



Contents lists available at SciVerse ScienceDirect

Autoimmunity Reviews

journal homepage: www.elsevier.com/locate/autrev

Review

Tools for critical appraisal of evidence in studies of diagnostic accuracy

Davide Giavarina*

Clinical Pathology Dept., St. Bortolo Hospital, Vicenza, Italy

ARTICLE INFO

Available online xxxx

Keywords:

Sensitivity
Specificity
Meta-analysis
Forest plot
Summary ROC curve
Diagnostic test accuracy

ABSTRACT

Studies of accuracy are often more complex to understand than clinical trials, since there can be more than one outcome and scope (screening, diagnosis, and prognosis) and because results have to be reported in more than one way, than in clinical trials (relative risk or odds ratio). Sensitivity and specificity are common terms for practitioners, but to remember that sensitivity is the “ratio between true positive rate and true positive rate plus false negative rate” may sometime cause some frustration. Moreover, likelihood ratio, predictive values, diagnostic odds ratio, and pre- and post-test probability complicate the framework. To summarize these indexes from multiple studies can be also a little more difficult. However, understanding diagnostic test accuracy from different study results and how to interpret systematic reviews and meta-analysis can help every practitioner improve critical appraisal of evidence about the best use of diagnostic tests. Avoiding complicated mathematical formulas, this paper attempts to explain the meaning of the most important diagnostic indexes and how to read a Forest plot and a summary Receiver Operating Characteristic curve.

© 2012 Elsevier B.V. All rights reserved.

Contents

1. Introduction	0
2. Evaluating the effectiveness of a diagnostic test	0
2.1. Differences between clinical effectiveness and diagnostic efficacy	0
2.2. Indexes of diagnostic efficacy	0
2.2.1. Sensitivity and specificity	0
2.2.2. Positive and negative predictive values	0
2.2.3. ROC curve	0
2.2.4. Positive and negative likelihood ratio	0
2.2.5. Diagnostic odds ratio	0
3. Confidence intervals	0
4. Systematic review and meta-analysis	0
4.1. Meta-analysis	0
4.2. Summary Receiver Operating Characteristic curve	0
5. Conclusions	0
Take-home messages	0
References	0

Abbreviations: AUC, Area under curve; CER, Control Event Rate; CI, Confidence intervals; DOR, Diagnostic odds ratio; DTA, Diagnostic test accuracy; EBM, Evidence-Based Medicine; EBP, Evidence-based practice; EER, Experimental Event Rate; FN, False negative; FP, False positive; LR, Likelihood ratio; OR, Odds ratio; PV−, Negative predictive value; PV+, Positive predictive value; RCT, Randomized controlled trial; ROC, Receiver Operating Characteristic; RR, Relative risk; Se, Sensitivity; Sp, Specificity; sROC, Summary Receiver Operating Characteristic; TN, True negative; TP, True positive.

* Clinical Chemistry and Haematology Laboratory, ULSS 6 Vicenza, Viale Rodolfi 37, I-36100 Vicenza, Italy. Tel.: +39 0444 753372; fax: +39 0444 752501.

E-mail address: davide.giavarina@ulssvicenza.it.

1. Introduction

“Do the right thing”. In a nutshell, this is what the Evidence-Based Medicine (EBM) movement would like to teach; that is: “At every opportunity, use the most accurate diagnostic test, the most effective treatment, for the right patient, at the right time and without wasting resources”.

The EBM movement was officially launched in 1992 in the Journal of the American Medical Association [1] (actually, the definition was used for the first time in an editorial of the American College of Pathologists

Club Journal by Guyatt, in 1991 [2]) and the statement said: “Use of strongest available data to make informed, unbiased decisions about the diagnosis and treatment of patients”. Deriving from the works of Archibald Cochrane, the main aim of EBM was to demonstrate the efficacy of health treatment, since there cannot be efficiency without effectiveness [3]. The quality improvement of primary clinical studies and a critical appraisal of the available evidence became the goal for EBM researchers.

Today, after 20 years, the EBM movement is more oriented to the clinical practice, to apply the principles in operational contexts of health, integrating the best evidence with the practitioners' experience and patients' expectations. While finding and appraisal of evidence are matters of study for methodologists and researchers working for systematic reviews, for most clinicians the challenge is to apply the best evidence for decision-making. However, an EBM process and some tools are needed also for good decision-making. The application of EBM to daily practice comprises five steps, referred to as the EBM A-5 cycle: Ask – structure the question that states the problem; Acquire – find the evidence; Appraise – determine whether the evidence is reliable and usable (since critical appraisal is the use of explicit, transparent methods to assess the data in published research, applying the rules of evidence to factors such as internal validity, adherence to reporting standards, conclusions and generalizability); Apply – take the evidence and apply it to your problem; Audit – assess the process you have used for effectiveness.

Systematic reviews remain the best publications giving complete summaries of all the available evidence, pooled and weighted. This type of publication responds to a defined question, reviews all the literature on the topic and summarizes the evidence using meta-analysis statistics to sum up the results. Although results of systematic reviews are usually explained and commented upon, understanding meta-analyses is a required skill for the practitioner so that he can make his own appraisal and make his own decisions for the care of his patients. More skills are needed to understand and evaluate studies of diagnostic accuracy than of clinical trials, because the statistics are more numerous and pooling statistics are more complex.

The aim of this paper is to present a review of the major indexes of accuracy and then to consider the opportunities and issues related to the synthesis of evidence of diagnostic accuracy.

2. Evaluating the effectiveness of a diagnostic test

2.1. Differences between clinical effectiveness and diagnostic efficacy

Beginning with available evidence, it is easier to evaluate the effectiveness of a given treatment than to evaluate a diagnostic test for accuracy. Two reasons can be suggested. Firstly, there are few randomized controlled trials (RCTs) on the diagnostic accuracy, due to less economic interest and less “culture” in comparing different tests for the same diagnosis. Secondly, the outcomes of a diagnostic test are not well defined but they depend on the actions arising from and following the diagnosis or lack thereof. Patient outcome rarely depends only on the efficacy of the diagnostic test; the action that follows the diagnosis is critical to the outcome (Fig. 1).

