Chapter 8
An Introduction to Diagnostic Meta-analysis

María Nieves Plana, Víctor Abraira, and Javier Zamora

Abstract Systematic review, and its corresponding statistical analysis, is becoming popular in the literature to assess the diagnostic accuracy of a test. When correctly performed, this research methodology provides fundamental data to inform medical decision making. This chapter reviews key concepts of the meta-analysis of diagnostic test accuracy data, dealing with the particular case in which primary studies report a pair of estimates of sensitivity and specificity. We describe the potential sources of heterogeneity unique to diagnostic test evaluation and we illustrate how to explore this heterogeneity. We distinguish two situations according to the presence or absence of inter-study variability and propose two alternative approaches to the analysis. First, simple methods for statistical pooling are described when accuracy indices of individual studies show a reasonable level of homogeneity. Second, we describe more complex and robust statistical methods that take the paired nature of the accuracy indices and their correlation into account. We end with a description of the analysis of publication bias and enumerate some software tools available to perform the analyses discussed in the chapter.

Introduction

Diagnosis is one of the most prestigious and intellectually appealing clinical tasks among physicians and, usually, the first step in clinical care. Furthermore, because a correct classification of patients according to the presence or absence of a specific clinical condition is essential for both prognosticating and choosing the right treatment, an accurate diagnosis is at the core of high-quality clinical practice. The use of diagnostic tests in clinical practice is generalized. However, introducing

M.N. Plana (✉) • V. Abraira • J. Zamora
Clinical Biostatistics Unit, Hospital Universitario Ramón y Cajal, CIBER en Epidemiología y Salud Pública (CIBERESP) and Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain
e-mail: nieves.plana@hrc.es

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_8,

a test into current diagnostic pathways must be preceded by a systematic assessment of its diagnostic performance.

Assessing the value of a diagnostic test is a multi-phase process involving the test's technical characteristics, its feasibility, accuracy, and impact on different dimensions (diagnostic thinking, treatment decisions and, most importantly, impact on patient outcomes). This assessment also includes the social and economic impact of incorporating the test into the diagnostic pathway. Evaluation studies of diagnostic accuracy, in general, and systematic reviews and meta-analyses of studies on test accuracy, in particular, are instrumental in underpinning evidence-based clinical practice. Meta-analysis is a statistical technique that quantitatively combines and summarizes the results of several studies that have previously been included as part of a systematic review of diagnostic tests. A quantitative synthesis of evidence is not always necessary or possible and it is not uncommon to find very high-quality systematic reviews of great informational value for clinical practice that do not include it. Even when a systematic review fails to provide a definite answer regarding the accuracy of a test, it may still contribute valuable information that fills existing scientific gaps and/or informs the design of future primary research studies.

Of the different evaluative dimensions of a diagnostic test, this chapter focuses on test accuracy. Assessing the diagnostic accuracy of a test consists of analysing its ability to differentiate, under the usual circumstances, between individuals presenting with a specific clinical condition (usually a pathology) and those without the condition. For the purpose of this chapter, we assume that diagnostic test results are reported either as positive or negative. This may reflect the actual outcome of the test (e.g. an imaging test result reported as normal or abnormal) or a simplification of a result reported in an ordinal or continuous scale that is then dichotomized into positive/negative using a pre-established cut-off point as with many laboratory results.

In the next section, we revisit the concept of diagnostic accuracy and how it is measured. In the third section, we describe the potential sources of heterogeneity present in systematic reviews of diagnostic test evaluation and how to explore it. The next two sections present two statistical methodologies to choose from according to the presence or absence of inter-study variability. The following section describes publication bias and its analysis. The last section provides a list of software programs available to perform the analyses discussed in the chapter. An appendix with the output of two examples is included.

Evaluation of Diagnostic Accuracy

In contrast with randomized clinical trials where the results regarding the effectiveness of an intervention are reported using a single coefficient (risk ratio, absolute risk reduction, number needed to treat, etc.), individual studies in evaluations of diagnostic test accuracy are summarized using two estimates, which are often

inter-related. The statistical methods used to summarize the systematic review results must take into account this dual measurement and report both statistical estimates simultaneously.

As mentioned in Chap. 5 on using and interpreting diagnostic tests, there are several diagnostic accuracy paired measures. These paired estimates are obtained from a 2×2 cross-classification table. The two specific indices conditioned to disease status are sensitivity (the proportion of test positives among people with the disease) and specificity (the proportion of test negatives among people without the disease). Predictive values, positive and negative, are measures conditioned to test results and are calculated as the proportion of diseased individuals among people with a positive test result and the proportion of non-diseased individuals among people with a negative test result, respectively. The well-known impact of the actual disease prevalence on these predictive values discourages their use as summary measures of test accuracy. Likelihood ratios (LRs) are another set of indices obtained directly from sensitivity and specificity. These ratios express how much more likely a specific result is among subjects with disease than among subjects without disease. Another measure of test accuracy is the diagnostic odds ratio (dOR). The dOR expresses how much greater the odds of having the disease are for the people with a positive test result than for the people with a negative test result. It is a single indicator of the diagnostic performance of a test because it combines the other estimates of diagnostic performance in one measure.

