ELSEVIER

European Journal of Radiology 27 (1998) 88–94

EJR
EUROPEAN JOURNAL OF RADIOLOGY

Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology

Arian R. van Erkel [a,b,*], Peter M. Th. Pattynama [a]

[a] *Department of Radiology, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, The Netherlands*
[b] *The Medical Decision Making Unit, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, The Netherlands*

Received 1 May 1997; received in revised form 6 October 1997; accepted 8 October 1997

Abstract

Receiver operating characteristic (ROC) analysis is a widely accepted method for analyzing and comparing the diagnostic accuracy of radiological tests. In this paper we will explain the basic principles underlying ROC analysis and provide practical information on the use and interpretation of ROC curves. The major applications of ROC analysis will be discussed and their limitations will be addressed. © 1998 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Sensitivity; Specificity; ROC analysis; Medical decision making; Radiology

1. Introduction

To address the clinical problems in everyday radiology practice, a large and ever expanding array of imaging modalities is available. This raises the question of what particular test to use for what purpose. There is thus, a need for a method to compare the diagnostic accuracy of the various tests in an objective manner. Over the last two decades, receiver operating characteristic (ROC) analysis has increasingly been used for this purpose, notably in radiology and clinical chemistry [1,2]. Originally developed in the early 1950s for the analysis of RADAR signal detection, ROC analysis was first applied in psychophysical research [1,3,4]. In the 1960s, Dr Lee Lusted was the first to recognize a possible role for ROC analysis in medical decision making [5,6].

In this article we will describe the principles underlying ROC analysis and explain the advantages of this method over conventional analysis, which uses comparison of sensitivity and specificity values. We will provide practical information on how to use and interpret ROC curves. The major applications of ROC analysis will be discussed and their limitations will be addressed.

2. Sensitivity and specificity: Need for ROC analysis

The traditional measures to quantify the diagnostic accuracy of a test are sensitivity and specificity. These parameters describe the fractions of patients (diseased and non-diseased) that are classified correctly. The sensitivity or true positive fraction (TPF) describes the fraction of diseased patients that actually has a positive test result. The specificity or true negative fraction (TNF) describes the probability of a negative test result in non-diseased individuals. Sensitivity and specificity describe the results of a test in a dichotomous way: a test result is either positive or negative. In this respect, there is an analogy to the dichotomous treatment decisions required in clinical practice. Should we operate on this patient suspected of appendicitis or not? Should we start antibiotic treatment for suspected pneumonia or not?

* Corresponding author. Tel.: +31 71 5262993; fax: +31 71 5248256; e-mail: vanerkel@radiology.azl.nl

PII S0720-048X(97)00157-5

By their nature, however, most radiological tests are not dichotomous; they contain much more detailed information. Basically, radiological tests provide one of the three following kinds of data:

1. Continuous quantitative data: The size of a lesion in centimeters or the CT density of a lesion in houndsfield units can, in some situations, indicate the histopathologic nature of the lesion [7]. Within a certain range these data can have all possible values.
2. Rating scale data: Some test information is expressed in an ordinal manner on a rating scale with a limited number of categories. The degree of renal artery stenosis ($<50\%$ stenosis, 50–74% stenosis, 75–99% stenosis and occlusion) can be used for further diagnostic and therapeutic work-up. Rotator cuff pathology can be expressed in terms of normal cuff, degenerative abnormalities, partial or complete tear.
3. Qualitative data: Often no quantitative data are provided. Many criteria used in radiology are of a morphologic nature. Evaluation of the margin and location of a lesion and of the presence and nature of calcifications can all contribute to a more definite diagnosis. These qualitative data can be integrated into a dichotomous diagnostic decision. The evidence of disease, however, is more convincing in some cases than it is in others. Morphologic data can also be integrated into an explicit confidence judgement regarding the probability of disease. In this way, qualitative data are converted to a continuous or ordinal scale of disease probability.

