Evidence-Based Medicine

Evidence-Based Medicine, Systematic Reviews, and Guidelines in Interventional Pain Management: Part 7: Systematic Reviews and Meta-Analyses of Diagnostic Accuracy Studies

Laxmaiah Manchikanti, MD[1], Richard Derby, MD[2], Lee Wolfer, MD[2], Vijay Singh, MD[3], Sukdeb Datta, MD[4], and Joshua A. Hirsch, MD[5]

From: [1]Pain Management Center of Paducah, Paducah, KY; [2]Spinal Diagnostics & Treatment Center, Daly City, CA; [3]Pain Diagnostics Associates, Niagara, WI; [4]Vanderbilt University Medical Center, Nashville, TN; and [5]Massachusetts General Hospital and Harvard Medical School, Boston, MA.

Dr. Manchikanti is Medical Director of the Pain Management Center of Paducah, Paducah, KY and Associate Clinical Professor of Anesthesiology and Perioperative Medicine, University of Louisville, Louisville, KY. Dr. Derby is Medical Director of Spinal Diagnostics & Treatment Center, Daly City, CA, and Associate Professor, Department of Physical Medicine and Rehabilitation, Stanford University, Stanford, CA. Dr. Wolfer is with the Spinal Diagnostics & Treatment Center, Daly City, CA. Dr. Singh is Medical Director of Pain Diagnostics Associates, Niagara, WI. Dr. Datta is Director, Vanderbilt University Interventional Pain Program, Associate Professor, Dept. of Anesthesiology, Vanderbilt University Medical Center, Nashville TN. Dr. Hirsch is Chief of Minimally Invasive Spine Surgery, Depts. of Radiology and Neurosurgery, Massachusetts General Hospital and Associate Professor of Radiology, Harvard Medical School, Boston, MA.

Address Correspondence:
Laxmaiah Manchikanti, M.D.
2831 Lone Oak Road
Paducah, Kentucky 42003
E-mail: drlm@thepainmd.co

Disclaimer: There was no external funding in preparation of this manuscript.
Conflict of Interest: Dr. Datta receives research support from Sucampo Pharmaceuticals and an honorarium from Smith and Nephew. Dr. Hirsch is a consultant for Cardinal Healthcare. He is a minor shareholder in Medtronic and Cardinal Healthcare. He serves on the Steering Committee for KAVIAR trial (volunteer position) and on the Data and Safety Monitoring Board (DSMB): CEEP trial (volunteer position).

Manuscript received: 10/19/2009
Accepted for publication: 10/30/2009

Free full manuscript:
www.painphysicianjournal.com

Appropriate diagnosis is essential in providing proper and effective therapy. The field of diagnostic accuracy tests is dynamic with new tests being developed at a fast pace along with improvement in technology of existing tests on a continuous basis. Well-designed diagnostic test accuracy studies can help in making appropriate health care decisions, provided that they transparently and fully report their participants, tests, methods, and results. Exaggerated and biased results from poorly designed and reported diagnostic test studies can trigger their premature dissemination and lead physicians into making incorrect treatment decisions. Consequently, a diagnostic test is useful only to the extent that it distinguishes between conditions or disorders that might otherwise be confused. Since it is unlikely that clinicians, patients, and policy makers have the time, skills, and resources to find, appraise, and interpret the evidence and incorporate it into their health care decisions, systematic reviews and meta-analysis provide an accurate and reliable synthesis of vast quantities of data.

A systematic review can identify what is known and what is unknown, giving guidance for future research. Systematic reviews have been considered as a vital link in the great chain of evidence that stretches from the laboratory to the bedside by helping to separate the insignificant, unsound, or redundant deadwood from the salient and critical studies that are worthy of reflection. A dangerous discrepancy exists between experts and evidence with all types of evidence.

Historically, it has been reported that in only 15% of all cases can a pathoanatomical explanation be found for patients with chronic low back pain of more than 3 months resulting in the assumption that very little can be done in our present state of ignorance to treat these patients and improve their natural histories. On the other end of the spectrum, due to lack of sound diagnostic information, excessive health care is utilized with exploding costs. The validity of all diagnostic techniques has been described with variable accuracy and reliability. Lack of understanding of reference standards and their unavailability with interventional diagnostic techniques and misinterpretation secondary to interpretation bias may adversely influence the applicability of diagnostic interventions.

This manuscript provides a review of the literature, a checklist, and a flow diagram describing the preferred way to present the abstract, introduction, methods, results, and discussion sections of the report of an analysis in a systematic review of diagnostic accuracy studies.

Key words: Diagnostic accuracy studies, evidence-based medicine, systematic reviews, meta-analysis, comparative effectiveness studies, interventional pain management, Standards for the Reporting of Diagnostic Accuracy Studies (STARD)

Pain Physician 2009; 12:929-963

Health care providers, consumers, researchers, and policy makers are inundated with unmanageable amounts of information, including evidence from health care research. It is unlikely that all will have the time, skills, and resources to find, appraise, and interpret this evidence and to incorporate it into health care decisions. Thus, a systematic review attempts to collate all empirical evidence that fits pre-specified eligibility criteria in order to answer a specific research question. It uses explicit, systematic methods that are selected with a view to minimizing bias, thus providing more reliable findings from which conclusions can be drawn and decisions made. The key characteristics of a systematic review include a clearly stated set of objectives with pre-defined eligibility criteria for studies; an explicit, reproductive methodology; a systematic search that attempts to identify all studies that would meet the eligibility criteria, as assessment of the validity of the findings of the included studies, for example through the assessment of risk of bias; and a systematic presentation and synthesis of the characteristics and findings of the included studies (1).

Evidence-based medicine (EBM) and comparative effectiveness research (CER) have taken a center stage in the United States. Researchers, policy-makers, insurers, and other stakeholders have voiced enthusiasm about the value of CER that rigorously evaluates 2 or more drugs or devices. The most recent boost for these efforts has been a U.S. congressional financial stimulus package that contains provisions for $1.1 billion to be devoted to this effort (2). While the appeal of CER is undeniable, what works and what does not work is arguable. Thus, EBM and CER have been described as analogous to religion and politics — meaning different things to different people. Over the past decade, 2 major movements have emerged in medicine, both intended to improve patient care. However, the CER may be in conflict with the medical humanism movement, which also seeks to improve patient care (3). While many aspects of EBM, including randomized trials, systematic reviews, meta-analyses, and clinical guidelines indicate signs of progress in the effort to keep pace with health care innovations, the medical profession continues to struggle with conflicts of humanism and evidence-based practice. EBM as a principle is considered to be a shift in medical paradigm, which acknowledges that intuition, unsystematic clinical experience, and pathophysiologic rationale are insufficient grounds for clinical decision-making (4-11).

Systematic reviews and meta-analyses of randomized trials are very common. However, systematic reviews and meta-analyses of diagnostic test accuracies while lagging behind randomized trials, are also becoming increasingly common (12-25). Further, multiple systematic reviews of diagnostic accuracy studies have been published in interventional pain management (26-41).

In contrast to systematic reviews and meta-analysis, a comparative effectiveness review is a unique type of systematic review, which synthesizes the available scientific evidence on a specific topic and expands the scope of the typical systematic review, which focuses on the comparison of the relative benefits and harms among a range of available treatments or interventions for a given condition (42). Consequently, CERs provide practical information for clinicians, patients, and policy makers.

1.0 An Introduction to Systematic Reviews, Meta-analyses, and Comparative Effectiveness Research

The history of systematic reviews has been described (4,7,10,11). The philosophical history of systematic reviews dates back to 1747 (43) with a description of early systematic review methods by social scientists during 1960s and 1970s (44). The terminology of systematic reviews was coined long before the terminology of EBM (45). The terminology of meta-analysis and systematic reviews is variable (4,7,10,11). Meta-analysis was described in 1904 (46), while CER has only been described in recent years (42).

A systematic review utilizes explicit methodology of clearly formulated questions and methods to identify, select, and critically appraise relevant research and then collect and analyze the data from the studies that are included in the review (1,4,7,10); whereas a meta-analysis incorporates the statistical pooling of data across studies to generate a summary in the form of a pool of estimated effects (14,47). In addition, a meta-analysis has been described as the final step in a systematic review, which ideally starts with an unbiased systematic review that incorporates articles chosen using predetermined selection or inclusion criteria. Thus, both systematic reviews and meta-analysis, despite their differences, share many similarities and provide a continuum of synthesis of an unmanageable and exponentially increasing body of literature with identification of beneficial or harmful interventions (47,48).

A comparative effectiveness review is a unique type of systematic review which synthesizes the available scientific evidence on a specific topic. CERs expand the scope of a typical systematic review (which focuses on the effectiveness of a single intervention) by comparing the relative benefits and harms among a range of available treatments or interventions for a given condition. Consequently, it is stated that in doing so, CERs more closely parallel the decisions facing clinicians, patients, and policy makers, who must chose among a variety of alternatives in making diagnostic, treatment, and health care delivery decisions (42). Further, in choosing topics for CERs, a number of criteria are considered including the burden of illness; evidence suggesting underuse or overuse; the cost of the intervention or of not treating the illness; controversy surrounding the treatment; and interventions intended to treat conditions that disproportionately affected women, traditionally underserved minorities, the elderly, and children. Prior to the establishment of CER in the United States, the Effective Health Care (EHC) Program research, originating from the Medicare Prescription Drug, Improvement, and Modernization Act (MMA) of 2003 and the Agency for Healthcare Research and Quality (AHRQ) conducted research on multiple topics. The first 14 CERs were conducted from 2005 through 2007 of which 13 were of therapy and one was of diagnosis. The comparative effectiveness study conducted for diagnostic purposes was effectiveness of non-invasive diagnostic tests for breast abnormalities (42).

Systematic reviews, meta-analyses, and CERs are labor intensive and require expertise in both the subject matter and review methodology. Further, they must follow the rules of EBM which suggests that a formal set of rules complement medical training and common sense for clinicians to interpret the results of clinical research effectively. Consequently, knowing the tools of evidence-based practice is necessary, but not sufficient for delivering the highest quality of patient care. However, expertise in a single area is not enough and may lead to inaccurate conclusions, which leads to inappropriate application of the results. A systematic review and meta-analysis must truly incorporate the definition which states that scientific strategies must be applied to limit bias by the systematic assembly, critical appraisal, and synthesis of relevant studies on a specific topic (49-54).

Systematic reviews and meta-analysis have been performed frequently for randomized trials (55-61). Similarly, systematic reviews may be performed for observational studies and diagnostic accuracy studies (4-41,62-83). Well-designed diagnostic test accuracy studies can help in making appropriate decisions if the testing improves outcome and identifies what test to use, purchase, or recommend in practice guidelines and how to interpret the results for clinicians, policy makers, and patients, provided that they transparently and fully report their participants, tests, methods, and results as facilitated, for example by the STARD (Standards for Reporting of Diagnostic Accuracy) statement (84,85). The STARD is a 25-item checklist which was published in many journals and is now adopted by more than 200 scientific journals worldwide. Thus, as in other areas, systematic reviews and meta-analyses of accuracy studies can be used to obtain more precise estimates when small studies addressing the same test and patients in the same setting are available. Further, systematic reviews can also be useful to establish whether and how scientific findings vary by particular subgroups, and may provide some re-estimates with a stronger generalizability than estimates from a single study (12). Further, systematic reviews may assist in identifying the risk for bias that may be present in the original studies and can be used to address questions that were not directly considered in the primary studies. The Cochrane Collaboration in 2003 decided to make preparations for including systematic reviews of diagnostic test accuracy in their Cochrane database of systematic reviews. Leeflang et al (12) reviewed methodological developments concerning problem formulation, location of literature, quality assessment, and meta-analysis of diagnostic accuracy studies. Devillé et al (13) described didactic guidelines for conducting systematic reviews of diagnostic studies.

