Research Article   ⬤ **Open Access**

Screening and Diagnostic tests for Health-related Quality of Life

Satyendra Nath Chakrabartty*

* Indian Statistical Institute, Indian Institute of Social Welfare and Business Management, Indian Ports Association.

Corresponding Author: Satyendra Nath Chakrabartty, *, Indian Statistical Institute, Indian Institute of Social Welfare and Business Management, Indian Ports Association.

Abstract

Objective: The paper describes conceptual framework and methodological issues to find cut-off points considering summative scores with emphasis on Receiver Operating Characteristic (ROC) curve and associated components like area under the curve (AUC), sensitivity and specificity along with scoring and their properties.

Material and Methods: To establish presence (or absence) of disease as a basis for treatment decisions, diagnostic tests require meaningful total/domain scores. Problems of evaluation of Health-related Quality of Life (HRQoL) scales and remedial measures by transforming item scores to continuous, monotonic scores following normal distribution satisfying desired properties, undertaking parametric analysis and better utilization of such tests are also suggested. Ethical approval of research is not required for the methodological paper involving no human beings or animals.

Results: Parametric method may result in improper ROC curve if data violate the normality assumption or within-group variations are dissimilar (heteroscedasticity). Equivalent non-parametric ROC curve, without involving assumption of normality can be obtained by ROCKIT package containing ROCFIT method along with estimate of AUC and partial area, under certain conditions. If multiple test results from a single case-condition combination are pooled in input, statistical significance of apparent difference between the ROC curves gets overestimated.

Conclusion: Transforming HRQoL scores to normally distributed scores with meaningful arithmetic aggregation have promise to improve current assessment practices.

Key words: Area under the curve, Cut-off point, Normal distribution, ROC curve, Sensitivity, Specificity

Introduction

Screening tests with high sensitivity and diagnostic tests with high specificity are different. While screening tests aim at early detection of disease or risk factors for disease, diagnostic tests establish presence (or absence) of disease as a basis for treatment decisions in symptomatic or screen positive individuals (confirmatory test). Thus, screening tests are unresponsive to some diseases (false negatives) and may raise suspicion of the disease where none exists (false positives)[1]. The common point is to decide a score called critical or cut-off score (C_0) in a specific population such that persons with scores less than C_0 are persons without the disease and scores exceeding C_0 indicate presence with the disease.

Quality of Life (QoL) and Health-related Quality of Life (HRQoL) questionnaires consist of finite number of domains and items in each domain which are binary (Yes-No type) or *K*-point Likert items (*K*= 3, 4, 5…) or combination of both. Number of domains, number of items per domain and number of response-category per item are different for different scales. HRQoL is a subjective, multidimensional concept based on self-perceptions of individuals on their own health covering dimensions like physical function, mental function, psychological well-being, social function, role function, and perceptions of functions and well-being [2]. Thus, HRQoL scales could be

generic for global assessment or disease-specific customizing specific problem areas of diseases.

Different methods are used to find cut-off points considering summative domain scores or scale scores as sum of domain scores. Domain-wise cut-off score are usually reported. Limitations of cut-off points were observed [3] based on arbitrary percentile values and suggested optimal cut-off points for six HRQoL scales in four prognostic classifications of patients with hepatocellular carcinoma (HCC) using the complex method of [4] which involves partitioning the sample into learning and validation sub-samples and obtaining optimal cut-off point for each sub-sample using the minimum p value associated with the maximal log-rank statistic and selecting final cut-off point as the value that minimizes the p value in the overall sample, using stratified log-rank test with sub-sample as the stratum. Popular method to find cut-off point is to use Receiver Operating Characteristic (ROC) curve and area under the curve (AUC) along with indication of sensitivity and specificity [5]. Four cut-off scores differed by treatment status for Cancer Core Questionnaire (EORTC QLQ-C30) was found [6].

Methodological issues:

- Pathological and clinical tests result in continuous data like bilirubin level and nominal data (categories) like Mantoux test for tuberculosis diagnosis. Data generated by HRQoL tests are usually ordinal and discrete (responses of Likert item). Each case requires obtaining total/domain scores and converting them into dichotomous groups to decide presence or absence of a disease. Combining continuous variables in ratio scale and nominal or discrete ordinal variables is problematic.