Both LRs and dOR index are calculated from the sensitivity and specificity indices and, except under special circumstances, although usually not affected by the disease prevalence, they are affected by the disease spectrum. The dOR index is very useful when comparing the overall diagnostic performance of two tests. Furthermore, because it is easily managed in meta-regression models, it is a valuable tool for analysing the effect of predictor variables on the heterogeneity across studies. However, its use for clinical decision making regarding individual patients is questionable given it is a single summary measure of diagnostic accuracy.

Heterogeneity

Before undertaking a meta-analysis of diagnostic accuracy studies as part of a systematic review, the researcher should ponder the appropriateness and significance of the task. Frequently, the large variability present in sensitivity and specificity indices across the individual studies puts into question the suitability of a statistical pooling of results. A preliminary analysis of the clinical and methodological heterogeneity of the studies should provide the necessary information regarding the appropriateness of synthesizing the results. The selection of potential sources of heterogeneity for further exploration must be done a priori, before starting data analyses, in order to avoid spurious findings. Meta-analysis should only be performed when studies have recruited clinically similar patients and have used comparable experimental and reference tests.

Sources of Heterogeneity

Clinical and Methodological Heterogeneity

In addition to the inherent expected random variability in the results, there can be additional sources of heterogeneity as a result of differences in the study populations (e.g. disease severity, presence of comorbidities), the tests under evaluation (differences in technology or among raters), the reference standards, and the way a study was designed and conducted.

In systematic reviews of treatment interventions, the individual studies usually share a standardized study design (randomized controlled trial or RCT), generally designed with comparable inclusion and exclusion criteria, similar interventions and methods to measure the intervention effect (i.e. similar clinical outcome). In contrast, systematic reviews of diagnostic accuracy studies have to contend with a great deal of variability regarding design, including some studies of questionable methodological quality (retrospective case series, non-consecutive case series, case–control studies, etc.). Empirical evidence shows that the presence of certain methodological shortcomings has a substantial impact on the estimates of diagnostic performance. Pooling results from studies with important methodological shortcomings that have recruited different patient samples may lead to biased or incorrect meta-analysis results.

Threshold Effect

A special source of heterogeneity present in the studies of diagnostic accuracy comes from the existence of a trade-off between sensitivity and specificity. This is a result of the studies using, implicitly or explicitly, different thresholds to determine test positivity.

When studies define different positivity criteria, the sensitivity and specificity change in opposite directions. This effect is known as the threshold effect. As we discuss later, the presence of this effect requires that the meta-analysis consider the correlation between the two indices simultaneously while discouraging analytical strategies based on simple pooling of the sensitivity and specificity measures. Consequently, the meta-analysis of diagnostic accuracy adds a certain level of complexity and requires fitting statistical models, taking into account the covariance between sensitivity and specificity.

Study of Heterogeneity

The first step in a meta-analysis is to obtain diagnostic performance estimates from the individual (or primary) studies included in the review. These data are used to estimate the level of consistency across the different studies (heterogeneity

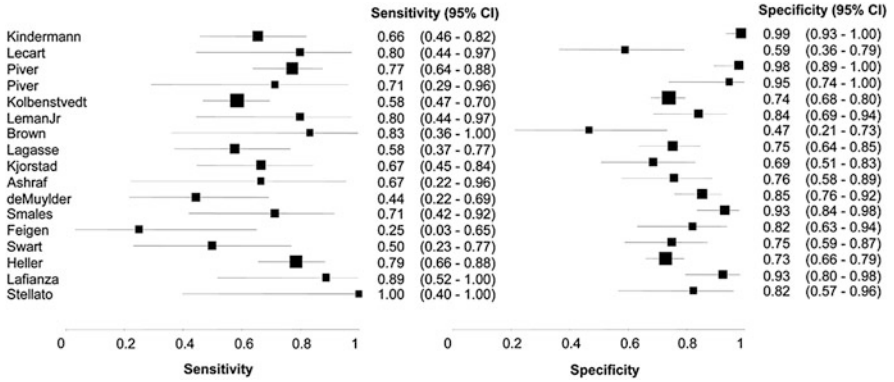

Fig. 8.1 Forest plot of sensitivity and specificity. The box sizes are proportional to the weights assigned and the *horizontal lines* depict the confidence intervals

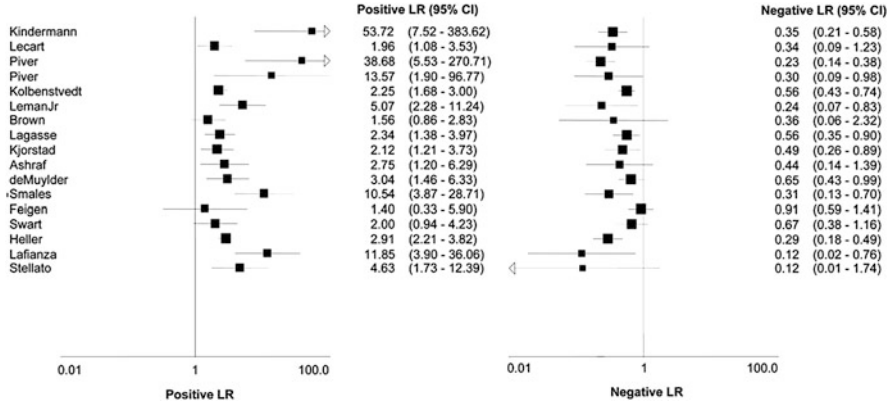

Fig. 8.2 Forest plot of positive and negative LRs. The box sizes are proportional to the weights assigned and the *horizontal lines* depict the confidence intervals

analysis). This description must provide the magnitude and precision of the diagnostic performance indices for every individual study. Given that these accuracy estimates are paired up and are frequently inter-related, it is necessary to report these indices simultaneously (sensitivity and specificity, or positive LR and negative LR). For this description one can use numerical tables of results or paired forest plots (Fig. 8.1) of sensitivity and specificity or of positive and negative LR (Fig. 8.2) for each study together with the corresponding confidence intervals.

