Clinical Research Methods

Systematic reviews and meta-analyses: An illustrated, step-by-step guide

MADHUKAR PAI, MICHAEL McCULLOCH, JENNIFER D. GORMAN, NITIKA PAI, WAYNE ENANORIA, GAIL KENNEDY, PRATHAP THARYAN, JOHN M. COLFORD, Jr

ABSTRACT

Systematic reviews and meta-analyses synthesize data from existing primary research, and well-conducted reviews offer clinicians a practical solution to the problem of staying current in their fields of interest. A whole generation of secondary journals, pre-appraised evidence libraries and periodically updated electronic texts are now available to clinicians. However, not all systematic reviews are of high quality, and it is important to be able to critically assess their validity and applicability. This article is an illustrated guide for conducting systematic reviews. A clear understanding of the process will provide clinicians with the tools to judiciously appraise reviews and interpret them. We hope that it will enable clinicians to conduct systematic reviews, generate high-quality evidence, and contribute to the evidence-based medicine movement.

Natl Med J India 2004;17:86-95

INTRODUCTION

Evidence-based medicine (EBM) is the process of 'integrating individual clinical expertise with the best available external clinical evidence from systematic research'.[1] The EBM approach requires healthcare decisions to be made on the basis of strong evidence generated by high-quality research studies.[1,2] In this context, 'evidence' derives from a state-of-the-art synthesis (review) of all research conducted regarding a particular clinical question.[2] Clinicians have always used review articles as sources of evidence, and these reviews can be useful tools if conducted properly. Unfortunately, empirical studies have shown that narrative review articles tend to be of poor quality.[3]

What is a narrative review and how is it different from a systematic review? Traditional, narrative reviews, usually written by experts, are qualitative summaries of evidence on a given topic. Typically, they involve informal, subjective methods to collect and interpret studies, and tend to selectively cite literature that reinforces preconceived notions.[3] Narrative reviews often do not explicitly describe how the reviewers searched, selected and appraised the quality of studies (Table I).[3] In contrast, a systematic review includes a comprehensive, exhaustive search for primary studies on a focused clinical question, selection of studies using clear and reproducible eligibility criteria, critical appraisal of studies for quality, and synthesis of results according to a pre-determined and explicit method (Table I).[4-7]

What is a meta-analysis? A meta-analysis is the statistical pooling of data across studies to generate summary (pooled) estimates of effects.[4,6] The term 'effect' refers to any measure of association between exposure and outcome (e.g. odds ratio). A meta-analysis is usually the final step in a systematic review. All meta-analyses should ideally start with an unbiased systematic review that incorporates articles chosen using predetermined inclusion criteria.[4,6] If the data extracted from these studies meet certain requirements (the most important being a high level of homogeneity of effect measures across studies), then the data can be combined using meta-analysis. However, if the effect measures are found to be heterogeneous, then it is still acceptable to present the work as a systematic review and not perform meta-analysis, or use statistical methods that can account for the heterogeneity. Indeed, there are situations when a meta-analysis is clearly inappropriate. Therefore, meta-analyses and systematic reviews are not synonymous.[4] Ideally, a meta-analysis should be performed as part of a systematic review (Fig. 1). In practice, meta-analyses are sometimes done without an initial systematic review. Within the set of meta-analyses, the investigators will sometimes choose to go beyond the analyses of published studies, contact authors of the primary studies for data on individual patients in their studies, and then combine the raw data. This is called an

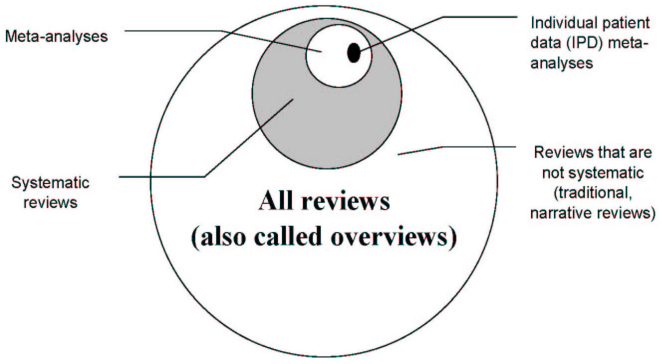


FIG 1. Types of review articles

University of California, Berkeley, CA 94720, USA
MADHUKAR PAI, MICHAEL McCULLOCH, JENNIFER D. GORMAN, NITIKA PAI, WAYNE ENANORIA, JOHN M. COLFORD, Jr
Berkeley Systematic Reviews Group, Division of Epidemiology

Cochrane HIV/AIDS Review Group, University of California, San Francisco, CA 94105 USA
MADHUKAR PAI, GAIL KENNEDY

Cochrane Schizophrenia Group and Christian Medical College, Vellore, 632002, India
PRATHAP THARYAN Department of Psychiatry

Correspondence to MADHUKAR PAI, Division of Epidemiology, University of California at Berkeley, 140 Warren Hall, Berkeley CA 94720; *madhupai@berkeley.edu*

TABLE I. Comparison of traditional and systematic reviews

Components of a review	Traditional, narrative reviews	Systematic reviews
Formulation of the question	Usually address broad questions	Usually address focused questions
Methods section	Usually not present, or not well-described	Clearly described with pre-stated criteria about participants, interventions and outcomes
Search strategy to identify studies	Usually not described; mostly limited by reviewers' abilities to retrieve relevant studies; usually not reproducible and prone to selective citation	Clearly described and usually exhaustive; transparent, reproducible and less prone to selective citation
Quality assessment of identified studies	Usually all identified studies are included without explicit quality assessment	Only high-quality studies are included using pre-stated criteria; if lower-quality studies included, the effects of this are tested in subgroup analyses
Data extraction	Methods usually not described	Usually undertaken by more than one reviewer onto pre-tested data forms; attempts often made to obtain missing data from authors of primary studies
Data synthesis	Qualitative description employing the 'vote counting' approach, where each included study is given equal weight, irrespective of study size and quality	Meta-analysis assigns higher weights to effect measures from more precise studies; pooled, weighted effect measures with confidence limits provide power and precision to results
Heterogeneity	Usually dealt with in a narrative fashion	Heterogeneity dealt with by graphical and statistical methods; attempts are often made to identify sources of heterogeneity
Interpreting results	Prone to cumulative systematic biases and personal opinion	Less prone to systematic biases and personal opinion

individual patient data (IPD) meta-analysis (Fig. 1).

Where can one find the best evidence for EBM? Systematic reviews and meta-analyses are widely considered the best sources of evidence.[4–7] A major challenge for clinicians today is to keep up with the literature.[2] Well-conducted systematic reviews offer busy clinicians a practical solution to the problem of staying up to date. In fact, a whole generation of secondary journals (e.g. *ACP Journal Club*, *Evidence Based Medicine*), pre-appraised evidence libraries (e.g. *Cochrane Library*), and periodically updated electronic textbooks (e.g. *UpToDate*) are now available to clinicians.[2] However, since not all reviews are of high quality, it is important to be able to critically assess their quality. In this article, we present the architecture of a systematic review. A clear understanding of the underlying process will, hopefully, help clinicians to critically appraise reviews. For those who plan to conduct reviews, we provide an illustrated, step-by-step guide.