2.0 Why Quality Systematic Reviews for Diagnostic Accuracy Studies?

A systematic review can identify what is known and what is unknown, giving guidance for further research. It has been stated "without a clear picture of where things stand now, simply adding one new study to the existing morass is unlikely to be very useful — for science to be cumulative, an intermediate step between the past and future research is necessary; synthesis of existing evidence" (86). Mulrow (87) described that systematic review is a search for the whole truth rather than just one part of it, and is thus, a fundamentally scientific activity (88). Mulrow (87) also has emphasized that systematic reviews are essential to efficiently integrate valid information and provide a basis for rational decision-making (89). Systematic reviews have been considered a vital link in the great chain of evidence that stretches from the laboratory bench to the bedside (90). Essentially, systematic reviews help separate the insignificant, unsound, or redundant deadwood from the salient and critical studies that are worthy of reflection (91). Systematic reviews also facilitate an important function of identifying the studies with weak designs, which tend to be frequently biased and misleading, often overestimating the benefits of the treatment being studied (92-96). Thus, systematic reviews serve multiple functions (97): reducting a large amount of information to a manageable size; helping to determine whether the results are consistent from study to study and to generalize the results; reducing the delay between publication of research findings and the implementation of new effective diagnostic strategies; combining information from individual studies so that its overall sample size is greater than that of any one study which leads to an increase in the power of the investigation; limiting bias and improving the reliability and accuracy of recommendations because of formalized and thorough methods of investigation; and finally a systematic review is less expensive and quicker to conduct than to embark on a new study.

Smidt et al (98), in a 2005 publication, evaluated the quality of reporting of 124 diagnostic accuracy studies published in 2000, prior to the publication of STARD. They concluded that the quality of reporting in diagnostic accuracy was less than optimal. They (99) also carried out a reproducibility study to assess the STARD checklist and to investigate the inter-assessment reproducibility of evaluating the quality of reporting of diagnostic accuracy studies published in 2000, using the items of STARD. They included 22 diagnostic accuracy studies published in 2000. The overall inter-assessment agreement for all items of the STARD statement was 85% and varied from 63% to 100% for individual items. The largest difference between the 2 assessments were found for the reporting of the rationale of the reference standard (kappa 0.37), number of included participants that underwent tests (kappa 0.25), distribution of the severity of the disease (kappa 0.33), a cross-tabulation of the results of the index test by the results of the reference standard (kappa 0.33), and how indeterminate results, missing data, and outliers were handled (kappa 0.25) (99). Within and between the reviewers, large differences were also observed for these items. They concluded that even though the overall reproducibility of the quality of reporting on diagnostic accuracy studies using the STARD statement was found to be good, substantial disagreements were found for specific items.

Among the 26 reviews on diagnostic tests published between 1996 and 1997, 19 were systematic reviews or meta-analyses (100). Even though multiple guidelines for critical appraisal of diagnostic research and meta-analysis have been published, these may be difficult to understand for clinical researchers or do not provide sufficient information (13-15,101-106).

There are several potential threats to the internal and external validity of a study of diagnostic accuracy. A survey of studies of diagnostic accuracy published in 4 medical journals between 1978 and 1993 revealed that the methodologic quality was mediocre at best. Further, the absence of critical information about the design and conduct of diagnostic studies has been confirmed in multiple meta-analyses (107,108). Diagnostic studies with specific design features may be associated with bias and optimistic estimates of diagnostic accuracy compared to studies without such deficiencies (100). In an evaluation of the assessment of neck pain and its associated disorders, it was shown that there was little information on the validity or utility of the self-reported history (109). Consequently, the study design can affect estimates of diagnostic accuracy (17-41,110-116). This can result in bias or variation (117). A recent manuscript evaluating systematic reviews of diagnostic test accuracy showed that the challenges that remained are the poor reporting of original diagnostic test accuracy studies and differences with the interpretation of the results of diagnostic test accuracy (12). In addition, a literature survey of sample sizes of studies on diagnostic accuracy concluded that few studies on diagnostic accuracy report considerations of sample size (118). The number of participants in most studies on diagnostic accuracy is probably too small to analyze the variability of the measure of accuracy across patient subgroups.

In a CER evaluation of the effectiveness of non-invasive tests for breast abnormalities from 2005 to 2008, including Positron Emission Tomography (PET) scans, scintimammograms, magnetic resonance imaging (MRIs), and ultrasound was conducted (42). They concluded that even though the technologies evaluated could reduce the need for biopsy in women with an abnormal mammogram who do not have cancer, each would miss some cancers. Further, they added that no literature will change this conclusion since there is no test with 100% sensitivity and they concluded that the previous conclusion was still valid.

3.0 Dangerous Discrepancies Between Experts and Evidence

Similar to randomized trials and systematic reviews and meta-analysis of randomized trials, a dangerous discrepancy exists between experts and evidence for diagnostic accuracy studies in interventional pain management (7). Precise anatomical diagnosis in low back pain has been described as elusive and the diagnostic evaluation is often frustrating for both physicians and patients (9,119-134). History, physical examination, and imaging provide limited information. Consequently, it has been stated that in low back pain, the diagnosis can be provided with certainty in only approximately 15% of the cases in patients without disc herniation or radiculitis. Precision diagnostic blocks have changed this substantially. Nachemson (129) reported that in only 15% of cases could a pathoanatomical explanation be found for patients with chronic low back pain of more than 3 months and he stated: "very little can be done at our present state of ignorance to treat these patients and improve their natural histories." Thus, when a source of pain is not obvious, diagnosis often depends on who makes the diagnosis and sets the reference standards by which the diagnosis is proven. Utilizing controlled diagnostic blocks, facet joint pain has been demonstrated in 36% to 67% in the cervical spine, 34% to 48% in the thoracic spine, and 16% to 40% in the lumbar spine (26-30,36,41,135-140); discogenic pain has been demonstrated in 26% to 39% of patients in the lumbar spine (34-36,67,135,141,142); and sacroiliac joint pain has been established in 10% to 26% of patients (31-33,135,143). Rubinstein and van Tulder (115) evaluated the scientific evidence for diagnostic procedures for neck and low back pain and concluded that there was strong evidence for facet joint nerve blocks in the diagnosis of spinal pain, whereas the evidence was moderate for sacroiliac joint injections in the diagnosis of sacroiliac joint pain. The validity of diagnostic interventional techniques with variable accuracy and reliability has been described in multiple studies yielding mixed results (26-40,67-79,81-83,120,144).

Even though extensive criticism has been focused against diagnostic interventional techniques and their validity (145-154), the literature is replete with the lack of validity of multiple diagnostic tests in spinal pain (115,116,155-157). Rubinstein and van Tulder (115) commented that it was quite remarkable that while many named orthopedic tests of the neck and low back are often illustrated in orthopedic text books, there is little evidence to support their diagnostic accuracy, and therefore their use in clinical practice. Consistent with clinical experience, many studies have demonstrated that the physical examination serves primarily to confirm suspicions that arose during the history. Further, they illustrated that individual red flags do not necessarily mean the presence of serious pathology. In fact, red flags have not been evaluated comprehensively in any systematic review. Even then, the incidence of spinal tumors is extremely low. In a systematic review (155) of the accuracy of diagnostic tests of lumbar spinal stenosis, it was concluded that the overall quality was poor; with only 5 studies scoring positive on more than 50% of the quality items — only 20% of the included studies. They also concluded that because of the heterogeneity and overall poor quality, no firm conclusions about the diagnostic performance of the differences can be drawn. In another study (156), strong conclusions were not permitted about the relative diagnostic accuracies of computed tomography (CT) and MRI, for the diagnosis of lumbar spinal stenosis due to a lack of methodological rigor. Similarly, a systematic review of diagnostic accuracy of the straight leg raising test in herniated disc (157) found a pooled sensitivity for the straight leg raising test of 0.91 with a pooled specificity of 0.26. Further, it was illustrated that discriminative power was lower in recent studies, with only the inclusion of primary disc herniation, and with blind assessment of both the index test, straight leg raising test, and the reference standard surgery. However, the cross straight leg raising test had a sensitivity of 0.9 with pooled specificity of 0.88. Finally, they concluded that the diagnostic accuracy of the straight leg raising test is limited by its low specificity. Even then, the reliability on these diagnostic tests is enormous in the practice of medicine and guidelines.

In contrast to tests with low accuracy, diagnostic interventional techniques have been shown to present with significant evidence of accuracy. Diagnostic spinal interventional techniques have presented concept validity, content validity, face validity, and construct validity. Multiple studies also have evaluated the false-positive rates of the diagnostic interventional techniques. Construct validity has been established by a controlled disc with discography and controlled comparative local anesthetic blocks for other interventions (26-30,36,41,67,135-144,158,159). Controlled comparative local anesthetic blocks in the lumbar spine have been validated with long-term follow-up (160,161). Multiple confounding factors such as psychological status, sedation, age, obesity, mode of injury, and smoking also have been evaluated (162-171). In addition, facet joint interventions have been proven to be significantly more effective with appropriate diagnosis (172-177). Further, the recent literature also has shown that sodium chloride may not be utilized as placebo, due to its effect on electrophysiology (178-182). Thus, considering the effect of local anesthetic or sodium chloride solution as placebo, leads to inaccurate and invalid conclusions (172-176,183-187).

In summary, there is a wide gap in the understanding and presentation of diagnostic accuracy studies based on the interest of the evaluator and conflicting research results.

4.0 Methodologic Quality Assessment of Systematic Reviews of Diagnostic Accuracy

It appears that, quite commonly, systematic reviewers seem to ignore the basic principles of EBM and the very different hierarchies necessary for issues of diagnosis, prognosis, and therapy. It has been stated that systematic reviews are only as complete and useful as the evidence that exists on a particular topic or the scope and nature of the evidence questions that guide the review. Even though there has been an explosion of systematic reviews and meta-analyses, empiric research on the quality of systematic reviews has shown that not all systematic reviews are truly systematic (188,189). The quality of systematic reviews of diagnostic accuracy studies is not only highly variable, but of low quality without following the methodological principles of assessment. In fact, authors have ignored or failed to realize the different aspects of diagnostic accuracy tests compared to therapeutic trials or evaluations. In studies of diagnostic accuracy, results from one or more tests are compared with the results obtained with the reference standard on the same subjects. Several factors threaten the internal and external validity of a study of diagnostic accuracy. These factors are all different from the quality of randomized or observational studies.

