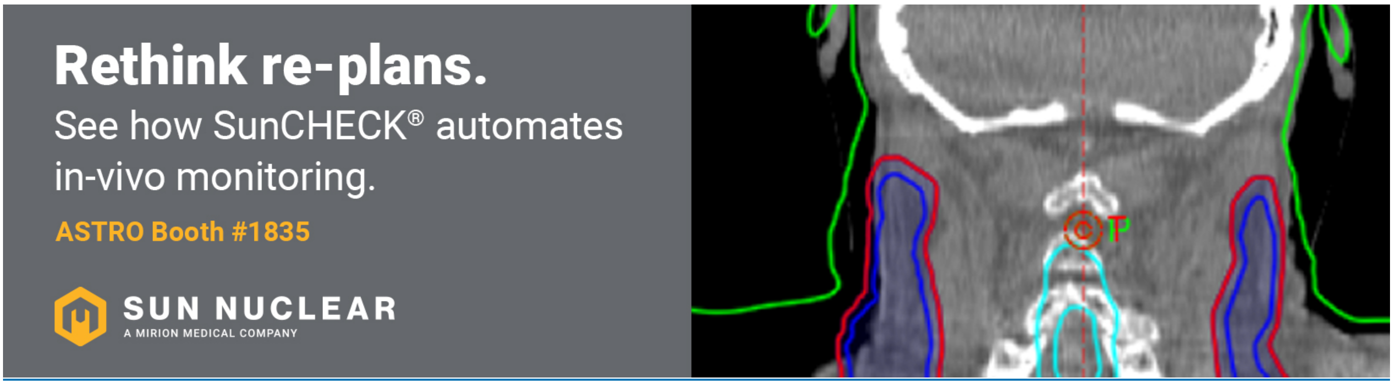
Physics in Medicine & Biology

Institute of Physics and Engineering in Medicine

TOPICAL REVIEW

Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine

To cite this article: Nancy A Obuchowski and Jennifer A Bullen 2018 *Phys. Med. Biol.* **63** 07TR01



Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
31 December 2017

REVISED
21 February 2018

ACCEPTED FOR PUBLICATION
7 March 2018

PUBLISHED
29 March 2018

TOPICAL REVIEW

Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine

Nancy A Obuchowski and Jennifer A Bullen

Quantitative Health Sciences, The Cleveland Clinic Foundation, Cleveland, OH, United States of America

Keywords: receiver operating characteristic curve, diagnostic medicine, ROC studies

Abstract

Receiver operating characteristic (ROC) analysis is a tool used to describe the discrimination accuracy of a diagnostic test or prediction model. While sensitivity and specificity are the basic metrics of accuracy, they have many limitations when characterizing test accuracy, particularly when comparing the accuracies of competing tests. In this article we review the basic study design features of ROC studies, illustrate sample size calculations, present statistical methods for measuring and comparing accuracy, and highlight commonly used ROC software. We include descriptions of multi-reader ROC study design and analysis, address frequently seen problems of verification and location bias, discuss clustered data, and provide strategies for testing endpoints in ROC studies. The methods are illustrated with a study of transmission ultrasound for diagnosing breast lesions.

1. Introduction

1.1. Role of diagnostic test accuracy assessment in medicine

Receiver operating characteristic (ROC) curve analysis is a statistical tool used extensively in medicine to describe diagnostic accuracy. It has its origins in WWII to detect enemy weapons in battlefields but was quickly adapted into psychophysics research (Peterson *et al* 1954, Tanner *et al* 1954, Van Meter *et al* 1954, Lusted 1971, Egan 1975, Swets 1996) due largely to the statistical methods and interpretation contributions of Hanley and McNeil (1982, 1983) and Metz (1978, 1986). Nowadays, ROC analysis is commonly used for characterizing the accuracy of medical imaging techniques (e.g. mammograms for detecting breast cancer, low-dose Computed Tomography to detect lung cancer), as well as non-imaging diagnostic tests (fasting glucose test for detecting diabetes, exercise stress test for detecting coronary artery disease), and prediction/risk scores (e.g. Framingham risk score to predict risk of cardiac events) in a variety of settings including screening, diagnosis, staging, prognosis, and treatment (Shiraishi *et al* 2009).

Before diagnostic tests (or prediction/risk scores) can be used in the clinical management of patients, they must be thoroughly evaluated. Fryback and Thornbury (1991) described six levels of diagnostic test efficacy: (1) technical efficacy, (2) diagnostic accuracy efficacy, (3) diagnostic thinking efficacy, (4) therapeutic efficacy, (5) patient outcome efficacy, and (6) societal efficacy. The second level, diagnostic accuracy efficacy, relies heavily on studies involving ROC analysis. Diagnostic accuracy efficacy evaluation typically starts with *exploratory* studies with easy-to-diagnose cases, followed by *intermediate* studies of difficult cases in order to challenge a new test. Finally, in *advanced* studies, large cohorts are examined to establish community-based estimates of accuracy (Zhou *et al* 2011). In each phase ROC analysis is used to characterize the test's ability to discriminate between groups of patients, where the groups are established by some highly accurate reference standard.

Table 1 lists a few applications of ROC analysis in medicine and biology. Examples of different diagnostic tests are given, along with the binary condition they are used to detect. The table is separated into applications where the diagnostic test provides an objective measurement (e.g. tumor heterogeneity based on lesion texture features measured quantitatively from a FDG-PET image to discriminate early vs advanced stage cervical cancer) vs subjective measurement (e.g. trained readers' subjective confidence scores to differentiate the presence/absence of a lesion on a mammogram). The latter applications involving human observers are prevalent in the medical imaging literature, accounting for almost 80% of published ROC studies (Shiraishi *et al* 2009). We will see that ROC studies of diagnostic tests with objective measures are easier to design and analyze because there are no confounding effects due to the human reader.

Table 1. Example applications of ROC analysis.

Authors	Application
Objective measures	
Mazzetti *et al* (2012)	Compare dynamic contrast enhanced MRI models to discriminate benign versus malignant prostate tissue
Mu *et al* (2015)	Measure accuracy of texture features to discriminate early versus advanced stage cervical cancer
Allen *et al* (2012)	Explore the utility of photoplethysmography as a diagnostic marker for chronic fatigue syndrome
Mahmoud *et al* (2014)	Detect hepatic steatosis using ultrasound thermal strain imaging
Subjective measures	
Bartkiewicz ***et al*** (1991)	Assess the effect of gamma camera variables on physicians' diagnostic performance
Lai *et al* (2006)	Compare physicians' ability to identify microcalcifications on three mammography systems

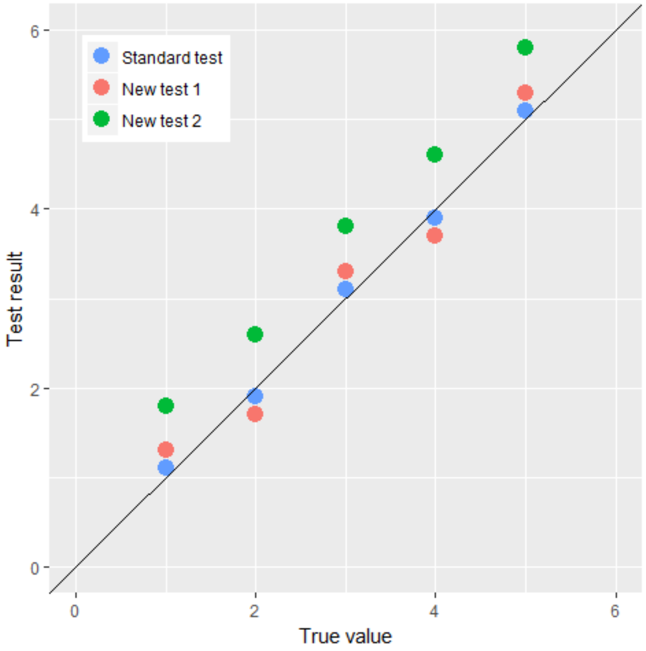


Figure 1. Distinguishing accuracy from correlation or agreement.

1.2. Accuracy versus correlation versus agreement

ROC curves characterize a diagnostic test's accuracy, in other words, how the test results discriminate between subjects with versus without a particular condition. In order to assess accuracy, a *reference standard* is needed to classify subjects into those who truly have the condition and those who truly do not. Zhou *et al* define a reference standard as 'a source of information, completely different from the test or tests under evaluation, which tells us the true condition status of the patient' (Zhou *et al* 2011). Surgical findings and pathology results are two common reference standards in diagnostic medicine, whereas long-term follow-up might be the reference standard for a prediction/risk score. Accuracy studies imply that a reference standard is available to distinguish correct and incorrect test results.

Too often investigators perform studies assessing the correlation or agreement of new tests with some established standard test, rather than performing an accuracy study. Figure 1 illustrates the problems with correlation and agreement type studies. The measurements of a fictitious standard test and two new diagnostic tests are shown, relative to the true value as determined by a reference standard. One of the new diagnostic tests has excellent agreement with the standard test while the other has excellent correlation. Based on these findings we might naively conclude that the new tests have similar accuracy as the standard test; however, both of the new tests have lower accuracy than the standard test. This would not be evident unless an accuracy study was performed.

The black line illustrates the relationship between the true value and the test result for a perfectly accurate test (i.e. where the test result is always exactly the true value). The standard test, with measurements illustrated as blue circles, has good accuracy based on the relatively short distance from this black line. New test 1, with measurements displayed as red circles, has good agreement with the standard test. New test 2, displayed as green circles,

Table 2. Sensitivity and specificity.

	Diagnostic test		
Reference standard	Positive	Negative	Total
Condition present	a	b	n_1
Condition absent	c	d	n_0

Classic 2 × 2 table illustrating a diagnostic test's sensitivity and specificity. The rows represent the results of the reference standard, while the columns represent the results of the diagnostic test being studied. a is the number of subjects with the condition who tested positive, i.e. the true positives. b is the number of subjects with the condition who tested negative, i.e. the false negatives. c is the number of subjects without the condition who tested positive, i.e. the false positives. d is the number of subjects without the condition who tested negative, i.e. the true negatives. Sensitivity is estimated as a/n_1. Specificity is estimated as d/n_0. The false positive rate (FPR) is estimated as c/n_0.

is perfectly correlated with the standard test. Both new tests have lower accuracy than the standard test though, which would not be evident without a reference standard to establish the true value.

1.3. Sensitivity and specificity as basic metrics of accuracy

In its simplest form the accuracy of a diagnostic test can be ascertained by two questions:

> *How well does the test perform among subjects with the condition?*
> *How well does the test perform among subjects without the condition?*

The two corresponding measures of accuracy are *sensitivity* and *specificity*. These accuracy metrics are defined in table 2 where N subjects have been classified by the reference standard as either condition present (total number of subjects with the condition is n_1) or condition absent (total number of subjects without the condition is n_0). Likewise, the subjects are classified by the diagnostic test as either positive or negative for the condition. Sensitivity is defined as a/n_1 and specificity as d/n_0.

Sensitivity and specificity, as metrics of accuracy, are limited in how they characterize accuracy:

1. Sensitivity and specificity must always be considered simultaneously. This makes it difficult to compare diagnostic tests (e.g. Test A can have higher sensitivity than Test B, while Test B has higher specificity than Test A).
2. Diagnostic test results must be dichotomized (positive or negative) in order to estimate sensitivity and specificity. The cut-point for defining the dichotomy is often arbitrary, unstandardized, and/or reduces information by pooling different findings together. Gatsonis (2009) calls this understanding that sensitivity and specificity are dependent on one another through a cut-point as 'ROC thinking'.
3. Sometimes we want to compare diagnostic tests whose results are defined differently (e.g. ordinal versus continuous scale), making it difficult to make equitable comparisons. We will illustrate this in the transmission ultrasound study in section 1.4.
4. For some conditions the reference test does not classify subjects neatly into two categories. The reference standard might classify subjects into more than two categories (e.g. normal, appendicitis, or intestinal obstruction (Obuchowski *et al* 2001)) or even provide an ordinal- or continuous-scaled value (e.g. scarring of heart muscle by positron emission tomography (PET) using an ordinal scale of normal, hibernating, ischemic, or necrotic tissue (Obuchowski 2006); biochemical analysis of serum ferritin concentration measured in μg l^{-1} (Obuchowski 2005)). This situation is further discussed in section 4.4.

ROC curves and their associated summary indices and modifications overcome these limitations.

1.4. Illustrative example: transmission ultrasound imaging of the breast

An example will help illustrate the concepts that we will present in sections 2–5. Transmission ultrasound has been developed to improve spatial and contrast resolution of breast tissue (Lenox *et al* 2015), and importantly to reduce false positive findings during breast screening and diagnostic follow-up. In addition to reflection B-mode imaging, the transmission ultrasound data provide a measure of the speed of sound, reflecting the stiffness of breast tissue or lesions (Iuanow *et al* 2017). With permission from its developers (Klock 2017), we present the results of a small pilot study of quantitative transmission ultrasound to illustrate the basic concepts of ROC analysis.

In this feasibility study breast transmission ultrasound images were collected from subjects undergoing diagnostic work-up at two institutions. Subjects underwent biopsy of suspicious lesions to discriminate cystic lesions

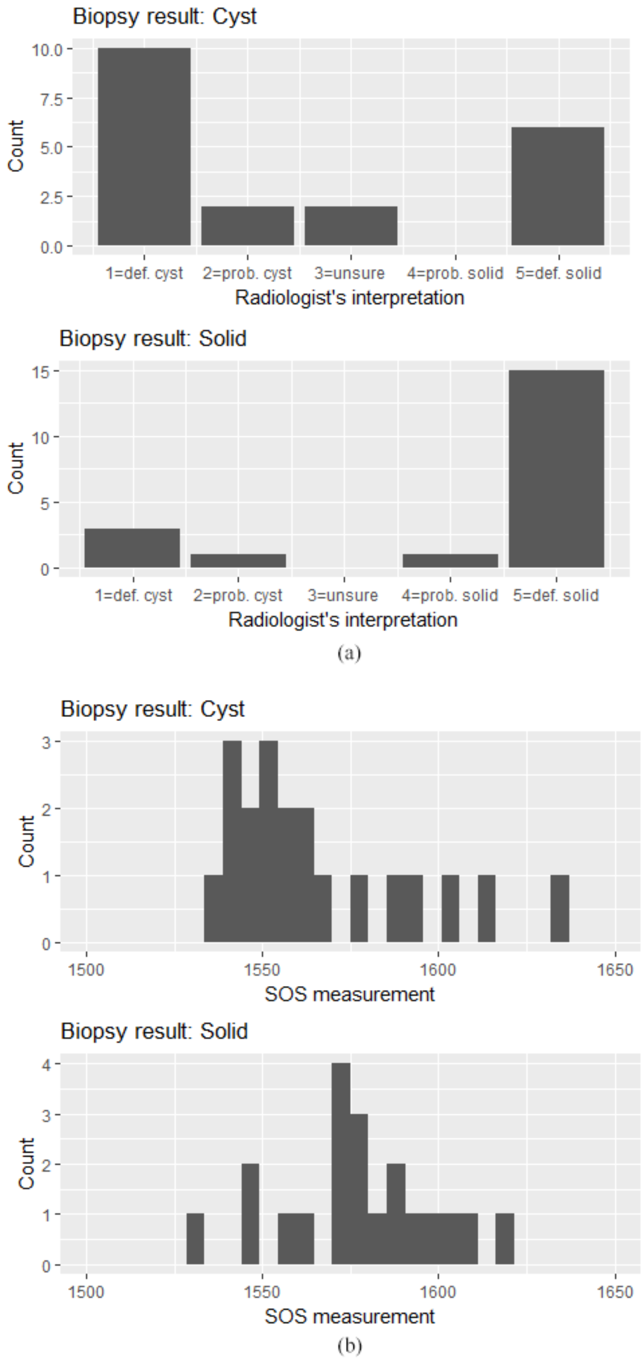


Figure 2. (a) Distribution of radiologist's subjective interpretation of the images in the transmission ultrasound study, stratified by biopsy result. (b) Distribution of the speed of sound (SOS) measurements in the transmission ultrasound study, stratified by biopsy result.

from solids. In this retrospective study design, 20 subjects with known cysts and 20 subjects with known solids, based on biopsy results, were selected for the study. A radiologist, blinded to the biopsy results, interpreted the images and scored the lesions as

1 = definitely a cyst (>90% confidence that lesion is a cyst)
2 = probably a cyst (60%–90% confidence that lesions is a cyst)
3 = unsure
4 = probably a solid (60%–90% confidence that lesion is a solid)
5 = definitely a solid (>90% confidence that lesion is a solid).