STEPS IN CONDUCTING A SYSTEMATIC REVIEW

Systematic reviews can be performed for questions relating to therapy, prevention, diagnosis, prognosis, aetiology and harm.[5] The key steps in a systematic review are: (i) formulation of a focused review question; (ii) a comprehensive, exhaustive search and inclusion of primary studies; (iii) quality assessment of included studies and data extraction; (iv) synthesis of study results (meta-analysis); and (v) interpretation of the results and report-writing.[4–7] Figure 2 presents the systematic review process. The core five steps of the process (shaded boxes in Fig. 2) are shown in greater detail. Based on our experience in conducting reviews[8,9] and developing training resources[10] (see *www.medepi.org/meta*), we present practical tips that we hope readers will find useful in performing reviews.

The central objective of a systematic review is to summarize the evidence on a specific clinical question.[4–7] Secondary objectives are to critically evaluate the quality of the primary studies, check for and identify sources of heterogeneity in results across studies, and, if necessary and possible, determine sources of heterogeneity.[4–7] Systematic reviews are also helpful in identifying new research questions. Ideally, every research study should begin with a systematic review and build upon the existing evidence base.

FORMULATION OF THE QUESTION

Because systematic reviews are time-consuming, it is important to first ascertain if a review is already available on the topic of interest. Reviewers could search sources of reviews (e.g. *Cochrane Library*), and PubMed (using filters for systematic reviews) before embarking on a new review. Once a decision is made to conduct a review, the first step is to formulate a clear, focused question[11] and prepare a protocol. The acronym PICO (patient, intervention, comparison and outcome) is often used to identify the four critical parts of a well-built clinical question.[6,11] The protocol should specify the patient population (or the disease of interest), the intervention (or exposure) being evaluated, the comparison intervention (if applicable), and the outcome. For example, consider a review on Chinese herbal medicines for the treatment of hepatitis B.[8] A focused question will be: among patients with chronic hepatitis B (patient), are Chinese herbal medicines (intervention) helpful in increasing the response to alpha-interferon (outcome) as compared to interferon therapy used alone (comparison)? A focused question will help in conducting more specific searches of databases, and also in creating unambiguous criteria for selecting studies.

SEARCH AND INCLUSION OF PRIMARY STUDIES

The next step is to conduct an exhaustive search for primary studies.[4–7,12] The search might include general databases (e.g. PubMed; Table II), subject-specific databases (e.g. Cancerlit; Table II), screening of bibliographies of included studies, handsearch of relevant journals, contact with authors and experts to locate ongoing and unpublished studies, and contact with pharmaceutical companies to identify studies.[12] Empirical research suggests that searching PubMed alone is inadequate.[13] It is, therefore, important to search databases other than PubMed. For identifying

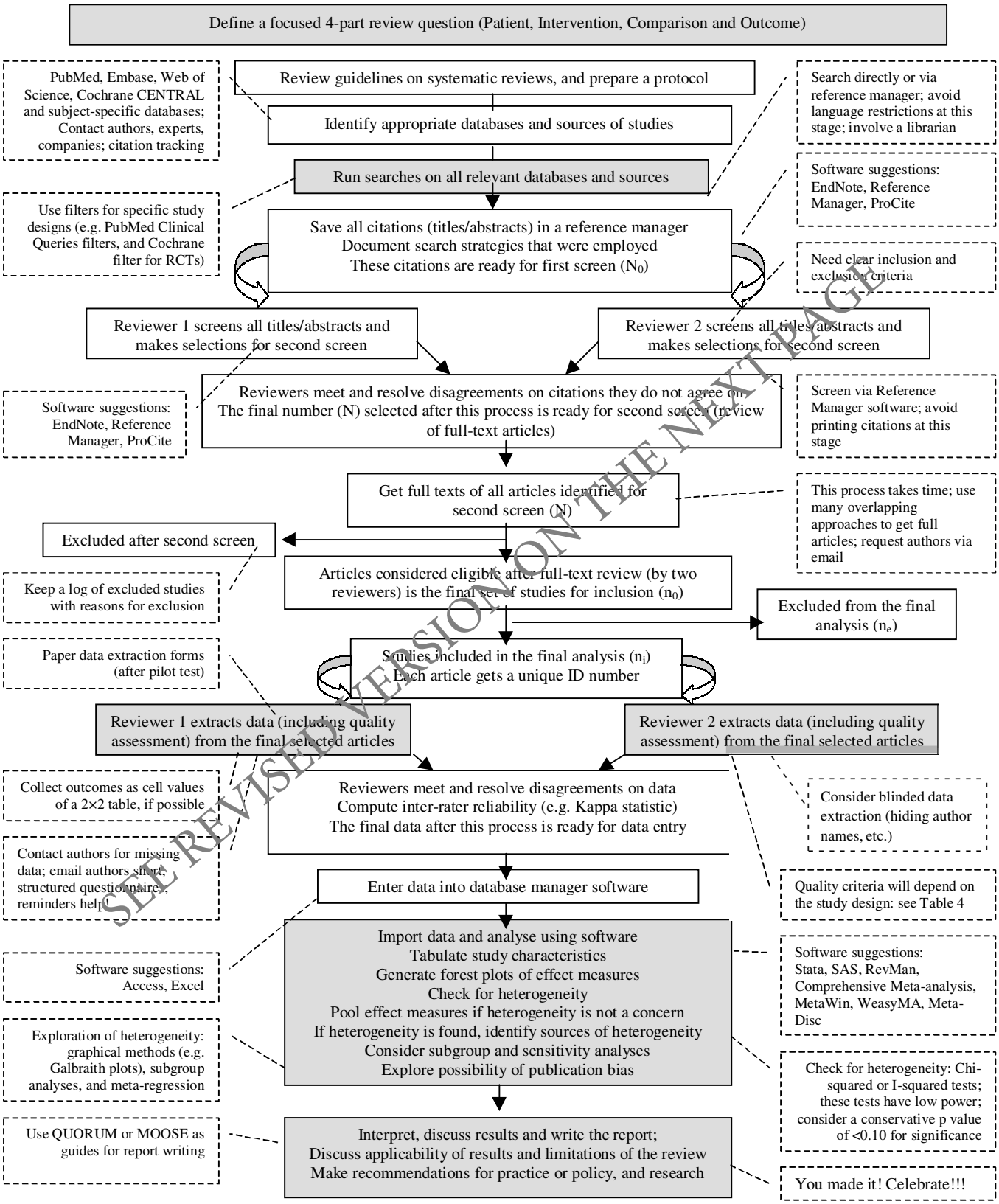
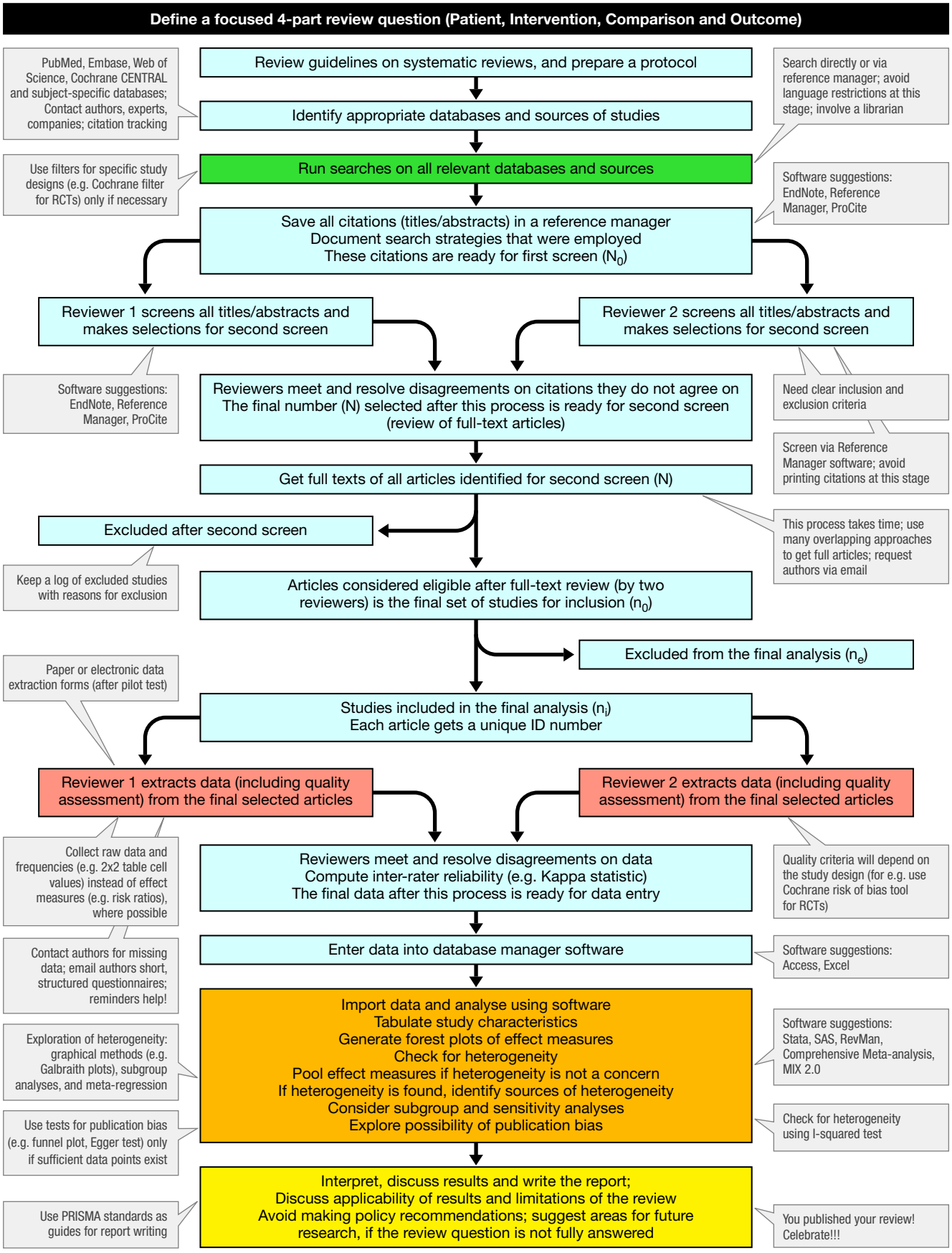