Diagnostic tests are also conducted in different phases: Phase I, II, III, and IV studies evaluate different questions yielding different answers (119). Phase I studies of diagnostic tests decide if the test results in affected patients differ from those in normal individuals. In contrast, Phase II studies of diagnostic tests are designed to answer if patients with certain test results are more likely to have the target disorder, comparing the range of test results of groups of patients who already have the established diagnosis. Phase III studies provide multiple answers to the question of whether or not the test results distinguish patients with and without the target disorder among those in whom it is clinically sensible to suspect the disorder. Phase III require showing the presence or absence of the disease, comparison with a gold standard, and blinded of whether or not of the test. Finally, Phase IV studies of diagnostic tests provide answers to the question if patients undergoing specific diagnostic tests fared better in their health outcomes than similar patients who have not been exposed to the test. A Phase IV study tests the clinical utility of the test. However, a test may be valid but not impact outcomes if there is no effective treatment available or may even adversely affect the patient who has the test done particularly if risky tests are performed and the available treatments are highly ineffective.

A rigorous evaluation process of diagnostic tests before introduction into clinical practice could not only reduce the number of unwanted clinical consequences related to misleading estimates of test accuracy, but also limit health care costs by preventing unnecessary testing. Various instruments have been developed to assess and report the quality of diagnostic accuracy studies (9,20,84,85,100,190-194). Poor methodologic quality has been empirically proven to affect the results of controlled trials and meta-analysis of diagnostic studies (100). Thus, the study quality should be assessed in any attempt to use results of published studies of diagnostic evaluation (101,102,192). Further, statistical methods should be provided to account for verification bias (195) and methods to evaluate tests for which there is no or only an imperfect reference standard available (196,197).

Multiple guidelines have been developed to evaluate the quality of systematic reviews. Oxman (198) noted the need for checklists analogous to flying an airplane. The most dangerous errors in reviews oare systematic ones (bias) rather than ones that occur by chance alone (random errors). Therefore, most important for doers and users of the review is to check

Table 1. *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

Source: Pai M et al. Systematic reviews and meta-analyses: An illustrated, step-by-step guide. *Natl Med J India* 2004; 17:86-95 (15).

its "validity," the extent its design and conduct are likely to have been protected against bias. Random errors and biases are of serious concern. In a properly performed systematic review with quantitative results, the confidence intervals (CIs) around the results should provide a good indicator of precision, the extent to which the results are likely to differ from truth because of chance alone (198-201). Oxman (198) provided guidance for the presentation of evaluation synthesis with a description of systematic review of 2 instruments critically appraising systematic reviews (202,203), and studies of how to present the results of a systematic review to policy-makers (204), the general public (205), and users of Cochrane reviews (206). One of the 2 reviews (203) utilized by Oxman (198) evaluated approximately 240 quality assessment instruments identified for systematic reviews, randomized controlled trials (RCTs), and observational studies, along with 50 evidence grading systems. Following this critical and extensive review, the AMSTAR 2005 was selected as the best instrument for appraising systematic reviews as illustrated in Table 1 (198,207).

The instrument developed by West et al (202) reviewed different instruments for critically appraising systematic reviews and found 20 systems concerned with the appraisal of systematic reviews or meta-analyses. They considered it important to appraise the study question, search strategy, selection or inclusion and exclusion criteria, data extraction, study quality, data synthesis and analysis, and funding or ownership as illustrated in Table 2 (14,56,208-211).

5.0 How to Conduct Systematic Reviews and Meta-Analyses of Diagnostic Accuracy Studies

Multiple documents describe steps for a systematic review or meta-analysis (15,48,56,57,188,212-214). While the central objective of a systematic review is to summarize the evidence on a specific clinical question (47,48,215,216), secondary objectives include critical evaluation of the quality of the primary studies, iden-

Table 2. *Domains in the Agency for Healthcare Research and Quality (AHRQ) criteria for evaluating systematic reviews.*

DOMAIN	ELEMENTS*
Study question	**• Question clearly specified and appropriate**
Search strategy	***• Sufficiently comprehensive and rigorous with attention to possible publication biases*** *• Search restrictions justified (e.g., language or country of origin)* • Documentation of search terms and databases used • Sufficiently detailed to reproduce study
Inclusion and exclusion criteria	**• Selection methods specified and appropriate, with a priori criteria specified if possible**
Interventions	**• Intervention(s) clearly detailed for all study groups**
Outcomes	**• All potentially important harms and benefits considered**
Data extraction †	• Rigor and consistency of process • Number and types of reviewers • Blinding of reviewers • Measure of agreement or reproducibility • Extraction of clearly defined interventions/exposures and outcomes for all relevant subjects and subgroups
Study quality and validity	***• Assessment method specified and appropriate*** • Method of incorporation specified and appropriate
Data synthesis and analysis	*• Appropriate use of qualitative and/or quantitative synthesis, with consideration of the robustness of results and heterogeneity issues* • Presentation of key primary study elements sufficient for critical appraisal and replication
Results	**• Narrative summary and/or quantitative summary statistic and measure of precision, as appropriate**
Discussion	**• Conclusions supported by results with possible biases and limitations taken into consideration**
Funding or sponsorship	***• Type and sources of support for study***

* Elements appearing in italics are those with an empirical basis. Elements appearing in bold are those considered essential to give a system a Yes rating for the domain.

† Domain for which a Yes rating required that a majority of elements be considered.

Adapted from West S et al. Systems to Rate the Strength of Scientific Evidence, Evidence Report, Technology Assessment No. 47. AHRQ Publication No. 02-E016. Rockville, MD: Agency for Healthcare Research and Quality, 2002. www.thecre.com/pdf/ahrq-system-strength.pdf (202).

tifying the sources of heterogeneity in results of cross studies, and determining sources of heterogeneity.

Guidelines illustrating the evaluation of systematic reviews of diagnostic accuracy (12) and meta-analysis (14) deliniate multiple steps in conducting a systematic review or meta-analysis. Leeflang et al (12) reviewed methodologic developments concerning problem formulation, location of literature, quality assessment, and meta-analysis of diagnostic accuracy studies by using their experience from the work on the Cochrane Handbook. Further, the information presented by Leeflang et al (12)is also based on the literature and updates previously published by Irwig et al (14). Leeflang et al (12) described multiple objectives.

5.1 Formulating the Question

As with any research, the first and most important decision in preparing a systematic review is to determine its focus (217). Clearly framed questions are essential for determining the structure of a systematic review or meta-analysis (217-220). Thus, the properly formulated question will guide much of the review process, including strategies for locating and selecting studies or data for critically appraising their relevance and validity, and for analyzing variations among their results.

Diagnostic test accuracy refers to the ability of a test to distinguish between patients with disease and those without. In a study of test accuracy, the results of the test and evaluation, the index test, are compared with those of the reference standard determined in the same patients. The reference standard is an agreed upon and accurate method for identifying patients who have the target condition. Test results are typically categorized as positive or negative for the target condition — a binary test outcome. The accuracy is most often described as the test sensitivity and the specificity; however, other measures have been proposed and are in use (221-223). Reviewers must understand that test accuracy is not a fixed property of a test (12). It can vary between patient subgroups with a spectrum of disease, with a clinical setting, or with the test interpreters, and may depend on the results of previous testing.

5.1.1 Key Components of a Question

A well formulated question consists of several key components which provide criteria for selecting studies (58,224). For diagnostic test accuracy, the study question should include multiple elements as subgroups, spectrum of the disease, clinical setting, the test interpreters, and the results of previous testing. In order to make a policy decision to promote the use of a new index test, evidence is required that using the new test increases test accuracy over the testing options, including current practice, or the new test has equivalent accuracy but offers other advantages (225-227).

As with the evaluation of interventions, systematic reviews of diagnostic accuracy need to include comparative analysis between alternative testing strategies and should not focus solely on evaluating the performance of a test in isolation.

5.1.1.1 Types of Participants and Settings

Test results can vary between patient subgroups, with a spectrum of disease, with a clinical setting, or with the test interpreters. Further, the results may also depend on the results of previous testing. Thus, selection or inclusion criteria for types of participants must be clear. First, define the disease or conditions that are of interest, such as facet joint pain, discogenic pain, or radicular pain. Second, the population of interest must be identified which involves deciding whether one is interested in a special population group determined on the basis of factors such as age, sex, race, educational status, or the presence of a particular condition such as low back pain or radiculitis. Third, setting may also be important such as community setting, ambulatory surgery setting, hospital outpatient setting, office setting, or inpatient setting.

Any restrictions with respect to specific population characteristics or settings should be based on sound evidence (217). For example, focusing a systematic review on the diagnostic accuracy of facet joint interventions in the Medicare population is justified based on controversial coverage policies and previously published opinions. However, focusing a review on a particular subgroup of people, based on some irrelevant factor, based on personal interest or bias when there is no underlying biological or sociological justification for doing so, is not acceptable.

5.1.1.2 Roles of a Test

Three potential roles for a new test include replacement, triage, and add-on (225). If a new test is to replace an existing test, comparison of the accuracy of both tests on the same population and with the same reference standard provides the most direct evidence. In a triage, the new test is used before the existing test or testing pathway, and only patients with a particular result on the triage test continue the testing pathway (12). When a test is needed to rule out disease in patients who then need no further testing, a test that gives a minimal proportion of false-negative results and thus a relatively high sensitivity should be used. The triage test may be less accurate than existing ones, but they have other advantages, such as simplicity or low cost. Finally, a third possible role of a new test is an add-on function. The new test is then positioned after the existing testing pathway to identify false-positive or false-negative results after the existing pathway. Thus, the systematic review should provide data to assess the incremental change in accuracy made by adding the new test.

While the potential roles for a new test are described, the question may also focus on existing tests with potential new roles of replacement of another test or by another test, triage, or to provide add-on function.

5.1.1.3 Types of Interventions

It is crucial to define the interventions in formulating a question, along with the specification of the interventions that are of interest. Intervention should be clearly described including placebo control or other controls. Various diagnostic tests from Phase I to IV require different designs. In contrast to randomized trials, the diagnostic studies in stages I to III are observational.

5.1.1.4 Types of Outcomes

The key component of well-formulated question is the delineation of particular outcomes that are of interest. While it is important to utilize primary outcomes (pain relief) and secondary outcomes such as improvement in functional status (ability to perform previously painful movements), trivial outcomes should not be included as they only overwhelm and confuse the readers by including data that is of little or no importance alongside the data that is important. Further, important data should never be left out. Consequently, explicit criteria for establishing the presence of appropriate outcomes and, if necessary, their combination must be specified. For example, outcome may be only a general statement of pain relief, or graduated pain relief of 80% or above, or a combination of pain relief with increased function or ability to perform previously painful movements without significant pain in a diagnostic study. Similarly, the outcome may be that pain production in the target disc with lack of pain whatsoever in contiguous discs.

5.1.1.5 Types of Study Designs

It is essential to identify the study designs. Phase I to III studies of diagnostic tests require observational studies, whereas Phase IV studies of diagnostic tests require a randomized design. Placebo control and its drawbacks in interventional pain management must be identified at present as the electrophysiologic effects of sodium chloride solution are recognized. The long-term effects of placebo over a nerve, nerve root, a closed space, or a disc are unknown.

