Chapter 6
Using and Interpreting Diagnostic Tests with Quantitative Results

The ROC Curve in Diagnostic Accuracy

Suhail A.R. Doi

Abstract Sensitivity and specificity, as defined previously, depend on the cut-off point used to define positive and negative test results. To determine the best cut-off point shift that optimizes sensitivity and specificity, the receiver operating characteristic (ROC) curve is often used. This is a plot of the sensitivity of a test versus its false-positive rate for all possible cut-off points. This chapter outlines its advantages, its use as a means of defining the accuracy of a test, its construction as well as methods for identification of the optimal cut-off point on the ROC curve. Meta-analysis of diagnostic studies is briefly discussed.

Introduction

As we have seen previously, diagnostic test results use a cut-off value based on either a binary or polychotomous scale to define positive and negative test outcomes. On a continuous or ordinal scale, the sensitivity (Se, the probability of a positive test outcome in a diseased individual) and specificity (Sp, the probability of a negative test outcome in a non-diseased individual) can also be computed for specific values. Since the diagnostic test considers the results in two populations, one population with a disease, the other population without the disease, therefore, for every possible cut-off point or criterion value there will be some cases with the disease correctly classified as positive (TP, true-positive fraction), and falsely classified as negative (FN, false-negative fraction). Similarly, cases without the disease can be correctly classified as negative (TN, true-negative fraction) or as positive (FP, false-positive fraction). Compared to the binary outcome, on a continuous scale, the choice of cut-off will affect the degree of false misclassification

S.A.R. Doi (✉)
School of Population Health, University of Queensland, Herston, QLD, Australia

Princess Alexandra Hospital, Brisbane, Australia
e-mail: sardoi@gmx.net

S.A.R. Doi and G.M. Williams (eds.), *Methods of Clinical Epidemiology*,
Springer Series on Epidemiology and Public Health,
DOI 10.1007/978-3-642-37131-8_6, © Springer-Verlag Berlin Heidelberg 2013

by the test and thus for the same test, different cut-offs will have different operating characteristics with an inversely related Se and Sp across cut-off values. Thus, for tests that have an ordinal or continuous cut-off, Se and Sp at a single cut-off value do not fully characterise the tests performance which varies across other potential cut-off values. In this situation, a comparison of such diagnostic tests requires independence from the selected cut-off value and this can be addressed via receiver operating characteristic (ROC) analysis. The ROC method has several advantages:

- Testing accuracy across the entire range of cut-offs thereby not requiring a predetermined cut-off point
- Easily examined visual and statistical comparisons across tests
- Independence from outcome prevalence

Example Data

Data from a prospective evaluation of an Australian pertussis toxin (PT) IgG and IgA enzyme immunoassay are used as an example (May et al. 2012). In this study, the accuracy of anti-PT IgG and anti-PT IgA (as normalized optical density) is examined for the diagnosis of pertussis infection with samples taken within 2–8 weeks after onset of symptoms. The gold standard was *Bordetella pertussis* polymerase chain reaction at the first visit.

Basic Principles

The continuous test result is viewed as a multitude of related tests each represented by a single cut-off with each considered to discriminate between two mutually exclusive states, so that we end up with a Se and Sp that are specific to a selected cut-off value. Each cut-off therefore generates a pair of Se and (1 − Sp) and it is these pairs that are then compared via ROC analysis and at each possible cut-off value for the test. Se and (1 − Sp) are essentially equivalent to the true-positive and false-positive proportions, respectively and when we plot Se against (1 − Sp) for various values of the cut-off across the measurement range, this generates the ROC curve. The ROC curve is therefore a plot of the FP probability on the x-axis and the TP probability on the y-axis across several thresholds of a continuous value measured in each subject, with the positive result being assumed for subjects above the threshold. Each point on the curve represents a Se/Sp pair corresponding to a particular cut-off; the latter are also known as the decision threshold or criterion values.