FIG 2. Steps in conducting a systematic review *Source:* Adapted from reference 10 and reproduced with permission from the BMJ Publishing Group and American College of Physicians

A ROADMAP FOR SYSTEMATIC REVIEWS & META-ANALYSES



Adapted from: Pai M et al. Natl Med J India 2004 Mar-Apr;17(2):86-95.

TABLE II. An overview of general and subject-specific electronic databases

General databases		Subject-specific databases	
Database	Access	Database	Access
PubMed (Medline), a database with over 14 million citations	Free access via: *http://www.pubmed.gov*	CANCERLIT, a cancer database from the National Cancer Institute, USA	Free access via: *www.cancer.gov/search/cancer_literature/*
Embase, Excerpta Medica, a database with over 9 million citations	Requires subscription; URL: *www.embase.com*	PsycINFO, a database on psychological and mental health literature	Requires subscription: *http://www.apa.org/psycinfo/*
Cochrane Controlled Trials Register (CENTRAL), a source of >400 000 trials	Requires subscription URL: *www.cochranelibrary.com/enter/* (selected developing countries have free access)	AIDSLINE, a database on HIV and AIDS literature	Free access via NLM Gateway: *http://gateway.nlm.nih.gov/gw/Cmd*
NLM Gateway, includes databases such as Medline, AIDSLINE, AIDS conference abstracts, etc.	Free access via: *http://gateway.nlm.nih.gov/gw/Cmd*	CINAHL: database for nursing, occupational therapy, physical therapy and other allied health fields	Requires subscription URL: *http://www.cinahl.com/*
DARE, Database of Abstracts of Reviews of Effects, is a source of systematic reviews	Free access via: *www.york.ac.uk/inst/crd/darehp.htm*	LILACS: a medical database on Latin American and Caribbean literature	Free access via: *http://www.bireme.br/bvs/I/ibd.htm*

For a more comprehensive checklist of sources of studies, see reference 12

randomized controlled trials (RCTs), the best single source is the Cochrane CENTRAL register, with more than 400 000 trials. This register is a part of the *Cochrane Library* that contains the Cochrane Database of Systematic Reviews (Table II).

What is the best strategy for searching databases? An effective strategy (Fig. 3) is to conduct separate, sensitive searches (using multiple, alternative terms combined with the Boolean operator 'OR') for each component of the PICO set, and then combine the separate searches using the operator 'AND'. Using 'OR' for each of the PICO searches will 'explode' the search and make it highly sensitive (i.e. likely to yield thousands of citations). Using 'AND' at the end of the process will dramatically narrow the search and select articles that contain all of the PICO terms (the intersection of PICO circles in Fig. 3). If reviewers decide to restrict the search to a specific study type (e.g. randomized controlled trials — RCT), then appropriate 'filters' (Table III) can be used to extract specific types of studies.[14,15]

After searching all sources, it is helpful to export all the citations into a reference manager software (e.g. EndNote: *www.endnote.com*). This allows reviewers to keep track of the included and excluded studies, maintain a log of why specific studies were excluded and eliminate the need to print out hundreds of abstracts for screening. The accumulated citations are then screened (electronically using the reference manager) independently by two reviewers who select those studies appropriate for inclusion in the review (Fig. 2). This process lessens the likelihood of missing relevant studies and reduces subjectivity in study selection. When the two reviewers disagree on the inclusion or exclusion of a specific study, they can resolve the disagreement by consensus, or request a third person to settle the disagreement.