5.1.2 Usefulness of Key Components of a Question

Properly focused questions should determine the initial searching strategies. This is related to the condition being studied, diagnostic intervention being assessed, and the population being studied. Above all, details relevant to the key components of the questions are what the authors will be collecting from individual studies. The questions that the review addresses may be broad or narrow in scope, both associated with certain advantages and disadvantages. However, the question may be refined based on the data which is available during the review. Further, it is essential to guard against bias in modifying questions, as post-hoc questions are more susceptible to bias than those asked a priori, and data-driven questions can generate false-conclusions based on spurious results. In addition, any changes to the protocol that results from revising the question for the review should be documented clearly.

5.2 Finding Relevant Studies

Finding the relevant studies is a complex and time-consuming process. The aim of research is to generate as comprehensive a list as possible of primary studies, both published and unpublished, which may be suitable to answer the question in review (228-232). However, NICE (233) has concluded that a more selective approach to database searching would suffice in most cases and would save resources. However, they have not specifically described the role in diagnostic accuracy studies. Due to paucity of diagnostic studies, specifically in interventional pain management, identification of relevant diagnostic accuracy studies by a thorough unbiased search strategy is crucial. A comprehensive search captures the relevant trials, leading to the validity of the systematic review. In addition, the level of precision of accuracy of a diagnostic test estimated by a systematic review depends on the volume of information included in the review. In essence, a comprehensive search for relevant diagnostic accuracy studies which seeks to minimize bias is one of the essential steps in doing a systematic review and one of the factors that distinguishes a systematic review from narrative or focused review. Unfortunately, the emerging CER and various organizations appear to have followed the philosophy of NICE (59,60,135,183,233-237). Recent analysis of Cochrane reviews of interventional pain management and other extensively quoted reviews (59,60,237-242) showed a lack of appropriate criteria and absence of many key manuscripts. The same was true with highly outspoken critics of interventional pain management in reviews (243,244).

5.2.1 Searching for Studies

Identifying test accuracy studies is more difficult than searching for randomized trials (245). There is not a clear, unequivocal keyword or indexing term for an accuracy study in literature databases comparable with the term "randomized, controlled trial." The Medical Subject Heading "sensitivity and specificity" may look suitable but is inconsistently applied in most electronic bibliographic databases (12). Furthermore, data on diagnostic test accuracy may be hidden in studies that did not have test accuracy estimation as their primary objective. Consequently, efficient identification of diagnostic test accuracy studies in electronic databases is complicated. Thus, searching for studies of diagnostic test accuracy will remain challenging and may require additional manual searches, such as screening reference lists (12).

A comprehensive search strategy may be developed for diagnostic accuracy studies by referring to the test(s) under evaluation, the target condition, and the patient description or subset of these. For tests with a clear name that are used for a single purpose such as provocation discography, searching for publications is easier; whereas, for other reviews, the addition of multiple terms may be necessary. Several methodological electronic search filters for diagnostic test accuracy studies have been developed, each attempting to restrict the search to articles that are most likely to be test accuracy studies (245-248). These filters generally rely on indexing terms for research methodology and text words used in reporting results, but they often miss relevant studies and are likely to decrease the number of articles one needs to screen. Therefore, they may not be recommended for systematic reviews of diagnostic accuracy studies (249,250). The value of searching in languages other than English and in the gray literature has not yet been fully investigated (12). In addition, investigating publication bias for diagnostic tests is problematic, because many studies are done without trial registration.

A quick and dirty search of, for example, MEDLINE, is generally not considered adequate. Studies have shown that even for RCTs, only 30% to 80% of all known published articles were identified using MEDLINE (251). Variations in the journals indexed in databases indicate a need to search more than one database to ensure optimal coverage of published literature, in subject, scope, and language of the report (252-254). It has been shown that there is significant value to adding EMBASE to MEDLINE in the search strategy (252). The overlap of EMBASE and MEDLINE has been estimated to be 10% to 87% depending on the topic under investigation (255-259). Further, comparison of databases has shown that relevant studies would be missed if only MEDLINE were searched for studies in multiple specialties. The results of many studies are never published, and most of these probably remain unknown. It has been the universal belief that studies showing an intervention to be effective are more likely to be published, thus any summary of only the published reports may result in an overestimate of the effect due to a publication bias (232,260). Contrary to this, in more recent years, due to the explosion of many journals and the bias exerted by them, it appears that negative trials are published more frequently than the positive trials based on self interest and turf protection. Even though there is no empirical evidence, on a pragmatic basis, a systematic review in interventional pain management at a minimum must have a comprehensive review using at least 3 sources and provide a description of efforts to identify all databases and journals, if not, unpublished trials. An effective combination of a comprehensive search includes a minimum of 3 bibliographic databases (MEDLINE, EMBASE, Cochrane library), a hand search of reference of eligible studies, and direct contact with the corresponding authors of eligible studies asking for additional published or unpublished information (261).

In summary, a search strategy must be developed and documented clearly. It is essential to strike a balance between comprehensiveness and precision. An electronic search strategy generally includes 3 sets of terms: terms to search for the health condition of interest, terms to search for the diagnostic interventions evaluated, and terms to search for the types of study design.

5.3 Study Selection and Quality Assessment

Once the search for potentially relevant studies is completed, the study should be retrieved and assessed for the relevance to the question posed in the review. The selection process should be explicit and should be conducted in such a way as to minimize the risk of errors of judgment (9,262-264). Quality assessment of primary studies is used at various stages in the review process, from study selection to generation of recommendations for practice and research (263,264).

Selection of the studies must be based on an explicit and standardized methodology. Such methodology ensures a selection of the highest-quality studies and demonstrates that the selection and assessment have been as free from bias as possible (230,265-269). The selection or inclusion or exclusion of the studies must be made according to predetermined written criteria as described in the protocol.

5.3.1 Selection or Inclusion Criteria

The review question determines both inclusion and exclusion criteria. They should be defined in terms of the population, the interventions, the outcomes, and the study design of interest. Thus, only studies that meet all of the predetermined selection or inclusion criteria and none of the exclusion criteria should be included in a review. The selection criterion specifying the type of the study design stems from the desire to base reviews on the highest quality evidence (270). In interventional pain management, multiple systematic reviews of diagnostic accuracies have not been evaluated with methodologically sound studies (59,60,183,234-236,241-244). Thus, studies of methodologically lower quality are inaccurately interpreted and included. By the same token, appropriate studies may be excluded without basis (59,60,234-236,242,243).

5.3.2 Study Selection Process

The study selection process is crucial and involves multiple stages. In searching for the manuscripts, the liberal selection criteria are applied and multiple manuscripts are generated. Thus, unless studies can be definitely excluded, the titles and abstracts identified as being potentially relevant from searches should be provisionally included for consideration on the basis of full text articles (264). However, the final inclusion or exclusion decision should be made only after retrieving the full text of all potentially relevant citations. Thus, many of the citations initially included may be excluded at later stages. In addition, a list of excluded studies may be made detailing the reason for each exclusion. A final report of the review may also include a flow chart or a table detailing the studies included and excluded from the review (212), which are described in the text.

5.4 Study Quality Assessment

Assessment of the methodologic quality of diagnostic accuracy studies and detailed reporting is crucial in a systematic review. Variability among diagnostic accuracy studies results is to be expected. However, some of this variability is due to chance, because many diagnostic studies have small sample sizes (118). The remaining heterogeneity may be due to differences in study population, but differences in study methods are also likely to result in differences in accuracy estimates (271). Test accuracy studies with design deficiencies can produce biased results (12,100,113,117). Table 3 describes some of the most important forms of bias. Sources of bias for which unambiguous evidence indicates that they lead to overestimation of diagnostic accuracy or the inclusion of healthy controlled participants and the differential use of referenced standards (100,113,117,272). Further, funnel-plot-based tests used to detect publication bias in reviews of RCTs have proven to be seriously misleading for diagnostic studies, and alternatives have poor power (272). In addition, because diagnostic accuracy studies frequently do not compare tests, they tend not to routinely report *P* values that dichotomize comparisons as significant or not significant (12). Without the same emphasis being given to statistical significance, the determinants for publication of diagnostic studies are unlikely to be the same as those of intervention studies (12).

5.4.1 Quality Assessment of Diagnostic Accuracy Studies

Quality assessment of individual studies in systematic reviews is therefore necessary to identify potential sources of bias and to limit the effects of these biases on the estimates and the conclusions of the review. There are several instruments for methodologic quality assessment of diagnostic studies. West et al (202) in the AHRQ Evidence Report of Technology Assessment provided pertinent evidence for rating the quality of individual articles, including studies of diagnostic tests. This panel identified approximately 20 systems, checklists, and developed 5 key domains for making judgments about the quality of diagnostic test reports as shown in Table 4. This scoring has been applied in multiple systematic reviews with weighted scoring (26-38,41,67,73,81,141). In addition, a tool for QUADAS (Quality Assessment Tool for Diagnostic Accuracy Studies) was developed by combining empirical evidence and expert opinion in a formal consensus method. The QUADAS tool is presented together with guidelines for scoring each of the items included in the tool as shown in Table 5. No weighted scoring system has been developed thus far for the QUADAS tool. While Leeflang et al (12) recommend the QUADAS tool, both tools appear very similar. The results of

Table 3. *Sources of bias in diagnostic test accuracy studies.*

Type of Bias	When Does It Occur?	Under-or Overestimation of Diagnostic Accuracy?*
Patients		
Spectrum bias	When included patients do not represent the intended spectrum of severity for the target condition or alternative conditions	Depends on difference between targeted and included part of spectrum
Selection bias	When eligible patients are not enrolled consecutively or randomly	Usually leads to overestimation
Index test		
Information bias	When the index test results are interpreted with knowledge of the results of the reference standard, or with more (or less) information than in practice	Usually leads to overestimation, unless less clinical information is provided than in practice, which may result in underestimation
Reference standard		
Misclassification bias	When the reference standard does not correctly classify patients with the target condition	Depends on whether both tests make the same mistakes
Partial verification bias	When a nonrandom set of patients does not undergo the reference standard	Usually leads to overestimation of sensitivity; effect on specificity varies
Differential verification bias	When a set of patients is verified with a second or third reference standard, especially when this selection depends on the index test result	Usually leads to overestimation
Incorporation bias	When the index test is incorporated in a (composite) reference standard	Usually leads to overestimation
Disease progression bias	When the patients' condition changes between administering the index test and the reference standard	Under-or overestimation, depending on change in patients' condition
Information bias	When the reference standard is interpreted knowing the index test results	Usually leads to overestimation
Data analysis		
Excluded data	When uninterpretable or intermediate test results and withdrawals are not included in the analysis	Usually leads to overestimation

* From refs. (100,113,117).

