

COMPARING SEVERAL DIAGNOSTIC PROCEDURES USING THE INTRINSIC MEASURES OF ROC CURVE

Vishnu Vardhan R* and Balaswamy S

* Department of Statistics, Pondicherry University, Puducherry – 605 014, India

Department of Statistics, Pondicherry University, Puducherry – 605 014, India

Abstract

Comparison of diagnostic tests is essential in medicine. Test procedures for comparing two or more ROC curves are all based on measures d' , AUC and the maximum likelihood estimates of binormal ROC curves. However, intrinsic measures such as sensitivity and specificity also play a pivotal role in assessing the performance of several diagnostic procedures. In this paper, a new methodology is proposed in order to compare several diagnostic procedures using the intrinsic measures of ROC curve

Keywords:

Diagnostic Procedure;

ROC curve; AUC;

Sensitivity

Introduction

“Receiver Operating Characteristic (ROC) Curve is a method of describing the intrinsic accuracy of a test apart from the decision thresholds”. Apart from providing accuracy of a diagnostic test, it is essential to validate its significance. Hence, there is a need to propose the inferential procedures to compare two diagnostic tests. Here, an outline is presented about the existing literature on comparison of diagnostic tests. In order to compare two diagnostic tests, Gourevitch and Galanter (1967) proposed the statistical test for pairwise comparison of ROC curves based on single parameter d' . Later, Marascuilo (1970) extended that test to compare a single ROC curves against chance and to perform multiple comparisons of three or more ROC curves by including an appropriate post-hoc procedure. A decade later, Metz and Kronman (1980) suggested statistical significance tests which help in evaluating the apparent differences between an obtained and an expected Binormal ROC curve, between two independent ROC curves and among groups of independent Binormal ROC curves. Appropriate chi-square tests were proposed for each of these tests by considering five-category rating scale data. Hanley and Mc. Neil (1982) proposed a test statistic which is based on Area Under the Curve (AUC). Wieand et. al (1989) gave a family of non-parametric statistics for comparing diagnostic markers with paired and unpaired data. But in practice, when more than two tests are involved, the performance of tests cannot be assessed all at a time. When comparing two or more diagnostic tests, it is often difficult to determine which test is superior for which the intrinsic measures sensitivity and specificity can be used instead of AUC. So, the work focuses on the use of sensitivities which are obtained from various diagnostic tests. Using these sensitivities, one can compare several diagnostic tests and can determine which test is providing more number of true positive cases.

The ROC curve analysis is used to test the performance to identify the abnormalities of diagnostic procedures. If there are several diagnostic procedures involved, then the data structure can be presented in the following form

Diagnostic Procedures (DT)	Observations (Diseased/Healthy)				
	1	2	3	...	j ... n _i
DT ₁	x_{11}	x_{12}	x_{13}	...	x_{1j} ... x_{1n_1}
DT ₂	x_{21}	x_{22}	x_{23}	...	x_{2j} ... x_{2n_2}
.
.
.
DT _i	x_{i1}	x_{i2}	x_{i3}	...	x_{ij} ... x_{in_i}
.
.

$$\begin{matrix} \cdot & & \cdot & & \cdot & & \cdot & & \cdot \\ \hline DT_k & & x_{k1} & x_{k2} & x_{k3} & \dots & x_{kj} & \dots & x_{kn_k} \end{matrix}$$

where x_{ij} represents the j^{th} observation from i^{th} diagnostic procedure. Every diagnostic procedure will be embedded with a status variable. Using this status variable the classification of observations such as healthy and diseased populations is done. On applying ROC classification procedure, we obtain a pair of Sensitivity and 1-Specificity at each particular threshold. From the obtained pairs or coordinates, the sensitivities are taken into account for further analysis. The major consideration in taking sensitivities alone is that the performance of diagnostic procedures can be identified with the true positive cases.

The data design so obtained after extracting the sensitivities from the ‘k’ diagnostic is presented in table 2.

