DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

INVESTIGATION OF DIAGNOSTIC TEST PERFORMANCE USING INFORMATION THEORY

by
Armağan KANCA

July, 2012
İZMİR

INVESTIGATION OF DIAGNOSTIC TEST PERFORMANCE USING INFORMATION THEORY

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for
the Degree of Master of Science in Statistics

by

Armağan KANCA

July, 2012
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"INVESTIGATION OF DIAGNOSTIC TEST PERFORMANCE USING INFORMATION THEORY"** completed by **ARMAĞAN KANCA** under supervision of **ASSOC. PROF. DR. ÖZLEM EGE ORUÇ** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. ÖZLEM EGE ORUÇ

Supervisor

Assist. Prof. Dr. EMEL KURUOĞLU

(Jury Member)

Assist. Prof. Dr. İLHAN KARAKILIÇ

(Jury Member)

Prof.Dr. Mustafa SABUNCU
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to thank everybody who supported me on my thesis for three years. I owe a debt of gratitude to my supervisor Assoc. Prof. Dr. Özlem EGE ORUÇ who generously shared her priceless knowledge and experiences during this whole process for her stimulating ideas and patience and for encouraging me to apply for student exchange program.

I also want to express my gratitude to Prof. Dr. Isabel NATARIO for her contributions on my research project that I conducted in Lisbon and for expanding my horizon. It was a privilege to conduct a research with her.

I thank Dr. Nilgün IŞIKSAÇAN for her support and help and for providing me with the opportunity to work with genuine data during application period.

I would like to extend my gratitude to Bahar YALÇIN for all her help and support and to my cousin Emin RODOSLU for his contributions and patience.

Last but not least, I want to express my greatest gratitude to my mother Gülsüm KANCA and my father Cemil KANCA who support me all the time for their love and patience, my sister Yasemin Özden KANCA who helped and supported me in every phase of this thesis and to my one and only grandmother Elvan RODOSLU for always making me feel special and lucky to have such a broad minded grandmother.

Armağan KANCA

INVESTIGATION OF DIAGNOSTIC TEST PERFORMANCE USING INFORMATION THEORY

ABSTRACT

For the application of this thesis, ASO values of 68 subjects who applied to Istanbul Mehmet Akif Ersoy Thoracic and Cardiovascular Surgery Training and Research Hospital for the diagnosis of rheumatic disorder were used. ASO is a value which is used to learn whether the patients have group A beta-hemolytic streptococcal infection which causes these diseases.

ASO values were evaluated according to Turbidimetric methods of two different firms. Since the names of the firms were kept secret, these methods were called as I. Turbidimetric method and II. Turbidimetric method. Both ROC and Information Theory Analyses were applied to the results. Therefore, both firms' Turbidimetric method diagnostic test performances were evaluated and which diagnostic test had better performance was determined.

The disease diagnosis is considered among the most important parts of the treatment process. The aim of this study is to demonstrate how basic concepts in Information Theory and in Receiver Operating Characteristics (ROC) apply to the problem of quantifying diagnostic test performance.

Before evaluation process, Receiver Operating Characteristic (ROC) analysis and Information Theory are described. The role of ROC analysis and Information Theory analysis in the medicine sector are presented in the context of diagnostic tests performance. Moreover, the notations of ROC analysis and Information Theory are introduced. In ROC analysis, it is indicated how to create ROC curves and their properties are explained in detail. After ROC curves are drawn, Area Under the Curve (AUC) concept which is one of the diagnostic test performance measures is evaluated. In Information Theory analysis, the concepts of entropy and conditional entropy are introduced. Afterwards by using these two concepts, Mutual Information

value which is one of the diagnostic test performance measures is evaluated. At the final stage, it is indicated which diagnostic test has the best performance based on the mentioned measures.

Keywords: Diagnostic tests, receiver operating characteristics, area under the curve, entropy, conditional entropy, mutual information.

TANI TESTİ PERFORMANSININ İNCELENMESİNDE BİLGİ TEORİSİ KULLANIMI

ÖZ

Bu tezin uygulamasında, İstanbul Mehmet Akif Ersoy Göğüs, Kalp ve Damar Cerrahisi Eğitim Araştırma Hastanesine romatizmal hastalıkların tanısı için gelen 68 kişinin ASO değerleri kullanılmıştır. ASO, hastaların bu hastalıklara sebep olabilecek A grubu beta-hemolitik streptokok enfeksiyonu geçirip geçirmediğini öğrenmek için kullanılan bir değerdir.

ASO değerleri, iki farklı firmaya ait Türbidimetrik yöntemlerle ölçülmüştür. Firma isimlerinin saklı tutulması sebebiyle bu yöntemlere I. Türbidimetrik yöntem ve II. Türbidimetrik yöntem adlarını vermek uygun bulunmuştur. Çıkan test sonuçlarına, hem ROC hem de Bilgi Teorisi analizi uygulanmıştır. Böylelikle her iki firmaya ait Türbidimetrik yöntem tanı testi performansları hesaplanmış ve hangi tanı testinin daha iyi performans sergilediği belirlenmiştir.

Hastalık tanısı tedavi sürecinin en önemli kısımlarından biri olarak kabul edilmektedir. Bu çalışmanın amacı Bilgi Teorisi ve Alıcı İşlem Karakteristiklerinin (ROC) temel kavramlarının tanı testi performansını belirlemek için nasıl uygulandığını göstermektir.

Değerlendirme sürecinden önce Alıcı İşlem Karakteristikleri (ROC) analizi ve Bilgi Teorisi tanımlanmıştır. Sağlık sektöründe ROC analizi ve Bilgi Teorisi analizinin rolü tanı testi performansı bağlamında ortaya konmuştur. Ayrıca, ROC analizi ve Bilgi Teorisi notasyonları tanıtılmıştır. ROC analizinde ROC eğrilerinin nasıl oluşturulduğu gösterilmiş ve bu eğrilerin özellikleri detaylı bir biçimde anlatılmıştır. ROC eğrileri çizildikten sonra, tanı testi performansının ölçütlerinden biri olan Eğri Altında Kalan Alan (AUC) kavramı hesaplanmıştır. Bilgi Teorisi analizinde entropi ve koşullu entropi kavramları tanıtılmıştır. Daha sonra bu iki kavramdan yararlanılarak, tanı testleri performansı ölçütlerinden Karşılıklı Bilgi

değeri hesaplanmıştır. En son aşamada, bahsedilen ölçütlere dayanarak hangi tanı testinin daha iyi performansa sahip olduğu belirlenmiştir.

Anahtar Kelimeler: Tanı testleri, alıcı işlem karakteristiği**,** eğri altında kalan alan**,** entropi, koşullu entropi, karşılıklı bilgi.

CONTENTS

Page

CHAPTER ONE
INTRODUCTION

Diagnostic tests are widely used in many areas. In particular, these tests have huge importance in medicine sector. By courtesy of early and accurate diagnosis, the morbidity and mortality of disease can be reduced. It is important to compare various diagnostics test with each other under specific clinical conditions in order to determine which one is the best to use.

One of the approaches used to analyze the performance of diagnostic tests is ROC theory. The roots of ROC theory are laid on statistical decision theory. ROC analysis was first used in the 1950's for radio signals, this use decreased gradually in the following decade. After the 1960's, the usage of ROC analysis was canalized to the medicine sector. Since that time, ROC played an essential role in medicine sector and it is still widely used in this sector. ROC curves became the standard approach to summarizing diagnostic test performance after published a medical application of this method (Lusted, 1971).

The other approach which is used to analyze the performance of diagnostic tests in recent years is Information Theory. The Information Theory was developed by Claude Shannon in 1948. In Shannon's theory, the information is associated with uncertainty. This theory of knowledge and uncertainty for the measurement is based on a mathematical basis. In 1973, Metz, Goodenough and Rossmann developed a formula used in assessing the performance of diagnostic tests by using information theory. After this work, Somoza and Mosmann developed a new mathematical and graphical method to evaluate and compare the performance of diagnostic tests for the value of any prevalence (P) by using the properties of the ROC analysis and Information Theory approach. In 1999, Lee obtained the distance between patients and healthy distributions by using the concept of relative entropy. In 2002, Benish investigated the concept of relative entropy with a different perspective.

In this study, the performance of the diagnosis tests on the field of rheumatic disorder is analyzed. The study consists of five chapters. In the first chapter, the introduction containing the historical development of the ROC theory and the historical development of the Information Theory concepts which are the measures of diagnosis test performance is introduced. In the second chapter, the notations of the ROC theory is explained in detail and at the end of the chapter a toy example about Down Syndrome is given. In the third chapter, the notations of the Information Theory are introduced and explained in detail. In the fourth chapter, the data of rheumatic disorder are analyzed according to ROC and Information Theory. In the last chapter, the deductions gathered from the analyses are interpreted.

CHAPTER TWO
BASIC CONCEPTS OF RECEIVER OPERATING CHARACTERISTIC (ROC)

The ROC curve is a fundamental tool for diagnostic test evaluation. In a ROC curve the true positive fraction (sensitivity) is plotted in function of the false positive fraction (1-specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/(1-specificity) pair corresponding to a particular decision threshold. The area under the ROC curve is a measure of how well a parameter can distinguish between two diagnosis groups.

2.1 Notations of ROC

When you consider the results of a particular test in two populations, one population with a disease, the other population without the disease, you will rarely observe a perfect separation between the two groups. Indeed, the distribution of the test results will overlap, as shown in the following Figure 2.1.

