Original Article

Interactive web-based software for evaluating diagnostic tests and roc curve analyses in health sciences

Seyma Yasar, Fatma Hilal Yagin, Ahmet Kadir Arslan, Cemil Colak, Saim Yologlu

Department of Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University, Malatya, Turkey

Abstract

Aim: This study aims to develop web-based and user-friendly DTROC software in which clinicians and researchers can perform two- and three-dimensional receiver operating characteristic (ROC) curve analyses and calculate diagnostic test performance metrics.
Materials and Methods: An illustrative example was presented to implement two- and three-dimensional ROC analyses in medicine. As a case study, three-dimensional ROC analysis in DTROC was applied to the ktemp variable (measurements on the neuropsychometric test for "temporal factor") in the open-access data set named "AL" (Alzheimer's disease neuropsychometric marker dataset). DTROC web-based software was developed using "plotly", "pROC", "dplyr", "shiny", "shinydashboard", "rhandsontable", "shinyBS", "DT", "epiR", "DiagTest3Grp", and "the "shiny" libraries.
Results: The developed web-based DTROC software allows clinicians and researchers to analyze two and three-dimensional ROC analyzes, ROC curve comparisons, optimum cut-off calculations, required sample size estimation for diagnosis, and calculate metrics for two/more than two-class diagnostic tests without any programming language knowledge or theoretical background about the analyzes mentioned above. When the results for the ktemp variable in the AL data set used in practice were examined, the volume under the surface was found to be 0.77. According to the application results in DTROC Three-dimensional ROC analysis, it can be said that the ktemp variable has a distinguishing feature in detecting AL.
Conclusion: All the comprehensive features of DTROC can be accessed free of charge through a graphical user interface that makes the analysis process very easy for users and automates the analysis process. With all these features, DTROC provides much more comprehensive features and various applications than commercial and free software to achieve two/three-dimensional ROC analyses and calculate diagnostic test metrics for two / more classes.

Keywords: Diagnostic tests; multi-class diagnostic tests; ROC analysis<; ROC curve<; three-dimensional ROC analysis

INTRODUCTION

The tests performed by physicians working in the field of health through laboratory techniques, clinical observations, or various measurements to determine whether individuals are healthy are called diagnostic tests. Diagnostic tests can vary in terms of the accuracy of the results, the procedures' cost, and the patient's risk. Although some diagnostic tests give definite results, they may not be preferred due to the application's difficulty or the risks they create on the patient. On the other hand, some diagnostic tests may not be used because the long completion time will delay the patient's intervention. Therefore, instead of applying high-cost and risky tests to diagnose patients with exact or near-exact accuracy to all patients, easy-to-apply, low-cost tests can be developed. The exact test, also called the newly developed test and reference test (gold-standard test), should be applied to a group of patients simultaneously, and the results should be compared with each other. Such a developed test's

success in distinguishing patients and healthy individuals from each other can be evaluated by sensitivity, specificity, and other combined measures obtained by combining these two measures (1,2).

While a single sensitivity and specificity value can be calculated from tests with two results, it is possible to calculate a different sensitivity and specificity value for each value taken by tests that perform a sequential or continuous measurement. Therefore, evaluating the success of tests that measure at the ordinal or continuous measurement level to distinguish the diagnosis is more complex and challenging than tests with two results. ROC (Receiver Operating Characteristics) analysis performed using sensitivity and (1-specificity) criteria are another diagnostic test methods. ROC curve plots the false positive ratio (1-specificity) (on the x-axis) of the test against the correct positive ratio (sensitivity) (on the y-axis) at different cut-off points. The area under the ROC curve is also used as an index showing test success. As the ROC curve is

based on sensitivity and specificity values, this criterion is independent of prevalence. It is also used to find the specificity corresponding to a specific sensitivity value or find the sensitivity value corresponding to a specific specificity value. Its most important use is to measure tests' discrimination success with different criteria obtained from the ROC curve and comparing tests. The most frequently used of these criteria is the area under the ROC curve calculated by various mathematical methods and can be interpreted as the new test's discrimination power (3,4).