But even if we evaluate the efficacy of the test based only on discriminative diagnostic ability, the evaluation shows important differences compared to studies of therapeutic efficiency.

Randomized control studies provide the strongest evidence for (relative) efficacy of treatment. In such studies, and generally in all the evaluation studies of medical and surgical therapy, a single measure of effectiveness is assessed: the Relative Risk (RR), or alternatively and more commonly, the Odds Ratio (OR). To calculate these indexes, we evaluate the presence of the desired (or adverse) outcome in the two arms of the study, where in the first it is applied to the treatment under study, and in the other, to placebo, current therapy or no treatment.

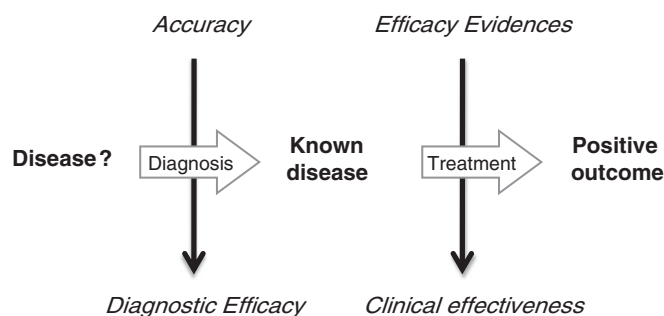


Fig. 1. Interaction between diagnosis and treatment to determine the outcome.

The percentage of outcomes in the treated subjects defines the Experimental Event Rate (EER); the percentage in the controls, defines the Control Event Rate (CER). The ratio of EER and CER is the RR.

The OR of a treatment is the ratio of the frequency of the outcome in the treated group and the frequency of the same event in the control group (Table 1). One single index, derived from many different studies, can be easily summarized in the Forest plot of meta-analysis.

In studies of diagnostic test accuracy (DTA) there are at least two, but frequently more than two, efficacy indexes: sensitivity (Se) and specificity (Sp), but also positive predictive value (PV+) and negative predictive value (PV−), positive and negative likelihood ratio (LR), diagnostic odds ratio (DOR), Receiver Operating Characteristic (ROC) curve. It can be a little more difficult to summarize these indexes; to do so, we first have to understand the meaning and the power of each index.

2.2. Indexes of diagnostic efficacy

2.2.1. Sensitivity and specificity

Every doctor has encountered Se and Sp many times during his working life, and sometimes everyone experiences frustration trying to remember the definitions: that $Se = TP / (TP + FN)$, where TP is the number of true positives and FN is the number of false negatives; and similarly, that $Sp = TN / (TN + FP)$, where TN is the number of true negatives and FP is the number of false positives.

Se and Sp were firstly used as indexes of the accuracy of a diagnostic test in 1947 [4]. Se is the ability to recognize the “positive”, the ill, while Sp is the ability to recognize the “negative”, the healthy people. If the population of healthy people and ill people were completely separated with respect to the concentration of a biomarker or a test, a correct decision level might be able to divide the healthy from the diseased. In reality, as shown in Fig. 2, the distribution of values in healthy and in ill people frequently overlaps. Any chosen cutoff level will cause a number of false positive and/or false negative results. In the example given in Fig. 2, among 100 healthy people, the chosen cutoff correctly classifies 93 people (TN), while seven are classified as positive (FP); this is the Sp of the test at this cutoff. The ability to recognize the negative in this case is $93/100 = 0.93$. The same cutoff correctly classifies 97 people with the disease (TP), while 3 are misclassified (FN). This is the Se of the test at this cutoff. The ability to recognize the positive in this case is $97/100 = 0.97$. The dashed squares represent the area of calculation: Sp is calculated in the healthy people, which are expected to be

Table 1

A simple way to calculate the relative risk and the odds ratio.

	Outcome present	Outcome absent
Diseased patients	A	B
Control patients	C	D

Relative risk: $RR = EER / CER$.

Odds ratio: $OR = (A/B) / (C/D)$.

Experimental Event Rate (EER) = $A / (A + B)$.

Control Event Rate (CER) = $C / (C + D)$.

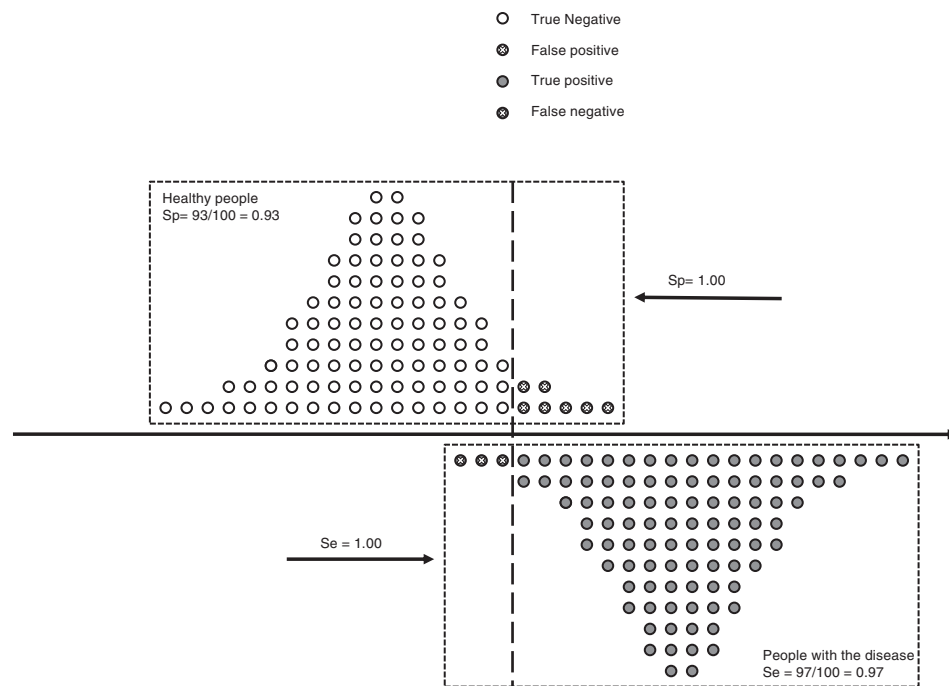


Fig. 2. Specificity measures the discriminating ability in healthy subjects (negative): sensitivity is the quality to recognize the “positive”, i.e. subjects with the disease.

negative; Se is calculated in people with the disease, which are expected to turn out positive.

