right)^k}{k!(b_0+b_1+1-k)!} \quad (13)$$

For all  $i$  from  $i=1$  to  $i=n$

Equations (12) and (13) give the final computational form of the expressions used to generate the quantized probability surface of a ROC point using the product in equation (5).

## 4. Multiple Points

The analysis can now be expanded to the general case of multiple ROC points. As it has been shown above that the surface can be treated as a product of two probability density vectors, one for Hit Rate and the other for False Alarm Rate, this discussion will examine only one vector, the  $X'$ , or False Alarm Rate vector, the identical method being applicable to the  $Y'$ , or Hit Rate Vector.

For a ROC curve of  $n$  points, there are  $n+1$  classifications of events (threshold ranges). Let there be  $a_i$  occurrences of event  $e_i$ , where  $i = 0, \dots, n$  (see table 2). Let the true probability of event  $e_i$  be  $x_i$ , where  $x_i$  is between 0 and 1. Now an extension of the hypothesis stated above can be applied, by asking the following question, for every point:

*'If this point,  $x_0, x_1, \dots, x_n$ , represents the true frequency of events,  $e_0, e_1, \dots, e_n$ , in the population, what would be the probability of getting the actual results  $a_0, a_1, \dots, a_n$  obtained?'*

If that question can be answered for every point, the relative probability of every point can be calculated and normalised so the sum of all points is 1, thereby giving the normalised probability density surface in  $n+1$  dimensional space.

By multinomial from the numerator of equation (1) for the single ROC point case, the relative probability  $p$ , can be written in as:

$$p \propto \prod_{i=0}^{n-1} x_i^{a_i} \left[ 1 - \sum_{i=0}^{n-1} x_i \right]^{a_n} \quad \text{where} \quad \sum_{i=0}^n x_i = 1$$

The relative probability is a point lying on a hyperplane in  $n+1$  dimensional space. To represent the probability distribution on a two dimensional ROC graph, the  $n+1$  dimensional normalized probability density surface must be mapped into one dimension, i.e. to a  $X'$  vector for the False Alarm Rate, or a  $Y'$  vector for the Hit Rate, and the two dimensional ROC surface formed as a dot product of the  $X'$  and  $Y'$  vectors.

Note that each point on the ROC curve represents a different combination of events. The first point represents  $e_0$  events only, but the second point represents the  $e_0$  plus the  $e_1$  events, the third point  $e_0, e_1$  plus  $e_2$  events, and so on. This adds a subtlety to the way the hyperplane is mapped to the linear probability distribution for each point.

Let  $s_0$  be the actual probability of the first ROC point. The first point represents only the true frequency of the  $e_0$  events. As  $s_0$  varies from 0 to 1, it is directly related to  $x_0$  by the relationship  $s_0 = x_0$ . The hyperplane is mapped into the line by integrating across slices at right angles to the  $x_0$  axis.

Let  $s_1$  be the actual probability of the second ROC point. The second point is the combined frequency of the  $e_0$  events, plus  $e_1$  events, in the population. It does not matter what the individual frequency of  $e_0$  events is, or what the individual frequency of  $e_1$  events is, only the combined frequency matters. As  $s_1$  varies from 0 to 1, the relative probability distribution of the second point can be calculated over the range. For any value of  $s_1$ ,  $x_0$  and  $x_1$  are bound by the relation  $s_1 = x_0 + x_1$ . The hyperplane is mapped into the line by integrating across two dimensional diagonal slices at right angles to the line  $x_0 + x_1 = 1$ ;

Similarly, let  $s_2$  be the actual probability of the third ROC point. The third point is the combined frequencies of  $e_0, e_1$  and  $e_2$  events. As  $s_2$  varies from 0 to 1,  $x_0, x_1$  and  $x_2$  are constrained by the relationship  $s_2 = x_0 + x_1 + x_2$ . The hyperplane is mapped into the line by integrating across the three dimensional diagonal slices at right angles to the line  $x_0 + x_1 + x_2 = 1$ ;

By way of example, the integrals for a four point ROC curve are therefore:

#### 4.1. Four ROC points, 1<sup>st</sup> Point

$$f(s_0) = \int_0^{1-s_0} \int_0^{1-s_0-x_1} \int_0^{1-s_0-x_1-x_2} s_0^{a_0} x_1^{a_1} x_2^{a_2} x_3^{a_3} (1 - s_0 - x_1 - x_2 - x_3)^{a_4} dx_3 dx_2 dx_1$$

The variable  $s_0$  ranges from 0 to 1 across the probability surface. In this case  $s_0$  is equivalent to  $x_0$ . Since the function is constrained to the hyperplane  $x_0+x_1+x_2+x_3+x_4 = 1$ ,  $x_1$  is therefore confined to the range  $1-s_0$ , which are therefore the limits of the outer integral. Similarly,  $x_2$  is then confined to the range 0 to  $1-s_0-x_1$ , the limits of the middle integral, and  $x_3$  is confined to the range 0 to  $1-s_0-x_1-x_2$ , the limits of the inner integral. The expression is kept on the hyperplane by substituting  $x_4 = 1 - s_0 - x_1 - x_2 - x_3$ , in the last term.

Substituting in equation (9), where  $s$  is in turn  $1-s_0-x_1-x_2$ ,  $1-s_0-x_1$ , and  $1-s_0$ :

$$f(s_0) = \int_0^{1-s_0} \int_0^{1-s_0-x_1} s_0^{a_0} x_1^{a_1} x_2^{a_2} (1-s_0-x_1-x_2)^{a_3+a_4+1} \frac{a_3! a_4!}{(a_3+a_4+1)!} dx_2 dx_1$$

$$f(s_0) = \int_0^{1-s_0} s_0^{a_0} x_1^{a_1} (1-s_0-x_1)^{a_2+a_3+a_4+2} \frac{a_2! (a_3+a_4+1)!}{(a_2+a_3+a_4+2)!} \frac{a_3! a_4!}{(a_3+a_4+1)!} dx_1$$

$$f(s_0) = s_0^{a_0} (1-s_0)^{a_1+a_2+a_3+a_4+3} \frac{a_1! (a_2+a_3+a_4+2)!}{(a_1+a_2+a_3+a_4+3)!} \frac{a_2! a_3! a_4!}{(a_2+a_3+a_4+2)!}$$

$$f(s_0) = s_0^{a_0} (1-s_0)^{a_1+a_2+a_3+a_4+3} \frac{a_1! a_2! a_3! a_4!}{(a_1+a_2+a_3+a_4+3)!}$$

#### 4.2. Four ROC points, 2<sup>nd</sup> point

$$f(s_1) = \int_0^{s_1} \int_0^{1-s_1} \int_0^{1-s_1-x_2} x_0^{a_0} (s_1-x_0)^{a_1} x_2^{a_2} x_3^{a_3} (1-s_1-x_2-x_3)^{a_4} dx_3 dx_2 dx_0$$

Again, the variable  $s_1$  ranges from 0 to 1 across the probability surface. In this case  $s_1 = x_0 + x_1$ , and therefore  $x_0$  is confined to the range 0 to  $s_1$ , which are therefore the limits of the outer integral,  $s_1 - x_0$  is substituted for  $x_1$  in the second term, and  $s_1$  is substituted for  $-(x_0+x_1)$  in the last term. The range of  $x_2$  is then confined to the range 0 to  $1-s_1$ , the limits of the middle integral, and  $x_3$  is confined to the range 0 to  $1-s_1-x_2$ , the limits of the inner integral.

$$f(s_1) = s_1^{a_0+a_1+1} \frac{a_0! a_1!}{(a_0+a_1+1)!} (1-s_1)^{a_2+a_3+a_4+2} \frac{a_2! a_3! a_4!}{(a_2+a_3+a_4+2)!}$$

