MAHADEVAPPA MAHESH, MS, PhD
JAMES M. HEVEZI, PhD

ROC, LROC, FROC, AFROC: An Alphabet Soup

Xin He, PhD, Eric Frey, PhD

Receiver operating characteristic (ROC) analysis is a well-established method for assessing binary classification task performance. It was first used by the US Army to increase the correct detection of aircraft from radar signals during World War II. Later, ROC was applied to medicine for evaluating diagnostic testing performance and has since been extensively applied in various areas of medical research.

RECEIVER OPERATING CHARACTERISTIC ANALYSIS

The Role of ROC Analysis in Task-Based Image Quality Assessment

In medical imaging, we often need to choose among different image acquisition or processing protocols on the basis of the resulting image quality. Thus, it is crucial to first quantify image quality.

Many figures of merit (FOMs) have been developed for quantifying image quality. Traditionally, FOMs such as uniformity, resolution, contrast, contrast/noise ratio (CNR), and so on, have been used. However, these physical metrics do not directly describe how the images serve with regards to their intended use.

Currently, a strong consensus is emerging in the medical imaging community on the necessity of using task-based image quality assessment [1], where image quality is quantified based on the performance of an observer on a clinically relevant task.

Clinical diagnostic tasks can be divided into two general categories: estimation tasks and classification tasks. Estimation of tumor volume is an example of an estimation task and a tumor detection task in breast cancer mammography is an example of a classification task, in which an observer (eg, a radiologist) classifies suspicious regions as cancerous or normal tissue. Because there are only two possible decisions (cancerous or normal), this task is often mathematically referred to as a binary diagnostic task, or a binary classification task.

Receiver operating characteristic analysis is a statistical tool for quantifying, visualizing, analyzing, and comparing the performance of binary diagnostic tasks.

The ROC Curve

In an ROC study, an observer (eg, a radiologist) assigns a rating for each image that represents the observer's confidence level that there is an abnormality in the patient [2]. Any number of responses may be used to rate this confidence level. For example, in many clinical studies, a set of 5 confidence level responses are used, with 1 indicating absolute certainty that a finding is normal and 5 indicating absolute certainty that a finding is abnormal. The rating values are pooled together to form rating value distributions of the two classes of patients. An example is shown in Figure 1.

Moving the threshold in Figure 1 and computing the true-positive fraction (TPF) and false-positive fraction (FPF) at each threshold location, an ROC curve can be traced. Examples of ROC curves are shown in Figure 2.

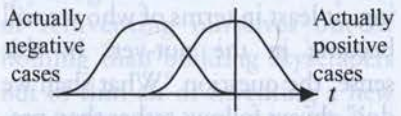

Fig 1. An example of rating value distributions.

Interpreting the ROC Curve

Figure 2 shows two ROC curves. The bold curve is above the other curve at every point, suggesting that the technique producing the bold curve is superior to the technique that produced the other curve. In particular, a system with a higher ROC curve suggests that 1) for any given set of costs or benefits for the various correct and incorrect decisions and prevalence of the two diagnostic classes, this system always results in better diagnosis in terms of expected utility 2) this system always has higher sensitivity for any given specificity; and 3) this system always results in a greater number of correct decisions, regardless of disease prevalence [3]. Thus, the higher the ROC curve, the better the system. Consequently, the area under the curve (AUC) is often used as a FOM for task perfor-

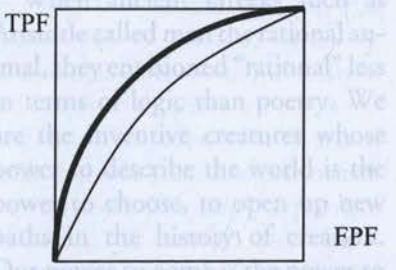

Fig 2. Comparing receiver operating characteristic curves. FPF = false-positive fraction; TPF = true-positive fraction.