2.2.2. Positive and negative predictive values

Positive predictive values (PV+) for a diagnostic test are conditional probabilities that a patient has a target disease, given that the test result is positive $[TP/(TP + FP)]$; and negative predictive value (PV−) is the probability that a patient does not have the target disease, given that the test result is negative $[TN/(TN + FN)]$ [5,6].

However, it should be remarked that the predictive value is not a good index of accuracy, since it is very close to the pre-test probability of the result. For instance, if we again consider the example in Fig. 2, that represents a test with $Se = 0.97$ and $Sp = 0.93$, the actual condition of the operating test on 100 healthy people and 100 people with the disease gives a pre-test probability equal to 0.50. When we have a positive result, it could be 1 of the 97 TP or 1 of the 7 FP; so, the PV+ is $97/(97 + 7) = 0.93$. But a pre-test probability of 0.5, which indicates a disease prevalence of 50%, is very high and uncommon; usually, disease prevalence is much lower and hence the pre-test probability. Let us hypothesize a disease prevalence of 9%, with 1000 healthy people and 100 subjects with the disease. In this situation, the pre-test probability is 0.09 (100/1100). Using the same test, with $Se = 0.97$ and $Sp = 0.93$, we will have the same number of TP, 97 out of 100, but 10-fold the number of false positives, that is 70 instead of 7. So, a positive result could be 1 of the 97 TP or 1 of the 70 FP; the PV+ will be $97/(97 + 70) = 0.58$. If instead the pre-test probability were 0.0009 (10,000 healthy and 100 ill), the PV+ will be $97/(97 + 700) = 0.12$. Similar considerations can be made for PV−.

Therefore, reporting PV+ and PV− is not recommended and when given, it has to be negatively appraised, since the pre-test probabilities during clinical studies are often higher than in the actual operational situation.

2.2.3. ROC curve

Most diagnostic tests are not dichotomous (positive or negative) but rather have continuous or categorical values and the definition of positive and negative results depends on the setting of the threshold. Consequently, Se and Sp depend on the cutoff level of the test.

The pattern of Se and Sp observed when the performance of a test is evaluated at several diagnostic thresholds is represented in the ROC curve [7]. Particularly, a ROC curve depicts the true positive rate (Se) as function of false positive rate (1-Sp); Se is on the y axis, while 1-Sp is on the x axis. Fig. 3 can help to convey the relationship between the threshold and the ROC curve. In this example, the positive results are lower than the negative ones (the positive population is on the left). Starting with the threshold in A, the true positive rate (Se) is 1.0, since all the positives are recognized, but also all the negatives are classified as positives, i.e. the false positive rate (1-Sp) is 1.0 too ($Sp = 0.0$). Moving the threshold from A to B, the accuracy in recognizing positives (Se) remains maximum (1.0) but the false positive rate is progressively reduced as Sp increases, in this example to 0.7. After the B threshold the improving of Sp causes a reduction of Se, to the threshold in C, when the false positive rate (1-Sp) becomes zero ($Sp = 1.0$). After the threshold C towards D there is an unhelpful loss of Se, since Sp is already maximum.

The curve generated by varying the threshold value for positivity over the entire possible range delimits an area (area under the curve, AUC) that is a measure of the general ability of the test to discriminate those subjects with the disease from those without. A straight line arising from the origin with a slope of 1 (angle of 45°) provides an AUC of 50%, that is like a discriminating ability equal to that of flipping a coin. The higher the AUC, the better is the test's performance, so the overall diagnostic performance of a test can be judged by measuring the area under the curve. Poor tests have ROC curves close to the diagonal, while the ROC curves for excellent tests rise steeply and pass close to the top left-hand corner, where both the Se and Sp are 1.

2.2.4. Positive and negative likelihood ratio

The ratio between the probability to have a particular test result in patients with disease and in those without disease defines the Likelihood Ratio (LR) [8]. For binary tests, where the results may take on only two values, plus or minus, the calculation formulas are:

$$LR(+) = \frac{\text{Probability of positive result in patient with disease}}{\text{Probability of positive result in patient without disease}} = \frac{\text{Sensitivity}}{1 - \text{Specificity}}$$

$$LR(-) = \frac{\text{Probability of negative result in patient with disease}}{\text{Probability of negative result in patient without disease}} = \frac{1 - \text{Sensitivity}}{\text{Specificity}}$$

