EDITORIAL

Meta-analysis for diagnostic tests

Anastasia M. Hartzes, MPH,[a] and Charity J. Morgan, PhD[a]

[a] Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL

See related article, pp. 59–67

Meta-analysis is a statistical method for pooling and analyzing the results from multiple studies.[1] A typical meta-analysis will focus on a single outcome measure, such as a treatment effect or rate of an adverse event. However, applying meta-analysis to studies of diagnostic test's accuracy is not straightforward. The accuracy of a diagnostic test is most often summarized with *two* outcomes, sensitivity and specificity, which cannot be expected to be independent and therefore must be analyzed together.[2] Thus, meta-analytic methods specific to diagnostic tests are needed. The paper by Lee et al., which reports the results of a meta-analysis of F-18 FDG PET for detection of disease activity, published in this issue demonstrates the use of some of these methods.[3]

We consider here the setting of a diagnostic test that yields a qualitative result (for example, a test designed to indicate the presence or absence of a disease). Let a positive test result indicate a patient probably has a disease, and let a negative test indicate a patient probably does not have disease. The *sensitivity* of a test describes the probability that the test predicts the disease, given the presence of the disease in the patient. Alternately, a sensitive test is one which detects the disease, when the patient has the disease. The *specificity* of a test describes the probability that the test does not detect the disease, given the patient does not have the disease. Alternately, a specific test is one which produces positive results for a small number of patients, when they do not actually have the disease. An ideal test would have both high sensitivity and high specificity.[4,5]

In addition to sensitivity and specificity, there are several measures that are used to describe the accuracy of a diagnostic test. The formulas for several of these measures are summarized in Table 1. Calculation of these measures requires consideration of a patient's actual disease status relative to the test result. A *true positive* is the situation when a patient has the disease, and the test is positive for it; a *true negative* is when a patient does not have the disease and the test result is negative. A *false positive* is when the patient does not have the disease, but has a positive test result; a *false negative* is when a patient has the disease, but has a negative test result.[4,5] A useful way to compare patients for a particular test result is via likelihood ratios (LR), which are ratios of the sensitivity and specificity. LRs describe how likely a diseased patient is to have a (positive or negative) result, compared to disease-free patients. Values greater than 1.0 provide evidence that the positive/negative test result is related disease presence; values less than 1.0 demonstrate the rest result is related to disease absence. A positive LR (LR+) represents the ratio of patients with the disease who test positive, to those who test positive but are disease free. A negative LR (LR-) represents the ratio of patients with the disease who test negative, compared to those who test negative and are disease free.[6] LR values ranging from 0.1 to 10 are considered substantial evidence to support a diagnosis as positive or negative, respectively.[6] The diagnostic odds ratio (DOR) is a ratio of the LR+ to the LR−. It is the odds that the test produces positive results compared to the odds of negative results. Values are all greater than 0; larger values indicate a better-performing screening test.[7]

We demonstrate these calculations with a simple example. Figure 1 displays the results of six hypothetical studies of a diagnostic test's accuracy. For one of these studies (Study 1), we summarize the results of disease status vs test results in Table 2. In this study, 350 subjects were assessed. Two hundred and fourteen subjects had the disease, with 94 testing positive (i.e., true positive) and 120 testing negative (false negative). Therefore, the estimate of the sensitivity of the test is

Table 1. Descriptive statistics for diagnostic tests

Test result	True disease status	
	Disease present	**Disease absent**
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

True positive rate (TPR) = Sensitivity = $\frac{TP}{TP+FN}$
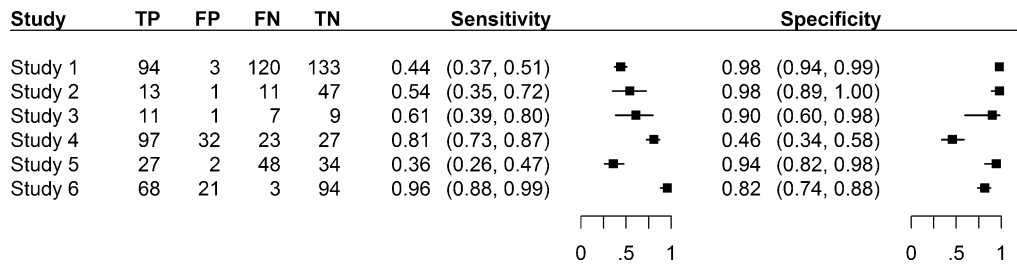
False negative rate (FNR) = $1 - $ TPR = $\frac{FN}{TP+FN}$

