DOI: 10.7860/JCDR/2018/32856.11609

Review Article

Statistics Section

Receiver Operator Characteristic Analysis of Biomarkers Evaluation in Diagnostic Research

KARIMOLLAH HAJIAN-TILAKI

ABSTRACT

Receiver Operator Characteristic (ROC) analysis is the choice of method in evaluation of biomarkers in bioinformatics research. However, there is no single method and also no single accuracy index in evaluating diagnostic tools. This review provides an extensive illustration of different methods of ROC curve analysis that can be used in clinical practice of diagnostic studies. It includes their early use for rating data and the recent developments for quantitative data with a discussion of choice of model selection in parametric ROC analysis compared with non-parametric approach. The relevant methodological issues of these two alternative approaches have been discussed in terms of bias and sampling variability of Area under the curve (AUC) index that may influence on the performance of diagnostic tests. The methods were illustrated with two relevant clinical examples. The semi-parametric and parametric model of mixture of Gaussian is comparable with purely nonparametric approach. The choice between methods depends on practical conveniences unless the presence of severe departure from binormality. The recent new development and the gaps in knowledge concerning their behaviours in actual applications for medical researches and a guideline for future research have been discussed.

Keywords: Area under the curve, Binormal model, Non-parametric, Parametric, ROC analysis, Semi-parametric

INTRODUCTION

Bioinformatics research deals with acquisition of new information to use in health care and patients' management [1]. Many diagnostic systems have been developed progressively during the two recent decades. In particular, the evaluation of these diagnostic imaging systems or biomarkers is primarily interested in bioinformatics research. ROC analyses have been used widely for evaluation of classifiers such as biomarkers/diagnostic technology [2,3]. There are several examples of the use of ROC analysis in bioinformatics medicine [4,5], in particular, when the diagnostic test results are recorded on continuous scales. For example a large number of computer based diagnostic program have been developed to advise a physician on patient diagnosis and to evaluate the extent to which these information systems can improve the patients' diagnosis and their management. Although ROC analysis is the choice of method in evaluation of bioinformatics systems, however there is no single method and alsono single accuracy index in evaluating diagnostic tools [3].

In diagnostic tests evaluation, when the test results are recorded in ordinal or continuous scales, the sensitivity (or True Positive Fraction-TPF) and the specificity (or True Negative Fraction-TNF) can be calculated across all possible cut-off values [2,3]. Then, the plot of sensitivity versus 1-specificity (or False Positive Fraction-FPF) is called ROC curve and the Area Under the Curve (AUC) has meaningful interpretation. It is also considered as a measure of overall accuracy index in diagnostic medicine [6-8]. The AUC is interpreted as the probability that a randomly chosen diseased subject is rated as more likely to be diseased than a randomly chosen non-diseased subject. This interpretation is based on nonparametric Mann-Whitney U-statistic that is used in calculating AUC [6]. The partial area at clinical relevant of false positive, as an index of accuracy, has been suggested [9]. This is particularly interesting when two ROC curves cross each other and produce a similar AUC. The slope of an ROC curve at any point is equal to the ratio of two density functions, describing the distribution of separator variables (or test results) in the "diseased" and "non-diseased" population called likelihood ratio[10]. A monotonically increasing likelihood ratio corresponds to a concave ROC curve.

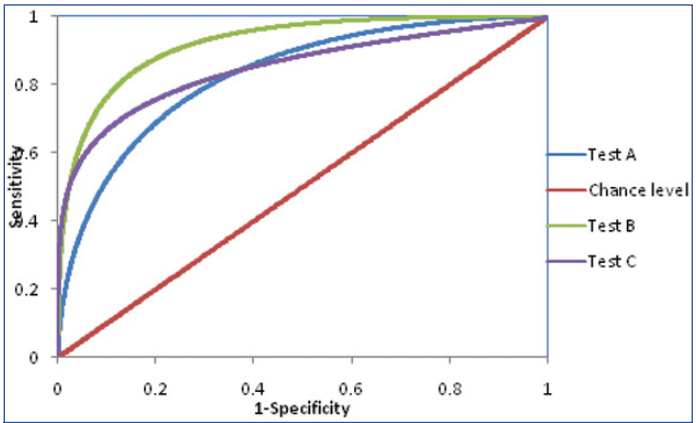
An advantage of ROC analysis is to determine the optimal cut-off values using equal or unequal weight for FPF and FNF [11]. The two conventional methods called Euclidian index and Youden index use an equal weight. The Euclidian index minimizes the square of distance between the point at (0,1) on the left hand corner of ROC space at any point on ROC curve. The Youden index maximizes the difference between sensitivity and FPF or 1-specificity that is defined as sensitivity+specificity-1. Thus, by maximization of sum of sensitivity and specificity across various cut-off values, the optimal cut-off is estimated. The third method incorporates the prevalence of diseases and the financial cost of correct and false diagnosis with decision utility based method. The fourth, so called product method uses the maximum product of sensitivity and specificity as a criterion in determining optimal cut- off point [12]. The three popular methods of Youden index, Euclidean index and product method yield the identical optimal cut-point in particular with symmetric condition of ROC curve [13] but the diagnostic odds ratio as other criterion produces an extreme cut-off value that is very far from the optimal relevant cut-off point for diagnostic biomarkers [13,14].

Several methods of parametric, semi-parametric non-parametric have been proposed for constructing of ROC curve and estimating AUC and testing of two ROC curves [15,16]. Although, these methods have been widely used in published diagnostic studies by clinician [17,18], the advantage and disadvantage of each approach were not well understood by clinical practitioner. Thus, this article provides a conceptual framework of full review of parametric, semi-parametric, and nonparametric, and re-sampling methods in ROC analysis for clinical practice.

Earliest Parametric Approaches for Rating Data

In the first applications of ROC methodology, curves were constructed from rating data by making certain distributional assumptions. The parametric assumptions specify the forms of

overlapping distributions on "separator" scale for the "diseased" and "nondiseased" populations, in order to estimate a smooth ROC curve. The interpretation was typically recorded on a five-point ordinal scale that indicated the degree of diagnostic certainty. Parametric ROC analysis is used when underlying distributions of test results for diseased and nondiseased are known. For example, binormal model, commonly used in ROC analysis assume the test results either Gaussian or after some monotonic transformation such as log, square or Box-Cox makes the data become Gaussian distribution for both diseased and non-diseased [3]. More than 40 years ago, by postulating that the ratings arose from the discretization of an underlying "latent" scale, and by making certain distributional assumptions, Dorfman and Alf (1969) and others were able to objectively fit smooth ROC curves to rating data using binormal assumption of two overlapping Gaussian distributions with different parameters [15]. [Table/Fig-1] shows the corresponding ROC curve was derived from the results of test A and test B from binormal model in “nondiseased” and “diseased” has equal variance, the corresponding ROC curve is symmetric but in test C more variation was ascertained for “diseased” than “nondiseased” and thus its



[Table/Fig-1]: ROC curves derived from different combinations of overlapping Gaussian distributions (Test A: ND: N (0, 1.5) and D: N (2, 1.5); Test B: ND: N (0, 1.5) and D: N (3, 2.5); Test C: ND: N (0, 1.5) and D: N (3, 1.5).

derived ROC curve is asymmetric. In addition, the ROC curves of tests A and C cross to each other.