Table 4. *Modified AHRQ methodologic assessment criteria for diagnostic interventions.*

Criterion	Weighted Score (points)
1. Study Population	15
Subjects similar to populations in which the test would be used and with a simlar spectrum of disease	
2. Adequate Description of Test	10
Details of test and its administration sufficient to allow for replication of study	
3. Appropriate Reference Standard	30
Appropriate reference standard (gold standard) used for comparison	
Reference standard reproducible	
4. Blinded Comparison of Test	30
Evaluation of test without knowledge of disease status, if possible	15
Independent, blind interpretation of test and reference	15
5. Avoidance of Verification Bias	15
Decision to perform reference standard not dependent on results of test under study	
Total Score	100

Adapted and modified from West S et al. S*ystems to Rate the Strength of Scientific Evidence, Evidence Report*, Technology Assessment No. 47. AHRQ Publication No. 02-E016 (202).

Table 5. *The QUADAS tool.*

Item	Yes	No	Unclear
1. Was the spectrum of patients representative of the patients who will receive the test in practice?	()	()	()
2. Were selection criteria clearly described?	()	()	()
3. Is the reference standard likely to correctly classify the target condition?	()	()	()
4. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	()	()	()
5. Did the whole sample or a random selection of a sample, receive verification using a reference standard of diagnosis?	()	()	()
6. Did patients receive the same reference standard regardless of the index test result?	()	()	()
7. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	()	()	()
8. Was the execution of the index test described in sufficient detail to permit replication of the test?	()	()	()
9. Was the execution of the reference standard described in sufficient detail to permit its replication?	()	()	()
10. Were the index test results interpreted without knowledge of the results of the reference standard?	()	()	()
11. Were the reference standard results interpreted without knowledge of the results of the index test?	()	()	()
12. Were the same clinical data available when the test results were interpreted as would be available when the test is used in practice?	()	()	()
13. Were uninterpretable/intermediate test results reported?	()	()	()
14. Were withdrawals from the study explained?	()	()	()

Adapted from Whiting P 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 (191).

quality appraisal can be summarized to offer a general impression of the validity of the available evidence. However, it also has been stated that the review author should not use an overall quality score because different shortcomings may generate different magnitudes of bias, even in opposing directions, which makes it very hard to attach sensible weights to each quality item (273). However, without such a display, the reviewer bias may impede the accurate analysis of the manuscripts. Various methods described address quality differences or sensitivity analysis, subgroup analysis, or meta-regression analysis, although the number of included studies may often be too small for meaningful investigation.

Incomplete reporting hampers any evaluation of study quality (98). Even though STARD guidelines are being utilized for reports of diagnostic accuracy studies this is a slow development (85).

The information gained from quality assessment is crucial in determining the strength of inferences and in assigning grades to recommendations generated within a review. Quality assessment can be used at various stages in a review, starting with the study selection to data synthesis and interpretation. While almost every systematic review has supporters and detractors, both groups agree on relevance of the dictum, “garbage in, garbage out” (274). It is always argued that if the study quality was assessed appropriately — if it was assessed at all — the expertise of various authors of reviews vary widely with some considering the quality assessment as an important strategy to identify and reduce bias and others who see assessment as a source of bias or as completely uninformative, whereas, some others criticize the criteria utilized with a multitude of personal biases (275,276).

5.4.2 Validity

In the context of a systematic review, the validity of a study is the extent to which its design and conduct are likely to prevent systematic errors or bias. An important issue that should not be confused with validity is precision. Precision is a measure of the likelihood of chance effect leading to random errors. It is reflected in the CI around the estimate of effect from each study and the weight given to the results of each study when an overall estimate of the effect or weighted average is derived. However, more precise results are given more weight.

Variation and validity can explain differences in the results studies included in a systematic review.

More rigorous studies may more likely yield results that are close to the truth. Quantitative analysis of results from studies of variable validity can result in false-positive conclusions (erroneously concluding a diagnostic intervention is positive), if the less rigorous studies are biased towards overestimating and interventions effect. They might also come to false-negative conclusions (erroneously concluding a negative result) if the less rigorous studies are biased towards underestimating an intervention's effect (277). Thus, it is important to systematically complete a critical appraisal of all studies in a review even if there is no variability in either the validity or results of the included studies. In a hypothetical situation, the results may be consistent among studies, but all the studies may be flawed, providing conclusions which are flawed and the conclusions would not be as strong as if a series of rigorous studies yielded consistent results about an intervention's effect.

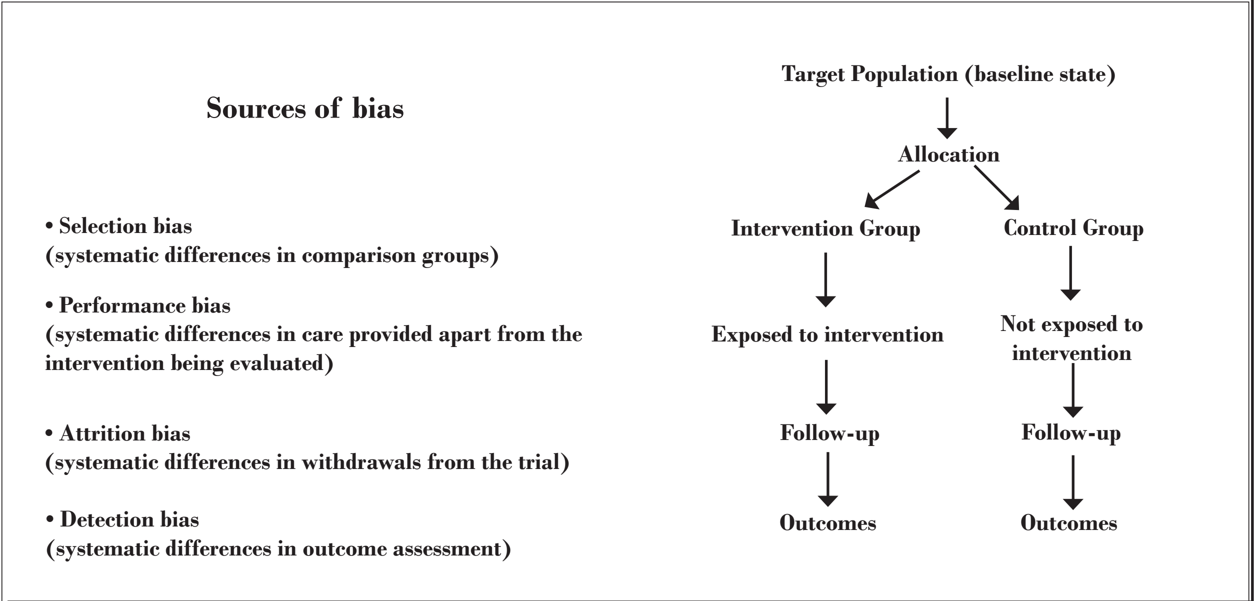
5.4.3 Assessment of Bias

Sources of bias in diagnostic test accuracy studies are shown in Figure 1. In a classic diagnostic accuracy study, a consecutive series of patients who are suspected of having the target condition undergo the index test. Bias is said to be present in a study if distortion is introduced from a consequence of defects in the design or conduct of the study. Therefore, a biased diagnostic accuracy study will produce estimates of the test performance that differ from the true performance of the test.

5.4.4. Assessment of Bias and Variation

In a classic diagnostic accuracy study a consecutive series of patients who are suspected of having the target condition undergo the index test. All patients are then verified by the same reference standard. The index test and reference standard are then read by persons blinded to the results of each and various measures of agreements are calculated, which include sensitivity, specificity, likelihood ratios, and diagnostic odds ratios. The classic design, however, has many variations, including differences in the way patients are selected for the study, in test protocol, in the verification of patients, and in the way the index test and reference standard are read. Some of the differences may bias the results of a study, whereas others may limit the applicability of the results (9,117).

Variations arise from the differences among studies, for example, in terms of population, setting, test protocol, or definition of the target condition (278). Variability limits the applicability of the results but does not lead to biased estimations. Even though bias and variation are different, the distinctions are not.



Sources of bias

- **Selection bias**
  (systematic differences in comparison groups)
- **Performance bias**
  (systematic differences in care provided apart from the intervention being evaluated)
- **Attrition bias**
  (systematic differences in withdrawals from the trial)
- **Detection bias**
  (systematic differences in outcome assessment)

Fig. 1. *Sources of bias in trials of health care interventions. Adapted and modified from Glasziou P et al.* Systematic Reviews in Health Care. A Practical Guide. *University Press, Cambridge, 2001 (216).*

Essentially, diagnostic accuracy is higher both with sensitivity and specificity when patients with a target condition and healthy volunteers are compared. Bias is higher sensitivity and specificity when patients with a target condition and healthy volunteers are compared in a diagnostic test which is considered as a spectrum bias. Others may argue that it is a form of variability; sensitivity and specificity have been measured correctly within the study, and thus there is no bias. Still, the results cannot be applied to the clinical setting as the results lack generalizability (279). Even then some have argued that when the goal of a study is to measure the accuracy of a test in the clinical setting, an error in the method of patient selection is made that will lead to biased estimates of test performance (117). The largest overestimation of accuracy was found in studies that included severe cases and healthy controls (113). In addition, the design features associated with significant overestimation of diagnostic accuracy included non-consecutive inclusion of patients and retrospective data collection. Further, random inclusion of eligible patients and differential verification also resulted in higher estimates of diagnostic accuracy. The selection of patients on the basis of whether they had been referred for the index test, rather than on clinical symptoms, was significantly associated with lower estimates of accuracy.

5.5 Data Collection

Data collection is considered as a bridge between what has been reported by primary investigators and what is ultimately reported by authors of systematic review. Collection of data either electronically or on a paper format serves 3 important functions. The data collection is directly linked to the formulated review question and planned assessment of included studies, and therefore, provides a visual representation of these (280). In addition, the data collection format is the historical record of the multitude of decisions and changes to decisions that occurred through the review process. Finally, the data collection format is a data repository from which the analysis will emerge. The components of data collection should include essential information and also methodologic quality assessment criteria.

5.6 Summarizing and Synthesizing Relevant Study Results

Systematic reviews are performed to summarize the findings of the best studies available (281,282). Thus, it is essential to provide a concise written summary of each of the relevant studies, often as a table of summaries. If a quantitative synthesis of results is described, the statistical method of meta-analysis is employed, and a summary result is produced, but this is not always necessary or appropriate. In larger studies that provide more precise diagnostic effects, diagnostic results are routinely given more weight in the meta-analysis calculations. One of the most common forms of a systematic review involves collaborating researchers pooling individual patient data from different studies. While not common, this method has been used in a number of studies. No studies have been produced in interventional pain management with meta-analysis for diagnostic accuracy studies.

In a randomized trial, the results are often reported by using a single measure of effect, such as difference in means, a risk difference, or a risk ratio. In contrast, diagnostic test accuracy studies report 2 or more statistics: the sensitivity and the specificity, the positive and negative predictive value, the likelihood ratios for the respective test results, or the receiver-operating characteristics (ROC) curve and quantities based on it (223,283).