ROC analysis is used to evaluate the accuracy of tests when the actual disease state determined by the gold-standard test is two-class (e.g., patient-control), and the diagnostic test result is numerical. However, although the gold standard test results are usually two-class in medicine, more than two diagnostic classes are frequently encountered. For the first time in 1999, Mossman developed the "Three Dimensional ROC Analysis" method to evaluate the performance of the tests developed to classify three-class disease states. In the three-dimensional ROC analysis, since the disease state has three classes, the ROC graph is analyzed in three dimensions. As a result of the analysis, the ROC surface is drawn as the test's performance indicator, and the volume under the ROC surface is calculated. The ROC surface is obtained in three dimensions inside a cube. Each axis in the three-dimensional system shows the three correct classification rates. The higher the ROC surface, the higher the test's discrimination power, and the flatter it is interpreted as useless (4).

There are many commercial [e.g., IBM SPSS Statistics (5), Stata (6), MedCalc (7), etc.] and open-source (e.g., R, SAS, so on) software packages used to assist researchers in the ROC analysis. Some of these software packages provide essential features for ROC analysis, while others provide advanced features. R software includes comprehensive packages for ROC analysis, such as ROCR, pROC, ROC, and OptimalCutpoint (8).

All the above R software packages perform ROC analysis using the available functions of these packages. Although these packages provide comprehensive and advanced features, they require at least basic knowledge of the R programming language. Working with the R software language, which has a command-based interface for clinicians and other healthcare professionals, can be quite tricky and time-consuming when a quick evaluation is required. Shiny, an R package, allows users to create interactive web tools with a well-designed, user-friendly, and easy-to-use interface (9).

This study aims to develop a web application named DTROC to perform ROC analysis for a two and three-dimensional diagnostic test and calculate performance metrics for this diagnostic test. DTROC software is available at http://biostatapps.inonu.edu.tr/DTROC/.

MATERIALS and METHODS

Data Set
The "ktemp" variable, one of the neuropsychometric test measures used for the "temporary factor" in the AL dataset, which is a subset of the longitudinal cohort of Washington University (WU) Alzheimer's Disease Research Center (ADRC) to illustrate the three-dimensional ROC analysis sub-module in the developed web-based software, DTROC (10).

ROC Analysis
ROC Analysis is a statistical evaluation method developed to eliminate the drawbacks of diagnosis using only one sensitivity and specificity value. ROC curve is a curve with a true positive rate (sensitivity) on the vertical axis and a false positive rate (1-specificity) on the horizontal axis for different threshold values. Each point on the ROC curve corresponds to sensitivity and (1-specificity) values corresponding to different threshold values. Threshold values, which generally give low false-positive rates, also have low correct positive rates. The higher the correct positive rate we get the higher the false positive rate. The most useful diagnostic test has a high rate of correct positive and a low rate of false positive. A near-perfect diagnostic test should have a ROC curve that passes almost vertically from (0.0) to (0.1) and then horizontally (1.1). In short, the test that gives the ROC curve closest to the upper left corner is the most useful test. As the ROC curve gets closer to the function Y = X, a failed test occurs. The diagonal line joining the points (0.0) to (1,1) is considered the reference line. A diagnostic test with a ROC curve close to this line is a useless diagnostic test for screening diseases (11,12).

Area under the ROC Curve
A practical method that can be used to determine a diagnostic test's diagnostic competence is the expression of performance with a single value. The most commonly used measurement is the area under the ROC curve. The higher the AUC value, the better the test is in predicting disease. Possible values of AUC range from 0.5 (undiagnosed) to 1.0 (perfectly diagnosable) (12,13). AUC has a physical interpretation. The probability that a person selected from the diseased population is greater than the criterion value of a person selected from a non-disease population (14).