In addition, the speed of sound (SOS) measurements were recorded for each of the 40 lesions. Figure 2(a) illustrates the distribution of the radiologist's subjective interpretation of the images, stratified by biopsy result, and figure 2(b) illustrates the distribution of the SOS measurements. In section 2 we will illustrate construction of ROC curves from each of these results.

2. Basics of the receiver operating characteristic (ROC) curve

2.1. Construction of ROC curve

An ROC curve is a plot of a diagnostic test's sensitivity (*y*-axis) versus its false positive rate (*x*-axis). Each (*x*,*y*) coordinate on the plot represents the sensitivity and false positive rate associated with a cut-point for defining positive and negative. Using the quantitative transmission ultrasound study in figure 2(a) we plot six coordinates associated with the following four cut-points: >1, >2, >3 and >4 (in addition to the non-informative cut-points of >0 and >5 with associated coordinates of (1,1) and (0,0), respectively) (see section 4.1 for details on calculating these coordinates). We connect the six points with line segments; the resulting curve is called the *empirical ROC curve* (see blue curve in figure 3). The diagonal line (dashed black line) from (0,0) to (1,1) is called the *chance diagonal*; it illustrates the accuracy of a diagnostic test based on pure guessing. Sometimes for a newly developed diagnostic test we might test the null hypothesis that its ROC curve is equivalent to the chance diagonal. We would expect to reject such a null hypothesis for diagnostic tests with any utility.

Figure 2(b) illustrated data from the quantitative biomarker SOS. The SOS has 31 informative cut-points. The red curve in figure 3 illustrates the empirical ROC curve for the SOS. As the red curve lies below the blue curve, we conclude that the diagnostic accuracy of the SOS alone is inferior to a radiologist's interpretation of the image. Also note that although the two tests were measured on different scales (i.e. subjective ratings by the radiologist and a quantitative measure for the SOS), both tests' accuracies can be depicted on the same ROC plot and compared on common metrics.

2.2. Associated summary measures and interpretation

While the ROC curve plot nicely illustrates diagnostic tests' accuracies, it is often desirable to summarize tests' accuracies by a single number. The most commonly used summary measure is the area under the ROC curve (i.e. AUC). The AUC of a diagnostic test with no diagnostic ability is 0.5 (i.e. the area under the chance diagonal, which is half of the unit square), while a test that perfectly discriminates between two conditions has an AUC of 1.0. The interpretation of the AUC is as follows: suppose there are two samples of subjects: one sample of subjects without the condition (e.g. cyst) and one sample of subjects with the condition (e.g. solid lesion). If we were to randomly select a subject from each sample, then the AUC is the probability that the diagnostic test assigned a higher score to the subject with the condition than to the subject without the condition.

It is easy to understand this interpretation if one considers the following simple formula for estimating the AUC:

$$\widehat{\mathrm{AUC}} = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \Psi \tag{1}$$

where ψ takes on one of three values: $\Psi = \begin{cases} 0, \text{ if subject with condition is rated lower than subject without the condition} \\ \frac{1}{2}, \text{ if subject with condition is rated the same as subject without the condition} \\ 1, \text{ if subject with condition is rated higher than subject without the condition} \end{cases}$

In words, the AUC can be estimated by pairing each subject with the condition against each subject without the condition. Each pair is assigned a score of 0 (if the subject with the condition has a lower test result than the subject without the condition), ½ (if the subject with the condition has the same test result as the subject without the condition), or 1 (if the subject with the condition has a higher test result than the subject without the condition). The average score, taken over all pairs, is the estimate of the AUC.

Note that the AUC describes the test's ability to discriminate between subjects with and without the condition. The AUC does not describe the *calibration* of the diagnostic test, i.e. the ability of the test to assign the 'correct' scores. In other words, even if the AUC is high, say 0.85, we cannot say that the SOS measurement by transmission ultrasound is providing the correct SOS value in 85% of cases. Rather, we can say that the measured values of SOS for solids are higher than the measured values of SOS for cysts in 85% of comparisons. Similarly, even if the radiologist's AUC is high, we cannot be sure that an assignment of 'definitely solid (>90%)' truly means that there is a >90% chance that the lesion is a solid. Rather, we conclude that the radiologist tends to assign higher scores to solids than to cysts.

There are several other summary measures of accuracy associated with the ROC curve, namely the partial area under the curve, and the sensitivity at a fixed FPR (see figure 4 for illustration). Although used less often than the AUC, these metrics have an important role, particularly in the advanced stage of a diagnostic test's

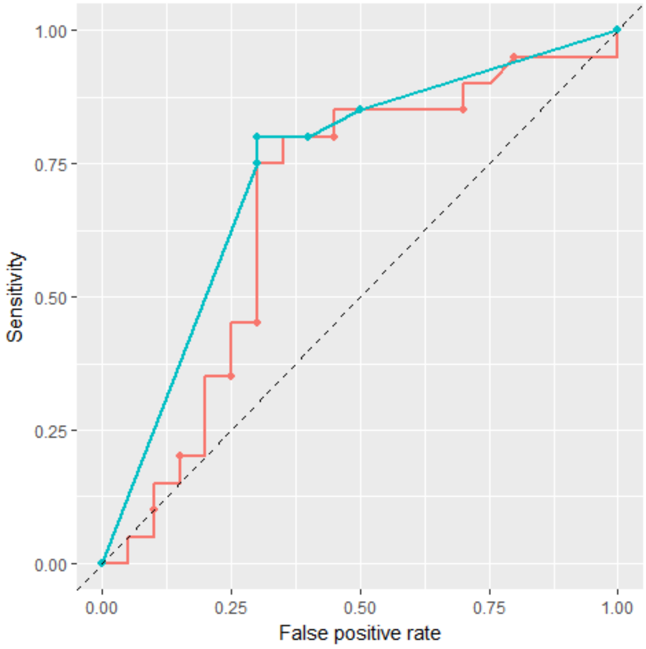


Figure 3. The empirical ROC curves for the radiologist's subjective interpretations (blue) and the SOS measurements (red) in the transmission ultrasound example.

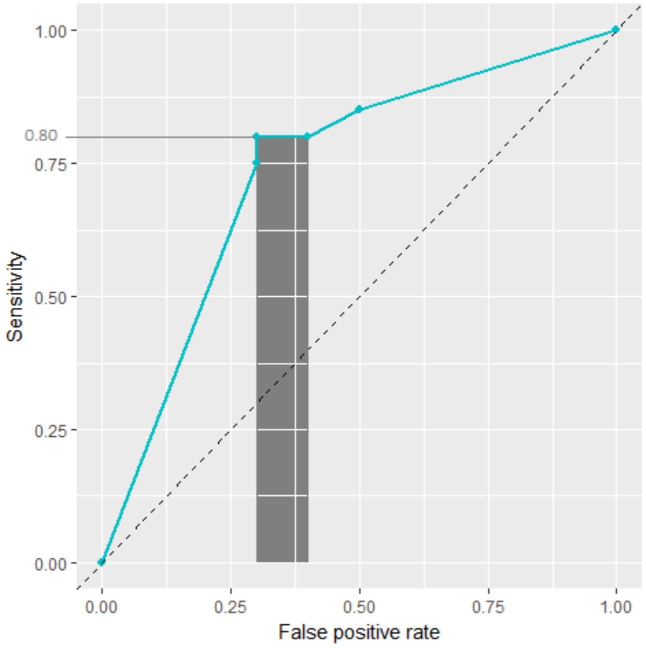


Figure 4. Looking at the empirical ROC curve for the radiologist's subjective interpretations in the transmission ultrasound example, the partial area under the curve when the FPR is between 0.30 and 0.40 is shown in gray. While the area under the full ROC curve is 0.74, this partial area is 0.08. At a fixed FPR of 0.30, the estimated sensitivity is 0.80.

assessment. These metrics focus on the part of the ROC space of most clinical relevance, in contrast with the AUC which considers all cut-points, even ones not clinically relevant. Estimation of these metrics is described in section 4. Zhou *et al* (2011) provide even greater detail, including resampling methods for estimation of the variances and confidence intervals. They also describe methods for identifying the optimal cut-point on the ROC curve. For example, for a given patient with a specified pre-test probability of disease and patient-specific aversions to false positive versus false negative results, the optimal cut-point for defining a positive test result for that patient can be calculated.

Table 3. Comparison of prospective and retrospective study designs for ROC studies.

Study features	Prospective design	Retrospective design
Reference standard	Allows the reference standard to be performed in a standardized fashion for all subjects as specified in the study protocol	Because the reference standard was performed in the past, it can often be non-standardized in terms of its performance and interpretation. Bias can be introduced when not all subjects undergo the reference standard.
Study subjects	Provides a generalizable sample, but large numbers are needed to ensure sufficient numbers of subjects with and without the condition.	Allows the investigators to select subjects for the study to ensure sufficient numbers and a broad spectrum of cases.
Study readers	Requires real-time interpretation by a clinical reader; however, retrospective interpretation by multiple readers is common.	Lends itself to standardized interpretation by multiple readers who are not connected to the patients' care.
Outcome measures	Diagnostic test accuracy (e.g AUC), effect of physician decision making on patient management and patient outcomes. Typically, more patient information (e.g. comorbidities) is available for subgroup analyses.	Diagnostic test accuracy
Efficiency	Tends to be long and expensive in order to recruit sufficient numbers of subjects.	Tends to be substantially shorter and less expensive because chart reviews are used to identify subjects.

3. Study design

Diagnostic accuracy studies are often more challenging to design than a typical treatment comparison study. First, there is the challenge of identifying an appropriate reference standard and ensuring that all study subjects are assessed by it. Second, diagnostic accuracy involves discriminating between (at least) two conditions (e.g. normal versus abnormal, benign versus malignant, cyst versus solid) and thus an appropriate number of subjects representing each condition must be included in the study sample. Third, for diagnostic tests that require subjective interpretation, more than one reader must be included in the study to properly characterize the test's accuracy. In this section we describe the main design issues and strategies for diagnostic accuracy studies. Note that the STARD group (Bossuyt 2015) has published a list of features (study design and analysis items) that serves to standardize the reporting of diagnostic test accuracy studies but is also a helpful aid during the planning phase.

3.1. Retrospective versus prospective

While prospective studies are often seen as the preferred study design, for assessing and comparing the accuracy of diagnostic tests, retrospective designs have many advantages. In a recent review by Shiraishi *et al* (2009) of the medical imaging literature, 75% of ROC studies used a retrospective design. Table 3 summarizes some of the advantages and disadvantages of these designs. Consider our transmission ultrasound example. In this retrospective design, 20 cases with solids and 20 with cysts were selected for the study in order to get a precise estimate of the ROC area. (The strategy of increasing the prevalence in the study sample to ensure enough cases with and without the condition is sometimes called **augmentation** (Zhou *et al* 2011).) Cases of solids and cysts were chosen to represent a broad spectrum of disease characteristics. (The strategy of increasing the prevalence of difficult or rare cases in the study sample is often referred to as **enrichment**.) In contrast, with a prospective design, many more than 40 subjects would need to be recruited and consented in order to achieve 20 cases with solids, since the prevalence of solids is much lower than the prevalence of cysts. Also, in a prospective design there is no oversight in the type of cases recruited, so there is no guarantee that there will be sufficient numbers of different types of solids (malignant versus benign solids) and different types of cysts (simple and complex).

3.2. Paired or unpaired designs

In ROC studies comparing two or more diagnostic tests, either a paired and unpaired design can be used. In an unpaired design, different subjects undergo each diagnostic test, whereas in a paired design the same subjects undergo all diagnostic tests under investigation. Paired designs are not only efficient in terms of reducing the number of subjects required for the study, but they ensure a fair comparison between tests because the same subjects undergo the tests. In fact, pairing is better than randomization. Over the long run, randomization makes two groups of subjects *similar*, but pairing *ensures* that two groups are the same. For these reasons, paired designs are extremely common in the diagnostic test literature.

3.3. Test results: quantitative, probability-based, or confidence ratings

Some diagnostic tests provide quantitative test results (e.g. SOS on ultrasound, or Hounsfield units on CT). Other tests (e.g. Framingham risk score) provide an estimated probability of disease from a statistical model that incorporates information from multiple factors or variables, each weighted according to their relative strength in

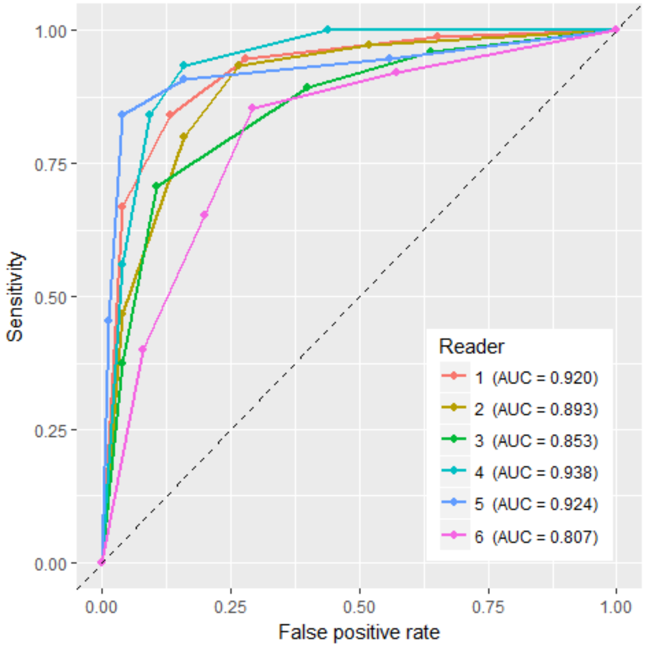


Figure 5. Hypothetical example of a diagnostic accuracy study with six readers. An ROC curve for each reader is plotted and the corresponding AUC is shown in the legend. The readers' mean AUC is 0.889.

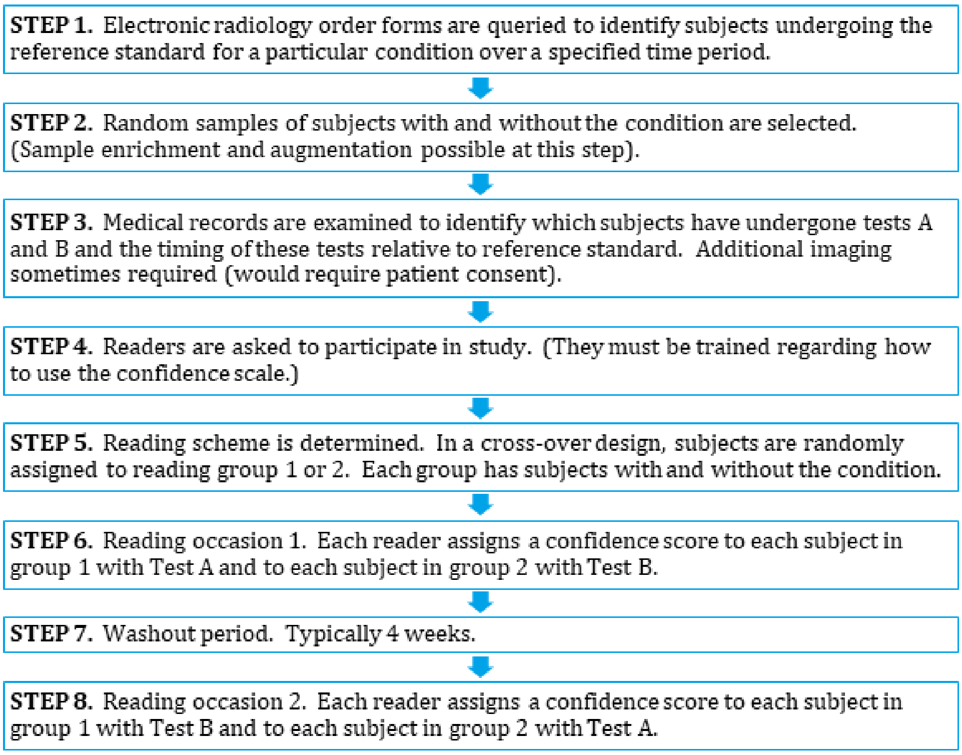
discriminating between different health states. Both quantitative diagnostic tests and probability risk scores lend easily to ROC analysis.

In contrast, in diagnostic radiology, where images must be interpreted by trained readers, the readers' subjective interpretation must be elicited in a fashion that is reproducible (at least to some degree, allowing for natural intra-reader variability) and adaptable to ROC analysis. A simple 'positive' and 'negative' will not allow for ROC analysis. There are two common approaches to eliciting readers' confidence: an ordinal scale and a semi-continuous probability scale. As described in section 1.4 in the transmission ultrasound study, a radiologist assigned a score from 1 to 5 to grade their confidence in the presence of a solid lesion. Probability-based confidence scores are also commonly used to indicate the probability that the condition is present (0% confidence to 100% confidence). Multiple investigators have compared these two approaches (Rockette *et al* 1992, Wagner *et al* 2001, Hadjiiski *et al* 2007) and found the probability scale to be preferable for ROC analysis in most situations because it provides more points on the ROC curve for better estimation.