- Different methods of finding cut-off scores using scale scores give different cut-off scores. Test validity i.e. association between HRQoL scale and clinical and pathological findings may be computed keeping in mind that both have limitations. A case study by [7] having missing core clinical features of Dementia with Lewy Body (DLB) and negative indicative biomarkers, but neuropsychological tests and Positron emission tomography imaging provided crucial evidence for DLB even in early stages, but core symptoms and indicative biomarkers appeared in later stages of the disease.

- Cut-off points depend on methods used and also on sample characteristics like proportion of participants in the poor/dissatisfied QoL group. It is not possible to compare cut-off points from different researches to evaluate QoL by specific tool [5], who suggested further investigations on different cut-off points for better discriminations and comparisons.

- Population estimations are not straight forward without knowledge of distribution of domain scores and test scores.

- For *Stroke-Adapted Sickness Impact Profile* (SA-SIP30) with 30 items in 8 subscales, cut-off score >33 and the same for *Sickness Impact Profile* (SIP136) with 136 "Yes–No" type items distributed over 12 domains > 22. How to establish those cut-off scores are equivalent?

- *Seattle Angina Questionnaire F*(SAQ), *Short Form Health Survey questionnaire*(SF-36) does not provide total score reflecting overall health status

- Higher score in *Nottingham Health Profile* (NHP), *Minnesota Living with Heart Failure* (MLHF) imply more significant health problems, unlike *Sickness impact Profile* (SIP). Thus, directions of scores are different for different HRQoL scales. If the test results for NHP or MLHF shows AUC = 0, then SIP will show AUC = 1 implying a perfectly inaccurate test can be transformed into a perfectly accurate test.

- Scales differ in number of items (length) and number of response-categories (width), scoring system, dimensions chosen, measurement properties, etc. Mean, variance of K-point Likert scale scores tend to increase as K increases.

- Sum of ordinal responses to items are not meaningful due to non-satisfaction of equidistant property and unjustified assigning of equal importance to items and domains, showing different correlations with total scores and different factor loadings [8].

Minnesota Multiphasic Personality Inventory (MMPI) was found to have poor measure of emotional adjustment 9 and was misscored at high rate primarily due to scoring procedure complexity and commitment to accuracy 10. Lack of methodological unanimity as far as measuring QoL is concerned was observed 11.

Likert scores are often skewed and do not satisfy normality, which is the common assumption of parametric analysis like AVOVA, regression, estimation and testing of population parameters, etc.

Distributions of item scores are unknown and different. Interpretation of X ± Y and further operations on X ± Y and

finding joint distribution of item scores are problematic when X and Y follow two different distributions, that too unknown. Method to transform ordinal item scores to normally distributed scores was suggested which also helps to find equivalent cut-off scores of two scales 12.

Item-wise raw scores are transformed differently by different scales to obtain domain scores. Domain score of MacNew Heart Disease Health–Related Quality of Life Questionnaire (MacNew) is average of the responses in that domain. Cardiovascular Limitations and Symptoms Profile (CLASP) scores are weighted to provide a total for each subscale. Each domain of Myocardial Infarction Dimensional Assessment Scale (MIDAS) is scored separately. Such domain scores create difficulties in meaningful computation of mean, variance, distribution of scale/domain scores for meaningful comparisons, ranking, classifying individuals, and statistical inferences.

Zero value in response-category lowers mean and S.D. Over 40% of the patients had zero score in 10 subscales of SIP and in one subclass of SF-36 13. Too many zeros to an item may artificially lower correlation with that item. When zero is used as an anchor value, expected values (score *corresponding probability) is not meaningful and computation of between group variance will be difficult for the sub-group endorsing zero value to an item, since mean = variance = 0 for that sub-group and correlation with that item is undefined.