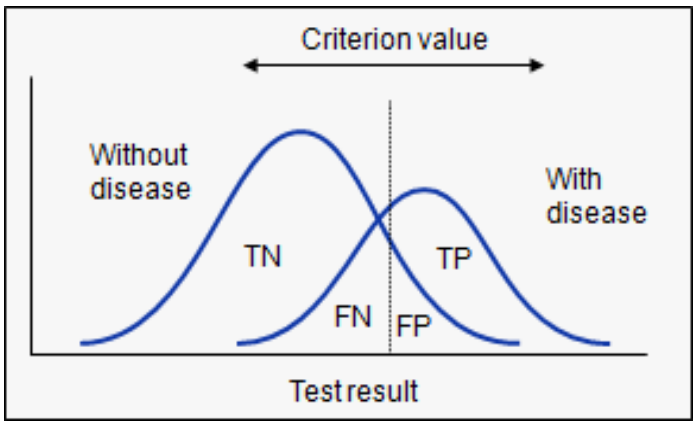

Figure 2.1 Two distributions of two groups

For every possible cut-off point or criterion value (threshold value) you select to discriminate between the two populations, there will be some cases with the disease correctly classified as positive (TP = True Positive fraction), but some cases with the disease will be classified negative (FN = False Negative fraction). On the other hand,

some cases without the disease will be correctly classified as negative (TN = True Negative fraction), but some cases without the disease will be classified as positive (FP = False Positive fraction).

The different fractions (TP, FP, TN, FN) are represented in the following table.

Table 2.1 Different fractions (TP, FP, TN, FN)

Test Result	**Diagnosis**	
	Positive	**Negative**
Positive	*TP*	*FP*
Negative	*FN*	*TN*

There are some criteria to measure the performance of the diagnostic test. Sensitivity, Specificity, Efficiency and Precision are some of the performance criteria. These criteria are based on the Table 2.1.

Sensitivity is a probability that a test result will be positive when the disease is present. It is equal to TP fraction. Specificity is a probability that a test result will be negative when the disease is not present. It is equal to 1-FP fraction. Efficiency is calculated by total number of TP and FP over sample size. It gives a clue about the accuracy of the diagnostic test. Precision is the positive predictive value, which is defined probability that the disease is present when the test is positive. It is calculated as the formula given below,

$$\text{Precision} = (\textit{Number of True Positives}) \;/\; (\textit{Number of True Positives} + \textit{Number of False Positives})$$

Alteration of specificity and sensitivity can be shown in Figure 2.2. When you select a higher criterion value (threshold value), the false positive fraction will decrease with increased specificity but on the other hand the true positive fraction and sensitivity will decrease.

When you select a lower criterion value (threshold value), then the true positive fraction and sensitivity will increase. On the other hand the false positive fraction will also increase, and therefore the true negative fraction and specificity will decrease.

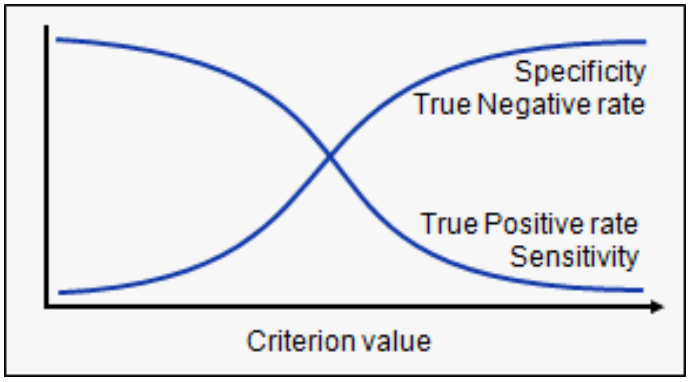

Figure 2.2 Alteration of specificity and sensitivity for different criterion values.

2.2 The Legitimacy of the Test

After constructing a Table 2.1, the legitimacy of the test (in order to check if a test is good or not) should be calculated. If a test is legitimate, a patient with a positive test is at least as likely to have a positive diagnosis as those with a negative test (Westin, 2001). The Chi-square test is used to check the legitimacy of the test. To use the common Chi-square test one has to compute a test statistic and compare it to a value. If the statistic is above a certain value (3.85 at 5% significance level, 6.63 at 1% significance level, and 10.83 at 0.1% significance level) the diagnostic test is a legitimate one. Chi-Square test depends on sample size of the test (N) and Kappa values. ($\kappa(1,0)$ and $\kappa(0,0)$) The test statistic can be computed as follows (Kraemer, 1992).

$$\chi^2 = N.K(1,0).K(0,0)$$

where

$$\kappa(1,0) = (SE - (TP\ fraction + FP\ fraction)) \ / \ (FN\ fraction + TN\ fraction)$$

$$\kappa(0,0) = (SP - (FN\ fraction + TN\ fraction)) \ / \ (TP\ fraction + FP\ fraction)$$

$\kappa(1,0)$ and $\kappa(0,0)$ are called quality indices (Kappa values) for the test. These values confirm the suspicion that the sensitivity has better quality than specificity (when $\kappa(1,0)$ is greater than $\kappa(0,0)$) or the specificity has better quality than sensitivity (when $\kappa(0,0)$ is greater than $\kappa(1,0)$).

2.3 ROC Graph

In a Receiver Operating Characteristic (ROC) curve the true positive fraction (TP or sensitivity) is plotted in function of the false positive fraction (FP or 1-specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/1-specificity pair corresponding to a particular decision threshold. FP fraction amounts to "*costs*" and TP fraction amounts to "*benefits*". A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left (northwest) corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left (northwest) corner, the higher the overall accuracy of the test (Zweig & Campbell, 1993).

In order to perform ROC curve analysis researcher should have a measurement of interest and an independent diagnosis which classifies the study subjects into two distinct groups: a diseased and non-diseased group. The latter diagnosis should be independent from the measurement of interest. For every study subject enter a code for the diagnosis as follows: 1 for the diseased, and 0 for the non-diseased. Therefore, each subject is mapped to one element of the set of positives and negatives. This kind of diagnosis coding is binary. After diagnosis process, we applied the diagnostic test. Our purpose is to predict subjects' condition. Consequently, the test result is referred to as predicted class. A classifier is a mapping from the subjects to the predicted classes. If the test result obtained from the diagnostic test is discrete, then the classifier is discrete classifier. Discrete classifier

produces only one point in the ROC graph. Therefore, all points in Figure 2.3 can be identified as corresponding to discrete classifiers.

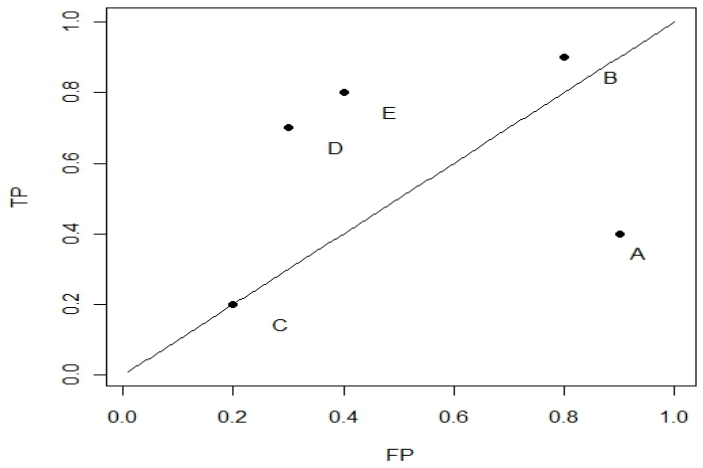

Figure 2.3 ROC graph for 5 points

The Figure 2.3 consists of 5 points. All of them have different benefits and costs. In that case, which point is the "*best*" to choose? At a first glance; A, B and C points can be eliminated quickly, as A point has high cost value and low benefit value, B point has high benefit value and high cost value and C point has low cost value and low benefit value. However, it cannot be said that E point or D point is the "*best*" choice, as E point has higher benefit value than D point has, but at the same time, D point has lower cost value than E point has. In this situation, the best decision is subjective.

If the test scores are continuous, ROC graph is called ROC curve. In this situation, the classifier is called a probabilistic classifier even if the test scores are not "*probabilities*". When the test scores are continuous, threshold values must be specified with respect to test scores. A threshold value is a cut-off point and their values range from minus infinity to plus infinity. Each threshold value generates different points of the ROC graph and these points build up the ROC curve.

In Table 2.2, an example of five subjects diagnosis and test scores are shown. The corresponding ROC curve is shown in Figure 2.4.

Table 2.2 Diagnosis and test scores for 5 subjects

Subjects	Diagnosis	Test Scores
1	Positive (1)	3.2
2	Positive (1)	2.4
3	Negative (0)	1.6
4	Positive (1)	0.9
5	Negative (0)	0.3

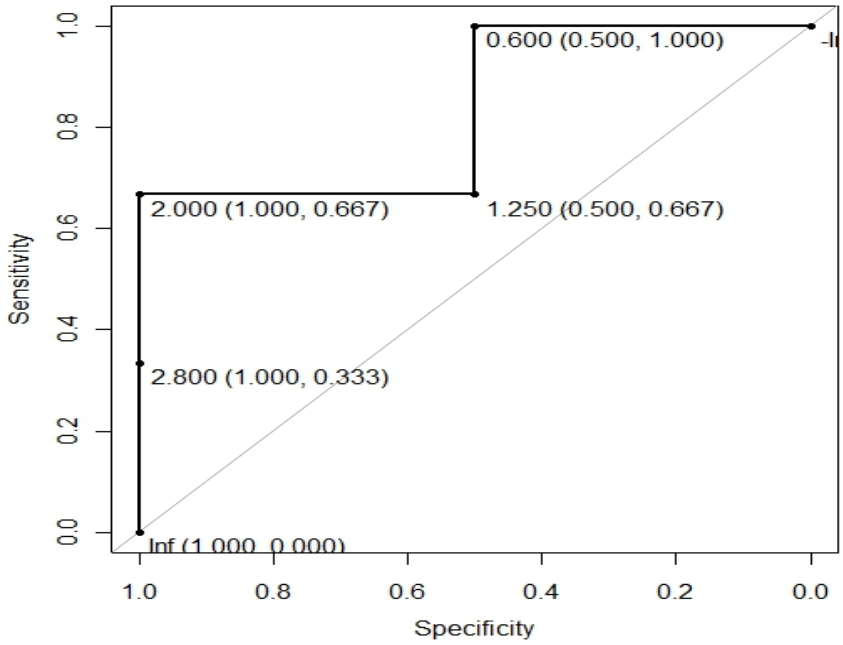

Figure 2.4 The ROC curve for the values shown in Table 2.2

In Figure 2.4, example with six different decision thresholds is shown. Before analysing them, we must keep in mind that if a score value is greater than the threshold value then we decide that the corresponding subject is positive. If score value is lower than the threshold value then we decide that the corresponding subject is negative, as we consider that a high score indicates a more likely ill person and a low score indicates a more likely healthy person.

i) Let the threshold value be plus infinitive, all scores will be lower than threshold value. So, we decide that all subjects are negative. Therefore, our contingency table becomes as below.