There are several important study design issues when assessing diagnostic tests that require subjective interpretation by human readers. First, readers will inevitably assign confidence scores differently. For example, one reader may rarely use the highest scores (i.e. rarely assign probabilities >90%, or rarely use the 5th rating score), while another reader may spread his confidence more evenly across the given range. Their AUCs may be equivalent, but the readers' scores for individual subjects will not agree. For ROC analysis it's important that each reader use the scale in a consistent manner, but the readers do not need to use the scale in the same way. For this reason, confidence scores from multiple readers should never be pooled together to construct an ROC curve. Rather, a ROC curve should be constructed for each reader separately; then, a summary measure of accuracy, such as the AUC, can be estimated for each reader and averaged across readers. The 18-reader study by Goenka *et al* (2014) illustrates this nicely. Unfortunately, pooling or averaging scores is still quite common, i.e. in about 17% of studies in the imaging literature (Shiraishi *et al* 2009).

Second, readers have different perceptual and cognitive skills, manifesting in considerable inter-reader variability in their performance (i.e. not just in how they use the confidence scale). (See the large reader study by Beam *et al* (1996).) Some of this variability is attributable to observable factors (e.g. reader experience and training) but some pure random noise (Shiraishi *et al* 2009). Inter-reader variability, and even inter-institution variability, should be accounted for when characterizing a diagnostic test's AUC. Instead of a single AUC to describe the diagnostic test's accuracy, these tests' accuracy is better represented by a distribution of AUCs (see figure 5 for an example). Investigators often summarize the test's accuracy by an estimate of the mean AUC of a population of readers (see section 5.2 for additional details). There are parametric, semi-parametric, and nonparametric regression methods available for modeling reader variability (Zhou *et al* 2011).

Third, readers need training to use the confidence scale. Readers need to understand that it's important to spread their confidence across the scale so that an ROC analysis can be performed and that they need to use the scale in a consistent manner. A few authors have discussed training, but more work is needed in this area (Gur *et al* 1990, FDA 2012).



STEP 1. Electronic radiology order forms are queried to identify subjects undergoing the reference standard for a particular condition over a specified time period.

STEP 2. Random samples of subjects with and without the condition are selected. (Sample enrichment and augmentation possible at this step).

STEP 3. Medical records are examined to identify which subjects have undergone tests A and B and the timing of these tests relative to reference standard. Additional imaging sometimes required (would require patient consent).

STEP 4. Readers are asked to participate in study. (They must be trained regarding how to use the confidence scale.)

STEP 5. Reading scheme is determined. In a cross-over design, subjects are randomly assigned to reading group 1 or 2. Each group has subjects with and without the condition.

STEP 6. Reading occasion 1. Each reader assigns a confidence score to each subject in group 1 with Test A and to each subject in group 2 with Test B.

STEP 7. Washout period. Typically 4 weeks.

STEP 8. Reading occasion 2. Each reader assigns a confidence score to each subject in group 1 with Test B and to each subject in group 2 with Test A.

Figure 6. Typical workflow for a retrospective MRMC study.

Table 4. Traditional design for MRMC study.

	Reader 1		Reader 2			Reader j			Reader J	
	Test A	Test B	Test A	Test B		Test A	Test B		Test A	Test B
Subject 1	Y_{11_A}	Y_{11_B}	Y_{12_A}	Y_{12_B}	…	Y_{1j_A}	Y_{1j_B}	…	Y_{1J_A}	$\mathrm{Y}_{\mathrm{1J\_B}}$
Subject 2	Y_{21_A}	Y_{21_B}	Y_{22_A}	Y_{22_B}	…	Y_{2j_A}	Y_{2j_B}	…	Y_{2J_A}	$\mathrm{Y}_{\mathrm{2J\_B}}$
Subject i	Y_{i1_A}	$Y_{\mathrm{i1\_B}}$	Y_{i2_A}	Y_{i2_B}	…	Y_{ij_A}	Y_{ij_B}	…	$Y_{\mathrm{i}J\_A}$	$\mathrm{Y}_{\mathrm{iJ\_B}}$
	…	…	…	…	…	…	…	…	…	…
Subject N	Y_{N1_A}	Y_{N1_B}	Y_{N2_A}	Y_{N2_B}	…	Y_{Nj_A}	Y_{Nj_B}	…	Y_{NJ_A}	$\mathrm{Y}_{\mathrm{NJ\_B}}$

Each of N subjects is imaged with both diagnostic tests A and B. The J readers each interpret all of the images of the N subjects. Thus, each reader interprets $2 \times N$ images, for a total of $J \times 2 \times N$ image interpretations. Note that in addition to Tests A and B, all subjects also undergo the reference standard. Y_{ij_A} and Y_{ij_B} denote the confidence scores assigned by reader j for the ith subject, using test A and B, respectively.

Lastly, special study designs are needed for ROC studies of diagnostic tests that require subjective interpretation. These studies are often referred to as multi-reader multi-case (MRMC) studies, the purpose of which is to elicit unbiased and generalizeable confidence scores from a random sample of readers. These studies commonly include two to several hundred readers, with a typical reader size of about 6 (Dendumrongsup *et al* 2014). (In section 3.5, we discuss sample size (in terms of both subjects and readers) in MRMC studies.) The most common MRMC design (used in more than 90% of medical imaging MRMC studies according to one review (Shiraishi *et al* 2009)) is the paired-reader, paired-case design, where a sample of subjects undergoes two or more competing tests, and a sample of readers interprets all of the tests for all subjects. Figure 6 illustrates a typical study workflow for this type of MRMC design, while table 4 illustrates the layout of confidence scores (i.e. the scores assigned by the J readers) to be used in the statistical analysis. Note that a typical time interval between readings of the two modalities is about 4 weeks (Dendumrongsup *et al* 2014) though the ideal interval for a particular study depends on a number of factors including the complexity and number of cases being read. The primary outcome for such a study is the difference in the readers' mean AUC of the two tests. This design is fairly efficient in terms of the number of readers and subjects that are required; however, a split-plot design can be more efficient (Obuchowski 2009). In the split-plot design the readers are split into small groups; each group interprets the test results of a different sample of subjects. This design is more efficient than the fully-paired design in table 4 because it reduces the between-reader correlation in the test results, thereby reducing the total number of interpretations required of each reader. Other MRMC designs include unpaired reader designs (for situations where readers are not equally skilled at interpreting competing tests) and unpaired case designs (for situations where subjects are not able to undergo all of the competing tests, e.g. when the tests are invasive) (Obuchowski 1995).

3.4. Common biases in ROC studies

We have recognized diagnostic test accuracy studies as being one of the more difficult studies to design. The consequences of poorly designed studies can be under-estimation of tests' accuracy, but more commonly, poorly designed studies lead to over-estimation of tests' accuracy. This can lead to inappropriate optimism, especially when the test is applied in real clinical decision making. Zhou *et al* (2011) list 13 common biases in diagnostic test accuracy studies. They include biases associated with selecting subjects and readers for a study, biases associated with the reference standard, and biases associated with interpretation of test results. In this section we present two common biases that are still often overlooked by investigators during the design and analysis phases of their study.

3.4.1. Verification bias

Verification bias is probably the most common bias in diagnostic accuracy studies. It is so prominent because it is counter-intuitive and difficult to avoid. While it is important to validate subjects' true condition status by means of a nearly perfect reference standard, many investigators take the step of omitting subjects who do not have meticulous validation. Often the omitted subjects are not a random sample of subjects, leading to selection bias. Verification bias is a type of selection bias which typically occurs when the results of the diagnostic test under investigation influence whether or not the reference test should be performed. Estimates of test accuracy can be seriously skewed, usually overestimating sensitivity and underestimating specificity, when the study sample is limited to the subjects who have undergone both the diagnostic test under investigation and the reference test. The effect on summary measures from the ROC curve depends on the severity of these over- and under-estimations.

We use the transmission ultrasound study as an illustration. Suppose subjects undergo breast cancer screening with mammography. Subjects with a positive finding might then be referred to transmission ultrasound for a diagnostic work-up to determine if the subject should undergo biopsy. Suppose we want to investigate the accuracy of transmission ultrasound; biopsy is the reference standard. For simplicity, we consider three possibilities:

- If the results of the transmission ultrasound suggest a simply cyst, the patient is unlikely to undergo biopsy and thus will not be included in a study estimating the accuracy of ultrasound. We set the probability of undergoing a biopsy for this group of subjects at X_1 (where X_1 is low, perhaps 5%).
- If the results are equivocal for a solid (potentially malignant) lesion, the patient may undergo additional imaging (e.g. MRI) before a decision is made about biopsy. We arbitrarily set the probability of undergoing a biopsy for this group of subjects at X_2 (where X_2 is moderate, perhaps 50%).
- If the results are very suspicious for a solid lesion, the patient is very likely to undergo biopsy and thus will be included in the study. We arbitrarily set the probability of undergoing a biopsy for this group of subjects at X_3 (where X_3 is high, perhaps 95%).

Figure 7 illustrates the effect on sensitivity, specificity, and the AUC for various values of X_1, X_2, and X_3. Starting with the situation where X_1, X_2, and X_3 all equal 100% (i.e. no verification bias), we see that sensitivity $= 0.90$, specificity $= 0.80$, and the AUC $= 0.885$. If we introduce a little verification bias (i.e. $X_1 = 50\%$ $X_2 = 80\%$ and $X_3 = 90\%$), then the AUC is underestimated by 1.2%. If we introduce more verification bias (i.e. $X_1 = 10\%$, $X_2 = 50\%$ and $X_3 = 90\%$), then the AUC is underestimated by 13.3%.

There are several strategies to correct for the effect of verification bias. We have summarized these strategies in table 5; the first two address study design modifications to avoid the bias (by creating a situation in which truly positive and truly negative cases are equally likely to be verified) and the last option offers statistical corrections for the bias. At a minimum, in the Discussion of their paper, investigators conducting diagnostic test accuracy studies should address the likelihood of verification bias in their study, its potential effect on estimates of accuracy, and interpretation of the study results in light of the bias.

3.4.2. Location bias

Location bias occurs in situations where multiple abnormalities are possible in the same subject (e.g. multiple malignant lesions) and the diagnostic test is used to locate any and all suspicious lesions. This occurs commonly for diagnostic tests that require subjective interpretation (e.g. a diagnostic image where multiple lesions can be detected by a radiologist) but it is also possible in studies of quantitative tests. There are two types of location bias (McGowan *et al* 2016): type I bias, where the true lesion is missed and another area in the same subject is identified, and type II bias, where the true lesion is located but another area in the same subject is identified and scored as more suspicious. Again we use the transmission ultrasound example to illustrate the problem.

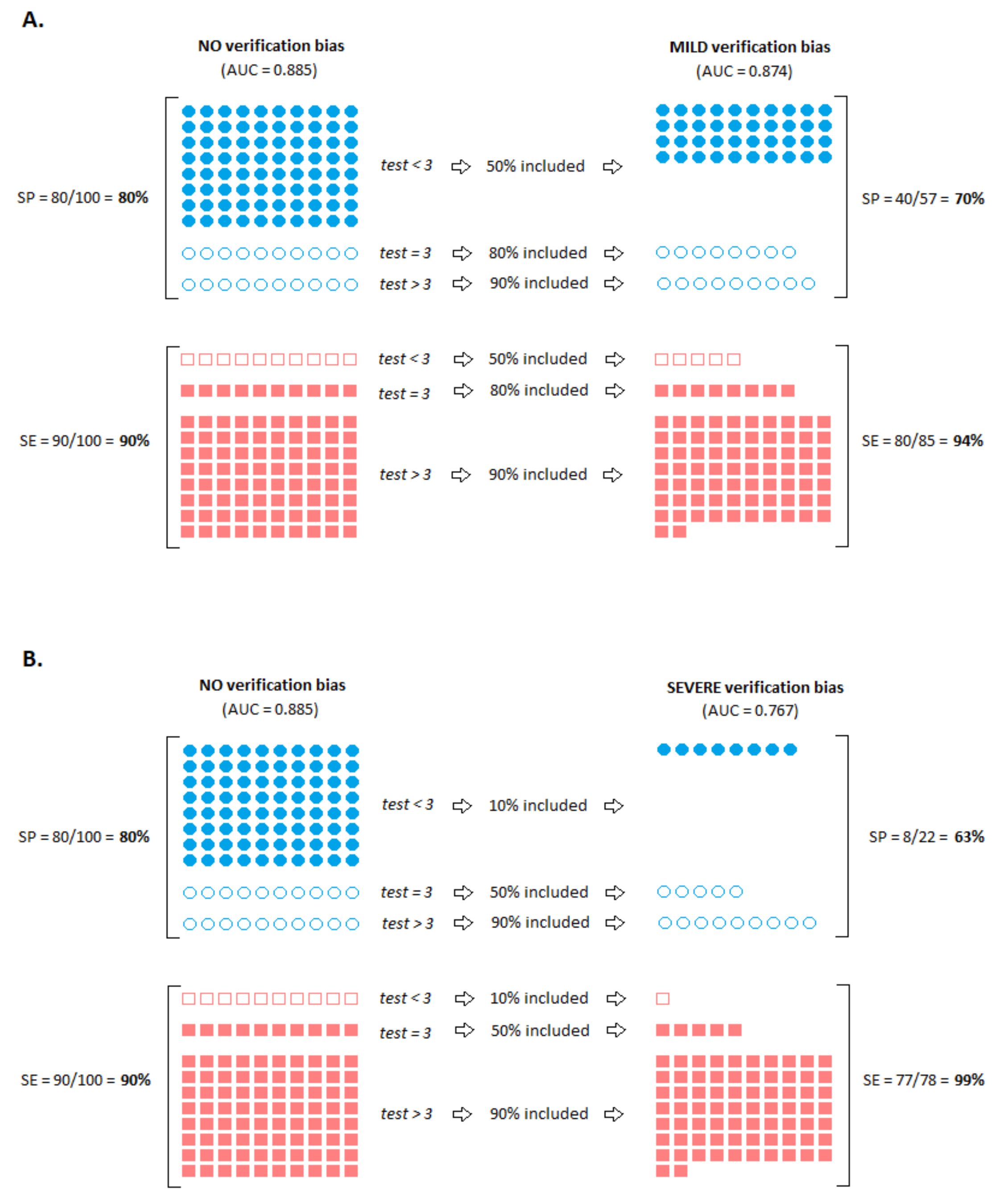


Figure 7. A hypothetical study of 100 cysts (blue circles) and 100 solids (red squares) is depicted, where filled in circles represent true negatives and filled in squares represent true positives. Verification bias is introduced and causes underestimation of specificity, overestimation of sensitivity, and underestimation of the AUC. The amount depends on the severity of the verification bias (A = mild verification bias, B = severe verification bias).

Table 5. Strategies to correct for verification bias.

1. *Recruit and consent subjects before testing.* Use a protocol whereby all subjects consent to undergo both the test(s) under investigation and the reference standard
2. *Use two reference standards*, one for patients with positive test results (often invasive) and one for patients with negative test results (often involving follow-up). The two reference standards should be equally valid
3. *Apply one of several available statistical corrections* to adjust estimates of test accuracy for verification bias (Begg and Greenes 1983, Greenes and Begg 1985, Zhou 1996, 1998, Toledano and Gatsonis 1999, Rodenberg and Zhou 2000, Zheng *et al* 2005, Alonzo and Pepe 2005, Rotnitzky *et al* 2006, He *et al* 2009, Liu and Zhou 2011, Zhou *et al* 2011). There are a number of factors to consider when selecting a correction method, including whether the data for the unverified patients can be considered missing-at-random and what type of data has been collected (i.e. binary or ordinal, correlated or uncorrelated, etc)

Suppose the transmission ultrasound is used for screening for breast cancer in asymptomatic subjects. In type I location bias the radiologist fails to find the true cancerous lesion but identifies a false lesion, assigning it a confidence score of 80% (figure 8(a)). In type II location bias the radiologist identifies two lesions: the true lesion,

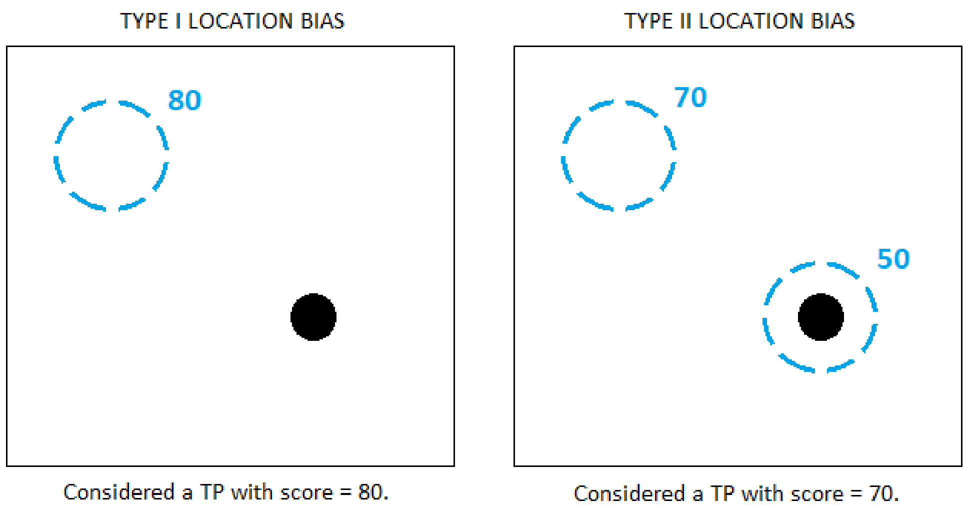

TYPE I LOCATION BIAS
80
Considered a TP with score = 80.
TYPE II LOCATION BIAS
70
50
Considered a TP with score = 70.