Hanley JA found that the distributional assumptions in these analyses of rating data were not critical [19]. In particular, the binormal model gave good fits even when rating data were generated by discretizing data generated from other models. Several other parametric forms such as bi-logistic and bi-exponential have also been suggested [20].

Fitting Binormal Model for Rating Data

In spite of many possible models for the overlapping distributions, the binormal model is by far the most commonly implemented in several packages (e.g., Metz's software) to fit ROC curves to rating data; it is also now used for laboratory-type data [21]. This functional assumption has been justified on theoretical and practical grounds for rating data [19]. The binormal assumption allows ROC curves to be transformed into straight lines on normal-deviate scales i.e., $\Phi^{-1}(TPF) = a+b\Phi^{-1}(FPF)$ where Φ is the cumulative density function of the Gaussian distribution [2]. The ROC curve is described by two parameters - a: the difference between the mean of the two distributions (standardized by the standard deviation of the diseased population), and b: the ratio of the two standard deviations [2,3]. Also, the distributions are described using the two parameters of interest rather than four (means and standard deviations of the two distributions).

The method of maximum likelihood, based on two sets of multinominal data is used to estimate binormal parameters “a” and “b” along with (c-1) "cut-off" parameters where c is the number of rating categories. From “a” and “b”, one can calculate the area under the curve and other indices along with their corresponding precision. The area under the ROC curve is the proportion of the Gaussian distribution to the left of Z_A where $Z_A=a/(1+b^2)^{1/2}$; i.e., $AUC=\Phi(a/(1+b^2)^{1/2})$ [2]. This methods also allows to estimate the variance of AUC under binomial assumption and to calculate its 95% confidence interval and to test the hypothesis AUC=0.5 versus AUC >0.5. Using Delta method with an approximation when the ratio of SD is close to one (i.e., b=1) the binormal estimator of AUC variance of $\widehat{AUC}$ is

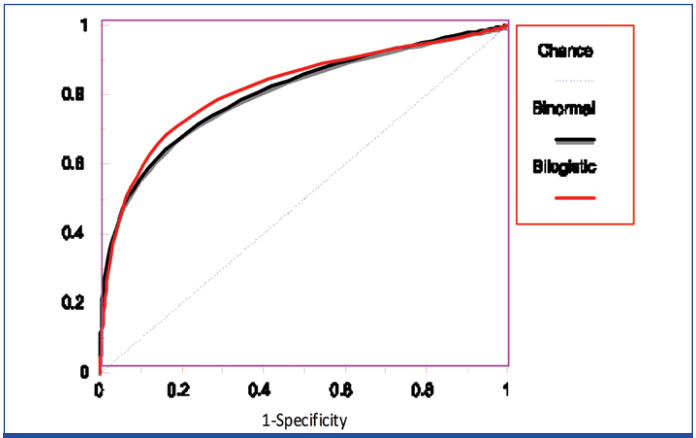
$$Var[\widehat{AUC}] = (0.0099 \times e^{-e^2}) \times (\frac{5a^2+8}{n_2} + \frac{a^2+8}{n_1})$$

where $a=\varphi^{-1}(AUC) \times 1.414$ and n_1 and n_2 are the sample size for nondiseased and diseased [21].

As of now, the fitting of the binormal model for rating data can be carried out using Dorfman and Alf's program (1968) and Metz's software (ROCFIT) [22]. One can obtain estimates of the ROC parameters, their variance covariance matrix, the relevant indices of accuracy and the chi-square test of goodness of fit. In addition, Metz CE et al., extended the binormal model to allow for testing for differences in the diagnostic accuracy of two systems on the basis of correlated data [23,24]. This bivariate binormal model was implemented in the CORROC software [22]. Although the estimates for each ROC curve are close to those obtained by analysing each set of data separately, the main advantage of CORROC procedure is that it measures the covariance of the estimates of accuracy for each test. This is particularly important if the covariance is strongly positive since it means that the standard error of the difference of the two estimates of accuracy indices will be decreased.

Bi-logistic Model

The only other serious competitor of binormal model is the bi-logistic model [20]. Bi-logistic model and bi-exponential can be fitted using PLUM software [25] but the software (PLUM) is not currently widely available. One may choose the bilogistic model



[Table/Fig-2]: ROC curves from bi-normal model and bi-logistic model with the same parameters.

rather than the binormal model. In fact, these two models with equal mean and standard deviation are close in shape. Compared with the normal density function, the logistic density function is slightly taller, narrower through the midsection and wider at the tails. The bilogistic model and binormal model with the same parameters of mean and standard deviation yield similar ROC curve and AUC {see [Table/Fig-2]- ND:L(0, 1), D:L(1.5, 1.4); ND: N(0, 1) and D:N(1.5, 1.4), ND:N(0, 1): $AUC_{Bilogistic}$=0.821, $AUC_{Binormal}$=0.808}.

Choice of Model in Fitting ROC Curves

One may ask whether a particular distributional form on the decision variable is preferred, especially when the data are collected in the rating scale, and the underlying decision variable is unobservable. Although any two underlying distributions determine an ROC

curve, the reverse does not hold: an ROC curve does not uniquely determine the underlying unobservable distributions of decision variables [3,23]. Any monotonic transformation of the decision axis generally yields different pairs of distributions, but does not change the ROC curve [3,23]. Because the decision variable is unobservable in many applications of ROC analysis, the underlying form of the latent distributions cannot be uniquely determined. Thus, one only can specify the underlying distribution up to a particular monotonic transformation of the decision axis.

However, among the parametric forms of overlapping pairs of distributions, the binormal model is more popular than any other because of practical convenience. Indeed, the binormal assumption concerns only the functional form of the ROC curve and not the form of the underlying distributions themselves. If one makes a monotonic transformation of decision axis, the underlying distributions are no longer binormal, but they yield the same ROC curve.