Table 2: Data design for Sensitivities observed from ‘k’ diagnostic procedures

Diagnostic Procedures	Sensitivities						Totals	Averages	
	1	2	3	...	j	...			n_i
1	Se_{11}	Se_{12}	Se_{13}	...	Se_{1j}	...	Se_{1n_1}	$Se_{1.}$	\overline{Se}_1
2	Se_{21}	Se_{22}	Se_{23}	...	Se_{2j}	...	Se_{2n_2}	$Se_{2.}$	\overline{Se}_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
I	Se_{i1}	Se_{i2}	Se_{i3}	...	Se_{ij}	...	Se_{in_i}	$Se_{i.}$	\overline{Se}_i
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
K	Se_{k1}	Se_{k2}	Se_{k3}	...	Se_{kj}	...	Se_{kn_k}	$Se_{k.}$	\overline{Se}_k
								$Se_{..}$	$\overline{Se}_{..}$

In the above table, ‘ Se_{ij} ’ represents the j^{th} Sensitivity of the i^{th} diagnostic procedure. The notation represented in the table 2 are defined as,

$$\overline{Se}_i = \frac{Se_i}{n_i} \quad ; \quad i=1, 2, \dots, k$$

where \overline{Se}_i is the average of the sensitivities under the i^{th} diagnostic procedure.

$$Se_{..} = \sum_{i=1}^k \sum_{j=1}^{n_i} Se_{ij}$$

where $Se_{..}$ is the grand total of the sensitivities.

$$\overline{Se}_{..} = \frac{Se_{..}}{N}$$

where $\overline{Se}_{..}$ is the grand average of sensitivities and $N = \sum_{i=1}^k n_i$ is the total number of sensitivities.

To assess these diagnostic procedures, the data design and model specifications are defined and brought into the framework of linear regression and ANOVA. Using which we can meet the requirement of comparing several diagnostic procedures. Once, the ANOVA procedure is carried out, it is feasible to identify the better diagnostic procedure which allows in providing more number of sensible cases. Further, multiple comparison tests are used to carry out the pair wise comparisons between several diagnostic procedures.

Materials and methods

The linear model for the specific problem is defined as,

$$Se_{ij} = \mu + \tau_i + \epsilon_{ij} \quad ; \quad i=1, 2, \dots, k \quad \& \quad j=1, 2, \dots, n_i \tag{1}$$

ϵ_{ij} is the random error i.e. $\epsilon_{ij} \sim N(0, \sigma^2)$.

where ‘ μ ’ is a parameter common to all diagnostic procedures usually referred to as the overall mean effect.

‘ τ_i ’ is the parameter unique to the i^{th} diagnostic procedure.

The above equation (1) is known as the “effects model”. The response variable Se_{ij} is a linear function of the model parameters. For hypothesis testing, the model errors are assumed to be normally and independently distributed random variables with mean zero and variance σ^2 . The variance is assumed to be constant for all levels of the factor.

$$Se_{ij} \sim N(\mu + \tau_i, \sigma^2)$$

and that the sensitivities are mutually independent. The proposed work is to verify which one of the DT's provides better accuracy with higher correct classification of diseased subjects.

The hypothesis based on average sensitivities can be defined as follows,

$$H_0: \overline{Se}_{1.} = \overline{Se}_{2.} = \dots = \overline{Se}_{i.} = \dots = \overline{Se}_{k.}$$

vs

$$H_1: \overline{Se}_{i.} \neq \overline{Se}_{j.} \text{ for atleast one pair } (i, j) \text{ where } i \neq j.$$

The total variability is partitioned into two components, namely variation between the diagnostic procedures and within the diagnostic procedures. The total corrected sum of squares,

$$SS_T = \sum_{i=1}^k \sum_{j=1}^{n_i} (Se_{ij} - \overline{Se}_{..})^2$$

is used as a measure of overall variability in the data with (N-1) degrees of freedom.

Note that the total corrected sum of squares SST may be written as,

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^{n_i} (Se_{ij} - \overline{Se}_{..})^2 &= \sum_{i=1}^k \sum_{j=1}^{n_i} [(\overline{Se}_{i.} - \overline{Se}_{..}) + (Se_{ij} - \overline{Se}_{i.})]^2 \\ \sum_{i=1}^k \sum_{j=1}^{n_i} (Se_{ij} - \overline{Se}_{..})^2 &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\overline{Se}_{i.} - \overline{Se}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (Se_{ij} - \overline{Se}_{i.})^2 \end{aligned}$$

i.e. $SS_T = SS_{DT} + SS_E$

where SS_T is called the sum of squares due to diagnostic test procedures with (k-1) degrees of freedom and SS_E is called the sum of squares due to error with (N-k) degrees of freedom.

To test the defined hypothesis, the test statistic is given by,

$$F = \frac{SS_{DT}/k-1}{SS_E/N-k} = \frac{MS_{DT}}{MS_E}$$

is distributed as F with (k-1) and (N-k) degrees of freedom.

Reject H_0 and conclude that mean differences exists between the diagnostic procedures if, $F > F_{\alpha, (k-1), (N-k)}$

The entire ANOVA procedure is summarized in the following table 3.