False positive rate (FPR) = $\frac{FP}{FP+TN}$

True negative rate (TNR) = $1 - $ FPR = Specificity = $\frac{TN}{FP+TN}$

Positive likelihood ratio (LR+) = $\frac{\text{Sensitivity}}{1-\text{Specificity}} = \frac{TPR}{FPR}$

Negative likelihood ratio (LR−) = $\frac{1-\text{Sensitivity}}{\text{Specificity}} = \frac{FNR}{TNR}$

Diagnostic odds ratio = $\frac{LR+}{LR-}$

| Study	TP	FP	FN	TN	Sensitivity	Specificity
Study 1	94	3	120	133	0.44 (0.37, 0.51)	0.98 (0.94, 0.99)
Study 2	13	1	11	47	0.54 (0.35, 0.72)	0.98 (0.89, 1.00)
Study 3	11	1	7	9	0.61 (0.39, 0.80)	0.90 (0.60, 0.98)
Study 4	97	32	23	27	0.81 (0.73, 0.87)	0.46 (0.34, 0.58)
Study 5	27	2	48	34	0.36 (0.26, 0.47)	0.94 (0.82, 0.98)
Study 6	68	21	3	94	0.96 (0.88, 0.99)	0.82 (0.74, 0.88)

Figure 1. Forest plots of sensitivity and specificity for six hypothetical studies.

Table 2. Descriptive statistics for diagnostic tests

Test result	True disease status		Total
	Disease present	**Disease absent**	
Positive	94	3	97
Negative	120	133	253
Total	214	136	350

0.44. Out of the remaining 136 subjects who did not have the disease, 3 tested positive (false positive) and 133 tested negative (true negative). Thus, the specificity of the test is estimated to be 0.98.

Once summary statistics for each of the studies to be included in the meta-analysis have been calculated, a forest plot can be used to display the estimates for each study (see Figure 1 for an example). A bivariate
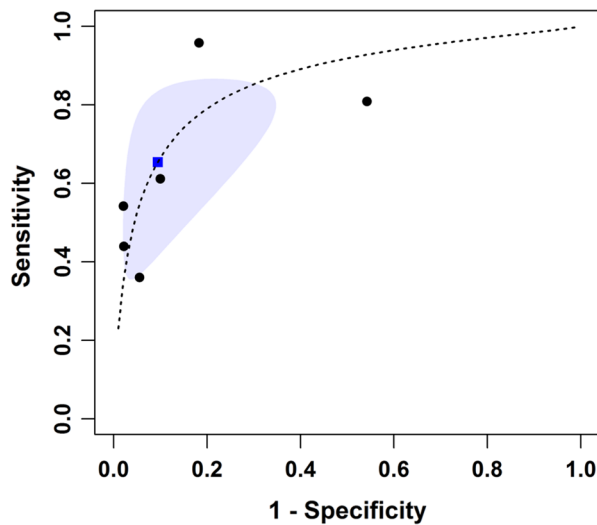
Figure 2. Summary ROC curve.

random-effects model (such as the one employed by Lee et al.[3]) can then be used to produce a summary point estimate of sensitivity and specificity.[8] Note that a bivariate model is necessary in order to take into account the likely correlation between sensitivity and specificity across studies.[9] Furthermore, while a common approach for meta-analysis of non-diagnostic studies is to consider both fixed-effect and random-effect models,[1] diagnostic studies should be expected to be heterogeneous, making a fixed-effect model inappropriate.[10]