Parametric and Semi-parametric ROC Method for Quantitative Data:

ROC techniques are no longer confined to subjective interpretations (rating data) from a "latent" decision scale. Most diagnostic tests in medicine yield quantitative data on direct or "observed" scales. Consequently, the ROC method has become popular for evaluating the performance of such tests [16,26,27]. In fitting binormal model to continuous data, two approaches have been proposed. The first directly fits the binormal model to raw continuous data using maximum likelihood methods to estimate ROC parameters and their SE. The validity of this approach depends that the results of tests for "diseased" and "nondiseased" follow Gaussian distribution or after some monotonic transformation such as log, square, Box-Cox makes the data binormal. In the latter case, the birormal model fits on the transformed data for diseased and nondiseased. Goddard MJ and Hinberg I (1990) warned that if the distribution of raw data from a biochemical test is far from Gaussian, indices of accuracy (and their corresponding standard errors) based on the directly fitted binormal model can be seriously distorted [28]. In cases of real or apparent non-binormality, one might be tempted to consider transformation of test results that the data to be nearly Gaussian or uses other pairs of distributions. Alternatively one may use an entirely non-parametric method [16].

The second parametric approach categorizes the data and uses an existing parametric method for rating data [29]. Since this approach does not fit directly binormal model to continuously distributed data, the discritization algorithm uses rank of data instead of raw data using LABROC algorithm in fitting binormal model, we call it as semi-parametric approach. The performance of binormal model implemented in LABROC (for now it is renamed as ROCKIT) has been evaluated by several investigators [29-32] and the summary of its performance was summarized in [Table/Fig-3]. Overall, this semi-parametric approach for estimating of AUC- as a global index- is robust from departure of binormality except for bimodal form.

Another method of semi-parametric ROC analysis developed by Pepe using Generalized Linear Model (GLM) [33,34]. This approach is similar to nonparametric that does not use to make any distributional assumptions about diagnostic test results but similar to parametric method it estimates the ROC parameters "a" and "b" and corresponding AUC. Colak E et al., showed this semi-parametric approach is reliable when distributions of test results are skewed and it provides a smooth ROC curve as well [35].

Extension of Parametric ROC Analysis with Mixture of Gaussian

Since the semi-parametric method may not yieldreliable and valid estimates of AUC and its SE with serious departure from binormality in particular with bimodal data [30], Hajian-Tilaki KO et al., proposed an extension of parametric method when distributions of biomarkers are Mixture Of Gaussian (MG) [36]. They algebraically showed that AUC and TPF are calculated by weighting parameters of each component of AUC's and TPF's of mixture elements. For a real problem, the parameters of mixture distribution will not be known and thus it is required to be estimated. By simulation investigation, they estimated ROC parameters with EM-algorithm for various configurations of bimodal data. Their results of simulation showed that the estimates of overall AUC are essentially unbiased and the bias of TPF was small and it was equally comparable with nonparametric counterparts [36]. However, the estimations of the parameters of mixture distributions are not simple and in some situations the MLE methods may not have solution because the likelihood function becomes singular toward infinity in particular for MG distributions with small variation [37]. With a mixture of Gaussian, in a case of strong bimodality, the simulated data was not encounter to singularity of likelihood function and the parameters of each of two components of mixture were estimated by MLE method using EM-algorithm [36]. Using penalty likelihood or Bayesian method avoids likelihood function becomes singular in maximization procedure of EM algorithm [38,39]. In addition, Sorribas A et al., showedthatthe distribution family known as the S-distribution is insensitive and flexible for parametric ROC analysis of non-binormal data [40].

Authors	Methods	Performance
Goddard MJ and Hinberg I [28]	Parametric: Directly fitting of binormal model to raw data	The estimated ROC parameters would be distorted seriously unless one uses some monotonic transformation such as log, Box-Cox.
Metz CE et al., [29]	Semi-parametric: Fitting binormal via discrtitization of LABROC of raw data using LABROC1 and LABROC4 algorithm=	They found that binormal model was fit satisfactory with binormal data.
Hajian-Tilaki KO et al., [30]	Semi-parametric: Fitting binormal via LABROC approach with Gaussian and non-Gaussian data	They showed that this semi-parametric approach is robust for estimating AUC from departure of binormality except for bimodal distribution for diseased. However, the estimate of TPF$^{\Phi}$ is more sensitive.
Faraggi D and Reiser B [32]	LABROC and Kernel methods	Transformation to normality is preferred except for bimodal distribution, kernel method would be more effective.
Pepe MS [33]	Semi-parametric: Using General linear model	The approach is rather similar to nonparametric but produced a reliable estimates of two ROC parameters of 'a' and 'b'.
Cai T and Moskowitz CS [34]	Semi-parametric of binormal ROC curve	Without making distributional assumption of test results yielded a valid estimate of ROC parameters using general linear model.
Colak E et al., [35]	Semi-parametric and nonparametric	Semi-parametric approach is reliable when distribution of test results is skewed.
Hajian-Tilaki KO et al., [36]	Extension of parametric ROC analysis with mixture of Gaussian	By simulated investigation, they showed that the estimated ROC parameter of AUC$^{£}$ using EM-algorithm was unbiased for various configurations of bimodal data and the bias for TPF$^{\Phi}$ was negligible.
Sorribas A et al., [40]	ROC analysis with family of S-distribution	The family of S-distribution is insensitive and flexible of parametric method in ROC analysis of non-normal data.

[Table/Fig-3]: The related studies in evaluation of performance of parametric and semi-parametric ROC methods of quantitative data.
=It was renamed for new as ROCKIT program, £Area under the curve, ΦTrue positive fraction

Non-parametric Approaches to ROC Analysis:

Alternatively, a purely nonparametric ROC analysis can be used for both rating and continuously distributed data [6,41,42]. This approach has obvious appeal when test results are recorded on the continuous scale. This does not require any distributional assumption

of test results in diseased and nondiseased to estimate AUC and other relevant index of accuracy and their standard errors. However, the trapezoidal role (i.e., nonparametric method) calculates the AUC by just joining ROC operating points (1-specicifity, sensitivity) will tend to underestimate AUC for rating data in particular when ROC operating points are not well spread out along the ROC curve [6] and the method itself does not yield smooth estimates of the entire ROC curve, unless one uses spline techniques. The area under the curve comprises several trapezoidals at each interval of observed operating points and the area of each segment can be easily calculated and summed as AUC. The area under the empirical ROC curve has been shown to be equal to the Mann-Whitney U-statistic (or Wilcoxon statistic) [6] which are widely used as a test of location for comparing distributions from two samples. This approached has become more attractive and less cumbersome and has the appeal of being simple to calculate AUC. The corresponding accuracy