5 dimensions of Euro-Quality of Life Questionnaire (EQ-5D-5L) are: Mobility, Self-care, Usual activities, Pain & discomfort, Anxiety and depression. Levels of each item are marked as 1, 2, 3, 4 and 5 where "1" means no problem and "5" means extreme problems. Health-profile of a person is a 5-digit number, minimum being 1-1-1-1-1(no problem in any dimension) and maximum 5-5-5-5-5 (max. problem in each dimension). Instead of score, a person is categorized in one of the possible 3125 = 5^5 categories. Frequency count of each such category is admissible. The scale showed great sensitivity to short-term changes, but the fluctuations of the results are significant, impeding the use of this tool 14.

The paper describes conceptual framework and methodological issues of ROC analysis and associated components like area under the curve (AUC), sensitivity and specificity along with cut-off points and their properties. Problems of evaluation of HRQoL scales and remedial measures by transforming item scores to continuous, monotonic scores following normal distribution for meaningful evaluation satisfying desired properties,

undertaking parametric analysis and better utilization of such tests are also suggested.

Measures of association avoiding contingency table:

Point bi-serial correlation between test score (X) and clinical findings (Y) for classifying individuals as "with disease" or "without disease" is given by

$$r_{pb} = \frac{M_1 - M_0}{S_n} \sqrt{\frac{n_1 n_2}{n}} \text{ where}$$

M_1: Mean of X for the group with Y=1 (sample size n_1)

M_0: Mean of X for the group with Y=0 (sample size n_2); $n = n_1 + n_2$

$$S_n = \sqrt{\frac{1}{n}\sum(X_i - \bar{X})^2} \text{ and } -1 \le r_{pb} \le 1$$

For testing the null hypothesis H_0: r_{pb}= 0, the test statistics is $t = \sqrt{\frac{n_1 - n_2 - 2}{1 - r_{pb}^2}} \sim t$-distribution with $(n_1 - n_2 - 2)$ degrees of freedom

If high test score indicates higher impairment which in turn indicates more intensity of the disease in question, a high positive value of r_{pb} will indicate that a person with high score on X is likely to be classified as "with disease".

Measures of association based on contingency table:

Diagnostic and Screening Tests aim at distinguishing persons who do and do not have the disease of interest. Since test results are recorded as dichotomous outcomes either positive or negative, test validity can be viewed as the measure of association between two outcome variables using contingency table. For a given diagnostic test consider 2 × 2 contingency table as follows:

Table 1: Contingency Table – Clinical evidences and HRQoL Test

		Clinical and pathological evidences showing existence/absence of the Disease		Total
		Yes	No	
HRQoL/ Cognitive tests	Positive	True positive (TP) $f_{(T+D+)}$	False positive (FP) $f_{(T+D-)}$	Row total (R_1)
	Negative	False negative (FN) $f_{(T-D+)}$	True negative (TN) $f_{(T-D-)}$	Row total (R_2)
Total		Column total (C_1)	Column total (C_2)	Grand Total (N)

Measures of association considering cell frequencies of contingency table are:

i) Chi-square measure of association $\chi^2 = \sum \frac{(f_{observed}-f_{expected})^2}{f_{observed}} \sim \chi^2$ distribution with $(r-1)(s-1)$ degrees of freedom where r denotes number of rows and s denotes number of columns of $r \times s$ contingency table and $f_{expected}$ for the (i–j)-th cell is $\frac{Total\ of\ i-th\ row\ \times Total\ of\ j-th\ column}{Grand\ total}$. Significance of value of χ^2 measure of association can be tested.

ii) Pearson's Contingency Coefficient(C) interpreted as a measure of the relative (strength) of an association between two variables and by $= \sqrt{\frac{\chi^2}{N+\chi^2}}$ where N denotes the sample size and C lies between 0 and $\sqrt{\frac{r-1}{r}}$. The measure C can be used for incidence and also prevalence studies but may not be applicable for paired data.

iii) Cramer's V-Coefficient: Value of χ^2 may get increased for large N, even if the variables may not have any substantive relationship. Cramer's V-Coefficient helps to improve association by $= \sqrt{\frac{\chi^2}{N(q-1)}}$ where $q = $ minimum (r, c) and $0 \leq V \leq 1$

iv) χ^2 measure of association, Contingency Coefficient (C),V-Coefficient, etc. can be applied to find validity of a test measuring cognitive impairments by its association with clinical findings related to the same disease along with test of significance of association.