A certain level of variability is expected by chance, but the presence of other sources of variation will increase the heterogeneity. These forest plots present the studies ordered from higher to lower sensitivity or specificity (see Fig. 8.4). This format may help analyse consistency among studies and the potential correlation between sensitivity and specificity. However, the best way of illustrating the

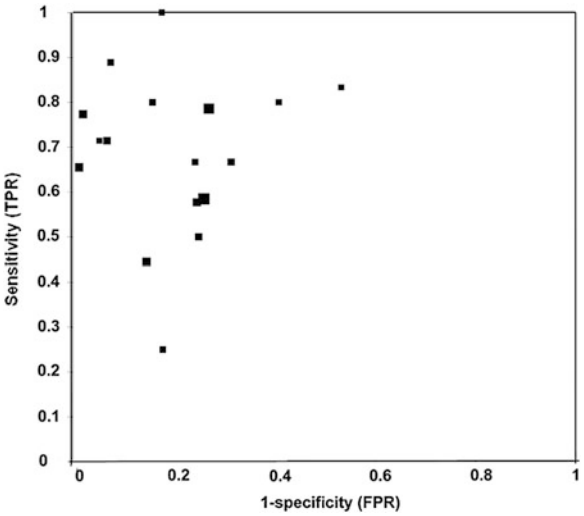

Fig. 8.3 The ROC plane: Plot of 1-specificity against sensitivity

covariance between these indices is to present the pairings of estimates for each study on a receiver operator characteristic (ROC) plot (Fig. 8.3). The *x*-axis of the ROC plot displays the false-positive rate (1-specificity). The *y*-axis shows the corresponding true-positive rate (sensitivity). The higher the diagnostic performance of a study, the closer it is to the upper left quadrant of the ROC plot where both sensitivity and specificity are close to 1. This graphical representation displays a shoulder arm pattern when sensitivity and specificity are correlated, for example, as a result of the presence of a threshold effect or as a result of a different spectrum of the disease among the patients included in the studies. In such situations, sensitivity and specificity are inversely correlated, that is, the true-positive rate (TPR) and the false-positive rate (FPR) are directly correlated.

Specific univariate statistical tests for homogeneity of accuracy estimates have been proposed. However, heterogeneity tests may lack the necessary statistical power to detect heterogeneity when a meta-analysis includes a small number of studies. Conversely, when a meta-analysis includes a large number of studies, heterogeneity tests may detect and interpret slight inter-study variations as strong evidence of heterogeneity by yielding highly significant values, especially when the studies include large sample sizes. In addition, these univariate approaches to heterogeneity analysis do not account for heterogeneity due to the correlation between sensitivity and specificity. The inconsistency index (I^2) may be used to quantify the proportion of total variation across studies beyond what would be expected by chance alone although these estimates must be interpreted with caution.

The results of the heterogeneity analysis must guide the researcher's next step in the completion of the meta-analysis. There are two alternative approaches: (1)

perform separate univariate analyses of the diagnostic accuracy indices; and (2) calculate a pooled estimate of both indices using the appropriate statistical model. Below we describe the two approaches and the circumstances under which one or the other is more appropriate.

Estimate of the Overall Summary Performance of a Diagnostic Test in the Absence of Variability Across Results

The first analytical approach may be used in the special circumstance in which measures of sensitivity, specificity (or both) of the individual studies show a reasonable level of homogeneity. In this scenario, summary estimates of diagnostic accuracy may be obtained through basic meta-analysis techniques with no need for more complex analytical models. Under this approach, two separate poolings of sensitivities and specificities are performed by univariate meta-analysis with fixed or random effects models as deemed appropriate. For added precision, we recommend the use of the logit transformation for sensitivity and specificity to perform the meta-analysis.[1] Once the estimates are averaged, they should be back-transformed to the original scale.

It is important to emphasize that the univariate analysis approach can only be used when there is evidence of homogeneity across estimates. Both sensitivity and specificity indices – and the explicit thresholds defining test positivity, if applicable – must be homogeneous. In this scenario, the correlation between these indices will approach zero and the results of simple pooling will be comparable with those from more advanced models such as bivariate and hierarchical models, discussed later in the chapter. An interesting study concluded that summary indices of diagnostic accuracy calculated with separate simple pooling did not differ significantly from those generated by more statistically robust methods and that the small differences were not clinically relevant.

In the absence of variability across thresholds for test positivity, positive and negative LRs could also be pooled using standard methods such as meta-analysis with fixed or random effects. However, there is some evidence that pooling diagnostic LRs in systematic reviews is not appropriate as the summary LRs generated could correspond to summary sensitivities and specificities outside the valid range from 0 to 1. Instead, it is recommended to calculate the LRs from summary sensitivity and specificity indices estimated using bivariate or hierarchical methods (see below).