For example, when AUC = 0.80, a randomly selected individual from the diseased group will yield an 80% probability of a higher threshold test result than an individual selected from the disease-free group. It does not indicate that the disease outcome will be seen with a probability of 80% or that the positive result will be associated with the disease with an 80% probability. Even if the threshold value remains the same, when the diseased and healthy populations' values diverge from each other, the ROC curve area increases as the criterion values (group distribution means) of the diseased and healthy populations move away from each other.
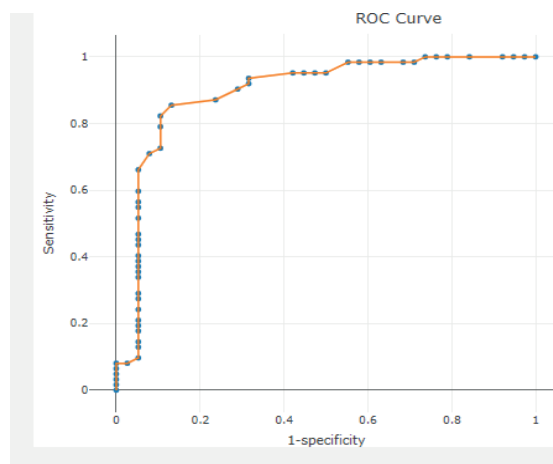
Figure 1. ROC Curve Plot in DTROC

Three Dimensional ROC Analyses
In the decision-making process in the field of medicine, disease states may not always consist of two classes as "present-absent". Sometimes the related disease state can be in three classes. For example; It can be in the form of positive-risk-negative or malignant-suspicious-benign. Clinicians often find it difficult to distinguish between three or more diagnostic options.

For example, when a radiologist examines mammograms, he may need to distinguish between three classes: "malignant", "benign", or "no nodule". The radiologist may want to measure the performance of the test he uses to distinguish these three classes. Three-class disease states are generally transformed into two-class diagnosis results such as "tumor present" - "no tumor" or "tumor present" - "no nodule", and the difficulty of the problems is tried to be reduced. However, changing the diagnostic result may cause the outcomes to diverge from reality or cause biased results (15).
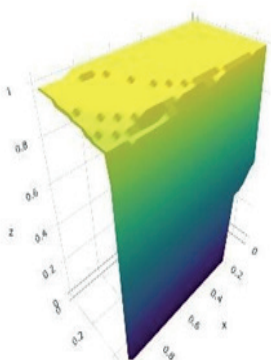

Figure 2. Three-dimensional ROC surface in DTROC

Three-dimensional ROC analysis, developed by Mossman in 1999, is used to measure the diagnostic test's performance to distinguish these three classes if the diagnostic result is three-class. It is a generalized version of two dimensional ROC analyses. In three-dimensional ROC analysis, the ROC surface is drawn as a performance criterion of the diagnostic test, and then the volume under the ROC surface is calculated (Figure 2). As a result of the three-dimensional ROC analysis, the best cut-off can

also be determined to distinguish the three classes. It is a generalized version of two dimensional ROC analyses. In three-dimensional ROC analysis, the ROC surface is drawn as a performance criterion of the diagnostic test, and then the volume under the ROC surface is calculated (16). As a result of the three-dimensional ROC analysis, the best Cut-offs can also be determined to distinguish the three classes.

Diagnostic Tests Performance Criteria
A diagnostic test's performance can be defined by the test's diagnostic adequacy or its capacity to subdivide cases accurately. Some statistical criteria are calculated to evaluate the performance of any tests. Among these criteria, the most frequently used ones appear in the form of cross-table statistics, as they are expressed by the rates of correct positive and false-positive cases (Table 1) (9).