Figure 8. The panel on the left illustrates Type I location bias, where the reader missed the true lesion (the black dot), and identified a false positive (the blue dotted circle), assigning it a confidence score of 80%. The reader is credited with a true positive. The panel on the right illustrates Type II location bias, where the reader detected the true lesion, but also identified a false positive, assigning it greater confidence than the true lesion. A confidence score of 70% is used in the estimation of test sensitivity.

assigning it a score of 50%, and a false lesion, assigning it 70% (figure 8(b)). If the location bias is ignored, the highest confidence score assigned to the subject is used in the ROC analysis, i.e. 80% in figure 8(a) and 70% in figure 8(b). In both situations, the diagnostic test is awarded higher sensitivity than is appropriate (i.e. credited for a true positive, when the actual confidence score is associated with a false positive finding). This can lead to over-estimation of test accuracy and incorrect comparisons of diagnostic tests' accuracies (e.g. when one test has more location bias than another). McGowan *et al* (2016) illustrate the problem with several real examples.

There are two statistical methods to correct for location bias. The ROI-ROC method (Zhou *et al* 2011) requires specifying regions of interest (ROIs); then the ROIs are treated as the unit of analysis instead of the subject. For example, in a breast cancer study, each breast would be considered an ROI. Estimation of the AUC would take into account the correlation between the two breasts from the same subject (Obuchowski 1997). The FROC approach (Zou *et al* 2011) treats lesions as the unit of analysis, thus offering a more refined analysis of each confidence score assigned by the radiologist. To avoid location bias, the two methods utilize the scores on individual lesions rather than relying on the highest confidence score assigned to a subject. Specifically, for subjects with a true lesion, both methods use the confidence score assigned to the actual true lesion when constructing the ROC curve; if the true lesion was not located by the reader, they use the default confidence score of 0% (i.e. false negative). For subjects without a true lesion, the ROI-ROC method uses the highest confidence score assigned to the ROI (e.g. the highest confidence score assigned to the breast), whereas the FROC approach accounts for the number of false positive findings and plots the number of false positive findings per subject. Both approaches provide bias-free estimates of accuracy, though with slightly different accuracy metrics (McGowan *et al* 2016). The typical accuracy metric for the ROI-ROC method is the area under the standard ROC curve. In the FROC approach, the area under the alternative FROC (AFROC) curve is often used. The AFROC curve is a plot of the lesion localization fraction (the number of correctly located lesions divided by the total number of lesions) versus the non-lesion localization fraction (the number of non-lesion localizations divided by the total number of subjects).

3.5. Sample size determination for ROC studies

As with any research study there are four basic steps to determining the appropriate sample size for an ROC study:

(1) Specify the study objective including study design, study populations, and diagnostic test accuracy endpoint. Most ROC studies determine the required sample size based on summary measures of the ROC curve, such as the AUC.

(2) Determine the statistical hypotheses (i.e. null and alternative hypotheses) and corresponding statistical analysis plan. Note that some studies are not designed to test hypotheses but rather to measure test accuracy; in these studies we must determine the appropriate sample size for constructing a confidence interval (CI) for diagnostic accuracy.

(3) Identify known parameters or plausible ranges for unknown parameters. The parameters needed for sample size calculation vary depending on the study objectives, but usually we need some sense of the magnitude of the AUC, the ratio of subjects without to with the condition, and the precision needed in the study in terms of the width of the CI or the detectable difference in accuracy of competing tests.

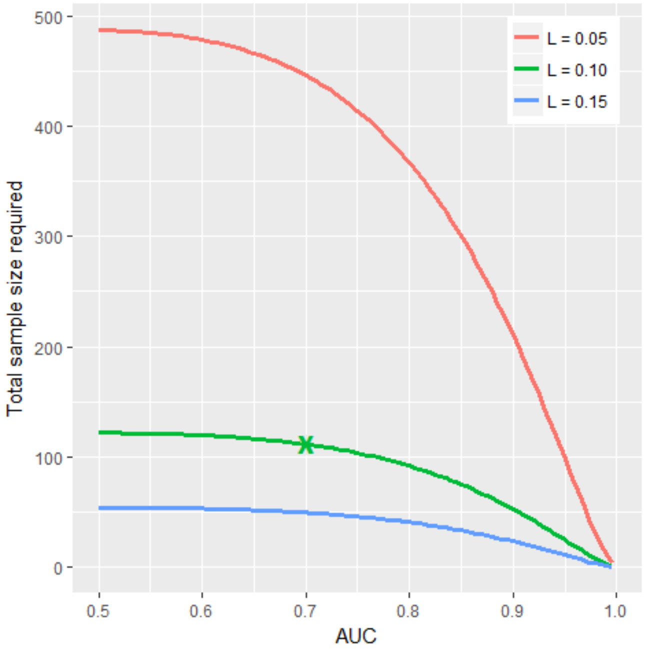


Figure 9. The total sample size required to estimate the AUC as a function of the magnitude of the AUC. L is the desired half width for a 95% confidence interval for the AUC. An equal number of subjects with and without the condition was assumed. The sample size under the specific conditions discussed in the text is marked by a green '*x*'.

(4) Calculate the required sample size. There are a variety of sample size methods for ROC studies, depending on the study objectives. Zhou *et al* (2011) include a chapter on sample size calculation that includes not only the most commonly used methods (which we illustrate here), but also methods for the partial area under the curve, the sensitivity at a fixed FPR, clustered data (i.e. multiple observations from the same patient), non-inferiority hypotheses, and sample size for studies to determine a suitable cut-point on the ROC curve.

We use the transmission ultrasound study to illustrate sample size calculation. Suppose we want to determine the sample size needed to estimate the AUC of the SOS measurements. Our study objective is to estimate the AUC and construct a 95% CI for the AUC. We will plan a retrospective study. For sample size calculation, we need to know (1) the ratio of subjects with cysts to subjects with solids in the study sample, and (2) the magnitude of the AUC. Since this is a retrospective design we can set the ratio of cysts to solids at 1:1, which is the most efficient design. We will consider a range for the AUCs of 0.60–0.90.

We use the following formula to estimate sample size.

$$n_1 = \frac{\left[(z_{\alpha/2})\sqrt{V}\ \right]^2}{L^2}$$

where V is the variance function for the AUC, given by $V = (0.0099\ \ \times \mathrm{e}^{-a^2/2}) \times ([5a^2 + 8] + [a^2 + 8]/K)$, and K is the ratio of non-diseased to diseased subjects (for our example we need the ratio of cysts to solids). a is a parameter of the (unknown) underlying distribution for the confidence scores; it is directly related to the magnitude of the AUC. A binormal distribution (i.e. two overlapping Gaussian distributions, one for subjects without the condition, and one for subjects with the condition) is often assumed for sample size calculation. For AUCs of 0.6, 0.7, 0.8, and 0.9, the parameter a takes on the values of 0.36, 0.74, 1.19, and 1.82 (assuming similar standard deviations for the two distributions) (Zhou *et al* 2011). L is the half-width of the desired 95% CI and controls the precision (tightness) of the CI.

Figure 9 illustrates the total sample size required $(n_1 + n_0)$ as a function of the magnitude of the AUC and the width of the 95% CI. Note that the sample size requirement decreases as the AUC increases and/or the width of the CI increases (i.e. lower precision). Since the AUC is unknown, choosing an AUC in the lower range of plausible values is advised so that the sample size is not too small for attaining the study's objective. For example, let us say we assume an AUC of 0.70 and our objective is to construct a 95% CI for the AUC with a width of 0.20 (i.e. $L = 0.10$). We have already decided that we will have an equal number of subjects with and without the condition (i.e. $K = 1$). Under these conditions, $a = 0.742$, $V = 0.145$, and $z_{\alpha/2} = 1.960$. Plugging these values into the equation above for n_1 yields $n_1 = 56$, suggesting that we would need 56 subjects with the condition (112 subject

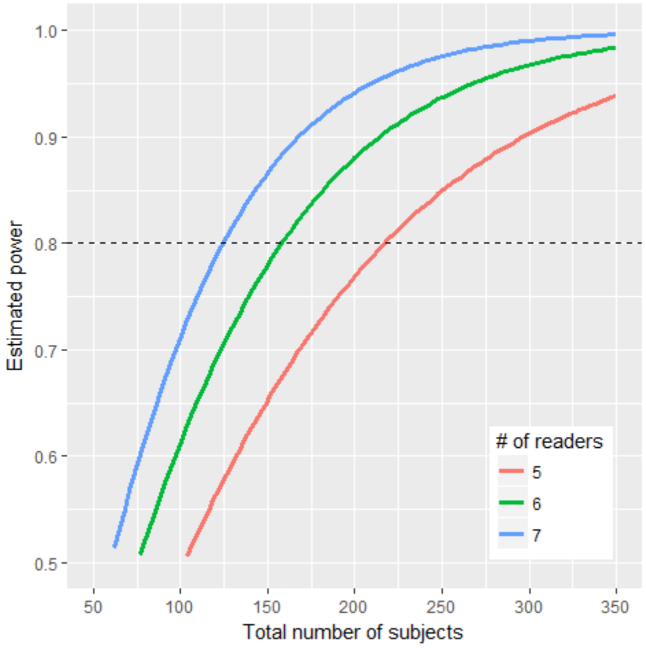


Figure 10. Estimated power to detect a difference of 0.05 in the readers' mean AUC for various subject and reader sample sizes using a traditional, paired-reader paired-case, design. Power is closest to but greater than 0.8 with 218 subjects and five readers, 159 subjects and six readers, or 125 subjects and seven readers (Assumptions based in part on a pilot study with forty subjects and one reader and in part on conjectured parameter values.).

total) to achieve our objective. The sample size under these particular assumptions is marked with a green '*x*' in figure 9.

Recognizing that a study with one or even a few readers cannot adequately describe the distribution of reader performance with transmission ultrasound, as a second example, suppose we want to conduct a MRMC study to compare the mean AUC of radiologists interpreting transmission ultrasound images versus the mean AUC of radiologists interpreting hand-held ultrasound (HHUS) images. The null and alternative hypotheses are

$$H_o : \mathrm{AUC}_{\mathrm{transmission}} = \mathrm{AUC}_{\mathrm{HHUS}} \text{ versus } H_A : \mathrm{AUC}_{\mathrm{transmission}} \neq \mathrm{AUC}_{\mathrm{HHUS}}.$$

We will plan a retrospective study. For sample size calculation for studies comparing AUCs, we need to know (1) the ratio of subjects with cysts to subject with solids in the study sample, (2) the magnitude of the AUCs under the null hypothesis, and (3) the expected difference in mean AUCs under the alternative hypothesis. We again set the ratio of cysts to solids at 1:1. For a conservative sample size we assume the mean AUC of 0.60 under the null hypothesis. We will determine sample size to detect a mean difference in AUCs of 0.05.

We apply the MRMC sample size method of Hillis *et al* (2011). The calculations (not shown here) are more extensive than for our first example, but there are published sample size tables for easy reference for investigators to use (Obuchowski 2004, 2011). Figure 10 illustrates the trade-off between the number of subjects (*N*) and number of readers (*J*). Observe that a study with 218 total subjects and 5 readers provides equivalent power as a study with 125 subjects and 7 readers, allowing investigators to choose a design that suits their available resources.

Finally, we note that diagnostic accuracy studies are often underpowered because of limited availability of subjects and/or readers, or due to the costs of these studies. Not surprisingly, there is a large literature base on statistical methods for combining and synthesizing results from multiple ROC studies. Zhou *et al* (2011) reviewed many excellent statistical methods for meta-analyses, and there are evidence-based guidelines (i.e. PRISMA (Moher *et al* 2009)) for conducting and reporting meta-analyses of diagnostic test accuracy (Irwig *et al* 1994, Leeflang *et al* 2008, The Cochrane Collaboration 2013).

4. Estimating ROC curve and associated summary measures

In this section we review methods for constructing an ROC curve and estimating its associated summary indices. We begin with the nonparametric methods as they are simple, require few assumptions, and are commonly used in the literature. We will then discuss parametric and semi-parametric methods, along with the relative advantages and disadvantages of these methods. We then present some special cases of ROC analysis.

Table 6. Sensitivity and false positive rate for each cut point, *k*, in the radiologist's subjective interpretations from the transmission ultrasound example. In this study, there were 20 truly positive cases (solids) and 20 truly negative cases (cysts) (i.e. according to the reference standard).

k	Sensitivity	False positive rate
1	0.85 (17/20)	0.50 (10/20)
2	0.80 (16/20)	0.40 (8/20)
3	0.80 (16/20)	0.30 (6/20)
4	0.75 (15/20)	0.30 (6/20)

4.1. Non-parametric methods

In section 2.1 we used the quantitative transmission ultrasound data to construct empirical ROC curves of a radiologist interpreting the images and of the SOS quantitative measurements. The scores assigned by the radiologist were illustrated and tabulated in figure 2(a). To construct these non-parametric ROC curves, we must first estimate the sensitivity and false positive rate at each possible cut-point. Since the radiologist used a *K*-point ordinal scale, there were $K - 1$ cut-points (here, the 5-point scale resulted in 4 cut-points: >1, >2, >3, and >4). To estimate sensitivity at a cut-point *k*, we must assume that scores greater than *k* are positive, and that scores $\leqslant k$ are negative. Sensitivity at cut-point *k* is the proportion of cases scored as positive among all cases with the condition, and the false positive rate at cut-point *k* is the proportion of cases scored as positive among all cases without the condition. Table 6 illustrates these calculations for the radiologist interpreting the quantitative transmission ultrasound images.

Figure 3 (blue curve) illustrated the resulting non-parametric ROC curve (aka 'empirical ROC curve') from the coordinates calculated in table 6. Note the long line segment from the last false positive rate (at the >1 cut-point) to the (1,1) point at the top right corner of the plot. This line segment extrapolates beyond the last known coordinate to the (1,1) point. If there is a lot of extrapolation used to construct the ROC curve, we might have less confidence in the curve, especially in the upper right section of the curve. The degree of extrapolation should be clear when presenting an ROC curve, but it is often overlooked (Dendumrongsup *et al* 2014). Illustrating the coordinates underlying the ROC curve, as indicated by the circles in figure 3, is probably the best approach for describing the degree of extrapolation.

In equation (1) we presented a simple formula for estimating the AUC non-parametrically; however, our estimate of the AUC needs a confidence interval (CI) to characterize the precision of our estimate. Calculation of the standard error of the AUC and construction of its CI are as follows (Delong *et al* 1988). For each subject *i* with the condition we calculate their T_1 variance component, and for each subject *j* without the condition, we calculate their T_0 variance component:

$$T_{1i} = \frac{1}{n_0}\sum_{j=1}^{n_0}\Psi \qquad T_{0j} = \frac{1}{n_1}\sum\nolimits_{i=1}^{n_1}\Psi \tag{2}$$

In words, for the *i*th subject with the condition, we compare their score to each of the n_0 subjects without the condition. Similarly, for the *j*th subject without the condition we compare their score to each of the n_1 subjects with the condition. As in equation (1), ψ takes on one of three values:

$$\Psi = \left[\begin{array}{l} 0,\text{ if subject with condition is rated lower than subject without the condition} \\ \frac{1}{2},\text{ if subject with condition is rated the same as subject without the condition} \\ 1,\text{ if subject with condition is rated higher than subject without the condition} \end{array}\right.$$

Next we need to calculate the variance estimates for T_{1i} and T_{0j}. The variance estimates are:

$$S_1 = \frac{1}{(n_1 - 1)}\sum_{i=1}^{n_1}\left(T_{1i} - \widehat{\mathrm{AUC}}\right)^2 \qquad S_0 = \frac{1}{(n_0 - 1)}\sum_{j=1}^{n_0}\left(T_{0j} - \widehat{\mathrm{AUC}}\right)^2 \tag{3}$$

where $\widehat{\mathrm{AUC}}$ is from equation (1), or can be calculated simply as

$$\widehat{\mathrm{AUC}} = \sum_{i=1}^{n_1} T_{1i}/n_1 \;= \sum_{j=1}^{n_0} T_{0j}/n_0.$$

Finally, the estimate of the variance of the AUC is

$$\hat{\mathrm{Var}}\left(\widehat{\mathrm{AUC}}\right) = \frac{1}{n_1}S_1 + \frac{1}{n_0}S_0. \tag{4}$$

Table 7. Illustration of the steps in calculating a 95% CI for the AUC using the radiologist's subjective interpretations in the transmission ultrasound example. In this study, there were 20 truly positive subjects (i.e. subjects with a solid) and 20 truly negative subjects (i.e. subjects with a cyst). The AUC was 0.738.