ROC Curve:

Contingency table also helps to find other qualities like Sensitivity and Specificity of the test, which are useful in quantifying the diagnostic accuracy of different tests/measures. Diagnostic accuracy in clinical settings was reviewed [15] who opined that sensitivity and specificity of a test in a clinical setting should not be interpreted in isolation but rather in the context of other diagnostic accuracy statistics like positive predictive power (PPP) and negative predictive power (NPP). However, most of the research studies report sensitivity and specificity statistics and not PPP, NPP and other diagnostic accuracy statistics [16].

Sensitivity indicates the percentage of people with disease that are predicted by the test, i.e. Probability of a person will test positive, given that they have the disease, is given by

$$Sensitivity = \frac{TP}{TP+FN} = \frac{Frequency\ of\ (T+D+)}{Frequency\ of\ (T+D+)+\ Frequency\ of\ (T-D+)} = \frac{f_{(T+D+)}}{C_1} \quad (1)$$

Similarly, the probability of a person will test negative given that they do not have the disease is Specificity which is defined as

$$Specificity = \frac{TN}{TN+FP} = \frac{Frequency\ of\ (T+D-)}{Frequency\ of\ (T+D-)+Frequency\ of\ (T-D-)} = \frac{f_{(T+D-)}}{C_2} \quad (2)$$

Sensitivity and Specificity together reflect clinical utility of a diagnostic test and also help in comparing a new test with the existing one.

Sensitivity and Specificity can be combined to a single measure either by

$$Odd\ Ratio\ (OR) = \frac{f_{(T+D+)} \times f_{(T-D-)}}{f_{(T+D-)} \times f_{(T-D+)}} \quad (3)$$

or $Relative\ Risk$ (RR) $= \frac{f_{(T-D-)}}{R_2} / \frac{f_{(T-D-)}}{R_2}$ (4)

While OR compares prevalence of exposure between two groups formed by outcome, RR compares incidence of disease between two groups formed by outcome. Both fail if the assumption of independence is violated. Confidence intervals (CI) of both OR and RR can be formed to reflect range of uncertainty.

Plot of Sensitivity versus (1-Specifity) is known as receiver operating characteristic (ROC) curve. The area under the curve (AUC) as a one-dimensional index summarizing the "overall" location of the entire ROC curve with meaningful interpretations [17] and help to compare alternative diagnostic tasks when each task is performed on the same sample [18]. Sensitivity and specificity can be computed across all possible threshold values even when the test results are reported on continuous scale [19]. Maximum possible value of AUC = 1 implies no overlapping in the distributions of diseased and non-diseased and the diagnostic test is perfect to differentiate diseased and non-diseased. Equal AUC of test A and B does not imply identical ROC curves for A and B. In fact, two ROC curves may even cross each other. AUC could be interpreted as the average value of sensitivity for all possible values of specificity or the average value of specificity for all possible values of sensitivity [20] or probability that a randomly selected patient with disease has positive test result that indicates greater suspicion than a randomly selected patient without disease [21]. Sensitivity

could be more important than specificity or vice versa. For example, [18] cited an hypothetical example where two diagnostic tests A (AUC = 0.686), and B (AUC=0.679) were applied on the same sample but, performance of Test B was better than test A where high sensitivity is needed, and test A performed better than B when high specificity is required.

In addition, ROC and AUC also help to find the optimal cut-off values by minimizing

$d^2 = (1 - Sensitivity)^2 + (1 - Specificity)^2$ or by commonly used Youden index which maximizes the difference between TPF (sensitivity) and FPF (1-sepicificity). Thus, by maximizing Sensitivity + Specificity across various cut-off points, the optimal cut-off point is calculated or by incorporating financial costs for correct and false diagnosis and the costs of further work up for diagnosis [22].