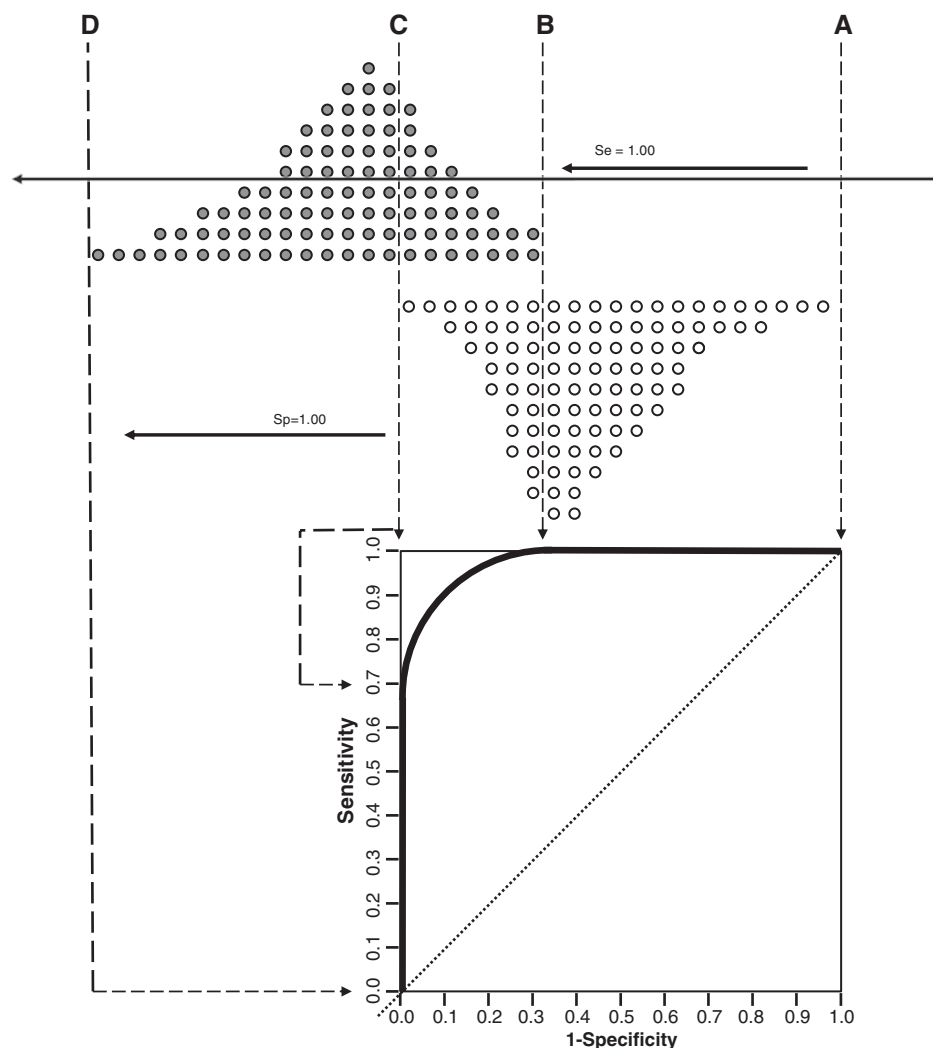


Fig. 3. Threshold changes the sensitivity and the specificity of a test. Relationship between sensitivity and the specificity can be observed in the Receiver Operating Characteristic curve.

A LR equal to 1 represents a completely not discriminating test, with equal probabilities in the diseased and non-diseased populations; the farther a likelihood ratio deviates from a value of 1, the more informative is the test result [9].

High likelihood ratios (e.g., > 10) indicate that the test can be used to rule-in the disease, while low likelihood ratios (e.g., < 0.1) can rule-out the disease. If a test can be used to rule-in the disease, it means that the post-test probability of the disease is very high. The Bayes' theorem can be applied to derive probabilistic statements regarding the likelihood of disease in an individual. Beyond the formulas deriving from Bayes' theorem, a simpler way to calculate post-test probability starting from pre-test probability is the Fagan nomogram [10]. As an example (Fig. 4), we may suppose that the prevalence of a positive ANA test in systemic lupus erythematosus (SLE) patients is 3%. Considering the use of an ANA test with a Se of 0.98 and a Sp of 0.93, the LR(+) would be 14, and the LR(–) 0.02.

As showed by the Fagan's nomogram, this ANA test is very accurate in the rule-out, with post-test probability for a negative result less than 0.1%, conversely, a positive result improves the probability of disease from 3 to 30%, meaning that 70% of the screened patients with a positive ANA test do not have SLE.

2.2.5. Diagnostic odds ratio

The diagnostic odds ratio (DOR) expresses the odds of positive test results in patients with disease compared to the odds of positive test

results in those without disease [11]. The formula to calculate the DOR is:

$$DOR = \frac{LR(+)}{LR(-)}$$

The DOR summarizes in a single value the accuracy of a test, while the other accuracy indexes are paired indicators. One important advantage of DOR is that it is not dependent on the prevalence of the target disease, and may be easier to understand, even if it is less common and intuitive than Se and Sp. It can be also used as the odds ratio in RCT studies, especially in meta-analysis, overcoming also the problem of different thresholds in different studies.

The value of a DOR ranges from 0 to infinity, with higher values indicating better discriminatory test performance. A value of 1 means that a test does not discriminate between patients with the disorder and those without. A value less than 1 deposes for a misleading test (more positive results in healthy people or more negative in diseased).

3. Confidence intervals

Another tool we have to know to be able to critically appraise clinical and diagnostic studies is confidence intervals (CI) [12]. CI is a range of values for a variable of interest (for instance: Se, Sp, LR, and DOR)