Authors	Methods	Performance
Bamber D [41]	Nonparametric ROC curve	He has shown that the AUC$^{\Phi}$ to be equal the Wilcoxon statistic and the nonparametric variance of AUC of a single diagnostic test was formulated.
Hanley JA and McNeil B [6]	Nonparametric estimates of AUC	He explained a clear interpretation of AUC$^{\Phi}$ and its link with Wilcoxon statistic.
DeLong ER et al., [42]	Nonparametric of variance-covariance matrix for correlated AUCs.	Without making any distributional assumption of test results, They formulated a variance-covariance matrix for two or more correlated AUCs$^{\Phi}$.
Hanley JA and Hajian-Tilaki KO [43]	Nonparametric estimates of AUC and its sampling variability	They restated Delong's method and introduced the interpretation of Delong pseudo accuracies and its link with jackknife pseudo values and thus the components of sampling variability of AUC$^{\Phi}$.
Hajian-Tilaki KO and Hanley JA [46]	Nonparametric SE of AUC with numerical investigation	They have shown that Delong's methods of SE of AUC$^{\Phi}$ reflects better the actual variation than semi-parametric approach with non-binormal data while for binormal data the semi-parametric method yields a more precise estimate of SE$^{=}$.
Hajian-Tilaki KO et al., [30]	Comparing nonparametric and semi-parametric	With numerical investigation it has been shown that nonparametric and semi-parametric approach yields a similar estimate of AUC$^{\Phi}$.

[Table/Fig-4]: The related studies of performance of nonparametric ROC method of quantitative data.
$^{\Phi}$Area under the curve, $^{=}$Standard error

indices are obtainable even for small sample sizes. The Mann-Witney U-statistic (or Wilcoxon statistic) allows to one estimate AUC [6,42,43]; the partial area for a clinically relevant range of FPF and TPF_{FPF} can also be estimated nonparametrically [43,44]. [Table/Fig-4] summarized the related studies in assessment of performance of nonparametric AUC and its SE. The practitioner can easily use SPSS software to depict the nonparametric ROC curve and to estimate AUC (Mann-Whitney U-statistic), the corresponding SE and 95% asymptotic confidence interval for AUC and it p-value in single modality.

The corresponding sampling variability of nonparametric estimates of AUC can also be calculated nonparametrically as well. The three methods have been proposed for SE of AUC. The first was Bamber method that estimates the null variance of Wilcoxon statistic [41]. The second was Hanley and McNeil method that used Bamber's formula with exponential approximation as follows [6]:

$$Var(\widehat{AUC}) = \frac{AUC(1-AUC)+(n_1-1)(Q_1-AUC^2)(n_2-1)(Q_2-AUC^2)}{n_1 n_2}$$

Where, n_1 and n_2 represent the sample size for nondiseased and diseased subjects respectively. The exponential approximation to estimate Q_1 and Q_2 as and

$$Q_1 = \frac{AUC}{2-AUC} \text{ and } Q_2 = \frac{2AUC^2}{1+AUC}$$

The third was Delong's method that has two components of pseudoaccuracy for diseased and nondiseased [42]. The advantage of DeLong's method is to compare the areas under two or more correlated ROC curves obtained from the same sample of subjects. The variance-covariance matrix is:

$$Var[\widehat{AUC_1}, \widehat{AUC_2}] = \frac{S_N}{n_1} + \frac{S_D}{n_2}$$

Where, S_N and S_D represent the estimated variance-covariance matrices of the paired pseudo-values in the nondiseased and diseased groups, respectively.

$$Var[\widehat{AUC}] = \frac{Var[V_N]}{n_1} + \frac{Var[V_D]}{n_2}$$

Hanley and Hajian-Tilaki KO [45] restated the DeLong's method of calculating the variance of the accuracy index in a single diagnostic test is as follows:

Where Var (V_N) and Var (V_D) are the variance of pseudoaccuracies in nondiseased and diseased subjects respectively. These components of pseudoaccuracy has been examined by Hanley and Hajian-Tilaki KO, and a full explanation of component of Delong's method was presented in particular for comparison of two AUCs and the link between Delong method of pseudoaccuracies and jackknife pseudovalues has been justified [43]. The correlation between Delong individual pseuaccuracies is used for multiple signal detection in ROC analysis [46]. For rating data, the Bamber method [41] and Delong method [42] give equally good results and better than Hanley and McNeil method [6]. In particular, the exponential approximation (Hanley JA and McNeil BJ method) slightly underestimates the empirical standard error of nonparametric estimate of the AUC for rating data when the ratio of the standard deviation of distributions for the "diseased" to the "nondiseased" was greater than two [45]. With continuously distributed data generated from binormal and nonbinormal models, Hajian-Tilaki KO and Hanley JA have shown that the Delong method SE of nonparametric AUC reflects the actual variation and give equally good results compared with parametric SE of AUC and exponential approximation slightly underestimates the SE of AUC compared with actual variation even for continuously distributed data [46]. In particular, they showed, for bimodal data, Delong method apparently reflects better the actual variation than semi-parametric approach while for binormal data, the semi-parametric approach yields a more precise estimates of AUC. In contrast, DeLong method provides an estimated covariance matrix for comparing the area under to or more correlated ROC curve using the theory of generalized U-statistic to provide an estimated covariance matrix [42]. In addition, nonparametric approach allows one to estimate AUC for degenerate data sets (i.e., the data set does not converge by MLE method in parametric approach). If the sample size is adequate, the nonparametric approach would avoid making distributional assumptions which would be viewed as somewhat restrictive. Overall, for continuous data, the binormal method implemented in LARROC procedure and nonparametric method yields a similar estimate of AUC and its SE [30]. Thus, the choice between semi-parametric and purely nonparametric methods for estimates of AUC depends on availability of software and practical conveniences.

Confidence Interval of AUC and Testing

Testing of AUC with a specific predetermined value in asingle modality or testing of equality of accuracy of two alternative diagnostic tasks is interesting for decision making in clinical practice. The p-value that the authors report in their results gives an intuition how likely the results reported is due to chance. However, it does not explainabout the precision of estimate and how the diagnosis parameter varied in plausible range with a confidence level. Instead SE of estimates is primary interested for precision purpose but Confidence Interval

(CI) of diagnosis accuracy reports a range of plausible results with a specific confidence level. Asymptotic properties of normal approximation of AUC in a single diagnostic task allows to derive confidence interval as

$$\widehat{AUC} \pm Z_{1-\alpha/2} SE(\widehat{AUC})$$

and in comparative diagnostic tasks, the confidence interval of difference in accuracy indexes is as

$$\widehat{AUC1} - \widehat{AUC2} \pm Z_{1-\alpha/2} SE(\widehat{AUC1} - \widehat{AUC2})$$

However, the asymptotic property of AUC may not be satisfied as AUC is close to 1. In this scenario, the exact binomial CI of AUC or bootstrapping CI is recommended[47,48].