When we translate diagnostic information into a dichotomous yes or no answer, we need decision criteria, or threshold values, to tell normal from abnormal. The choice of this threshold value is subject to both inter- and intra-observer variation. We can distinguish under- and overreaders, who apparently use different threshold values for their decisions. Depending on the clinical situation, even a single radiologist will use different threshold values for the same radiological test. This illustrates that the diagnostic accuracy of a test is inadequately described by a single pair of sensitivity and specificity values. To obviate this problem, we need to compare diagnostic tests by means that are independent of the chosen threshold value.

3. The ROC curve: Basic principles

The choice of the threshold value influences both sensitivity and specificity. For the ideal diagnostic test, the probability distributions of test results indicating presence or absence of disease do not overlap and the chosen threshold value is in between these distributions (Fig. 1). The resulting sensitivity and specificity are both 100%. For most diagnostic tests, however, the probability distributions of diseased and normal overlap. Any threshold value will lead to the misclassification of some diseased patients as normal, or of some individuals without the disease as diseased, or to both (Fig. 2).

Applying a lower threshold value decreases the number of false-negative results (higher sensitivity; Fig. 2a), but increases the number of false positives (lower specificity; Fig. 2b). Raising the threshold value, on the other hand, will increase the number of false negatives (lower sensitivity; Fig. 2a) and decrease the number of false positives (higher specificity; Fig. 2b). There is thus, a reciprocal relationship between sensitivity and specificity. A higher sensitivity is associated with a decrease in specificity and a lower sensitivity with an increased specificity.

The ROC curve is the graphic representation of this reciprocal relationship between sensitivity and specificity, calculated for all possible threshold values (Fig. 3). The vertical axis of the graph shows the sensitivity or TPF. The horizontal axis represents the false-positive fraction (FPF = 1 − specificity). Each operating point on the ROC curve represents the combination of sensitivity and specificity at a given threshold value. At unrealistically high threshold values, all patients are classified as normal, resulting in a TPF of 0 and a FPF of 0 (specificity = 1). This corresponds to the operating point in the lower left-hand corner of the ROC graph. Lowering the threshold will increase both the TPF and FPF (lower specificity). For the lowest possible threshold, the TPF and FPF are both 1 (specificity = 0), corresponding to the upper right-hand corner of the ROC graph.

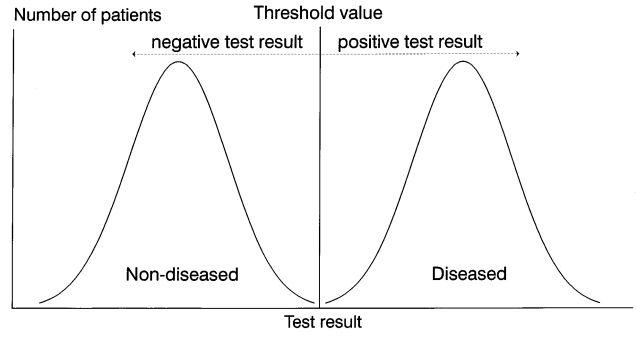

Fig. 1. The probability distributions of the results of a hypothetical perfect diagnostic test. The results of the diseased and non-diseased individuals show no overlap and the chosen threshold value is in between these distributions. If the test result is higher than the threshold value, the test is considered positive. Below the threshold value the test is negative. All diseased and non-diseased patients are classified correctly. Sensitivity and specificity are both 100%.