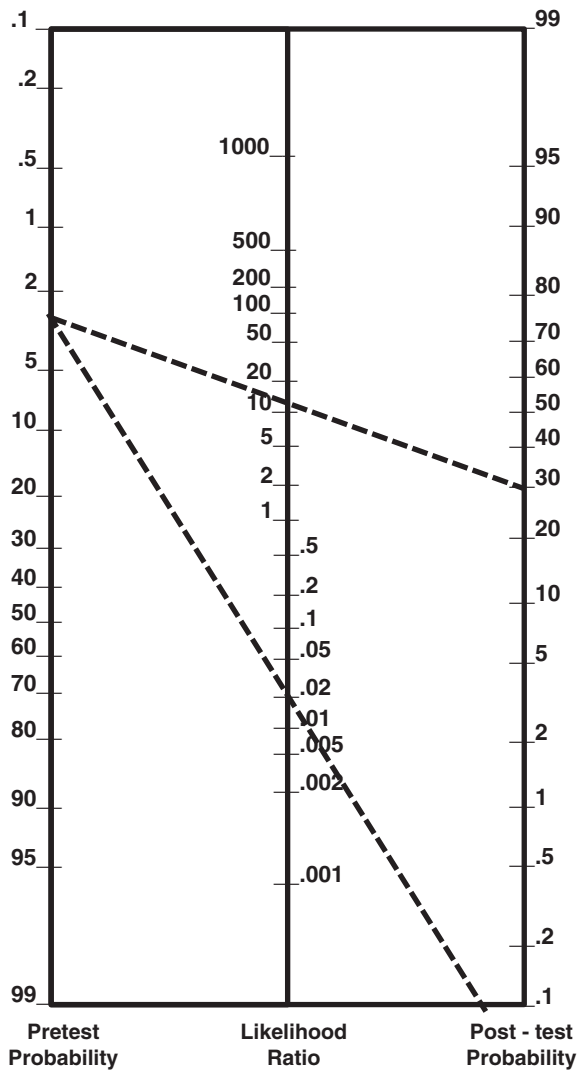


Fig. 4. Fagan's nomogram. Dotted lines represent the estimation of post-test probability for the ANA test ($Se = 0.98$ and $Sp = 0.93$), starting from a pre-test probability of 3%.

constructed so that this range has a specified probability (usually 95%) of including the true value of the variable.

In statistical inference, one wishes to estimate population parameters using observed sample data. This holds true also for nearly all the accuracy test evaluations. Even excluding the problem that a sample may be not representative of the population (i.e., for selection or

other bias), there is still a possibility of random variation in the sample. Results obtained from only a sample of the entire population are subjected to statistical uncertainty, which is related to the size of the sample. The confidence intervals indicate the imprecision of the measure. The limits of confidence define a range of values which contain the true value, and the width of the CI is inversely related to the size of the sample. Accuracy indexes obtained from larger samples will have a narrower confidence interval and the evidence will be stronger.

4. Systematic review and meta-analysis

Medical professionals feel confused about the huge available medical literature, both for difficulties in finding and reading all those information and for the lack of a critical comparative appraisal. Reviews, as summary of the available literature on an argument, have become an important tool to keep up-to date and seek for evidence. However, narrative reviews are often not well done, since they may include a number of biases (publication, language, indexing inclusion) [13] and may be influenced by the opinion of the review writer. A well-conducted systematic review may overcome many of these problems. Provided by the foundational work of Archibald Cochrane in the 1980s, a formal method to conduct systematic reviews has been proposed. As stated in the Cochrane Library Glossary [14], systematic review is “a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyze data from the studies that are included in the review. Statistical methods (meta-analysis) may or may not be used to analyze and summarize the results of the included studies.” A systematic review must have a clearly focused question, a comprehensive search to find all original reports on the topic of interest, a critical evaluation and appraisal of the reports, and conclusions based on a synthesis of studies that meet preset quality criteria. Hence, knowledge about the statistical tools used in the systematic reviews is helpful to understand and critically appraise even this type of publication.

4.1. Meta-analysis

Meta-analysis is undertaken to summarize the results of multiple studies addressing a single research issue. The first stage in any analysis should be to plot the individual study results. Like clinical studies, accuracy studies can be pooled in a Forest plot graph [15]. It is possible to plot, in different Forest plots, Se , Sp , $LR(+)$, $LR(-)$ and DOR . The statistic from each individual study is represented by a dot, cut by a horizontal line, wide as the associated confidence interval.

In Fig. 5, a meta-analysis on 10 studies is represented for Se and Sp statistics. The area of every square represents the dimension of the study; the horizontal lines are the CIs, and their lengths clearly show the inverse relationship between the size of the sample and its CIs.

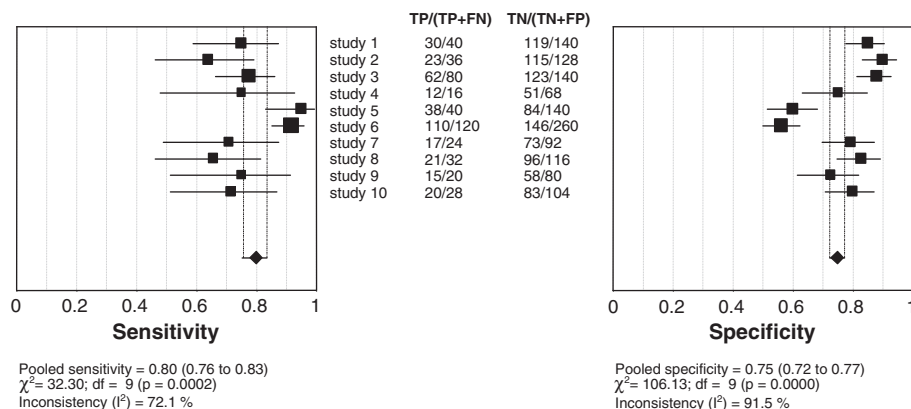


Fig. 5. Forest plot of 10 hypothetical studies, for sensitivity and specificity.

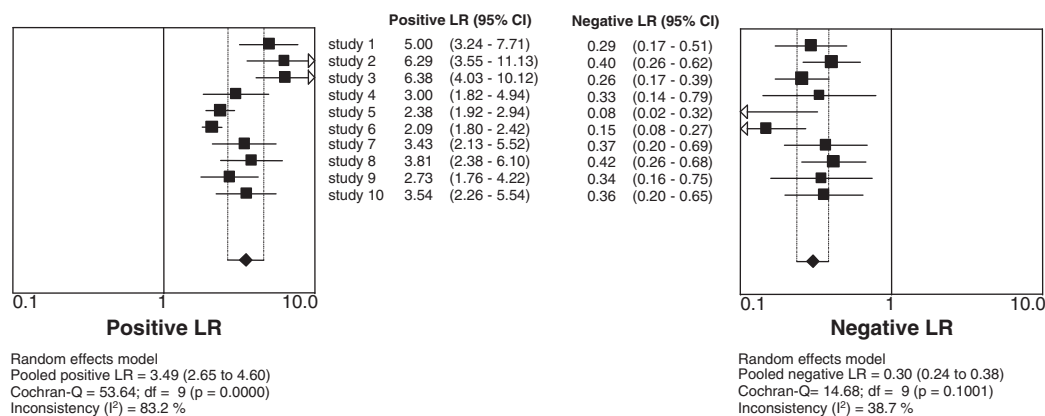


Fig. 6. Forest plot of the same 10 hypothetical studies, for positive and negative likelihood ratio.

