Frontline Learning Research
An official journal of EARLI

Frontline Learning Research Vol.5 No. 3 Special Issue (2017) 41 - 52
ISSN 2295-3159

Receiver Operating Characteristic (ROC) Analysis

Elizabeth A. Krupinski

Emory University, USA

Article received 13 April / revised 23 March / accepted 23 March / available online 14 July

Abstract

Visual expertise covers a broad range of types of studies and methodologies. Many studies incorporate some measure(s) of observer performance or how well participants perform on a given task. Receiver Operating Characteristic (ROC) analysis is a method commonly used in signal detection tasks (i.e., those in which the observer must decide whether or not a target is present or absent; or must classify a given target as belonging to one category or another), especially those in the medical imaging literature. This frontline paper will review some of the core theoretical underpinnings of ROC analysis, provide an overview of how to conduct an ROC study, and discuss some of the key variants of ROC analysis and their applications.

Keywords: Receiver Operating Characteristic analysis, ROC, observer performance, visual expertise

Elizabeth A. Krupinski, PhD, Department of Radiology & Imaging Sciences, Emory University 1364 Clifton Rd NE D107 Atlanta, GA30322, United States. Email: ekrupin&emory.edu. DOI: http://dx.doi.org/10.14786/flr.v5i3.250

1. Introduction

Visual expertise can be measured or assessed in a variety of ways, but in many cases it is the behavioral outcome related to visual expertise that is of ultimate concern. For example, in medical imaging (e.g., radiology, pathology, telemedicine) a physician must view an image (e.g., x-ray exam, pathology slide, photograph of a skin lesion) and render a diagnostic decision (e.g., tumor present or absent), prognosis (e.g., malignant vs benign) and/or a recommendation for a treatment plan (e.g., additional exams or surgery). In the military, radar screens need to be monitored for approaching targets (e.g., missiles) and decisions made as to whether they are friend or foe and whether to escalate the finding to the next level of action. There are many other situations where these types of visual detection and/or classification tasks take place in real life, and where investigations take place to assess how expertise impacts these decisions and what we can do to improve them. In medicine this is particularly important as when decision errors are made they have direct and significant impacts on patient care and well-being [1-2].

The problem with “real life” is that is often very difficult to determine how well the interpreter is actually performing. Feedback is often not provided and when it is it is often separate in time from the actual decision making event – often quite disparate in time, making the feedback less impactful. Compounding the problem is that there is often not a single correct answer. Thus, the majority of performance assessments are done in the context of research and are often focused on assessing observer performance in the context of comparing a new technique or technology to an existing one. To deal with the complex nature of decision interpretation tasks where it is important to understand and balance the consequences both correct and incorrect decisions, Receiver Operating Characteristic (ROC) analysis is a very valuable tool.

It is important throughout this review to keep in mind that performance is not a constant. One’s performance on any given task will vary as a function of a number of factors and thus the metrics and principles discussed will reflect those changes and differences. For example, when someone first learns a decision task, their criteria for rendering a decision may be based more on didactic learning that they have engaged in and the exact examples of the task they have encountered to date. As expertise grows and they encounter more and varied examples, the criteria they use in their decisions is likely to change as their knowledge and skill develops. Even when one is an expert, changes in the environment, the stimuli and the consequences of the decisions rendered may change, making it necessary to adjust one’s criteria to the new situation. For example, on a chest x-ray image cancer looks like a white spot(s) on the lung and even residents in training quickly learn to detect and diagnose lung cancer. However, in the Southwest United States and other desert regions there is a condition known as Valley Fever (in infection caused by the fungus Coccidioidomycosis) that appears in the lungs as a white spot(s). Radiologists who move to places like Arizona where Valley Fever is quite common go through a period of criteria adjustment – initially calling nearly everything cancer (high false positive rate) but soon learning to distinguish Valley Fever from lung cancer as they see more cases and learn the discriminating features.

2. Basics of decision making for ROC

ROC analysis was developed in the early 1950s based on principles from signal-detection theory for evaluation of radar operators in the detection of enemy aircraft and missiles [3-4], and additional contributions were thereafter made by researchers in engineering, psychology, and mathematics [5-7]. ROC was introduced into medicine in the 1960s by Lee Lusted [8-11], with significant efforts devoted to gaining a better understanding of decision-making [12-15]. This entrée into medicine was the result of a series of studies in radiology that began soon after World War II to determine which of four radiographic and fluoroscopic techniques (e.g., radiographic film vs slides) was better for tuberculosis (TB) screening [16-17]. The goal was to find a single imaging technique that would outperform all the others (in other words allow the radiologists to reach the correct decision). Instead what they found was that the intra-observer and inter-observer variation was unexpectedly so high that it was impossible to determine which one was best. This was unexpected as until then it was presumed that given the same image data all radiologists looking at the images would be seeing the same diagnostic features and information, detecting the TB if it was present, and rendering the same diagnostic decision. The idea that everyone may “see” something different in the images, perhaps as a function of experience or expertise, had never been considered. Thus, it was necessary to build systems that could generate better images so radiologists’ performance could improve (i.e., reduce observer variability), and develop methods to evaluate these new systems and assess their impact on observer performance.