#### 4.3. Four ROC points, 3<sup>rd</sup> point

$$f(s_2) = \int_0^{s_2} \int_0^{s_2-x_0} \int_0^{1-s_2} x_0^{a_0} x_1^{a_1} (s_2-x_0-x_1)^{a_2} x_3^{a_3} (1-s_2-x_3)^{a_4} dx_3 dx_1 dx_0$$

Here,  $s_2 = x_0 + x_1 + x_2$ . This is used in the third and fifth term. The outer and middle integrals are limited by this expression. The inner integral is constrained by the hyperplane.

$$f(s_2) = s_2^{a_0+a_1+a_2+2} \frac{a_0! a_1! a_2!}{(a_0+a_1+a_2+2)!} (1-s_2)^{a_3+a_4+1} \frac{a_3! a_4!}{(a_3+a_4+1)!}$$

#### 4.4. Four ROC points, 4<sup>th</sup> point

$$f(s_3) = \int_0^{s_3} \int_0^{s_3-x_0} \int_0^{s_3-x_0-x_1} x_0^{a_0} x_1^{a_1} x_2^{a_2} (s_3-x_0-x_1-x_2)^{a_3} (1-s_3)^{a_4} dx_2 dx_1 dx_0$$

Here,  $s_3 = x_0 + x_1 + x_2 + x_3$ . This is used in the fourth and fifth term. All the integrals are limited by this expression. The hyperplane constraint is only evident in the last term.

$$f(s_3) = s_3^{a_0+a_1+a_2+a_3+3} \frac{a_0! a_1! a_2! a_3!}{(a_0 + a_1 + a_2 + a_3 + 3)!} (1 - s_3)^{a_4}$$

## 4.5. The General Result for Multiple ROC Points

Since  $f(s_n)$  is a ratio, the factorial terms cancel out. Integration of the expressions for each point of one, two, three, and four point ROC curves reveals a pattern, which by induction generalises to:

$$f(s) = s^{\sum_{i=0}^n a_i + n - 1} (1 - s)^{\sum_{j=0}^m a_j + m - 1}$$

Where  $n$  is the number of the point, and  $n+m$  the total number of points in the curve.

If:

$$a'_0 = \sum_{i=0}^n a_i + n - 1 \quad \text{and} \quad a'_1 = \sum_{j=0}^m a_j + m - 1$$

The multi-point ROC equations are in exactly the same form as the numerator of single point ROC equation (1) and since the denominator is the intergal over the whole hyper volume used to normalise the distribution to sum to 1.0 the same method can be applied to calculate the normalised relative probability surface.

It should be noted that the probability distribution of  $n$  ROC points actually exists in  $2n$  dimensional space. The mapping to  $n$  2D probability surfaces, overlaid on one ROC curve, is merely a convenient representation of this single multidimensional probability distribution.

## 5. Plotting the Surface

A computer program was written in C++ to perform the calculation and plot the 95% confidence interval contour. The surface was quantized to 256 by 256 elements for the graphical presentation. The actual code optimised the mathematical expression (12) (and (13)) by only calculating, and storing in a vector, the boundary values:

$$BoundaryValue = (a_0 + a_1 + 1)! \sum_{k=0}^{a_1} \frac{\left(\frac{i}{n}\right)^{a_0+a_1+1-k} \left(1 - \frac{i}{n}\right)^k}{k!(a_0 + a_1 + 1 - k)!}$$

Each element was then calculated as the difference between two boundary values. Because of the range of the exponent required in the calculation, the excess exponent was held in a long integer as each value was calculated. Many terms in the expression were pre-calculated and accessed from lookup tables.

The surface was then calculated by a dot product of the two vectors. The 'tiles' of the surface were then sorted by normalised probability, largest first, and marked in order, from the largest, as being

inside the confidence boundary until the sum of the marked 'tiles' equalled  $z\%$  of the sum of all the tiles. A boundary drawing algorithm was then applied to draw around the marked area to give the  $z\%$  confidence boundary.

## 6.Simulation Study

To validate the method and the algorithm a Monte Carlo simulation was performed. The L'Ecuyer [30] pseudo random number generator with a period in excess of  $2 \times 10^{18}$ , incorporating a Bays–Durham shuffle with added safeguards, was used. Samples with  $2^n$ ,  $n=0..10$  cases were simulated. Each sample size was simulated with a different set of frequencies of disease in the population. Samples with one case were simulated with a frequency of disease of  $1/2$ , samples with two cases with frequencies of disease  $1/2$  and  $1/4$ , through to samples with 1024 cases being simulated with frequencies of  $1/2^n$ ,  $n=1..11$ . It was not considered worthwhile simulating situations where the frequency of disease, in relation to the number of cases, would often result in no diseased cases at all. For each sample size, at each frequency of disease, ROC curves with 1, 2, 4, 8 and 16 points were simulated. Each point of the multi point ROC curves were simulated independently of the other points to avoid correlation effects as each 2D distribution of a individual point is a different view of the same multidimensional probability distribution of all the points combined. For each test, a population hit rate and false alarm rate was generated for all points, whatever the actual point under test. Then each event within the sample was randomly assigned to the diseased or healthy groups according to the frequency of disease in the population and categorised according to the previously generated population hit rates and false alarm rates. For instance, a 4 point ROC curve has 5 categories. This synthesised sample data was used to generate the probability distribution of each sample for the chosen point. The position of the actual population ROC point within the probability distribution was then recorded to produce a histogram of 20 bins giving the number of times the points fell in the 5%, 10%, ... 95%, 100% confidence interval. Each test was run 2,000 times with the expectation that an average of 100 points would be found in each of the 20 confidence intervals. A chi-squared ( $\chi^2$ ) measured was taken of the 2,000 tests, and the experiment repeated 200 times to obtain a histogram of the chi-squared values. The total simulation thus generated  $(1 + 2 + \dots + 11) \times (1 + 2 + 4 \dots + 16)$  chi-squared histograms of  $200 \times 2,000$  simulated ROC curves, and ran for over a month of a powerful Unix workstation. If the experiment was working, each histogram of chi-squares would approximate the chi-squared distribution for 19 degrees of freedom. The results of the Monte Carlo simulation are discussed in the section 7.

The population hit rate and false alarm rate for multi point ROC curves were produced by generating a uniformly distributed random number between 0 and 1 for the population hit rate of each point and sorting them into ascending order. The same was done for the false alarm rate. The hit rates and false alarm rates were then paired together in the sorted order. This produced data compatible with the ROC curve format, but without parametric assumptions.

It should be noted that the simulation study is unusual in the method of picking the population ROC curves. Many studies [17][23][31] fix the parameters of the curve, e.g. to a binormal curve with an AUC of 0.8, generate random data samples from that population curve, generate sample ROC curves from the data, and verify that the confidence limits (e.g. 95%) of the sample curves, contains the population curve the correct percentage of the time. The current study simulated population ROC points occurring anywhere on the surface, generated data samples from the point, plotted the probability distribution from the data sample, and verified that the point was within given percentiles of the distribution the correct percentage of the time.