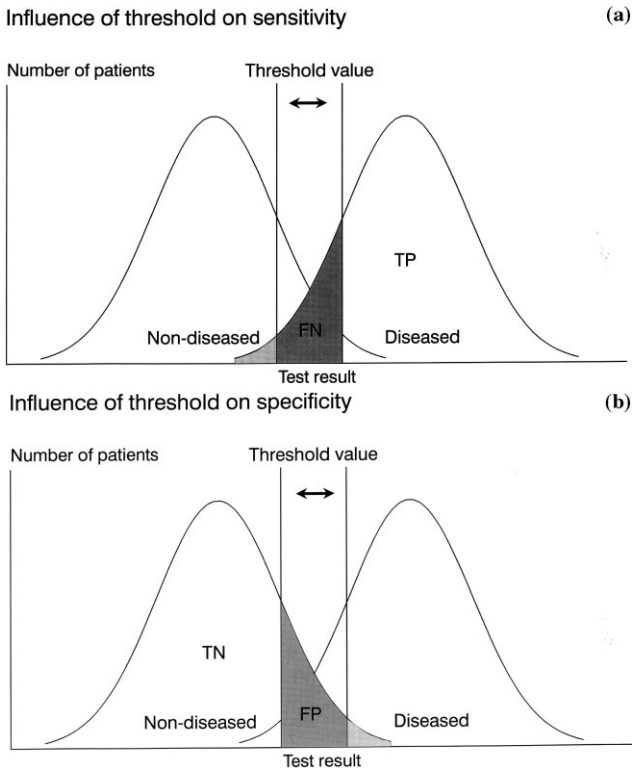

Fig. 2. More realistic probability distributions of the results of a diagnostic test. The results of diseased and non-diseased individuals show some overlap. Variation of the threshold value will change both sensitivity and specificity. (a) Demonstrates the influence of the threshold value on the sensitivity. Applying a lower threshold value results in fewer false-negative (FN) results (small grey area). The fraction of true positives (TP) will be higher (higher sensitivity). Raising the threshold will lead to a higher number of FN (large grey area) and thus to a lower sensitivity. (b) Shows the influence of the threshold value on the specificity. A lower threshold value will increase the number of false-positive (FP) results (large grey area), while the fraction of true negatives (TN) will be reduced (lower specificity).

4. Practical aspects of ROC analysis

ROC analysis can be performed for tests that provide either continuous data or rating-scale data. A rating scale for confidence judgements will generally produce a meaningful curve if five rating categories are used [8,9]. Several computer programs are available to estimate a smooth ROC curve through the observed operating points. The most widely used computer software package is the one developed by Metz et al. [10]. These computer programs estimate a binormal ROC curve. This binormal model is the generally accepted model for ROC analysis and has been shown to be robust for practical purposes [1,11].

In planning experiments to construct ROC curves, special care must be taken to avoid selection bias in the case sample. Two sources of selection bias are of particular importance.

1. Spectrum bias: Owing to the natural spectrum of pathologic, clinical and comorbidity aspects of both normal and diseased individuals, the correct diagnosis is more difficult to make in some individuals than it is in others [12]. The case-mix of patients in the sample influences the position of the ROC curve and must therefore, be representative of the population for which the test is intended [3]. The case-mix is of special importance if the test is preceded by another test that is more or less based on the same principle. Consider the following example. A large pulmonary embolism with a segmental defect on perfusion scintigraphy will probably be detected by pulmonary angiography. On the other hand, an embolism that does not result in a perfusion defect is likely to be smaller and is less readily detected by angiography. The sensitivity of pulmonary angiography is said to be conditionally dependent on the result of perfusion scintigraphy. The ROC curve of pulmonary angiography for patients with a normal perfusion scan differs from the curve for patients with segmental perfusion defects (lack of a true standard of reference for pulmonary embolism precludes the construction of an ROC curve for pulmonary angiography).
2. Disease verification bias: Any diagnostic experiment requires a diagnostic truth provided by a standard of reference. The sensitivity and specificity of a test will be influenced when the test result affects the search for the diagnostic truth and its outcome [3,8,12]. Consider again the example of pulmonary embolism, for which pulmonary angiography is considered the standard of reference. The intensity of the search for pulmonary embolism may be affected by the results of the perfusion scintigraphy. If the perfusion scan is normal, the patient may be considered free of pulmonary embolism and angiography will not be performed. Even if all patients are evaluated with the reference test, there is still a chance of verification bias if the radiologist performing the test is not blinded. If the perfusion scintigraphy shows segmental perfusion defects, we may look more closely for an embolism on angiography, than if the perfusion scintigraphy is normal. When we subsequently construct the ROC curve for perfusion scintigraphy, we may overestimate the sensitivity of perfusion scintigraphy and underestimate its specificity.