Table 3: The ANOVA table based o Sensitivites

Source of Variation	Sum of Squares	d.f.	Mean Squares	F- ratio
Between Diagnostic Procedures	$SS_{DT} = n_i \sum_{i=1}^k (\overline{Se}_{i.} - \overline{Se}_{..})^2$	k-1	$MS_{DT} = \frac{SS_{DT}}{k-1}$	$F = \frac{MS_{DT}}{MS_E}$
Error	$SS_E = SS_T - SS_{DT}$	N-k	$MS_E = \frac{SS_E}{N-k}$	
Total	$SS_T = \sum_{i=1}^k \sum_{j=1}^{n_i} (Se_{ij} - \overline{Se}_{..})^2$	N-1		

Post Hoc Tests

Post Hoc tests are used to determine which mean or group of means are significantly different from the others. These tests are done only when ANOVA yields a significant F ratio.

Tukey's Test in terms of Sensitivity measure

Suppose that following an ANOVA in which the null hypothesis gets rejected means, it is necessary to test all pair wise comparisons of mean sensitivities.

$$H_0: \bar{Se}_i = \bar{Se}_j \quad \text{Vs} \quad H_0: \bar{Se}_i \neq \bar{Se}_j \quad \forall i \neq j$$

Tukey (1949b) proposed a procedure for testing hypotheses for which the overall significance level is exactly ' α ' when the sample sizes are equal and at most ' α ' when the sample sizes are unequal. The Tukey's procedure controls the experiment wise or family error rate at the selected level ' α '. Tukey's procedure makes use of the distribution of the studentized range statistic,

$$q = \frac{\bar{Se}_{\max} - \bar{Se}_{\min}}{\sqrt{\frac{MS_E}{n}}}$$

where \bar{Se}_{\max} and \bar{Se}_{\min} are the largest and smallest sample means respectively.

For equal sample sizes, Tukey's test declares two means significantly different if the absolute value of their sample differences exceeds,

$$T_\alpha = q_\alpha(k, f) \sqrt{\frac{MS_E}{n}}$$

where $q_\alpha(k, f)$ the upper $\alpha\%$ points of 'q' from the table and 'f' is the number of degrees of freedom associated with the MSE.

When sample sizes are not equal, then

$$T_\alpha = \frac{q_\alpha(k, f)}{\sqrt{2}} \sqrt{MS_E \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

The unequal sample size version is sometimes called the *Tukey-Kramer Procedure*.

Results and discussion

To illustrate the above proposed methodology, a data set of Parathyroid Disease is used (Zhou. et. al. (2011)). Parathyroid glands are small endocrine glands usually located in the neck or upper chest that produce a hormone which controls the body's calcium levels. Most people have four parathyroid glands. There are three types of nuclear medicine imaging test used to detect and localize parathyroid lesions prior to surgical investigation is called SPECT (Single photon emission computed tomography). Three types of investigation namely SPECT with no attenuation (No SPECT), SPECT with attenuation and SPECT / CT along with a status variable for surgical result like positive or negative is used to identify the number of lesions. Total of 97 samples were considered in the study.

For all the three diagnostic procedures, ROC analysis has been carried out and the intrinsic measures are reported. From these results, sensitivity values are taken for further analysis, since the main objective of the paper is to compare and assess the performance of three diagnostic procedures with respect to their average sensitivities. The results so obtained from the proposed methodology are presented in the following tables and a proper discussion is reported.

Source of Variation	Sum of Squares	df	Mean Square	F	Sig.
Between Diagnostic Tests	1.202	2	0.601	10.002	0.000*
Error	17.300	288	0.060		
Total	18.502	290			

From table 4, the value of F ratio is 10.002 and is significant ($p < 0.05$), which means that all three diagnostic test means differ significantly ($p < 0.05$) in identifying the lesions. It is reasonable to conclude that at least one of the diagnostic tests mean sensitivities is significantly different from the others. Beyond this conclusion; there is a need to conduct a post hoc test to determine which diagnostic test differs from the other two diagnostic tests.

Table 5: Multiple Comparisons for Sensitivity (Tukey)

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
NO SPECT	SPECT	-0.01281296	0.03519291	0.930	-0.0957238	0.0700979
	SPECT CT	-0.14226804*	0.03519291	0.000	-0.2251789	-0.0593572
SPECT	NO SPECT	0.01281296	0.03519291	0.930	-0.0700979	0.0957238
	SPECT CT	-0.12945508*	0.03519291	0.001	-0.2123659	-0.0465443
SPECT CT	NO SPECT	0.14226804*	0.03519291	0.000	0.0593572	0.2251789
	SPECT	0.12945508*	0.03519291	0.001	0.0465443	0.2123659

*. The mean difference is significant at the 0.05 level.