We also discourage the practice of averaging predictive values (positive and negative) due to the well-documented effect the prevalence of the disease has on the results. To make matters worse, this prevalence may vary across studies adding an

[1] The standard error of a logit transformed proportion p is computed as the square root of $1/(np(1-p))$.

additional source of heterogeneity to the estimates. The summary predictive value is estimated for unknown average disease prevalence. However, in some cases, it is the only index available given the design characteristics of the studies in which reference standards were performed on test positives but not on test negatives (partial verification bias). A typical example of this scenario is when histopathology is used to confirm imaging findings, and no histological sample can be obtained after a negative image result.

Estimate of the Overall Summary Performance of a Diagnostic Test in the Presence of Variability Across Results (sROC Curve)

It is common for researchers performing meta-analyses to run into substantial variability in diagnostic accuracy indices. This second analytical approach addresses the issue of heterogeneity across individual studies. Part of this variability could well originate in differences in the thresholds of positivity used, either explicit or implicit, across studies. Other source of variation could be a differential spectrum of patients across studies. In these cases, separate pooling is not the appropriate method to calculate a summary measure of test accuracy. Instead, the analysis must start by fitting a summary ROC (sROC) curve modelling the relationship between test accuracy measures. Of the different parameters that have been proposed to summarize a sROC curve, the most common is the area under the curve (AUC). This statistic summarizes the diagnostic test performance with only one figure: a perfect test achieves an AUC close to 1, whereas the AUC is near 0.5 for a useless test. This figure may be interpreted as the probability of the test correctly classifying two random individuals, a diseased and a non-diseased subject. Thus, the AUC may be also a useful tool to compare the performance of various diagnostic tests. Another statistic suggested for this task is the Q^* index, defined as the point of the curve in which sensitivity equals specificity. In a symmetric curve, this is the point closest to the upper left corner of the ROC space. The fitted sROC curve may also be used to calculate a sensitivity estimate from a given specificity or vice versa. Two methods for fitting a sROC curve are discussed below.

Moses–Littenberg Model

The Moses–Littenberg method, initially developed to generate sROC curves easily, is the simplest and most popular method for estimating test performance as part of meta-analyses of diagnostic tests. The shape of the ROC curve depends on the underlying distribution of the test results in patients with and without the disease.

There are two methods of fitting the ROC curve. Diagnostic tests where the dOR is constant regardless of the diagnostic threshold have symmetrical curves around the sensitivity = specificity line. When the dOR changes with diagnostic threshold, the ROC curve is asymmetrical. The Moses–Littenberg method is used to study dOR variation according to threshold and thereby generates symmetrical or asymmetrical curves.

The method consists of studying this relationship by fitting the straight line:

$$D = a + bS$$

$$D = \text{logit}(\text{sensitivity}) - \text{logit}(1 - \text{specificity})$$

$$S = \text{logit}(\text{sensitivity}) + \text{logit}(1 - \text{specificity})$$

where D is the natural logarithm (ln) of the diagnostic odds ratio (dOR) and S is a quantity related to the overall proportion of positive test results. S can be considered as a proxy for the test threshold because S will increase as the overall proportion of test positives increases both in the diseased and non-diseased groups. The contrast in test performance variability (measured by dOR) according to threshold is equivalent to the contrast on the model's parameter b. When $b = 0$ there is no variation and the model generates a symmetrical sROC curve; whereas when $b \neq 0$, performance varies according to the threshold and the resulting sROC curve is asymmetrical. The fitting of the previous linear model can be weighted using the inverse variance of ln(dOR) to account for inter-study differences in the sampling error in D.

The model may be expanded to analyse the effect of other factors on diagnostic performance (dOR) as a supplement to the exploration of heterogeneity described here. Such factors, which would be included in the model as covariates, may capture characteristics related to the study design, the patients, or the test.

The Moses–Littenberg model, although very useful for studies of an exploratory nature, is not adequate for drawing statistical inferences. Thus, it should be used keeping in mind some important limitations. First, the model does not take into account either the correlation between sensitivity and specificity or the different precision with which the indices were estimated. In addition, the model's independent variable is random and, thus, its inherent measurement error violates the basic assumption of linear regression models. Finally, the model must be empirically adjusted to avoid empty cells by adding an arbitrary correction factor (0.5).

Bivariate and Hierarchical Models

Two models have been put forward to overcome the limitations ascribed to the Moses–Littenberg model: the bivariate model and the hierarchical sROC model (HSROC). These random effects models are substantially more robust from the

statistical point of view than the Moses–Littenberg model. The methodological literature relevant to this area of research proposed these models as the gold standard in meta-analyses of diagnostic accuracy studies. The differences between the two models are small and, in the absence of covariates, both approaches simply amount to different parameterizations of the same model.

The bivariate model is a random effects model based on the assumption that logit (sensitivity) and logit (specificity) follow a normal bivariate distribution. The model allows for the potential correlation between the two indices, manages the different precision of the sensitivity and specificity estimates, and includes an additional source of heterogeneity due to inter-study variance. The second model the methodological literature proposes is known as the hierarchical model or HSROC. It is similar to the previous model except that it explicitly addresses the relationship between sensitivity and specificity using the threshold. Similar to the previous model, this one also accounts for the inter-study heterogeneity.

