# *Summarizing Studies of Diagnostic Test Performance*

In this issue of *Clinical Chemistry*, Brown et al. *(1)* report a metaanalysis of studies of the test characteristics of the latex turbidimetric D-dimer test for the diagnosis of pulmonary embolism. In this editorial, we will discuss the importance of conducting systematic reviews of diagnostic evidence and their contribution to the practice of evidence-based laboratory medicine.

Evidence-based practitioners complement, or at times substitute, diagnostic intuition with the explicit use of the best available quantitative evidence about the power of symptoms, signs, and laboratory tests to increase or decrease the probabilities associated with alternative diagnoses. Summarizing studies that have high validity will yield unbiased results, and pooling across studies will reduce the random error associated with individual smaller studies. In addition to generating more precise, accurate summaries, pooling across different patient groups will, if tests perform similarly in those groups, yield results that apply to a broader population than the individual studies. Thus, systematic summaries of valid diagnostic evidence are at the top of the hierarchy of diagnostic evidence.

Summaries of evidence will yield misleading results if they try to combine results across patient groups or test methods that are too heterogeneous; if they assemble an incomplete, biased sample of potentially available studies; or if they use results from studies that are themselves methodologically weak and very susceptible to bias. To avoid these sources of error, authors of systematic reviews should (*a*) ask a sensible question; (*b*) conduct a detailed and exhaustive search for relevant studies; (*c*) if possible, focus on studies of high methodologic quality; and (*d*) use reproducible approaches to assess the limitations in the methodologic quality of the studies on which they focus *(2)*.

Brown et al. *(1)* asked a narrowly focused and sensible question and translated their review question into appropriate eligibility criteria. Using procedures to minimize bias (e.g., use of two reviewers working independently), they applied these criteria to studies identified through a systematic search for published and unpublished evidence. These researchers also assessed eligible articles for the extent to which they included safeguards against two threats to validity: spectrum of disease and verification of test results with the reference standard. As we review below, empirical evidence suggests that these two criteria constitute important markers of bias in diagnostic studies.

Clinicians are rarely interested in the ability of a test to sort out definitely ill patients from apparently healthy volunteers. Studies that choose the severely affected as their target positive population and apparently healthy individuals as the target negative are likely to overestimate the power of the test when used in the right patients. The right patients are those in whom, before obtaining the test results, clinicians were unsure whether the patients did or did not have the target condition. To determine whether the estimates of diagnostic accuracy are unbiased, clinicians should therefore judge whether the population studied really represents a group in which genuine diagnostic uncertainty existed.

In an evaluation of 184 studies of diagnostic tests, Lijmer et al. *(3)* quantified the effect of spectrum bias, which arises when clinicians enroll very ill patients and compare their results with those from healthy controls. Studies with inadequate disease spectrum overestimated diagnostic performance threefold [relative diagnostic odds ratio (RDOR) = 3.0; 95% confidence interval (CI), 2.0–4.5] relative to those that recruited patients in whom genuine diagnostic uncertainty existed *(3)*.

Studies of diagnostic test properties can also yield biased estimates when investigators do not conduct, in all patients, a blind comparison of tests results with an independent reference standard. By blind we mean that those judging the results of the reference standard are unaware of the results of the test under evaluation and vice versa. By independent we mean that information from the test under evaluation should not affect the interpretation of the reference standard. Finally, the reference standard should be applied to all patients regardless of the results of the test under evaluation. Lijmer et al. *(3)* found that lack of blinding and verification bias (the error of using a different reference standard depending on the test result) overestimated test performance [RDOR = 1.3 (95% CI, 1.0–1.9) and 2.2 (95% CI, 1.5–3.3), respectively].

Brown et al. *(1)* determined the methodologic quality of the included studies, found that all were free from verification bias, that two had enrolled patients with less than ideal spectrum of disease, and that two had not optimally blinded the individual judging the reference standard. In summary, the authors of this systematic review asked a clinically sensible question, conducted a detailed and thorough search, and included studies of high methodologic quality. Thus, clinicians can draw strong inferences from this review.

Statistical pooling of results from individual studies, also called metaanalysis, can provide a single best estimate of diagnostic test performance *(4)*. Brown et al. *(1)* pooled the sensitivities and specificities from the studies by use of a random-effects model, a statistical technique that yields conservative estimates (i.e., wider confidence intervals) because it takes into account between-study differences. The metaanalysis yielded very precise estimates for sensitivity (93%; 95% CI, 89–96%) and specificity (51%; 95% CI, 42–59%).

Clinicians obtain diagnostic tests to lower the probability of the target condition below the testing threshold (i.e., stopping testing for it and eliminating it from further consideration) or to increase this probability above the treatment threshold (i.e., stopping testing for it and initiating appropriate therapy). The likelihood ratio (LR) best captures the direction and magnitude of this change from

pretest to posttest probability. Brown et al. *(1)* reported a LR for D-dimer >500 μg/L of 1.9 and a LR for D-dimer <500 μg/L of 0.14.