5.6.1 Descriptive or Non-Quantitative Synthesis

The objective of a descriptive or non-quantitative review is to correlate and present the extracted data in a manner such that information about the characteristics (population, adequate description of test, appropriate reference standard, blinded comparison of test, and avoidance of verification bias) and results of the studies included in the review are summarized in a meaningful way. This is best done by tabulation, which allows readers to look at the evidence, its methodological rigor, and the differences between the studies. The descriptive overview is an essential part of the data on which an understanding of the data, planning and quantitative data synthesis, and preventing errors in its interpretation are dependent. Thus, the process of carrying out the descriptive part of the data synthesis should be explicit and rigorous (266,284). In general, the results of a diagnostic test are dependent on a large number of factors, some known and others unknown relating to who receives it, who delivers it and how, and in what context. The key elements in the descriptive approach to data synthesis may include multiple characteristics such as population; interventions; settings where the technology was applied; environmental, social, and cultural factors that may influence compliance; nature of the outcome measures used, their relative importance and robustness; the validity of evidence; the sample sizes; and results of the studies included in the review.

Data synthesis involves computation of an average effect where the results of each study are weighed according to some measure of the studies importance. The weight of each study usually relates to the study population, appropriate reference standard, and avoidance of bias.

5.6.2 Quantitative Synthesis or Meta-Analysis

The meta-analysis is performed to increase the power, to improve precision, and to answer the questions not posed by the individual studies, and to settle controversies arising from conflicting studies or to generate new hypotheses (285).

The first step in the meta-analysis of diagnostic test accuracy is to draft the results of the individual studies. The paired results for sensitivity and specificity in the included studies should be plotted as points in ROC space, which can highlight the covariation between sensitivity and specificity. However, one of the disadvantages is that forest plots do not display the covariation between sensitivity and specificity. The ability to estimate underlying summary ROC curves and average sensitivities and specificities allows flexibility in testing a hypothesis and estimating diagnostic accuracy. Analysis based on all included studies facilitates well-powered comparison between different tests or between subgroups of studies, which are not restricted to investigating accuracy at a particular threshold. The judgements about the validity of pooling data should be informed by considering the quality of the studies, the similarity of patients and tests being pooled, and whether the results may consequently be misleading. Where there is statistical heterogeneity in results, random-effects models are used to account for the variability and derive suitably conservative assessments of the uncertainty in the estimates. Naturally, increased uncertainty about the estimates may make it more difficult to draw from conclusions about the accuracy of a particular test.

5.6.3 Interpreting the Results

The interpretation of the results offered in the systematic review should help readers to understand the implications for practice (12). The interpretation should consider whether the evidence derived from the review suitably addresses the objectives of the review. The interpretation of the findings should consider the consequences of the false-positive and false-negative results and whether the estimates of the accuracy are sufficiently high for the foreseen role that the test will have in practice. However, some systematic reviews may not result in useful summary estimates of sensitivity and specificity due to various issues such as large variability in the individual study estimates.

Table 6 shows the essential elements in a systematic review of diagnostic test accuracy.

5.6.4 Strength of Evidence

One of the major goals of interpretation is to try to explain the strength of the evidence from the different studies that the review summarized. In other words, for a clinical question for a diagnostic intervention, the user of the review needs to know whether the best available evidence comes from the study designs at a high level in the hierarchy of evidence. Conclusions regarding the strength of inferences about the accuracy of diagnostic studies are essentially causal inferences.

5.6.5 Level of Evidence

The National Health and Medical Research Council (NHMRC) of Australia considered scientific data to be at the core of evidence-based approaches to clinical or public health issues, emphasizing that evidence needs to be carefully gathered and collated from a systematic literature review of each particular issue in question (286). West et al (202) also published systems for grading of the strength of a body of evidence. Thus, in determination of level of evidence, grading the quality of individual studies and rating the strength of body of evidence are both crucial elements. However, the systems for grading the strength of a body of evidence are less uniform and consistent than those rating the study quality (202). Selecting the evidence to be used in grading systems depends on the reason for measuring evidence strength, the types of studies that are being summarized, and the structure of the review panel. Table 7 illustrates domains for rating the overall strength of a body of evidence (286).

Table 8 illustrates panel ratings of available evidence supporting guideline statements developed by AHRQ (formerly AHCPR) (287). However, the document publishing these ratings is considered as outdated. The United States Preventive Services Task Force (USPSTF) (288) describes level of evidence for therapy rather than diagnosis. These have been modified to incorporate diagnostic studies in multiple systematic reviews performed (Table 9) (26-41,67-78,81-83) .

Table 6. *Essential elements in a systematic review of diagnostic test accuracy.*

Phase in Review Process	Key Issues
1. Definition of the review objectives	To identify the review question:
	State the patient group and define presenting condition(s), previous test results, and health care setting.
	Describe the tests (or test strategies) under evaluation, specifying their intended roles. Identify tests and test strategies currently used in practice for comparison, if available.
	Define the target condition to be diagnosed and reference standards to be used.
2. Study identification and selection	Search several electronic databases.
	Use a search strategy built around terms for the index test, target condition, and possibly patient characteristics.
	Do not use restrictive methodological search filters.
3. Quality assessment	Identify biases for which the included studies are at risk.
	Use the QUADAS checklist as a tool for identifying many common deficiencies.
	Comment on the adequacy of each aspect of study design. Do not use summary quality scores.
4. Data extraction, analysis, and presentation	Extract paired estimates of test sensitivity and specificity from each study overall and, if available, for patient subgroups.
	Plot studies in ROC space to identify the location, variability, and correlations.
	The hierarchical summary ROC and bivariate random-effects models provide a sound statistical framework for analysis, accounting for sampling variability, unexplained heterogeneity, and covariation between sensitivity and specificity.
	Compute average values of sensitivity and specificity when the data combined share a common threshold.
	Use summary ROC curves to describe test performance and to compare tests without restricting to particular thresholds.
	Obtain estimates of summary likelihood ratios from average values of sensitivity and specificity and not through separate pooling of likelihood ratios.
	Global tests for heterogeneity before data synthesis or tests for publication bias are typically not useful.
	Meta-analyze and present studies that compare tests by using randomized or within-patient designs separately from the results of indirect comparisons.
5. Interpretation	Consider the consequences of using the test, in terms of (changes in) the numbers of true-positive, false-positive, true-negative, and false-negative test results with the expected prevalence of the target disorder.
	Address the applicability of the results in terms of whether the patients in the primary studies were similar to those outlined in the objective, and whether tests and test strategies evaluated and compared were representative of test strategies that are used in practice.
	Address to what extent the original studies were biased and how these biases could influence the results and the degree to which comparisons between tests may be confounded.
	Consider complementing the interpretation with decision modeling by using results of the review.

QUADAS = Quality Assessment of Diagnostic Accuracy Studies; ROC = receiver-operating characteristic.

Leeflang MM et al. Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008; 149:889-897 (12).

Table 7. *Criteria for rating the overall strength of a body of evidence.*

Domain	Definition
Quality	• The quality of all relevant studies for a given topic, where "quality" is defined as the extent to which a study's design, conduct, and analysis has minimized selection, measurement, and confounding biases
Quantity	• The magnitude of treatment effect
	• The number of studies that have evaluated the given topic
	• The overall sample size across all included studies
Consistency	• For any given topic, the extent to which similar findings are reported from work using similar and different study designs

Adapted from How to use the evidence: Assessment and application of scientific evidence. National Health and Medical Research Council, Canberra, Commonwealth of Australia, 2000, pp 1-84 (286).

Table 8. *Panel ratings of available evidence supporting guideline statements.*

A	Strong research-based evidence (multiple relevant and high-quality scientific studies).
B	Moderate research-based evidence (one relevant high-quality scientific study or multiple adequate scientific studies*).
C	Limited research-based evidence (at least one adequate scientific study* in patients with low back pain).
D	Panel interpretation of information that did not meet inclusion criteria as research-based evidence.

* Met minimal formal criteria for scientific methodology and relevance to population and specific method addressed in guideline statement.

Note: These criteria were derived from Bigos SJ et al. Acute low back problems in adults. Clinical Practice Guideline No.14, AHCPR Publication No. 95-0642. Rockville, Maryland. U.S.A., Agency for Health Care Policy and Research, Public Health Service, U.S., Department of Health and Human Services, December, pp. 1-60, 1994 (287). AHCPR was extinguished by Congress in 1995, changing AHCPR to AHRQ. Acute Low Back Pain Guidelines (287 provides a disclaimer "not for patient care.")

5.6.6 Grading Recommendations

Grading recommendations have been provided for guidelines for therapy based on the RCTs and observational studies (289). However, no such recommendations are available for diagnostic accuracy studies. Guyatt et al (289) developed grading strength of recommendations and quality of evidence in clinical guidelines based on randomized and observational studies for therapy. They recommended that guideline panels should make recommendations to administer or not to administer an intervention on the basis of a trade-off between benefits on one hand and the risks, burdens, and potential costs on the other. They provided only 2 levels of recommendations, either strong or weak, with 3 subcategories. However, a number of factors in grading recommendations must be considered. These include 1) methodologic quality of evidence reporting estimates of likely benefit and likely risk, inconvenience, and costs; 2) importance of the outcome; 3) magnitude of the treatment effect; 4) estimate of treatment effect; 5) risks associated with therapy; 6) burden of therapy; 7) risk of target event; 8) costs; and finally 9) circumstances, patients' or societal values. In contrast to much of the literature, they have provided strong recommendations for exceptionally strong evidence derived from observational studies. As illustrated in Table 10, which shows grading recommendations of Guyatt et al (289), methodologic quality of supporting evidence may be converted to incorporate diagnostic accuracy studies.

Table 11 illustrates modified methodologic quality of supporting evidence for diagnostic accuracy studies.

5.6.7 Applicability

Generalizability of the results also known as applicability to the general population of diagnostic accuracy tests are crucial. Decisions about applicability depend on knowledge of particular circumstances in which decisions about health care are being made; however, authors of systematic reviews should cautiously approach the issue of applicability and should not assume that their own circumstances, or the circumstances reflected in the included studies, are necessarily the same as those of others. Further, system-

Table 9. *Modified quality of evidence developed by USPSTF.*

I:	Evidence obtained from at least one properly randomized controlled trial or multiple properly conducted diagnostic accuracy studies.
II-1:	Evidence obtained from one well-designed controlled trial without randomization or at least one properly conducted diagnostic accuracy study of adequate size
II-2:	Evidence obtained from at least one properly designed small diagnostic accuracy study.
II-3:	Evidence obtained from multiple time series with or without the intervention. Dramatic results in uncontrolled experiments (such as the results of the introduction of penicillin treatment in the 1940s) could also be regarded as this type of evidence.
III:	Opinions of respected authorities, based on clinical experience descriptive studies and case reports or reports of expert committees.

Adapted from the U.S. Preventive Services Task Force (USPSTF) (288).

Table 10. *Grading recommendations*.