Neg. subject	T_{0j}	Neg. subject	T_{0j}	Pos. subject	T_{1i}	Pos. Subject	T_{1i}
S1	0.83	S11	0.38	S21	0.85	S31	0.85
S2	0.38	S12	0.38	S22	0.70	S32	0.85
S3	0.80	S13	0.83	S23	0.85	S33	0.85
S4	0.93	S14	0.38	S24	0.85	S34	0.85
S5	0.93	S15	0.93	S25	0.85	S35	0.85
S6	0.80	S16	0.93	S26	0.85	S36	0.85
S7	0.38	S17	0.93	S27	0.25	S37	0.85
S8	0.93	S18	0.93	S28	0.85	S38	0.25
S9	0.93	S19	0.93	S29	0.85	S39	0.85
S10	0.38	S20	0.93	S30	0.55	S40	0.25
$S_0 = 0.061$				$S_1 = 0.049$			
$\hat{\mathrm{Var}}\left(\widehat{\mathrm{AUC}}\right) = 0.006$							
95% CI for AUC: (0.592, 0.883)							

Table 8. Illustration of the steps in calculating the partial AUC between two FPRs using the radiologist's subjective interpretations in the transmission ultrasound example.

C_1	3	$\widehat{\mathrm{AUC}}_{\mathrm{FPR}_1-\mathrm{FPR}_2}$	0.080
C_2	4	MAX	0.100
FPR_1	0.40	MIN	0.035
FPR_2	0.30	$\widehat{\mathrm{AUC}}^*_{\mathrm{FPR}_1-\mathrm{FPR}_2}$	0.846

The 95% CI for the AUC is $[\widehat{\mathrm{AUC}} - 1.96 \times \sqrt{\hat{\mathrm{Var}}(\widehat{\mathrm{AUC}})},\ \widehat{\mathrm{AUC}} + 1.96 \times \sqrt{\hat{\mathrm{Var}}(\widehat{\mathrm{AUC}})}]$. Note that these calculations are illustrated in table 7 using our transmission ultrasound example. Commonly used statistical software packages have incorporated these calculations; see section 6. It's important to note that the Delong *et al* variance estimates have been shown to be slightly biased (i.e. too small) for small and moderate sample size studies (Zou *et al* 2011, Tcheuko *et al* 2016); a two-way random-effects ANOVA method, based on U-statistics, should be used as an alternative, especially for studies with small sample sizes.

In the next section we will see that estimation of the partial area under the ROC curve and the sensitivity at a fixed false positive rate is straightforward with the parametric approaches. The reason is that in the parametric approaches we assume a specific distribution underlying the data, so we can extrapolate the partial area or sensitivity for any false positive rate points. In contrast, with the non-parametric method, without making any assumptions underlying the data, we are restricted to the observed operating points (though Dodd and Pepe have suggested a linear extrapolation method (Dodd and Pepe 2003a)). For example, in table 6 we observed that at the cut-points of >2 and >3, the operating points were ($\mathrm{FPR}_1 = 0.40$, sensitivity $= 0.80$) and ($\mathrm{FPR}_2 = 0.30$, sensitivity $= 0.80$). We will now discuss how to estimate the partial area under the ROC curve between FPR_1 and FPR_2 non-parametrically.

Let C_1 and C_2 denote the cut-points associated with FPR_1 and FPR_2, respectively. The non-parametric estimate of the partial area under the ROC curve between FPR_1 and FPR_2 is (Zhang *et al* 2002, He and Escobar 2008):

$$\widehat{\mathrm{AUC}}_{\mathrm{FPR}_1-\mathrm{FPR}_2} = \frac{1}{n_1 n_0} \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \Psi',$$

where Ψ' is defined as in equation (1) except that it is restricted to subjects without the condition whose test score is contained in the FPR interval (i.e. between C_1 and C_2). Note that the area under the full ROC curve is bounded by 0.5 and 1.0. The partial area under the ROC curve also has bounds, as follows (McClish 1989):

$$\mathrm{MAX} = \ \mathrm{FPR}_1 - \mathrm{FPR}_2$$

$$\mathrm{MIN} = \frac{1}{2}(\mathrm{FPR}_1 - \mathrm{FPR}_2)(\mathrm{FPR}_1 + \mathrm{FPR}_2).$$

These bounds can be used to help interpret the partial area. McClish (1989) suggested a transformation such that the partial area can be interpreted on the same scale as the full area under the curve:

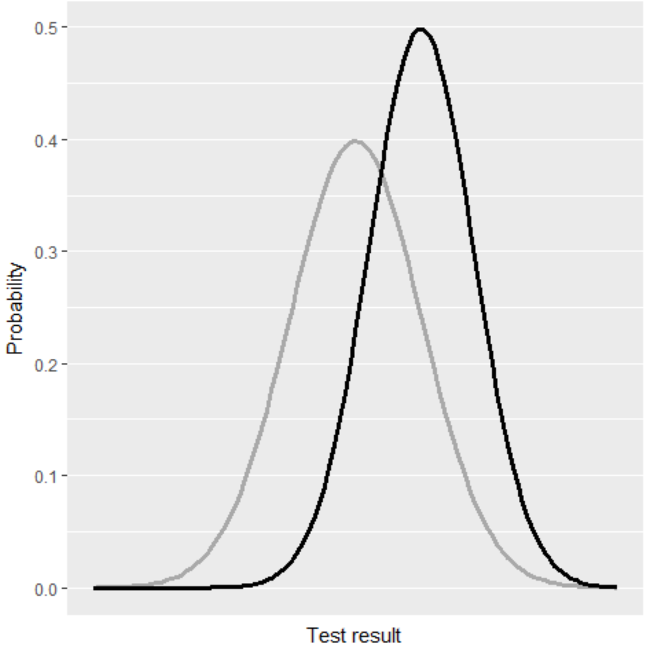


Figure 11. Hypothetical distribution of test results for subjects with the condition (black curve) and without the condition (gray curve).

$$\widehat{\mathrm{AUC}}^{*}_{\mathrm{FPR}_1-\mathrm{FPR}_2} = \frac{1}{2}[1 + \frac{(\widehat{\mathrm{AUC}}_{\mathrm{FPR}_1-\mathrm{FPR}_2} - \mathrm{MIN})}{(\mathrm{MAX} - \mathrm{MIN})}].$$

See table 8 for illustration of these calculations with the transmission ultrasound example.

4.2. Parametric methods

The empirical ROC curves in figure 3 are jagged because the curves drawn between the observed points are crudely fit; this is particularly true of the blue curve drawn from the ordinal scores assigned by the radiologist. While the jagged curve is unattractive, a bigger issue is the underestimation of the area under the ROC curve. The empirical curve is not concave between the observed points and thus the area under the curve will be less than the area under a concave curve. To fit a smooth concave curve, there are two main approaches: (1) assume that the test results follow a bi-normal distribution (one distribution for those with the condition and one distribution for those without the condition), and (2) assume that the test results are monotonically related to a (latent) variable that follows a bi-normal distribution. The former assumption (i.e. 'parametric') is strong and rarely used (Metz 2008). The latter assumption ('semi-parametric') is quite flexible and commonly used (Metz 2008); we describe it here.

We hypothesize a model that describes the test results, or more specifically a latent variable monotonically related to the test results, as a function of the reference test results. Since there are two truth states (i.e. without the condition and with the condition), there are two, usually overlapping, normal distributions (i.e. the bi-normal distribution). The test results are hypothesized to be ordinal-scale, spanning the bi-normal distribution (see figure 11). We can completely describe the distribution by two parameters: a and b:

$$a = (\mu_1 - \mu_0)/\sigma_0 \text{ and } b = \sigma_1/\sigma_0,$$

where test results from subjects without the condition (or a transformation of the test results) come from a normal distribution with mean μ_0 and variance σ_0^2, and similarly test results from subjects with the condition (or a transformation of the test results) come from a normal distribution with mean μ_1 and variance σ_1^2. A smooth ROC curve is then given by

$$\mathrm{ROC}(t) = \Phi(a + b\Phi^{-1}(t)) \quad (5)$$

where Φ is the cumulative normal distribution. It is important to note that the binormal ROC curve in equation (5) represents a functional form for the curve and not the form of the underlying distribution of the test results. The test results themselves do not need to, and usually will not, follow a bi-normal distribution (Dorfman and Alf 1969, Metz 1984, Metz *et al* 1998, Hanley 1988, Zhou *et al* 2011).

There are multiple software packages available for fitting and estimating functional forms for the ROC curve and its parameters (see section 6). Given estimates of the parameters, we can then estimate the area under the fit-

Table 9. Parametric estimates of the AUC, partial AUC (FPR between 0.3 and 0.4), and sensitivity at FPR of 0.30 from the transmission ultrasound example (for the radiologist's subjective interpretations).

$\widehat{\text{AUC}}$	0.805
$\widehat{\text{AUC}}_{0.3\leqslant\text{FPR}\leqslant 0.4}$	0.078
$\widehat{\text{Sensitivity}}_{\text{FPR}=0.3}$	0.752

ted ROC curve (equation (6)), the sensitivity at a fixed FPR (equation (7)), and the partial area under the curve (equation (8)). Note that the ROC software packages also provide estimates of the standard errors and CIs, which often require iterative algorithms.

$$\widehat{\text{AUC}} = \Phi(\hat{a}/\sqrt{1+\hat{b}^2}) \tag{6}$$

$$\widehat{\text{Sensitivity}}_{\text{FPR}=e} = \Phi(\hat{a}+\hat{b}\Phi^{-1}(e)) \tag{7}$$

$$\widehat{\text{AUC}}_{e_1\leqslant\text{FPR}\leqslant e_2} = \int_{e_1}^{e_2} \Phi(\hat{a}+\hat{b}\Phi^{-1}(t))\text{d}t. \tag{8}$$

Table 9 provides estimates of the area, partial area, and sensitivity at fixed FPRs for our transmission ultrasound example.

The main advantage of using parametric or semi-parametric approaches is that they fit a smooth ROC curve from which the AUC and partial AUC can be easily estimated. These approaches require some assumptions about the distribution of the test results and/or the ROC curve. Additionally, parametric estimates are typically generated algorithmically and the computer programs that execute these algorithms will occasionally not converge (often due to a small sample size). Nonparametric approaches, on the other hand, do not require distributional assumptions and involve simple equations that can be solved analytically. However, the main disadvantage of these approaches is that they produce rough or jagged curves. Nonparametric estimates of the AUC also tend to underestimate the true AUC. When deciding whether to take a parametric or nonparametric approach, one should consider the ROC metrics of interest in the study, the computational resources available, and whether distributional assumptions can reasonably be made.

4.3. Clustered data

Clustered data are quite common in the medical literature according to a review by Dendumrongsup *et al* (2014), where they found that almost 50% of studies involved measurements on more than one organ, organ segment, or lesion in the same subject. For example, in the transmission ultrasound study, each subject has two breasts. If each breast is interpreted and scored, then we have clustered data, i.e. multiple correlated observations from the same subject. These organs/segments/lesions, clustered within the same subject, become the unit of analysis for estimating and comparing ROC curves.

A naïve approach to clustered data would be to treat each observation (e.g. each breast) as an independent observation, i.e. as if each breast came from a different patient. Unfortunately, observations within the same subject are usually positively correlated, at least to some degree. Consequently, this naïve approach will lead to standard errors that are too small, CIs that are too narrow, and p-values that are misleading (Obuchowski 1997). Another approach sometimes used by investigators is to combine or synthesize all of the findings within a subject into a single measurement (i.e. a subject-level score) and construct an ROC curve from the single subject-level score. For example, one might take the highest confidence score assigned to all findings in a subject. The limitations of this approach are (1) the synthesized single score may not be clinically relevant, (2) there is no correction for the location of findings (i.e. see Location bias in section 3.4.2), and (3) there is a loss of statistical power. In this subsection, we describe and cite several alternative approaches to handling clustered data.

One simple non-parametric approach utilizes the variance components in equations (2) and (3) (Obuchowski 1997). Let X_{ik} denote the test result on the k-th unit with the condition for subject i and Y_{ik} denote the test result on the kth unit without the condition for subject i, where n_{1i} and n_{0j} denote the number of units with and without the condition in subject i. For example, in the transmission ultrasound study, where the unit is a breast, for a subject without any solid lesions: $n_{1i}=0$ and $n_{0j}=2$; for a subject with a solid lesion in one breast: $n_{1i}=1$ and $n_{0j}=1$. The total number of units with the condition is n_1 and the total number of units without the condition is n_0. The total number of subjects with at least one unit with the condition (e.g. total number of subjects with solid lesions) is denoted I_1 and the total number of subjects with at least one unit without the condition is I_0.

We then calculate the T_1 variance component for each unit with the condition and the T_0 variance component for each unit without the condition as:

Table 10. Illustration of the steps in calculating a 95% CI for the AUC with clustered data. In the transmission ultrasound example there was only one breast per subject in the sample. To illustrate a clustered data set, we have added fictitious data for the bilateral breast for half of the subjects. Now, there are 31 truly positive units (i.e. breasts with a solid) and 29 truly negative units (i.e. breasts with a cyst). The breast-level AUC is 0.669.

Subject	T_{0i}	T_{1i}	Subject	T_{0i}	T_{1i}	Subject	T_{0i}	T_{1i}	Subject	T_{0i}	T_{1i}
S1	1.5	0.0	S11	0.3	0.0	S21	0.0	1.6	S31	0.0	0.8
S2	0.3	0.8	S12	0.3	0.0	S22	0.3	0.6	S32	0.0	0.8
S3	1.1	0.0	S13	0.8	0.0	S23	0.0	1.3	S33	0.0	0.8
S4	0.9	0.4	S14	0.3	0.0	S24	0.0	1.6	S34	0.0	0.8
S5	0.9	0.4	S15	0.9	0.0	S25	0.7	0.8	S35	0.0	0.8
S6	1.1	0.0	S16	0.9	0.0	S26	0.0	1.6	S36	0.0	0.8
S7	0.3	0.2	S17	0.9	0.0	S27	0.3	0.2	S37	0.0	0.8
S8	0.9	0.5	S18	0.9	0.0	S28	0.3	0.8	S38	0.0	0.2
S9	0.9	0.8	S19	0.9	0.0	S29	0.9	0.8	S39	0.0	0.8
S10	1.2	0.0	S20	0.9	0.0	S30	0.0	1.2	S40	0.0	0.2
$S_0 = 0.052$						$S_1 = 0.067$					
$S_{11} = 0.172$											
$\hat{\mathrm{Var}}\left(\widehat{\mathrm{AUC}}\right) = 0.004$											
95% CI for AUC: (0.539, 0.799)											

$$T_{1ik} = \frac{1}{n_0}\sum_{i'=1}^{I_0}\sum_{k=1}^{n_{0i'}}\Psi \quad T_{0ik} = \frac{1}{n_1}\sum_{i'=1}^{I_0}\sum_{k=1}^{n_{0i'}}\Psi. \tag{9}$$

In words, for the kth unit with the condition, we compare their score to each of the n_0 units without the condition. Similarly, for the kth unit without the condition we compare their score to each of the n_1 units with the condition. Let I denote the total number of subjects. The AUC with clustered data is estimated as

$$\widehat{\mathrm{AUC}} = \sum_{i=1}^{I}\sum_{k=1}^{n_{1i}} T_{1ik}/n_1 = \sum_{i=1}^{I}\sum_{k=1}^{n_{0i}} T_{0ik}/n_0.$$

Next we calculate the variance estimates and a covariance estimate. Let T_{1i} and T_{0i} denote the sums of the T_{1ik} and T_{0ik} components for the ith subject. Then

$$S_1 = \frac{I_1}{(I_1-1)n_1}\sum_{i=1}^{I_1}\left(T_{1i.} - n_{1i}\widehat{\mathrm{AUC}}\right)^2,$$

$$S_0 = \frac{I_0}{(I_0-1)n_0}\sum_{i=1}^{I_0}\left(T_{0i.} - n_{0i}\widehat{\mathrm{AUC}}\right)^2, \text{ and } S_{11} = \frac{I}{(I-1)}\sum_{i=1}^{I}\left[(T_{1i.} - n_{1i}\widehat{\mathrm{AUC}})(T_{0i.} - n_{0i}\widehat{\mathrm{AUC}})\right]. \tag{10}$$

Finally, the estimate of the variance of the AUC is

$$\hat{\mathrm{Var}}\left(\widehat{\mathrm{AUC}}\right) = \frac{1}{n_1}S_1 + \frac{1}{n_0}S_0 + \frac{2}{n_1 n_0}S_{11}. \tag{11}$$

These calculations are illustrated in table 10 for the transmission ultrasound data.