The proposed validation method will not work when the parameters of the curve are fixed. This can be explained by considering the following thought experiment. Fig. 3. shows the probability distribution of a ROC point as a contour map with the 33%, 66%, and 100% contour marked. The 100% contour covers the whole graph. The interpretation of the contours is that 33% of the population points that might have produced this sample are inside the 33% contour, 33% of the points are between the 33% and the 66% contour, and 34% are between the 66% and 100% contour. For the sake of the thought experiment, the contours should be regarded as steps with even distributions within each contour. Now consider running a Monte Carlo experiment where the sample that produced this probability distribution happens to be generated 100 times. If the contours are correct, about 33, 33 and 34 population points will land within each contour respectively. Now consider only drawing the test population ROC points from the grey area which shows a region of hypothetical ROC curves. The easiest way to do this is to regard the grey area as a mask. If we repeat the experiment with the mask, only generated population points that happen to lie in the grey area are used. Since the whole of the 33% contour is grey, we will again get about 33 population points in the 33% contour. However, the 66% contour is only about 40% grey, so 60% of the cases that would fall in this area are masked out, leaving about 13 population points in the contour. Similarly, the 100% contour is only about 10% grey, so 90% of the population points will get masked out leaving about 3 population points. A chi-squared test against the expected result of 33, 33, 34 will therefore fail. In other words, there is no point in a confidence interval including any area that cannot have produced a population point, or conversely, all points on the surface have to be able to produce a population point. The thought experiment can be extended to the situation where there are multiple 'grey' regions with various probabilities of masking population points, and any number of step contours. At the limit, the mask becomes an arbitrary continuous probability distribution and the step contour approximation becomes to a continuous probability distribution. It can be seen that the experiment is unlikely to work if the distribution of population points is uneven. The method used in the experiment preserves the uniform distribution along the hit rate and false alarm axis, though the sorting to give a valid ROC curve produces a non-uniform, but still valid, distribution across the ROC graph surface.

## 7.Simulation Results

The simulation run produced 2046 histograms in total, of which only 63 are shown here. Fig. 4. gives the histograms for 1 point ROC curves over 6 sample sizes and 6 frequencies of disease. Fig 5. gives the same histograms for the 3<sup>rd</sup> point of 4 point ROC curves, and Fig. 6. for the 5<sup>th</sup> point of a 16 point ROC curve. The theoretical chi-squared distribution is plotted as a curve on each histogram so a visual comparison can be made between the results expected in theory, and those obtained in practice. As can be seen from the diagrams, the experimental results give the expected chi-squared distributions, which given the number of cases simulated, 400,000 in each histogram, indicates the method is working within the limitations of the quantization of the ROC graph into 256 by 256 elements and the stochastic nature of Monte Carlo simulations.

## 8.Application

The method detailed above was applied to two examples of ROC analysis published in the literature in order to investigate the confidence surfaces produced on real data, and to illustrate how the proposed method could enhance the analysis.

The first example is taken from Swets [10] in which the author recommended the use of ROC analysis for measuring the accuracy of diagnostic systems in general. An example was presented to illustrate the use of ROC analysis and the area under the curve as the preferred single-valued measure of accuracy. A study had previously been carried out in which six radiologists were asked to examine 118 mammograms (58 malignant, 60 benign) and classify them into one of five categories according to likelihood that the lesion was malignant. The radiologists diagnosed the mammograms firstly unaided (denoted as ‘standard’), and then using two diagnostic aids (denoted as ‘enhanced’). The raw data for pooled categorisations were given in the paper, allowing the ROC graph for the standard and enhanced diagnoses (as shown in Swets’ Fig. 2 [10]) to be reproduced here as Fig. 7. From the raw data the 95% confidence boundary of each point was calculated by the program described in Section 5. and these confidence boundaries are shown in Fig. 8. The entire process of calculating the ROC probability distributions and determining the 95% boundaries took roughly 5 seconds on a 100 MHz Pentium PC.

It should be noted that each confidence boundary is a different two dimensional view of the same 8 dimensional probability distribution. Each co-ordinate in the 8 dimensional space represents the probability of the population ROC curve passing through the 4 pairs of hit rate and false alarm rate that describe that 8 dimensional co-ordinate. A 95% confidence boundary can thus be described in the 8 dimensional hyper volume, which is the actual 95% confidence interval for the ROC curve joining the four points. A full theoretical analysis and methodology has yet to be completed, but preliminary analysis has shown that the process can be approximated by joining line segments through tangents to the 95% confidence boundaries of each ROC point such that the maximum and minimum areas are enclosed. It should also be noted that use of a smooth curve or straight line segments to join points is an arbitrary choice outside the theory of the method. An example using straight line segments for Swets’ ‘standard’ ROC data is shown in Fig. 9. In

order to give a comparison with ROC curves in the literature, this approximation has been used to estimate the confidence interval for the AUC.

Calculating AUC from straight line segments as shown in Fig. 7 and the 95% CI's as just described gives non-parametric values for AUC (with 95% CI) of 0.79 (0.739 to 0.839) for the 'standard' points of Swets. Similarly, values for AUC of 0.86 (0.815 to 0.898) are obtained for the 'enhanced' points. In his paper Swets used a parametric method to estimate a 'maximum-likelihood' curve through the points and a corresponding parametric estimate of the AUC and its standard error. The values he obtained were 0.81 and 0.87 with standard errors of 0.017 and 0.014 for the 'standard' and 'enhanced' diagnoses respectively. Parametrically, the 95% CI is given by the mean value  $\pm 2 \times$  standard error, which leads to values for AUC (with 95% CI) of 0.81 (0.776 to 0.844) and 0.87 (0.842 to 0.898). It can be seen that the non-parametric estimates obtained here agree well with Swets' parametric estimates, albeit with slightly lower AUC's and fractionally larger confidence intervals.

The second example is taken from Adlassnig & Scheithauer [4] in which an expert system, known as CADIAG-2/PANCREAS, for the differential diagnosis of ten different types of pancreatic disease is described. The performance of the system was compared to an histologically or clinically confirmed 'gold-standard' diagnosis. Forty seven patient records were available in which one or more of a subset of six pancreatic diseases had been diagnosed. Four patients had dual diagnoses, giving a total of fifty one diagnoses of one of six diseases. A series of ROC graphs were presented, illustrating the performance of the CADIAG-2 system in the differential diagnosis of specific diseases, both using what was described as a limited set of patient data and with the full set of available patient data. Their Fig. 9 and Fig. 10 presented the evaluations of 8 diagnoses of acute pancreatitis from the 51 by CADIAG-2 compared to the 'gold-standard' using limited patient data and full patient data respectively [4]. Although the raw data was not given, because the number of cases were so small, it can be accurately reconstructed from the ROC graphs. These data are here combined and reproduced as Fig. 9 and the 95% confidence boundaries calculated from the data are shown in Fig. 10. In this case the process took only 3 seconds on a 100 MHz Pentium PC.

Although Adlassnig and Scheithauer described the use of AUC and testing its differences statistically for comparison of ROC curves in their paper, no results were given for the curves obtained. From Fig. 10 it is obvious that there is considerable uncertainty in the results due to the limited number of cases used. Using the method described above gives an AUC of 0.79 (0.476 to 0.936) for the diagnoses based on 'limited' patient data and 0.94 (0.617 to 0.981) for the diagnoses based on 'full' patient data. In their analysis of evaluation results the authors go on to state that accuracy was always increased by adding the 'full' patient data, in accordance with anticipation. While the ROC curves presented in Fig. 10 appear to support this common sense conclusion, the large confidence boundaries in Fig. 11 suggest that this conclusion was probably premature given the data. Further, given that a random classifier has an AUC of 0.5 and a perfect classifier has an AUC

of 1.0, the fact that the 95% confidence intervals of the AUC for both sets of data are so close to these extremes illustrates how much caution should be placed in a study of such limited numbers.