Although there are newer methods that allow for more than two decisions in the ROC task environment [18-19], ROC is traditionally a binary decision task – target/signal (e.g., lesion, disease, missile) present versus target/signal absent, or in the case of classification rather than detection the target/signal belongs to class 1 (e.g., cancer, enemy) or class 2 (e.g., not cancer, friend). For ROC analysis, these two conditions must be mutually exclusive. There must also be some sort of “truth” or gold standard for each option. In radiology for example, pathology is often used as the gold standard. In cases where there is no other definitive test or method for determining the truth, panels of experts are often used to establish the gold standard [20-21]. Given the truth and the decisions of the observers in the study, a 2x2 table readily summarizes all four possible decisions: true positive (TP) (target present, observer reports as present), false negative (FN) (target present, observer reports as absent), false positive (FP) (target absent, observer reports as present), and true negative (TN) (target absent, observer reports as absent). The TP and TN decisions are correct while the FN and FP decisions are incorrect.

2.1 Common performance metrics

Suppose you have an observer who is an expert skilled at visually detecting a specific poisonous frog in the jungle versus a similar but non-poisonous frog. Knowing that she can make the correct decision is important because she is your guide on an expensive eco-jungle tour and if she says a given cute little frog is not poisonous but in reality is, you might reach out to touch it with potentially fatal consequences. In radiology, one of the common sources of litigation is mammography – mammographers either missing potential breast cancers or overcalling benign conditions as malignant causing undo stress and anxiety. Real life examples of important decision tasks abound, all of which require careful assessment of correct and incorrect decisions.

From the basic 2x2 matrix of 4 decisions described above come some key metrics often used to assess performance in visual expertise and other observer performance studies. The two most commonly used are sensitivity and specificity. Sensitivity reflects the ability of the observer to correctly classify the target present stimuli (e.g., x-ray or other images) and is calculated as:

$$\text{Sensitivity} = TP/(TP + FN) \tag{2.1}$$

Specificity reflects the ability of the observer to correctly classify the target absent stimuli and is calculated as:

$$\text{Specificity} = TN/(TN + FP) \tag{2.2}$$

When you combine these decisions, a measure of accuracy can be obtained as:

$$\text{Accuracy} = (TP + TN)/\ (TP + FN + TN + FP) \tag{2.3}$$

Two other common metrics that are used in medicine are positive and negative predictive value:

$$\text{Positive Predictive Value (PPV)} = TP/(TP + FP) \tag{2.4}$$

$$\text{Negative Predictive Value (NPV)} = TN/(TN + FN) \tag{2.5}$$

In general, there is a trade-off between sensitivity and specificity – as you increase one you decrease the other. In other words, if you want to detect more targets (high sensitivity) it often occurs as a result of making more false positives (decreased specificity). Why would you want to use sensitivity/specificity versus PPV/NPV? Basically, the former are independent of the prevalence of targets in the case sample while the latter are not. An example might be useful. Suppose you have an observer who is an expert skilled at visually detecting a specific poisonous frog in the jungle versus a similar but non-poisonous frog and her sensitivity is 95% and specificity 80%. In jungle #1 there are 1000 frogs total with a prevalence of 50% poisonous (n = 500). In jungle #2 there are also 1000 frogs total but only 25% are poisonous (n = 250).

Based on this, our observer has the following performance levels. It can be seen that depending on which jungle the observer study is conducted, the performance even of an expert observer will differ.