Table 2.3 2x2 Contingency table when the threshold value is plus infinitive

Test Result	**Diagnosis**	
	Positive	**Negative**
Positive	0	0
Negative	3	2

For Table 2.3, Specificity= (Number of TN) / (Number of TN + Number of FP) =2/2=1 and Sensitivity= (Number of TP) / (Number of TP + Number of FN)=0/3=0.

ii) Let threshold value be 2.8 (Midpoint between 3.2 and 2.4 test scores), first subject score is greater than this threshold value and the other subject scores are lower than this threshold value. So, we decide that the first subject is positive and the other subjects are negative. Therefore, our contingency table becomes as below.

Table 2.4 2x2 Contingency table when the threshold value is 2.8

Test Result	**Diagnosis**	
	Positive	**Negative**
Positive	1	0
Negative	2	2

For Table 2.4, Specificity=2/2=1 and Sensitivity=1/3=0.33.

iii) Let threshold value be 2.0 (Midpoint between 2.4 and 1.6), first and second subjects scores are greater than this threshold value and the other subject scores are

lower than this threshold value. So, we decide that the first and second subjects are positive and the other subjects are negative. Therefore, our contingency table becomes as below.

Table 2.5 2x2 Contingency table when the threshold value is 2.0

Test Result	**Diagnosis**	
	Positive	**Negative**
Positive	2	0
Negative	1	2

For Table 2.5, Specificity=2/2=1 and Sensitivity=2/3=0.66.

iv) Let threshold value be 1.25 (Midpoint between 1.6 and 0.9), first, second and third subjects score are greater than this threshold value and the other subjects scores are lower than the threshold value. So, we decide that first, second and third subject are positive and the other subjects' are negative. Therefore, our contingency table becomes as below**.**

Table 2.6 2x2 Contingency table when the threshold value is 1.25

Test Result	**Diagnosis**	
	Positive	**Negative**
Positive	2	1
Negative	1	1

For Table 2.6, Specificity=1/2=0.5 and Sensitivity=2/3=0.66.

v) Let threshold value be 0.6 (Midpoint between 0.9 and 0.3), fifth subject score is lower than threshold value and the other subjects scores are greater than the threshold

value. So, we decide that the fifth subject is negative and the other subjects' are positive. Therefore, our contingency table becomes as below.

Table 2.7 2x2 Contingency table when the threshold value is 0.6

Test Result	**Diagnosis**	
	Positive	**Negative**
Positive	3	1
Negative	0	1

For Table 2.7, Specificity=1/2=0.5 and Sensitivity=3/3=1.

vi) Let threshold value be minus infinitive, all scores are greater than threshold value. So, we decide that all subjects are positive. Therefore, our contingency table becomes as below.

Table 2.8 2x2 Contingency table when the threshold value is minus infinitive

Test Result	**Diagnosis**	
	Positive	**Negative**
Positive	3	2
Negative	0	0

For Table 2.8, Specificity=0/2=0 and Sensitivity=3/3=1.

2.4 Some Comments On ROC Curves Shapes

In this section we analyse some possible scenarios, concerning ROC curves. The curves presented here are only for illustration purposes and are not base on any

practical examples. They were drawn to show possible different situations regarding ROC curves.

Assume that there are two different ROC curves passing point A and point B separately (Figure 2.5). B curve has a higher slope for some FP values and is all above A curve. Hence, B curve is better because it has higher TP fraction for the same FP fraction values. Similarly, it has lower FP fraction for the same TP fraction.

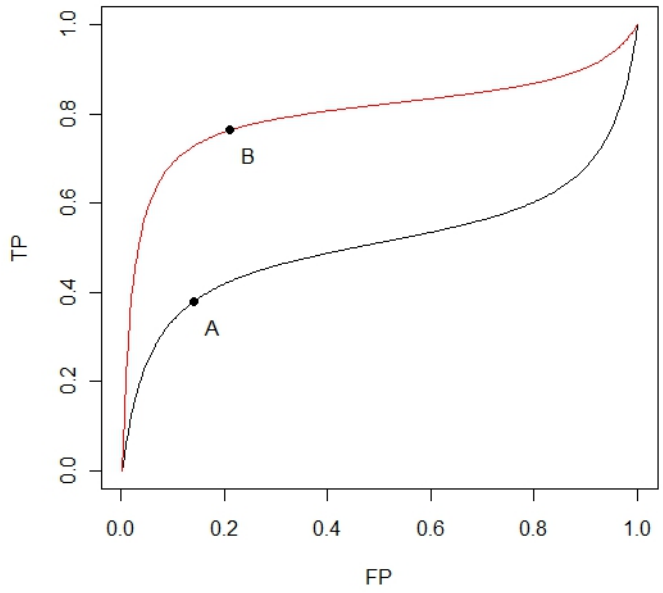

Figure 2.5 No intersection between two ROC curves

A second situation is when we have curve intersection. What if two curves intersect? (Figure 2.6) It is not easy to interpret Figure 2.6 because both curves have an advantage for certain TP fractions and FP fractions. This situation is not always a problem because the intrinsic characteristics of the disease that is under study can help in determining which curve is more suitable. (See Down Syndrome Example to be presented in Section 2.7).

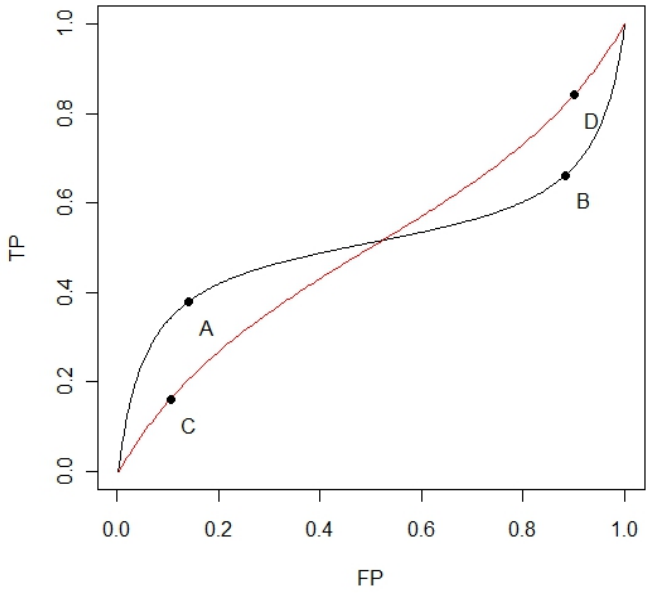

Figure 2.6 Intersection between two ROC curves

2.5 Area Under the Curve, AUC

The total area under the ROC curve is a measure of the performance of the diagnostic test since it reflects the test performance at all possible cut-off levels. The area lies in the interval [0.5, 1] and the larger area, the better performance. There are several ways to calculate the area under a ROC curve. First, the trapezoidal rule can be used but gives an underestimation of the area. Second, it is possible to get a better approximation of the curve by fitting the data to a binormal model with maximum-likelihood estimates. After that it is possible to get a better estimate of the area. This is done, for example, in the program Rockit (Rockit, 2002). A third way to calculate the area is to use the Mann-Whitney U statistic (also known as the non-parametric Wilcoxon statistic). That is, no assumptions on the distributions of the data are done since Wilcoxon is a distribution-free statistic (Bamber, 1975, Hanley and McNeil, 1982).

2.6 AUC Use

Our purpose is to learn the difference between the distributions of the tests scores for healthy persons and ill persons. Roughly, if these two distributions have large overlapping, the difference between these two distributions is low. Small overlapping between the distributions is better than large overlapping, because we want to detect the difference between actually healthy people and actually ill people.

There are three different curves in Figure 2.7. The black curve is the most desirable one. Because it is the closest curve to the northwest point and it also corresponds to a small overlapping of the test distributions (It can be seen in Section 2.8). The blue curve is the most undesirable one. AUC value of the blue curve is almost 0.5.

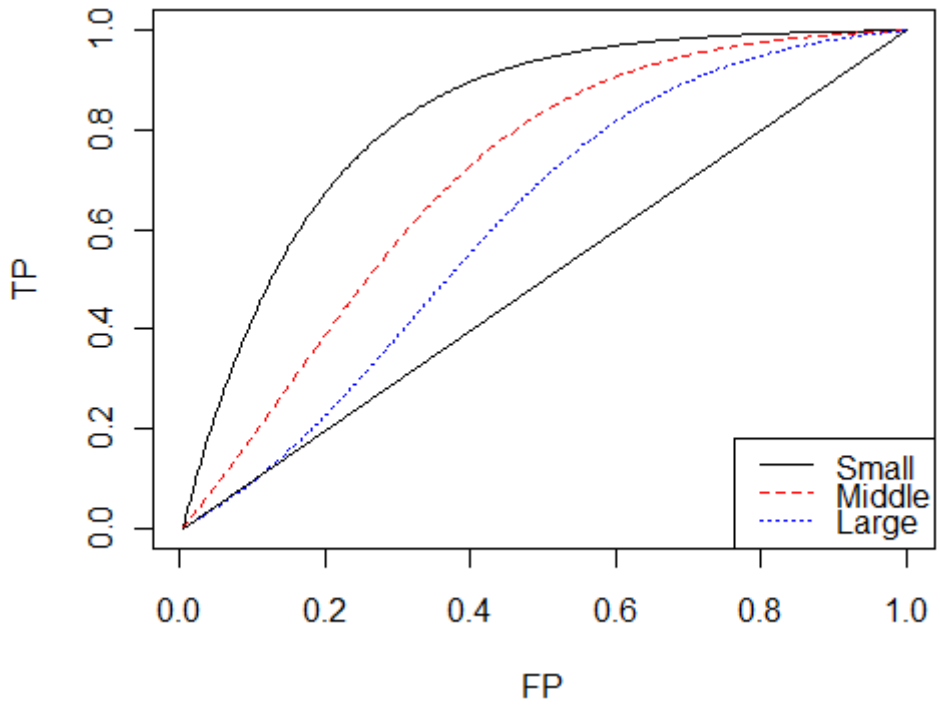

Figure 2.7 Overlapping of three ROC curves

2.7 Down Syndrome Example

AUC value may not be helpful at all times. This happens when the aim of the study is more important than AUC value, for example, if we study about Down Syndrome, the detection of a healthy fetus will be more vital than the detection of an ill fetus, for Down Syndrome can be determined before birth and it is not desirable to have an abortion of a healthy child. In this situation, specificity value will be more important than sensitivity value.