The diamond in the bottom represents the pooled Se and Sp, with their CIs. Se, Sp and LR (Fig. 6) should only be pooled in the absence of variability of the diagnostic threshold. The pooled statistics correspond to weighted averages in which the weight of each study is represented by its sample size. Likelihood ratios and diagnostic odds ratios (Fig. 7) can be pooled by different mathematical methods, to incorporate variation among studies [16,17]. Simplifying, “fixed effect” is chosen as the calculation method when there is no or low heterogeneity among the studies; “random effect” should be chosen instead when heterogeneity exists. In addition, this latter type of computation gives more importance to smaller trials [17].

Homogeneity among study results should first be evaluated graphically by plotting Se and Sp values from each study on a Forest plot. The consistency of each statistic can be gauged from the overlap of the individual CIs [18].

There are two numerical options to measure inconsistency: the test of homogeneity (Cochran Q test) and the test of inconsistency.

The first one yields a χ^2 statistic computed as the weighted sum of the squared differences between the overall summary estimate and the results of individual studies, with a p value for the null hypothesis (the lower is p under 0.05, the lower is homogeneity). The second method (inconsistency test, I^2), expresses heterogeneity as a percentage, with higher values indicating more between-study heterogeneity. I^2 is preferred when considering only a few small studies, since Cochran Q test is quite insensitive for detecting heterogeneity in these situations [19–21].

Heterogeneity is to be expected in a meta-analysis: it would be surprising if multiple studies, performed by different teams in different places with different methods, all ended up estimating the same underlying parameter [22]. Although some variation can be expected by chance, other factors, such as variation in study designs, heterogeneity of analytical methods used for diagnostic testing, and patient selection, may increase the observed heterogeneity [23]. An important extra source of heterogeneity is the variation introduced by changes in diagnostic threshold, which can be a casual or deliberate choice of the investigators. Looking at a Forest plot we can recognize a possible threshold effect by the observation of the opposite distribution of Se and Sp values. In Fig. 5, for instance, we see that studies 5 and 6 are quite different from the others, as they show a higher Se and a corresponding lower Sp. This is a typical effect obtained by lowering the cutoff threshold.

4.2. Summary Receiver Operative Characteristic curve

Plotting the TP rate (Se) against the FP rate (1-Sp), the threshold effect can be emphasized (Fig. 8). In this kind of graphic, Se is plotted as function of 1-Sp for each study, like a usual ROC plot. Then a summary ROC (sROC) curve can be fitted. The simplest approach has been

developed by Moses et al. [24], but other approaches have been proposed [25,26]. Computations are quite difficult and beyond the scope of this review. However, sROC curves are proposed in many systematic reviews of diagnostic accuracy, so practitioners have to understand the primary significance, for a better appraisal. Firstly, we can observe if the fitted curve is symmetric or asymmetric. Diagnostic thresholds should be similar for each study, so threshold effect does not influence the shape of the curve. Moreover, if the statistics for the several studies under meta-analysis are distributed along and close to the sROC curve defined as $Se = f(1-Sp)$, this is evidence for different cut-offs in tests with similar diagnostic accuracy. In Fig. 8 this is the case of the studies no. 5 and no. 6.

The area under the curve (AUC) reflects overall performance of one given test. The perfect test will have an AUC of one. The sROC curves and their AUCs are useful to compare the accuracy of multiple/numerous diagnostic tests [27].

The Q^* point is the intercept on the sROC curve by the antidiagonal line, across the unit square (the asterisk is used to differentiate the Q^* from the Q Cochran test). (It is very hard to understand why statisticians love to complicate the life of clinicians with misleading symbols like this one.) Its value indicates overall accuracy by giving the point where Se and Sp are the same. The closer the curve to the top left corner, the better the accuracy. In Fig. 8, the pooled Se is 0.80, the pooled Sp is 0.75 (see data from Fig. 4), while Q^* is 0.79.

5. Conclusions

The five steps (the five “A’s”) of the process of evidence-based practice (EBP) has recently arrived in the arena of diagnostic medicine [28].

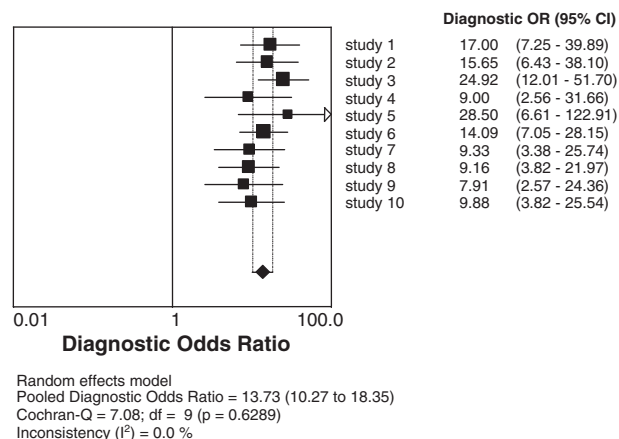


Fig. 7. Forest plot of the same 10 hypothetical studies, for diagnostic odds ratio.