The multiple comparisons table 5 is showing the results for the Tukey's test. The assumption of homogeneity of variance is met. The mean difference for the two diagnostic tests namely NO SPECT and SPECT is found to be -0.0128 and this pair is found to be insignificant ($p > 0.05$) whereas the other two pairs of combinations NO SPECT, SPECT/CT and SPECT, SPECT/CT are having the mean differences -0.1422 and -0.1294 respectively with significant pairs of diagnostic tests. From this discussion, it is reasonable to conclude that the diagnostic test SPECT/CT is significantly different from the other two diagnostic test procedures with the high average Sensitivity. i.e. the diagnostic test SPECT/CT is more appropriate in finding the lesions of gland of the Parathyroid disease than the other diagnostic tests.

Table 6: Descriptive Statistics for Sensitivity

Diagnostic Test	Mean \pm Std. Error	Tukey's Sig.
NO SPECT	0.632 ^a \pm 0.028	0.930 ^{NS}
SPECT	0.645 ^a \pm 0.026	
SPECT / CT	0.775 ^b \pm 0.020	1.000 ^{NS}

NS = Not Significant

The descriptive statistics such as Mean, Standard Error for the three diagnostic tests NO SPECT, SPECT and SPECT/CT are reported in table 6. The SPECT/CT diagnostic test has the highest mean 0.775 than the other two diagnostic tests NO SPECT and SPECT i.e., 0.632 and 0.645 respectively. The homogeneous subsets can be given by the Tukey's sign value, which represents in the descriptive table 6 as a superscript for the mean sensitivities and which provides the information of post hoc tests. The diagnostic tests which are having the same superscripts do not differ significantly ($p > 0.05$). So, the diagnostic tests NO SPECT and SPECT are not significantly different. But the diagnostic test SPECT/CT is different from the other two diagnostic tests with different homogeneous subset shown with the superscript 'b'. The same information can be explained graphically in figure 1 using line whisker's plot.

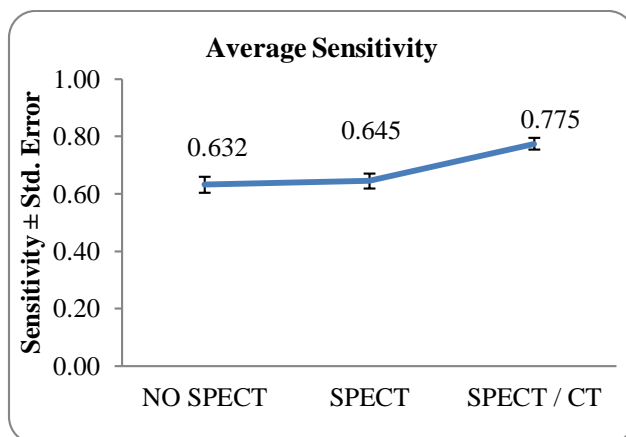


Figure 1: Plot of average Sensitivities of three Diagnostic Tests

Using the proposed methodology, it can be concluded that the SPECT/CT diagnostic test is most important in finding the number of glands of the Parathyroid disease among all the three diagnostic tests. To support our proposed methodology the accuracy measures of ROC curve are also reported in table 7.

Diagnostic Test	Threshold for Lesion	Sensitivity	Specificity	ROC Area
NO SPECT	5	0.6142	0.6666	0.6404
SPECT	5	0.7142	0.5925	0.6534
SPECT / CT	4	0.7	0.9629	0.8314

From table 7, it is observed that the accuracy measure Area under the curve (AUC) is 0.8314 with the cutoff 4 with SPECT/CT. i.e. an individual's lesions score exceeding 4 can easily be identified as the one with parathyroid disease with 83% accuracy. Whereas the AUC is found to be 0.6404 and 0.6534 for NO SPECT and SPECT diagnostic tests respectively with the cutoff value 5 (test score greater than or equal to 5 are positive cases of disease). Therefore, the two diagnostic tests NO SPECT and SPECT are almost identical in identifying the number of glands of Parathyroid disease. From this it can be concluded that the diagnostic test SPECT/CT is superior in identifying the positive cases of disease than the other two diagnostic tests. Further, the ROC Curves (Figure 2) are also drawn in supporting the interpretation of results among the three diagnostic tests.