Jungle #1: TP = 475 FN = 25 FP = 100 TN = 400

Accuracy = (475 + 400)/(475 + 25 + 100 + 400) = 0.88 or 88%

PPV = 475/(475 + 100) = 0.83 or 83%

NPV = 400/(400 + 25) = 0.94 or 94%

Jungle #2: TP = 238 FN = 12 FP = 150 TN = 600

Accuracy = (238 + 600)/(238 + 12 + 150 + 600) = 0.84 or 84%

PPV = 238/(238 + 150) = 0.61 or 61%

NPV = 600/(600 + 12) = 0.98 or 98%

In many cases sensitivity and specificity are more than adequate measures of performance for visual search tasks, but it becomes complicated when the test sets contain cases with a range of difficulty levels. For example, in radiology some bone fractures are very obvious and thus easy to detect but others are quite subtle and can readily be missed. In the frog example, a bright red frog in a green jungle is likely easier to detect than a light green frog in a dark green jungle. In cases that are not obvious the decision as to whether or not the target is actually there becomes less certain and observers may not be willing to give a binary yes/no present/absent decision, but may be more willing to report their decision as a function of confidence, for example reporting a target (or lack of target) as definitely present, probably present or possibly present. In other words, the observer’s decision threshold for reporting can change as a function of many things, including but not limited to the nature of the target, target prevalence, background complexity within which the target is embedded, number and type of targets, and observer experience or expertise. This is where ROC analysis becomes useful.

2.2 The ROC curve

Even the visual expert may not perform as one would think without delving further into the nature of the task. For example, in radiology decision thresholds can change within and between observers as a function of the nature of the task and its consequences. In chest CT exam interpretation a radiologist may adopt a very conservative threshold for reporting possible abnormalities in the lungs that could be cancer nodules but are less than 5 mm in size because they know that CT is typically the best imaging exam to do (i.e., no other follow-up imaging options) and obtaining a biopsy on such a small target is very difficult (potentially leading to a pneumothorax or puncture of the lung) and unlikely to yield a specimen large enough to get an accurate biopsy on. Instead of reporting it, the radiologist may recommend a 6-month watch period and another CT exam. In mammography however, small lesions are easier to biopsy and there is no risk of puncturing a lung or other vital structure, so mammographers tend to be more liberal in their reporting of potential cancerous findings. This comes at the risk of more false positives but before doing a biopsy additional x-ray images or an ultrasound is often recommended, reducing the impact of a false positive even further.

ROC analysis and the resulting ROC curve is a method that captures the relationship (i.e., trade-offs) between sensitivity and specificity as well as the range of decision thresholds that every observer has no matter what their level of expertise and experience. The ROC curve (Figure 1) is a graphical representation of this relationship, plotting sensitivity (the TP fraction) versus 1-specificity (the FP fraction or 1 – TN/(TN + FP) = FP/(FP + TN)) for all possible decision thresholds. The axes go from 0 to 1 since sensitivity and specificity are typically represented as proportions. The diagonal line is chance performance or guessing and the curves indicate increasingly better performance moving to the upper left corner which represents perfect performance. Thus in terms of visual expertise, one would expect that those with more expertise would tend to have curves closer to the upper left corner.

The fact that those with more expertise tend to have curves closer to the upper left corner is true, but as noted above much depends on the decision thresholds that the individual observer has for a given task. Each point on a given curve represents a specific TP and FP fraction or decision threshold setting. For example, on the "good" curve, the plus sign represents a rather conservative decision maker – not reporting a lot of targets (low sensitivity) but also not making a lot of false positives (high specificity or low 1-specificity). The cross sign represents a more liberal decision maker – reporting a lot of targets (high sensitivity) but making a lot of false positives (low specificity or high 1-specificity). Referring back to the example of the radiologist moving to Arizona and overcalling Valley Fever as cancer, the Figure could just as well represent his/her learning or adjustment period, in which the lower curve is when they first move and overall cancers and the upper curve is their performance after a few months of seeing more exemplars of Valley Fever so better discrimination is possible.

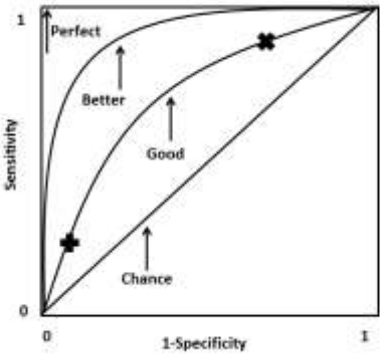


Figure 1. Example of a typical ROC curve plotting sensitivity vs 1-specificity. The diagonal line is chance performance or guessing and the curves indicate increasingly better performance moving to the upper left corner which represents perfect performance.