All relevant statistical software can calculate the 95% CI for AUC that is the most commonly used in medical reports. For example, 95% CI of AUC: 0.75-0.85 i.e. 95% of times through the sampling replication, the true AUC would be within 0.75 and 0.85. In another word, if the study was performed with 100 times with sampling replication, 95% of times the true accuracy index is within the range such as 0.75-0.85. The clinical researchers may only report the p-value but reporting CI is more informative rather p-value alone.

In comparative diagnostic studies, a common problem is to test the accuracy of two diagnostic tasks that perform on the same subjects. Many ROC methods and also some software do not take into account the correlation between two ROC curves that estimates on the same subjects. Thus, the investigator should clearly say for their methods whether they accounted such correlation between two ROC curves. The CORROCFIT program provides such as unbiased comparison and the SE of difference of AUC's is estimated by taking into account of the correlation between two ROC curves [24]. Also, Delong methods of SE estimation allows for pairwise comparison that was implemented in Stata software [42]. Additionally, taking into account the correlation between two ROC curves, in estimating SE of difference in AUC's, the power of statistical test would be increased.

Re-Sampling Methods in ROC Analysis

The sampling distributions of ROC indices (AUC and TPF) are rather complex and asymmetric. Thus, asymptotic theory of confidence interval for AUC and TPF may yield invalid estimates, in particular for small sample size. The two attractive approaches of re-sampling techniques (jackknife and bootstrap methods) have been adapted to estimate sampling variability of AUC and TPF and their confidence intervals.

Jackknife Method: The main concept of jackknife method in ROC analysis is pseudovalues as a replacement of raw data value by an equivalent [43]. The pseudovalue for each observation can be defined as the contribution of that observation to the summary index. The average of these pseudovalues gives the same summary index as raw data. The variation of these pseudovalues allows one to calculate the sampling variability of ROC indices. The jackknife pseudovalues can be calculated for any indices. For example, if AUC is summary index, the AUC pseudovalues (pAUC) corresponding to i^{th} observation is defined as:

$$pAUC=(m+n)\ AUC-(m+n-1)AUC_{(-i)}$$

where AUC is calculated for all m+n observation (m for diseased and n for nondiseased) and AUC(-i) calculated for m+n-1 observations when the ith observation omitted from samples. Then, the jackknife variance of summary index is: Var(AUC)= Var(pAUC's)/m+n

The method allows one to compare the two ROC curves derived from the same subjects [10]. The covariance between two pseudovalues from the same subject can be calculated by correlation between them. Hanley and Hajian-Tilaki KO algebraically showed the link between jackknife pseudovalues and Delong pseudoaccuracies and they showed that two approaches gave equally results in SE of nonparametric AUC [42]. This approach was adapted for assessment of multiple reader variation in ROC analysis [49]. Yao W et al., developed the nonparametric method for estimation of AUC in order properly into account the sampling scheme in complex sample survey data [50]. They used jackknife and balanced repeated replication methods to estimate variance of nonparametric AUC and to account for complex sample design and sample weighting.

Bootstrapping method: Another re-sampling technique called bootstrapping has been used to obtain the SE and confidence interval of AUC and TPF. This technique is a statistical procedure for estimation of distribution function and it has been used for ROC analysis by several authors [51,52] - see Appendix for bootstrap algorithm. The method allows to derive the density distribution of AUC (or TPF or partial area) and to obtain the percentiles of these distributions and their 95% confidence interval.

Regression Approach in ROC Analysis

In most diagnostic studies, ROC curve for diagnostic test have been reported without taking into account the possible effect of covariate on test results [53]. However, the test results might be influenced by age and severity of disease [54]. Thus pooling all data into single ROC curve might distort the accuracy of diagnostic test. This occurs when covariate is associated with both test results and the true state of disease [55,56]. Janese H and Pepe MS graphically showed that the overall ROC curve and corresponding AUC substantially differ from stratum specific ROC curve and their AUC when confounding is present [53]. Thus, the covariate adjustment is required in ROC analysis. For example, for evaluation the accuracy of PSA (prostate serum antigen) concentration, the influence of patient age had been reported [56]. Nevertheless, few clinical investigators adjusted the effect of covariates in clinical practice of diagnostic test evaluation. At least three methods to regression for adjustment of covariate in ROC analysis have been developed [25,33,57,58]. The first, Tosteson ANA and Begg CB proposed the use of ordinal regression model for covariate adjustment into ROC curve [25]. Second, models that quantify covariate effects on summary measure of ROC curve such as AUC have been proposed by Thompson ML and Zucchini W [59]. The third uses the General Linear Model (GLM) thathas been suggested by Pepe MS [58]. The latter can be used a broader range of test results and covariates [33,56]. The main concern of this direct modeling is that the estimating equation might be difficult to solve numerically for estimation of model parameters [33]. This analysis can be performed using standard software for fitting GLM. Stata software has a feature for adjustment of covariates in ROC analysis. Moreover, Cai T and Zheng Y extended cumulative residual based approach for graphical and numerical assessment of model checking for ROC regression analysis [60]. With this framework, the goodness of fit of some commonly model used in ROC analysis can be evaluated. This regression methodology is an active area of ongoing research in ROC analysis from direct to indirect method of modeling of covariate adjustment.

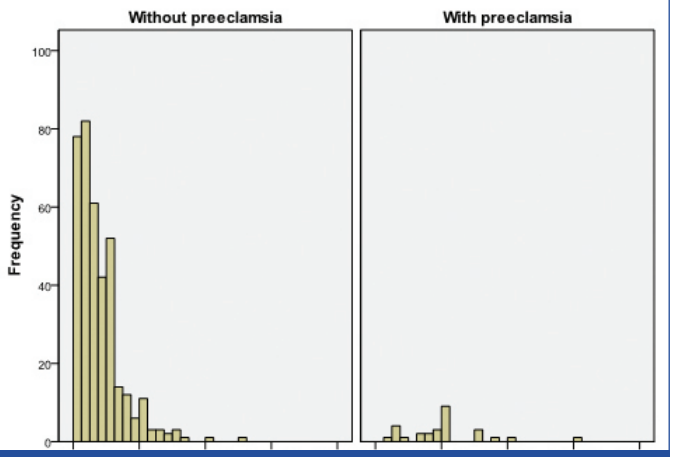
Measuring of DiagnosticAccuracy with Three Diagnostic Categories

ROC methodology is no longer confined to only two diagnostic categories "diseased" and 'nondiseased" but also the method proposed for estimating the diagnostic accuracy when there are three ordinal diagnostic groups [61]. There aremotivative examples where patients with Alzheimer's disease are usually classified into three groups and should receive different categories specific treatment. In this situation, the ROC curve surface describes the probability of correct classification in the three diagnostic groups based on various set of diagnostic thresholds used and it was proposed the

entire or partial volume under the ROC surface measures diagnostic accuracy [62]. The description of details of this method is beyond the scope of this article. The interested readers are referred to some recent published articles [61,62].