5. Applications of ROC analysis: Comparing tests or observers

Which test or observer discriminates best between presence and absence of disease? The discriminative ability of a test is determined by the amount of overlap

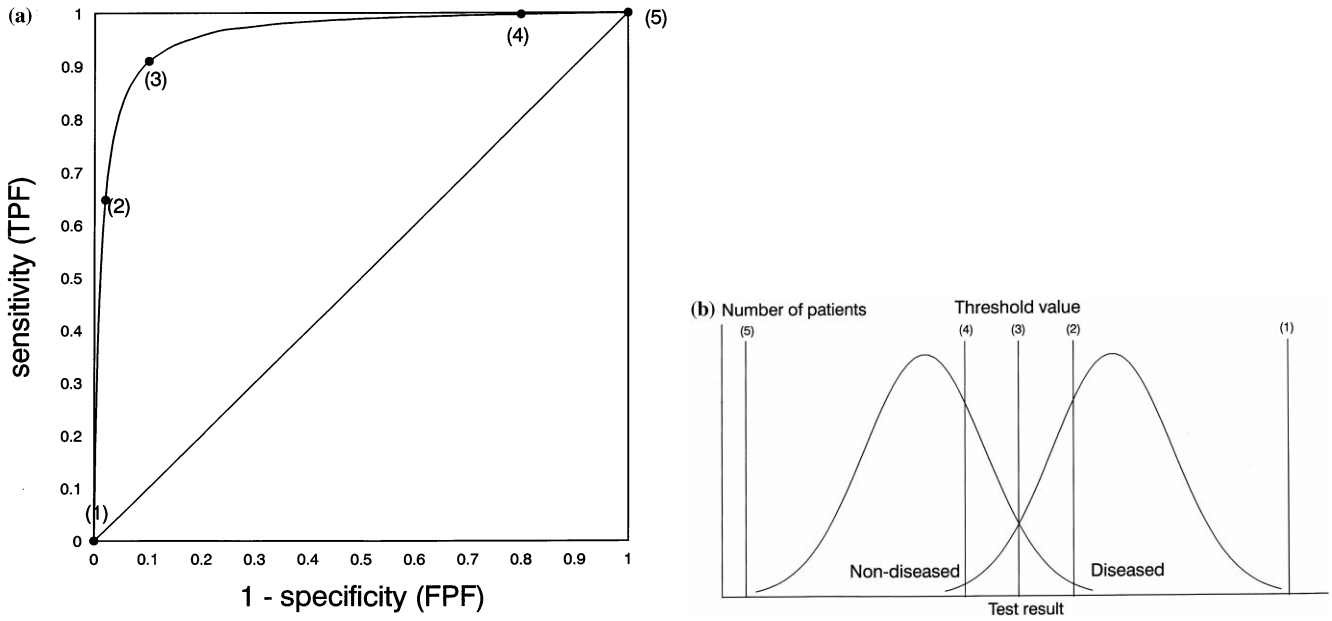

Fig. 3. The ROC curve (a) graphically represents the reciprocal relationship between sensitivity and specificity for all possible threshold values (b). The vertical axis of this graph shows the true-positive fraction (TPF) or sensitivity. The horizontal axis represents the false-positive fraction (FPF) or 1 − specificity. Each threshold value corresponds to an operating point on the ROC curve which represents a combination of sensitivity and specificity. With high threshold values [1], all patients are classified as non-diseased, resulting in a TPF of 0 and a FPF of 0 (specificity = 1). This corresponds to the operating point in the 'lower left-hand corner' of the graph. Lowering the threshold will increase both the TPF and FPF (i.e. decrease specificity). For the lowest possible threshold [5] the TPF and FPF are both 1 (specificity = 0), corresponding to the 'upper right-hand corner' of the ROC graph.