2.3 How to plot an ROC curve

It is rare that someone actually plots an ROC curve by hand as there are a number of software programs available both in commercial statistical packages and freely downloadable from research web sites. However, it is useful to see how the confidence ratings discussed above lead to the generation of an ROC curve. Thus, suppose you ran a visual detection experiment and the observer was required to identify whether or not a given image had a target (n = 50) or not (n = 50) and then report confidence in that decision as definite, probable or possible. This yields 6 categories of responses where 1 = absent, definite and 6 = present, definite. You can then create a table (Table 1) showing the distribution of responses as a function of truth (whether or not the image actually contains a target). It should be noted that continuous rating scales (e.g., 0 – 100) can be used as well and methods exist for generating operating points (decision thresholds) and plotting these as well [22].

Table 1

An example of the distribution of confidence scores for a subject in an observer performance study with a 6-point confidence scale and images with a target present or absent (truth).

Truth	**1**	**2**	**3**	**4**	**5**	**6**
Present	2	3	2	5	20	18
Absent	16	15	10	4	3	2

The sensitivity, specificity and FP fraction can then be determined at each threshold or cutoff point as in Table 2.

Table 2

Sensitivity, specificity and FP fraction can then be determined at each threshold or cutoff point.

Result positive is ≥	**Sensitivity**	**Specificity**	**FP fraction**
2 - probably absent	0.96 (48/50)	0.32 (16/50)	0.68
3 - possibly absent	0.90 (45/50)	0.62 (31/50)	0.38
4 - possibly present	0.86 (43/50)	0.82 (41/50)	0.18
5 - probably present	0.76 (38/50)	0.90 (45/50)	0.10
6 - definitely present	0.36 (18/50)	0.96 (48/50)	0.04

The ROC plot can then be generated plotting Sensitivity on the y-axis and 1-Specificity (FP fraction) on the x-axis.

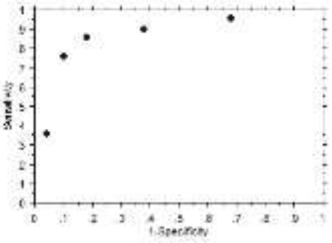


Figure 2. ROC curve generated from the data in Table 2.

In terms of fitting an actual curve to the data points in the ROC plot, there are a variety of methods. Simply "connecting the dots" is the empirically based version but it creates a stepped or jagged plot. A smooth curve reflecting the theoretical "true" curve is much more desirable. There are basically two ways to approach generating the curve – parametric and non-parametric [22-28]. The non-parametric approach does not have any assumptions about the structure of the underlying data distribution and essentially smooths the histograms of the output data for the two classes. The parametric methods do rely on the validity of the underlying distribution assumptions. Most researchers prefer the parametric approaches and much of the available software uses these approaches as well.

2.4 Interpreting the ROC curve

There are a few key metrics used to interpret the ROC curve and characterize observer performance. The most common one is the area under the curve (AUC or Az). As noted above the diagonal line in the ROC plot represents chance or guessing performance and it clearly divides the ROC space into two halves thus representing an AUC of 0.5. The top left corner is perfect performance and encompasses all of the area below it, thus AUC is 1.0. Any curve lying between chance and perfect performance will have a value between 0 and 1.0, with better performance having values closer to 1.0. As with generation of the curve itself, there are a variety of methods to calculate AUC [22-29] and most programs use one of these methods.

Partial AUC acknowledges that the more traditional AUC is often not appropriate, as not all decision thresholds or operating points are equally important [30-31]. In other words, in real life observers may not actually operate at certain thresholds for one reason or another. In medicine for example, a diagnostic test with low specificity (a high false positive rate) may not be clinically acceptable. In this case it may be useful to select a given (acceptable) FP rate, determine its associated sensitivity (TP rate), and then calculate the area under the curve only up to that operating point (i.e., capturing only a part of the total AUC). Partial AUC is very common in the development of computer-aided detection and discrimination algorithms for medical imaging. Other metrics used less often are d', d_e', Δm, B and Z_k [32-33].

2.5 Comparing ROC curves

Although a single ROC curve is common, quite often studies are designed to compare performance, for example between experts and novices on a given visual task. Thus there are two curves – one for experts and one for novices – and the question is whether there is a significant difference in performance (AUC typically) or not. Visually it is not always possible to tell if the difference is significant. This is especially true when the ROC curves cross at some point (usually the upper right quadrant/corner) as in Figure 3 [34-35]. Therefore, statistical methods have been developed, some for comparing only two curves and some for comparing multiple curves. Again, there are parametric and non-parametric options [22-23, 36-40]. One of the most common methods for comparing multiple observers and multiple cases is the Multi-Reader Multi-Case method developed by Dorfman, Berbaum and Metz [39].