Another approach to clustered data is with bootstrap methods (Zhou *et al* 2011). In this approach samples of subjects (not units) are drawn repeatedly from the observed data. The process is performed with replacement, meaning that a selected subject is returned to the sample and thus is available to be selected again. Typically we select ⩾1000 such samples. The AUC is estimated from each of the bootstrap samples. If there are B bootstrap samples, then there would be B AUC estimates. The 95% CI for AUC is constructed by ordering the B AUC estimates from lowest to highest, and identifying the 2.5th and 97.5th ranked estimates as the lower and upper bounds of the CIs.

4.4. Modifications to standard ROC curve

The standard ROC curve, where the reference standard classifies subjects as condition present versus condition absent, has been modified by many authors in order to account for more complex diagnostic situations (Shiraishi *et al* 2009). We mention three of these here.

Several authors have expanded the traditional ROC summary measure and the ROC curve for the situations where the reference standard scale is not binary but has > 2 truth categories (Kijewski *et al* 1989; Mossman 1999;

Dreiseitl *et al* 2000; Edwards*et al* 2004, and Obuchowski *et al* 2005), or is continuous (Obuchowski 2006). For example, in diagnosing children with acute abdominal pain, Obuchowski *et al* (Obuchowski *et al* 2001) asked radiologists to assign a differential diagnosis to subjects based on the subject's CT image. Readers were given a list of potential diagnoses and asked to assign confidence scores such that the sum of the scores equals 100. A new ROC summary measure of accuracy was then computed based on pairwise comparisons of diagnoses, weighted by their relative prevalence. As a second example, Nakas *et al* (Nakas *et al* 2010) illustrated how proton magnetic resonance spectroscopy can be used to distinguish three groups of patients: negative, HIV-positive non-symptomatic, and HIV-positive with AIDS dementia complex. They estimate the three distributions of test results, illustrate a three-dimensional ROC surface, and identify optimal cut-points for classifying patients.

Time-dependent ROC curves, often used to assess the accuracy of imaging biomarkers, is an important modification to the traditional ROC curve. Here the reference standard is observed at some time in the future (Heagerty *et al* 2000). For example, the change in volume in a pulmonary lesion from baseline to post-treatment may be considered an imaging biomarker; its accuracy can be assessed using time-dependent ROC curves where the binary reference standard is the occurrence of death, observed at some time in the future, or the subject is followed until the end of the study (i.e. censored observation).

Recently Shiu and Gatsonis described the predictive ROC curve (Shiu and Gatsonis 2008). It is a plot of (1-negative predictive value, or 1-NPV) versus (positive predictive value, or PPV), as compared to the traditional ROC curve of (1-true negative rate, or 1-TNR) versus (true positive rate, or TPR). Its advantage over an ROC curve is that predictive values like NPV and PPV, and now summary measures of the predictive ROC curve, are sometimes more clinically relevant. Although not commonly applied, this approach is potentially useful.

Lastly, we note that there has been considerable work in developing models that characterize the ROC curve as a function of subjects' characteristics (e.g. age, gender) and disease characteristics (e.g. severity, location). Readers interested in ROC regression analysis should consult these references (Tosteson and Begg 1988, Pepe 1998, 2003, Pepe and Thompson 2000, Alonzo and Pepe 2002, Dodd and Pepe 2003b, Janes and Pepe 2009, Zhou *et al* 2011). These methods are extended to methods for combining imaging and non-imaging characteristics, or biomarkers, to find optimally performing prediction tools using ROC analysis. Some key papers in this area include (Pepe and Thompson 2000, Pepe 2003, Jiang and Metz 2006, Janes and Pepe 2008, 2009, Liu *et al* 2011, Liu and Zhou 2013).

5. Comparing diagnostic tests using ROC analysis

Probably the most important role of ROC analysis is in the comparison of the accuracy of diagnostic tests. Consider the transmission ultrasound example. Suppose we want to compare the accuracy of the SOS measurements to the radiologist's accuracy. We could try to find cut-points where the SOS measurements and the radiologist's scores have similar specificity, then we could compare sensitivities to see which has better accuracy. Conversely, we could try to find cut-points where the sensitivities are similar, then compare their specificities. Examining figure 3, we might choose a FPR of 0.30 since they both have an observed operating point here. The SOS curve has two empirical points at a FPR = 0.30: one with a low sensitivity of 0.45 and the other with sensitivity at 0.75, similar to the radiologist's sensitivity. It's not clear then how to compare their sensitivities at a FPR = 0.30. In other situations there may not exist observed cut-points where two tests' FPR or sensitivity align, and furthermore if such cut-points exist, they may occur at specificities (or sensitivities) that are not relevant (i.e. too low specificity for clinical relevance). Returning to figure 3, it is not at all evident that a comparison of the tests' sensitivities at a FPR = 0.30 would appropriately capture the overall differences in accuracy between the two tests. Of course, ROC analysis, particularly the comparison of ROC summary measures, overcomes these limitations.

5.1. Statistical hypotheses

We begin our discussion of the comparison of ROC curves with a review of statistical hypothesis testing. In section 4 we discussed estimation of ROC summary measures and construction of 95% CIs. In many studies, the primary objective is often to compare the accuracy of two or more tests, in addition to reporting the tests' accuracies. One approach is to construct a 95% CI for the difference in accuracy of two tests. Such a CI is extremely useful in understanding the magnitude of the difference and the precision with which the difference was estimated. However, investigators often want to formally test a hypothesis about the tests' accuracies. The two most common types of hypotheses are superiority and non-inferiority. We discuss each of these next.

5.1.1. Superiority

Suppose we want to determine if the accuracy of the SOS measurements differed from the accuracy of radiologist's subjective interpretation. Since we are hoping to find a difference, we use this as our alternative hypothesis. We start with the null hypothesis of no difference, then examine the study results to determine if we have sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis. The hypotheses can be written as:

$$H_o:\ \theta_{\text{SOS}} = \theta_{\text{radiologist}} \text{ versus } H_A:\ \theta_{\text{SOS}} \neq \theta_{\text{radiologist}} \tag{12}$$

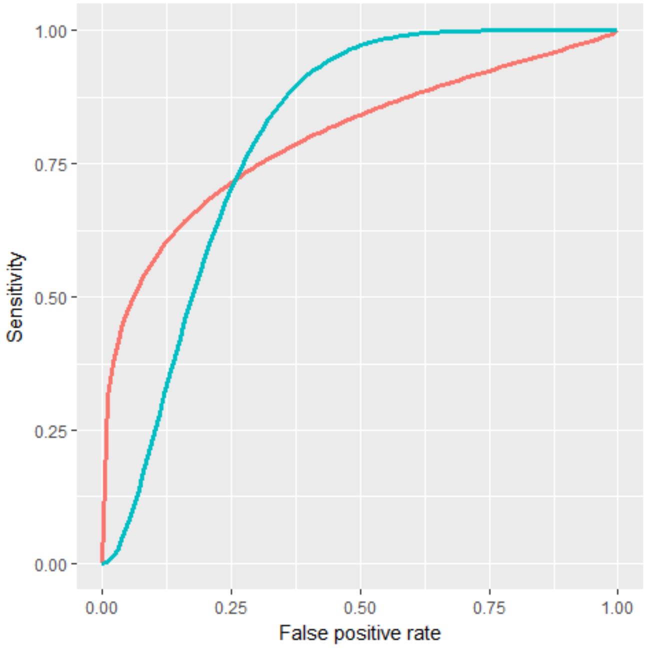


Figure 12. Two hypothetical ROC curves which intersect at FPR = 0.26. The AUC is 0.80 for both curves, but the red curve may be viewed more favorably if lower false positive rates are needed (because sensitivity is greater for the red curve for FPRs less than 0.26).

where θ_{SOS} denotes the accuracy of the SOS measurements and $\theta_{\text{radiologist}}$ denotes the accuracy of the radiologist. Accuracy might be characterized by the area under the ROC curve, the partial area in a pre-defined region (e.g. area where the FPR $\leqslant 0.10$), or the sensitivity at a pre-defined FPR (e.g. sensitivity at a FPR $= 0.10$). A Wald test can be used to test the set of hypotheses in equation (12):

$$z = \frac{(\hat{\theta}_{\text{SOS}} - \hat{\theta}_{\text{radiologist}})}{\sqrt{\widehat{\text{Var}}(\hat{\theta}_{\text{SOS}} - \hat{\theta}_{\text{radiologist}})}} \tag{13}$$

where $\hat{\theta}$ denotes the estimate of the accuracy, e.g. estimated area under the ROC curve, as described in section 4. The estimated variance of the difference in accuracy takes on the general form:

$$\widehat{\text{Var}}\left(\hat{\theta}_{\text{SOS}} - \hat{\theta}_{\text{radiologist}}\right) = \widehat{\text{Var}}\left(\hat{\theta}_{\text{SOS}}\right) + \widehat{\text{Var}}\left(\hat{\theta}_{\text{radiologist}}\right) - 2\widehat{\text{Cov}}_{\text{SOS,Rad}}. \tag{14}$$

Nonparametric estimates of the variance of the area under the ROC curve for a single diagnostic test were given in equations (4) and (11). Cov denotes the covariance between the two tests. In an unpaired study design (i.e. one set of subjects measured with the SOS and a different set interpreted by the radiologist) the covariance is zero. For a paired design, as in the transmission ultrasound study, the covariance must be estimated. A nonparametric estimate of the covariance for the area under the ROC curve is (DeLong *et al* 1988):

$$\widehat{\text{Cov}} = \frac{1}{n_1}S_{10} + \frac{1}{n_0}S_{01}, \tag{15}$$

where the covariance terms are:

$$S_{10} = \frac{1}{(n_1-1)}\sum_{i=1}^{n_1}(T_{1i_{\text{test1}}} - \widehat{\text{AUC}}_{\text{test1}})(T_{1i_{\text{test2}}} - \widehat{\text{AUC}}_{\text{test2}}) \text{ and}$$
$$S_{01} = \frac{1}{(n_0-1)}\sum_{j=1}^{n_0}(T_{0j_{\text{test1}}} - \widehat{\text{AUC}}_{\text{test1}})(T_{0j_{\text{test2}}} - \widehat{\text{AUC}}_{\text{test2}})$$

and $T_{1i_{\text{test1}}}$ is the T_1 variance component for test 1 (e.g. SOS) and $T_{0j_{\text{test1}}}$ is the T_0 variance component for test 1 (see equation (2)). Parametric estimates are also available (Metz *et al* 1980, 1984). Most of the ROC software packages perform these calculations for you. For small studies, the two-way random-effects ANOVA method of Tcheuko *et al* (2016) should be used as an alternative to the Delong *et al* variance estimates.

Once the z statistic in equation (13) is calculated, it is compared against values from a standard normal distribution. For example, to test the hypotheses in equation (12) using a pre-determined significance level of 5% (that is, allowing for a 5% chance of wrongly rejecting the null hypothesis when in fact the null hypothesis is true), we would reject the null hypothesis if our calculated z is < -1.96 or if z is $> +1.96$. It's important to first examine the two tests' ROC curves before making conclusions about the superior accuracy of one test over another. If

Table 11. Using the transmission ultrasound data, we compare the area under the ROC curve for the radiologist's subjective interpretations and the SOS measurements.

$\hat{\theta}_{\text{radiologist}}$	0.737 50
$\widehat{\text{Var}}\left(\hat{\theta}_{\text{radiologist}}\right)$	0.005 532 90
$\hat{\theta}_{\text{SOS}}$	0.671 25
$\widehat{\text{Var}}\left(\hat{\theta}_{\text{SOS}}\right)$	0.008 403 13
$\widehat{\text{Cov}}_{\text{SOS,Rad}}$	0.006 398 03
Superiority test	$z = \frac{(0.671\,25-0.737\,50)}{\sqrt{0.005\,532\,90+0.008\,403\,13-2*0.006\,398\,03}} = -1.9622$ Associated p-value: 0.0497
Non-inferiority test	$z = \frac{([0.671\,25+0.05]-0.737\,50)}{\sqrt{0.005\,532\,90+0.008\,403\,13-2*0.006\,398\,03}} = -0.481\,29$ Associated p-value: 0.6303

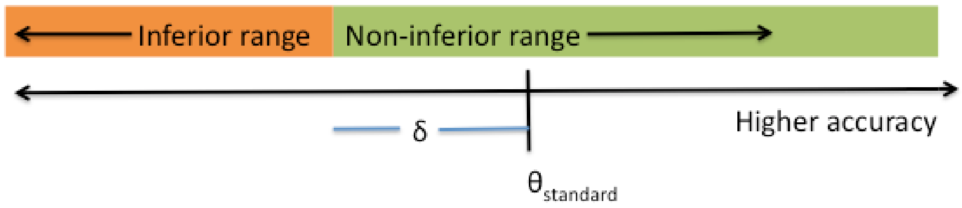


Figure 13. Illustration of non-inferiority in diagnostic test accuracy of a new test compared with a standard test. If the new test has accuracy equal to the standard test or worse than the standard by no more than δ, then it is non-inferior (green zone). If the new test has accuracy of more than δ worse than the standard test, then it is inferior to the standard test (orange zone).

the ROC curves cross, then we need to consider where they cross. In figure 12, the curves have the same area, i.e. 0.80, but we might not value the two tests equally if, for example, low false positive rates are needed clinically. In a recent review by Dendumrongsung *et al* (2014), they report that 41% of diagnostic accuracy studies fail to present the ROC curves when comparing two tests. The partial area under the curve might be a more appropriate metric to compare two tests when curves cross. McClish (1990) described a test to determine the range of FPRs where two tests have the same accuracy and the range where the tests have different accuracy. This test is useful in an exploratory type of study, but when hypothesis testing, the range for the partial area should be specified apriori, not after examining the data.

Table 11 illustrates the calculations for comparing the area under the ROC curve using the transmission ultrasound data.

5.1.2. *Non-inferiority*

Now suppose we want to determine if the accuracy of the SOS measurements is as good as the accuracy of the radiologist's subjective interpretation. Since we are hoping to show that the SOS measurements are at least as accurate, maybe even better (i.e. 'non-inferior'), we put this in our alternative hypothesis. The null hypothesis is that the accuracy of the SOS measurements is worse than the radiologist's, as follows:

$$H_o: \theta_{\text{SOS}} + \delta \leqslant \theta_{\text{radiologist}} \text{ versus } H_A: \ \theta_{\text{SOS}} + \delta > \ \theta_{\text{radiologist}} \tag{16}$$

where δ denotes the non-inferiority margin. Figure 13 illustrates the notion of non-inferior. Let θ_{new} denote the accuracy of a new diagnostic test, such as the SOS measurements, which might be less costly or have fewer complications than the standard test, θ_{standard}, i.e. the radiologist's interpretations. The accuracy of the new test could be a little less than the accuracy of the standard test and still be useful because of its lower cost and/or lower risk of complications. δ is the maximum reduction in the new test's accuracy that is allowed.

A Wald test can be used to test the set of hypotheses in equation (16):

$$z = \frac{([\hat{\theta}_{\text{SOS}} + \delta] - \hat{\theta}_{\text{radiologist}})}{\sqrt{\widehat{\text{Var}}(\hat{\theta}_{\text{SOS}} - \hat{\theta}_{\text{radiologist}})}}$$

where the null hypothesis is rejected if z is >1.96 (for a one-tailed test with 5% type I error rate). The estimated variance of the difference in accuracy takes on the same form as in equation (14) and the covariance as in equation (15). See Chen *et al* (2012) for further details on testing non-inferiority of diagnostic test accuracy. Table 11 illustrates testing the non-inferiority hypothesis using the transmission ultrasound data where the non-inferiority margin is arbitrarily set at $\delta = 0.05$.