Table 1. 2x2 Classification Matrix

Predict	Actual		Total
	Positive	Negative	
Positive	TP	FP	TP + FP
Negative	FN	TN	FN + TN
Total	TP + FN	FP + TN	n

TP: True positive; Sensitivity = TP/(TP + FN); FP: False positive; Specificity = TN/(FP + TN); TN: True negative ; PPV = TP/(TP + FP); FN: False negative ;   NPV = TN/(TN + FN); NPV: Negative predictive value; PLR = Sensitivity/(1–Specificity); PPV: Positive predictive value ; NLR = (1–Sensitivity)/Specificity ;PLR: Positive likelihood ratio ; PLR: Positive likelihood ratio

For example, the Optimal Cut-off is calculated using one or more performance metrics together. Determining the optimal Cut-off is critical to obtaining accurate results. Extensive methods are available in the literature to determine the optimal Cut-off. Many of these methods are based on sensitivity and specificity measures. Besides, there are different methods for determining the optimal Cut-off based on cost-benefit, prevalence, predictive values, and diagnostic odds ratios. Two popular methods used for this purpose are the Youden index and minimizing the distance of the point on the curve from the upper left corner. The sensitivity and specificity of a classifier are maximized when choosing the optimal Cut-off. For a classifier with excellent discrimination, the sensitivity and specificity criteria will be equal to 1. Therefore, the area under the curve for this classifier will be equal to 1. For detailed information on the optimal Cut-off and the mathematical background, please see the reference (17).

Web-Based Software Developed
To construct DTROC software, the Shiny version 1.0.5 package was used based on the R programming language. DTROC contains separate sections for ROC Analysis, Comparison of ROC Curves, Three-dimensional ROC Analysis, and calculation of Diagnostic Tests. The "Introduction" menu contains basic information about the application. In the "File Upload" menu, the data files can be uploaded with .xls / .xlsx (MS Excel) and .sav (SPSS)

extensions. In this menu, the type (levels of measurement) of variables in the uploaded data set can also be determined. The "Download" menu allows researchers to download all analysis outputs as "PDF", "Word," or "HTML" reports.

DTROC, can be accessed free of charge at http://biostatapps.inonu.edu.tr/DTROC/. This web-based application has been developed using "plotly" (18), "pROC" (19), "dplyr" (20), "shiny" (21), "shinydashboard" (22), "rhandsontable" (23), "shinyBS" (24), "DT" (25), "DiagTest3Grp" (26), and "epiR" (27) R software packages. Tables and Properties in DTROC Interface are located in Table 2 and Table 3.

Table 2. The sections in DTROC Interface and Their Properties	
Modules (Sections)	**Features**
ROC Analysis	**Descriptive Statistics**
	Arithmetic Mean (Standard Deviation)
	Median (IQR)
	Area Under the Curve
	AUC
	95% Confidence Interval
	z-Test
	p-Value
	ROC Curve
	ROC Curve Comment
	Statistics Regarding the Cut-Off
	Youden Index
	Cut-Off
	Sensitivity
	Specificity
	Performance
	Cut-Off
	Sensitivity
	Specificity
	95% Confidence Interval
Comparison of ROC Curves	**ROC Statistics for Variables**
	AUC
	95% Confidence Interval
	Binary Comparisons of ROC Curves
	AUC of Differences
	z-Test
	p-Value
	Statistical Significance
	Multiple ROC Chart
Three Dimensional ROC Analysis	**Descriptive Statistics**
	Arithmetic Mean (Standard Deviation)
	Median (IQR)
	Statistics on Cut-Off
	Cut-Off
	Correct Classification Rate
	Overall Accuracy Ratio
	Statistics Regarding the Volume Under the Surface
	Volume Under the Surface
	95% Confidence Interval
	p-Value
	Statistical Significance
	ROC Surface Plot
	ROC Surface Plot Comment

Table 3. Tabs in DTROC Interface and Their Properties	
Modules (Sections)	**Features**
Diagnostic Tests Criteria	Accuracy
	F1-Score
	MCC
	Prevalence
	Real Prevalence
	Sensitivity
	Specificity
	Diagnostic Accuracy
	Diagnostic Odds Ratio
	Number of People Required for Diagnosis
	Youden Index
	Positive Prediction Value
	Negative Prediction Value
	Positive Likelihood Ratio
	Negative Likelihood Ratio
Multi-Class Diagnostic Tests Criteria	**Performance Metrics**
	Accuracy
	Precision
	Sensitivity
	Specificity
	F1-Score
	MCC
	G-mean
Download	**PDF**
	Word
	HTML