Software Development

Several software programs have been implemented for parametric and nonparametric ROC analysis. As we already mentioned, ROCKIT software developed by Metz CE et al., at university of Chicago for parametric and semi-parametric ROC analysis that is widely available. It is used for comparison of two ROC curves for paired and unpaired subjects but not for partial area. Stata software can be used for comparison of nonparametric and parametric two partial areas and it has all features of ROC analysis. SPSS software produces nonparametric ROC curve and AUC and testing with the null value (AUC=0.5) but it does not test for comparison of two ROC curves and the partial area. Stat 2010 software provides nonparametric ROC analysis of both paired and unpaired subjects for AUC but not partial area. Additionally, Sigma plot program also is available for nonparametric ROC analysis. Furthermore, SAS releases (version 9.2 or higher) and several other SAS ROC macros have been built for parametric and nonparametric ROC analysis. As we already pointed out, bi-logistic model and bi-exponential



[Table/Fig-5]: The distribution of CRP among subjects with and without preeclampsia.

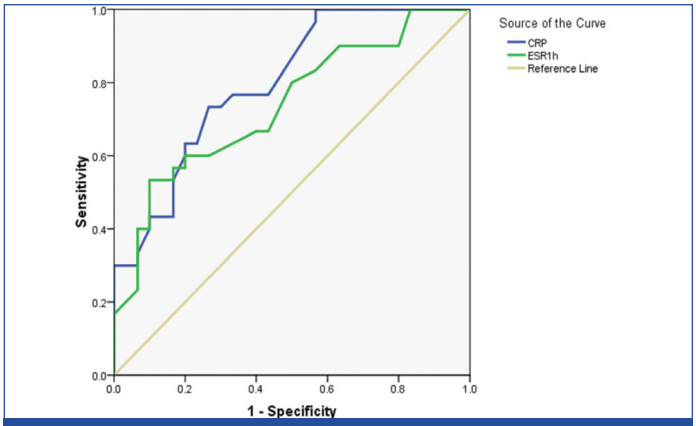
Methods	AUC	SE(AUC)			
		Delong	Bamber	Hanley	Binormal
Nonparametric	0.844	0.0402	0.0401	0.0396	-
Parametric	0.850	-	-	-	0.0389

[Table/Fig-6]: The parametric and nonparametric estimates of diagnostic accuracy of CRP for predicting of preeclampsia and its SE with different methods.
AUC: Area under the curve; SE: Standard error; CI: Confidence interval

model can be fit using PLUM software but this software is not currently widely available. Moreover, function of R software so called "roc.test" (pROC) compares the AUC or partial area AUC of two correlated or uncorrelated ROC curves and three methods: "Delong", "bootstrap"and "Venkatraman" are used.

Examples and Illustrations

Example 1: A single diagnostic task: We applied both parametric and nonparametric methods with real clinical data to show how different methods yield rather similar results. In a clinical study, 400 pregnant women were recruited to demonstrate the predictive ability of serum C-Reactive Protein (CRP) levelfor preeclampsia. The distribution of data of CRP has been shown in [Table/Fig-5] among subjects with and without preeclampsia (n=28 and n=372 respectively). As shown the distribution of CRP rather deviates from normality and both t -test and Wilcoxon rank test



[Table/Fig-7]: ROC curve of CRP and ESR1h in predicting IBD patients.

Biomarkers	n	AUC	SE	95% CI	p-value*
CRP	60	0.798	0.056	0.689-0.908	0.48
ESR1h	60	0.737	0.064	0.611-0.864	

[Table/Fig-8]: Comparison of diagnostic accuracy of CRP and ESR1h for prediction of IBD patients.
H_0: area (CRP) = area(ESR1h), *Chi-square test; AUC: Area under the curve; SE: Standard error; CI: Confidence interval

showed a significant difference of the mean of CRP between two groups (10.03±6.16 and 3.77±3.45 respectively, p-value=0.001). We applied ROC analysis using Stata 13.0 software. This software has a special feature to calculate the diagnostic accuracy (AUC) and its SE both parametrically and nonparametrically. Using binormal model the rocfit program were performed. As results shown in [Table/Fig-6], the parametric and non-parametric AUC yielded rather similar (the estimates of AUC were 0.850 and 0.844 respectively). We also calculated the three methods of nonparametric SE of AUC, Hanley JA and McNeil BJ, Bamber D, and DeLong ER et al., [6,41,42]. The estimates of SE of Delong's and Bamber's methods were identical but Hanley's method produced slightly smaller. However, the estimate of SE was lower than other methods with binormal fit using Delta method approximation.

Example 2. A comparative study of two diagnostic tasks: In a clinical study of 30 patients of Inflammatory Bowel Diseases (IBD) and 30 healthy controls, the data of two biomarkers: CRP and ESR1h were measured on the same subjects and thus two sets of outcomes are correlated (the ROC curves are correlated). Roccomp function in Stata software was used to compare the classification ability of two modalities for predicting of true disease status. We plotted two ROC curves corresponding to CRP and ESR 1 hr and their area were compared in [Table/Fig-7]. For each biomarker, the summary statistics were reported in [Table/Fig-8] and the test for the equality of area under the curves was performed using the methods proposed by DeLong ER et al., [42]. Although the AUC for CRP is slightly larger than ESR 1 hr, the chi square test yields a non-significant p-value of 0.48.

DISCUSSION

Despite the flexibility of binormal model for quantitative data in ROC analysis, using the LABROC procedure produces an unbiased estimate in AUC when underlying distribution is far from Gaussian [19,27]. However, several numerical investigations have shown that in a case of severe departure from binormality, especially bimodal forms, the calculated SE of AUC overestimates its empirical SE and the bias in estimates of TPF is considerable [30-32]. A more flexible parametric form of mixture of Gaussian for underlying distributions of test results yields an unbiased estimate of AUC and TPF and their SE [36]. However, a little attentionhas been donein clinical practice because of practical inconvenience and the lack of availability of software.