The ROC method is therefore an overall measure (across all possible cut-offs) of diagnostic performance of a test and can be used to compare the diagnostic performance of two or more laboratory or diagnostic tests. The perfect test with perfect discrimination (no overlap in the diseased and healthy distributions) has a

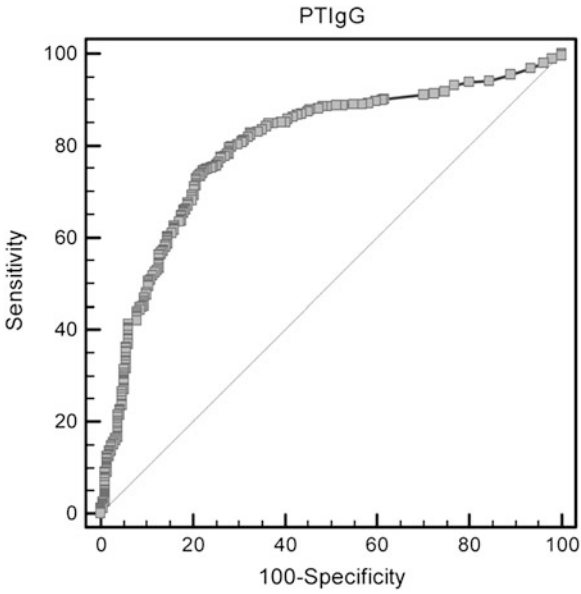

Fig. 6.1 The ROC curve for PT IgG. Each point on the *curve* represents a single cut-off, with sensitivity plotted against 1-specificity (false positive rate). The *central diagonal* is the line of equality

ROC curve that passes through the upper left corner (100 % sensitivity, 100 % specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test. Also, the slope of the tangent line at a cut-point gives the likelihood ratio (LR) for that value of the test. The ROC curve for PT IgG is depicted in Fig. 6.1.

Discrimination Between Diseased and Non-diseased

The area under such an ROC curve is used as a measure of how well a test can distinguish between two diagnostic groups (diseased/normal), independent of any particular cut-off. The closer the curve follows the left-hand border and then the top border of the ROC space, the more area there is under the curve and the more accurate the test. The closer the curve follows the 45° diagonal of the ROC space, the less accurate the test.

The area under the curve (AUC) is therefore a global (i.e. independent of the cut-off value) summary statistic of diagnostic accuracy. The AUC is also known as the c statistic or c index, and can range from 0.5 (random chance or no predictive ability, which would follow the 45° line on the ROC plot) to 1 (perfect discrimination/accuracy; the ROC curve reaches the upper left corner of the graph). The greater the AUC, the more able is the test to capture the trade-off between Se and Sp over a continuous range. According to an arbitrary guideline, one could then use the AUC to classify the accuracy of a diagnostic test (Table 6.1).

Table 6.1 Interpretation of the AUC in terms of accuracy of a test

Accuracy	AUC
Non-informative	AUC = 0.5
Less accurate	0.5 < AUC < 0.7
Moderately accurate	0.7 < AUC < 0.9
Highly accurate	0.9 < AUC < 1
Perfect test	AUC = 1
Results for PT IgG	
Area under the ROC curve (AUC)	0.798
Standard error[a]	0.0177
95 % confidence interval[b]	0.763–0.832
Z statistic	16.836
Significance level P (area = 0.5)	<0.0001

[a]Hanley and McNeil (1982)
[b]AUC ± 1.96 SE

The AUC is interpreted as a probability that a randomly drawn individual from the diseased or abnormal reference sample has a greater test value than a randomly drawn individual from the healthy or normal reference sample. In is clear that if we interpret this as a single member of each group (diseased and non-diseased) taking the test, the probability of a correct answer (the AUC) is not influenced by the prevalence of disease within the sample because each member of the selected pair represents a fixed prevalence at 50 %.

The AUC summarizes the whole of the ROC curve, and therefore all parts of the curve are represented within the AUC. Some parts of the curve can be vertical (lower left part) or horizontal (upper right part) and their contribution to the AUC is less useful because they include cut-off values with increasing Se (without loss of Sp) or increasing Sp (without loss of Se), respectively. Also, we may want to have a fixed Se or Sp for diagnosis (e.g. that Se is at least 80 %), in which case the AUC may not be the best way to compare two tests since the part of the ROC curve below this threshold still contributes to the AUC making this method less optimal for our diagnostic situation.