For a single diagnostic test, the hypothesis $H_0: AUC = AUC_0$ can be tested by considering $Z = \frac{\widehat{AUC} - AUC_0}{SE(\widehat{AUC})}$ where $\widehat{AUC}$ and its standard error (SE) can be estimated by parametric (binormal model) or by non-parametric approaches. Similarly, for two diagnostic tests $H_0: AUC_1 = AUC_2$ can be tested for the same group of subjects using normal approximation under H_0 and defining $Z = \frac{\widehat{AUC_1} - \widehat{AUC_2}}{SE(\widehat{AUC_1} - \widehat{AUC_2})}$. Estimate of $SE(\widehat{AUC_1} - \widehat{AUC_2})$ can best be obtained by method given by [21] which considers covariance between AUC_1 and AUC_2 also.

Properties:

ROC curve based on the notion of a "separator" scale is frequently used in clinical epidemiology to find accuracy of diagnostic tests in discriminating "diseased" and "non-diseased" and if the criterion for positivity is changed, ROC curve shows the tradeoff between the true positive fraction (TPF) and false positive fraction (FPF) [23]. Comparison of two tests requires consideration of the entire ROC curve rather than at a particular point [24].

Measure of accuracy, in terms of AUC reflects ability of the test to discriminate between the diseased and healthy populations [25]. AUC helps to compare individual tests or judge whether the various combination of tests (e.g. combination of imaging techniques) can improve diagnostic accuracy.

AUC ranges from 0 (⟹test fails to correctly classify subjects with diseased as negative and subjects with non-diseased as positive; extremely unlikely in clinical practice)

to 1 (⟹the diagnostic test is perfect in differentiating diseased and non-diseased i.e. no overlapping in the distribution of test results for the diseased and non-diseased). AUC is scale-invariant and also classification-threshold-invariant.

AUC could be interpreted as average value of sensitivity for all possible values of specificity [26] or probability of a randomly chosen diseased person is likely to be rated as diseased than a randomly chosen non-diseased subject [17]. AUC is not affected by decision criterion and is independent of prevalence of disease since it is based on sensitivity and specificity.

Limitations of ROC - AUC:

- Does not consider prevalence of the disease.

- Odd ratio and Relative risk fail if the assumption of independence is violated

- AUC depended heavily on the method used for curve fitting and lacks clinical interpretability

- AUC measures performance over all thresholds, including both clinically relevant and clinically irrelevant thresholds.

- Different tests can have identical/similar AUC but different performance at clinically important thresholds

-Confidence scores used to build ROC curves may be difficult to assign. False-positive and false-negative diagnoses have different misclassification costs. Excessive ROC curve extrapolation is undesirable. Net benefit methods may provide more meaningful and clinically interpretable results than ROC - AUC.

- Perfect test with 100 % sensitivity and 100 % specificity, across all thresholds does not exists

-Change in ROC or AUC has little direct clinical meaning for clinicians

- Sensitivity and specificity are equally important when averaged across all thresholds. However, in CT colonography, poor sensitivity implies missed cancer resulting in delayed treatment or even death, whereas poor specificity may imply unnecessary colonoscopy. Patients and healthcare professionals agreed to accept large number of false-positive diagnoses in lieu of one additional true-positive cancer for mammographic and colorectal cancer screening [27; 28]. In radiological tests with known clinical context, net benefit method, based on changes in sensitivity and specificity at clinically relevant thresholds are more useful to assess clinical impact [29].

-Unjustified arithmetic aggregation of item scores of HRQoL scales to obtain domain scores and scale scores with equal importance may distorts the results. Discrete, ordinal item score were transformed to normally distributed scares in a desired score range, where domain score equals sum of the transformed item scores in that domain and scale score is taken as sum of domain scores [12]. The proposed scores offer meaningful arithmetic aggregations and normal distribution of domain scores and scale scores. Normality facilitates estimation of population parameters and finding equivalent score combinations to integrate two HRQoL scales.

Discussion and Conclusion:

ROC curve is a statistical technique for binary classification assumes continuous random variable X representing scores or function of scores of the diagnostic tests. For a given threshold parameter T, the score is classified as "positive" if X>T, and "negative" if X< T where X follows known probability distribution with density $f_1(x)$ if X>T and $f_0(x)$ if X<T. The true positive rate (sensitivity i.e. probability of detection) is given by TPR (T) = $\int_T^\infty f_1(x)dx$ and the false positive rate (1-Specificity i.e. probability of false alarm) FPR(T) = $\int_T^\infty f_0(x)dx$. The ROC curve plots TPR (T) versus FPR (T) with T as the varying parameter.