Suppose that there are two curves in a toy example are about Down Syndrome (Figure 2.8). AUC value of the blue curve is 0.77 and AUC value of the green curve is 0.61. So, the blue curve has greater AUC value than the green curve. On the other hand, the green curve has greater specificity value while sensitivity is under 0.25. For example; while sensitivity value is almost 0.25 for both curves, specificity value of the green curve (AUC=0.61) is 0.89 and specificity value of the blue curve (AUC=0.77) is 0.88. In the circumstances, the green curve is more suitable than the blue curve.

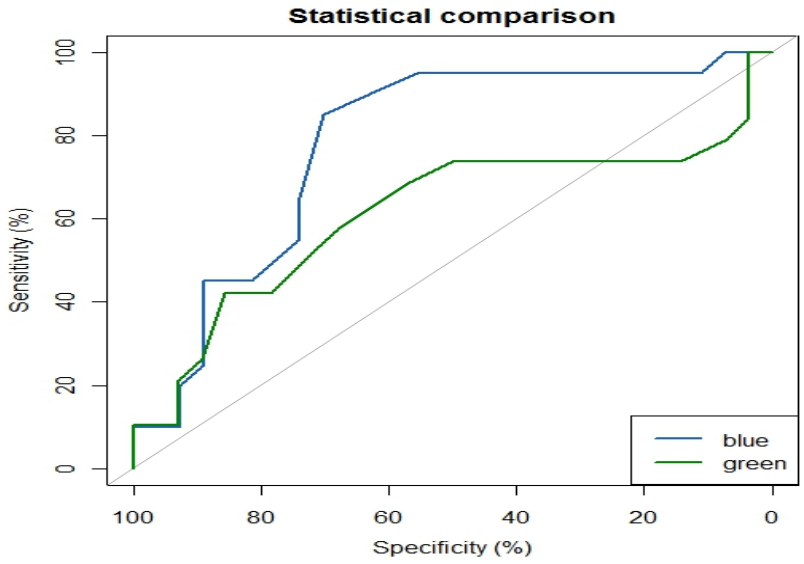

Figure 2.8 Two ROC curves about Down Syndrome

2.8 Demonstration For Overlapping and Cut-Off

Test results of the healthy person and the sick person have district distributions. In ROC analysis, it can be determined how similar the distributions of the test results for healthy people and sick people are.

The area under the curve for the distribution of test results of healthy people is equal to 1. This area can be decomposed in the FP fraction and TN fraction (Hence, FP+TN=1). The area under the curve for the distribution of test results of sick people is also equal to 1, consisting of FN fraction and TP fraction (Hence, FN+TP=1).

Threshold value is a cut-off point. The notion of threshold is related to test scores. Therefore, threshold values are specified with respect to test scores. When the threshold values change, alteration of TP fraction and FP fraction can be observed (Figure 2.9 and Figure 2.10). The change of the threshold does not affect FP fraction + TN fraction and FN fraction + TP fraction. These values are always equal to 1. Besides, the change of the test threshold does not affect the value of AUC as well. The curve separation and AUC value do not change, although test threshold changes (Figure 2.9 and Figure 2.10).

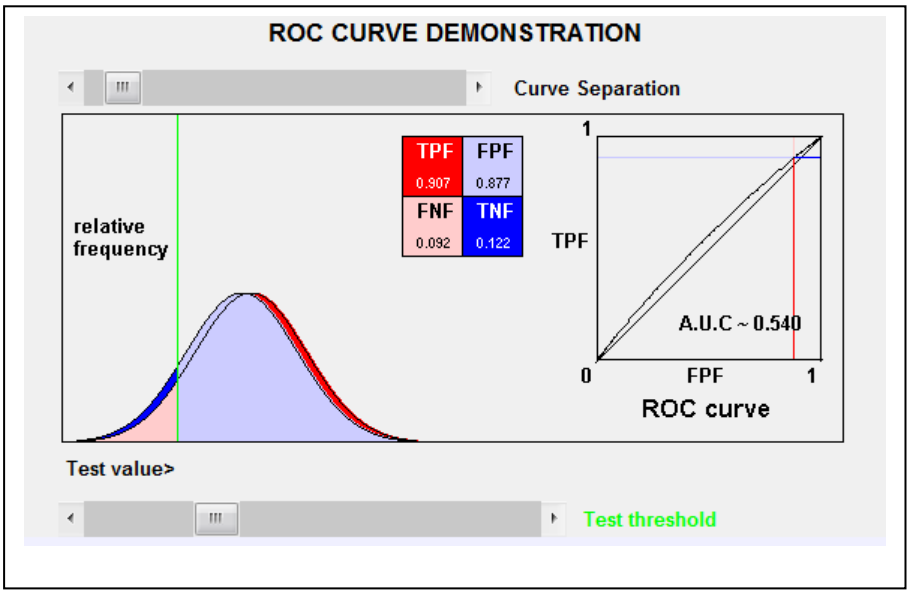

Figure 2.9 ROC curve demonstration

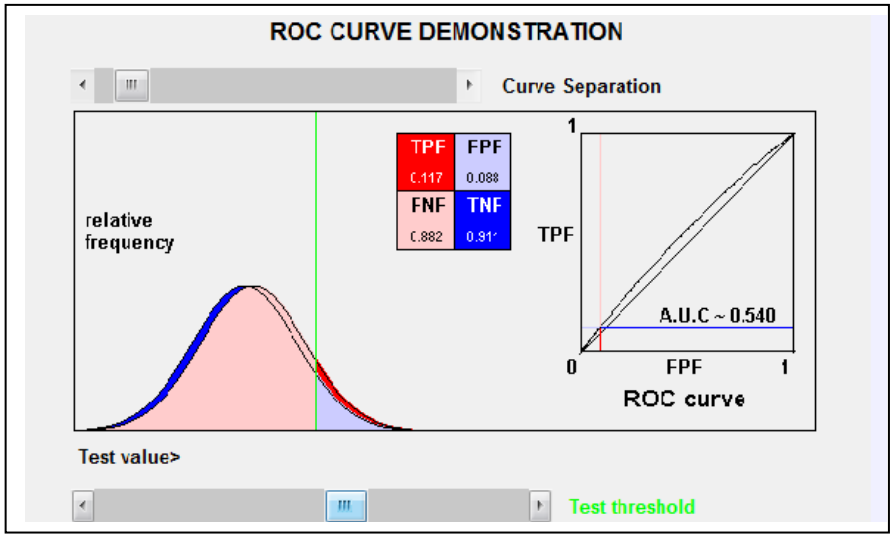

Figure 2.10 ROC curve demonstration

Cut-off points (threshold values) are selected arbitrary. Whether the cut-off point (threshold value) is an optimal point or not depends on the disease that we are studying on. However, the optimal cut-off point is generally accepted as the nearest point to the northwest corner.

As it can be seen on Figure 2.11, Figure 2.12 and Figure 2.13, after the threshold value was changed, TP fraction and TN fraction values were changed too. If we assume TP fraction and TN fraction value on the either pan of the scales, we can conclude that while the one increases the other one decreases at the same time.

When threshold value goes to the left side, TP fraction increases but TN fraction decreases (Efficiency is Efficiency1= (0.967+0.414) / N). When threshold value goes to the right side, TP fraction decreases but TN fraction increases. (Efficiency is Efficiency2=(0.295+0.985) / N) When threshold value goes to the intersection point of these two distributions, Efficiency is Efficiency3=(0.782+0.798) / N. It is greater than Efficiency1 and Efficiency2.