The 95 % confidence interval is the interval in which the true (population) area under the ROC curve lies with 95 % confidence. The *P* value is the probability that the sample AUC is found when the true (population) AUC is 0.5 (null hypothesis: area = 0.5). If *P* is low (<0.05) then it can be concluded that the AUC is significantly different from 0.5 and therefore there is evidence that the laboratory test does have an ability to distinguish between the two groups. This probability of a correct ranking is the same quantity that is estimated by the non-parametric Wilcoxon statistic and can be used to provide rapid closed-form expressions for the approximate magnitude of the sampling variability, i.e. standard error that one uses to accompany the area under a smoothed ROC curve. Finally, concerning sample size, it has been suggested that meaningful qualitative conclusions can be drawn from ROC experiments performed with a total of about 100 observations. A minimum of 50 cases may be required in each of the two groups, so that one case represents not more than 2 % of the observations.

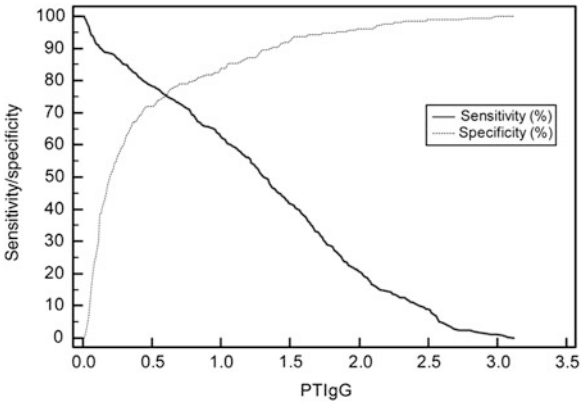

Fig. 6.2 Plot of Se and Sp for different cut-off values of PTIgG (also known as plot vs. criterion values). The criterion was the PCR result

Determining an Optimum Cut-Off Value

If we use the normal distribution to define a value two standard deviations (2SD) above the mean of the normal reference sample, this would result in a cut-off value with a Sp of 97.5 % (since 2SD encompasses 95 % of the population, i.e. 2.5 % on either side of the distribution). This would however not work for skewed or multimodal distributions, and also, it ignores the Se, which is a disadvantage. A better option therefore is to create a table or a plot of Se and Sp for different cut-off values (plot versus criterion values in MedCalc) which can then provide a useful visualization and can also be used to derive two optimal cut-off values: One where good sensitivity is retained and the other where good specificity is retained. This is depicted in the Fig. 6.2 and Table 6.2. Also, it should be kept in mind that the slope of the ROC curve gives us the LR of the test value at that particular cut-off and a table of LR against the cut-off values (see Table 6.2) is an alternative way a cut-off can be selected. Where we choose to place our optimal cut-off will eventually depend on the prevalence of disease in the target population and the consequences of FN versus FP test results (which may differ for every different scenario). For example, a very low prevalence disease with a high cost of false-positive diagnoses may require us to select a cut-off that maximizes Sp. If, on the other hand, for a high prevalence disease where missing a diseased individual has serious consequences, a cut-off value would be selected to maximize Se.

Another alternative is to select the point on the ROC curve closest to the upper left corner of the unit square as this would optimize prevalence-independent summary measures of Se and Sp. The Youden index (Se + Sp − 1) attempts to do this and gives us the optimal or criterion value J corresponding to the maximum of the Youden index; i.e. $J = \max[SE_i + SP_i - 1]$ where SE_i and SP_i are the sensitivity and specificity over all possible threshold values. This value corresponds with the point on the ROC curve farthest from the diagonal line. The MedCalc manual (www.medcalc.org) indicates the following pointers for interpretation of the criterion value:

Table 6.2 Criterion values and coordinates of the ROC curve

Criterion	Sensitivity	95 % CI	Specificity	95 % CI	+LR	95 % CI	–LR	95 % CI
≥0.01	100.00	99.1–100.0	0.00	0.0–1.7	1.00			
>0.05	95.40	92.8–97.2	10.90	7.0–15.9	1.07	0.7–1.6	0.42	0.3–0.7
>0.1	91.30	88.1–93.9	27.49	21.6–34.0	1.26	1.0–1.6	0.32	0.2–0.4
>0.15	89.26	85.8–92.1	41.71	35.0–48.7	1.53	1.3–1.8	0.26	0.2–0.4
>0.2	88.49	84.9–91.5	50.71	43.8–57.6	1.80	1.6–2.1	0.23	0.2–0.3
>0.3	85.17	81.2–88.5	60.66	53.7–67.3	2.17	1.9–2.4	0.24	0.2–0.3
>0.45	79.80	75.5–83.7	71.56	65.0–77.5	2.81	2.5–3.1	0.28	0.2–0.4
>0.64[a]	74.42	69.8–78.7	77.73	71.5–83.2	3.34	3.0–3.7	0.33	0.2–0.4
>0.7	72.89	68.2–77.2	79.15	73.0–84.4	3.50	3.2–3.8	0.34	0.3–0.5
>0.8	69.31	64.5–73.8	80.09	74.1–85.3	3.48	3.2–3.8	0.38	0.3–0.5
>1.35	48.08	43.0–53.2	90.05	85.2–93.7	4.83	4.3–5.4	0.58	0.4–0.9
>1.91	23.53	19.4–28.1	95.26	91.5–97.7	4.96	4.1–6.0	0.80	0.4–1.5
>2.72	2.56	1.2–4.7	99.05	96.6–99.9	2.70	1.5–5.0	0.98	0.2–3.9
>2.8	2.56	1.2–4.7	99.53	97.4–100.0	5.40	2.9–10.0	0.98	0.1–6.9
>2.95	1.28	0.4–3.0	99.53	97.4–100.0	2.70	1.1–6.4	0.99	0.1–7.0
>2.99	1.28	0.4–3.0	100.00	98.3–100.0			0.99	
>3.12	0.00	0.0–0.9	100.00	98.3–100.0			1.00	

[a]Cut-off via the Youden index

- When you select a lower criterion value, then the true-positive fraction and the sensitivity increases. On the other hand, the false-positive fraction also increases, and therefore the true-negative fraction and specificity decrease.
- When you select a higher criterion value, the false-positive fraction decreases with increased specificity but, on the other hand, the true-positive fraction and sensitivity decrease.
- If a test is used for the purpose of screening, then a cut-off value with a higher sensitivity and negative predictive value must be selected. In order to confirm the disease, the cases positive in the screening test can be tested again with a different test. In this second test, a high specificity and positive predictive value are required.

ROC analysis can also be used to evaluate the diagnostic discrimination' of logistic regression models in general as they have binary outcomes. In such an analysis, the power of the model's predicted values to discriminate between positive and negative cases is quantified by the AUC, which is sometimes referred to as the c statistic (or concordance index), and varies from 0.5 (discriminating power not better than chance) to 1.0 (perfect discriminating power). Essentially, we can save the predicted probabilities and use this new variable in ROC curve analysis. The dependent variable used in logistic regression then acts as the classification variable in the ROC curve analysis.

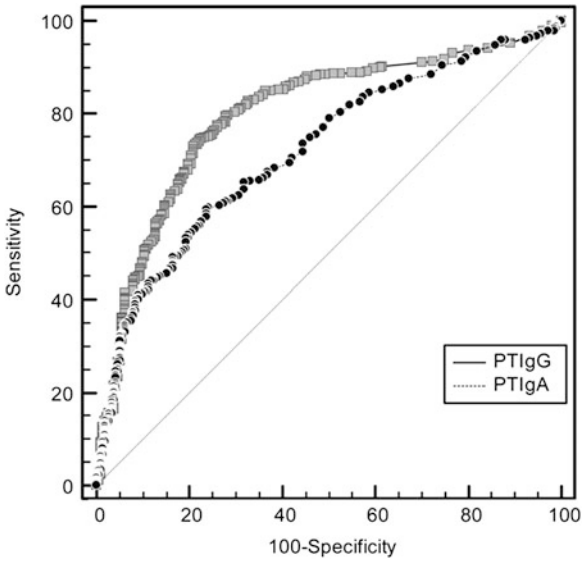

Fig. 6.3 ROC curves comparing plots for PT IgG and PT IgA. Each point on the *curve* represents a single cut-off. The *diagonal* is the line of equality and the higher the plot above this line, the higher its discriminative capacity

Table 6.3 Area under the ROC curves (AUC) for PT IgG and PT IgA

	AUC	SE[a]	95 % CI[b]
PT IgG	0.798	0.0177	0.763–0.832
PT IgA	0.720	0.0206	0.680–0.761

[a]Hanley and McNeil (1982)
[b]AUC ± 1.96SE

Test Comparison

As described above, the AUC represents a summary statistic of the overall diagnostic performance of a test. It makes sense therefore to use the AUC to compare the discriminatory abilities of different tests overall and independent of any specific cut-offs they may have. However, the AUC gives equal weights to the entire ROC curve and it could happen that two tests that differ in terms of optimal sensitivity and specificity have a similar AUC. This may happen when ROC curves cross each other though they may have similar AUC estimates.