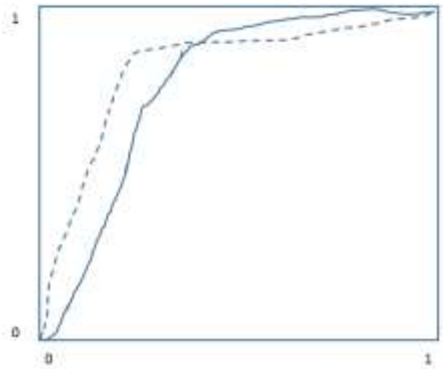
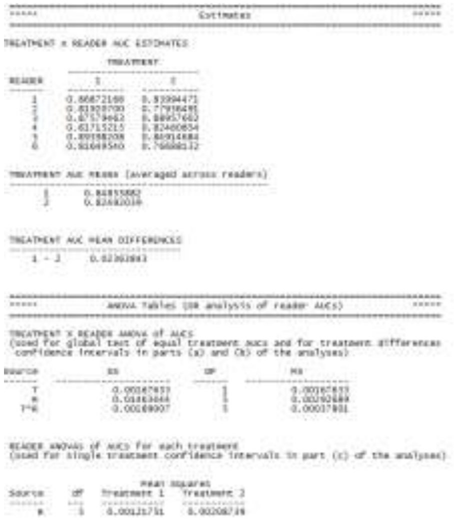


Figure 3. Example of 2 ROC curves that cross.

Figure 4 is an example of the output from an MRMC analysis on a study that had 6 observers viewing a series of images in 2 conditions (different computer monitors for displaying medical images). The visual task was to search for subtle fractures in bone x-ray images. Readers 1, 3 and 5 were expert radiologists and 2, 4 and 6 were resident trainees. The upper portion provides the AUC values for each observer in each condition, followed by the difference between the two conditions. AUC values are usually reported in publications out to three decimal points maximum. The lower portion shows the results of the Analysis of Variance (ANOVA) comparing the AUCs in the standard ANOVA output format. The actual output document provides more information than provided here, such as the variance components and error-covariance estimates, and different ways of treating the various variables (e.g., random readers and random cases, fixed readers and random cases, random readers and fixed cases), but this example shows how many available programs output relevant data comparing ROC curves. The analyses can also take into account level of expertise by comparing the two groups of observers as described above (in this case there were differences but it did not reach statistical significance).



```
*****                          Estimates                              *****

TREATMENT x READER AUC ESTIMATES

                 TREATMENT
             -------------------------
READER           1            2
------       ----------   ----------
    1        0.86872166   0.83994471
    2        0.81920700   0.77916491
    3        0.87579462   0.88957682
    4        0.81715215   0.82460654
    5        0.89198208   0.86914684
    6        0.81649540   0.79888132

TREATMENT AUC MEANS (averaged across readers)
----------------------------------------------
       1        0.84855882
       2        0.82482036

TREATMENT AUC MEAN DIFFERENCES
------------------------------
    1 - 2       0.02363883

***************************************************************************
*****               ANOVA Tables (OR analysis of reader AUCs)         *****
***************************************************************************

TREATMENT X READER ANOVA of AUCs
(Used for global test of equal treatment AUCs and for treatment differences
 confidence intervals in parts (a) and (b) of the analyses)

Source            SS               DF               MS
------    -------------------    ------    -------------------
     T             0.00167833         1             0.00167833
     R             0.01463444         5             0.00292689
   T*R             0.00189007         5             0.00037801

READER ANOVAs of AUCs for each treatment
(Used for single treatment confidence intervals in part (c) of the analyses)

                          Mean Squares
Source      DF    Treatment 1   Treatment 2
------     ---    -----------   -----------
     R       5     0.00121751    0.00208739
```

Figure 4. An Example of the output from an MRMC analysis.

3. Other types of ROC

The discussion of ROC analysis up to this point has been about tasks that involve the detection of one target and for the most part assess FPs only as they occur in target absent images (again 1 per). In real life however, scenes and other visual stimuli often contain multiple targets and FPs can occur in both target present and target absent stimuli. Traditional ROC analysis also typically does not ask or require the observer to locate the target once it has been detected. Thus there is always some question (unless the target appears in a specific location (e.g., center of the display) every time) as to whether the observer actually detected the true target (TP) or called something else in the image (FP) thereby actually missing the true target (FN).