RESULTS

In the DTROC, the information page is opened, shown in Figure 3a as first. Then, the AL data set was uploaded into the DTROC interface via the File Upload tab (Figure 3b). DTROC transfers the data to a panel where it can be viewed, and here the type of variables in the data set can be determined. Three-dimensional ROC analysis is entered into the sub-module (Figure 4a). Here the confidence level was chosen as 95%. The results for the ktemp variable in the AL data set are as follows (Figure 4b). Downloading the results obtained is as in Figure 5a. The screen with the information for reference to the DTROC software used in the studies is given in Figure 5b.
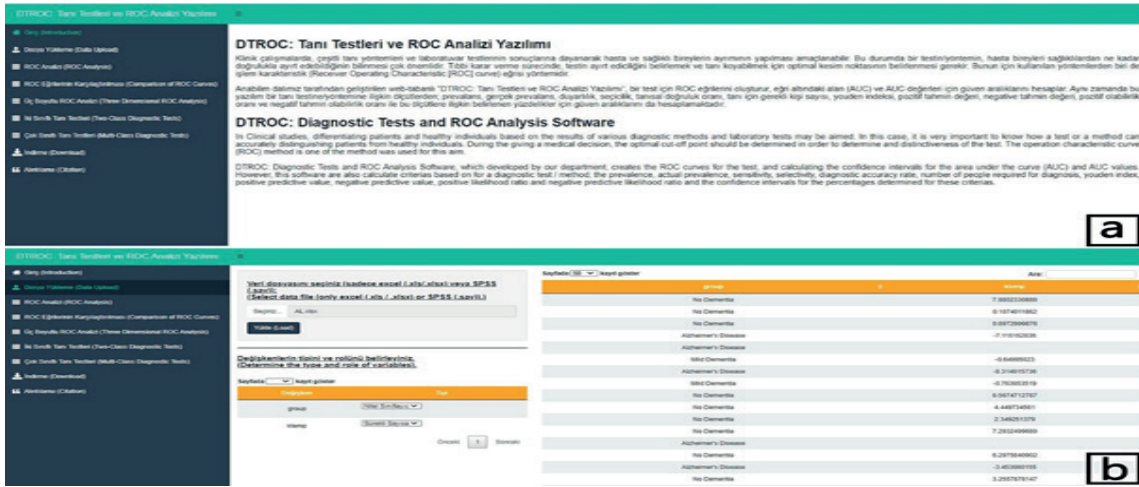
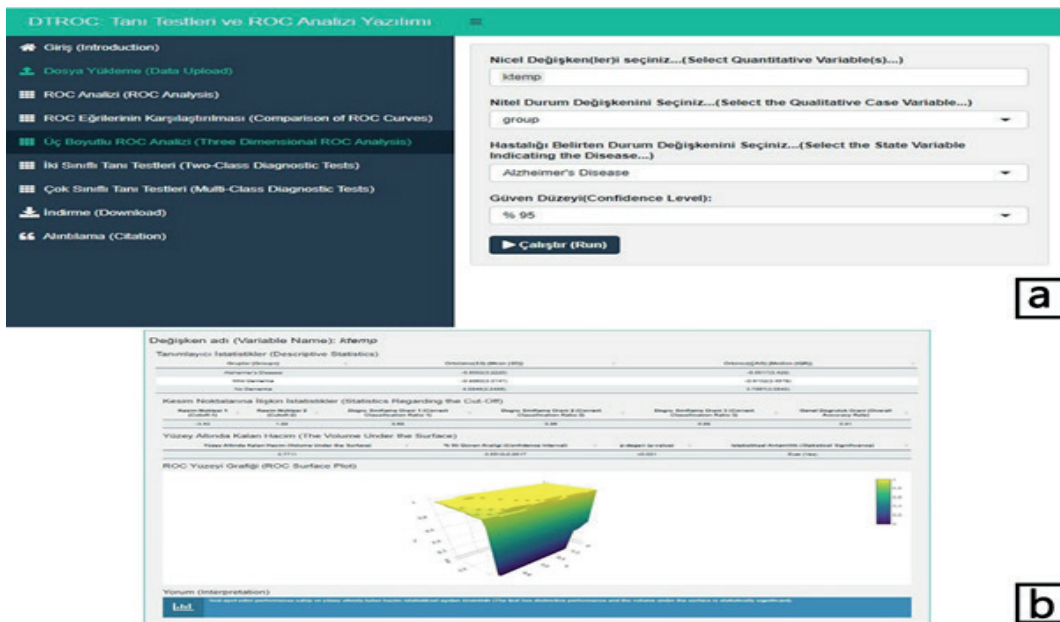**Figure 3.** a) The DTROC Information Page; b) The DTROC File Upload