between the probability distributions of the test results of diseased and non-diseased patients. This overlap determines the shape and position of the ROC curve. If the probability distributions of diseased and non-diseased are identical, i.e. they overlap completely, the TPF and FPF are equal at any threshold value. The test has no discriminative power and is essentially worthless. The ROC curve of this test is a straight diagonal from the lower left-hand corner to the upper right-hand corner of the graph (Fig. 3a). The area under this 'curve' is 0.5 (50% of the total area). An ideal test, on the other hand, has no overlap between the distributions. The ROC curve contains the optimal operating point (i.e., TPF = 1 and FPF = 0), corresponding to the upper left-hand corner of the ROC graph. The area under this ROC curve is 1.0 (100% of the total area).

The area under the ROC curve is a measure for the diagnostic accuracy of a test and is often used to make comparisons between diagnostic tests or observers [1,3]. With the appropriate computer software, the areas under the ROC curves can be computed and tested for significant differences with a univariate z-score test [10,13]. Although the binormal assumption of the ROC curve has proven to be valid, some authors recommend using the non-parametric Wilcoxon statistic in analyzing for differences between areas under the ROC curves [14–17]. With non-parametric calculation of the area under the curve, non-parametric methods for comparing the area under the ROC curve are more appropriate than the z-score test. It should therefore be understood, that the significance of the differences between the areas under ROC curves may vary with the method of analysis used.

Comparison of two tests using the area under the ROC curve is quite straightforward if each test is interpreted by a single reader or observer. Statistical testing becomes much more cumbersome whenever the comparison involves ROC curves that are generated by pooling data from multiple observers. Pooling data, however, is often desirable, as it is a useful method to average out the within-reader differences [8]. Several approaches have been proposed, to deal with this problem. The commonly used paired t-test has the distinct disadvantage that it does not take account of the case sample variation [8,18]. The non-parametric method recently suggested by Swaving et al. accounts for both the within-reader and the case sample variation. All methods however, have their own specific methodological pros and cons, a detailed discussion of which is beyond the scope of this paper. Interested readers may wish to consult the papers by Metz and by Swaving et al. [8,18].

The major advantage of comparing tests by means of the area under the ROC curve is that this is done independently of decision criteria, thus eliminating the influence of the threshold value on sensitivity and specificity values. However, by doing so, another problem