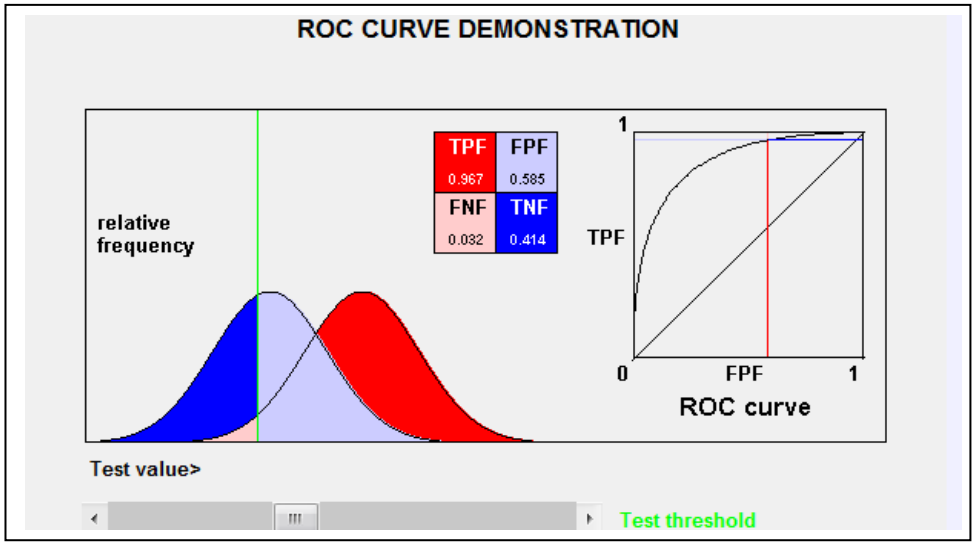

Figure 2.11. ROC curve demonstration

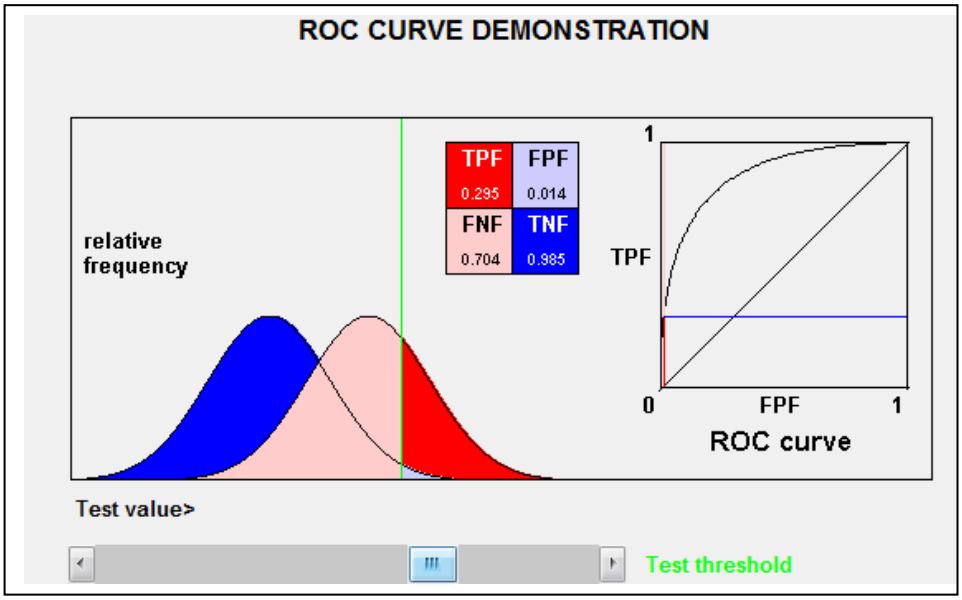

Figure 2.12 ROC curve demonstration

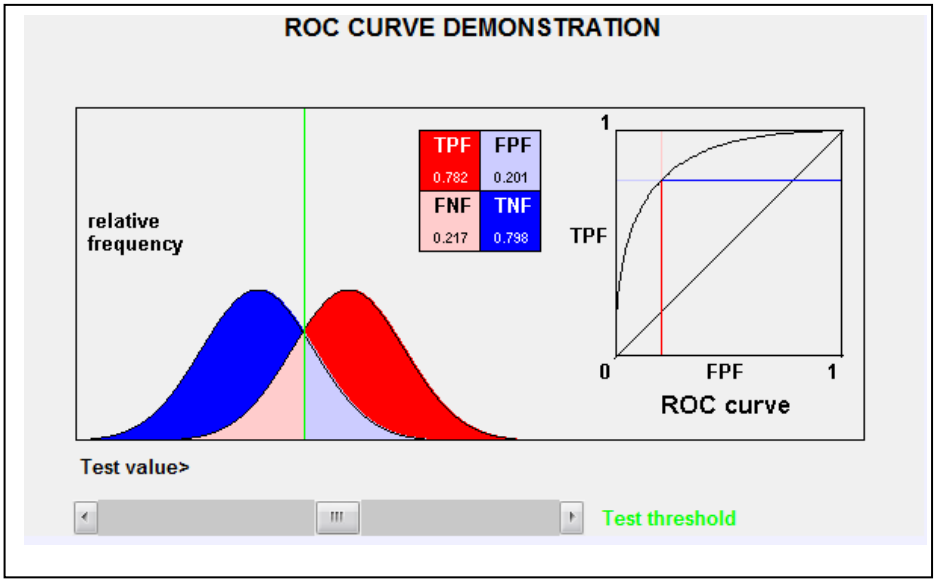

Figure 2.13 ROC curve demonstration

CHAPTER THREE
BASIC CONCEPTS OF INFORMATION THEORY

The performance of a diagnostic test is frequently described in terms of the amount of information it provides. A fundamental concept of information theory, relative entropy and mutual information, is directly applicable to evaluation of diagnostic test performance. In this chapter we introduce most of the basic definitions in information theory required for evaluation of diagnostic test performance.

3.1 Entropy, Relative Entropy and Mutual Information

This section briefly defines Shannon entropy, relative entropy (Kullback Leibler divergence) and mutual information. The entropy of a random variable is a measure of the uncertainty of the random variable. It is the number of bits on average required to describe the random variable. Let X be a discrete random variable, taking a finite number of possible values $X_1,...,X_n$ with respective probabilities $p_i \geq 0$ for $i = 1,...,n$ and $\sum_{i=1}^{n} p_i = 1$ The Shannon entropy $H(X)$ is defined by Cover and Thomas (2006).

$$H(x) = -\sum_{x} p(x) \log_2 p(x)$$

The joint entropy measures how much uncertainty is contained in a joint system of two random variables. If the random variables are X and Y, the joint entropy H(X, Y) given in Cover and Thomas (2006) is

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y)$$

If $(X,Y) \sim p(x,y)$, the conditional entropy $H(X \mid Y)$ is defined as

$$H(Y \mid X) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y \mid x)$$

The relative entropy (Kullback Leibler divergence) $D(p \backslash\backslash q)$ is an appropriate measure of the similarity of the underlying distribution. It may be calculated from

$$D(p \backslash\backslash q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

The properties of the relative entropy equation make it non-negative and it is zero if both distributions are equivalent namely, $p = q$. The smaller the relative entropy is the more similar the distribution of the two variables and vice versa (Cover and Thomas, 2006).

Mutual information is a special case of a more general quantity called relative entropy, which is a measure of the distance between two probability distributions. The mutual information of two random variables is a quantity that measures the mutual dependence of the two variables. The interpretation is that when mutual information is absent, marginal distributions are independent and their entropies add up to total entropy. Mutual information $I(X;Y)$ for the random variables X and Y is evaluated by the formula as below.

$$I(X;Y) = H(X) - H(X \backslash Y)$$

3.2 Information-Based Measures of Diagnostic Test Performance

The performance of a diagnostic test is frequently described in terms of the amount of information it provides. The purpose of this chapter is to demonstrate how basic concepts in information theory apply to the problem of quantifying diagnostic test performance.

On the condition that there is a bilateral class result as disease positive / negative and the test result positive / negative, a crosstab is created. The crosstab is as below.

Table 3.1 Relations between prevalence and the level of a test.

Test Result	**Diagnosis**		
	Positive	**Negative**	
Positive	Number of TP	Number of FP	$Total\ C$
Negative	Number of FN	Number of TN	$Total\ D$
	$Total$ A	$Total\ B$	$Total\ E$

Let p_i be the probability that patient i will get a positive diagnosis and q_i be patient i's probability of a positive test.

P is called as the prevalence and it is equal to $mean(p_i)$ value. It is also called as pretest probability of disease.

$$P = (Total\ A)\ /\ (Total\ E)$$

Q is called as the level of the test and it is equal to $mean(q_i)$ value.

$$Q = (Total\ C)\ /\ (Total\ E)$$

We also define $P' = 1 - P$ and $Q' = 1 - Q$

TP: If both diagnosis and test are positive, it is called a true positive.

FP: If diagnosis is negative and test is positive, it is called a false positive.

FN: If diagnosis is positive and test is negative, it is called a false negative.

TN: If both diagnosis and test are negative, it is called a true negative.

We are going to mention about 5 criteria for the calculation of the test quality. These criteria are also used in the ROC theory.

Sensitivity: The possibility of finding sick people among actually sick people.

$$SE = number\ of\ TP\ /\ (number\ of\ TP + number\ of\ FN) = number\ of\ TP\ /\ Total\ A$$

Specificity: The possibility of finding health people among actually healthy people.

$$SP = number\ of\ TN\ /\ (number\ of\ FP + number\ of\ TN) = number\ of\ TN\ /\ Total$$

Efficiency is defined as follows.

$$EFF = (number\ of\ TP + number\ of\ TN)\ /\ (Total\ E)$$

The Predictive Value of A Positive Test: Positive test result gives the possibility of being ill.

$$PVP = number\ of\ TP\ /\ (number\ of\ TP + number\ of\ FP) = number\ of\ TP\ /\ Total\ C$$

The Predictive Value of A Negative Test: Negative test result gives the possibility of not being ill.

$$PVN = number\ of\ TN\ /\ (number\ of\ TN + number\ of\ FN) = number\ of\ TN\ /\ Total\ D$$

While evaluating the performance of the diagnosis test using the information theory, we need to explain the concepts of test results and disease statement. Disease statement is denoted by D. On the condition that there are two statements such as the existence or the non-existence of a disease, we can specify the disease statement as follows,

$$D = \{D_i\} \qquad i : \{+, -\}$$

$D+$ = Get ill before diagnosis test

$D-$ = Get not ill before diagnosis test

The probability distribution of the disease statement before the test is defined with Prevalence ($P(D+)$) and 1-Prevalence ($P(D-)$) values. In this case, the entropy before the test is calculated as below.

$$H(D) = P(D+)\log_2 P(D+) + P(D-)\log_2 P(D-)$$

After the diagnosis test is applied, the uncertainty of the disease statement changes. On the condition that the diagnosis test results are known, the entropy of the disease statement is called conditional entropy and is calculated according to the formula below.

$$H(D\backslash T) = P(T+)[P(D+\backslash T+)\log_2 P(D+\backslash T+) + P(D-\backslash T+)\log_2 P(D-\backslash T+)] + P(T-)[P(D+\backslash T-)\log_2 P(D+\backslash T-) + P(D-\backslash T-)\log_2 P(D-\backslash T-)]$$

If $H(D)$ is defined as pretest entropy, we need to define $H(D\backslash T)$ as the expected value of posttest entropy (Benish, 2009). Besides, the difference between $H(D)$ and $H(D\backslash T)$ is called as Mutual Information. Mutual Information is denoted by $I(D;T)$. It is the reduction in the uncertainty of D due to the knowledge of T. Mutual Information is the general criterion of what the diagnosis test will tell us.