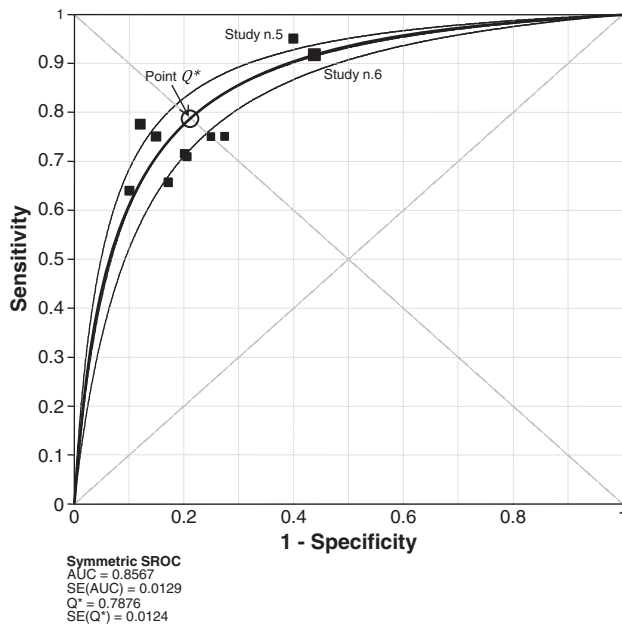


Fig. 8. Summary ROC curve of the same 10 hypothetical studies (thick line). Thin lines mark confidence limits.

Not all clinicians have the same levels of responsibility within evidence-based organizations and require the same skill for EBP [29]. But even if not all practitioners need to appraise evidence from scratch, all need some skills [30]. Considering the real-time setting of practice, searching and appraisal need to be done by many clinicians in minutes rather than hours or days. Secondary publications, especially systematic review and meta-analysis, could be powerful instruments to keep up-to date our knowledge and practice [31–39]. Understanding the summarized results presented in this

type of publications can be very useful for clinical and diagnostic decisions (Table 2).

Se and Sp are the first and most important quality indexes of accuracy evaluation, from which all the other indexes can be derived. They represent, respectively, the ability of a given test in recognizing people with the disease and healthy people.

PV+ and PV− are poor indexes, as they are dependent on the pre-test probability of disease. Usually, the experimental setting is often different and frequently better than in the real-life practice: the study population is more homogeneous; complicated patients are often excluded by the study; processes are standardized; personnel is well trained and motivated; and especially, the number of positive patients is much higher than in routine setting.

LR(+) and LR(−) are good probabilistic indexes. Respectively, they are an estimate of the probability of disease if the test is positive or negative. Knowing the pre-test probability of disease in a patient, LR allows for estimating the post-test probability.

DOR is the ratio of LR(+) and LR(−). It is used to describe the diagnostic characteristics of the test as a single term.

In meta-analyses, similar information from different studies is combined to give a more accurate estimate of the accuracy of a test. Pooled Se, Sp, LR(+), LR(−) and DOR are usually calculated. Forest plots for these statistics, especially for the DOR, permit an appraisal of homogeneity among the considered studies by comparing the overlap between the confidence intervals line. Statistical indexes such as the Cochran Q test and the Inconsistency index can effectively display this information. If inconsistency exists, pooled statistic results have to be evaluated with caution. A threshold bias can affect the homogeneity between different studies. A sROC curve plot summarizing true positive rate (Se) against false positive rate (1-Sp), as represented by the distribution of the scattering, may indicate whether heterogeneity is related or not to the thresholds used in the different studies.

All practitioners involved in the clinical interpretation of diagnostic test data should be more confident with the principles of EBM and

Table 2

Accuracy indexes and remarks.

Index	Acronym	Formula	Meaning	Remarks
Sensitivity	Se	$\frac{TP}{TP+FN}$	Quality to recognize the “positive”, the illness	SPIN: high specificity and positive results rule IN the diagnosis
Specificity	Sp	$\frac{TN}{TN+FP}$	The quality to recognize the “negative”, the healthy people	SNOUT: high sensitivity and negative results rule OUT the diagnosis
Positive predictive values	PV+	$\frac{TP}{TP+FP}$	Probability that a patient has a target disease given that the test result is positive	Not a good index, as it depends on pre-test probability (prevalence)
Negative predictive value	PV−	$\frac{TN}{TN+FN}$	Probability that a patient does not have the target disease given that the test result is negative	Not a good index, as it depends on pre-test probability (prevalence)
Receiver Operator Characteristic curve	ROC curve	Plot: $Se = f(1 - Sp)$	Depicts the true positive rate (Se) as function of false positive rate (1-Sp), evaluated at several diagnostic thresholds	Measure of the general ability of the test to discriminate those with disease from those without
Area Under Curve ROC	AUC	$\int_0^1 Se = f(1 - Sp)$	Measure of the general ability of the test to discriminate those with disease from those without	The higher the AUC, the better is the test's performance
Positive Likelihood Ratio	LR(+)	$LR(+) = \frac{Se}{1-Sp}$	Ratio between the probability to have a positive test results in patients with disease and in those without disease	High likelihood ratios (e.g., $LR > 10$) indicate that the test can be used to rule in the disease
Negative Likelihood Ratio	LR(−)	$LR(-) = \frac{1-Se}{Sp}$	ratio between the probability to have a negative test results in patients with disease and in those without disease	low likelihood ratios (e.g., $LR < 0.1$) can rule out the disease
Diagnostic Odds Ratio	DOR	$DOR = \frac{LR(+)}{LR(-)}$	Odds of positive test results in patients with disease compared to the odds of positive test results in those without disease	Accuracy in a single value. It ranges from 0 to infinity. 1 is not discriminating value. Higher values indicate better discriminatory test performance

meta-analyses. Some simple tools and knowledge presented in this paper may help clinicians to better interpret the results of a study of accuracy and of meta-analytic studies and to understand the strengths and limitations of such studies.

Take-home messages

- Using the five-step process of evidence-based practice (Ask, Acquire, Appraise, Apply, Assess), all practitioners should be more confident with evaluating the accuracy of diagnostic tests.
- Sensitivity, specificity, and likelihood ratios are the most informative statistics, while predictive values may be misleading, being largely dependent on the pre-test probability (prevalence of the disease).
- Like odds ratio for randomized controlled trial, diagnostic odds ratio is used in diagnostic accuracy studies to describe the diagnostic characteristics of the test as a single term.
- Confidence intervals are the best way to indicate the relative imprecision of the statistics: they provide a range of plausible values for the unknown parameter (Se, Sp, LR, DOR, etc.).
- Forest plot and sROC curve are the graphs of meta-analysis; in appraising a Forest plot, pay attention to homogeneity; if inconsistency is suspected, examination of the sROC curve can help to uncover a threshold bias.