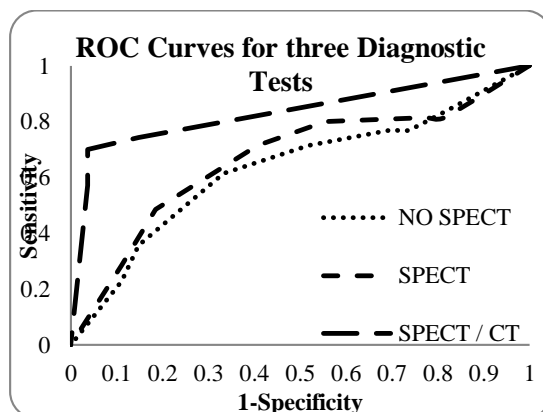


Figure 2: ROC Curves for three Diagnostic Tests

Conclusion

When three or more diagnostic tests are involved, there is a need to identify the better diagnostic test among them. The accuracy measures like AUC and pAUC are single numeric values that explain about the extent of correct classification in diagnostic tests. Using this single numerical value of accuracy, one cannot conclude which diagnostic test is better. Sometimes, AUC will not provide a clear information in identifying the better one and hence the focus is made on the use and importance of sensitivity measure in identifying the better procedure. To find out the better diagnostic test, ANOVA procedure based on Sensitivities of ROC curve is proposed. In the data considered for the procedure, it is identified that the diagnostic test SPECT/CT is found to be more appropriate in identifying the number of lesions of Parathyroid glands among other two diagnostic tests with 83% of accuracy and threshold of 4. Even though several researchers have proposed test procedures for comparing two or more ROC curves, are all based on measures d' , AUC and the maximum likelihood estimates of Binormal ROC curves. However, intrinsic measures such as sensitivity and specificity also play a pivotal role in assessing the performance of several diagnostic procedures. It is shown that using sensitivities also, one can compare several diagnostic procedures and post-hoc comparisons is also suggested using Tukey's test.



Acknowledgements

The first author would like to acknowledge University Grants Commission (UGC) for funding [F.No. 42 – 1002 / 2013 (SR) dated 22 March 2013 and 18 April 2013].

References

1. John W. Tukey, Comparing Individual Means in the Analysis of Variance, Biometrics, vol. 5, pp. 99 – 114, 1949b
2. Xiao-Hua Zhou, Nancy A. Obuchowski and Donna K. McClish, Statistical Methods in Diagnostic Medicine – Second Edition . Wiley Series in Probability and Statistics, ISBN: 978 – 0470 – 18314 – 4, 2011
3. James A Hanley, Barbara J Mc Neil, A Meaning and Use of the area under a Receiver Operating Characteristics (ROC) Curves, Radiology; 143; 29 – 36, 1982
4. Metz and Kronman, Statistical Significance Tests for Binormal ROC curves , Journal of Mathematical Psychology, 22, 218-243, 1980
5. Gourevitch and Galanter, A significance test for one parameter isosensitivity functions, Psychometrika, 32, 25-33, 1967
6. Leonard A. Marascuilo, Extensions of significance test for one-parameter signal detection hypotheses, Psychometrika, 35, 237-243, 1970
7. Sam Wieand, Mitchell H. Gail, Barry R. James and Kang L. James, A Family of Nonparametric Statistics for Comparing Diagnostic Markers with Paired and Unpaired data, Biometrika, 76 (3), 585-592, 1989

Author Bibliography

	<p>R Vishnu Vardhan is currently working as Assistant Professor in the Department of Statistics, Pondicherry University (A Central), Puducherry. His areas of research are Biostatistics, Statistical Process Control and Statistical Computing. He has published 44 research papers in reputed National and International journals. He has organized three national workshops, one national conference and one international conference. He is a recipient of Ms Bhargavi Rao and Padma Vibhushan Prof. C R Rao Award for best Poster Presentation in an International Conference in the year 2010, Indian Society for Probability and Statistics (ISPS) Young Statistician Award during December 2011 and Young Scientist Award from Indian Science Congress in the year 2014. He is a life member of several professional bodies. He serves as referee for two reputed journals and is an editorial member in an international journal.</p>
	<p>Balaswamy S Pursued Ph.D. in Biostatistics in the Department of Statistics, Pondicherry University (A Central), Puducherry. His areas of research are Biostatistics and Statistical Computing. He has published 12 research papers in reputed National and International journals. He is a recipient of Indian Society for Probability and Statistics (ISPS) Young Statistician Award during December 2014</p>