A notable source of heterogeneity for diagnostic studies is the result of what is known as the ''threshold effect''.[9] Many diagnostic tests compare a result or measurement to a pre-specified threshold, or reference standard. The choice of threshold will affect both the sensitivity and specificity of the test. For example, consider a simple diagnostic test that measures the amount of an antibody in a blood sample, with high levels of the antibody resulting in a positive test result. If the threshold for a positive result is lowered (meaning a smaller amount of the antibody is required to be present in the sample in order to diagnose illness), we would expect the test to return more false positives (and therefore have lower specificity) and fewer false negatives (greater sensitivity). If the threshold were raised, we would observer fewer false positives (greater specificity) and more false negatives (lower sensitivity). A receiver operating characteristic (ROC) curve plots the true positive rate (or sensitivity) against the false positive rate (1-specificity) for a diagnostic test under varying thresholds. The area under the ROC curve provides an overall summary of diagnostic test's accuracy, independent of the threshold effect.[4,5]

For meta-analysis of diagnostic tests, the studies being pooled may use varying thresholds, leading to heterogeneity in the estimates of sensitivity and specificity. In order to examine such heterogeneity, the estimated sensitivities and specificities can be plotted against each other, a summary ROC (sROC) curve can be estimated and the area under the sROC curve calculated.[11] Figure 2 provides an illustration of this approach. The sensitivities and specificities for the six hypothetical studies are represented as black circles, and the summary estimate of sensitivity and specificity is denoted with a blue square; the shaded oval represents a 95% confidence region for the summary estimate. The dotted line indicates the estimated sROC curve. For this particular example, the pooled estimate of sensitivity is 0.65, and the pooled estimate of specificity is 0.91. The area under the sROC curve is 0.86.

Screening or diagnostic tests are useful when needing to determine the presence or potential development of a disease in question; they are particularly valuable when the confirmatory procedure is invasive, cost-prohibitive, time-intensive, or only available upon autopsy.[5] When synthesizing evidence from different studies of a diagnostic test's accuracy is of interest, meta-analysis may be used. However, meta-analytic methods specific to diagnostic tests must be used in order to properly summarize the study results.

Disclosure

Authors have no conflicts of interest to disclose.

References

1. Kalra R, Arora P, Morgan C, Hage FG, Iskandrian AE, Bajaj NS. Conducting and interpreting high-quality systematic reviews and meta-analyses. J Nucl Cardiol. 2017;24:471–81.
2. Liu Z, Yao Z, Li C, Liu X, Chen H, Gao C. A step-by-step guide to the systematic review and meta-analysis of diagnostic prognostic test accuracy evaluations. Br J Cancer. 2013;108:2299–303.
3. Lee S-W, Kim S-J, Seo Y, Jeong SY, Ahn BC, Lee J. F-18 FDG pet for assessment of disease activity of large vessel vasculitis: A systematic review and meta-analysis. J Nucl Cardiol. 2018. https://doi.org/10.1007/s12350-018-1406-5.
4. Rosner B. Fundamentals of biostatistics. Boston, MA: Brooks/Cole; 2011.
5. van Belle G, Fisher LD, Heagerty PJ, Lumley T. Biostatistics: A methodology for the health sciences. Hoboken: Wiley; 2004.
6. Deeks JJ, Altman DG. Diagnostic tests 4: Likelihood ratios. BMJ. 2004;329:168–9.
7. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM. The diagnostic odds ratio: A single indicator of test performance. J Clin Epidemiol. 2003;56:1129–35.
8. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity

produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58:982–90.

9. Leeflang MMG. Systematic reviews and meta-analysis of diagnostic test accuracy. Clin Microbiol Infect. 2013;20:105–13.

10. Lee J, Kim KW, Choi SH, Huh J, Park SH. Systematic review and meta-analysis of studies evaluating diagnostic test accuracy: A practical review for clinical researchers-part ii. Statistical methods of meta-analysis. Korean J Radiol. 2015;16:1188–96.

11. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20:2865–84.