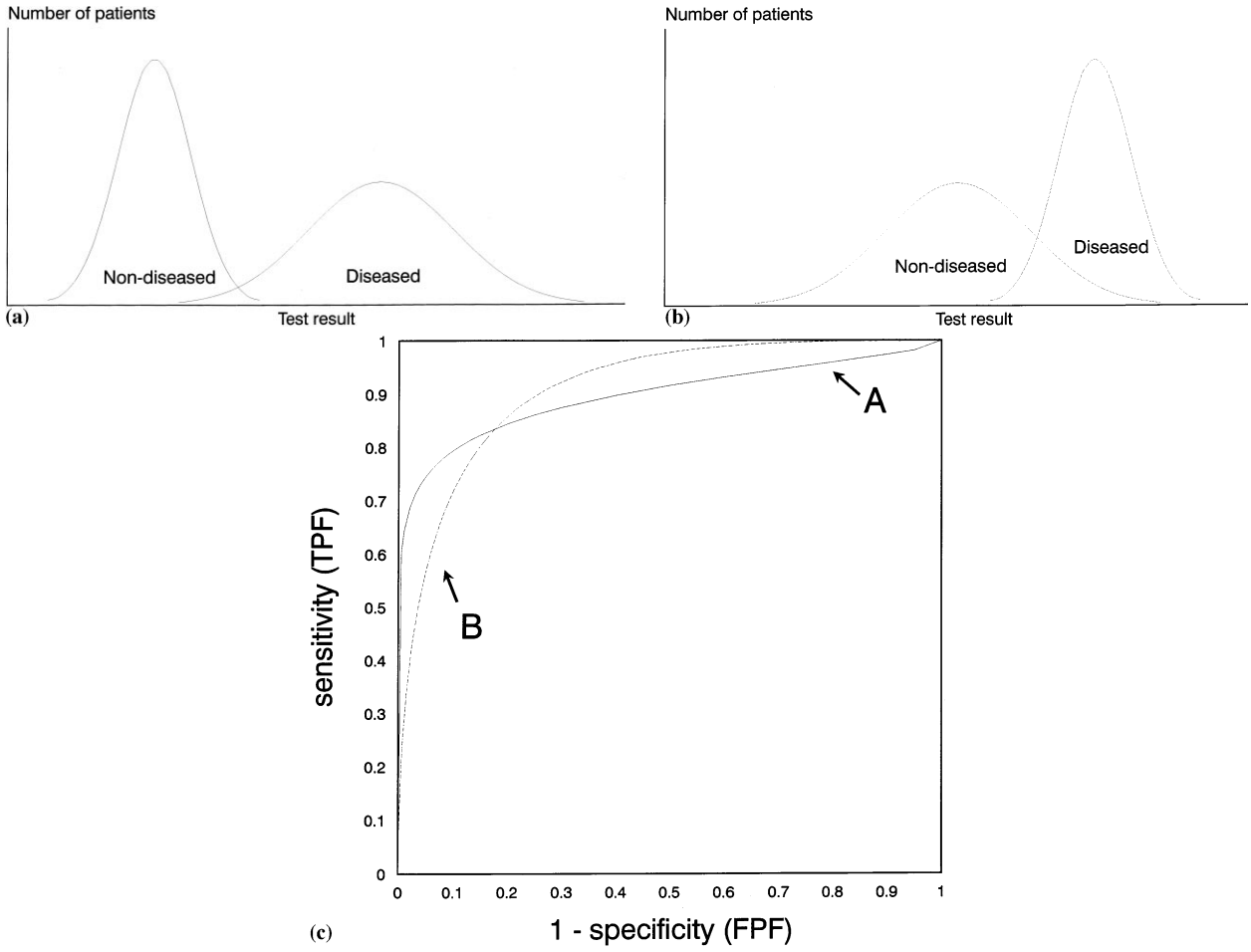

Fig. 4. How can ROC curves intersect? The position of the ROC curve depends on the shape and position of the probability distributions of diseased and non-diseased individuals. When the highest threshold value in (a) is reduced, the TPF will rise immediately, while the FPF will change only after further reduction of the threshold value. These probability distributions will result in a curve that is positioned on the left-hand side of the graph (c: curve A). If the probability distributions are oriented as in (b), the TPF will rise when the highest threshold value is lowered, but the FPF will rise sooner compared to the situation in (a). The ROC curve is positioned more in the upper right-hand quadrant of the graph (curve B)

is created. A large part of the ROC curve consists of clinically irrelevant TPF and FPF combinations. The extreme corners of the ROC curve represent combinations of either very high sensitivity with very low specificity or vice versa. In most clinical situations these combinations are not useful. This problem is especially relevant if the ROC curves of two different tests intersect. Fig. 4 demonstrates how different probability distributions of test results result in intersecting ROC curves. Whenever the clinical situation demands high sensitivity, we want to use the test that results in a higher ROC curve in the high sensitivity value range. In Fig. 4c, for example, test B is preferred over test A, because it has a higher specificity at the same high sensitivity values. The total areas under the curves, however, are equal and fail to reveal the superiority of test B in this respect. Therefore, the area under the curve is of limited use in comparing the diagnostic accuracy of tests with intersecting ROC curves [19,20].

