Appendix: Test Performance Metrics

The basic definitions of test performance are based upon a 2×2 cross tabulation of the “true disease status” and the test results (Appendix Table). Ostensibly simple, the information in a 2×2 table can be summarized in several ways—some of which are mathematically equivalent (e.g., sensitivity/specificity vs. positive/negative likelihood ratios)—and these measures and their application can be confusing in practice.[1] The basic measures are described briefly below; for a discussion on their relative merits and drawbacks, see Tatsioni et al.[2]

Appendix Table. 2×2 table used in the calculation of test performance measures

		True Disease Status	
		Disease*	**No Disease***
(Index) Medical Test	Suggestive of disease (“positive”)	“TP” (=”true positives”)	“FP” (=”false positives”)
	Not suggestive of disease (“negative”)	“FN” (=“false negatives”)	“TN” (= “true negatives”)

“TP” = true positive; “FN” = false negative; “FP” = false positive; “TN” = true negative. The quotation marks are retained to stress that, in the calculation of the basic measures reviewed here, we assume that the reference standard test has negligible misclassification rates for practical purposes.

- Sensitivity: “TP”/(“TP” + “FN”)
- Specificity: “TN”/(“FP” + “TN”)
- Positive likelihood ratio (LR+): Sensitivity/(1–Specificity)
- Negative likelihood ratio (LR-): (1–Sensitivity)/Specificity
- Diagnostic odds ratio: (“TP” * “TN”) / (“FP” * “FN”)
- Positive predictive value: “TP”/(“TP” + “FP”) = (Sensitivity * Prevalence)/(Sensitivity * Prevalence + (1-Prevalence)*(1-Specificity))
- Negative predictive value: “TN”/(“TN” + “FN”) = (Specificity * (1-Prevalence))/(Specificity *(1- Prevalence) + Prevalence*(1-Sensitivity))

*This is typically ascertained by a reference test. The reference test is assumed to have negligible misclassification of the true disease status.

Sensitivity and Specificity

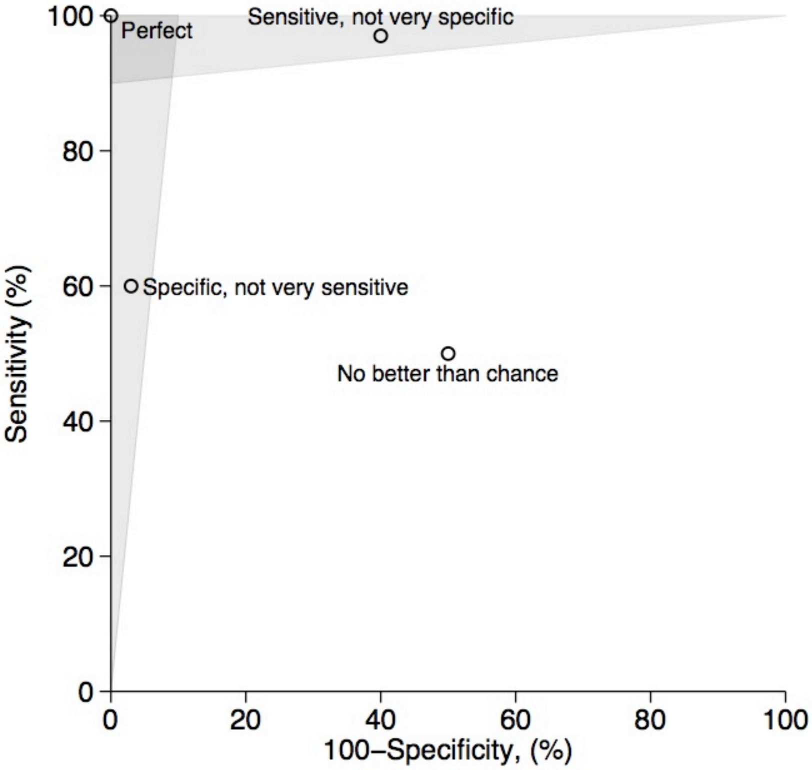
Sensitivity, also known as the true positive rate, is the probability of testing positive for diseased patients. It expresses the ability of a medical test to maximize true positives. Specificity, or true negative rate, is the probability of testing negative for non-diseased patients. It expresses the ability of a test to minimize false positives.

The two measures have clear clinical interpretations, but are not as useful clinically as the positive and negative predictive values (see below). Sensitivity and specificity are negatively correlated with each other with respect to diagnostic thresholds. If the threshold (cutpoint) for test positive is set higher—say when the test provides a continuously valued result—sensitivity will decrease while specificity will increase. On the other hand, if the threshold is lower, we will see an increase in sensitivity and a corresponding decrease in specificity. This non-independence of sensitivity and specificity across explicit or implicit diagnostic thresholds poses challenges for quantitative synthesis.