0091-2182/09/$36.00 • DOI 10.1016/j.jacr.2009.06.001

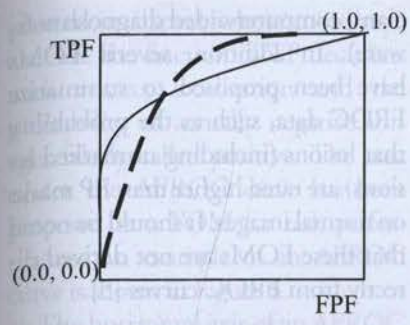

Fig 3. Comparing two crossing receiver operating characteristic curves. FPF = false-positive fraction; TPF = true-positive fraction.

mance when ROC curves do not cross.

When ROC curves cross, the AUC value only characterizes the average task performance. In this case, ranking systems requires considering the part of the curve that is relevant to the diagnostic task. Figure 3 shows such an example. The system producing the dashed ROC curve is more specific when high sensitivities are required and that the system producing the solid ROC curve is more sensitive when high specificities are desirable. So the former system would be better for a first line screening test and the latter would be better in cases where the costs of making a false positive decision were very high.

EXTENSIONS OF RECEIVER OPERATING CHARACTERISTIC ANALYSIS

In a classic ROC study, an observer gives one rating for each image representing the observer's confidence that the image belongs to one of the classes. This experimental paradigm might be problematic when applied to cases in which 1) the location of the abnormality is unknown to the observer; 2) multiple abnormalities are present in the same image; or 3) more than two diagnostic alternatives are involved. Below, we describe extensions to ROC methods that have been proposed to address these limitations of the classic ROC paradigm.

Localization ROC

Consider the case of myocardial perfusion single photon-emission computed tomographic imaging, in which the task is to detect perfusion defects. An observer may specify an FP detection on the anterior wall while missing a true defect on the septal wall. Thus, this image will be reported as positive, a correct decision, but for the wrong reason: the defect is on the septal wall instead of on the anterior wall. The clinical consequence of such a mistake could be serious. To correct for this kind of mistakes, an ROC paradigm that can assess an observer's ability to locate an abnormality in addition to providing a confidence rating is desired [4].

Localization ROC (LROC) analysis is used to account for the localization and detection task. In a typical LROC experiment, an observer is required to mark the location of a suspicious region and then provide the confidence level of defect presence. The mark provided by the observer is considered a correct localization if it is "close enough" to the true location. Presently, the definition of closeness is rather ad hoc.

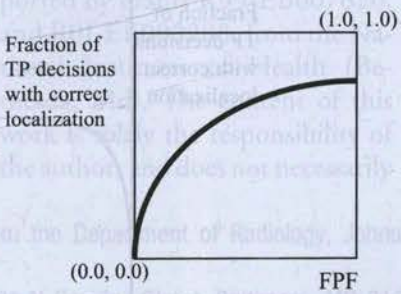

Fig 4. An example of a localization receiver operating characteristic curve. FPF = false-positive fraction; TP = true positive.

An LROC curve is plotted in a space whose coordinate axes are the fraction of TP decisions with correct localization vs FP fraction. An example of an LROC curve is shown in Figure 4. It can be seen that in contrast to the ROC curve LROC curve does not necessarily pass the point (1, 1). In other words, when the FP fraction is 1, the fraction of TP decisions with correct localization may well be less than 1 because of incorrect localizations. The area under the LROC curve is considered a FOM for task performance, and a higher area under the LROC curve indicates better image quality. However, this is strictly true only when the LROC curves do not cross, and similar considerations apply as for the case of ROC curves.

Free-Response ROC

Again, consider the case of myocardial perfusion SPECT (MPS) imaging, in which the task is to detect perfusion defects. An observer may be able to detect one defect on the anterior wall while missing defect on the septal wall. Consequently, this image will be reported as positive—a correct decision—but for an incomplete reason: there were a total of two defects and only one was detected. Such mistakes might also have serious clinical consequences. To account for this kind of mistake, a different ROC experimental paradigm, which allows an observer to specify an arbitrary number of defects on an image in addition to providing a confidence rating for each defect, is desired.