Unless compelling reasons exist to think otherwise, clinicians should feel comfortable applying the results of this review to their own clinical settings. To illustrate how clinicians could apply the results, consider the following cases. A 30-year-old healthy patient without a personal or family history of clotting disorders presents with pleuritic chest pain and palpitations. He has no leg symptoms, hypoxemia, or fever, and according to a clinical prediction rule, he has a pretest probability of pulmonary embolism of 5% *(5)*. His D-dimer test is 120 μg/L. Another patient, a 62-year-old woman with a previous deep vein thrombosis after trauma 10 years ago presents with dyspnea and palpitations and without hypoxemia or fever. Her pretest probability is 30%. Her D-dimer test is 1000 μg/L. Using the LR estimates from the metaanalysis and a nomogram *(6)* or a formula {posttest probability = pretest probability × LR ÷ [1 + pretest probability × (LR − 1)]} *(7)*, the clinician calculates the posttest probability of pulmonary embolism for each patient. The test result in the 30-year-old man lowers the probability of pulmonary embolism from 5% to 0.7%, virtually removing this condition from the differential diagnosis. The positive D-dimer test in the 62-year-old woman increases her probability of pulmonary embolism from 30% to 45%, a minimally increased posttest probability likely below the clinician's threshold to start anticoagulation, which therefore warrants further testing. As in this case, LRs of 1–2 or 0.5–1 alter probability to a small degree. On the other hand, LRs >10 or <0.1 generate large and often conclusive changes from pre- to posttest probability *(8)*.

Unfortunately, the information available to Brown et al. *(1)* did not allow them to derive LRs for narrow intervals of D-dimer results. Consequently, clinicians are left with the awkward interpretation of a D-dimer test result of 480 μg/L as negative and one of 520 μg/L as positive, when both values likely provide similar information about changes from pretest to posttest probability. Similarly, results of 480 and of 0 μg/L are given the same interpretation (negative test) when they should lead to different changes in probability (a value of 480 μg/L is likely to lower the probability of embolism but little, whereas a value of 0 μg/L will lower the probability markedly). Often, individual studies cannot fully characterize LRs at narrow result intervals because some of these intervals will not have enough patients with those test results. Ideally, evidence summaries could accumulate enough patients to estimate LRs for every possible test result *(9)*. Such results would inform clinicians of the extent to which each test result increases the probability of the target condition given the pretest probability.

Clinicians using the literature to guide practice, and reviewers such as Brown et al., have much to gain from standardized reporting of diagnostic studies. Experts in diagnostic studies and journal editors have proposed such a standard, the STARD Initiative *(10)*. STARD recommends the explicit reporting of the methodologic details critical to evaluating validity, as well as optimally informative presentation of results, allowing calculation of interval LRs in both individual studies and pooled analyses, an opportunity Brown et al. did not have.

We have suggested that the first step toward evidence-based laboratory medicine will be the routine reporting of the best available LR estimates for a given test result in laboratory reports *(8)*. Systematic reviews with meta-analyses of diagnostic tests are likely to provide these best estimates. Clinicians interested in the practice of evidence-based medicine require more and better reviews of diagnostic test accuracy, such as the one published in this issue of *Clinical Chemistry*.



## References


1. Brown MD, Lau J, Nelson RD, Kline JA. Turbidimetric D-dimer test in the diagnosis of pulmonary embolism: a metaanalysis. Clin Chem 2003;49: 1846–53.
2. Oxman AD, Guyatt GH, Cook D, Montori V. Summarizing the evidence. In: Guyatt GH, Rennie D, eds. Users' guides to the medical literature: a manual for evidence-based clinical practice. Chicago, IL: AMA Press, 2002:155–73.
3. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061–6.
4. Deeks J. Systematic reviews of evaluations of diagnostic and screening tests. In: Egger M, Smith G, Altman D, eds. Systematic reviews in health care: meta-analysis in context. London, UK: BMJ Publishing Group, 2001: 248–82.
5. Wells PS, Ginsberg JS, Anderson DR, Kearon C, Gent M, Turpie AG, et al. Use of a clinical model for safe management of patients with suspected pulmonary embolism. Ann Intern Med 1998;129:997–1005.
6. Fagan TJ. Nomogram for Bayes theorem [Letter]. N Engl J Med 1975;293: 257.
7. Prasad K. Communicating calculation of posttest probability using likelihood ratio in one step. http://bmj.com/cgi/eletters/324/7341/824#21308 (Accessed July 29, 2003).
8. Montori V, Guyatt GH. Evidence-based medicine and the diagnostic process. In: Price C, Christenson R, eds. Evidence-based laboratory medicine. Washington, DC: AACC Press, 2003:1–19.
9. Guyatt GH, Oxman AD, Ali M, Willan A, McIlroy W, Patterson C. Laboratory diagnosis of iron-deficiency anemia: an overview. J Gen Intern Med 1992; 7:145–53.
10. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Clin Chem 2003;49:1–6.


**Victor M. Montori[1,2]**
**Gordon H. Guyatt[2]***

[1] *Department of Medicine*
*Mayo Clinic*
*Rochester, MN 55905*

[2] *Departments of Clinical Epidemiology and Biostatistics and Medicine*
*McMaster University*
*Hamilton, Ontario L8N 3Z5, Canada*

*Author for correspondence.

DOI: 10.1373/clinchem.2003.024935