Both models allow fitting an sROC curve and provide a summary estimate of sensitivity and specificity with the corresponding confidence and prediction intervals. After fitting either of these models, we have to select the most appropriate result to report. It depends on the variability of the results of the individual studies. When sensitivities and specificities of these studies vary substantially, it is advisable to forego average indices and, instead, report the sROC curve. In contrast, when the variability across indices is small, the recommendation is to report the average sensitivity and specificity as calculated based on the bivariate (or the hierarchical) model with its 95 % confidence interval. This is much preferred to the alternative, which would entail risking extrapolating to the ROC space a curve that may misrepresent the test diagnostic accuracy. Summary LRs can be calculated from the pooled estimates of sensitivity and specificity generated by these models. It is worth noting that when an average sensitivity and specificity point is reported over the sROC curve, the position represents the midpoint of the results of the studies calculated based on the average threshold for test positivity, or the average spectrum of the disease, observed in the sample.

Publication Bias

Identifying articles about diagnosis is more cumbersome than finding published clinical trials for a review of intervention performance. Although the MeSH (Medical Subject Heading) term “randomized controlled trial” effectively describes and leads researchers to studies describing clinical interventions, there is no comparable term for the specific literature describing the design of such studies. We should take into account, however, that many studies on diagnosis are based on observations of actual clinical practice in the absence of protocols recorded and/or approved by research ethics committees. For this reason, it is difficult not only to follow up these studies but also to get their results published at the level of detail necessary to be fully useful. If the studies identified in the search were to differ

systematically from unpublished manuscripts, the meta-analysis would yield biased estimates that would fail to reflect the real value of diagnostic accuracy.

Similarly, it is also more complex to assess publication bias regarding studies about diagnosis than about treatment. Graphical tools (funnel plots) and the traditional statistical comparisons to evaluate the asymmetry of these graphs were developed to assess publication bias in systematic reviews of clinical trials. Thus, their validity to assess bias in reviews of diagnostic tests is questionable. Deeks and colleagues have adapted the statistical tests of asymmetry of funnel plots to address the issues inherent to meta-analyses of test accuracy. In this version, the funnel plot represents the dOR versus the inverse of the square root of the effective sample size (ESS), which ultimately is a function of the number of diseased and non-diseased individuals. The degree of asymmetry in the plot is statistically evaluated by a regression of the natural logarithm of dOR against $1/ESS^{1/2}$, weighted by ESS.

Software

There is a great variety of statistical packages able to perform the analyses described. Some, like SAS and STATA, are packages for general statistical purposes that facilitate the calculations mentioned by means of a series of macros and user-written commands. The best known user-written commands are the STATA commands MIDAS and METANDI and the SAS macro named METADAS. In addition, the package DiagMeta (http://www.diagmeta.info) was developed for the R environment and it also performs the analyses described.

Additional programs specific to the meta-analysis of diagnostic test accuracy studies are Meta-DiSc and Review Manager (RevMan) by the Cochrane Collaboration. Both perform the basic analyses described in this chapter and RevMan also allows the user to enter parameters obtained from bivariate and hierarchical models and produce corresponding sROC plots.

Appendix

Example 1

For this example we selected the 17 studies included in Scheidler et al.'s meta-analysis (Table 8.1). In their meta-analysis, they evaluated the diagnostic accuracy of lymphangiography (LAG) to detect lymphatic metastasis in patients with cervical cancer.

First, the indices of diagnostic accuracy, sensitivity and specificity (Fig. 8.1) or the positive and negative LRs (Fig. 8.2) of the reviewed studies are described for exploratory purposes using paired forest plots as obtained with Meta-DiSc.

Table 8.1 The studies included in Scheidler et al.'s meta-analysis

id	Study	Year	Test	tp	fp	fn	tn
1	Kindermann	1970	LAG	19	1	10	81
2	Lecart	1971	LAG	8	9	2	13
3	Piver	1971	LAG	41	1	12	49
4	Piver	1973	LAG	5	1	2	18
5	Kolbenstvedt	1975	LAG	45	58	32	165
6	Leman Jr	1975	LAG	8	6	2	32
7	Brown	1979	LAG	5	8	1	7
8	Lagasse	1979	LAG	15	17	11	52
9	Kjorstad	1980	LAG	16	11	8	24
10	Ashraf	1982	LAG	4	8	2	25
11	deMuylder	1984	LAG	8	12	10	70
12	Smales	1986	LAG	10	4	4	55
13	Feigen	1987	LAG	2	5	6	23
14	Swart	1989	LAG	7	10	7	30
15	Heller	1990	LAG	44	50	12	135
16	Lafianza	1990	LAG	8	3	1	37
17	Stellato	1992	LAG	4	3	0	14

From Scheidler et al. (1997)