The earliest attempt to account for location in ROC tasks was the LROC (Location ROC). In this method, the observer provides a confidence rating that somewhere in the image there is a target, then marks the location of the most suspicious region [41-42]. LROC is an advantage over ROC in that it takes location into account, but it still only allows for a single target. To account for multiple targets, Free Response ROC (FROC) was developed in which observers mark different locations and provide a confidence rating for each mark [43-46]. The problem with FROC is that when the ROC is plotted the x-axis, rather than going from 0 to 1 like the traditional RFOC curve, goes from 0 to infinity (based on however many FPs are reported). This makes calculating the area under curve quite difficult and the comparison of two curves even more challenging. The Alternative FROC (AFROC) method was developed to address issue, creating a plot that has both axes going from 0 to 1 [47]. The Jackknife AFROC (JAFROC) method was then developed to allow for generalization to the population of readers and cases, in the same way that MRMC ROC does [44-46].

4. Other Considerations

In addition to deciding which type of ROC analysis is best (which really depends on the hypotheses, nature of the task, types and number of images and targets), there are two other aspects that are typically important. As already discussed above, the truth or gold standard for cases must be known in advance. With simple psychophysical studies this is quite easy but for real life images (scenes, medical images, industrial images, satellite images) this can be more difficult. Other considerations when selecting cases include: how subtle or obvious the targets are, how much and what type of background "noise" is in the image, where the targets are located (random or in specified locations), the size(s) of the targets and how much background image is included, how long the images will remain available for viewing, whether the images can be manipulated (e.g., zoom/pan or window/level) by the observers, and target prevalence as noted above.

Sample size is the other key issue with respect to setting up an ROC study – how many images and observers are required to achieve adequate power once the study is completed. As with any other power calculation, sample size will depend on a number of factors including the metric under consideration (e.g., AUC or partial AUC) and the design (e.g., repeated measures with the same observers viewing the same images in two or more conditions or different readers viewing the images in different conditions). There are a number of key papers describing methods to calculate sample sizes for various study designs, many of which include representative tables showing sample sizes required for different power estimates [48-51]. Some of the available ROC programs also include a power calculator and some will provide power in the analysis output.

5. Software Programs

It is not possible to list all of the available software programs for ROC analysis as there are always new ones being released. There are however some reliable sites where the more commonly used programs can be found. The Medical Image Perception Laboratory web site [52] at the University of Iowa and the ROC Software site at the University of Chicago [53] contain the programs developed by that team (Dorfman, Berbaum, Metz, Hillis) including the MRMC ROC, ROCFIT, LABROC4, CORROC, INDROC, ROCKET, LABMRMC, PROPROC, RSCORE, BIGAMMA, RSCORE-J and SAS programs to perform sample size estimates. Software for FROC, AFROC and JAFROC (Chakraborty) analyses are available as well [54]. Some commercial statistical software also has modules for ROC analysis [55-59].

Keypoints

- ROC analysis provides metrics of observer performance in visual detection and discrimination tasks
- Area Under the Curve (AUC) is the most commonly used metric of performance
- Common variants of ROC analysis allow for multiple targets and location accuracy
- Key study design issues include target characteristics and establishing a gold standard
- Software is available to conduct ROC analyses

References

Analyse-It. http://analyse-it.com/docs/220/method_evaluation/roc_curve_plot.htm Last accessed April 13, 2016.

Birkelo, C.C., Chamberlain, W.E., Phelps, P.S. (1947). Tuberculosis case finding. A comparison of the effectiveness of various roentgenographic and photofluorographic methods. *JAMA, 133*, (6), 359-366. PMID: 20281873

Bunch, P.C., Hamilton, J.F., Sanderson, G.K. Simmons, A.H. (1978). A free-response approach to the measurement and characterization of radiographic-observer performance. *J Appl Photogr Eng*, *4*,166–171.

Chakraborty, D.P., Berbaum, K.S. (2004). Observer studies involving detection and localization: modeling, analysis, and validation. *Med Phys*, *31* (8), 2313-2330. DOI: 10.1118/1.1769352

Chakraborty, D.P. (2005). Recent advances in observer performance methodology: jackknife free-response ROC (JAFROC). *Rad Protect Dosim, 114* (1), 26-31. DOI: 10.1093/rpd/nch512

Chakraborty, D.P. (2006). Analysis of location specific observer performance data: validated extensions of the jackknife free-response (JAFROC) method. *Acad Radiol, 13* (10), 1187-1193. DOI: 10.1016/j.acra.2006.06.016

Chakraborty, D.P., Winter, L.H.L. (1990). Free-response methodology: alternate analysis and the new observer-performance experiment. *Radiol, 174* (3), 873-881. DOI: 10.1148/radiology.174.3.2305073

Delong, E.R., Delong, D.M., Clarke-Pearson, D.L. (19880. Comparing the areas under two or more correlated receiver operating characteristics curves: a non-parametric approach. *Biometrics, 44* (3), 837-845. PMID: 3203132

Dev Chakraborty's FROC Web Site. http://perception.radiology.uiowa.edu/Software/ReceiverOperatingCharacteristicROC/tabid/120/Default.aspx Last accessed April 13, 2016.