In diagnosis tests, the distance between positive posttest entropy and pretest entropy is relative entropy. This distance is called as the information content which is provided with positive result and it is denoted by $I\ pos.$ Similarly, the distance between negative posttest entropy and pretest entropy is called information content which is provided with negative result and it is denoted by $I\ neg.$ The concepts of $I\ pos.$ and $I\ neg.$ are calculated according to the formulas below.

$$I\ pos. = (1/B)\ P\ SE\ \log_2 SE + (1/B)(1-P)(1-SP)\ \log_2 (1-SP)\ -\ \log_2 B$$

$$I\ neg. = (1/B)^{-1}\ P\ (1-SE)\ \log_2(1-SE) + (1/B)^{-1}(1-P)(SP)\ \log_2 (SP)\ -\ \log_2(1-B)$$

Where B is calculated by the formula of $SE.P + (1-SP).(1-P)$. B is directly equal to Q (the level of the test) value.

The value of mutual information can be finding by using I. pos. and I. neg. formulas. The relationship between mutual information MI and $I\ pos.$/ $I\ neg.$ is shown the formula given below.

$$MI = B\ I\ pos.\ +\ (1 - B)\ I\ neg.$$

Kullback Leibler divergence is helpful for clinicians in interpreting diagnostic test result. It measures the distance between the diased $(d+)$ and the non-diased $(d-)$ distributions.

Taking the non-diased distribution as the reference, Kullback Leibler divergence is denoted by the following equation:

$$D(d+ // d-) = \sum_{i=1}^{k} p(d+)_i \log \frac{p(d+)_i}{p(d-)_i}$$

Taking the diased distribution as the reference, The Kullback Leibler divergence is denoted by the following equation:

$$D(d- // d+) = \sum_{i=1}^{k} p(d-)_i \log \frac{p(d-)_i}{p(d+)_i}$$

$D(d+ \backslash\backslash d-)$ and $D(d- \backslash\backslash d+)$ are the criteria for evaluating diagnostic test performances. These values are affected by the $p(d+)$ and the $p(d-)$. Therefore they can be interpreted as the "*before test*" potentials of ruling in and ruling out disease, respectively (Lee, 1999).

Positive results from a test, which can rule a diagnosis in, means a high probability of the presence of disease, if the test is highly specific. The value of $D(d+ \backslash\backslash d-)$ which is called "*rule in potential*" is the criterion showing the existence of the disease.

Negative results from a test, which can rule a diagnosis out, means a high probability of the absence of disease, if the test is highly sensitive. The value of

$D(d-\backslash\backslash d+)$ which is called "*rule out potential*" is the criterion showing the non-existence of the disease.

Provided that test results have binary class, the formulas for the concepts of $D(d+\backslash\backslash d-)$ and $D(d-\backslash\backslash d+)$ are as below (Lee, 1999).

$$D(d+\backslash\backslash d-) = (1-SE)\log\frac{1-SE}{SP} + (SE)\log\frac{SE}{1-SP}$$

$$D(d-\backslash\backslash d+) = (SP)\log\frac{SP}{1-SE} + (1-SP)\log\frac{1-SP}{SE}$$

CHAPTER FOUR
APPLICATION

Turbidimetric test and Nefelometric test are used for the diagnosis of rheumatic disorder. Both tests are based on the principal of impurity in the blood. Nefelometric test is accepted as "*the gold standard*" in the analysis of plasma protein with micro molecule of which molecule massiveness is measured with milligram. New Turbidimetric tests are alternatives to Nefelometric test and they are becoming more precise day by day for the specific proteins such as ASO. Furthermore, while the unit cost of the Nefelometric test is more than the unit cost of the Turbidimetric test, there are disadvantages such as the requirement of more space in the laboratory, occupying additional personnel and orientation of them. There are no significant differences between those two tests with regard to the duration of test results. Each laboratory is required to decide to work with whether Turbidimetric test and Nefelometric test due to its substructure, patient potential and establishment requirement.

ASO values are used for the diagnosis of rheumatic disorder. ASO is a value which is used to learn whether the patients have group "*A*" beta-hemolytic streptococcal infection which causes these diseases. In this study, ASO values of 68 subjects (September/December-2011) who applied to Istanbul Mehmet Akif Ersoy Thoracic and Cardiovascular Surgery Training and Research Hospital for the diagnosis of rheumatic disorder were used. ASO values were evaluated according to Turbidimetric tests of two different firms. Since the names of the firms were kept secret, these tests were called as I. Turbidimetric test and II. Turbidimetric test. Both ROC and Information Theory analyses were applied to the results. Therefore, both firms' Turbidimetric test diagnostic test performances were evaluated and which diagnostic test had better performance was determined.

4.1 ROC Curves and AUC Values for Two Tests

At the beginning of the study, diagnoses are determined with respect to Nefelometric test results. If Nefelometric test results are in the range of 0-200 IU/ml

reference interval, the diagnosis is resulted as "*healthy*" for the person. If Nefelometric test results are over 0-200 IU/ml reference interval, the diagnosis is resulted as "*ill*" for the person.

Diagnosis values, I. Turbidimetric test results and II. Turbidimetric test results are coded as vectors in R programme (See Appendix 1). After coding process, ROC curves of the both tests are generated in the Figure 4.1 (See Appendix 2).

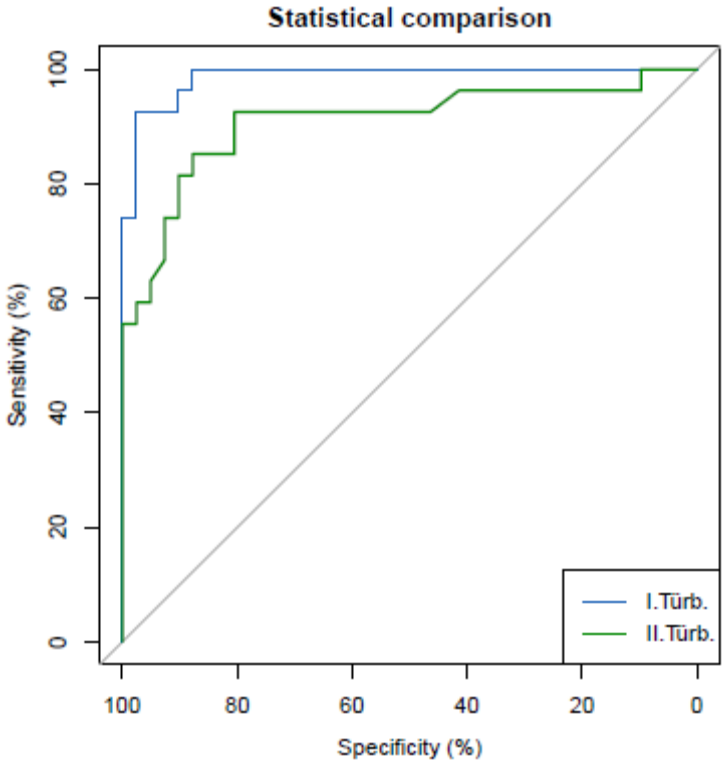

Figure 4.1 Statistical comparison for two ROC curves

In the Figure 4.1, it is observed that I. Turbidimetric test dominated II. Turbidimetric test for all sensitivity and specificity values. This observation is verified in the Table 1 comparing of AUC values (See Appendix 3 and Appendix 4). According to these results, AUC is calculated 0.98 for I. Turbidimetric test, 0.90 for II. Turbidimetric test.

Table 4.1 AUC of two tests

Tests	**AUC**
I. Turbidimetric	0.98
II. Turbidimetric	0.90

4.2 Sensitivity, Specificity and Chi-Square Results for I. Turbidimetric Test

In the Table 4.2; sensitivity, specificity and chi-square values are compared for three threshold values.

Table 4.2 SE, SP and Chi-Square of first Turbidimetric test for different threshold values

I. Turbidimetric			
Threshold V.	***SE***	***SP***	**Chi-Square**
165	0.96	0.90	49.45
197	0.77	0.97	42.21
173	0.92	0.97	56.02

According to the results deduced from Table 4.2, the threshold value that makes SE maximum (except the value "*1.00*") is for 165 (See Appendix 5) and the threshold values that make SP maximum (except the value "*1.00*") are for 173 and 197 (See Appendix 6).

SE value of I. Turbidimetric test for the threshold 165 is bigger than SE values of I. Turbidimetric test for the threshold 173 and the threshold 197. It can be said that I. Turbidimetric test for the threshold 165 can select actually ill people better than I. Turbidimetric test for the threshold 173 and the threshold 197.

SP value of I. Turbidimetric test for the threshold 173 and the threshold 197 is bigger than SP value of I. Turbidimetric test for the threshold 165. It can be said that I. Turbidimetric test for the threshold 173 and the threshold 197 can select actually healthy people better than I. Turbidimetric test for the threshold 165.

SP value for the threshold 173 and the threshold 197 are both equal. In that case, the problem is to determine the optimal choice. The threshold 173 has bigger *SE* and Chi-Square values than the threshold 197 which is the closest value to "*200 utmost reference point*". In the circumstances, the threshold 173 can be advised to the clinicians to replace with the threshold 197 as the upmost reference point.

For the legitimacy of both tests, the results in Table 4.2 should be interpreted. It can be concluded that Chi-square values of all threshold values are legitimate in Table 4.2 (See Appendix 7). Since all Chi-square values are bigger than certain values (3.85 at 5% significance level, 6.63 at 1% significance level, and 10.83 at 0.1% significance level).