To address this problem, two methods for regional assessment of the ROC curve have been suggested. McGlish has advocated the analysis of a portion of the ROC curve that is determined by a range of FPF values [21]. This method allows comparison of the diagnostic capacity of tests for a preset area of specificity values. The partial area index (PAI), suggested by Jiang et al. uses a preset area of sensitivity values [22]. The PAI is calculated by dividing the estimated area under the portion of the ROC curve of interest (the range of TPF values above a preselected TPFo) by the maximum possible area under this part of the curve. In this way an index value is created, ranging from 0 to 1, that can be used for comparisons analogous to the total area under the curve [2,22].

The following example illustrates the use of ROC analysis in comparing diagnostic tests. Jiang et al. used ROC analysis to compare the diagnostic accuracy of a group of five radiologists with that of a computer-aided

diagnostic scheme in differentiating between benign and malignant microcalcifications in mammography [22]. The probability of malignancy was estimated on a continuous scale and the areas under the ROC curves were calculated. The total areas under the ROC curves of the individual radiologists did not differ significantly from the area of the computer, as calculated with the univariate z-score test. The combined area under the curve of the five radiologists was 0.89 and for the computer the total area under the curve was 0.92. The difference was not significant (Student's t-test). The ROC curves intersected and therefore, the PAI was calculated for the sensitivity range between 90 and 100%, which was considered clinically relevant. This resulted in significantly different values for the combined radiologists and the computer of 0.42 and 0.82, respectively.

6. Applications of ROC analysis: Optimizing the threshold value

Another potential use of the ROC curve is in optimizing the threshold value of a test. The ROC curve comprises all possible combinations of sensitivity and specificity at all possible threshold values. This offers the opportunity to assess the optimal threshold value to be used in clinical practice.

In practice, choosing an optimal threshold value based on ROC analysis is practicable only for continuous data, e.g. Doppler velocity parameters for carotid artery stenosis or CT-density for characterization of adrenal masses [7,23]. For continuous data, all operating points on the curve correspond to realistic threshold values. For ordinal test results, the smooth ROC curve is falsely suggestive of continuity [2]. In this situation, the ROC curve is a theoretical estimation based on a limited number of observed operating points. Most of the operating points on the ROC curve consist of sensitivity and specificity combinations that do not correspond to realistic threshold values. A similar problem occurs whenever a categorical rating scale of disease probability is used in generating the ROC curve. This artificial ordinal scale is mostly used for scientific purposes only. The actual threshold values that we use in clinical practice are unclear and cannot be related to the scientifically observed operating points.

In our group we have used ROC curves to compare the diagnostic accuracy of CT and MRI in distinguishing between adenomas and nonadenomas of the adrenal gland [7]. Using the CT-density of the lesion resulted in a significantly larger area under the curve compared with all MRI-parameters. The continuous character of the attenuation values allowed determination of the optimal threshold value. Using a threshold value of 16.5 houndsfield units resulted in a sensitivity and specificity of 100 and 95%, respectively.

Intuitively, one would identify the 'optimal' operating point as the point on the ROC curve that is closest to the ideal upper left-hand corner. Determining the optimal operating point on the ROC curve, however, involves both clinical and financial issues. For instance, pneumonia is a disease in which a large therapeutic gain can be achieved at relatively little cost and with few complications of the antibiotic treatment. Thus, a certain amount of false positive results is acceptable, whereas false negative results are less desirable. As a result, in testing for pneumonia, the threshold value is usually set low. In other words, the optimal operating point will move towards the upper right-hand part of the ROC curve. On the other hand, when we are considering a costly and potentially harmful treatment with only little therapeutic benefit, it is the false positive results that are to be limited. The optimal range of the operating point will thus, shift towards the lower left-hand corner of the ROC graph. Clearly, in determining the optimal threshold value, we have to take into account all the clinical and financial consequences of the different test results. Ideally, such decisions should be made by linking the constructed ROC curve to explicit decision analysis [24,25].