Fig. 8 and Fig. 10 clearly illustrate the difference that sample size makes to the confidence that can be placed in the location of each point. While the ROC curve of the mammogram diagnoses in Fig. 7 do not look as accurate as the ROC curve for the diagnosis of acute pancreatitis in Fig. 9, examination of the confidence boundaries in Fig. 8 and Fig. 10 shows that the 708 (six opinions of 118) mammograph cases are sufficient to give good confidence of the location of the ROC points and hence in the curve, while the 51 pancreatic cases give a much larger confidence boundary. In particular, it can be seen by consideration of pairwise points in Fig. 8 that the points are outside of each others' confidence boundaries in all cases, and that the confidence boundaries are mutually exclusive in one case. In contrast, in Fig. 10, the points with false alarm rate of 0.093 lie within each other's confidence boundaries, the 'limited' data point with false alarm rate of 0.638 lies well within the confidence boundary of the 'full' point, and that there is a high degree of overlap in confidence boundaries in all cases.

## 9. Discussion

This focus of this research is to improve the formalisation of the evaluation of intelligent medical systems. ROC analysis is an established method of measuring diagnostic performance that has frequently been applied to the evaluation of intelligent medical systems [4]. In this context there are two common characteristics of evaluation data: (i) a set of 'difficult' or 'interesting' cases are usually selected for the evaluation task rather than cases being picked at random from population data, and (ii) the sample sizes are often quite small in order to be manageable by the human experts. In any method of measuring diagnostic performance, including ROC analysis, it is highly desirable to accurately quantify confidence intervals. Because of the described properties it is important therefore to have a method of generating ROC confidence intervals based on non-parametric assumptions which is robust for small sample size. However, none of the previously derived methods satisfy these requirements [22]. This paper sets out a novel method for generating robust and accurate probability distributions for all points of a non-parametric ROC curve based on samples of any size. From these probability distributions it is straight forward to derive confidence boundaries as required.

The theoretical analysis indicates that the method derived above should be robust and accurate over any sample size, for any frequency of disease, and for any number of points. The method would appear to overcome the limitation stated by Zou et al. [22] that all non-parametric models are unreliable for corners of the curve since there is not much information at the extreme points. At the limit, with samples consisting of (an impractical) zero cases, the method is robust and accurate in producing a flat probability distribution over the whole surface. In other words, each point on the ROC surface is regarded as equally likely for the population a priori. The method is also robust with samples that do not have any false positive or false negative cases. Other methods

[29] fail on these samples that occur with increasing frequency the further the diseased and normal distributions are separated and the fewer the number of cases in the sample.

The method described was implemented in software and was thoroughly tested by Monte Carlo simulation. It was seen to produce accurate probability distributions over a wide range of sample sizes, from 1 to 1024 in powers of 2, over a wide range of frequencies of disease in the population,  $1/2$  to  $1/2048$  in powers of  $1/2$ , for ROC plots with a wide range of number of plotted points (1, 2, 4, 8 and 16). The Monte Carlo simulation used a novel technique in randomly picking the population points on the surface with an even distribution across each axis. This contrasts with existing methods that pick a small number of fixed curves. It has been shown that this method is theoretically justified.

The use of confidence boundaries to enhance the analysis of real data has been illustrated. The separation of points and their confidence boundaries in Fig. 8 visually emphasises Swets' analysis [10] that the difference between AUC for the two diagnostic situations is statistically significant. In the case of Adlassnig and Scheithauer, the large confidence boundaries obtained from such limited data clearly indicate the limitations of their study. ROC analysis is being used with increasing popularity in the evaluation of medical intelligent systems. If ROC curves are to be of real benefit, rather than simply being pretty drawings, the errors must be properly calculated and represented. This work establishes an important new method for generating probability distributions for all such studies.

This work has a further, potentially very important, application to the evaluation of medical intelligent systems. For ROC analysis to be valid and, in particular, for area under the curve to be a meaningful measure, successive points on an ROC curve must be generated by altering the perceived cut-off value in the single (multi-valued) output. However, often in intelligent systems such as expert systems, fuzzy logic models, and artificial neural networks, internal model parameters may be varied or 'tuned' in order to alter performance. Such alterations produce alternative outputs which can be plotted as points on an ROC chart, but the points are not related to each other as points on a single curve. Another example would be different expert opinion of a single fixed diagnostic test. The method presented here allows a probability distribution to be calculated for each such point independently and hence could allow meaningful comparisons between points.

In future, extensions of the method to other aspects of ROC analysis will be investigated. These include comparison of ROC curves produced by different diagnostic tests, either using paired, unpaired or partially paired data, and, for the present research, comparing ROC curves produced by intelligent medical systems with ROC curves produced by experts when examining the same cases [23]. Potentially, from the probability distributions of ROC curves it is conjectured that it should also be possible to give the exact probability that one 'expert' is better than another 'expert'. This improves on the present situation where only a qualitative comparison can be made,

and would be significant because then exact risks could be calculated for the deployment of an intelligent medical system in clinical use.

## Acknowledgement

Thanks to Julian Stander for advice on the presentation of the results of the Monte Carlo simulation and for his editorial suggestions.

## References

- 1 J. M. Garibaldi, E. C. Ifeachor, 'Application of Simulated Annealing Fuzzy Model Tuning to Umbilical Cord Acid-Base Interpretation', *IEEE Transactions on Fuzzy Systems*, Vol.7, No.1, pp.72-84, 1999
- 2 J. W. Huang, Y. Lu, A. Nayak and R. J. Roy, 'Depth of Anesthesia Estimation and Control', *IEEE Transactions on Biomedical Engineering*. Vol.46, No.1, pp.71-81, 1999.
- 3 D. A. Cairns, J. H. L. Hansen and J. E. Riski, 'A Noninvasive Technique for Detecting Hypernasal Speech Using a Nonlinear Operator', *IEEE Transactions on Biomedical Engineering*, vol.43, no.1, pp.35-45, 1996.
- 4 K. P. Adlassnig and W. Scheithauer, 'Performance evaluation of medical expert systems using ROC curves', *Computers and Biomedical Research*, Vol.22, No.4, pp.297-313, 1989.
- 5 L. G. Koss, M. E. Sherman, M. B. Cohen, A. R. Anes, T. M. Darragh, L. B. Lemos, B. J. McClellan, D. L. Rosenthal, S. Keyhani-Rofagha, K. Schreiber, P. T. Valente, 'Significant Reduction in the Rate of False Negative Cervical Smears With Neural Network-Based Technology (PAPNET Testing System)', *Human Pathology*, Vol 28, No 10, pp.1196-1203, 1997.
- 6 S. Andreassen, A. Rosenfalck, B. Falck, K. G. Olesen, S. K. Andersen, 'Evaluation of the diagnostic performance of the expert EMG assistant MUNIN', *Electroencephalography and clinical Neurophysiology*, Vol 101, pp.129-144, 1996.
- 7 B. V. Ambrosiadou, D. G. Goulis, C. Pappas, 'Clinical evaluation of the DIABETES expert system for decision support by multiple regimen insulin dose adjustment', *Computer Methods and Programs in Biomedicine*, Vol.49, pp.105-115, 1996.
- 8 H. D. Cheng, Y. M. Lui, R.I. Freimanis, 'A novel approach to microcalcification detection using fuzzy logic technique', *IEEE Transactions on Medical Imaging*, Vol.17, No.3, pp.442-450, 1998.
- 9 R. D. F. Keith, S. Beckley, J.M. Garibaldi, J.A. Westgate, E.C. Ifeachor, K. R. Green, 'A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram', *British Journal Obstetrics Gynaecology*, Vol.102, pp.688-700, 1995.
- 10 J. A. Swets, 'Measuring the Accuracy of Diagnostic Systems', *Science*, Vol.240, pp.1285-1293, 1988.