Table 12. Results from a hypothetical example of a multi-reader study. Estimates of the AUC (and the standard error) are displayed for each reader and test. The difference between the readers' mean AUCs with test A and B was 0.061. The 95% confidence interval for the difference was (0.008, 0.115).

Reader	AUC for Test A estimate (SE)	AUC for Test B estimate (SE)	Difference estimate (SE)	p-value
1	0.810 (0.034)	0.920 (0.022)	0.109 (0.043)	0.011
2	0.801 (0.035)	0.893 (0.026)	0.092 (0.041)	0.026
3	0.763 (0.040)	0.853 (0.031)	0.090 (0.053)	0.088
4	0.883 (0.028)	0.938 (0.019)	0.055 (0.034)	0.109
5	0.869 (0.030)	0.924 (0.024)	0.055 (0.039)	0.165
6	0.840 (0.032)	0.807 (0.036)	−0.033 (0.048)	0.495
Mean	0.828 (0.019)	0.889 (0.021)	0.061 (0.021)	0.032

5.1.3. Testing multiple hypotheses

Sometimes investigators want to test multiple hypotheses in the same study. When these hypotheses involve measures of test accuracy, we must account for the correlation among them. For example, we know that summary measures of the ROC curve (e.g. AUC) are a function of the test's sensitivity and specificity. Suppose we conduct a study where we want to compare two tests' diagnostic accuracy; suppose we want to compare their AUCs, as well as their sensitivities and specificities at a specific cut-point. We have three measures of test accuracy that we are interested in: AUC, sensitivity, and specificity. Bullen *et al* (2017) describe several scenarios of multiple testing in diagnostic accuracy studies. They show that if we compare the tests' AUC, sensitivity, and specificity, each at the α significance level, then we will have an elevated risk of a type I error. In other words, if we conduct all three statistical comparisons at $\alpha = 0.05$, and if there really is no difference in the two tests' accuracy, then we will wrongly conclude that there is a difference about 10% of the time. Since we have stated that we are using a significance level of 0.05 (5% risk of type I error), our study results and conclusions could be misleading.

For many studies the following recommendation for multiple testing would apply (Bullen *et al* 2017). First, compare the tests' AUCs at a significance level of α; usually we choose $\alpha = 0.05$. If we fail to reject the null hypothesis, then no further testing is performed, i.e. we do not conduct the statistical tests for sensitivity and specificity. On the other hand, if we reject the null hypothesis for the AUC, then we simultaneously compare the tests' sensitivities and specificities, controlling the significance level at α. We can control the type I error by using the Holm's correction (Holm 1979).

5.2. Multi-reader studies

In section 3 we discussed study designs for MRMC studies, as well as sample size calculation. We pointed out that readers' findings should never be pooled together to construct an ROC curve, but rather each reader's findings should be used to construct reader-specific ROC curves. From each reader's curve we can estimate the area or partial area under the curve and use the methods described earlier to characterize individual readers' accuracy (section 4) and to perform reader-specific comparisons between diagnostic tests (section 5.1). In this subsection we focus on analyses that synthesize the results over all the study readers in order to generalize the results to a population of readers.

In figure 5 we illustrated the hypothetical results of 6 readers interpreting the same set of images. Each reader had his/her own ROC curve and area estimate. Table 12 summarizes the hypothetical readers' ROC areas and standard errors (i.e. the square root of the variance calculated from equation (4)) for two tests. For each reader, there is an estimate of the difference in the areas of the two tests, the standard error of the difference (i.e. the square root of the variance calculated from equation (14)), and the p-value associated with the Wald test (given in equation (13)) (last column of table 12). Note that for two readers, the diagnostic accuracy of test B is significantly higher than the accuracy of test A; for three readers there is a trend for test B to have higher accuracy than test A but the difference is not statistically significant; and there is one reader where the accuracy was higher with test A than test B. Figure 14 is a simple bar chart illustrating the gain in accuracy with test B for the 6 readers. These results describe the performance of the six specific readers in the study; however, we want to generalize the results to the population from which these six readers came from. Several statistical models and methods have been developed to address this. We describe here one popular model and method and cite several others.

Obuchowski and Rockette (1995) described a model to explain the sources of variability observed in readers' diagnostic test accuracies. The model has been extended by Hillis (Hillis *et al* 2005, Hillis 2007, 2012, 2014) to improve model estimation and allow generalizability to many different study designs. The basic model is given in equation (17). The estimated diagnostic accuracy of reader j with diagnostic test i on reading occasion k (denoted as $\hat{\theta}_{ijk}$) is a function of an overall mean accuracy (μ), a fixed average increase or decrease in accuracy due to diagnostic test i (μ_i), a random effect due to reader j (r_j) which is assumed to follow a normal distribution with mean zero and some variance attributable to inter-reader differences, an interaction between diagnostic test and reader ($(\mu r)_{ij}$) which is assumed to follow a normal distribution with mean zero and some variance attributable to how

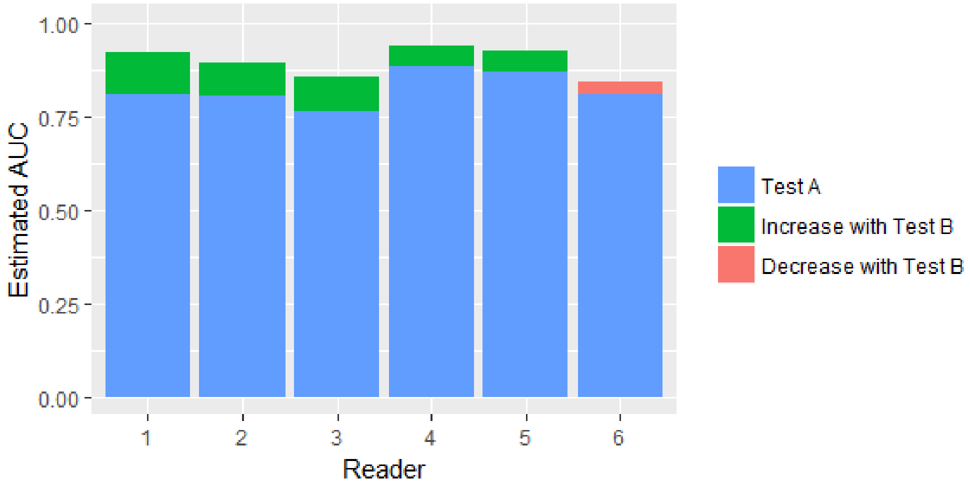


Figure 14. Bar chart displaying the results from a hypothetical example of a multi-reader study. Five readers had higher AUC with test B and one reader had lower AUC with test B.

reader *j* uniquely interprets images with test *i*, and finally a random effect ($\in_{ijk}$) which also is assumed to follow a normal distribution with mean zero and variance attributable to both intra-reader effects and variability due to the particular sample of subjects in the study (i.e. these two effects cannot be separated in MRMC studies unless the study design includes replicate interpretations by the same reader):

$$\hat{\theta}_{ijk} = \mu + \mu_i + r_j + (\mu r)_{ij} + \in_{ijk} \tag{17}$$

where there are *I* total tests ($I = 2$ in our example), *J* total readers ($J = 6$ in our example), and *K* replicate reading occasions (usually $K = 1$).

To characterize the mean accuracy of a population of readers with diagnostic test *i*, we estimate the readers' mean accuracy and its variance. Let $\hat{\theta}_{i..}$ denote the readers' mean accuracy for test *i*, i.e. the average area under the ROC curve for test *i* over the *J* readers, and let $\hat{\theta}_{ij.}$ denote the average area under the ROC curve for reader *j* with the *i*th test over the *K* reading occasions. Equation (18) shows the construction of the CI for the mean accuracy for test *i* for the population of readers.

$$\hat{\theta}_{i..} \pm t_{\frac{\alpha}{2},\mathrm{DF}} \sqrt{\frac{1}{J(J-1)} \sum_{j=1}^{J} \left(\hat{\theta}_{ij.} - \hat{\theta}_{i..}\right)^2 + \widehat{\mathrm{Cov}}(\hat{\theta}_{ij}, \hat{\theta}_{ij'})} \tag{18}$$

where $t_{\frac{\alpha}{2},\mathrm{DF}}$ denotes a random variable from a Student's t distribution at $\alpha/2$ with degrees of freedom (DF) given by Hillis (2014):

$$\mathrm{DF} = \frac{(J-1)(\mathrm{MS}_i + J[\widehat{\mathrm{Cov}}\left(\hat{\theta}_{ij}, \hat{\theta}_{ij'}\right)])^2}{(\mathrm{MS}_i)^2}$$

where $\mathrm{MS}_i = \sum_{j=1}^{J} \left(\hat{\theta}_{ij.} - \hat{\theta}_{i..}\right)^2 / (J)$. For a 95% CI, α is set to 0.05. The last term in equation (18) is an estimate of the covariance between the estimated diagnostic accuracies of any random pair of different readers (i.e. reader *j* and reader *j'*); it will be zero in an unpaired-reader study design and non-zero when readers interpret the same sample of subjects, as in the usual paired-reader, paired-subject MRMC design (depicted in table 4). The covariance can be estimated using equation (15), by taking the average over all pairwise estimates between the *J* readers. In table 12 in the bottom row (middle two columns) are estimates of readers' mean accuracy with Test A and B, and the standard error of the mean.

Now suppose that we want to compare the mean accuracy of the two tests. We want to generalize the results to the population of readers (rather than just the 6 readers in our study). The null hypothesis might be that the readers' mean accuracy with Test A is the same as with Test B, versus the alternative hypothesis that the readers' mean accuracies differ (i.e. superiority hypothesis). The statistical test is given in equation (19). Note that this test is for the classical paired-reader paired-subject design depicted in table 4; the statistical test would be slightly different if another MRMC design had been used (Obuchowski 1995).

$$F = \frac{J\left(\hat{\theta}_{1..} - \hat{\theta}_{2..}\right)^2 / 2}{\mathrm{MS}_{ij} + J\{\widehat{\mathrm{Cov}}\left(\hat{\theta}_{ij}, \hat{\theta}_{ij'}\right) - \widehat{\mathrm{Cov}}\left(\hat{\theta}_{ij}, \hat{\theta}_{i'j'}\right)\}} \tag{19}$$

Table 13. ROC Software[a].

Software	Description	Platform(s)	Location
Metz ROC Software	ROC-kit includes a variety of programs for fitting and comparing binormal-fit ROC curves	The interface allows calls from R, SAS, Matlab, and IDL, or from an external program or scripting language	http://metz-roc.uchicago.edu/MetzROC/software
Medical Image Perception Laboratory	Single reader and MRMC analysis—parametric and nonparametric; sample size estimation	.NET Framework Version 2.0 or later is required on XP, Vista, and Windows 7. Some sample SAS programs also available to call various routines. Point-and-click sample size software	http://perception.radiology.uiowa.edu/Software/ReceiverOperatingCharacteristicROC/tabid/120/Default.aspx
ROC Analysis from Quantitative Health Sciences	Directory of software packages for non-parametric single and MRMC ROC analysis; sample size estimation	SAS, R, S-Plus, Fortran	https://lerner.ccf.org/qhs/software/roc_analysis.php
Pepe lab	Estimation of ROC curves, sample size calculation,	Stata	http://research.fhcrc.org/diagnostic-biomarkers-center/en/software/brasoft.html
iMRMC	analyzing and sizing MRMC reader studies	license-free Java applications and R package	https://github.com/DIDSR/iMRMC
Software for time-dependent ROC	Time-dependent ROC	R/S-plus package survivalROC	http://faculty.washington.edu/heagerty/Software/SurvROC/
Analyzing ROC curves with SAS	SAS code and macros for estimation and comparison of ROC curves—parametric and nonparametric	SAS	https://sas.com/sas/books/authors/mithat-gonen.html
Web-based Calculator for ROC curves	Calculates an ROC curve	Web-based calculator; requires a web browser that supports HTML5	www.jrocfit.org
MedCalcAnalyin	Estimation and comparison of ROC curve and sample size calculation for single reader studies.	Written inWYSIWYG text editor. It imports from Excel, Excel 2007, SPSS, DBase and Lotus files, and files in SYLK, DIF or plain text format	https://medcalc.org/manual/roc-curves.php
SAS	Estimation and comparison of nonparametric ROC curve and plots	SAS and bridge to R	http://research.fhcrc.org/content/dam/stripe/diagnostic-biomarkers-statistical-center/files/ROC_Curve_Plotting_in_SAS_9_2.pdf
pROC	Display and analysis of ROC curves	R	https://cran.r-project.org/web/packages/pROC/pROC.pdf

[a] Please note that it is beyond the scope of this paper to test, compare, and validate these software programs. Users should check that the software has been validated by the software authors.

where $\mathrm{MS}_{ij} = \sum_{i=1}^{2}\sum_{j=1}^{J}\left(\hat{\theta}_{ij.} - \hat{\theta}_{i..} - \hat{\theta}_{.j.} + \hat{\theta}_{...}\right)^2/(J-1)$, $\hat{\theta}_{.j.}$ denote the average area under the ROC curve for reader j over both tests and over the K reading occasions, the covariance $\widehat{\mathrm{Cov}}\left(\hat{\theta}_{ij}, \hat{\theta}_{ij'}\right)$ describes the mean covariance in estimated diagnostic test accuracies of any random pair of different readers using test i, and the covariance $\widehat{\mathrm{Cov}}\left(\hat{\theta}_{ij}, \hat{\theta}_{i'j'}\right)$ describes the mean covariance in estimated diagnostic test accuracies of any random pair of different readers each using different tests. The statistic in equation (19) follows an F distribution with numerator degrees of freedom of $(I-1)$ and denominator degrees of freedom given by Hillis (2007):

$$\text{denominator DF} = \frac{(J-1)\left(\mathrm{MS}_{ij} + J\left[\widehat{\mathrm{Cov}}\left(\hat{\theta}_{ij}, \hat{\theta}_{ij'}\right) - \widehat{\mathrm{Cov}}\left(\hat{\theta}_{ij}, \hat{\theta}_{i'j'}\right)\right]\right)^2}{\left(J\left(\hat{\theta}_{1..} - \hat{\theta}_{2..}\right)^2/2\right)^2}.$$

In table 12 in the bottom row of the last column, an estimate of the difference in the readers' mean accuracy is given, along with the standard error of the difference. The 95% confidence interval for the difference is (0.008, 0.115). The p-value associated with the test in equation (19) is $p = 0.032$, which is <0.05 indicating that we reject the null hypothesis and conclude that readers' mean accuracy with Test B is superior to their mean accuracy with Test A. Note that this result applies to the populations of similar readers and subjects. It is referred to as the random-reader approach (Obuchowski and Rockette 1995). It is also possible to test the hypothesis using a fixed-reader approach, which applies only to the specific readers in the study and thus is less generalizeable.

There are several other MRMC analyses methods. A bootstrap method (Beiden *et al* 2000) can be used to compare tests and is useful for estimating the various sources of variance. A regression approach involves building a subject-level model and can accommodate subject and reader covariates (Song and Zhou 2005). Ishwaran and Gatsonis (2000) describe a hierarchical approach for diagnostic tests with ordinal data. More recently, a model-free U-statistics method was proposed and validated by simulation by Gallas (Gallas 2006, Gallas *et al* 2009). Several authors (Zhou *et al* 2011, Obuchowski *et al* 2004, Gallas *et al* 2009) have compared the various MRMC methods, identifying strengths of each. There is also a recent systematic review on data reporting practices of multi-reader ROC studies (Dendumrongsup *et al* 2014). See section 6 for software options for MRMC analyses.

6. ROC software

There is a wide variety of ROC software available. Table 13 outlines many of the available programs, with links to information about the software. Many of the programs are free.

7. Conclusion

Methods for ROC analysis, including study design, estimation of the curve and its summary measures, hypothesis testing, and comparison methods have advanced tremendously over the last 35 years due to the early groundwork contributions of Dorfman and Alf (1969), Swets and Pickett (1982), Hanley and McNeil (1982, 1983) and Metz (1978, 1980, 1984, 1986). While new methods continue to advance the field, the groundwork principles and methods established by these authors are still widely used today.

