Systematic reviews of diagnostic test evaluations: What's behind the scenes?

As readers of *ACP Journal Club,* you are aware that systematic reviews are considered the best source of evidence for evidence-based clinical practice. Systematic reviews synthesize data from existing primary research and bring some order and sanity to the otherwise-stressful process of sorting out a plethora of studies and staying up to date. However, since not all reviews are created equal, it is important to be able to critically assess their quality. In this editorial, we take you behind the scenes of a systematic review, using diagnostic test accuracy as an illustration. We hope that a clear understanding of the process will guide what you look for in a review. Further, if you can't find an existing diagnostic review and decide to do one yourself, we've provided you with a "road map" (Figure) for navigation.

STEPS IN THE SYSTEMATIC REVIEW PROCESS

Systematic reviews are done on a range of clinical questions, such as therapy, diagnosis, prognosis, etiology, harm, and disease prevalence. All systematic reviews follow the same critical steps:

1. Formulation of the review question
2. A comprehensive, systematic search and selection of primary studies
3. Critical appraisal of included studies for quality and data extraction
4. Synthesis and summary of study results
5. Interpretation of the results

These steps resemble those of the evidence-based medicine (EBM) process, but are more thorough. In the EBM process, our objective is to quickly hunt down a valid source of evidence (e.g., a high-quality systematic review) on a focused clinical question and get to the bottom line (i.e., clinically meaningful results) within minutes. In contrast, the systematic review involves a comprehensive search for all published and unpublished primary studies on a focused question, critical appraisal of the relevant studies, and synthesis of these studies to generate evidence for clinical practice. This process typically takes months, not minutes.

The core steps of the systematic review process (shaded boxes in the Figure) can be broken down further into more discrete steps. Based on our experience in conducting reviews and developing training material (see www.medepi.org/meta), we have provided some helpful tricks and tips for surviving the process. For all the major steps, we have provided references to important articles and resources.

SYSTEMATIC REVIEWS OF DIAGNOSTIC TEST ACCURACY

Although not as common as systematic reviews on therapeutic questions (i.e., of randomized controlled trials [RCTs]), an increasing number of diagnostic reviews are being published in the medical literature. The main objective of a diagnostic review is to summarize the evidence on the accuracy of a test or instrument (in this case, accuracy refers to such measures as sensitivity [Se], specificity [Sp], and likelihood ratios [LRs]). The other objectives are to critically evaluate the quality of primary studies, check for heterogeneity (variability) in results across studies, and determine sources of heterogeneity, where necessary.

FORMULATION OF THE REVIEW QUESTION

The first step is to formulate a clear, focused review question. It is important to specify the patient population (or the disease of interest) and setting, the index test (or tests) being evaluated, the reference standard (comparison), and the outcomes (e.g., sensitivity and specificity). For example, consider a review on ultrasonography for suspected deep venous thrombosis. A focused question would be: *Is ultrasonography [test] a sensitive and specific [outcomes] test compared with venography [reference standard] in the diagnosis of suspected deep venous thrombosis in adults [patients]?* A focused question will help in searching databases and with formulating explicit eligibility criteria for selecting studies.

COMPREHENSIVE SEARCH AND SELECTION OF PRIMARY STUDIES

The second step is to conduct an exhaustive search for primary studies. The search might include general databases (e.g., MEDLINE and EMBASE/Excerpta Medica), subject-specific databases (e.g., MEDION, a database of diagnostic literature, http://www.mediondatabase.nl), scanning bibliographies of included studies, contacting authors and experts to locate ongoing and unpublished studies, and contacting test manufacturers. It is important to extend the search beyond MEDLINE and cover other databases as well. Once all sources have been searched, the accumulated citations are screened independently by 2 reviewers who select those studies that will be included in the review. This process reduces missed studies and bias in study selection.

CRITICAL APPRAISAL OF INCLUDED STUDIES FOR QUALITY AND DATA EXTRACTION

The third step is to critically appraise included studies. Quality assessment, again, is ideally done independently by 2 reviewers. Several quality criteria need to be considered when evaluating diagnostic studies. These include the clinical spectrum of included patients; blinded interpretation of test and reference standard results; potential for verification bias; consecutive patient sampling; prospective design; and adequate description of the index test, reference standard, and study population. Often, several of these features may not be reported in the primary studies. Reviewers might need to contact authors of the studies and seek additional information. Reviewers might choose to exclude low-quality studies from the review at this stage. An alternative approach would be to stratify studies by quality at the time of analysis and examine the effect of study quality on test accuracy.

Data extraction is done in parallel with quality assessment. The outcomes reported in diagnostic reviews are the measures of accuracy: Se, Sp, LR, diagnostic odds ratios (DORs), and receiver operating-characteristic (ROC) curve data. Where possible, reviewers should extract raw data to fill the 4 cell values of a diagnostic 2×2 table: true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs).