- 11 K. Clarke, R. O'Moore, R. Smeets, J. Talmon, J. Brender, P. McNair, P. Nykanen, J. Grimson, B. Barber, 'A methodology for evaluation of Knowledge-based systems in medicine', *Artificial Intelligence in Medicine*, Vol.6, pp.107–121, 1994.
- 12 R. Engelbrecht, A. Rector, W. Moser, 'Verification and Validation', in 'Assessment and Evaluation of Information Technologies', E. M. S. J. van Gennip, J. L. Talmon, Eds, IOS Press, pp.51–66, 1995.
- 13 A. R. Henderson, 'Assessment of clinical enzyme methodology: a probabilistic approach', *Clinica Chimica Acta*, Vol 257, pp.25–40, 1997.
- 14 A. A. Renshaw, K. R. Lee, S. R. Granter, 'Use of Statistical Analysis of Cytologic Interpretation to Determine the Causes of Interobserver Disagreement and in Quality Improvement', *Cancer Cytopathology*, Vol.81, No.4, pp.212–219, 1997.
- 15 S. Y. Jang, R. J. Jaszczak, B. M. W. Tsui, C. E. Metz, D. R. Gilland. T. G. Turkington, R. E. Colman, 'ROC evaluation of SPECT myocardial lesion detectability with and without single iteration non-uniform Chang attenuation compensation using an anthropomorphic female phantom', *IEEE Transactions on Nuclear Science*, Vol.45. No.4, Pt.2, pp.2080–2088, 1998.
- 16 D. J. Goodenough, K. Rossmann, L. B. Lusted, 'Radiographic applications of signal detection theory', *Radiology*, Vol.105, pp.199–200, 1972
- 17 J. A. Hanley, and B. J. McNeil, 'The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve', *Diagnostic Radiology*, Vol.143, No.1, pp.29–36, 1982.
- 18 B. J. McNeil and J. A. Hanley, 'Statistical approaches to the analysis of receiver operating characteristic (ROC) curves', *Medical Decision Making*, Vol.4, pp.137–150, 1984.
- 19 S. C. Kazmierczak, P. G. Catrou, F. Van Lente, 'Diagnostic Accuracy of Pancreatic Enzymes Evaluated by Use of Multivariate Data Analysis', *Clinical Chemistry*, Vol.39, No.9, pp.1960–1965, 1993.
- 20 L. Ohno-Machado, 'A Comparison of Cox proportional hazards and artificial neural network models for medical prognosis', *Computers in Biology and Medicine*, Vol.27, No.1, pp.55–65, 1997.
- 21 D. Bamber, 'The area above the ordinal dominance graph and the area below the receiver operating graph', *J. Math. Psychol.*, Vol.12, pp.387–415, 1975.
- 22 N. A. Obuchowski, M. L. Lieber, 'Confidence Intervals for the Receiver Operating Characteristic Area in Studies with Small Samples', *Academic Radiology*, Vol.5, pp.561–571, 1998.
- 23 C. E. Metz, 'Statistical analysis of ROC data in evaluating diagnostic performance', in D. Herbert and R. Myers (eds), 'Multiple Regression Analysis: Applications in the Health Sciences', New York, American Institute of Physics, pp.365–384, 1986.
- 24 E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson, 'Comparing the areas under two or more correlated receiver operating characteristic

- curves: a nonparametric approach', *Biometrics*, Vol.44, pp.837–845, 1988.
- 25 B. Efron, R. J. Tibshirani, 'An Introduction to the bootstrap', New York, Chapman & Hall, 1993.
  - 26 S. W. Greenhouse and N. Mantel, 'The evaluation of diagnostic tests', *Biometrics*, Vol.6, pp.399–412, 1950.
  - 27 H. Schafer, 'Efficient confidence bounds for ROC curves', *Statistics in Medicine*, Vol.13, pp.1551–1561, 1994.
  - 28 R. A. Hilgers, 'Distribution-free confidence bounds for ROC curves', *Methods of Information in Medicine*, Vol.30, pp.137–150, 1991.
  - 29 C. E. Metz, B. A. Herman, J. Shen, 'Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data', *Statistics in Medicine*, Vol.17, pp.1033–1053, 1998.
  - 30 P. L'Ecuyer, 'Efficient and Portable Combined Random Number Generators', *Communications of the ACM*, Vol.31, No.6, pp.742–774, 1988.
  - 31 K. H. Zou, W. J. Hall, and D. E. Shapiro, 'Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests', *Statistics in Medicine*, Vol.16, pp.2143–2156, 1997

## Captions

Fig. 1. The underlying model for ROC curves

Table 1. Single threshold contingency table

Table 2. Three threshold contingency table, where  $\phi$  is the measurement.

Fig. 2. Example ROC Curve

Fig 3. Thought experiment on the distribution of ROC curves tested.

Fig. 4. Histograms of chi-squares for 1 point ROC curve

Fig. 5. Histograms of chi-squared for 3<sup>rd</sup> point of 4 point ROC curve

Fig. 6. Histograms of chi-squared for 5<sup>th</sup> point of 16 point ROC curve

Fig. 7. ROC curve of diagnosis of 708 mammograms (from Swets)

Fig. 8. The 95% confidence boundaries of ROC points in Fig. 7.

Fig. 9. ROC curve for 51 diagnoses for acute pancreatitis (from Adlassnig & Scheithauer)

Fig. 10. The 95% confidence boundary of points in Fig. 9.

# Figures

Frequency

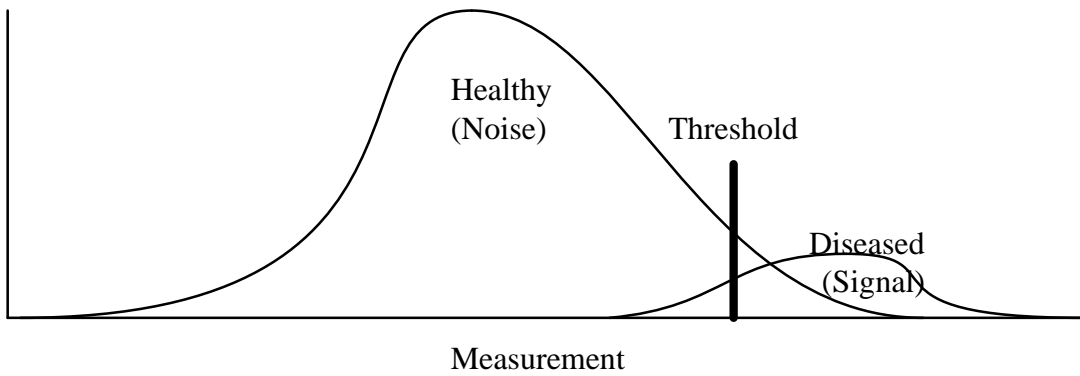


Fig. 1.

		Standard	
		Signal	Noise
Test	Signal	True +ve $b_0$	False +ve $a_0$
	Noise	False -ve $b_1$	True -ve $a_1$

Table 1.

		Standard	
		Signal	Noise
Test	$\phi \geq \text{Threshold } 0$	$b_0$	$a_0$
	$\text{Threshold } 0 > \phi \geq \text{Threshold } 1$	$b_1$	$a_1$
	$\text{Threshold } 1 > \phi \geq \text{Threshold } 2$	$b_2$	$a_2$
	$\text{Threshold } 2 > \phi$	$b_3$	$a_3$

Table 2.

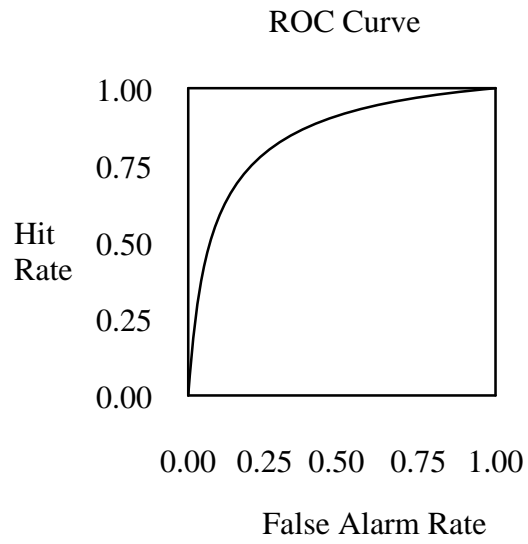


Fig. 2.