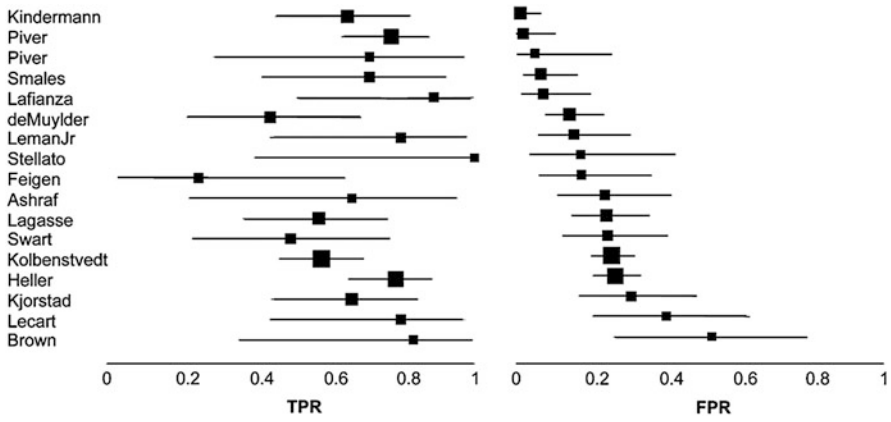

Fig. 8.4 Forest plot with studies sorted by FPR: Heterogeneity is evident

Second, still within the graphical data exploration, we can illustrate the TPR or sensitivity, the FPR (i.e. 1 − specificity), and the LRs (LR + and LR−) organized by one of these indices (Fig. 8.4) or illustrate the pairing indices on a ROC space (Fig. 8.3). At this exploratory phase, all graphical representations should not include pooled estimates of accuracy.

To perform these exploratory analyses, we can use free software (Meta-DiSc, RevMan or the DiagMeta package in the R environment) or any other commercial software.

In this example, and looking at the forest plot generated, we cannot rule out the presence of heterogeneity across the studies included in the review; thus, the analysis should focus on fitting an sROC model.

Given the limitations of the Moses–Littenberg model, we fit a bivariate model using the DiagMeta package. The output is presented below:

```
> bivarROC(Scheidler)
```

	ML	MCMC	lower limit	upper limit
average TPR%	**67.38561**	67.59189	60.52091	74.75159
average FPR%	**16.22516**	16.05203	9.25013	25.49491
SD logit TPR	0.34943	0.31889	0.04271	0.87571
SD logit FPR	0.90087	1.06136	0.63934	1.84290
correlation	−0.23882	−0.53898	−0.99999	0.59240

Because the estimated correlation between logit (sensitivity) and logit (specificity) is small and it cannot be ruled out that it is not different from zero, the results estimated by the bivariate model do not significantly differ from those obtained through separate pooling of sensitivity and specificity. Based on the same example, the results using a simple pooling with a fixed or random effects model according to the variability of each of the indices are as follows:

```
> twouni(subset(Scheidler,GROUP=='LAG'))
```

TPR	TPR	lower limit	upper limit
Fixed effects	0.6711590	6.218139e-01	0.7169960
Random effects from ML	0.6763973	6.056993e-01	0.7398633
Random effects from MCMC	0.6729242	6.148178e-01	0.7327660
SD of REff	0.0692814	5.935713e-07	0.7516062
FPR	FPR	lower limit	upper limit
Fixed effects	0.1996143	0.1764147	0.2250311
Random effects from ML	0.1619847	0.1059149	0.2397768
Random effects from MCMC	0.1631190	0.1035210	0.2426649
SD of REff	0.9576528	0.5716529	1.5829222

Figure 8.5 shows the sROC curve fitted with a STATA bivariate model, together with the estimated summary point and confidence and prediction intervals.

Example 2

For this illustration we used Fahey et al.'s data (Table 8.2). The goal of their study was to estimate the accuracy of the Papanicolaou (Pap) test for detection of cervical cancer and precancerous lesions.

The sensitivity and specificity forest plots (data not shown) confirm the presence of substantial heterogeneity, in both indices, across the studies included in the review. Figure 8.6 shows the representation of the studies in the ROC space.

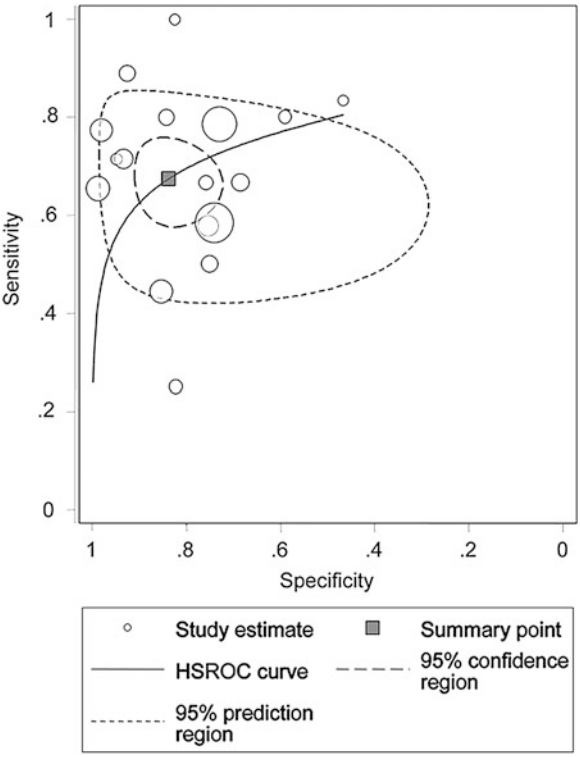

Fig. 8.5 Fitted SROC curve: Study estimates are shown as *circles* sized according to the total number of individuals in each study. Summary sensitivity and specificity are depicted by the *square marker* and the 95 % confidence region for the summary operating point is depicted by the *small oval* in 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

The slight curvilinear pattern of their distribution suggests the presence of a correlation between sensitivity and specificity.