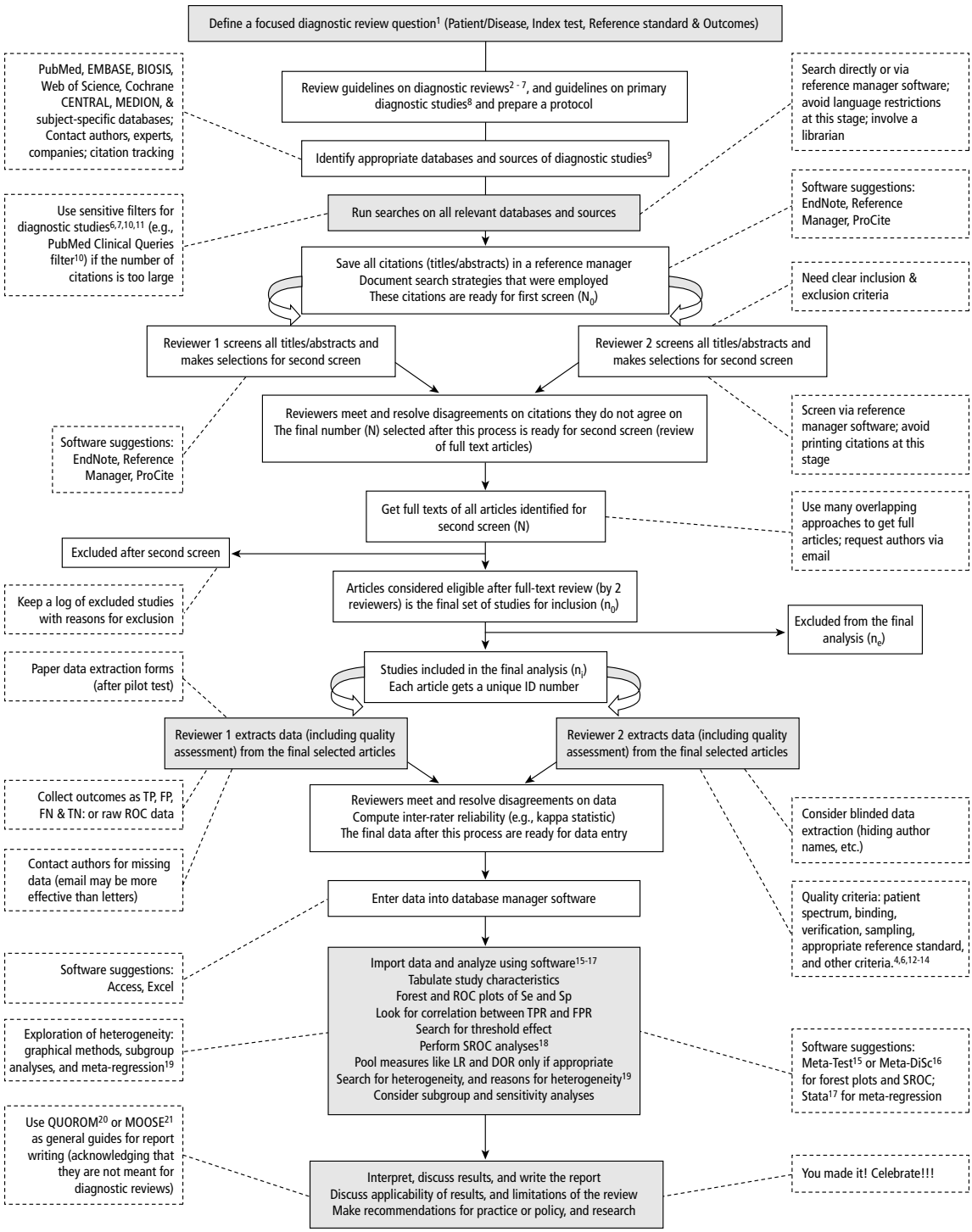

Figure. **"Road Map" for systematic reviews of diagnostic test evaluations.** Se = sensitivity; Sp = specificity; LRs = likelihood ratios; DORs = diagnostic odds ratios; ROC = receiver operating characteristic; SROC = summary receiver operating characteristic; TPs = true positives; FPs = false positives; TNs = true negatives; FNs = false negatives; TPR = true-positive rate; FPR = false positive rate. Superscripts indicate reference numbers.

SYNTHESIS AND SUMMARY OF STUDY RESULTS (META-ANALYSIS)

Analysis begins with simple tabulation of study characteristics and results. Forest plots of accuracy measures (e.g., Se and Sp) show estimates from each study with their confidence intervals. These plots provide a useful visual summary of the data. Although, as with intervention studies, all measures of accuracy can be statistically pooled using random- or fixed-effects methods, this may not always be appropriate. Each study in the meta-analysis contributes a pair of numbers: true-positive rate (Se) and false-positive rate (1 – Sp). Because these measures are correlated and vary with the thresholds (cutpoints for determining test positives) used, it is important to analyze them as pairs and to explore the effect of threshold on study results. Simple pooling of accuracy measures does not address these important issues. A more meaningful approach is to summarize the joint distribution of Se and Sp using the summary ROC curve. Unlike a traditional ROC plot that explores the effect of varying thresholds on Se and Sp in a single study, each data point in the summary ROC space represents a separate study. The summary ROC curve is obtained by fitting a regression curve to pairs of Se and Sp. The summary ROC curve and the area under it present a global summary of test performance and show the trade-off between Se and Sp. A symmetric, shoulder-like ROC curve suggests that variability in the thresholds used could, in part, explain variability in study results.