Grade of Recommendation/ Description	Benefit vs Risk and Burdens	Methodological Quality of Supporting Evidence	Implications
1A/strong recommendation, high-quality evidence	Benefits clearly outweigh risk and burdens, or vice versa	RCTs without important limitations or overwhelming evidence from observational studies	Strong recommendation, can apply to most patients in most circumstances without reservation
1B/strong recommendation, moderate quality evidence	Benefits clearly outweigh risk and burdens, or vice versa	RCTs with important limitations (inconsistent results, methodological flaws, indirect, or imprecise) or exceptionally strong evidence from observational studies	Strong recommendation, can apply to most patients in most circumstances without reservation
1C/strong recommendation, low-quality or very low-quality evidence	Benefits clearly outweigh risk and burdens, or vice versa	Observational studies or case series	Strong recommendation but may change when higher quality evidence becomes available
2A/weak recommendation, high-quality evidence	Benefits closely balanced with risks and burden	RCTs without important limitations or overwhelming evidence from observational studies	Weak recommendation, best action may differ depending on circumstances or patients' or societal values
2B/weak recommendation, moderate-quality evidence	Benefits closely balanced with risks and burden	RCTs with important limitations (inconsistent results, methodological flaws, indirect, or imprecise) or exceptionally strong evidence from observational studies	Weak recommendation, best action may differ depending on circumstances or patients' or societal values
2C/weak recommendation, low-quality or very low-quality evidence	Uncertainty in the estimates of benefits, risks, and burden; benefits, risk, and burden may be closely balanced	Observational studies or case series	Very weak recommendations; other alternatives may be equally reasonable

Adapted from Guyatt G et al. Grading strength of recommendations and quality of evidence in clinical guidelines. Report from an American College of Chest Physicians task force. *Chest* 2006; 129:174-181 (289).

atic reviews may assist with recommendations about applicability by drawing attention to the spectrum of circumstances to which the evidence is likely to be applicable (290).

5.6.8 Limitations

The interpretation may also discuss the trade-offs between the benefits and harms, and, less often, costs. The cost-effective analysis or economic evaluation is important for policy decisions.

5.7 Updating Reviews

Updating and improving access to diagnostic accuracy systematic reviews is crucial in modern medicine, similar to randomized and observational studies. The updating requirements have been described as variable from 2 to 4 years. The requirements for updates illustrated that a qualitative or quantitative signal for updating occurred in 57% of reviews with a median duration of survival free of a signal for updating of

Table 11. *Modified methodologic quality of supporting evidence for diagnostic accuracy studies.*

Grade of Recommendation/ Description	Benefit vs Risk and Burdens	Methodological Quality of Supporting Evidence (modified to incorporate diagnostic accuracy studies based on Guyatt et al's grading strength of recommendation)	Implications
1A/strong recommendation, high-quality evidence	Benefits clearly outweigh risk and burdens, or vice versa	1A/controlled diagnostic accuracy studies without important limitations.	Strong recommendation, can apply to most patients in most circumstances without reservation
1B/strong recommendation, moderate quality evidence	Benefits clearly outweigh risk and burdens, or vice versa	1B/controlled diagnostic accuracy studies with important limitations, either with inconsistent results, methodological flaws, indirect or imprecise results.	Strong recommendation, can apply to most patients in most circumstances without reservation
1C/strong recommendation, low-quality or very low-quality evidence	Benefits clearly outweigh risk and burdens, or vice versa	1C/diagnostic accuracy studies, which have not been controlled.	Strong recommendation but may change when higher quality evidence becomes available
2A/weak recommendation, high-quality evidence	Benefits closely balanced with risks and burden	2A/controlled diagnostic accuracy studies without important limitations.	Weak recommendation, best action may differ depending on circumstances or patients' or societal values
2B/weak recommendation, moderate-quality evidence	Benefits closely balanced with risks and burden	2B/controlled diagnostic accuracy studies with important limitations, either with inconsistent results, methodological flaws, indirect or imprecise results.	Weak recommendation, best action may differ depending on circumstances or patients' or societal values
2C/weak recommendation, low-quality or very low-quality evidence	Uncertainty in the estimates of benefits, risks, and burden; benefits, risk, and burden may be closely balanced	2C/ diagnostic accuracy studies, which have not been controlled.	Very weak recommendations; other alternatives may be equally reasonable

Modified from Guyatt G et al. Grading strength of recommendations and quality of evidence in clinical guidelines. Report from an American College of Chest Physicians task force. *Chest* 2006; 129:174-181 (289).

5.5 years (291). Even then, 7% of the reviews required revision at the time of the publication, 15% required a review within one year, and 23% of reviews required a review within 2 years. Considering that interventional pain management is an evolving specialty, longevity and survival of the diagnostic accuracy systematic reviews and the related topics may be shorter than other subjects.

6.0 Reporting of Systematic Reviews

Even though Quality of Reporting of Meta-analyses QUOROM (212) and Meta-analysis of Observational Studies in Epidemiology (MOOSE) (292) have described proposals for reporting systematic reviews and meta-analysis, no such guidelines have been established for systematic reviews of diagnostic accuracy studies. Further, diagnostic accuracy studies may also be reported in conjunction with evaluation of treatment. Consequently, based on QUOROM and MOOSE, the following checklist is proposed for diagnostic accuracy studies (Table 12).

6.1 Title

The title should identify the report as a systematic review or meta-analysis of diagnostic accuracy studies.

6.2 Abstract

The structured abstract must provide a series of headings pertaining to the design, conduct, and analysis of diagnostic intervention with standardized information appearing under each heading. The literature is replete with reports illustrating that structured abstracts are of higher quality than the more traditional descriptive abstracts (293-298). These headings include background, study design, objective, methods, level of evidence, outcome measures, results, limitations, and conclusions.

Table 12. *A proposed reporting checklist of systematic review or meta-analysis of diagnostic accuracy studies.*

Heading	Subheading	Descriptor	Reported? (Y/N)	Page number
Title		Identify the report as a systematic review or meta-analysis of diagnostic accuracy studies.		
Abstract		The abstract must utilize a structured format.		
	Background	Problem definition Hypothesis statement		
	Study Design	Clearly specify the nature of the design, either it is systematic review or meta-analysis or both.		
	Objective(s)	Identify the objective(s) of the systematic review.		
	Methods	Identify the methods related to literature search, inclusion criteria, method of review or validity assessment.		
	Level of Evidence	Identify the nature of determination of level of evidence.		
	Outcome Measures	Identify the outcome measures utilized in the accuracy studies.		
	Results	Identify characteristics of the diagnostic accuracy studies included and excluded; qualitative and quantitative findings; and subgroup analysis.		
	Limitations	Identify the limitations of the systematic review.		
	Conclusion	Identify the main results.		
Introduction		Describe the explicit clinical problem, biological rationale for the intervention, and rationale for the review		
Methods	Literature Search	Provide description of comprehensive literature search with databases, registers, personal files, expert informants, agencies, hand searching and any restrictions (years considered, publication status, language of publications).		
	Selection Criteria	Describe the inclusion and exclusion criteria with definition of the population, intervention, and outcomes.		
	Method of review or validity assessment	Describe the methodologic quality assessment and the instrument utilized to achieve such an assessment.		
	Data extraction	Describe the process of data extraction and the process utilized (e.g., completed independently, in duplicate, or other methodology utilized).		
	Study characteristics	Describe the type of study design, participants' characteristics, details of intervention, outcome definitions, etc., and how clinical heterogeneity was assessed Describe the reference standard.		
	Quantitative data synthesis	Describe validity, assessment bias and variation, synthesis of relevant study results, descriptive or non-qualitative synthesis, and quantitative synthesis or meta-analysis.		
	Analysis of Evidence	Provide the results of analysis of evidence with strength and level of evidence and if applicable, with grading of recommendations.		

continued on page 951

Table 12 (cont.). *A proposed reporting checklist of systematic review or meta-analysis of diagnostic accuracy studies.*

Heading	Subheading	Descriptor	Reported? (Y/N)	Page number
Results	Trial flow	Provide a systematic review or meta-analysis profile summarizing the trial flow.		
	Study characteristics	Present descriptive data of methodological quality assessment and diagnostic accuracy, including prevalence, confounding factors, study designs, and criterion standard.		
	Quantitative data synthesis	Report agreement on the selection and validity assessment with presentation of simple summary results with statistical analysis descriptions.		
Discussion		Summarize key findings; discuss clinical inferences based on internal and external validity; interpret the results in light of the totality of available evidence; describe potential biases in the review process (e.g., publication bias, selection bias, etc.; describe limitations and weaknesses; and suggest a future research agenda.		

Adapted, revised and modified from Moher et al. Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. Quality of reporting of met-analyses. *Lancet* 1999; 354:1896-1900 (212).

6.3 Introduction

The introduction includes the scientific background and explanation of rationale for any type of systematic review. Typically it includes free-flowing text without a structured format in which the authors explain the scientific background of the clinical problem, biological rationale for the intervention, and rationale for the systematic review. In addition to this, readers will benefit from an appropriate explanation for how the systematic review might work and the research involving people should be based on the thorough knowledge of the scientific literature (299,300). Further, explanation in the introduction section with regards to if the systematic review is limited for the review itself or if the meta-analysis is planned.

6.4 Methods

Methods include literature search, selection, method of review or validity assessment, data extraction, study characteristics, quantitative data synthesis, and analysis of evidence.

6.4.1 Literature Search

The information sources must be described in detail with databases, registers, personal files, expert informants, agencies, hand searching, and any restriction such as years considered, publication status, or language of publications (301-303).

Authors should provide details of the literature search including the search strategy and terminology utilized.

6.4.2 Selection or Inclusion Criteria

The authors should clearly describe the inclusion and exclusion criteria with a definition of the population, intervention, principle outcomes, and outcomes (304).

6.4.3 Methodologic Quality Assessment of Individual Articles

The multiple criteria and processes to the validity or methodologic quality of assessment must be described. These may include appropriate selection criteria of the patients, allocation, quality assessment, the instruments utilized, and the results.

6.4.4 Data Extraction

Data extraction should be described clearly whether it was completed independently or in duplicate.

6.4.5 Study Characteristics

Under this section, the type of study design, participants' characteristics, details of intervention, outcome definitions, and the assessment of clinical heterogeneity must be described.

6.5 Quantitative Data Synthesis

The principle measure of effect is to describe validity, assessment bias and variation, synthesis of relevant study results, descriptive or non-qualitative synthesis, and quantitative synthesis or meta-analysis.

6.6 Results

The results section includes trial flow, study characteristics, and quantitative data synthesis.

6.6.1 Trial Flow

A trial flow figure should be inserted which shows how the literature was searched and inclusion/exclusion criteria were met, this is illustrated in Figure 2.

6.6.2 Study Characteristics

Authors should present descriptive data for each diagnostic accuracy study, along with sample size, intervention, dose, duration, and follow-up periods, etc.

6.6.3 Quantitative Data Synthesis

Results should show the principle measure of effect (method of combining results), statistical testing, and confidence intervals; handling of missing data; the results of statistical heterogeneity; results of subgroup analysis performed; and the results of publication bias if they were assessed.

It may be less complicated for the reviewers if information is provided of agreement on the selection and validity assessment in the form of simple summary results for each group in each trial for each primary outcome; data needed to calculate the effect sizes and confidence intervals; confidence intervals; and tables and means and standard deviations or proportions.