Figure 4. a) Three-dimensional ROC Analysis Screen; b) Three-dimensional ROC Analysis Results for the ktemp variable
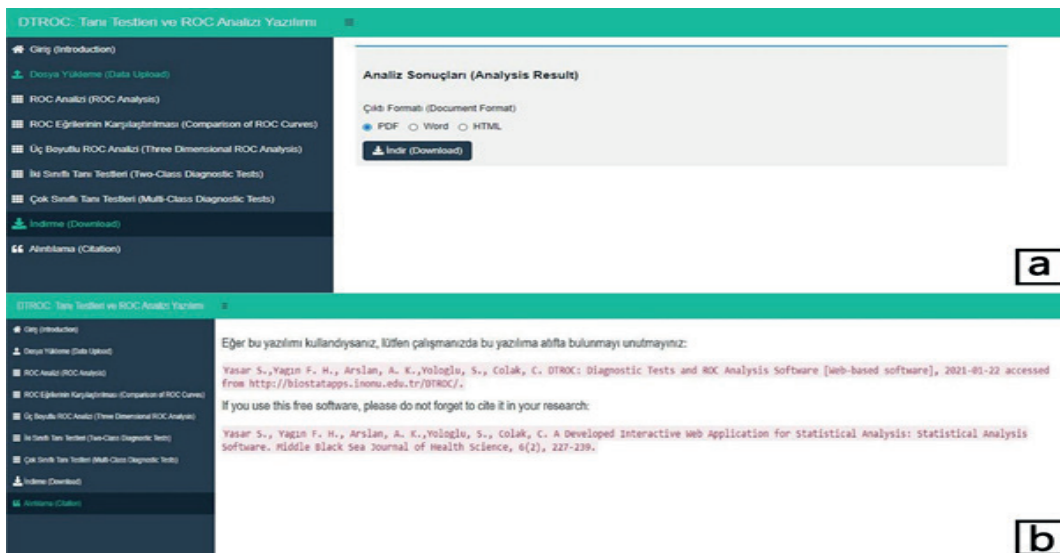

Figure 5. a)The Download Screen; b) The Cite Screen

When the results for the ktemp variable in the AL data set used in practice were examined, the volume under the surface was found to be 0.77. According to the application results in DTROC Three-dimensional ROC Analysis, it can be said that the ktemp variable has a distinguishing feature in detecting AL.

DISCUSSION

In the medical field, researching the effectiveness of diagnostic methods is essential and has become necessary. Distinguishing sick and healthy individuals or examining correct diagnosis methods for patients is always a topic of interest and research. There are diagnostic tests for certain diseases that precisely distinguish between sick and healthy individuals. ROC analysis is a method that is frequently used to evaluate the accuracy of tests when the actual disease state determined by the gold-standard test has two / more classes, and the diagnostic test result is numerical. ROC analysis is commonly implemented as it is one of the most basic statistical analyzes (1,2).