Using Meta-DiSc we calculated the Spearman correlation coefficient between the TPR and FPR logits and obtained a positive and statistically significant correlation of 0.584 ($p < 0.001$) which confirms the results of the bivariate adjustment obtained using the package DiagMeta:

Estimates and 95 % confidence intervals from mcmc samples				
	ML	MCMC median	lower limit	upper limit
average TPR%	65.56718	64.93881	57.58497	72.49102
average FPR%	25.38124	25.27866	18.74132	32.57494
SD logit TPR	1.21834	1.27374	1.04000	1.59237
SD logit FPR	1.22834	1.27623	1.02968	1.60834
correlation	0.77408	0.77709	0.61593	0.87730
Posterior probability that rho positive 1				
Correlation positive - threshold model appropriate				

Table 8.2 Data from Fahy et al.'s study

id	Study	tp	fp	fn	tn	id	Study	tp	Fp	fn	tn
1	Ajons-van K	31	3	43	14	31	Morrison BW	23	50	10	44
2	Alloub	8	3	23	84	32	Morrison EAB	11	1	1	2
3	Anderson 1	70	12	121	25	33	Nyirjesy	65	13	42	13
4	Anderson 2	65	10	6	6	34	Okagaki	1,270	927	263	1,085
5	Anderson 3	20	3	19	4	35	Oyer	223	22	74	83
6	Andrews	35	92	20	156	36	Parker	154	30	20	237
7	August	39	7	111	271	37	Pearlstone	6	2	12	81
8	Bigrigg	567	117	140	157	38	Ramirez	7	4	3	4
9	Bolger	25	37	11	18	39	Reid	12	5	11	60
10	Byrne	38	28	17	37	40	Robertson	348	41	212	103
11	Chomet	45	35	15	48	41	Schauberger	8	4	11	34
12	Engineer	71	87	10	306	42	Shaw	12	2	6	0
13	Fletcher	4	0	36	5	43	Singh	95	9	2	1
14	Frisch	2	2	3	21	44	Skehan	40	18	20	19
15	Giles 1	5	9	3	182	45	Smith	71	13	20	18
16	Giles 2	38	21	7	62	46	Soost	1205	186	454	250
17	Gunderson	4	2	16	31	47	Soutter 1	40	20	17	27
18	Haddad	87	13	12	9	48	Soutter 2	35	9	12	12
19	Hellberg	15	3	65	15	49	Spitzer	10	31	5	32
20	Helmerhorst	41	1	61	29	50	Staff	3	5	3	15
21	Hirschowitz	76	12	11	12	51	Syrjanen	118	40	44	183
22	Jones 1	3	0	5	1	52	Szarewski	13	3	82	17
23	Jones 2	10	4	48	174	53	Tait	38	14	13	62
24	Kashimura 1	28	11	28	77	54	Tawa	16	25	67	291
25	Kashimura 2	79	26	13	182	55	Tay	12	14	6	12
26	Kealy	61	20	27	35	56	Upadhyay	238	52	2	16
27	Kooning 1	62	20	16	49	57	Walker	111	44	20	39
28	Kooning 2	284	31	68	68	58	Wetrich	491	164	250	702
29	Kwikkel	66	25	20	44	59	Wheelock	48	16	38	31
30	Maggi	40	43	12	47						

From Fahey et al. (1995)

With this information in hand, we conclude that the most appropriate method to summarize the results of the meta-analysis is using an sROC curve (Fig. 8.7). This curve was fitted using the bivariate model produced by the macro METANDI in STATA. Figure 8.8 shows the results of a comparable analysis with Meta-DiSc using the Moses–Littenberg model which, in this case, has generated a practically identical sROC curve to that in Fig. 8.7.