In summary, ROC analysis is a useful technique to compare the diagnostic accuracy of radiological tests and observers. The area under the curve provides an objective parameter of the diagnostic accuracy of a test, which is superior to comparing single combinations of sensitivity and specificity values, since the influence of the threshold value is eliminated. Because only part of the ROC curve represents clinically relevant combinations of sensitivity and specificity, comparing the ROC curves in the relevant sensitivity or specificity ranges is to be preferred over comparing the total area under the curve.

In addition, ROC analysis can be used to determine the optimal threshold value for tests that generate continuous quantitative data. Choosing the optimal operating point on the ROC curve involves both clinical and financial issues and is ideally done by combining ROC analysis with a formal cost-effectiveness analysis.

References

[1] Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986;21:720–33.
[2] Dwyer AJ. In pursuit of a piece of the ROC. Radiology 1997;202:621–5.
[3] Hanley JA. Receiver operating characteristic analysis (ROC) methodology: The state of the art. Crit Rev Diagn Imaging 1989;29:307–35.
[4] Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. Invest Radiol 1979;14:109–21.

[5] Lusted LB. Logical analysis in roentgen diagnosis. Radiology 1960;74:178–93.
[6] Lusted LB. Signal detectability and medical decision making. Science 1971;171:1217–9.
[7] Van Erkel AR, Van Gils APG, Lequin M, Kruitwagen C, Bloem JL, Falke THM. CT and MR distinction of adenomas and nonadenomas of the adrenal gland. J Comput Assist Tomogr 1994;18:432-438.
[8] Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989;24:234–45.
[9] Rockette HE, Gur D, Metz CE. The use of continuous and discrete judgements in receiver operating characteristic studies of diagnostic imaging techniques. Invest Radiol 1992;27:169–72.
[10] Metz CE, Kronman HB. Computer programs ROCFIT, LABROC1, CORROC2, CLABROC, INDROC, ROCPWRPC. Available from CE Metz, Department of Radiology, University of Chicago, Chicago, IL.
[11] Hanley JA. The robustness of the 'binormal' assumptions used in fitting ROC curves. Med Decis Mak 1988;8:197–203.
[12] Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. New Engl J Med 1978;299:829–30.
[13] Metz CE, Wang PL, Kronman HB. A new approach for testing the significance of differences between ROC curves measured from correlated data. In: Deconinck F, editor. Information Processing in Medical Imaging. The Hague: Nijhoff, 1984:432–45.
[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] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839–43.
[16] Vida S. A computer program for non-parametric receiver operating characteristic analysis. Comput Methods Programs Biomed 1993;40:95–101.
[17] DeLong ER, DeLong DM, Clarke Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 1988;44:837–45.
[18] Swaving M, Van Houwelingen H, Ottes FP, Steerneman T. Statistical comparison of ROC curves from multiple readers. Med Decis Mak 1996;16:143–52.
[19] Centor RM. Signal detectability. The use of ROC curves and their analyses. Med Decis Mak 1991;11:102–6.
[20] Mann FA, Hildebolt CE, Wilson AJ. Statistical analysis with receiver operating characteristic curves. Radiology 1992;184:37–8.
[21] McClish D. Analyzing a portion of the ROC curve. Med Decis Mak 1989;9:190–5.
[22] Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology 1996;201:745–50.
[23] Hunink MGM, Polak JF, Barlan MM, O'Leary DH. Detection and quantification of carotid artery stenosis: Efficacy of various Doppler velocity parameters. Am J Radiol 1993;160:619–25.
[24] Halpern EJ, Albert M, Krieger AM, Metz CE, Maidment AD. Comparison of receiver operating characteristic curves on the basis of optimal operating points. Acad Radiol 1996;3:245–53.
[25] DeNeef P, Kent DL. Using treatment-tradeoff preferences to select diagnostic strategies. Linking the ROC curve to threshold analysis. Med Decis Mak 1993;13:126–32.