The non-parametric area under the plots for the above example (Fig. 6.3) is shown in Tables 6.3 and 6.4. The difference (and 95 % confidence interval) from MedCalc output is shown. The confidence interval for the differences between the tests does not include zero and it can be concluded that there is a statistically significant difference in the AUC estimates and thus the performance of the two tests for pertussis (PT IgG and PT IgA). The better test (PT IgG) is the test with the higher dome and thus greater AUC. It can be seen however that the curves overlap at both ends, suggesting that at these cut-offs, the tests characteristics are identical.

Table 6.4 Pairwise comparison of ROC curves

PT IgG ~ PT IgA	
Difference between areas	0.0775
Standard error[a]	0.0183
95 % confidence interval	0.0417–0.113
Z statistic	4.244
Significance level	$P < 0.0001$

[a]Hanley and McNeil (1982)

Meta-Analysis of Diagnostic Studies

The ROC approach can also be applied to combine multiple estimates of Se and Sp for one test across several primary evaluation studies. The procedure is known as meta-analysis of diagnostic tests. This sort of summary ROC pooling for meta-analysis occurred due to the explosion in the discussion surrounding the implicit threshold across studies of the same radiological investigation, so much so that diagnostic meta-analysis moved from univariate pooling of sensitivity and specificity to summary ROC curves as first defined by Moses et al. (1993) and methods for these have subsequently evolved into hierarchical and bivariate sROC models (see Arends et al. 2008; Reitsma et al. 2005; Rutter and Gatsonis 2001). The basic reasoning was that sensitivity and specificity across studies are negatively correlated and thus study investigators must be using different implicit diagnostic thresholds and thus fit in at different points on an ROC curve. These models were thought to account for the potential presence of a (negative) correlation between sensitivity and specificity within studies and address this explicitly by incorporating this correlation into the analysis. However, Simel and Bossuyt (2009) demonstrate that results from univariate and bivariate methods may be quite similar.

The problem with such an approach is that in reality there may be no implicit diagnostic threshold at play. On the contrary, radiologists might make a diagnosis based on an implicit information size threshold based on the amount of obvious information available to the average radiologist on the image. If images are from very sick persons, they will tend to have a lot more information, thus making it both more likely for a true diagnosis to be made as well as for a false diagnosis to be made. On the other hand, subjects that are not as sick have less information on the image and thus the radiologist will meet the implicit information threshold with difficulty. Thus, while the true-positive rate decreases, so too does the false-positive rate. If we have a set of studies from a varying spectrum of subjects, the Se and Sp are negatively correlated simply on the basis of the varying spectrum of disease – a spectrum effect. There is no change in the implicit diagnostic threshold and chasing such a threshold using sROC models is a questionable pursuit since the goal is ill-defined.

A recent study by Willis (2012) that grouped images by high probability or not according to a trained radiographer and then interpreted by junior doctors revealed exactly this phenomenon. Images with more information content (high probability group) were interpreted with higher Se and lower Sp than the low probability group.

The two groups were said to fit perfectly on the ROC curve and were thus incorrectly interpreted as representing doctors changing their implicit diagnostic threshold rather than a change in the information size from the images making it more difficult for low probability radiographs to meet the information threshold required of the doctors. In such a scenario, the real Se and Sp would actually be the combined Sp and Se based on all radiographs, not high versus low probability on an ROC curve if there are no implicit thresholds.

The same author has previously stated that the Se and Sp may vary between different patient subgroups even when the test threshold remains constant, and this lies at the heart of the concept of the spectrum effect (Willis 2008). The latter effect has not only been mixed up with the concept of an implicit threshold but has also been misleadingly called the spectrum bias (Mulherin and Miller 2002). Such subgroup variation is not a bias and just contributes to heterogeneity across studies; these will lead to estimates of test performance that are not generalizable if the studies are mostly non-representative of their relevant clinical populations.