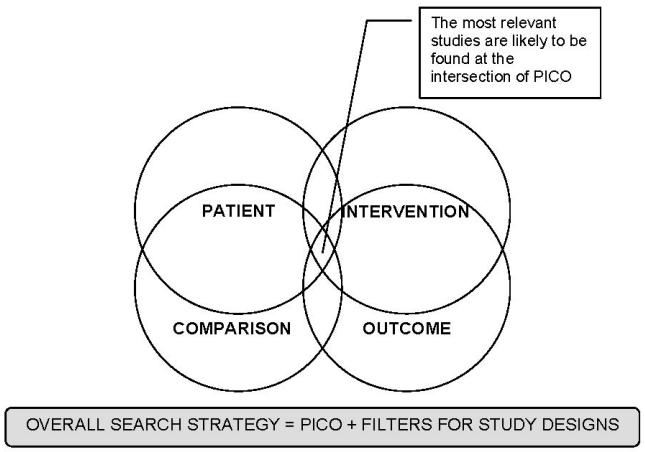


FIG 3. An overview of the literature search strategy

QUALITY ASSESSMENT AND DATA EXTRACTION

The next step is quality assessment of the included studies. This should also be performed independently by two reviewers (Fig. 2). Quality refers to internal validity of the studies (i.e. lack of bias). The quality criteria used will depend on the study design (Table IV).[2] For example, issues such as randomization, concealment of allocation and blinding are important quality features of RCTs.[2] Often, these features may not be reported in the primary studies. For example, a trial report may not mention anything about blinding. In this case, it is not clear if the trial should be coded as 'unblinded' or as 'not reported' for that criterion. In such situations, reviewers could contact the study authors for clarification. If no further information is received, we recommend classifying the study as 'not reported' with respect to blinding; at times, reviewers will classify such studies as 'unblinded' in the absence of information but we do not believe that this is appropriate. After quality assessment is complete, reviewers might decide to exclude low-quality studies from the review. An alternative and useful approach is to stratify studies by quality at the time of meta-analysis, and examine the impact of study quality on summary effect measures.

Data extraction, along with quality assessment, is done using data extraction forms developed after pilot testing (for sample data forms see *www.medepi.org/meta*). Reviewers usually extract information on study characteristics, methodology, population, interventions and outcomes. The outcomes reported in systematic reviews vary, depending on the types of studies included. If RCTs are included, the outcomes are usually expressed as risk ratios (RR), odds ratios (OR) or difference between means for continuous outcomes. In systematic reviews of diagnostic studies, the

TABLE III. Search filters for specific study designs in PubMed

Filter and purpose	PubMed search string	References
Cochrane highly sensitive search filter for controlled trials in PubMed	(randomized controlled trial[pt] OR controlled clinical trial[pt] OR randomized controlled trials[mh] OR random allocation[mh] OR double-blind method[mh] OR single-blind method[mh] OR clinical trial[pt] OR clinical trials[mh] OR ("clinical trial"[tw]) OR ((singl*[tw] OR doubl*[tw] OR trebl*[tw] OR tripl*[tw]) AND (mask*[tw] OR blind*[tw])) OR ("latin square"[tw]) OR placebos[mh] OR placebo*[tw] OR random*[tw] OR research design[mh:noexp] OR comparative study[mh] OR evaluation studies[mh] OR follow-up studies[mh] OR prospective studies[mh] OR cross-over studies[mh] OR control*[tw] OR prospectiv*[tw] OR volunteer*[tw]) NOT (animal[mh] NOT human[mh])	15
PubMed Clinical Queries sensitive filter for diagnostic studies	"sensitivity and specificity" [MESH] OR "sensitivity" [WORD] OR "diagnosis" [SH] OR "diagnostic use" [SH] OR "specificity" [WORD]	14
PubMed Clinical Queries sensitive filter for aetiological studies (e.g. case–control)	"cohort studies" [MESH] OR "risk" [MESH] OR ("odds" [WORD] AND "ratio*" [WORD]) OR ("relative" [WORD] AND "risk" [WORD]) OR "case-control*" [WORD] OR case-control studies [MESH]	14
A sensitive search strategy for systematic reviews in PubMed	((meta-analysis [pt] OR meta-analysis [tw] OR metanalysis [tw]) OR ((review [pt] OR guideline [pt] OR consensus [ti] OR guideline* [ti] OR literature [ti] OR overview [ti] OR review [ti]) AND ((Cochrane [tw] OR Medline [tw] OR CINAHL [tw] OR (National [tw] AND Library [tw])) OR (handsearch* [tw] OR search* [tw] OR searching [tw]) AND (hand [tw] OR manual [tw] OR electronic [tw] OR bibliographi* [tw] OR database* OR (Cochrane [tw] OR Medline [tw] OR CINAHL [tw] OR (National [tw] AND Library [tw])))))) OR ((synthesis [ti] OR overview [ti] OR review [ti] OR survey [ti]) AND (systematic [ti] OR critical [ti] OR methodologic [ti] OR quantitative [ti] OR qualitative [ti] OR literature [ti] OR evidence [ti] OR evidence-based [ti]))) BUT NOT (case* [ti] OR report [ti] OR editorial [pt] OR comment [pt] OR letter [pt])	16

For additional filters in PubMed: *www.ncbi.nlm.nih.gov/entrez/query/static/clinical.html*

TABLE IV. Important quality features of selected study designs

Study design	Questions for ascertaining quality (validity)	References
Therapy (e.g. randomized controlled trial)	1. Were patients randomized? 2. Was concealment of allocation adequate? 3. Were patients analysed in the groups to which they were randomized? 4. Were patients aware of group allocation? 5. Were clinicians aware of group allocation? 6. Were outcome assessors aware of group allocation? 7. Was follow up complete?	2, 17
Diagnosis (e.g. cross-sectional diagnostic study)	1. Was there a comparison with an independent, appropriate gold standard? 2. Did the included patients cover a wide patient spectrum likely to be encountered in a usual clinical practice setting? 3. Was the index test result interpreted without the knowledge of gold standard, and vice-versa? 4. Did the study prospectively recruit consecutive patients suspected to have the disease of interest?	2, 17, 18
Harm (e.g. cohort or case–control study)	1. Did the investigators demonstrate similarity in all known determinants of outcome (e.g. confounders)? Did they adjust for differences in the analysis? 2. Were exposed patients equally likely to be identified in the two groups? 3. Were the outcomes measured in the same way in the groups being compared? 4. Was follow up sufficiently complete?	2, 17, 19
Prognosis (e.g. cohort study)	1. Was the sample of patients representative? 2. Were the patients sufficiently homogeneous with respect to prognostic risk? 3. Was follow up sufficiently complete? 4. Were objective and unbiased outcome criteria used?	2, 17, 19

Source: Adapted mainly from reference 2; for other quality checklists and scales, please see *www.medepi.org/meta*

