EBM notebook

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

INTRODUCTION

As readers of *Evidence-Based Medicine,* you are aware that systematic reviews are considered the best source of evidence for evidence-based clinical practice. Systematic reviews synthesise 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. A clear understanding of the process will, hopefully, guide what you look for in a review. Furthermore, if you can't find an existing diagnostic review and decide to do one yourself, we provide 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, aetiology, 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 (eg, a high quality systematic review) on a focused clinical question and get to the bottom line (clinically meaningful results) within minutes. In contrast, the systematic review involves doing a comprehensive search for all published and unpublished primary studies on a focused question, critical appraisal of the relevant studies, and synthesis of these studies to generate evidence for clinical practice.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 http://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 (ie, of randomised controlled trials [RCTs]), diagnostic reviews are increasingly published in the medical literature. The main objective of a diagnostic review is to summarise the evidence on the accuracy of a test or instrument (in this case accuracy refers to such measures as sensitivity, specificity, and likelihood ratios). 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 (eg, 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 also with formulating explicit eligibility criteria for selecting studies.
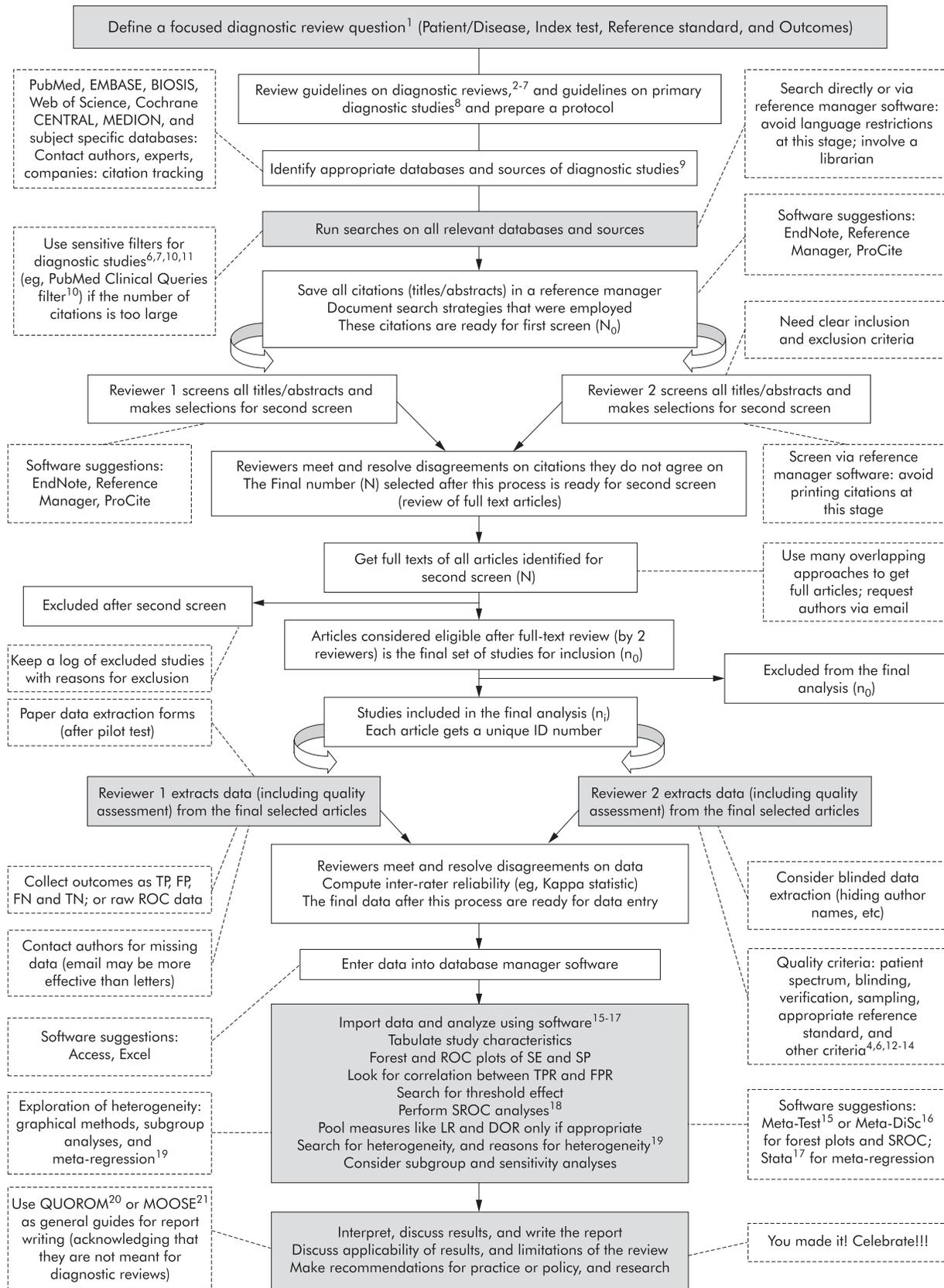
COMPREHENSIVE SEARCH AND SELECTION OF PRIMARY STUDIES

The second step is to conduct an exhaustive search for primary studies. The search might include general databases (eg, Medline and EMBASE/Excerpta Medica), subject specific databases (eg, 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 by 2 reviewers independently. 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

A ''road map'' for systematic reviews of diagnostic test evaluations

Define a focused diagnostic review question[1] (Patient/Disease, Index test, Reference standard, and Outcomes)