Parametric method may result in improper ROC curve if data violate the normality assumption or within-group variations are dissimilar (heteroscedasticity).Nonparametric ROC curve (empirical method) can be used without involving assumption of normality and distribution of the data. Sensitivity and false positive rates are calculated from contingency table based on each cut-off value and are plotted to get a staircase like (jagged line) rather than a smooth curve. Non-parametric AUC is closely related to the Mann–Whitney U statistic which is used to test whether positives are ranked higher than negatives. It is also equivalent to another non-parametric method using Wilcoxon rank test [30]. ROCKIT package containing ROCFIT method (www-adiology.uchicago.edu/krl/KRL_ROC/software_index6.htm) gives non-parametric ROC curve and estimate of AUC, for comparing two tests and to calculate partial area, with the restriction that input data must *not* include more than one test result from each case for each condition, unless there is strong evidence that these test results can be considered independent. If multiple test results from a single case-condition combination are pooled in input,

statistical significance of apparent difference between the ROC curves gets overestimated.

Additionally, a semi-parametric ROC curve is sometimes used to overcome the shortcomings of non-parametric and parametric methods. This method has the advantage of presenting a smooth curve without requiring assumptions about the distribution of the diagnostic test results. However, many statistical packages do not include this method, and is not widely used in the medical research.

ROC and AUC analysis in diagnostic test evaluation, designing of test with broad spectrum of case avoiding bias are necessary for a valid and reliable conclusion in the assessment of performance of diagnostic tests.

ROC- AUC is useful in the early stages of test assessment. But in radiological tests with known clinical context, methods based on net benefit, are more useful to assess clinical impact. Net benefit incorporates estimates of prevalence and misclassification costs, and it is clinically interpretable since it reflects changes in correct and incorrect diagnoses when a new diagnostic test is introduced.

Transforming HRQoL scores to normally distributed scores with meaningful arithmetic aggregation have promise to improve current assessment practices. ROC – AUC analysis considering such normally distributed scores could be investigated empirically with multi-data set along with empirical relationship between Point bi-serial correlation (r_{pb}) and AUC, since high value of r_{pb} indicates that a person with high score is likely to be classified as"with disease"

Declaration:

Acknowledgement: Nil

Conflicting interests: The Author declares that there is no conflict of interest

Funding: This research received no specific grant from any funding agency in the public, commercial, or not for-profit sectors.

Approval of Institutional Review Board: Not applicable

Statement of human and animal right: Not applicable

Data sharing: No data used in this methodological paper

References:

1. McGovern PM, Gross CR, Krueger RA, Engelhard DA, Cordes JE, Church TR. False-Positive Cancer Screens

and Health-related Quality of Life. *Cancer Nursing* 2004; 27(5); 347-352

2. Wilson IB and Cleary PD. Linking clinical variables with health-related quality of life: a conceptual model of patient outcomes. *JAMA* 1995; 273 (1), 59-65

3. Diouf M, Bonnetain F, Barbare JC, Bouché O, Dahan L, Paoletti X, *et al.* Optimal cut points for quality of life questionnaire-core 30 (QLQ-C30) scales: utility for clinical trials and updates of prognostic systems in advanced hepatocellular carcinoma. *Oncologist* 2015; 20(1):62-71. doi: 10.1634/theoncologist.2014-0175

4. Faraggi D. and Simon R. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. *Stat Med* 1996; 15:2203–2213.

5. Silva PA, Soares SM, Santos JF, Silva LB. Cut-off point for WHOQOL-bref as a measure of quality of life of older adults. *Rev Saude Publica* 2014; 48(3):390-397. 10.1590/s0034-8910.2014048004912.