TABLE V. Software for meta-analysis

Software	Description	Applications
Review Manager (RevMan)	Free Windows-based software from the Cochrane Collaboration. Can be downloaded from: *www.cochrane.org/software/revman.htm*	Primarily designed for Cochrane reviews; can perform meta-analyses of RCTs; graphics options available
Stata	General statistical software; not designed exclusively for meta-analysis. Stata commands for meta-analysis are user-written, add-on programs that can be freely downloaded and added to Stata. Can be purchased via: *www.stata.com*	Powerful and versatile; at least 14 meta-analysis commands are available, and they can perform: meta-analyses, cumulative meta-analyses, forest and funnel plots, publication bias, meta-regression, and sensitivity analyses
SAS	General statistical software package; not designed exclusively for meta-analysis. SAS can be used for meta-analysis by adding special macros created for meta-analysis. Can be purchased via: *www.sas.com*	Can perform a wide range of analyses: meta-analyses, meta-regression, sensitivity analyses, etc.
Comprehensive Meta-analysis	A Windows-based software designed specifically for meta-analysis. Can be purchased via: *www.meta-analysis.com/*	Can perform a wide range of analyses, including forest plots and subgroup analyses
MetaWin	A Windows-based software. Can be purchased via: *www.metawinsoft.com/*	Can perform common routines such as random and fixed effects meta-analyses and forest plots
WeasyMA	A Windows-based software. Can be purchased via: *www.weasyma.com/*	Easy-to-use software for analysis and forest plots, but very expensive
Meta-DiSc	A free Windows-based package, exclusively designed for diagnostic meta-analysis. Can be downloaded via: *http://www.hrc.es/investigacion/metadisc.html*	Can perform diagnostic meta-analyses, summary ROC analyses, meta-regression and generate forest plots

For a comprehensive overview of software, including free DOS-based packages, see reference 20

outcomes are the measures of test performance (e.g. sensitivity and specificity). It is important that reviewers extract raw data from studies where possible (cell values to fill a 2×2 table necessary to compute measures such as RR or OR). If 2×2 table data cannot be obtained, reviewers should extract the effect measure (e.g. OR) along with some measure of its variance (e.g. confidence intervals [CI]). Meta-analysis software packages (Table V) often require variance measures for weighting and pooling effects.

SYNTHESIS AND SUMMARY OF STUDY RESULTS (META-ANALYSIS)

Most reviewers begin analysis with simple tabulation of study characteristics (e.g. year, setting, study design, quality) and results, and this should be done for all systematic reviews, even if no meta-analysis is performed. Forest plots display effect estimates from each study with their CI, and provide a visual summary of the data. The results of each component study are shown as boxes centred on the point estimate, with the horizontal line representing the CI. The pooled estimate is usually displayed at the bottom of the plot as a diamond. Figure 4 shows the forest plot of a meta-analysis on Chinese herbal medicine and interferon therapy compared to interferon alone in the treatment of hepatitis B.[8] In diagnostic reviews, forest plots of sensitivity and specificity can be generated. Figure 5 displays the forest plot for a meta-analysis of polymerase chain reaction (PCR) for tuberculous meningitis.[9]

The next step in the analysis is pooling of effect measures across studies. Pooling is essentially a process of computing weighted averages.[21] In the absence of weighting, all studies are assigned the same weight, irrespective of their sample sizes. An unweighted average, therefore, would be the simple average (e.g. arithmetic mean). In meta-analyses, typically, larger studies (with larger sample sizes and more events) are assigned more weight in the computation of averages. Pooling is accomplished using two

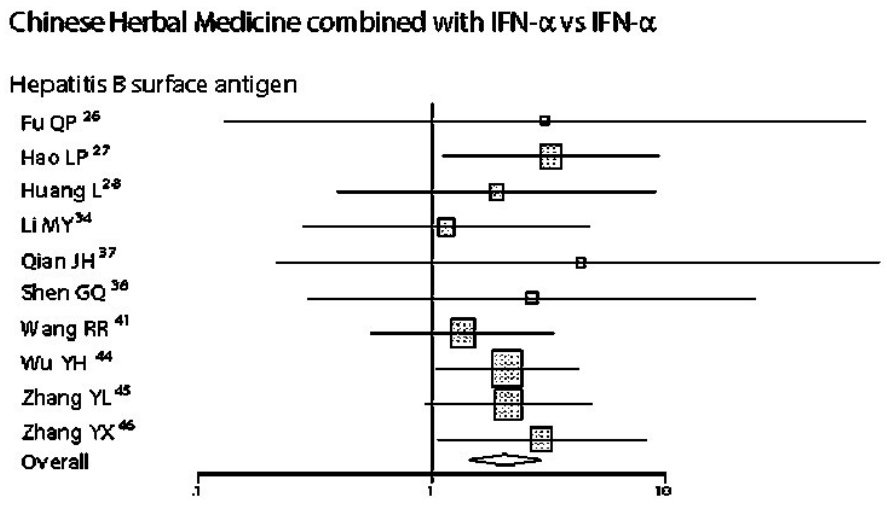


FIG 4. Forest plot of a meta-analysis on efficacy of Chinese herbal medicine for hepatitis B *Source:* Reference 8, reproduced with permission from the American Public Health Association (*Am J Public Health*)

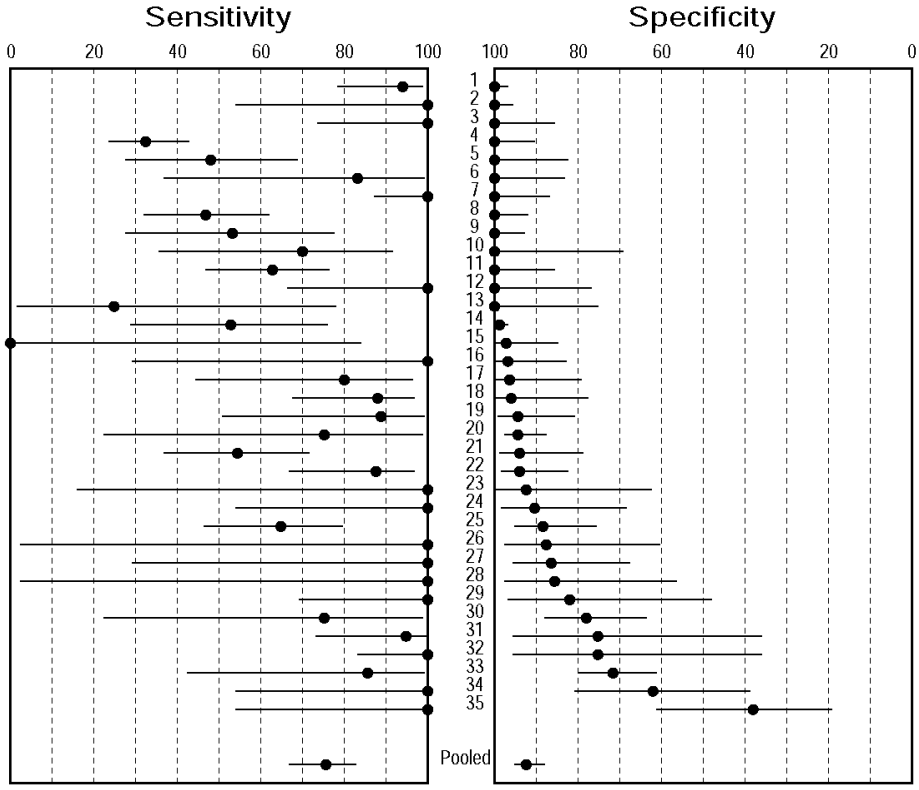


FIG 5. Forest plot of sensitivity and specificity estimates from a meta-analysis on accuracy of polymerase chain reaction tests for tuberculous meningitis *Source:* Adapted from reference 9 and reproduced with permission from Elsevier (*Lancet Infect Dis*)

statistical models: the random effects model or the fixed effects model.[21] Both models can be used to pool a variety of effect measures (discrete and continuous): OR, RR, risk differences, p values, differences in means, sensitivity, specificity, etc. Examples of fixed effects models are: Mantel–Haenszel, Peto and Inverse Variance methods.[4] The most popular random effects model is the DerSimonian–Laird model.[4]