It has been suggested by Goehring et al. (2004) that in some situations this spectrum effect may lead to a spectrum bias, that is, a distortion of the posterior probability, which can potentially affect the clinical decision. It has been shown that spectrum bias on either a positive or a negative test result can be expressed as the subgroup-specific LR divided by the LR in the overall population of patients (ratio of LR or RLR) and this assessment of spectrum bias has been shown to be independent of the pretest probability. In the usual situation in which sensitivity increases from one patient subgroup to another but the specificity simultaneously decreases, the LRs remain constant and thus while spectrum effects are quite common, spectrum bias is usually not an issue. Nevertheless, despite the absence of bias, sensitivity and specificity on their own may not reflect values that are generalizable to the overall populations that the studies are trying to represent. It has therefore been suggested by Moons et al. (2003) that Se and Sp may have no direct diagnostic meaning because they vary across patient populations and subgroups within populations and thus there is no advantage for researchers in pursuing estimates of a test's Se and Sp rather than post-test probabilities. However, the study by Goehring et al. (2004) clearly demonstrates that the subgroup/population RLR (and thus post-test probability) will not change across subgroups simply due to a spectrum effect because Se and Sp change simultaneously. There is an advantage in pursuing Se and Sp over and above post-test probabilities, and that is to determine, for an average subject of the types represented in the trials, what the expected false-positive and false-negative rates likely to be.

If what we need is the best estimate of Se and Sp across studies that reflects a generalizable population value, then a weighted average of the spectrum of effects across the studies themselves is not necessarily bad. What may lead to spectrum bias, however (as opposed to the spectrum effect), is the methodological rigueur with which the study was conducted and thus a quality assessment is necessary. This is preferable to simply considering the varying effects across studies as random changes because the set of studies cannot be visualized as a random subset from a population of all studies, and therefore the random effects model does not

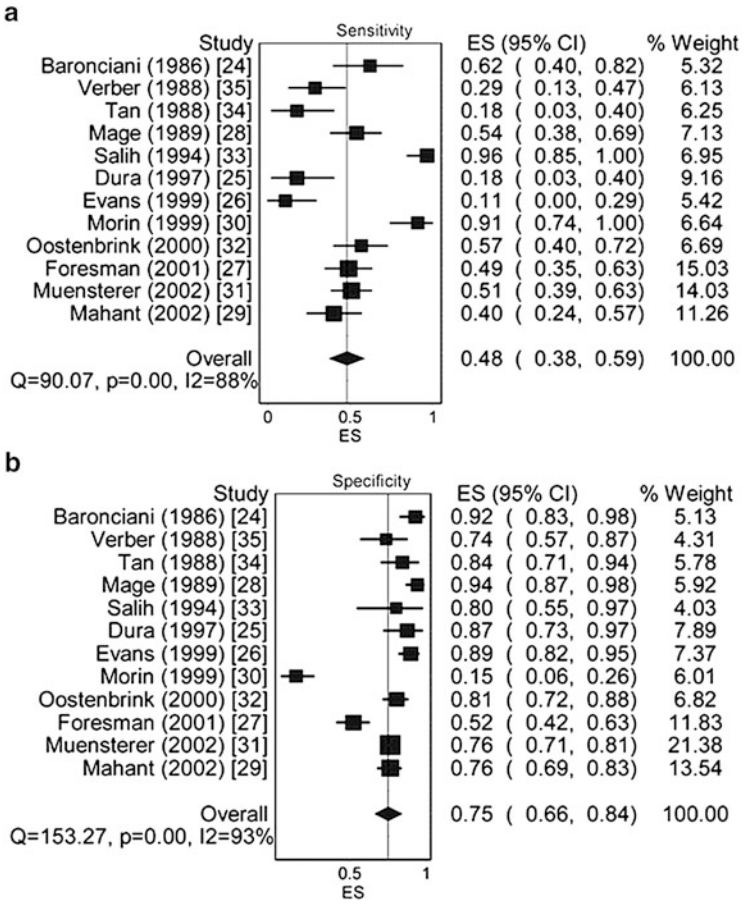

Fig. 6.4 Forest plots depicting the results of univariate bias adjusted meta-analyses of sensitivity and specificity. The box sizes are proportional to the weight given to each study for pooling. The *horizontal lines* are the confidence intervals and the *diamond* depicts the summary sensitivity or specificity. Q = the Cochran Q statistic I2 = the I^2 statistic

apply, at least according to a strict interpretation of randomization in statistical inference. Until recently, there was no simple way of bias adjustment in meta-analysis, but there is currently a quality effects method and software (MetaXL) that implements this (http://www.epigear.com). MetaXL uses a double arcsin square root transformation to stabilize variances of proportions for meta-analysis and provides a method for bias adjustment in addition to the usual inverse variance adjustment.