Dorfmam, D.D., Alf, E. (1968). Maximum likelihood estimation of parameters of signal detection theory: a direct solution. *Psychometrika, 33* (1), 117-124. DOI: 10.1007/BF02289677

Dorfman, D.D., Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals – rating method data. *J Math Psychol*, *6* (3), 487-496. DOI: http://dx.doi.org/10.1016/0022-2496(69)90019-4

Dorfman, D.D., Berbaum, K.S., Metz, C.E. (1992). Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. *Invest Radiol*, *27*(9), 723-731. PMID: 1399456

Dorfman, D.D., Berbaum, K.S., Metz. C.E., Lenth, R.V., Hanley, J.A., Dagga, H.A. (1997). Proper receiver operating characteristic analysis: the bigamma model. *Acad Radiol, 4* (2), 138-149. PMID: 9061087

Edwards, D.C. (2013). Validation of Monte Carlo estimates of three-class ideal observer operating points for normal data. *Acad Radiol, 20* (7), 908-914. DOI: 10.1016/j.acra.2013.04.002

Egan, J.P. (1975). *Signal detection theory and ROC analysis*. New York, NY: Academic Press.

Faraggi, D., Reiser, B. (2002). Estimation of the area under the ROC curve. *Stats Med, 21* (20), 3093-3106. DOI: 10.1002/sim.1228

Garland, L.H. (1949). On the scientific evaluation of diagnostic procedures. *Radiol, 52*, (3), 309-328. DOI: http://dx.doi.org/10.1148/52.3.309

Green, D.M., Swets, J.A. (1974). *Signal Detection Theory and Psychophysics*. Huntington, NY: Krieger Publishers.

Hajian-Tilaki, K.O., Hanley, J.A., Joseph, L., Collet, J.P. (1997). A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Med Decis Making, 17* (1), 94-102. DOI:10.1177/0272989X9701700111

Hanley, J.A. (1988). The robustness of the "binormal" assumptions used in fitting ROC curves. *Med Decis Making, 8* (3), 197-203. DOI: 10.1177/0272989X8800800308

Hanley, J.A., McNeil, B.J. (1983). A method for comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiol, 148* (3), 839-843. DOI: 10.1148/radiology.148.3.6878708

Hanley, J.A., McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiol, 143* (1), 29-36. DOI: http://dx.doi.org/10.1148/radiology.143.1.7063747

Institute of Medicine. (1999). *To Err is Human: Building a Safer Health Care System*. Washington, DC: National Academy Press.

Jiang, Y., Metz, C.E., Nishikawa, R.M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiol, 201* (3), 745-750. DOI: 10.1148/radiology.201.3.8939225

Kundel, H.L., Polansky, M. (1997). Mixture distribution and receiver operating characteristic analysis of bedside chest imaging with screen-film and computed radiography. *Acad Radiol, 4* (1), 1-7. PMID: 904086.

Lusted, L.B. (1960). Logical analysis in roentgen diagnosis. *Radiol, 74*,178-193. DOI: http://dx.doi.org/10.1148/74.2.178

Lusted, L.B. (1968). *Introduction to Medical Decision Making*. Springfield, IL: Charles C. Thomas Publishers.

Lusted, L.B. (1969). Perception of the Roentgen image: Applications of signal detection theory. *Rad Clin N Am, 7*, 435-459.

Lusted, L.B. (1971). Signal detectability and medical decision making. *Science, 171*, (3977), 1217-1219. DOI: 10.1126/science.171.3977.1217

McClish, D.K. (1989). Analyzing a portion of the ROC curve. *Med Decis Making, 9* (3), 190-195. DOI: 10.1177/0272989X8900900307

McNeil, B.J., Adelstein, S.J. (1976). Determining the value of diagnostic and screening tests. *J Nuc Med, 17*, (6), 439-448. PMID:1262961

McNeil, B.J., Hanley, J.A. (1984). Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Med Dec Making, 4*, (2), 137-150. DOI:10.1177/0272989X8400400203

McNeil, B.J., Keeler, E., Adelstein, S.J. (1975). Primer on certain elements of medical decision making. *NE J Med, 293*, (5), 211-215. DOI: 10.1056/NEJM197507312930501