The free-response ROC (FROC) paradigm has been proposed to account for the localization and detection of abnormalities on images containing an arbitrary number of abnormalities. In a typical FROC experiment, the observer can report an

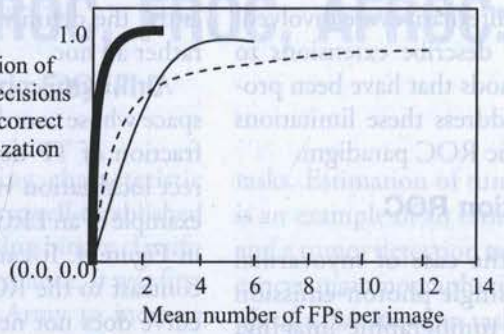

Fig 5. Examples of free-response receiver operating characteristic curves. FP = false positive; TP = true positive.

arbitrary number of mark-rating pairs for each image. When the rating of a mark is above threshold, it is considered a TP decision with correct localization if it is within a tolerance range around the true location and an FP decision otherwise.

Free-response ROC curves are plotted in a space where the coordinate axes are the fraction of TPs with correct localization and the mean number of FPs per image. Examples of FROC curves are shown in Figure 5. It can be seen that an FROC curve starts at the point (0, 0), corresponding to the use of a very high decision threshold such that all decisions are negative. The FROC curve then rises up quickly, corresponding to marks with high confidence, where only a few FP decisions result per image. The curve then reaches a plateau after a shoulder. This plateau corresponds to marks with a low confidence level, where there will be a larger number of FP decisions.

Based on the definitions of the axes of an FROC curve, we see that a better diagnostic system will have an FROC that is higher in the vertical direction (ie, more TP decisions with correct localization) and be more compact in the horizontal direction (ie, fewer FP, per image) [6]. For example, in Figure 5, the bold, solid curve characterizes a system that is better than those characterized by the other two FROC curves. As a result, the area under the curve can no longer be an FOM to summarize the FROC curve. This is because a larger area under the curve can result either from an increase in TPs with correct localization or an increase in the number of FPs on each image.

Compared with the ROC experimental paradigm, the use of location information in FROC experiments increases statistical power, allowing more precise rankings of observer performance. However, FROC studies present significant challenges in terms of data analysis. To address these challenges, several mathematical models have been proposed for fitting FROC curves for various practical observers (eg, physicians, computer-aided diagnosis software). In addition, several FOMs have been proposed to summarize FROC data, such as the probability that lesions (including unmarked lesions) are rated higher than FP marks on normal images. It should be noted that these FOMs are not derived directly from FROC curves [5].

Alternative Free-response ROC

In an FROC study, the raw data consists of mark-rating pairs. An observer may report an arbitrary number of mark-rating pairs per image. In FROC analysis, one counts the total number of FPs per image at each decision threshold and then plots the fraction of TPs with correct localizations against the number of FPs per image at each decision threshold.

Alternative FROC (AFROC) analysis provides an alternative way to analyze FROC data. Alternative FROC and FROC curves differ in the definitions of the horizontal axis. In particular, for a decision threshold, no matter how many FPs are in one image, only one FP decision is considered: the one with the highest rating. Then the fraction of actually negative images falsely called positive are computed. The AFROC curve is then

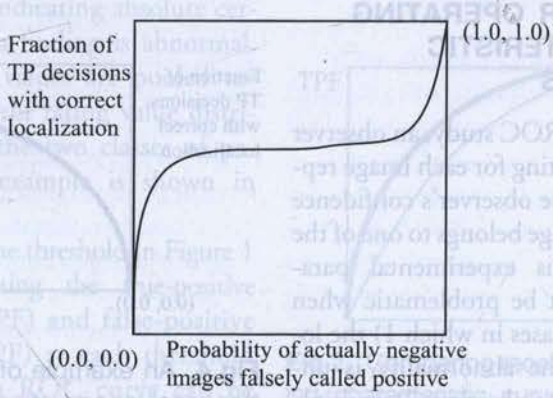

Fig 6. An example of an alternative free-response receiver operating characteristic curve. TP = true positive.

plotted in a space whose coordinate axes are the fraction of TP decisions with correct localization and the probability of actually negative images falsely called positive. The area under the AFROC curve has been proposed as a FOM for task performance. An example of an AFROC curve is shown in Figure 6.