4.3 Sensitivity, Specificity and Chi-Square Results for II. Turbidimetric Test

In the Table 4.3; sensitivity, specificity and chi-square values are compared for three threshold values.

Table 4.3 SE, SP and Chi-Square of second Turbidimetric test for different threshold values

II. Turbidimetric			
Threshold V.	***SE***	***SP***	**Chi-Square**
48	0.96	0.41	11.92
202	0.51	1.00	26.77
101	0.92	0.78	32.49

According to the results deduced from Table 4.3, the threshold value that makes *SE* maximum (except the value "*1.00*") is for 48 (See Appendix 8) and the threshold values that make *SP* maximum is for 202 (See Appendix 9).

SE value of II. Turbidimetric test for the threshold 48 is bigger than *SE* values of II. Turbidimetric test for the threshold 202 and the threshold 101. It can be said that II. Turbidimetric test for the threshold 48 can select actually ill people better than II. Turbidimetric test for the threshold 202 and the threshold 101.

SP value of II. Turbidimetric test for the threshold 202 is bigger than *SP* value of II. Turbidimetric test for the threshold 48 and the threshold 101. It can be said that II. Turbidimetric test for the threshold 202 can select actually healthy people better than II. Turbidimetric test for the threshold 48 and the threshold 101.

The threshold 101 has bigger *SE* and Chi-Square values than the threshold 202 which is the closest value to "*200 utmost reference point*". In the circumstances, the threshold 101 can be advised to the clinicians to replace with the threshold 202 as the upmost reference point.

For the legitimacy of both tests, the results in Table 4.3 should be interpreted. It can be concluded that Chi-square values of all threshold values are legitimate in Table 4.3 (See Appendix 10). Since all Chi-square values are bigger than certain values (3.85 at 5% significance level, 6.63 at 1% significance level, and 10.83 at 0.1% significance level).

4.4 PPV, PVN and Efficiency Results for I. Turbidimetric Test

In the Table 4.4; positive predictive value, negative predictive value and efficiency values are compared for three threshold values.

Table 4.4 PVP, PVN and EFF of first Turbidimetric test for different threshold values

I. Turbidimetric			
Threshold V.	***PVP***	***PVN***	***EFF***
173	0.96	0.95	0.95
165	0.86	0.97	0.91
197	0.95	0.86	0.89

According to the results deduced from Table 4.4, *EFF* value of I. Turbidimetric test for the threshold 173 is bigger than *EFF* values of I. Turbidimetric test for the threshold 165 and the threshold 197. It can be said that I. Turbidimetric test for the

threshold 173 is more efficient and has less error than I. Turbidimetric test for the threshold 165 and the threshold 197 (See Appendix 11).

On the condition that the test result is positive, the possibility to diagnose the person as ill is the highest for the threshold 173; however, if the test result is negative, the possibility to diagnose the person as healthy is the highest for the threshold 165 (See Appendix 12 and Appendix 13).

4.5 PPV, PVN and Efficiency Results for II. Turbidimetric Test

In the Table 4.5; positive predictive value, negative predictive value and efficiency values are compared for three threshold values.

Table 4.5 PVP, PVN and EFF of second Turbidimetric test for different threshold values

II. Turbidimetric			
Threshold V.	***PVP***	***PVN***	***EFF***
101	0.73	0.94	0.83
202	1.00	0.75	0.80
48	0.52	0.94	0.63

According to the results deduced from Table 4.5, *EFF* value of II. Turbidimetric test for the threshold 101 is bigger than *EFF* values of II. Turbidimetric test for the threshold 202 and the threshold 48. It can be said that II. Turbidimetric test for the threshold 101 is more efficient and has less error than II. Turbidimetric test for the threshold 202 and the threshold 48 (See Appendix 14).

On the condition that the test result is positive, the possibility to diagnose the person as ill is the highest for the threshold 202; however, if the test result is negative, the possibility to diagnose the person as healthy is the highest for the threshold 101 and the threshold 48 (See Appendix 15 and Appendix 16).

4.6 Mutual Information Results of Two Tests

Before examining the results of the mutual information for four threshold values, the best threshold values are shown for two tests in Figure 4.2, respectively (See Appendix 17 and Appendix 18). The best threshold value is resulted as the threshold 172 for I. Turbidimetric test and as the threshold 101.5 for II. Turbidimetric test, as expected, because these values are the midpoint of the threshold 171 and the threshold 173 for I. Turbidimetric test and the midpoint of the threshold 101 and the threshold 102 for II. Turbidimetric test (In between section 4.2 and section 4.5, it is mentioned that the threshold 173 for I. Turbidimetric test and the threshold 101 for II. Turbidimetric test has the biggest Chi-square values and Efficiency values, respectively).

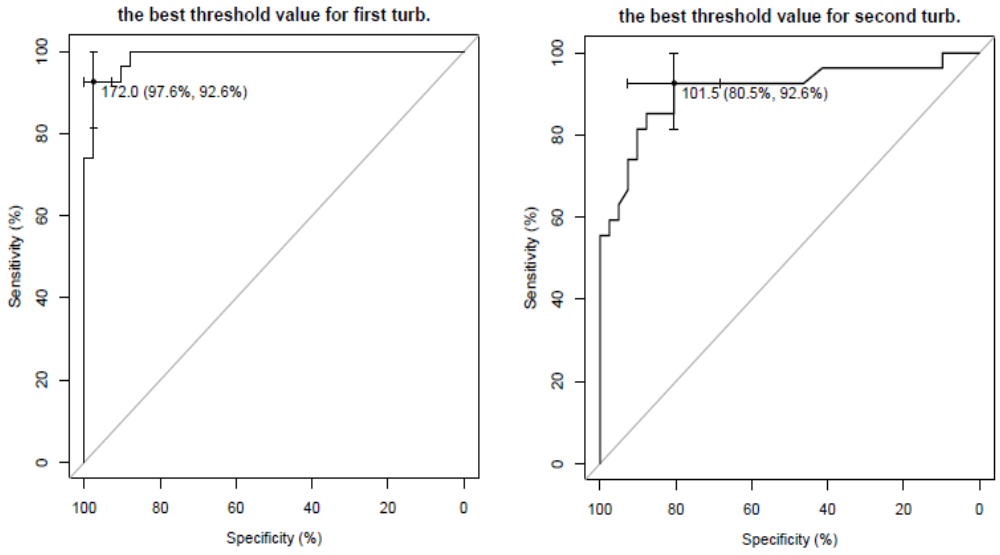

Figure 4.2 The best threshold values for two tests

After showing the best threshold values in Figure 4.2, Table 4.6 is generated. Table 4.6 represents four threshold values that make mutual information maximum for each test (See Appendix 19 and Appendix 20). Table 4.6 doesn't contain the threshold 165 for I. Turbidimetric test and the threshold 48 for II. Turbidimetric test. These threshold values have the biggest SE and PVN, they *don't* have the biggest mutual information values. These results prove that, for the overall quality, neither sensitivity nor specificity but the results of mutual information should be examined.

Table 4.6 MI of two tests for different threshold values

Threshold V.		***MI***	
I. Turb.	**II. Turb.**	**I. Turb.**	**II. Turb.**
173	102	0.70	0.41
142	118	0.67	0.41
185	124	0.64	0.40
171	101	0.64	0.39

4.7 Entropy, Conditional Entropy and Mutual Information Results of Two Tests

For constructing Table 4.9, ‘a vector’ and ‘b vector’ (See Appendix 1) are converted to binary class as ‘e vector’ and ‘f vector’, respectively. ‘d vector’ remained the same (See Appendix 21). After converting process; Table 4.7 and Table 4.8 are generated. With regard to Table 4.7 and Table 4.8; entropy, conditional entropy and mutual information are calculated in Excel programme using the formulas in Chapter Three (Table 4.9).

Table 4.7 2x2 contingency table of first Turbidimetric test

Test Result	**Diagnosis**	
	Positive	**Negative**
Positive	20	1
Negative	7	40

Table 4.8 2x2 contingency table of second Turbidimetric test

Test Result	**Diagnosis**	
	Positive	**Negative**
Positive	14	0
Negative	13	41

If the disease statement is taken as the random variable, the random variable is indicated either as the presence or the absence of the disease before the diagnostic test. According to this situation the entropy of the disease is only affected with the possibility of disease existence or disease non-existence. Since these possibilities are equal in both tests, the entropy of the disease is the same. In Table 4.9, the entropy value is the same but mutual information values are different. According to the result of mutual information, I. Turbidimetric test provides more diagnostic information than II. Turbidimetric test. Therefore I. Turbidimetric test dominates II. Turbidimetric test. Based on these results, it can be verified that mutual information value is parallel to AUC value. While mutual information can be measured for all threshold values (See Table 4.6), AUC isn't measured for all threshold values. Therefore, mutual information value has an advantage to AUC value.

Table 4.9 Entropy, Conditional Entropy and Mutual Information of two tests

Tests	***H(D)***	***H(D\T)***	***I(D;T)***
I. Turbidimetric	0.96	0.50	0.46
II. Turbidimetric	0.96	0.63	0.33

4.8 I Positive, I Negative and Kullback Leibler Divergences of Two Tests

In the Table 4.10; I positive, I negative and Kullback Leibler divergences are compared for two tests.

Table 4.10 I Positive, I Negative and Kullback Leibler divergences of two tests

Tests	***I pos.***	***I neg.***	***D(d + //d-)***	***D(d - //d+)***
I. Turbidimetric	1.02	0.21	0.94	0.52
II. Turbidimetric	1.29	0.07	Non-Defined	Non-Defined

According to the results of Table 4.10, a positive test result provides more information than a negative test result for both tests. The positive test result of II. Turbidimetric test provides more information than the positive test result of I. Turbidimetric test. On the contrary, the negative result of I. Turbidimetric test provides more information than the negative result of II. Turbidimetric test.