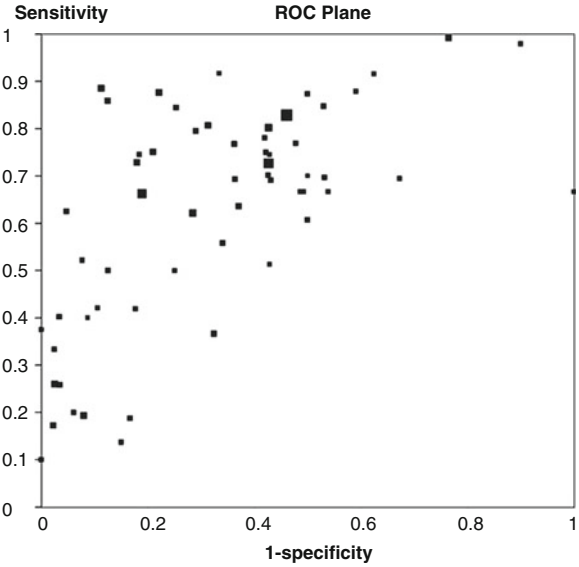

Fig. 8.6 ROC plane: Plot of 1-specificity versus sensitivity

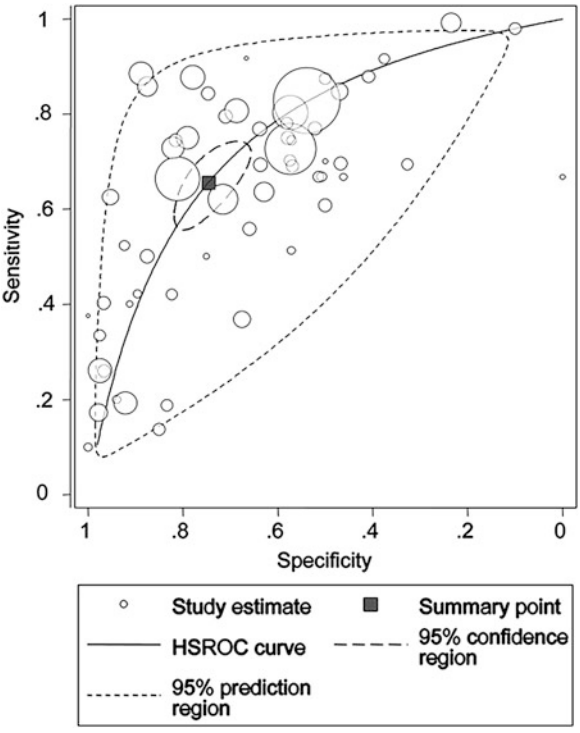

Fig. 8.7 Fitted SROC curve (bivariate model)

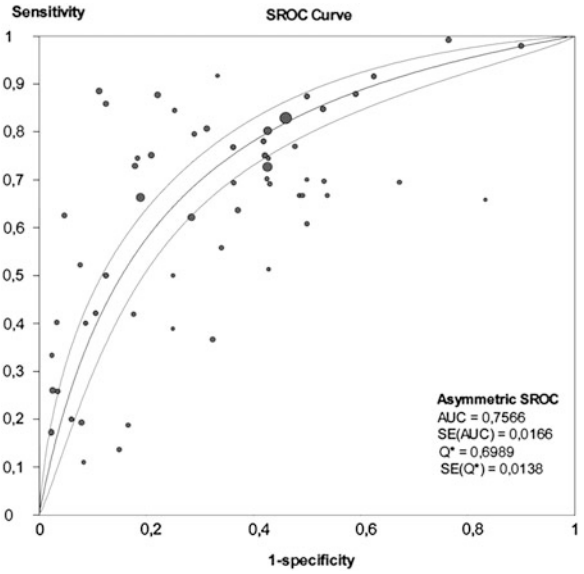

Fig. 8.8 Fitted SROC curve using the Moses–Littenberg model

Bibliography

Begg CB (1994) Publication bias. In: Cooper J, Hedges LV (eds) The handbook of research synthesis. Sage Foundation, New York

Chappell FM, Raab GM, Wardlaw JM (2009) When are summary ROC curves appropriate for diagnostic meta-analyses? Stat Med 28:2653–2668

Deeks JJ, Macaskill P, Irwig L (2005) The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. J Clin Epidemiol 58:882–893

Dwamena BA (2007) Midas: a program for meta-analytical Integration of diagnostic accuracy studies in Stata. Division of Nuclear Medicine, Department of Radiology, University of Michigan Medical School, Ann Arbor

Fahey MT, Irwig L, Macaskill P (1995) Meta-analysis of Pap test accuracy. Am J Epidemiol 141:680–689

Gatsonis C, Paliwal P (2006) Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. AJR Am J Roentgenol 187:271–281

Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM (2003) The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol 56:1129–1135

Harbord RM (2008) Metandi: Stata module for meta-analysis of diagnostic accuracy. Statistical software components. Boston College Department of Economics. Chestnut Hill MA, USA

Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA (2007) A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 8:239–251

Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A, Bachmann LM (2008) An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. J Clin Epidemiol 61:1095–1103

Higgins JP, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. BMJ 327:557–560

Honest H, Khan KS (2002) Reporting of measures of accuracy in systematic reviews of diagnostic literature. BMC Health Serv Res 2:4

Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS (1992) Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Ann Intern Med 117:135–140

Leeflang MM, Bossuyt PM, Irwig L (2009) Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol 62:5–12

Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM (1999) Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 282:1061–1066

Lijmer JG, Bossuyt PM, Heisterkamp SH (2002) Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med 21:1525–1537

Lijmer JG, Leeflang M, Bossuyt PMM (2009) Proposals for a phased evaluation of medical tests. Medical tests-white paper series. Agency for Healthcare Research and Quality, Rockville. Bookshelf ID: NBK49467

METADAS (2008) A SAS macro for meta-analysis of diagnostic accuracy studies. User guide version 1.0 beta. http://srdta.cochrane.org/en/clib.html. Accessed 3 July 2009

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

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

Review Manager (RevMan) (2008) Version 5.0. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration

Rutjes AW, Reitsma JB, Di NM, Smidt N, van Rijn JC, Bossuyt PM (2006) Evidence of bias and variation in diagnostic accuracy studies. CMAJ 174:469–476

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

Scheidler J, Hricak H, Yu KK, Subak L, Segal MR (1997) Radiological evaluation of lymph node metastases in patients with cervical cancer: a meta-analysis. JAMA 278:1096–1101

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

Walter SD (2005) The partial area under the summary ROC curve. Stat Med 24:2025–2040

Whiting PF, Sterne JA, Westwood ME, Bachmann LM, Harbord R, Egger M, Deeks JJ (2008) Graphical presentation of diagnostic information. BMC Med Res Methodol 8:20

Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A (2006) Meta-DiSc: a software for meta-analysis of test accuracy data. BMC Med Res Methodol 6:31

Zwinderman AH, Bossuyt PM (2008) We should not pool diagnostic likelihood ratios in systematic reviews. Stat Med 27:687–697