Heterogeneity in meta-analysis refers to a high degree of variability in study results, a fairly common finding in diagnostic meta-analyses. For example, one might find reviews with Se estimates ranging from 0% to 100%. Such heterogeneity could be due to variability in thresholds, disease spectrum, test methods, and study quality. In the presence of significant heterogeneity, the pooled summary estimate is not meaningful. Reviewers should then focus on finding sources of heterogeneity. This can be accomplished by looking at the details of the studies (e.g., selection of patients or test procedure), examining subgroups to look for homogeneous populations, and by meta-regression, to statistically assess the differences in study design that might explain variation in findings. Graphical methods can also be used to identify sources of heterogeneity.

INTERPRETATION OF THE RESULTS

The final steps in the systematic review process are interpretation of the results, including discussion of such issues as applicability, and writing the report for publication. Reviewers also need to discuss the limitations of the primary studies reviewed and limitations of the review itself. The review usually concludes with a discussion on implications for clinical practice and the need for further research on the clinical question.

CONCLUSION

Just as systematic reviews of high-quality clinical trials are considered to be at the top of the hierarchy of evidence for treatment, properly conducted systematic reviews of valid diagnostic studies are at the top of the hierarchy of diagnostic evidence. A clear understanding of how systematic reviews are done will help clinicians appreciate the strengths and limitations of the reviews they read.

Madhukar Pai, MD
Michael McCulloch, LAc, MPH
Wayne Enanoria, MPH
John M. Colford Jr., MD, PhD
For the Berkeley Systematic Reviews Group
Division of Epidemiology
University of California, Berkeley
Berkeley, California, USA

References

1. **Richardson WS, Wilson MC, Nishikawa J, Hayward RS.** The well-built clinical question: a key to evidence-based decisions [Editorial]. ACP J Club. 1995 Nov-Dec;123:A12-3.
2. **Irwig L, Tosteson AN, Gatsonis C, et al.** Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med. 1994;120:667-76.
3. **Irwig L, Macaskill P, Glasziou P, Fahey M.** Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol. 1995;48:119-30.
4. Cochrane Methods Group on Systematic Review of Screening and Diagnostic Tests: Recommended Methods. Updated 6 June 1996. www.cochrane.org/cochrane/sadtdoc1.htms
5. **Deeks JJ.** Systematic reviews of evaluations of diagnostic and screening tests. In: Egger M, Smith GD, Altman DG, eds. Systematic reviews in health care. Meta-analysis in context. London: BMJ Publishing Group; 2001:248-82.
6. **Deville WL, Buntinx F, Bouter LM, et al.** Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol. 2002;2:9.
7. **Battaglia M, Bucher H, Egger M, et al.** The Bayes Library of Diagnostic Studies and Reviews. 2d edition. 2002. www.ispm.unibe.ch/downloads
8. **Bossuyt PM, Reitsma JB, Bruns DE, et al.** Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ. 2003;326:41-4.
9. Centre for Reviews and Dissemination. Finding Studies for Systematic Reviews: a Checklist for Researchers. Updated January 2004. York, UK: University of York. www.york.ac.uk/inst/crd/revs.htm
10. **Haynes RB, Wilczynski NL.** Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. BMJ. 2004;328:1040-4.
11. **McKibbon A, Eady A, Marks S.** PDQ Evidence-based Principles and Practice. Hamilton, Canada: BC Decker; 1999:61-82.
12. **Lijmer JG, Mol BW, Heisterkamp S, et al.** Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-6.
13. **Jaeschke R, Guyatt G, Lijmer J.** Diagnostic tests. In: Guyatt G, Rennie D, eds. Users' guides to the medical literature. A manual for evidence-based clinical practice. Chicago: AMA Press; 2002:121-40.
14. **Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J.** The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003;3:25.
15. **Lau J.** Meta-Test version 0.6. 1997. www.cochrane.org/cochrane/sadt.htm
16. **Zamora J, Muriel A, Abraira V.** Meta-DiSc for Windows: A Software package for the Meta-analysis of Diagnostic Tests. XI Cochrane Colloquium. Barcelona, 2003. www.hrc.es/investigacion/metadisc.html
17. **Sterne JA, Bradburn MJ, Egger M.** Meta-analysis in Stata. In: Egger M, Smith GD, Altman DG, eds. Systematic reviews in health care. Meta-analysis in context. London: BMJ Publishing Group; 2001:347-69.
18. **Littenberg B, Moses LE.** Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decis Making. 1993; 13:313-21.
19. **Lijmer JG, Bossuyt PM, Heisterkamp SH.** Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med. 2002;21:1525-37.
20. **Moher D, Cook DJ, Eastwood S, et al.** Improving the quality of reports of meta-analyses of randomized controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet. 1999;354:1896-1900.
21. **Stroup DF, Berlin JA, Morton SC, et al.** Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA. 2000;283:2008-12.