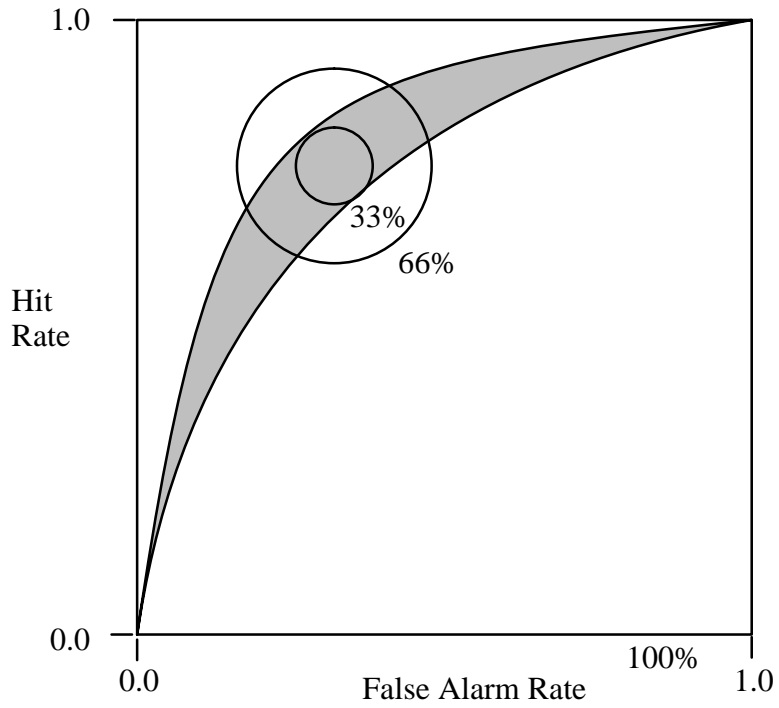


Fig. 3.

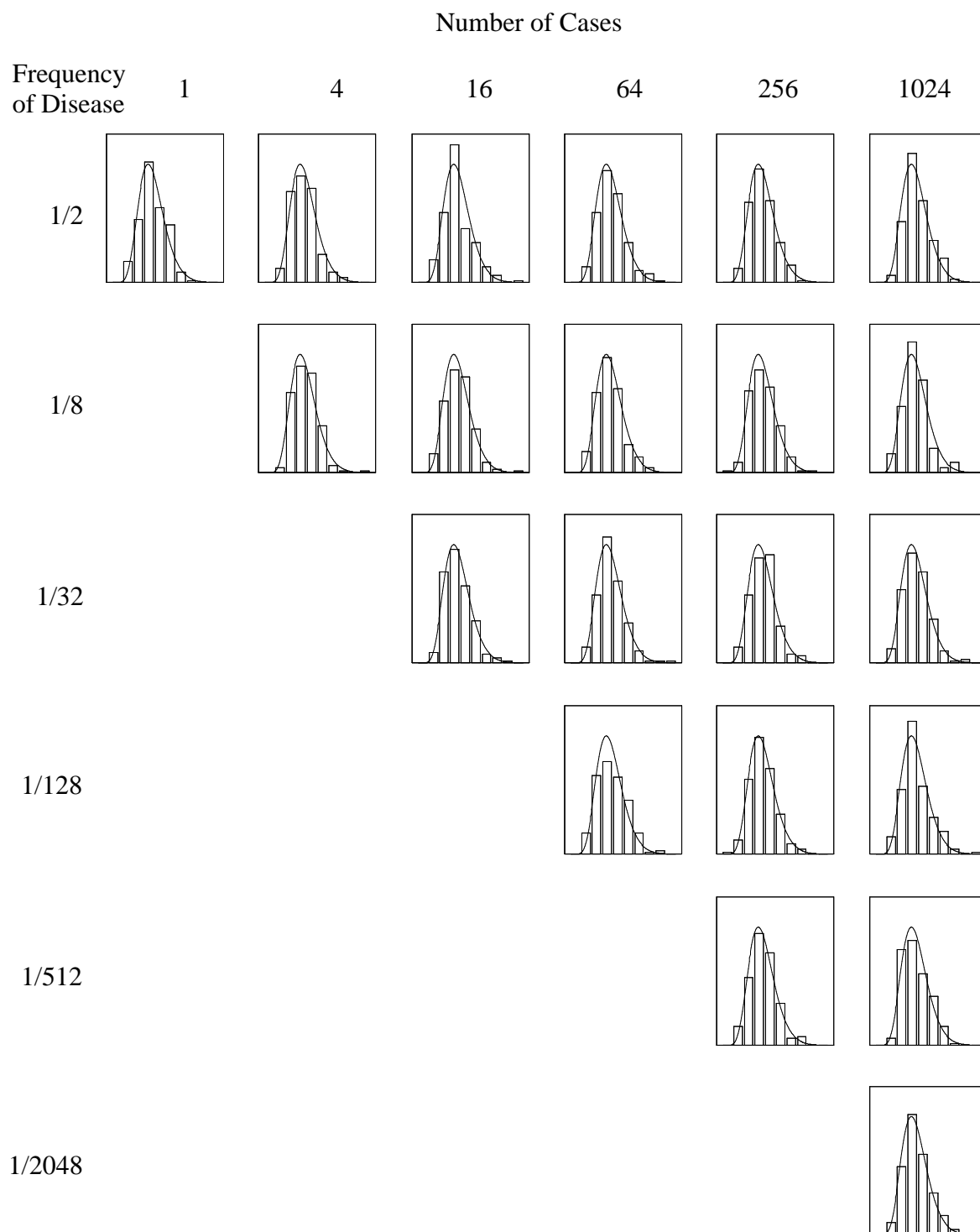


Fig. 4.