If we take the example of standard ultrasound data presented by Whiting et al. (2005) for the diagnosis of vesico-ureteral reflux in children, the results of a univariate analysis of Se and Sp are shown in Fig. 6.4. Using the *metandi* procedure in Stata, we can also compute bivariate results and the resulting sROC plot is shown

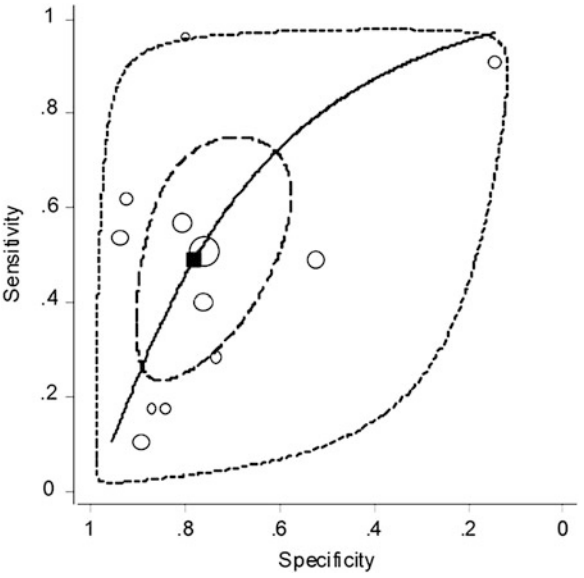

Fig. 6.5 sROC curve generated using the *metandi* procedure in Stata. Study estimates are shown as *circles* sized according to the total number of individuals in each study. Summary (square marker) Se was 49 % (95 % CI 30.6–67.7) and Sp was 78.1 % (95 % CI 64.8–87.3) and the 95 % confidence region for the summary operating point is depicted by the small oval in the centre. The larger oval is the 95 % prediction region (confidence region for a forecast of the true sensitivity and specificity in a future study). The summary curve is from the HSROC model

in Fig. 6.5. The bivariate summary sensitivity was 49 % (95 % CI 30.6–67.7) and the summary specificity was 78.1 % (95 % CI 64.8–87.3). Clearly, these are quite similar to the univariate results in Fig. 6.4 and the added advantage of the univariate plots would be the obvious depiction of the spectrum of effects as well as their correlation with bias, if any.

Bibliography

Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T (2008) Bivariate random effects meta-analysis of ROC curves. Med Decis Mak 28:621–638

Gardner IA, Greiner M (2006) Receiver-operating characteristic curves and likelihood ratios: improvements over traditional methods for the evaluation and application of veterinary clinical pathology tests. Vet Clin Pathol 35:8–17

Goehring C, Perrier A, Morabia A (2004) Spectrum bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance. Stat Med 23:125–135

Griner PF, Mayewski RJ, Mushlin AI, Greenland P (1981) Selection and interpretation of diagnostic tests and procedures. Ann Intern Med 94:555–600

Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36

Harbord RM, Whiting P (2009) Metandi: meta-analysis of diagnostic accuracy using hierarchical logistic regression. Stata J 9:211–229

May ML, Doi SA, King D, Evans J, Robson JM (2012) Prospective evaluation of an Australian pertussis toxin IgG and IgA enzyme immunoassay. Clin Vaccine Immunol 19:190–197

Moons KG, Harrell FE (2003) Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Acad Radiol 10:670–672

Moses LE, Shapiro D, Littenberg B (1993) Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 12:1293–1316

Mulherin SA, Miller WC (2002) Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 137:598–602

Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH (2005) Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 58:982–990

Rutter CM, Gatsonis CA (2001) A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 20:2865–2884

Simel DL, Bossuyt PM (2009) Differences between univariate and bivariate models for summarizing diagnostic accuracy may not be large. J Clin Epidemiol 62:1292–1300

Whiting P, Harbord R, Kleijnen J (2005) No role for quality scores in systematic reviews of diagnostic accuracy studies. BMC Med Res Methodol 5:19

Willis BH (2008) Spectrum bias – why clinicians need to be cautious when applying diagnostic test studies. Fam Pract 25:390–396

Willis BH (2012) Empirical evidence that disease prevalence may affect the performance of diagnostic tests with an implicit threshold: a cross-sectional study. BMJ Open 2:e000746