PubMed, EMBASE, BIOSIS, Web of Science, Cochrane CENTRAL, MEDION, and subject specific databases: Contact authors, experts, companies: citation tracking

Review guidelines on diagnostic reviews,[2-7] and guidelines on primary diagnostic studies[8] and prepare a protocol

Search directly or via reference manager software: avoid language restrictions at this stage; involve a librarian

Identify appropriate databases and sources of diagnostic studies[9]

Run searches on all relevant databases and sources

Software suggestions: EndNote, Reference Manager, ProCite

Use sensitive filters for diagnostic studies[6,7,10,11] (eg, PubMed Clinical Queries filter[10]) if the number of citations is too large

Save all citations (titles/abstracts) in a reference manager
Document search strategies that were employed
These citations are ready for first screen (N_0)

Need clear inclusion and exclusion criteria

Reviewer 1 screens all titles/abstracts and makes selections for second screen

Reviewer 2 screens all titles/abstracts and makes selections for second screen

Software suggestions: EndNote, Reference Manager, ProCite

Reviewers meet and resolve disagreements on citations they do not agree on
The Final number (N) selected after this process is ready for second screen
(review of full text articles)

Screen via reference manager software: avoid printing citations at this stage

Get full texts of all articles identified for second screen (N)

Use many overlapping approaches to get full articles; request authors via email

Excluded after second screen

Articles considered eligible after full-text review (by 2 reviewers) is the final set of studies for inclusion (n_0)

Excluded from the final analysis (n_0)

Keep a log of excluded studies with reasons for exclusion

Paper data extraction forms (after pilot test)

Studies included in the final analysis (n_i)
Each article gets a unique ID number

Reviewer 1 extracts data (including quality assessment) from the final selected articles

Reviewer 2 extracts data (including quality assessment) from the final selected articles

Collect outcomes as TP, FP, FN and TN; or raw ROC data

Reviewers meet and resolve disagreements on data
Compute inter-rater reliability (eg, Kappa statistic)
The final data after this process are ready for data entry

Consider blinded data extraction (hiding author names, etc)

Contact authors for missing data (email may be more effective than letters)

Enter data into database manager software

Quality criteria: patient spectrum, blinding, verification, sampling, appropriate reference standard, and other criteria[4,6,12-14]

Software suggestions: Access, Excel

Import data and analyze using software[15-17]
Tabulate study characteristics
Forest and ROC plots of SE and SP
Look for correlation between TPR and FPR
Search for threshold effect
Perform SROC analyses[18]
Pool measures like LR and DOR only if appropriate
Search for heterogeneity, and reasons for heterogeneity[19]
Consider subgroup and sensitivity analyses

Software suggestions: Meta-Test[15] or Meta-DiSc[16] for forest plots and SROC; Stata[17] for meta-regression

Exploration of heterogeneity: graphical methods, subgroup analyses, and meta-regression[19]

Use QUOROM[20] or MOOSE[21] as general guides for report writing (acknowledging that they are not meant for diagnostic reviews)

Interpret, discuss results, and write the report
Discuss applicability of results, and limitations of the review
Make recommendations for practice or policy, and research

You made it! Celebrate!!!

Se = sensitivity; Sp = specificity; LR = likelihood ratios; DOR = diagnostic odds ratios; ROC = receiver operating characteristic; SROC = summary receiver operating characteristic; TP = true positives; FP = false positives; TN = true negatives; FN = false negatives; TPR = true positive rate; FPR = false positive rate. Superscripts indicate reference numbers.

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: sensitivity, specificity, likelihood ratios, diagnostic odds ratios, 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, false positives, true negatives, and false negatives.

SYNTHESIS AND SUMMARY OF STUDY RESULTS (META-ANALYSIS)

Analysis begins with simple tabulation of study characteristics and results. Forest plots of accuracy measures (eg, sensitivity and specificity) 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 (sensitivity) and false positive rate (1 − specificity). Because these measures are correlated and vary with the thresholds (cutpoints for determining test positives) used, it is important to analyse them as pairs and to also 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 summarise the joint distribution of sensitivity and specificity using the summary ROC curve. Unlike a traditional ROC plot that explores the effect of varying thresholds on sensitivity and specificity 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 sensitivity and specificity. The summary ROC curve and the area under it present a global summary of test performance and show the trade off between sensitivity and specificity. 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 sensitivity estimates ranging from 0–100%. 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 (eg, 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 of 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
Berkeley Systematic Reviews Group
University of California, Berkeley
Berkeley, California, USA

1 Richardson WS, Wilson MC, Nishikawa J, *et al*. The well-built clinical question: a key to evidence-based decisions [editorial]. *ACP J Club* 1995;**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, *et al*. 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. http://www.cochrane.org/cochrane/sadtdoc1.htm.

5 Deeks JJ. Systematic reviews of evaluations of diagnostic and screening tests. In: Egger M, Smith GD, Altman DG, editors. *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. 2nd edition. 2002. http://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, http://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, editors. *Users' guides to the medical literature. A manual for evidence-based clinical practice*. Chicago: AMA Press, 2002:121–140.

14 Whiting P, Rutjes AW, Reitsma JB, *et al*. 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. http://www.hrc.es/investigacion/metadisc.html

17 Sterne JA, Bradburn MJ, Egger M. Meta-analysis in Stata. In: Egger M, Smith GD, Altman DG, editors. *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.