ROC analysis is one of the standard features found in commercial and free software packages such as IBM SPSS (28), MedCalc (7), Stata (6), R (29), etc. Each of these tools has different features for ROC analysis. The commercial software package IBM SPSS calculates some basic statistics such as AUC and standard error, confidence interval, and statistical significance by drawing the ROC curve. However, it does not provide any calculations for Cut-off determination or three-dimensional ROC analysis. Stata offers various ROC analysis calculations, including partial AUC, multiple ROC curve comparisons, determining the optimum Cut-off using the Youden index, and various performance criteria, but does not include any calculations for three-dimensional ROC analysis. Another software package for ROC analysis is MedCalc, which has more comprehensive features than many of the different commercial software packages available. MedCalc (7) provides sample size estimation for a single diagnostic test but has no three-dimensional ROC calculation option.

R is an open-source and free software package that includes all commercial software packages and even more with various packages; ROC, ROCR, pROC, and OptimalCutpoints. The ROC package is an R software package that can generate the ROC curve and calculate AUC. ROCR is a very comprehensive R package that provides more than 25 performance criteria. It allows researchers to create two-dimensional performance curves. Although the ROCR package is one of the comprehensive packages for evaluating performance criteria for diagnostic tests, it offers limited options for selecting the optimum breakpoint. Any of the two-dimensional performance charts can be used to determine the optimum Cut-off graphically. Besides, the ROCR package calculates the AUC and confidence interval but does not provide a statistical test for performance criteria (30). Like the ROCR, the pROC package has the Youden index and the point properties closest to the top left corner to determine the optimum Cut-off. So the pROC package also provides limited features to determine

the optimum breakpoint (31). The OptimalCutpoint package is a comprehensive R package developed to find a diagnostic test / /biomarker's optimum Cut-off. It includes 34 different Cut-off calculation methods based on sensitivity/specificity criteria, cost-benefit analysis, p-values, and diagnostic probability ratios (17). Although these R software packages, especially the pROC package, are quite the right package for ROC analysis, none of these packages have a graphical user interface, requiring advanced coding knowledge. Therefore, performing ROC analysis with these packages can be very difficult and time-consuming, especially for researchers with limited or no R programming knowledge.

The DTROC web application developed in this study combines the features of several ROC packages in the R software, allowing researchers to perform ROC analysis with an easy-to-use interface without the need to write any R code. This application is free and web-based, created with the Shiny package and HTML codes. DTROC includes quite extensive options for performance criteria analysis of two / three-dimensional ROC analysis and two / more than two class diagnostic tests that other software do not have (or share some options). In the DTROC ROC analysis tab, Descriptive Statistics on a variable basis, Statistics on Area Under the Curve (AUC, Confidence Interval, z value, p-value), ROC Curve and interpretation of the generated curve, Cut-off Statistics (Youden Index, Cut-off, Sensitivity, Specificity) calculates Performance metrics according to different Cut-offs. The three-dimensional ROC analysis tab includes Descriptive Statistics, Statistics on Cut-offs, Statistics on Volume Below Surface, and ROC Surface Graph and its interpretation. Clinicians may wish to distinguish between a diagnostic test and each condition of a tri-state disease. In three-state problems, trying to convert three disease states into two disease states is an additional standard solution in such cases. However, this solution causes loss of information and inaccurate prediction of the relevant diagnostic test result.

Three-dimensional ROC analysis, because of the insufficient knowledge of and the limited computer software to perform the analysis, has caused it not to be widely used in studies with three states of the disease. Therefore, three-dimensional ROC analysis in the DTROC web-based application distinguishes DTROC from many other analysis tools. Considering that there are more than two diagnostic classes of many diseases, especially in health, the developed DTROC software is more comprehensive than many software packages. DTROC includes quite comprehensive performance criteria in the diagnostic test criteria section, and these criteria can be calculated for diagnostic tests with two or more than two classes. In this respect, it is clear that DTROC has more options than existing softwares. All the comprehensive features of DTROC can be accessed free of charge through a graphical user interface that makes the analysis process very easy for users. With all these features, DTROC provides much more comprehensive features and various applications for two and three-dimensional ROC analysis compared to commercial and free software.