Another result that can be deduced from Table 4.10 is the comparison of $D(d+\backslash\backslash d-)$ and $D(d-\backslash\backslash d+)$ values. $D(d+\backslash\backslash d-)$ value is higher than $D(d-\backslash\backslash d+)$ for I. Turbidimetric test. I. Turbidimetric test is the most specific diagnostic test to rule in disease (existence of the disease). Neither $D(d+\backslash\backslash d-)$ nor $D(d-\backslash\backslash d+)$ values can be measured for II. Turbidimetric test (See Page 26).

CHAPTER FIVE
CONCLUSION

In this study; ROC which is a long-standing method for the evaluation of the diagnostic test performance and Information Theory which has been used recently to evaluate the diagnostic test performance are presented in detail.

ASO values being the first phase for the diagnosis of rheumatic disorder are measured using Turbidimetric tests which belong to two different firms. The tests' performances are examined using ROC and Information Theory. This study aims to investigate which Turbidimetric test has better performance and this performing test is going to be conducted during the study in order to demonstrate whether this performing test can be an alternative to Nefelometric test which is currently the gold standard for the diagnosis of rheumatic disorder. With regard to ASO values, it is concluded that I. Turbidimetric test is more likely to show the similarity to Nefelometric test in comparison with II. Turbidimetric test. Using I. Turbidimetric test has financial benefits to clinicians, since it is less expensive in contrast with Nefelometric test.

As a result of Information Theory analysis, the threshold value of 173 is the optimal value which maximizes mutual information. Based on this optimal threshold value, it can be deduced that 0-200 UI/ml reference interval which is mentioned in the medicine literature for Nefelometric test can be replaced with a "*new*" 0-173 UI/ml reference interval. The use of this new reference interval provides more accuracy and leads to less error in the diagnosis of ASO values. As a conclusion of the study, it is recommended to the clinicians to implement I. Turbidimetric test with a new reference interval for the diagnosis of rheumatic disorder.

It is aimed that this study will hopefully give various points of view to the researchers who want to make research on this subject by explaining how the tests used for the diagnosis of various diseases are evaluated with this way.

REFERENCES

Bamber, D. (1975). The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph. *Journal of Mathematical Psychology, 12,* 387-415.

Benish, W. A. (2002). The use of information graphs to evaluate and compare diagnostic tests. *Methods Inf Med, 41*, 114-118.

Benish, W. A. (2009). Intuitive and axiomatic arguments for quantifying diagnostic test performance in units of information. *Methods Inf Med, 48*, 552-557.

Boyko, E. J. (1994). Ruling Out or Ruling In Disease with the Most Sensitive or Specific Diagnostic Test Short Cut or Wrong Turn? *Medical Decision Making, 14*, 175-179.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory 2nd Edition.* New Jersey: John Wiley & Sons, Inc.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861-87.

Hanley, J. A., & McNeil, B. J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (R OC) Curve. *Radiology, 143* (1), 29-36.

Hanley, J. A., & McNeil, B. J. (1983). A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases[1]. *Radiology, 148* (3), 839-843.

Kraemer, H. C. (1992). *Evaluating Medical Tests.* Sage Publications.

Lee, W. C. (1999). Selecting Diagnostic Tests for Ruling Out or Ruling In Disease: The Use of the Kullback-Leibler distance. *International Journal of Epidemiology, 28*, 521-525.

Lusted, L. B. (1971). Signal detectability and medical decision-making. *Science, 171*, 1217-1219.

Metz, C. E., Goodenough, D. J., & Rossmann, K. (1973). Evaluation of receiver operating characteristic curve data in terms of information theory, with applications in radiography. *Radiology*, *109*, 297-303.

Metz, C. E. (1998). *ROCKIT (Version 0.9.1b) [Computer software]*. Retrived July 30, 2002, from http://www-radiology.uchicago.edu/krl/toppage11.htm.

Mossman, D., & Somoza, E. (1989). Maximizing diagnostic information from the dexamethasone suppression test: An approach to criterion selection using receiver operating characteristic analysis. *Archives of General Psychiatry, 46*, 653-660.

The Magnificent ROC, (n.d.). Retrieved November 20, 2011, from http://www.anaesthetist.com/mnm/stats/roc/Findex.htm

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*, 561-577.

APPENDIX
R PROJECT CODES

Appendix 1 (Vectors)

Package: pROC

d= Diagnosis
a= I. Turbidimetric test results
b= II. Turbidimetric test results

```
d=c(0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,1,
1,0,1,1,0,1,1,0,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)

a=c(1,3,17,17,370,50,31,47,165,47,53,47,65,79,72,61,53,35,66,75,786,93,53,72,75,9
1,82,47,106,101,111,119,103,78,140,145,192,129,168,135,142,119,194,230,171,197,
315,166,234,259,108,185,213,173,275,281,325,292,410,388,447,611,572,884,669,98
8,1113,1437)

b=c(11,11,14,14,23,24,25,27,102,34,35,36,39,39,39,42,42,42,46,48,48,48,51,57,60,6
2,67,68,75,76,77,78,86,94,96,101,105,109,113,115,118,121,124,129,143,149,150,15
2,152,156,166,180,187,192,202,205,211,257,268,280,367,368,413,550,706,742,832,
832)
```

Appendix 2 (Statistical Comparison)

```
rocobj1 <- plot.roc(d,a, main="Statistical comparison", percent=TRUE,
col="#1c61b6")
rocobj2 <- lines.roc(d,b, percent=TRUE, col="#008600")
testobj <- roc.test(rocobj1, rocobj2)
text(50, 50, labels=paste(("p-value =",format.pval(testobj$p.value)), adj=c(0, .5))
legend("bottomright", legend=c("I.Türb.", "II.Türb."), col=c("#1c61b6", "#008600"),
lwd=2)
```

Appendix 3 (AUC)

Package: ROCR

```
pred.1<-prediction(a,d)
perf.1.3<-performance(pred.1,'auc')
perf.1.3
```

Appendix 4 (AUC)

Package: ROCR

```
pred.1<-prediction(b,d)
perf.1.3<-performance(pred.1,'auc')
perf.1.3
```

Appendix 5-6-7 (SE, SP and Chi-Square for First Turbidimetric Method)

Package: ROCR

SE

```
pred.1<-prediction(a,d)
perf.1.5<-performance(pred.1,'sens')
perf.1.5
```

SP

```
pred.1<-prediction(a,d)
perf.1.6<-performance(pred.1,'spec')
perf.1.6
```

Chi-Square

```
pred.1<-prediction(a,d)
perf.1.8<-performance(pred.1,'chisq')
perf.1.8
```

Appendix 8-9-10 (SE, SP and Chi-Square for Second Turbidimetric Method)

Package: ROCR

SE

```
pred.1<-prediction(b,d)
perf.1.5<-performance(pred.1,'sens')
perf.1.5
```

SP

```
pred.1<-prediction(b,d)
perf.1.6<-performance(pred.1,'spec')
perf.1.6
```

Chi-Square

```
pred.1<-prediction(b,d)
perf.1.8<-performance(pred.1,'chisq')
perf.1.8
```

Appendix 11- 12-13 (EFF, PVP and PVN for First Turbidimetric Method)

Package: ROCR

EFF

```
pred.1<-prediction(a,d)
perf.1.2<-performance(pred.1,'acc')
perf.1.2
```

PVP

```
pred.1<-prediction(a,d)
perf.1.9<-performance(pred.1,'ppv')
perf.1.9
```

PVN

```
pred.1<-prediction(a,d)
perf.1.10<-performance(pred.1,'npv')
perf.1.10
```

Appendix 14-15-16 (EFF, PVP and PVN for Second Turbidimetric Method)

Package: ROCR

EFF

```
pred.1<-prediction(b,d)
perf.1.2<-performance(pred.1,'acc')
perf.1.2
```

PVP

```
pred.1<-prediction(b,d)
perf.1.9<-performance(pred.1,'ppv')
perf.1.9
```

PVN

```
pred.1<-prediction(b,d)
perf.1.10<-performance(pred.1,'npv')
perf.1.10
```

Appendix 17 (The Best Threshold Value for First Turbidimetric Method)

Package: pROC

```
d=c(0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,1,
1,0,1,1,0,1,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)

a=c(1,3,17,17,370,50,31,47,165,47,53,47,65,79,72,61,53,35,66,75,786,93,53,72,75,9
1,82,47,106,101,111,119,103,78,140,145,192,129,168,135,142,119,194,230,171,197,
315,166,234,259,108,185,213,173,275,281,325,292,410,388,447,611,572,884,669,98
8,1113,1437)

plot.roc(d, a, main="the best threshold value for first turb.", percent=TRUE,
ci=TRUE, of="thresholds", thresholds="best", print.thres="best")
```

Appendix 18 (The Best Threshold Value for Second Turbidimetric Method)

Package: pROC

```
d=c(0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,1,
1,0,1,1,0,1,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)

b=c(11,11,14,14,23,24,25,27,102,34,35,36,39,39,39,42,42,42,46,48,48,48,51,57,60,6
2,67,68,75,76,77,78,86,94,96,101,105,109,113,115,118,121,124,129,143,149,150,15
2,152,156,166,180,187,192,202,205,211,257,268,280,367,368,413,550,706,742,832,
832)

plot.roc(d, b, main="the best threshold value for second turb.", percent=TRUE,
ci=TRUE, of="thresholds", thresholds="best", print.thres="best")
```

Appendix 19-20 (Mutual Information)

Package: ROCR

```
pred.1<-prediction(a,d)
perf.1.7<-performance(pred.1,'mi')
```

```
perf.1.7
pred.1<-prediction(b,d)
perf.1.7<-performance(pred.1,'mi')
perf.1.7
```

Appendix 21 (Binary Vectors)

d= Diagnosis
e= I. Turbidimetric test results (binary)
f= II. Turbidimetric test results (binary)

```
d=c(0,0,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,1,
1,0,1,1,0,1,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)

e=c(0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,0,0,1,0,1,1,0,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)

f=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
```