The fixed effects model assumes that the studies included in the meta-analysis estimate the same underlying 'true' effect that is 'fixed', and that the observed differences across studies are due to random error (chance).[4,21] On the other hand, the random effects model assumes that the studies included in the meta-analysis are only a random sample of a theoretical universe of all possible studies on a given research question, and that the effects for the individual studies vary around some overall average effect.[4,21] Random effects models incorporate two sources of variability: random error and between-study variability. Therefore, the random effects model is preferred when the data are heterogeneous, since it allows for between-study and within-study variability, and provides a more conservative estimate with a wider CI.[4,6,21] In the absence of heterogeneity, both models produce similar results. Several software packages (Table V) can perform both fixed and random effects meta-analyses.

Cumulative meta-analysis can be performed to evaluate how summary estimates change over a time period.[21] In a cumulative meta-analysis, the summary estimate is calculated repeatedly through meta-analysis as if it had been done each time a new study had been reported. At each calculation, the meta-analysis summary estimate to that point in time is shown. Such a cumulative meta-analysis can retrospectively identify the point in time when a treatment effect first reached statistical significance (e.g. $p<0.05$). Figure 6 displays the cumulative meta-analysis plot for trials of beta-blockers after acute myocardial infarction.[22] The plot shows that a significant protective effect of beta-blockers was achieved by the early 1980s, many years and many trials before its general adoption in clinical practice.[22] Thus, cumulative meta-analyses have the potential to provide information that could reduce the need for further large and expensive trials.

Heterogeneity refers to a high degree of variability in results across studies and is not uncommon in meta-analyses.[23] For example, consider a meta-analysis on oral zinc for common cold.[24] The authors reported a summary OR for the incidence of 'any' cold symptom at 1 week: 0.52 (95% CI 0.25, 1.2), indicating a 50% risk reduction. However, the forest plot (Fig. 7) displays a great degree of variability in the effect of zinc; some studies show protection, while others suggest harm. This heterogeneity raises concerns about the interpretation of the summary measure. Heterogeneity in diagnostic reviews may be manifest as widely varying estimates of sensitivity and specificity. For example, Fig. 5 shows sensitivity estimates ranging from 0% to 100%.[9] Reviewers, therefore, should routinely test for heterogeneity and common approaches include the use of χ^2 and I^2 tests.[25] Most software packages routinely generate heterogeneity test values along with summary estimates.

In the presence of significant heterogeneity, the pooled, summary estimate is not meaningful, since it is an average of extreme values and does not adequately describe the data.[23] In fact, reviewers may choose not to force the results into a single summary estimate. In the presence of heterogeneity, reviewers should focus instead on finding potential sources of variability in effect estimates.[23] This may be accomplished by methods such as subgroup analyses, meta-regression and graphical methods.[23] Figure 8 illus-

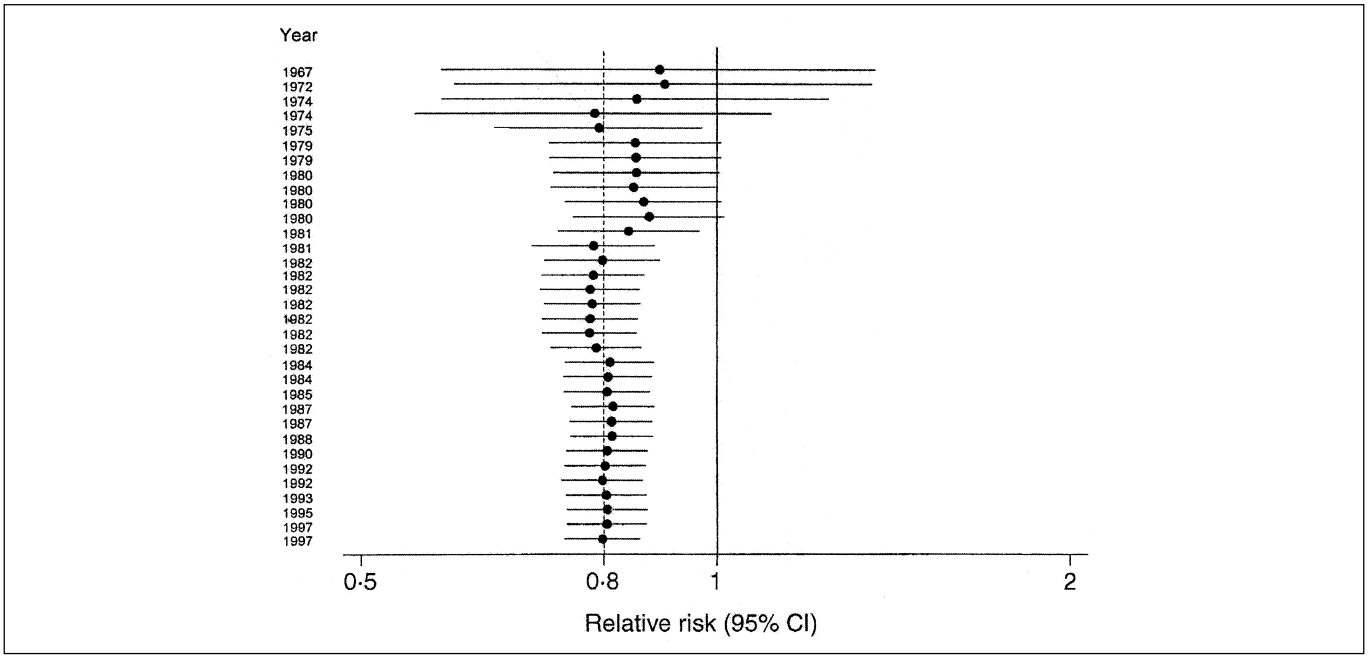


FIG 6. Cumulative meta-analysis of trials on beta-blockers after acute myocardial infarction *Source:* Reference 22, reproduced with permission from BMJ Books