CONCLUSION

According to the results of the current study, it is thought that the web-based DTROC software developed will provide clinicians and researchers with much more comprehensive features and facilities in ROC and diagnostic tests analysis compared to other software.

REFERENCES

1. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology 2003;229:3-8.
2. Pepe MS. The statistical evaluation of medical tests for classification and prediction: Medicine; 2003.
3. Obuchowski NA. Estimating and comparing diagnostic tests' accuracy when the gold standard is not binary1. Acad Radiol 2005;12:1198-204.
4. Zhou X, Obuchowski N, McClish D. Comparing the accuracy of two diagnostic tests. Statistical Methods in Diagnostic Medicine New York: John Wiley & Sons 2002:165-94.
5. Nie NH, Bent DH, Hull CH. SPSS: Statistical package for the social sciences: McGraw-Hill New York; 1975.
6. StataCorp L. Stata data analysis and statistical Software. Special Edition Release 2007;10:733.
7. Schoonjans F, Zalata A, Depuydt C, et al. MedCalc: a new computer program for medical statistics. Comput Methods Programs Biomed 1995;48:257-62.
8. Team R. R Studio: integrated development for R. RStudio, Inc, Boston, MA URL http://www rstudio com. 2015;42:14.
9. Goksuluk D, Korkmaz S, Zararsiz G, et al. easyROC: an interactive web-tool for ROC curve analysis using R language environment. R J 2016;8:213.
10. Storandt M, Hill RD. Very mild senile dementia of the Alzheimer type: II. Psychometric test performance. Arch Neurol 1989;46:383-6.
11. Søreide K. Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. J Clin Pathol 2009;62:1-5.
12. Goehring C, Perrier A, Morabia A. Spectrum bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance. Stat Med 2004;23:125-35.
13. Obuchowski NA. ROC analysis. AJR Am J Roentgenol 2005;184:364-72.
14. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.
15. Xiong C, van Belle G, Miller JP, et al. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. Stat Med 2006;25:1251-73.
16. Mossman D. Three-way rocs. Medical Decision Making 1999;19:78-89.
17. López Ratón M, Rodríguez Álvarez MX, Cadarso Suárez CM, et al., editors. OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests2014: American Statistical Association.
18. Sievert C, Parmer C, Hocking T, et al. plotly: Create Interactive Web Graphics via 'plotly. js'. R package version 2017;4:110.
19. Robin X, Turck N, Hainard A, et al. Package 'pROC'. 2012-09-10 09: 34, 2020.
20. Wickham H, Francois R, Henry L, et al. dplyr: A grammar of data manipulation. R package version 04. 2015;3.
21. Chang W, Cheng J, Allaire J, et al. Package 'shiny'. See http://citeseerx ist psu edu/viewdoc/download. 2015.
22. Chang W. shinydashboard: create dashboards with 'Shiny'. R package version 0.5. 3. 2016.
23. Owen J. rhandsontable: Interface to the 'Handsontable. js' library. R package version 03. 2018;7.
24. Bailey E. shinyBS: twitter bootstrap components for Shiny. R package version 061. 2015.
25. Xie Y, Cheng J, Tan X. DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.2. 2016.
26. Luo J, Xiong C, Luo MJ. Package 'DiagTest3Grp'. 2012.
27. Stevenson M, Stevenson MM, BiasedUrn I. Package 'epiR'. 2020.
28. Frey F. SPSS (Software). The International Encyclopedia of Communication Research Methods 2017:1-2.
29. Chambers J. Software for data analysis: programming with R: Springer Science & Business Media; 2008.
30. Sing T, Sander O, Beerenwinkel N, et al. ROCR: visualizing classifier performance in R. Bioinformatics 2005;21:3940-1.
31. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics 2011;12:1-8.