6. Lidington E, Giesinger JM, Janssen SHM, Tang S, Beardsworth S, Darlington AS *et al*. Identifying health-related quality of life cut-off scores that indicate the need for supportive care in young adults with cancer. *Qual Life Res* 2022; 31, 2717–2727 https://doi.org/10.1007/s11136-022-03139-6

7. Bouter Caroline, Hansen Niels, Timäus Charles, Wiltfang Jens, Lange Claudia. Case Report: The Role of Neuropsychological Assessment and Imaging Biomarkers in the Early Diagnosis of Lewy Body Dementia in a Patient with Major Depression and Prolonged Alcohol and Benzodiazepine Dependence. *Frontiers in Psychiatry* 2020; Vol. 11, https://doi.org/10.3389/fpsyt.2020.00684

8. Jamieson S. Likert scales: How to (ab) use them. *Medical Education* 2004; 38, 1212-1218

9. Cripe LI. The MMPI in neuropsychological assessment: a murky measure. *Appl Neuropsychol* 1996*;* 3 (3-4): 97-103.10.1080/09084282.1996.9645373.

10. Allard G. and Faust D. Errors in Scoring Objective Personality Tests. *Assessment* 2000; 7 (2); https://doi.org/10.1177/107319110000700203

11. Boixadós M, Pousada M, Bueno J. and Valiente L. Quality of Life Questionnaire: Psychometric Properties and Relationships to Healthy Behavioral Patterns, *The Open Psychology Journal* 2009; 2, 49-57.

12. Chakrabartty SN. Integration of various scales for measurement of insomnia. *Research Methods in Medicine & Health Sciences* 2021; 2(3), 102-111. 10.1177/26320843211010044

13. Stucki G, Liang MH, Phillips C, Katz JN. The Short Form-36 is preferable to the SIP as a generic health status measure in patients undergoing elective total hip arthroplasty. *Arthritis Care Res*. 1995; 8(3):174-181.10.1002/art.1790080310

14. Klocek M, Brzozowska-Kiszka M, Rajzer M, Kawecka-Jaszcz K. Changes in Quality of Life in Hypertensive patients during home blood pressure Telemonitoring. *Journal of Hypertension* 2010; Vol.28, Issue-pe453 10.1097/01.hjh.0000379556.11680.52

15. Lange R. & Lippa S. Sensitivity and specificity should never be interpreted in isolation without consideration of other clinical utility metrics. *The Clinical Neuropsychologist* 2017; 31(6-7), 1015–1028. doi:10.1080/13854046.2017.1335438

16. Marshall P, James Hoelzl J and Nikolas M. Diagnosing Attention- Deficit/Hyperactivity Disorder (ADHD) in young adults: A qualitative review of the utility of assessment measures and recommendations for improving the diagnostic process, *The Clinical Neuropsychologist* 2021; 35:1, 165-198. 10.1080/13854046.2019.1696409

17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29-36.

18. Kumar R. and Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatr* 2011; 48: 277-289.

19. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diagn Imaging* 1989; 29: 307-335.

20. Zhou Xh, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. New York: John Wiley and Sons 2002

21. 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: 837-845.

22. Christensen E. Methodology of diagnostic tests in hepatology. *Ann Hepatol* 2009; 8: 177-183.

23. Metz CE. ROC methodology in radiological imaging. *Invest Radiol* 1986; 21: 720-33.

24. Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. *Invest Radiol* 1979; 14:109-121.

25. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978; 8: 283-98.

26. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med* 2013; 4(2):627-635

27. Schwartz LM, Woloshin S, Sox HC, Fischhoff B, Welch HG. US women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: cross sectional survey. *BMJ.* 2000*;* 320:1635–1640. Doi: 10.1136/bmj.320.7250.1635.

28. Boone D, Mallett S, Zhu S, Yao GL., Bell N. Ghanouni Alex, *et al.* Patients' & healthcare professionals' values regarding true- & false-positive diagnosis when colorectal cancer screening by CT colonography: discrete choice experiment. *PLoS One* 2013*;* 8:e80767. 10.1371/journal.pone.0080767.

29. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. *Eur Radiol* 2015; 25(4):932-9. 10.1007/s00330-014-3487-0.

30. Mason SJ. & Graham NE. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society 2002; 128*, 2145-2166.