Metz, C.E., Herman, B.A., Shen, J.H. (1998). Maximum-likelihood estimation of ROC curves from continuously-distributed data. *Stats Med, 17* (9), 1033-1053. PMID: 9612889

Metz, C.E., Kronman, H.B. (1980). Statistical significance tests for binormal ROC curves. *J Math Psych, 22* (3), 218-243. DOI: http://dx.doi.org/10.1016/0022-2496(80)90020-6

Metz, C.E., Pan, X. (1999). "Proper" binormal ROC curves: theory and maximum-likelihood estimation. *J Math Psych, 43* (1), 1-33. DOI: 10.1006/jmps.1998.1218

Nakas, C.T. (2014). Developments in ROC surface analysis and assessment of diagnostic markers in three-class classification problems. *REVSTAT – Stat J, 12* (1), 43-65.

NCSS Statistical Software. http://www.ncss.com/software/ncss/procedures/ Last accessed April 13, 2016.

Obuchowski, N.A. (1997). Testing for equivalence of diagnostic tests. *Am J Roentgen, 168* (1), 13-17. DOI: 10.2214/ajr.168.1.8976911

Obuchowski, N.A. (1994). Computing sample size for receiver operating characteristic studies. *Invest Radiol, 29* (2), 238-243. DOI:10.2214/ajr.175.3.1750603

Obuchowski, N.A. (2000). Sample size tables for receiver operating characteristic studies. *Am J Roentgen, 175* (3), 603-608. DOI:10.2214/ajr.175.3.1750603

Obuchowski, N.A. (2004). How many observers care needed in clinical studies of medical imaging? *Am J Roentgen, 182* (4), 867-869. DOI: 10.2214/ajr.182.4.1820867

Peterson, W.W., Birdsall, T.L., Fox, W.C. (1954). The theory of signal detectability. *IRE Prof Gp In Theory Trans PGIT*, 4, (4), 171-212. DOI: 10.1109/TIT.1954.1057460

Petrick, N., Gallas, B.D., Samuelson, F.W., Wagner, R.F., Myers, K.J. (2005). Influence of panel size and expert skill on truth panel performance when combining expert ratings. *Proc SPIE Med Imag, 5749*, 596286. DOI: 10.1117/12.596286

Schulman, K.A., Kim, J.J. (2000). Medical errors: how the US government is addressing the problem. *Curr Control Trials Cardiovasc Med, 1*(1), 35-37. DOI: 10.1186/cvm-1-1-035

SPSS Statistics. http://www-03.ibm.com/software/products/en/spss-statistics Last accessed April 13, 2016.

Starr, S.J., Metz, C.E., Lusted, L.B., Goodenough, D.J. (1975). Visual detection and localization of radiographic images. *Radiol, 116* (3), 533-538. DOI:10.1148/116.3.533

STATA Data Analysis and Statistical Software. http://www.stata.com/features/overview/receiver-operating-characteristic/ Last accessed April 13, 2016

Swensson, R.G. (1996). Unified measurement of observer performance in detecting and localizing target objects on images. *Med Phys*, *23* (10), 1709-1725. DOI:10.1118/1.597758

Swets, J.A. (1979). ROC analysis applied to the evaluation of medical imaging techniques. *Radiol, 14* (2), 109-121. PMID: 478799

Swets, J.A., Dawes, R.M., Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psych Sci Public Interest, 1* (1), 1-26. DOI: 10.1111/1529-1006.001

Swets, J.A., Pickett, R.M. (1982). *Evaluation of Diagnostic Systems. Methods from signal detection theory*. New York, NY: Academic Press.

Tanner, W.P., Swets, J.A. (1954). A decision-making theory of visual detection. *Psych Rev*, *61*, (6), 401-409. PMID: 13215690

University of Iowa Medical Image Perception ROC Software. http://perception.radiology.uiowa.edu/Software/ReceiverOperatingCharacteristicROC/tabid/120/Default.aspx Last accessed April 13, 2016.

University of Chicago ROC Software. http://metz-roc.uchicago.edu/ Last accessed April 13, 2016.

Wald, A. (1950). *Statistical Decision Functions*. New York, NY: Wiley, Inc.

Zou, K.H., Hall, W.J., Shapiro, D.E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Stats Med*, *16* (19), 2143-2156. PMID: 9330425

Zhou, X.H., Obuchowski, N.A., McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine*. New York, NY: Wiley.

MedCalc Statistical Software. https://www.medcalc.org/manual/roc-curves.php Last accessed April 13, 2016.