References

- [1] Evidence-Based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA* 1992;268:2420–5.
- [2] Guyatt G. Evidence-based medicine. *ACP J Club* 1991;114:A-16.
- [3] Cochrane AL. Effectiveness and efficiency: random reflections on health services. London: Nuffield Provincial Hospitals Trust; 1973.
- [4] Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Rep* 1947;62:1432–49.
- [5] Vecchio TJ. Predictive value of a single diagnostic test in unselected populations. *N Engl J Med* 1966;274:1171–3.
- [6] Galen RS, Gambino SR. Beyond normality: the predictive value and efficiency of medical diagnoses. New York: John Wiley; 1975.
- [7] Henderson AR, Bhayana V. A modest proposal for the consistent presentation of ROC plots in clinical chemistry. *Clin Chem* 1995;41:1205–6.
- [8] Albert A. On the use and computation of likelihood ratios in clinical chemistry. *Clin Chem* 1982;28:1113–9.
- [9] Jaeschke R, Guyatt GH, Sackett DL, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature. VI. How to use an article about a diagnostic test. B: What are the results and will they help me in caring for my patients? *JAMA* 1994;271:703–7.
- [10] Fagan TJ. Nomogram for Bayes theorem. *N Engl J Med* 1975;293:257.
- [11] Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;56:1129–35.
- [12] Altman DG, Machin D, Bryant TN, Gardner MJ, editors. *Statistics with confidence*. 2nd ed. Bristol: BMJ Books; 2000.
- [13] Egger M, Smith GD. Meta-analysis bias in location and selection of studies. *BMJ* 1998;316:61–6.
- [14] <http://www.cochrane.org/resources/glossary.htm> [Last access January 17, 2012].
- [15] Lewis S, Clarke M. Forest plots: trying to see the wood and the trees. *BMJ* 2001;322:1479–80.
- [16] Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of diseases. *J Natl Cancer Inst* 1959;22:719–48.
- [17] DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177–88.
- [18] Pereira R, Heneghan C. Systematic review and meta-analyses. In: Price CP, Christenson RH, editors. *Evidence-based laboratory medicine – principle, practice, and outcomes*. 2nd ed. Washington, DC: AACC Press; 2007. p. 245–74.
- [19] Normand ST. Tutorial in biostatistics. Meta-analysis: formulating, evaluating, combining, and reporting. *Stat Med* 1999;18:321–59.
- [20] Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M, Davey Smith G, Altman DG, editors. *Systematic reviews in health care: meta-analysis in context*. 2nd ed. London: BMJ Books; 2001. p. 285–312.
- [21] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557–60.
- [22] Higgins JPY. Commentary: heterogeneity in meta-analysis should be expected and appropriately quantified. *Int J Epidemiol* 2008;37:1158–60.
- [23] Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994;309:1351–5.
- [24] Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 1993;12:1293–316.
- [25] Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007;8:239–51 [Erratum in: *Biostatistics* 2008].
- [26] Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med* 2002;21:1237–56.
- [27] Boyd JC. Statistical analysis and presentation of data. In: Price CP, Christenson RH, editors. *Evidence-based laboratory medicine – principle, practice, and outcomes*. 2nd ed. Washington, DC: AACC Press; 2007. p. 113–40.
- [28] Price CP, Glenn JL, Christenson RH. *Applying evidence-based laboratory medicine. A step-by-step guide*. Washington, DC: AACC Press; 2009.
- [29] Dawes M, Summerskill W, Glasziou P, Cartabellotta A, Martin J, Hopayian K, et al. Second international conference of evidence-based health care teachers and developers. Statement on evidence-based practice. *BMC Med Educ* 2005;5:1.
- [30] Guyatt GH, Meade MO, Jaeschke RZ, Cook DJ, Haynes RB. Practitioners of evidence based care. Not all clinicians need to appraise evidence from scratch but all need some skills. *BMJ* 2000;320:954.
- [31] Heijstek MW, Ott de Bruin LM, Borrow R, van der Klis F, Koné-Paut I, Fasth A, et al. Vaccination in paediatric patients with auto-immune rheumatic diseases: a systemic literature review for the European League against Rheumatism evidence-based recommendations. *Autoimmun Rev* 2011;11:112–22.
- [32] Zettl UK, Stüve O, Patejdl R. Immune-mediated CNS diseases: a review on nosological classification and clinical features. *Autoimmun Rev* 2012;11:167–73.
- [33] Petříková J, Lazúrová I. Ovarian failure and polycystic ovary syndrome. *Autoimmun Rev* 2012;11:A471–8.
- [34] Rutgers A, Kallenberg CG. Refractory vasculitis. *Autoimmun Rev* 2011;10:702–6.
- [35] Polido-Pereira J, Vieira-Sousa E, Fonseca JE. Rheumatoid arthritis: what is refractory disease and how to manage it? *Autoimmun Rev* 2011;10:707–13.
- [36] Scoble T, Wijetilleka S, Khamashta MA. Management of refractory anti-phospholipid syndrome. *Autoimmun Rev* 2011;10:669–73.
- [37] Pietrogrande M, De Vita S, Zignego AL, Pioltelli P, Sansonno D, Sollima S, et al. Recommendations for the management of mixed cryoglobulinemia syndrome in hepatitis C virus-infected patients. *Autoimmun Rev* 2011;10:444–54.
- [38] van Assen S, Elkayam O, Agmon-Levin N, Cervera R, Doran MF, Dougados M, et al. Vaccination in adult patients with auto-immune inflammatory rheumatic diseases: a systematic literature review for the European League Against Rheumatism evidence-based recommendations for vaccination in adult patients with auto-immune inflammatory rheumatic diseases. *Autoimmun Rev* 2011;10:341–52.
- [39] Selmi C, Tsuneyama K. Nutrition, geoepidemiology, and autoimmunity. *Autoimmun Rev* 2010;9:A267–70.