When several thresholds have been considered for a single set of data, a receiver operating characteristic (ROC) curve could be obtained by plotting sensitivity vs. 1–specificity. (Appendix Figure). The ROC curve depicts the observed patterns of sensitivity and specificity at different thresholds, as well as the negative correlation between the two measures. As we will discuss below, one way to summarize diagnostic accuracy data is to calculate a summary ROC curve.

Note that the closer a study point is to the upper left corner of the plot, the better its diagnostic ability.

Appendix Figure. Typical plot of sensitivity versus 100 percent specificity



Four hypothetical studies are depicted in the square sensitivity/100 percent–specificity plot. The closer a study is to the upper-left corner of the plot, the better its diagnostic ability. Studies lying on the major diagonal of the plot have no diagnostic ability (no better than chance). Studies lying on the left shaded area have positive likelihood ratio (LR+) of 10 or more. Studies lying on the top shaded area have negative likelihood ratio (LR-) of 0.1 or less. Studies lying on the intersection of the grey areas (darker grey polygon) have both LR+>10 and LR-<0.1. Screening tests typically operate in the less shaded areas, whereas confirmatory tests used to rule out a diagnosis often operate near or in the top shaded area. The systematic reviewer must be familiar with the mentioned measures and their interpretation: the same medical test can be used in different settings and roles, and different measures will best capture its performance each time.

Positive and Negative Likelihood Ratios

The *positive* and *negative likelihood ratios* (LR+ and LR-, respectively) quantify the change in the certainty of the “diagnosis” conferred by test results. More specifically, the likelihood ratios transform the *pretest odds* to the *posttest odds* of a given (positive or negative) diagnosis:

$$posttest\ odds = pretest\ odds \times LR$$

For a positive result with the medical test, the positive likelihood ratio would be used in the above relationship; for a negative result with the medical test portable monitor, the negative likelihood ratio would be used.

If a given medical test has very good ability to predict the “true disease status,” its positive likelihood ratio will be high (i.e., will greatly increase the odds of a positive diagnosis) and its negative likelihood ratio will be low (i.e., will diminish substantially the likelihood of the positive diagnosis). A completely non-informative portable monitor would have likelihood ratios equal to 1 (i.e., does not transform the pre-test odds substantially in the equation above). Typically, a positive likelihood ratio of 10 or more and a negative likelihood ratio of 0.1 or less are considered to represent informative tests.[3] We note that other, more lenient boundaries for LR+ and LR- can be used[3] and that the choice of the boundaries is a subjective decision. It is interesting to note that studies with high LR+ and low LR- can be readily identified in the square sensitivity/100 percent-specificity plot, as shown in the Appendix Figure above.

Diagnostic Odds Ratio

The diagnostic odds ratio (DOR) describes the odds of a positive test in those with disease relative to the odds of a positive test in those without disease.[4] It can be computed in terms of sensitivity and specificity as well as in terms of positive and negative likelihood ratios (DOR = LR+/LR-). Thus this single measure includes information about both sensitivity and specificity and tends to be reasonably constant despite diagnostic threshold. However, it is impossible to use diagnostic odds ratios to weigh sensitivity and specificity separately, and to distinguish between tests with high sensitivity and low specificity and tests with low sensitivity and high specificity.

Another disadvantage is that it is difficult for clinicians to understand and apply, limiting its clinical value. This is partly because they are not often exposed to diagnostic odds ratios. A diagnostic odds ratio is similar to an odds ratio that measures strength of association in an observational study or effect size in a trial. However, contrary to the typical effect size magnitudes of such odds ratios (often between 0.5 and 2), diagnostic odds ratios can attain much larger values (often greater than 100).

Positive and Negative Predictive Values

Positive predictive value is the probability of disease given a positive test and negative predictive value is the probability of no disease following a negative test. These values are highly useful for clinical purposes because they give the clinician an indication of the likelihood of disease or a specific event such as death following the results of the medical test.

Positive and negative predictive values depend upon disease prevalence, which is unlikely to be consistent among studies. Therefore, they are often calculated for a range of plausible prevalence values and tabulated or plotted in graphs.

References

1. Loong TW. Understanding sensitivity and specificity with the right side of the brain. BMJ 2003; 327(7417):716-9.

2. Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C, et al. Challenges in systematic reviews of diagnostic technologies. Ann Intern Med 2005; 142(12 Pt 2):1048-55.

3. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994; 271(9):703-7.

4. 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(11):1129-35.