Alternatively, nonparametric ROC analysis has been used widely for

numerical test results [41-44]. An attractive feature of this approach is the pseudoaccuracy that is interpreted the accuracy of test at individual level. The average of these pseudoaccuracies is equivalent to Wilcoxon statistic (or AUC) with meaningful interpretation [42]. The variance of AUC is the sum of variance of two components of pseudoaccuracies in diseased and nondiseased divided by the number of subjects respectively [43]. As we already noted, Hanley JA and Hajain-Tilaki KO restated Delong's method of SE of AUC in a single diagnostic task [45] but the attractive future of Delong's methods is the covariance of two correlated ROC can be calculated from the structure of variance-covariance matrix of pseudoaccuracy elements. The results of our illustrated example show that in spite of departure of data from binormality, the semi-parametric and nonparametric methods yielded similar results in AUC and its SE. While Hajian-Tilaki et al., showed that the Delong's method of estimates of SE better reflects the actual variation than binormal based and exponential approximation with numerical investigation [46].

In comparing two diagnostic tasks, the clinician almost use the same subjects. The analysis should consider the covariance of two correlated AUC or any other accuracy indexsuch a paired design. The lack of consideration of pair design in analysis, the distinction between two diagnostic tasks may not be detected and the power of statistical test would be decreased. The relevant issues must be considered in sample size calculation in the design of diagnostic studies and analysis [63].

Despite some methods presented for covariate adjustment in ROC analysis [25,57,58], little attention has been focused in regression approach inanalysis of medical published articles of diagnostic studies. This might be due to the lack of availability of established methods and relevant software. Additionally, there are several recent new development of ROC methodologies with involving multiple signals (three or more diagnostic categories) [61,62] and also combing the information of multiple biomarkers to optimize the ROC curve [64-67]. The research in this area is still continuing and these new methods were not used widely in clinical investigations yet.

CONCLUSION

In spite of attractive performance of semi-parametric approach of binormal model implemented in LABROC, in some circumstance the fit may be less appropriate with severe departure from binormality. Thus, a more flexible parametric model such as bi-mixture of Gaussian needs to be implemented in LABROC. This model is more robust to departure from binormality. In addition, more numerical investigations are needed to compare the performance of LABROC and GLM semi-parametric with purely nonparametric method for quantitative diagnostic test data.

ACKNOWLEDGEMENTS

The author acknowledges the Deputy of Research of Babol University of Medical Sciences for their support.

REFERENCES

[1] Perry GP, Roderer NK, Asnar SA. Current perspective of medical informatics and health sciences librarianship. J Med Libr Assoc. 2005;33:199-206.
[2] Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. Invest Radiol. 1979;14:109-21.
[3] Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. Crit Rev Diagn Imaging. 1989;29:307-35.
[4] Lu C, Van Gestel T, Suykens JAK, Van Huffel SV, Vergote I, Timmerman D. Preoperative prediction of malignancy of ovarian tumors using least squares support vector machines. Artificial Intelligence in Medicine. 2003;28:281-306.
[5] Faraser HSF, Lonc WJ, Naimi S. Evaluation of a cardiac diagnostic program in a typical clinical setting. J Am Med Inform Assoc. 2003;10:373380.
[6] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29-36.
[7] Hanley JA, McNeil BJ. A method of comparing the area under receiver operating characteristic curves derived from the same cases. Radiology. 1983;148:839-43.
[8] Kumar R, Indraya A. Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatr. 2011;48(4):277-89.
[9] McClish DK. Analysing a portion of the ROC curve. Med Decis Making. 1989;9:190-95.
[10] Choi BC. Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. Am J Epidemiol. 1998;148(11):1127-32.
[11] Hajian-Tilaki K. Receiver operating characteristic ROC) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med. 2013;4(2):627-35.
[12] Liu X. Classification accuracy and cut point selection. Stat Med. 2012;31:2676–86.
[13] Hajian-Tilaki K. The choice of methods in determining theoptimal cut-off value for quantitativediagnostic test evaluation. Stat Meth Med Res. 2017;DOI: 10.1177/0962280216680383.
[14] Böhning D, Holling H, Patilea V. A limitation of the diagnostic-odds ratio in determining an optimal cut-off value for a continuous diagnostic test. Stat Meth Med Res. 2011;20:541–50.
[15] Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. J Math Psychol. 1969;6:487-96.
[16] Wieand S, Gail, MH, James BR, James KL. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika. 1989;76:585-92.
[17] Reddy S, Dutta S, Narang A. Evaluation of lactate dehydrogenase, creatine kinase and hepatic enzymes for retrospective diagnosis of perinatal asphyxia among sick neonates. Indian Pediatr. 2008;45:144-47.
[18] Hajian-Tilaki KO, Gholizadehpasha AR, Bozogzadeh S, Hajian-Tilaki E. Body mass index and waist circumference are predictor biomarkers of breast cancer risk in Iranian women. Med Oncol. 2011;28(4):1296-301.
[19] Hanley JA. The robustness of the 'binormal' assumptions used in fitting ROC curves. Med Decis Making. 1988;8:197-203.
[20] Grey DR, Morgan BGT. Some aspect of ROC curve-fitting: normal and logistic models. J Math Psychol. 1972;9:128-39.
[21] Obuchowisky NA. Sample size calculation in studies of test accuracy. Statistical Methods in Medical Research. 1996;7:371-92.
[22] FORTRAN programs ROCFIT, CORROC2, LABROC1, LABROC4 and ROCPWR. Available from Metz CE. Department of Radiology, University of Chicago, Chicago 1990.
[23] Metz CE. ROC methodology in radiological imaging. Invest Radiol. 1986;21:720-33.
[24] Metz CE, Wang PL, Kronman HB. A new approach for testing the significance of differences between ROC curves from correlated data. In: Deconick F, Eds. Information processing in medical imaging, The Hague, Nijhoff, 1984; pp 432-445.
[25] Tosteson ANA, Begg CB. A general regression methodology for ROC curve estimation. Med Decis Making. 1988;8:207-15.
[26] Begg CB. Advances in statistical methodology for diagnostic medicine in the 1980's. Stat Med. 1991;10:1887-95.
[27] Hanley JA. The use of the 'binormal' model for parametric ROC analysis of quantitative diagnostic tests. Stat Med. 1996;15(14):1575-85.
[28] Goddard MJ, Hinberg I. Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. Stat Med. 1990;9:325-77.
[29] Metz CE, Herman BA, Shen J-H. Maximum-likelihood estimation of ROC curves from continuously-distributed data. Stat Med. 1998;17:1033-53.
[30] Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP. A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. Med Decis Making. 1997;17:94-102.
[31] Hajian-Tilaki KO, Hanley JA. Robustness of LABROC procedure to fit binormal model for estimating TPF in ROC analysis of quantitative diagnostic data. Proceeding of 6th International Congress of Iranian Statistical Society, Allameh-Tabatabaie University, Tehran, 2004: Pp.169-196.
[32] Faraggi D, Reiser B. Estimation of the area under the ROC curve. Stat Med. 2002;21:3093-106.
[33] Pepe MS. An interpretation for the ROC curve and inference using GLM procedures. Biometrics. 2000;56(2):352–59.
[34] Cai T, Moskowitz CS. Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. Biostatistics. 2004;5(4):573–86.
[35] Colak E, Mutlu F, Bal C, Oner S, Ozdamar K, Gok B, et al. Comparison of Semiparametric, parametric and nonparametric ROC analysis for continuous test using a simulation studies and acute coronary syndrome data. Comput Math Methods Med. 2012;2012;698320.
[36] Hajian-Tilaki KO, Hanley JA, Nassiri V. An extension of parametric ROC analysis when the underlying distributions are mixture of Gausian. J Appl Stat. 2011;38(9):2009-22.
[37] Render RA, Walker HF. Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev. 1984;26:195-239.
[38] Wang BO, Titterington DM. Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. Bayesian Analysis. 2006;1(3):625-50.
[39] Diebolt I and Robert CP. Estimation of finite mixture distributions through Bayesian sampling. J R Statist Soc B. 1994;56:363-75.
[40] Sorribas A, March J, Trujllano J. A new approach parametric based on S-distributions for computing receiver operating curves for continuous tests. Stat Med. 2002;21(9):1213-35.
[41] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating graph. J Math Psychol. 1975;12:387-415.
[42] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two