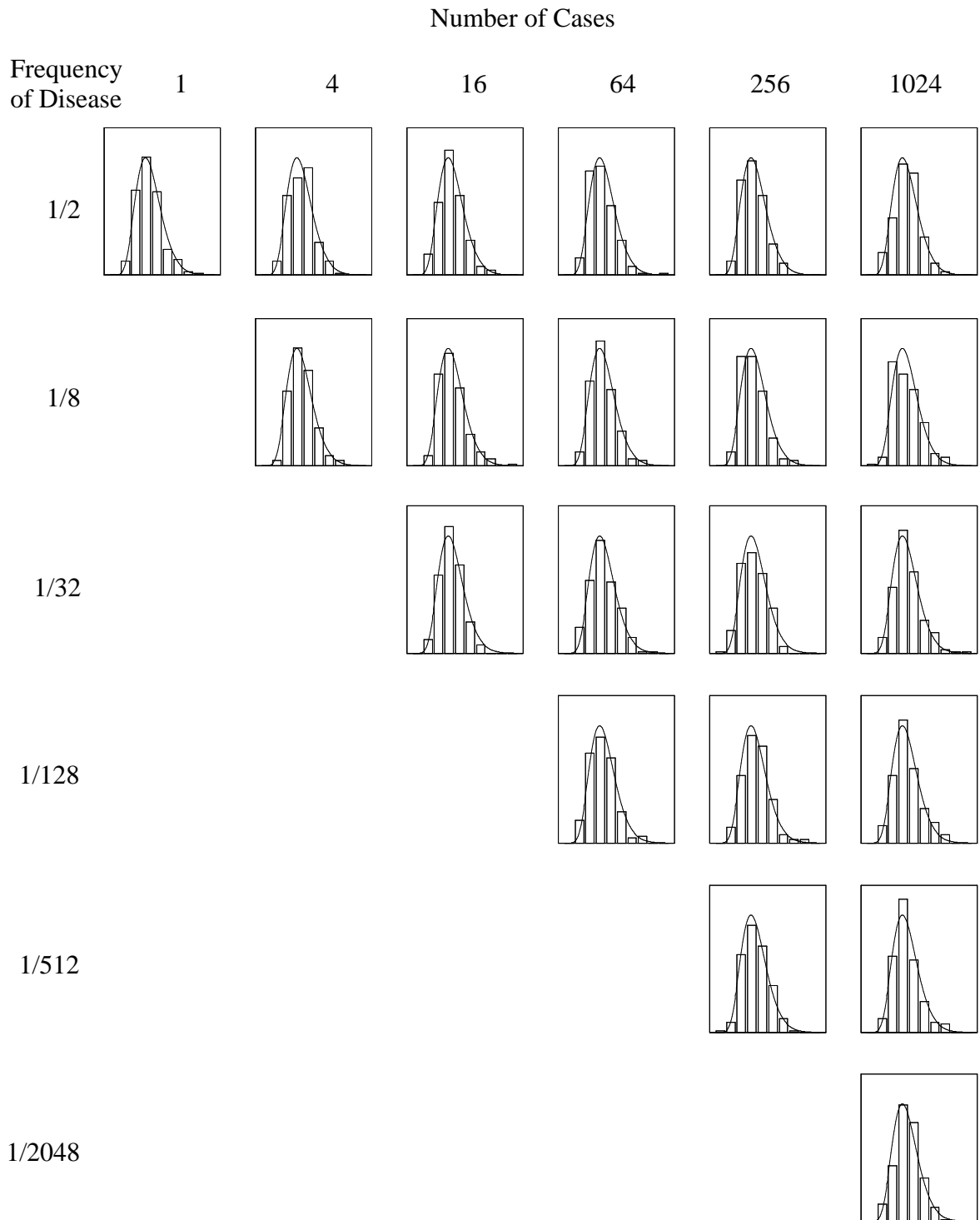


Fig. 5.

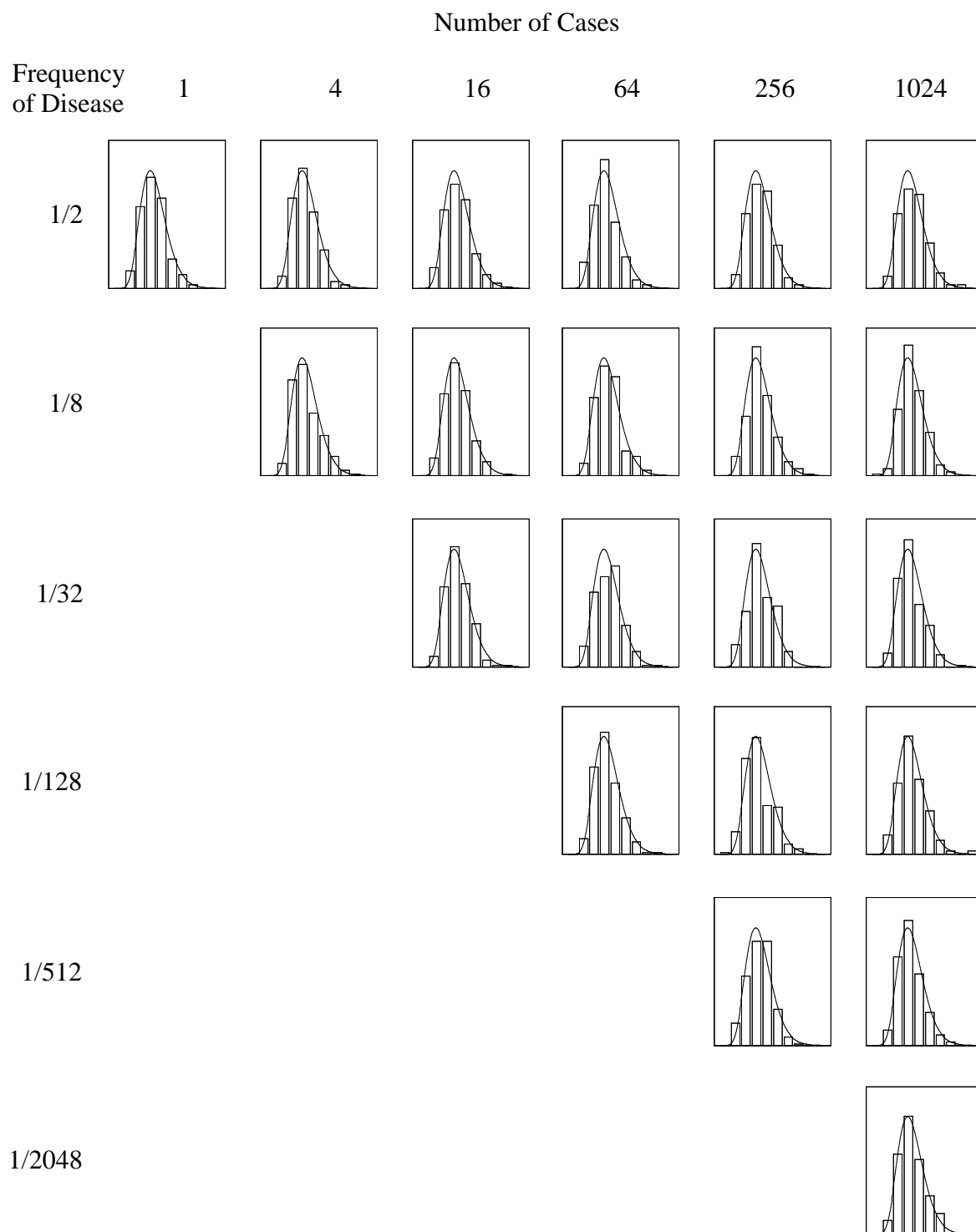


Fig. 6.

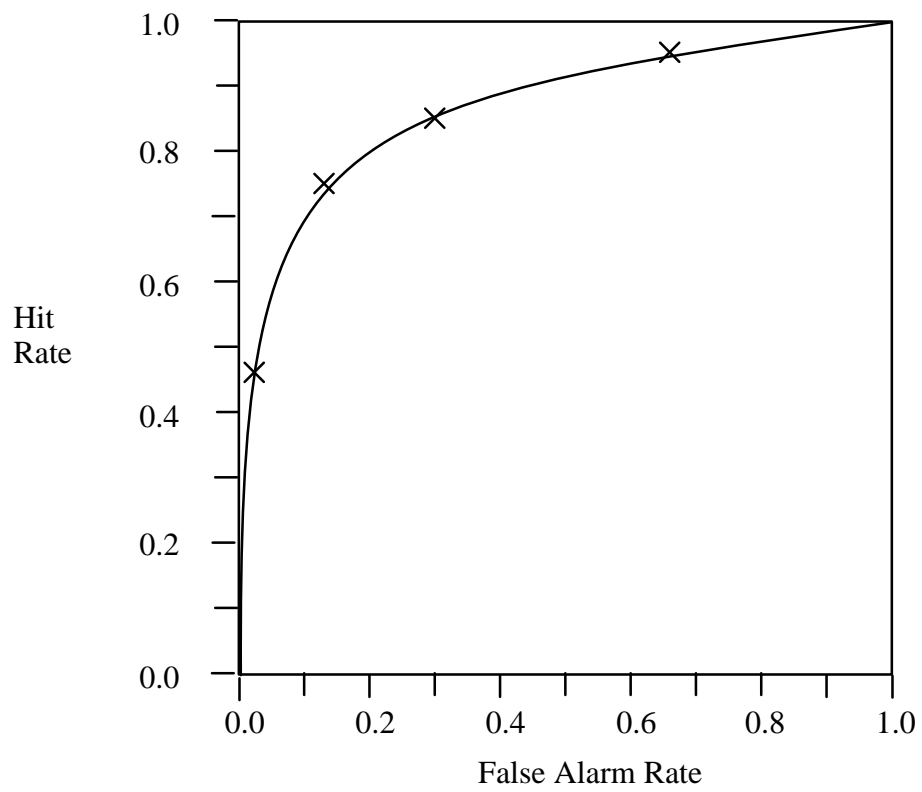


Fig. 7.