The horizontal axis of an AFROC curve is similar to those of ROC and LROC curves. The only difference is that AFROC analysis allows for marking of more than one location on an image.

Multiclass ROC Analysis

Although ROC analysis has become the standard method for assessing diagnostic performance, it is limited to binary diagnostic tasks. Many diagnostic tasks involve more than two classes. Examples include breast cancer diagnosis using mammography, in which the diagnostic classes are normal, benign tumor, and malignant tumor, and diagnosis of cardiac disease diagnosis using MPS imaging, where the classes are normal, reversible defect, and fixed defect. To assess multiclass diagnostic tests, multiclass ROC analysis is required. Many methods have been proposed to evaluate multiclass classifications [7-9]. However, research in this area is still in its infancy.

DISCUSSION AND CONCLUSIONS

A strong consensus is emerging in the medical imaging community that medical image quality should be assessed using task-based methods. Receiver operating characteristic analysis and its various extensions have been developed to answer this need for the case of classification tasks. Many software packages are readily available to analyze ROC, LROC, and FROC data [10,11]. Binary ROC analysis methods are well established and in common use. Methodology for analyzing LROC experiments is also well established. However, these methods do not allow analysis of data from all experimental paradigms and classification tasks.

As mentioned above, several analysis methods and FOMs have been proposed for both FROC and multiclass ROC. It should be noted that different FOMs may rank systems in different orders, there is currently no consensus about which FOMs are most appropriate or relevant. Like all other techniques, evaluation techniques themselves are techniques that must be evaluated. Currently, research is underway in the medical imaging community relating to evaluate the appropriateness of the competing approaches for analyzing both FROC data and multiclass ROC data.

ACKNOWLEDGMENT

The authors express their sincere appreciation for helpful comments with regard to this article to their colleague Dr Antonio J. Machado. This work was supported by grants K99-EB007620, and R0I-EB000288, from the National Institutes of Health (Bethesda, Md). The content of this work is solely the responsibility of the authors and does not necessarily represent the official view of the National Institutes of Health or its various institutes.

REFERENCES

1. Metz CE, Wagner RF, Doi K, Brown DG, Nishikawa RM, Myers KJ. Toward consensus on quantitative assessment of medical imaging systems. Med Phys 1995;22:1057-61.
2. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283-98.
3. Barrett HH, Myers KJ. Foundations of image science. Hoboken, NJ: John Wiley; 2003.
4. Gifford HC, King MA, Wells RG, Hawkins WG, Narayanan MV, Pretorius PH. LROC analysis of detector-response compensation in SPECT. IEEE Trans Med Imaging 2000; 19:463-73.
5. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis, and validation. Med Phys 2004;31:2313-30.
6. Chakraborty DP, Winter LHL. Free Response methodology alternate analysis and a new observer-performance experiment. Radiology 1990;174:873-81.
7. He X, Song X, Frey EC. Application of three-class ROC analysis to task-based image quality assessment of simultaneous dual-isotope myocardial perfusion SPECT (MPS). IEEE Trans Med Imaging 2008;27: 1556-67.
8. Xiong CJ, Van Belle G, Miller JP, Morris JC. Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. Stat Med 2006;25: 1251-73.
9. Mossman D. Three-way ROCs. Med Decis Making 1999;19:78-89.
10. Kurt Rossman Laboratories for Radiologic Image Research. Software programs available from the Kurt Rossman Laboratories. Available at: http://home.uchicago.edu/~junji/KRL_HP/KRL_ROC/software_index.htm. Accessed June 15, 2009.
11. Chakraborty DP. Dev Chakraborty's FROC Web site. Available at: http://www.devchakraborty.com. Accessed June 15, 2009.

Xin He, PhD, and Eric Frey, PhD, are from the Department of Radiology, Johns Hopkins School of Medicine, Baltimore, Maryland.

Xin He, PhD, Johns Hopkins University, 601 N Caroline Street, Baltimore, MD 21287-0856; e-mail: xinhe@jhmi.edu.