or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837-45.
[43] Hanley JA, Hajian-Tilaki KO. Sampling variability of nonparametric estimates of the area under receiver operating characteristic curves: An update. Accad Rdiol. 1977;4(1):49-58.
[44] Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP. An extension of ROC analysis to data concerning multiple signals. Acad Radiol. 1977;4(5):225-29.
[45] Obuchowski NA. Computing sample size for receiver operating characteristic studies. Invest Radiol. 1994;29:238-43.
[46] Hajian-Tilaki KO, Hanley JA. Comparison of three methods for estimation the standard error of the area under the curve in ROC analysis of quantitative data. Acad Radiol. 2002;9:1278-85.
[47] Wunderlich A, Noo F, Gallas BD, Heilbrun ME. Exact confidence intervals for channelized Hotelling observer performance in image quality studies. IEEE Trans Med Imaging. 2015;34(2):453-64.
[48] Platt RW, Hanley JA, Yang H. Bootstarp confidence intervals for the sensitivity of a quantitative diagnostic test. Stat Med. 2000;19(303):313-22.
[49] Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the Jackknife method. Invest Radiol. 1992;27:723-31.
[50] Yao W, Li Z, Graubard BI. Estimation of ROC curve with complex survey data. Stat Med. 2015;34(8):1293-303.
[51] Mossman D. Re-sampling techniques in the analysis of non-binormal ROC data. Med Decis Making. 1995;15(4):358-66.
[52] Margolis DJ, Bilker W, Boston R, Localio R, Berlin JA. Statistical characteristic of area under the receiver operating characteristic curve for simple prognostic model using traditional and bootstrapped approaches. J Clin Epidemiol. 2002;55(5):518-24.
[53] Janese H, Pepe MS. Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: An old concept in a new setting. Am J Epidemiol. 2011;168:89-97.
[54] Medeiros FA, Sample PA, Zangwill LM, Liebmann JM, Christopher A, Girkin CA. A statistical approach to the evaluation of covariate effects on the receiver operating characteristic curves of diagnostic tests in glaucoma. Invest Ophthalmol Vis Sci. 2006;47:252027.
[55] Hajian-Tilaki K. Methodological issues of confounding in analytical epidemiologic studies. Caspian J Intern Med. 2012;3(3):488-95.
[56] Oesterling JE, Cooner WH, Jaconsen SJ, Guess HA, Lieber MM. Influence of patient age on serum PSA concentration. An important clinical observation. Urol Clin North Am. 1993;20:671-80.
[57] Zou KH, Liu A, Bandos AL, Bandos AI, Ohno-Machado L, Rockette HE. Statistical evaluation of diagnostic performance: Topic in ROC analysis. Boca Raton FL: Chapman & Hall CRP Press 2011.
[58] Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. Biometrics. 1998;54:124-35.
[59] Thompson ML, Zucchini W. On the statistical analysis of ROC curves. Statistics in Medicine. 1989;81277-1290.
[60] Cai T, Zheng Y. Model checking for ROC regression analysis. Biometric. 2007;63(1):152-63.
[61] Li J, Fine JP. ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. Biostatistics. 2008;9(3):566-76.
[62] Xiong C, van Belle G, Miller JP, Yan Y, Gao F, Yu K, et al. A parametric comparison of diagnostic accuracy with three ordinal diagnostic groups. Biomedical Journal. 2007;49(5):682-93.
[63] Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform. 2014;48:193-204.
[64] Fong Y, Yin S, Huang Y. Combining biomarkers linearly and nonlinearly for classification using the area under the ROC curve. Stat Med. 2016. Doi:10.1002/sim.6956.
[65] Yin J, Tian l. Optimal linear combinations of multiple diagnostic biomarkers based on Youden index. Stat Med. 2014;33(8):1426-40.
[66] Yan L, Tian L, Liu S. Combining large number of weak biomarkers based on AUC. Stat Med. 2015;34(29):3811-30.
[67] Kang LE, Liu A, Tian L. Linear combination methods to improve diagnostic/prognostic accuracy on future observations. Stat Method Med Res. 2013. doi:10.1177/0962280213481053.

PARTICULARS OF CONTRIBUTORS:
1. Professor, Department of Biostatistics and Epidemiology, Babol University of Medical Sciences, Babol, Iran.

NAME, ADDRESS, E-MAIL ID OF THE CORRESPONDING AUTHOR:
Dr. Karimollah Hajian Tilaki,
Professor, Department of Biostatistics and Epidemiology, Babol University of Medical Sciences, Babol, Iran.
E-mail: drhajian@yahoo.com

FINANCIAL OR OTHER COMPETING INTERESTS: None.

Date of Submission: **Sep 23, 2017**
Date of Peer Review: **Dec 05, 2017**
Date of Acceptance: **Mar 24, 2018**
Date of Publishing: **Jun 01, 2018**