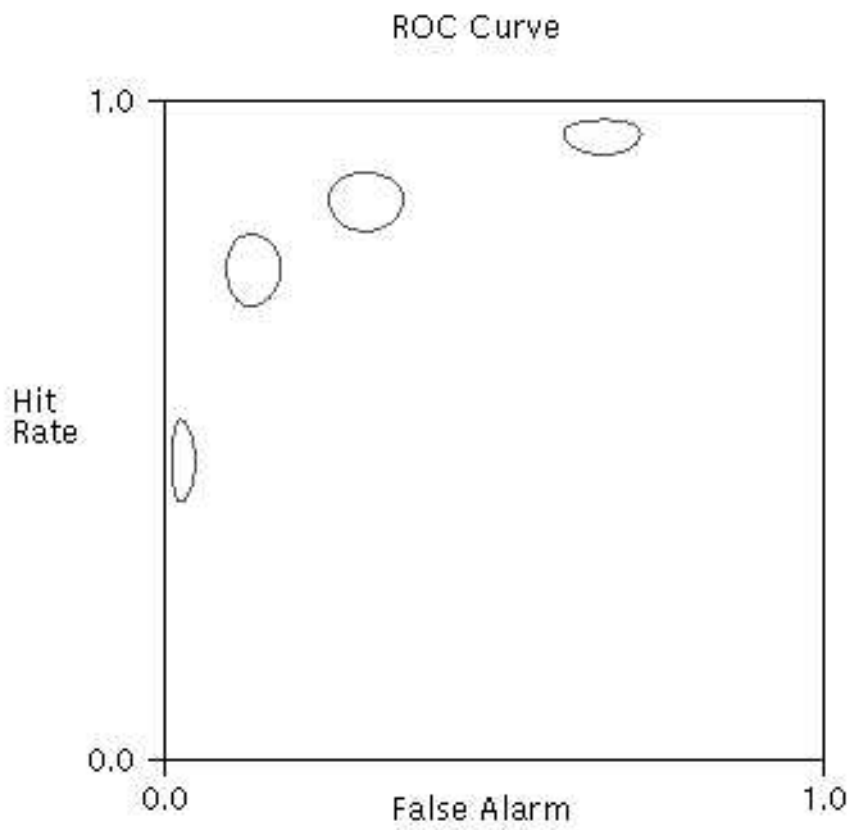


Fig. 8.

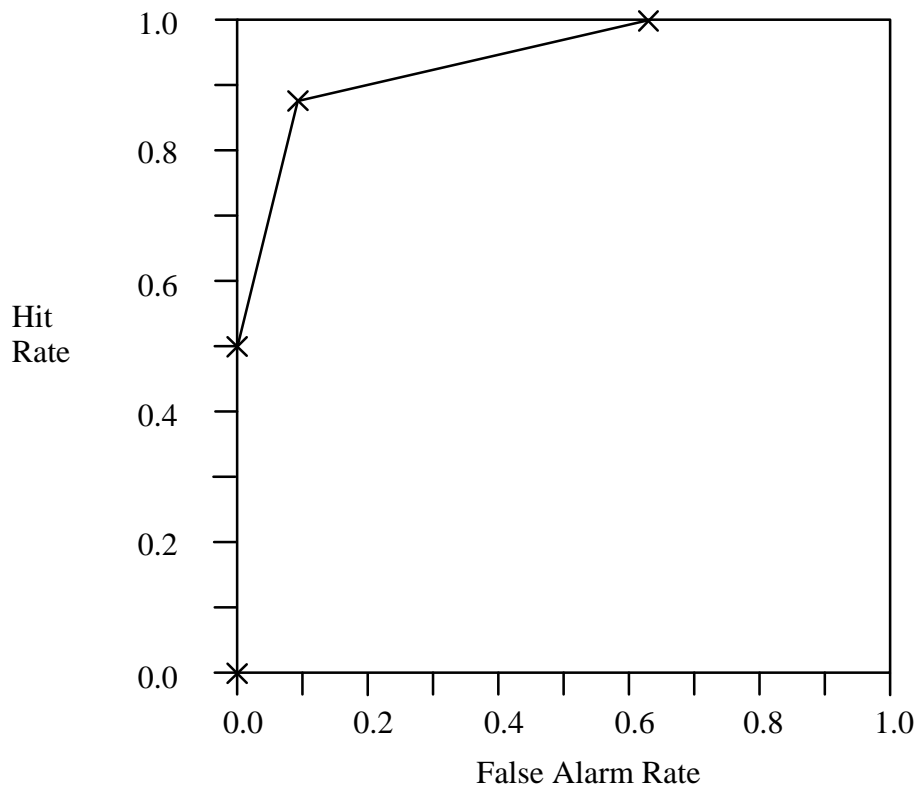


Fig. 9.

ROC Curve

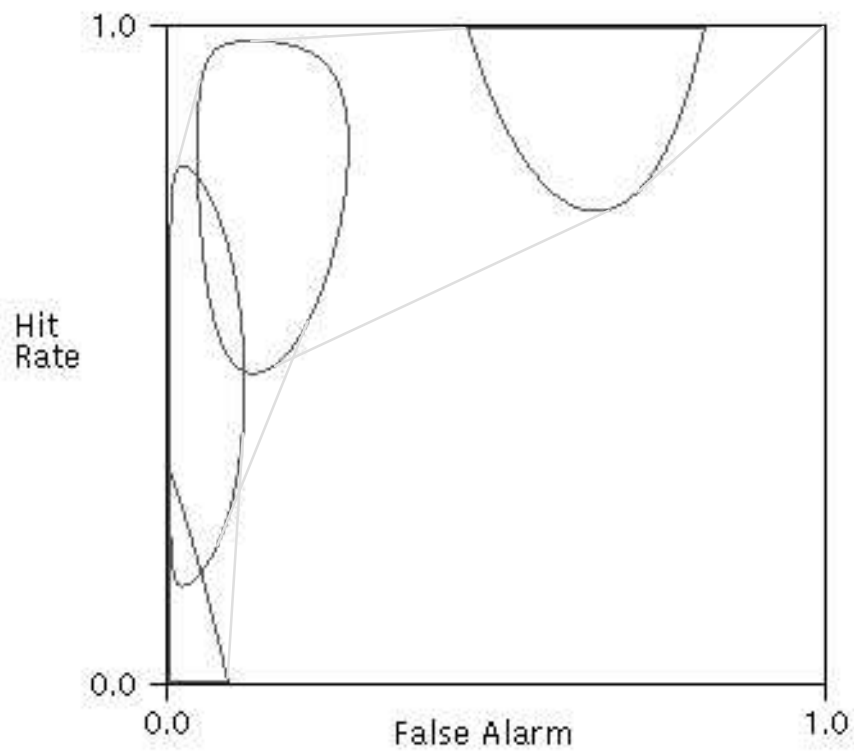


Fig. 10.