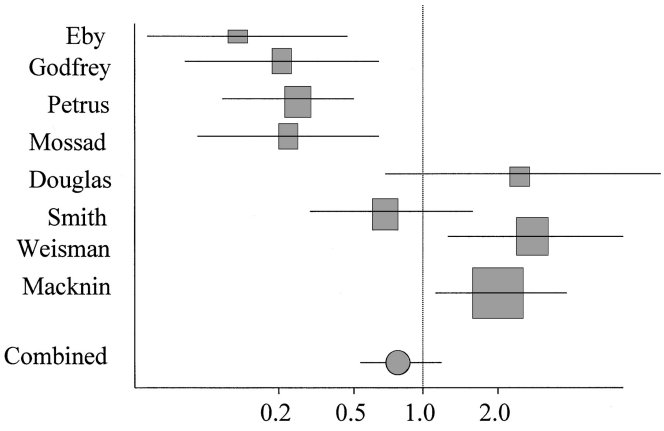


FIG 7. Meta-analysis of randomized controlled trials on oral zinc for common cold: Example of heterogeneity *Source:* Reference 24, reproduced with permission from the American Society for Nutritional Sciences (*J Nutr*)

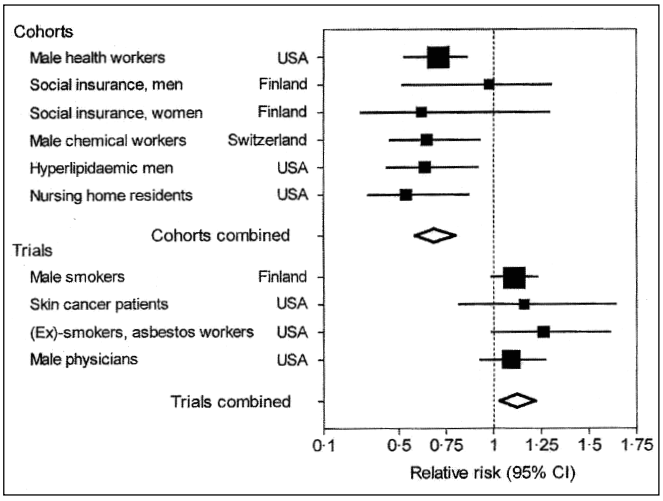


FIG 8. Meta-analysis on beta-carotene intake and cardiovascular mortality: example of subgroup analysis and exploration of heterogeneity *Source:* Reference 26, reproduced with permission from BMJ Books

trates the use of graphical methods and subgroup analysis. In a meta-analysis on beta-carotene intake and cardiovascular mortality, for example, observational studies showed considerable benefit, whereas RCTs showed harm.[26] Given this heterogeneity, it would be inappropriate to combine the effects from observational and experimental studies. This plot illustrates an approach to evaluating the impact of study quality on results. Since well-done RCTs are considered to be stronger designs for causal inference (as compared to observational studies), this analysis is stratified by study design, a surrogate for study quality.

Another critical element of a well-conducted meta-analysis is the evaluation of publication bias.[27] Publication bias is just one type of a family of biases called 'reporting bias'. Reporting biases tend to occur when statistically significant ('positive') studies are more likely to be submitted and accepted for publication (publication bias), more likely to be published in English (language bias), more likely to be published rapidly (time-lag bias) and cited more often (citation bias).[6,27] Also, studies that are easily accessible as electronic, full-text reports may be identified more often than those that are not. If a meta-analysis summarizes only published studies prone to these biases, the overall summary effect might be spuriously exaggerated.[27] Since it is very hard to identify unpublished studies, there is no easy method to overcome this problem. Reviewers can check for the presence of publication bias using graphical methods (e.g. funnel plots), and statistical tests (e.g. Egger test).[27] Figure 9 illustrates the use of funnel plots in the evaluation of publication bias in a meta-analyses on PCR for the diagnosis of tuberculous pleuritis.[28] The funnel graph plots the log of the diagnostic OR (DOR; a measure of diagnostic accuracy)

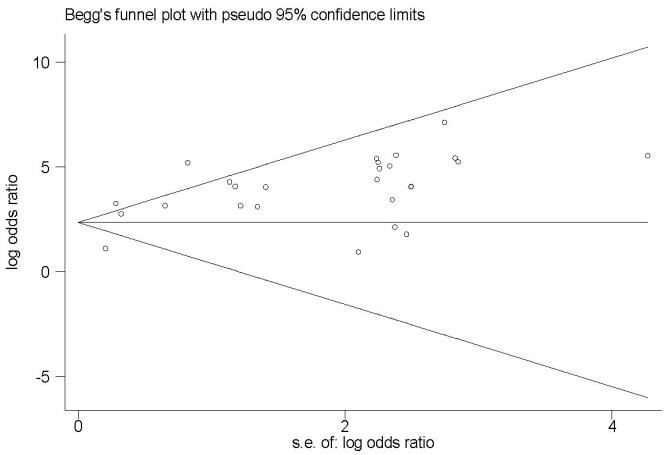


FIG 9. Funnel plot for evaluation of publication bias in a meta-analysis on polymerase chain reaction for the diagnosis of tuberculous pleuritis *Source:* Reference 28, © 2004 Pai *et al*. licensee BioMed Central Ltd.

against the standard error of the log of the DOR (an indicator of sample size). Each open circle represents each study in the meta-analysis. The line in the centre indicates the summary DOR. In the absence of publication bias, the DOR estimates from smaller studies are expected to be scattered above and below the summary estimate, producing a triangular or funnel shape.[27] The funnel plot appears asymmetric—smaller studies with low DOR estimates (poor diagnostic accuracy) are missing—indicating a potential for publication bias. The Egger test for publication bias was statistically significant in this analysis.

INTERPRETATION OF THE RESULTS

The last step is interpretation of the results, discussion of issues such as clinical applicability and writing of the manuscript for publication. Reviewers need to discuss the limitations of the primary studies included in their review, and limitations in how the review itself was conducted.[6] Limitations of the primary studies, for example, may include issues relating to design flaws. Limitations of the review itself may include issues such as inclusion of only English language studies or inability to accurately interpret the summary estimates due to heterogeneity. A discussion of these limitations will enable readers to judge the strength of the evidence presented in the review. The review usually concludes with a discussion on the implications for clinical practice, and need for further research. If the evidence is strong and unequivocal, reviewers might recommend no further trials on that clinical question.[6] Some reviews (e.g. reviews on screening tests such as mammography) may have important public health or policy implications that merit discussion.

For writing the manuscript for publication, reviewers have two useful guides: the QUOROM guidelines[29] for meta-analyses of controlled trials, and the MOOSE guidelines[30] for meta-analyses of observational studies. Many journals now encourage authors to submit manuscripts formatted according to these guidelines. Moreover, these guidelines can serve as practical tools for the critical reader in assessing the quality of an individual meta-analysis. In addition to these guidelines, reviewers can find a variety of outstanding resources for conducting reviews on the internet (Table VI).

CONCLUSION

Systematic reviews of high-quality studies are considered to represent the pinnacle of evidence. However, to trust the evidence presented in a systematic review, it is imperative that the review is a comprehensive assessment of the existing literature and that the final interpretation incorporates information regarding features of the individual studies (e.g. quality) and the review process (e.g. publication bias). Due to the increasing dependence of clinicians upon reviews to identify and amass relevant information quickly, the ability to assess the quality of evidence is critical. In this paper, we discussed the design and conduct of systematic reviews. A clear understanding of how to conduct systematic reviews will enable clinicians to critically appraise and use such evidence in practice. We also hope that it will encourage clinicians to conduct systematic reviews and contribute to evidence-based clinical practice in their areas of expertise.