ROC analysis plays an essential role in answering critical questions about the diagnostic accuracy of new and existing tests by providing a tool that incorporates both sensitivity and specificity and is independent of the decision threshold. It is particularly important for comparing diagnostic tests' accuracies without dependence on decision thresholds. It is widely applicable, as it can be used to describe the relative cost and benefit of any binary decision-making process. Yet, there are shortcomings to ROC analysis. The complexity of the summary metrics and their estimation and comparison can be challenging to clinicians, regulators, payors, and the public (Gatsonis 2009). Translating ROC study results into clinical use is sometimes difficult when the metrics are not directly applicable and when diagnostic test findings are reported in ways that are not generalizable to clinical practice. Furthermore, ROC analysis does not address outcomes that are most important to patients, i.e. morbidity and mortality, but rather assess an intermediary outcome, i.e. test accuracy. Finally, while today's users of ROC software have many choices of platforms, there still remains the need for more adoption of the methods, especially the more complex methods such as ROC regression, into standard statistical packages. For example, verification bias is a common problem in diagnostic test accuracy studies, yet few studies correct for this bias. Similarly, clustered data is quite prevalent, yet investigators continue to ignore intra-cluster correlation in their analyses. By incorporating ROC methods into standard statistical software packages, it is hoped that these methods would be more accessible to investigators. Acknowledging these strengths and limitations, ROC analysis continues to be a vital tool in the assessment of diagnostic tests.

ORCID iDs

Nancy A Obuchowski https://orcid.org/0000-0003-1891-7477

References

Allen J, Murray A, Di Maria C and Newton J L 2012 Chronic fatigue syndrome and impaired peripheral pulse characteristics on orthostasis–a new potential diagnostic biomarker *Physiol. Meas.* **33** 231–41

Alonzo T A and Pepe M S 2002 Distribution-free ROC analysis using binary regression techniques *Biostatistics* **3** 421–32

Alonzo T A and Pepe M S 2005 Assessing accuracy of a continuous screening test in the presence of verification bias *Appl. Stat.* **54** 173–90

Bartkiewicz B, Huda W and McLellan Y 1991 Impact of gamma camera parameters on imaging performance, evaluated by receiver operating characteristics (ROC) analysis *Phys. Med. Biol.* **36** 1065–74

Beam C A, Layde P M and Sullivan D C 1996 Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample *Arch. Intern. Med.* **156** 209–13

Begg C B and Greenes R A 1983 Assessment of diagnostic tests when disease verification is subject to selection bias *Biometrics* **39** 207–15

Beiden S V, Wagner R F and Campbell G 2000 Components-of-variance models and multiple bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis *Acad. Radiol.* **7** 341–9

Bossuyt P M *et al* (for the STARD Group) 2015 STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies *BMJ* **351** h5527

Bullen J A and Obuchowski N A 2017 Correcting for multiple testing during diagnostic accuracy studies *J. Biopharm. Stat.* **9** 243–8

Chen W, Petrick N A and Sahiner B 2012 Hypothesis testing in noninferiority and equivalence MRMC ROC studies *Acad. Radiol.* **19** 1158–65

DeLong E R, DeLong D M and Clarke-Pearson D L 1988 Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach *Biometrics* **44** 837–45

Dendumrongsup T, Plumb A A, Halligan S, Fanshawe T R, Altman D G and Mallett S 2014 Multi-reader multi-case studies using the area under the ROC curve as a measure of diagnostic accuracy: a systematic review with focus on quality of data reporting *PLoS One* **9** e116018

Dodd L E and Pepe M S 2003a Partial AUC estimation and regression *Biometrics* **59** 614–23

Dodd L E and Pepe M S 2003b Semiparametric regression for the area under the receiver operating characteristic curve *J. Am. Stat. Assoc.* **98** 409–17

Dorfman D D and Alf E 1969 Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data *J. Math. Psychol.* **6** 487–96

Dreiseitl S, Ohno-Machado L and Binder M 2000 Comparing three-class diagnostic tests by three-way ROC analysis *Med. Decis. Mak.* **20** 323–31

Edwards D C, Metz C E and Kupinski M A 2004 Ideal observers and optimal ROC hypersurfaces in n-class classication *IEEE Trans. Med. Imaging* **23** 891–5

Egan J P 1975 *Signal Detection Theory and ROC Analysis* (New York: Academic)

FDA 2012 *Guidance for Industry and FDA Staff—Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data—Premarket Approval (PMA) and Premarket Notification [510(k)] Submissions* (www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments)

Fryback D G and Thornbury J R 1991 The efficacy of diagnostic imaging *Med. Decis. Making* **11** 88–94

Gallas B D 2006 One-shot estimate of MRMC variance: AUC *Acad. Radiol.* **13** 353–62

Gallas B D, Bandos A, Samuelson F and Wagner R F 2009 A framework for random-effects ROC analysis: biases with the bootstrap and other variance estimators *Commun. Stat.* A **38** 2586–603

Gatsonis C 2009 Receiver operating characteristic analysis for the evaluation of diagnosis and prediction *Radiology* **253** 593–6

Goenka A H, Herts B R, Obuchowski N A, Primak A N, Dong F, Karim W and Baker M E 2014 Effect of reduced radiation exposure and iterative reconstruction on detection of low-contrast low-attenuation lesions in an anthropomorphic liver phantom: an 18-reader study *Radiology* **272** 154–63

Greenes R and Begg C 1985 Assessment of diagnostic technologies: methodology for unbiased estimation from samples of selective verified patients *Investigative Radiol.* **20** 751–6

Gur D, Rockette H E, Good W F, Slasky B S, Cooperstein L A, Straub W H, Obuchowski N A and Metz C E 1990 Effect of observer instruction on ROC study of chest images *Investigative Radiol.* **25** 230–4

Hadjiiski L, Chan H P, Sahiner B, Helvie M A and Roubidoux M A 2007 Quasi-continuous and discrete confidence rating scales for observer performance studies. Effects on ROC analysis *Acad. Radiol.* **14** 38–8

Hanley J A 1988 The robustness of the binormal assumption used in fitting ROC curves *Med. Decis. Making* **8** 197–203

Hanley J A and McNeil B J 1982 The meaning and use of the area under a receiver operating characteristic (ROC) curve *Radiology* **143** 29–36

Hanley J A and McNeil B J 1983 A method of comparing the areas under receiver operating characteristic curves derived from the same cases *Radiology* **148** 839–43

He H, Lyness M L and McDermott M P 2009 Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias *Stat. Med.* **28** 361–76

He Y and Escobar M 2008 Nonparametric statistical inference method for partial areas under receiver operating characteristic curves, with applications to genomic studies *Stat. Med.* **27** 5291–308

Heagerty P J, Lumley T and Pepe M S 2000 Time-dependent ROC curves for censored survival data and a diagnostic marker *Biometrics* **56** 337–44

Hillis S L 2007 Comparison of denominator degrees of freedom methods for multiple observed ROC analysis *Stat. Med.* **26** 596–610

Hillis S L 2012 Simulation of unequal-variance binormal multireader ROC decision data: an extension of the Roe and Metz simulation model *Acad. Radiol.* **19** 1518–28

Hillis S L 2014 A marginal-mean ANOVA approach for analyzing multireader multicase radiological imaging data *Stat. Med.* **33** 330–60

Hillis S L, Obuchowski N A and Berbaum K S 2011 Power estimation for multireader ROC methods: an updated and unified approach *Acad. Radiol.* **18** 129–42

Hillis S L, Obuchowski N A, Schartz K M and Berbaum K S 2005 A comparison of the Dorfman–Berbaum–Metz and Obuchowski–Rockette methods for receiver operating characteristic (ROC) data *Stat. Med.* **24** 1579–607

Holm S 1979 A simple sequentially rejective multiple test procedure *Scand. J. Stat.* **6** 65–70

Irwig L, Tosteson A N, Gatsonis C, Lau J, Colditz G, Chalmers T C and Mosteller F 1994 Guidelines for meta-analyses evaluating diagnostic tests *Ann. Intern. Med.* **120** 667–76

Ishwaran H and Gatsonis C 2000 A general class of hierarchical ordinal regression models with applications to correlated ROC analysis *Can. J. Stat.* **28** 731–50

Iuanow E, Smith K, Obuchowski N A, Bullen J and Klock J C 2017 Accuracy of cyst versus solid diagnosis in the breast using quantitative transmission (QT) ultrasound *Acad. Radiol.* **24** 1148–53

Janes H and Pepe M S 2008 Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting *Am. J Epidemiol.* **168** 89–97

Janes H J and Pepe M S 2009 Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve *Biometrika* **96** 371–82

Jiang Y and Metz C E 2006 A Quadratic model for combining quantitative diagnostic assessments from radiologist and computer in computer-aided diagnosis *Acad. Radiol.* **13** 140–51

Kijewski M F, Swensson R G and Judy P F 1989 Analysis of rating data from multiple-alternative tasks *J. Math. Psychol.* **33** 428–51

Klock J 2017 personal communication

Lai C J, Shaw C C, Whitman G J, Yang W T, Dempsey P J, Nguyen V and Ice M F 2006 Receiver operating characteristic analysis for the detection of simulated microcalcifications on mammograms using hardcopy images *Phys. Med. Biol.* **51** 3901–19

Leeflang M M, Deeks J J, Gatsonis C, Bossuyt P M and Cochrane Diagnostic Test Accuracy Working Group 2008 Systematic reviews of diagnostic test accuracy *Ann. Intern. Med.* **149** 889–97

Lenox M W *et al* 2015 Imaging performance of quantitative transmission ultrasound *Int. J. Biomed. Imaging* **2015** 454028

Liu C, Liu A and Halabi S 2011 A min–max combination of biomarkers to improve diagnostic accuracy *Stat. Med.* **30** 2005–14

Liu D and Zhou X H 2011 A model for adjusting for nonignorable verification bias in estimation of ROC curve and its area with likelihood based approach *Biometrics* **66** 1119–28

Liu D and Zhou X H 2013 ROC analysis in biomarker combination with covariate adjustment *Acad. Radiol.* **20** 874–82

Lusted L B 1971 Signal detectability and medical decision-making *Science* **171** 1217–9

Mahmoud A H, Ding X, Dutta D, Singh V P and Kim K 2014 Detecting hepatic steatosis using ultrasound-induced thermal strain imaging: an *ex vivo* animal study *Phys. Med. Biol.* **59** 881–95

Mazzetti S, Gliozzi A S, Bracco C, Russo F, Regge D and Stasi M 2012 Comparison between PUN and Tofts models in the quantification of dynamic contrast-enhanced MR imaging *Phys. Med. Biol.* **57** 8443–53

McClish D K 1989 Analyzing a portion of the ROC curve *Med. Decis. Mak.* **9** 190–5

McClish D K 1990 Determining a range of false positives for which ROC curves differ *Med. Decis. Mak.* **10** 283–7

McGowan L D, Bullen J A and Obuchowski N A 2016 Location bias in ROC studies *Stat. Biopharm. Res.* **8** 258–67

Metz C E 1978 Basic principles of ROC analysis *Semin. Nucl. Med.* **8** 283–98

Metz C E 1986 ROC methodology in radiologic imaging *Investigative Radiol.* **21** 720–33

Metz C E 2008 ROC analysis in medical imaging: a tutorial review of the literature *Radiol. Phys. Technol.* **1** 2–12

Metz C E and Kronman H B 1980 Statistical significance tests for binormal ROC curves *J. Math. Psychol.* **22** 218–43

Metz C E, Herman B A and Shen J 1998 Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data *Stat. Med.* **17** 1033–53

Metz C E, Wang P L and Kronman H B 1984 A new approach for testing the significance of differences measured from correlated data *Information Processing in Medical Imaging VIII* ed F Deconick (The Hague: Martinus Nijhof) pp 432–45

Moher D, Liberati A, Tetzlaff J, Altman D G and Group P 2009 Preferred reporting items for systematic reviews and meta-analysis: the PRISMA statement *PLoS Med.* **6** e1000097

Mossman D 1999 Three-way ROCs *Med. Decis. Mak.* **19** 78–89

Mu W, Chen Z, Liang Y, Shen W, Yang F, Dai R, Wu N and Tian J 2015 Staging of cervical cancer based on tumor heterogeneity characterized by texture features on ^{18}F-FDG PET images *Phys. Med. Biol.* **60** 5123–39

Nakas C T, Alonzo T A and Yiannoutsos C T 2010 Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index *Stat. Med.* **10** 2946–55

Obuchowski N A 1995 Multi-reader ROC studies: a comparison of study designs *Acad. Radiol.* **2** 709–16

Obuchowski N A 1997 Nonparametric analysis of clustered ROC curve data *Biometrics* **53** 567–78

Obuchowski N A 2004 How many observers are needed in clinical studies of medical imaging? *Am. J. Roentgenol.* **182** 867–9

Obuchowski N A 2005 Estimating and comparing diagnostic tests' accuracy when the gold standard is not binary *Acad. Radiol.* **12** 1198–204

Obuchowski N A 2006 An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale *Stat. Med.* **25** 481–93

Obuchowski N A 2009 Reducing the number of reader interpretations in MRMC studies *Acad. Radiol.* **16** 209–17

Obuchowski N A and Rockette H E 1995 Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations *Commu. Stat. Simul. Comput.* **24** 285–308

Obuchowski N A, Beiden S, Berbaum K S, Hillis S L, Ishwaran H, Song H H and Wagner R F 2004 Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods *Acad. Radiol.* **11** 980–95

Obuchowski N A, Goske M J and Applegate K A 2001 Assessing physicians' accuracy in diagnosing paediatric patients with acute abdominal pain: measuring accuracy for multiple diseases *Stat. Med.* **20** 3261–78

Obuchowski N and Hillis S 2011 Sample size tables for computer-aided detection studies *Am. J. Roentgenol.* **197** 821–8

Pepe M S 1998 Three approaches to regression analysis of receiver operating characteristic curves for continuous test results *Biometrics* **54** 124–35

Pepe M S 2003 *The Statistical Evaluation of Medical Tests for Classification and Prediction* (New York: Oxford University Press)

Pepe M S and Thompson M L 2000 Combining diagnostic test results to increase accuracy *Biostatistics* **1** 123–40

Peterson W W, Birdsall T G and Fox W C 1954 The theory of signal detectability *IRE Trans.* **PGIT-4** 171–212

Rockette H E, Gur D and Metz C E 1992 The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques *Investigative Radiol.* **27** 169–72

Rodenberg C and Zhou X H 2000 ROC curve estimation when covariates affect the verification process *Biometrics* **56** 131–6

Rotnitzky A, Faraggi D and Schisterman E 2006 Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias *J. Am. Stat. Assoc.* **101** 1276–88

Shiraishi J, Pesce L L, Metz C E and Doi K 2009 Experimental design and data analysis in receiver operating characteristic studies: lessons learned from reports in Radiology from 1997 to 2006 *Radiology* **253** 822–30

Shiu S Y and Gatsonis C 2008 The predictive receiver operating characteristic curve for the joint assessment of the positive and negative predictive value *Phil. Trans.* A **366** 2313–33

Song X and Zhou X H 2005 A marginal model approach for analysis of multi-reader multi-test receiver operating characteristic (ROC) data *Biostatistics* **6** 303–12

Swets J A 1996 *Signal Detection theory and ROC Analysis in Psychology and Diagnostics: Collected Papers* (Mahwah, NJ: Lawrence Erlbaum Associates)

Swets J A and Pickett R M 1982 *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory* (New York: Academic)

Tanner W P Jr and Swets J A 1954 A decision-making theory of visual detection *Psychol. Rev.* **61** 401–9

Tcheuko L, Gallas B and Samuelson F 2016 Using ANOVA/random-effects variance estimates to compute a two-sample U-statistic of order (1,1) estimate of variance *J. Stat. Theory Pract.* **10** 87–99

The Cochrane Collaboration 2013 *Diagnostic Test Accuracy Working Group* (http://srdta.cochrane.org)

Toledano A Y and Gatsonis C A 1999 GEEs for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate *Biometrics* **55** 488–96

Tosteson A A N and Begg C B 1988 A general regression methodology for ROC curve estimation *Med. Decis. Mak.* **8** 204–15

Van Meter D and Middleton D 1954 Modern statistical approaches to reception in communication theory *IRE Trans.* **PGIT-4** 119–41

Wagner R F, Beiden S V and Metz C E 2001 Continuous versus categorical data for ROC analysis: some quantitative considerations *Acad. Radiol.* **8** 328–34

Zhang D D, Zhou X H, Freeman D H and Freeman J L 2002 A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets *Stat. Med.* **21** 701–15

Zheng Y, Barlow W E and Cutter G 2005 Assessing accuracy of mammography in the presence of verification bias and intrareader correlation *Biometrics* **61** 259–68

Zhou X H 1996 Nonparametric ML estimate of an ROC area corrected for verification bias *Biometrics* **52** 310–6

Zhou X H 1998 Comparing the correlated areas under the ROC curves of two diagnostic tests in the presence of verification bias *Biometrics* **54** 349–66

Zhou X H, Obuchowski N A and McClish D K 2011 *Statistical Methods in Diagnostic Medicine* (Hoboken, NY: Wiley)

Zou K H, Liu A, Bandos A I, Ohno-Machado L and Rockette H E 2011 *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis* (Boca Raton, FL: CRC Press)