TABLE VI. Internet resources for systematic reviews

Name	Description	URL
Berkeley Systematic Reviews Group	Website with several useful guidelines, checklists, data forms, and software for conducting reviews	*www.medepi.org/meta*
Cochrane Collaboration	Prepares, maintains and promotes the accessibility of systematic reviews of the effects of healthcare interventions	*http://www.cochrane.org*
Cochrane Library	Contains: Cochrane Database of Systematic Reviews, the Cochrane Controlled Trials Register and other databases	*http://www.update-software.com/clibng/cliblogon.htm*
Centre for Reviews & Dissemination (CRD)	The CRD offers rigorous and systematic reviews on selected topics, a database of high-quality reviews and useful resources on how to conduct reviews	*http://www.york.ac.uk/inst/crd*
CONSORT	CONSORT comprises a checklist and flow diagram to help improve the quality of reports of RCTs. The website also contains QUOROM, MOOSE and STARD guidelines	*http://www.consort-statement.org*
Users' Guides to the Medical Literature	Book/CD versions of the popular *Users' Guides* series—provide the most detailed exposition of the concepts necessary to critically appraise the medical literature	*http://www.usersguides.org*
Centre for Evidence Based Medicine	Oxford Centre for Evidence Based Medicine, aims to promote EBM and provide training resources	*http://www.cebm.net*

ACKNOWLEDGEMENTS

MP and NP receive training support from the National Institutes of Health, Fogarty AIDS International Training Program (1-D43-TW00003-15). None of the authors have any conflicts of interest with regard to this publication. The views expressed in this article do not necessarily state or reflect those of the Cochrane Collaboration.

REFERENCES

1 Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: What it is and what it isn't. *BMJ* 1996;**312:**71–2.
2 Guyatt GH, Rennie D (eds). *Users' guides to the medical literature. A manual for evidence-based clinical practice.* Chicago:AMA Press; 2002.
3 McAlister FA, Clark HD, van Walraven C, Straus SE, Lawson FM, Moher D, *et al.* The medical review article revisited: Has the science improved? *Ann Intern Med* 1999;**131:**947–51.
4 Egger M, Smith GD, Altman DG (eds). *Systematic reviews in health care. Meta-analysis in context.* London:BMJ Publishing Group; 2001:347–69.
5 Glasziou P, Irwig L, Bain C, Colditz G. *Systematic reviews in health care. A practical guide.* Cambridge:Cambridge University Press, 2001.
6 Clarke M, Oxman AD (eds). *Cochrane reviewers' handbook 4.2.0* [updated March 2003]. In: *The Cochrane Library*, Issue 2, 2003. Oxford:Update Software. Available at: *http://www.cochrane.dk/cochrane/handbook/hbook.htm*
7 CRD Centre for Reviews and Dissemination, University of York, York, UK. *Undertaking systematic reviews of research on effectiveness.* CRD's guidance for carrying out or commissioning reviews. CRD Report Number 4 (2nd), March 2001. Available at: *http://www.york.ac.uk/inst/crd/report4.htm*
8 McCulloch M, Broffman M, Gao J, Colford JM Jr. Chinese herbal medicine and interferon in the treatment of chronic hepatitis B: A meta-analysis of randomized, controlled trials. *Am J Public Health* 2002;**92:**1619–28.
9 Pai M, Flores LL, Pai N, Hubbard A, Riley LW, Colford JM. Diagnostic accuracy of nucleic acid amplification tests for tuberculous meningitis: A systematic review and meta-analysis. *Lancet Infect Dis* 2003;**3:**633–43.
10 Pai M, McCulloch M, Enanoria W, Colford JM. Systematic reviews of diagnostic test evaluations: What's behind the scenes? *Evid Based Med* & *ACP Journal Club* 2004 (in press).
11 Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: A key to evidence-based decisions. *ACP J Club* 1995;**123** (3):A12–A13.
12 Centre for Reviews and Dissemination, University of York. *Finding studies for systematic reviews: A basic checklist for researchers.* University of York, 2002. Available at: *http://www.york.ac.uk/inst/crd/revs.htm*
13 Suarez-Almazor ME, Belseck E, Homik J, Dorgan M, Ramos-Remus C. Identifying clinical trials in the medical literature with electronic databases: MEDLINE alone is not enough. *Control Clin Trials* 2000;**21:**476–87.
14 Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *J Am Med Inform Assoc* 1994;**1:**447–58.
15 Robinson KA, Dickersin K. Development of a highly sensitive search strategy for the retrieval of reports of controlled trials using PubMed. *Int J Epidemiol* 2002;**31:** 150–3.
16 Shojania KG, Bero LA. Taking advantage of the explosion of systematic reviews: An efficient MEDLINE search strategy. *Eff Clin Pract* 2001;**4:**157–62.
17 Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 2001;**323:**42–6.
18 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.
19 Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, *et al*. *The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses.* Department of Epidemiology and Community Medicine, University of Ottawa, Canada. Available at: *http://www.lri.ca/programs/ceu/oxford.htm*
20 Sterne JAC, Egger M, Sutton AJ. Meta-analysis software. In: Egger M, Smith GD, Altman DG (eds). *Systematic reviews in health care. Meta-analysis in context.* London:BMJ Publishing Group; 2001:336–46.
21 Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med* 1997;**127:**820–6.
22 Egger M, Smith GD, O'Rourke K. Rationale, potentials, and promise of systematic reviews. In: Egger M, Smith GD, Altman DG (eds). *Systematic reviews in health care. Meta-analysis in context.* London:BMJ Publishing Group; 2001:3–19.
23 Glasziou PP, Sanders SL. Investigating causes of heterogeneity in systematic reviews. *Stat Med* 2002;**21:**1503–11.
24 Jackson JL, Lesho E, Peterson C. Zinc and the common cold: A meta-analysis revisited. *J Nutr* 2000;**130** (5S Suppl):1512S–1515S.
25 Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;**327:**557–60.
26 Egger M, Smith GD, Schneider M. Systematic reviews of observational studies. In: Egger M, Smith GD, Altman DG (eds). *Systematic reviews in health care. Meta-analysis in context.* London:BMJ Publishing Group; 2001:211–27.
27 Sterne JAC, Egger M, Smith GD. Investigating and dealing with publication and other biases. In: Egger M, Smith GD, Altman DG (eds). *Systematic reviews in health care. Meta-analysis in context.* London:BMJ Publishing Group; 2001:189–208.
28 Pai M, Flores LL, Hubbard A, Riley LW, Colford JM. Nucleic acid amplification tests in the diagnosis of tuberculous pleuritis: A systematic review and meta-analysis. *BMC Infect Dis* 2004;**4:**6.
29 Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF, *et al*. Improving the quality of reports of meta-analyses of randomized controlled trials: The QUOROM statement. Quality of reporting or meta-analyses. *Lancet* 1999;**354:**1896–900.
30 Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, *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.