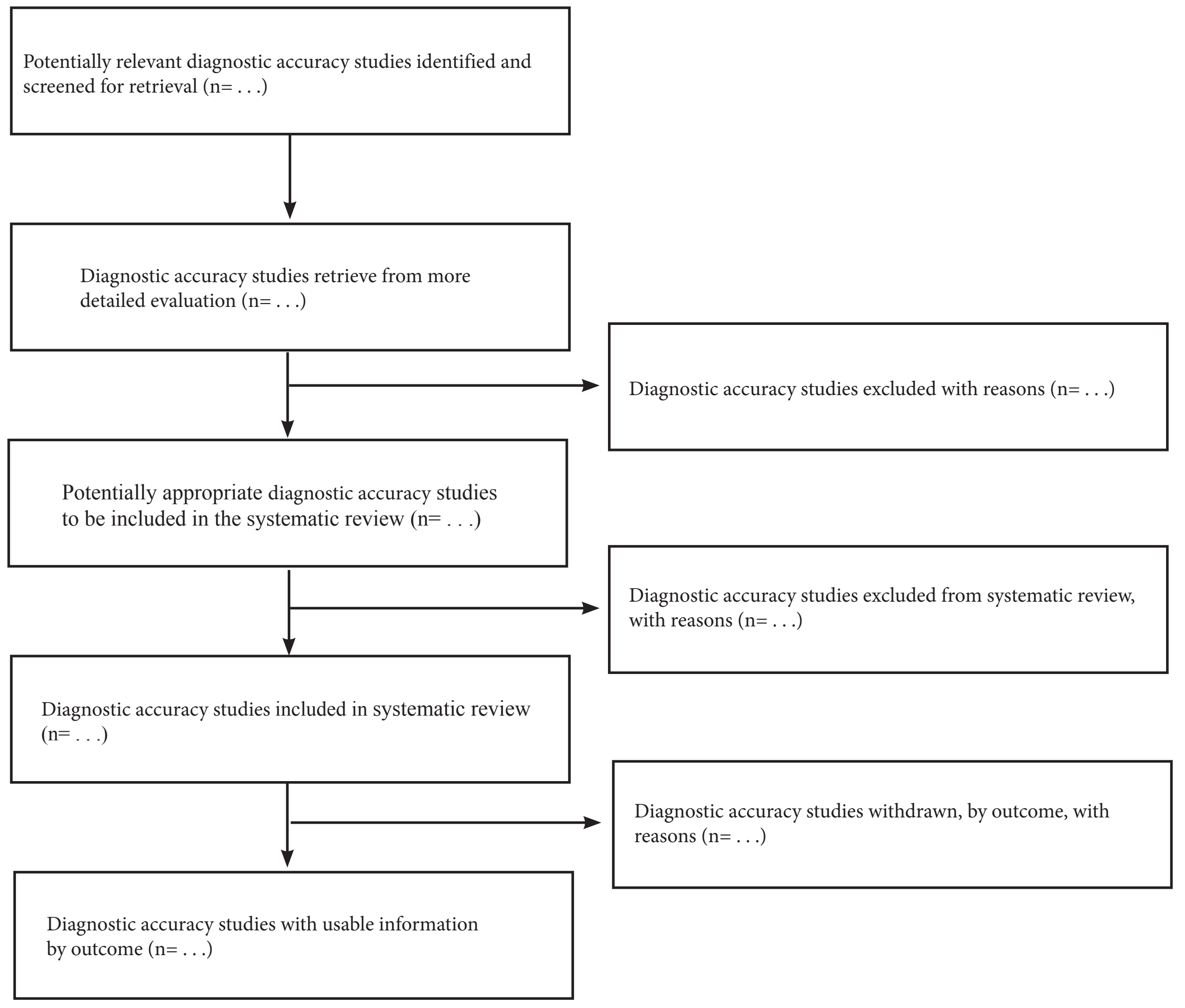

Potentially relevant diagnostic accuracy studies identified and screened for retrieval (n= . . .)
Diagnostic accuracy studies retrieve from more detailed evaluation (n= . . .)
Diagnostic accuracy studies excluded with reasons (n= . . .)
Potentially appropriate diagnostic accuracy studies to be included in the systematic review (n= . . .)
Diagnostic accuracy studies excluded from systematic review, with reasons (n= . . .)
Diagnostic accuracy studies included in systematic review (n= . . .)
Diagnostic accuracy studies withdrawn, by outcome, with reasons (n= . . .)
Diagnostic accuracy studies with usable information by outcome (n= . . .)

Fig. 2. Progress through the stages of asystematic review of observational studies.

6.7 Level of Evidence

Level of evidence may be presented based on the results, however, this is not considered as a requirement either for QUOROM or for MOOSE.

6.8 Recommendations

An additional recommendation is with regards to grading of recommendations which may be provided, however, once again this is not a requirement of either QUOROM or MOOSE. In addition, a cost-effective analysis may also be provided.

6.9 Discussion or Comment

The discussion or comment should summarize key findings; discuss clinical inference based on internal and external validity; interpret the results in light of the totality of available evidence; describe potential biases in the review process such as publication bias; and suggest a future research agenda. Table 13 illustrates the proposed approach for reporting of systematic review of diagnostic accuracy studies. Some journals have encouraged a structured format in reporting the discussion of the results (305) (Table 13) .

Table 13. *Proposed approach for reporting of systematic review of diagnostic accuracy studies.*

1) A brief synopsis of the key findings

2) Consideration of possible mechanisms and explanation

3) Comparison with relevant findings from other published studies

4) Limitations of the present study and methods used to minimize and compensate for those limitations

7.0 Discussion

Diagnosis is a critical component of health care, and clinicians, policy makers, and patients routinely face a range of questions regarding diagnostic tests. Well-designed diagnostic test accuracy studies can help in making these decisions, provided that they transparently and fully report their participants, tests, methods, and results. Reviews performed systematically can assist clinicians, patients, and policy makers.

Interventional pain management is an evolving specialty with multiple limitations in performing diagnostic accuracy studies including the criterion standard without tissue biopsy. However, clinical follow-up can be used for validation of interventional diagnostic studies. Thus, it is essential for the methodologists and clinicians to accurately follow the requirements of EBM in conducting systematic reviews of diagnostic accuracy studies.

Acknowledgments

The authors wish to thank Vidyasagar Pampati, MSc, statistician; Sekar Edem for assistance in the literature search; and Tonie M. Hatton and Diane E. Neihoff, transcriptionists, for their assistance in the preparation of this manuscript.

References

1. Green S, Higgins JPT, Alderson P, Clarke M, Mulrow CD, Oxman AD. Chapter 1: Introduction. In: Higgins JPT, Green S (eds). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.0.1 (updated September 2008), pp 8-13. The Cochrane Collaboration, 2008. Available from www.cochrane-handbook.org.
2. Alexander GC, Stafford RS. Does comparative effectiveness have a comparative edge? *JAMA* 2009; 301:2488-2490.
3. Hartzband P, Groopman J. Keeping the patient in the equation—humanism and health care reform. *N Engl J Med* 2009; 361:554-555.
4. Manchikanti L, Boswell MV, Giordano J. Evidence-based interventional pain management: Principles, problems, potential, and applications. *Pain Physician* 2007; 10:329-356.
5. Manchikanti L. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: Part 1: Introduction and general considerations. *Pain Physician* 2008; 11:161-186.
6. Manchikanti L, Hirsch JA, Smith HS. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: Part 2: Randomized controlled trials. *Pain Physician* 2008; 11:717-773.
7. Manchikanti L, Benyamin RM, Helm S, Hirsch JA. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: Part 3: Systematic reviews and meta-analysis of randomized trials. *Pain Physician* 2009; 12:35-72.
8. Manchikanti L, Singh V, Smith HS, Hirsch JA. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: Part 4: Observational studies. *Pain Physician* 2009; 12:73-108.
9. Manchikanti L, Derby R, Wolfer LR, Singh V, Datta S, Hirsch JA. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: Part 5: Diagnostic accuracy studies. *Pain Physician* 2009; 12:517-540.

10. Manchikanti L, Datta S, Smith HS, Hirsch JA. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: Part 6: Systematic reviews and meta-analyses of observational studies. *Pain Physician* 2009; 12:819-850.

11. Guyatt G, Drummond R. Part 1. The basics: Using the medical literature. 1A. Introduction: The philosophy of evidence-based medicine. In: Guyatt G, Rennie D (eds). *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. AMA Press, Chicago, 2002, pp 3-12.

12. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM; Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Ann Intern Med* 2008; 149:889-897.

13. Devillé WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, Bezemer PD. Conducting systematic reviews of diagnostic studies: Didactic guidelines. *BMC Med Res Methodol* 2002; 2:9.

14. Irwig L, Tosteson AN, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994; 120:667-676.

15. Pai M, McCulloch M, Gorman JD, Pai N, Enanoria W, Kennedy G, Tharyan P, Colford JM Jr. Systematic reviews and meta-analyses: An illustrated, step-by-step guide. *Natl Med J India* 2004; 17:86-95.

16. Hockings RL, McAuley JH, Maher CG. A systematic review of the predictive ability of the Orebro Musculoskeletal Pain Questionnaire. *Spine (Phila Pa 1976)* 2008; 33:E494-E500.

17. Jarvik JG, Deyo RA. Diagnostic evaluation of low back pain with emphasis on imaging. *Ann Intern Med* 2002; 137:586-597.

18. Vlaar AM, van Kroonenburgh MJ, Kessels AG, Weber WE. Meta-analysis of the literature on diagnostic accuracy of SPECT in parkinsonian syndromes. *BMC Neurol* 2007; 7:27.

19. Szadek KM, van der Wurff P, van Tulder MW, Zuurmond WW, Perez RS. Diagnostic validity of criteria for sacroiliac joint pain: A systematic review. *J Pain* 2009; 10:354-368.

20. Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? *BMC Med Res Methodol* 2005; 5:20.

21. Honest H, Khan KS. Reporting of measures of accuracy in systematic reviews of diagnostic literature. *BMC Health Serv Res* 2002; 2:4.

22. Siddiqui MA, Azuara-Blanco A, Burr J. The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. *Br J Ophthalmol* 2005; 89:261-265.

23. Mirkhil S, Kent PM. The diagnostic accuracy of brief screening questions for psychosocial risk factors of poor outcome from an episode of pain: A systematic review. *Clin J Pain* 2009; 25:340-348.

24. Pearson SD, Knudsen AB, Scherer RW, Weissberg J, Gazelle GS. Assessing the comparative effectiveness of a diagnostic technology: CT colonography. *Health Aff (Millwood)* 2008; 27:1503-1514.

25. Henschke N, Maher CG, Refshauge KM. A systematic review identifies five "red flags" to screen for vertebral fracture in patients with low back pain. *J Clin Epidemiol* 2008; 61:110-118.

26. Sehgal N, Shah RV, McKenzie-Brown A, Everett CR. Diagnostic utility of facet (zygapophysial) joint injections in chronic spinal pain: A systematic review of evidence. *Pain Physician* 2005; 8:211-224.

27. Sehgal N, Dunbar EE, Shah RV, Colson JD. Systematic review of diagnostic utility of facet (zygapophysial) joint injections in chronic spinal pain: An update. *Pain Physician* 2007; 10:213-228.

28. Atluri S, Datta S, Falco FJE, Lee M. Systematic review of diagnostic utility and therapeutic effectiveness of thoracic facet joint interventions. *Pain Physician* 2008; 11:611-629.

29. Falco FJE, Erhart S, Wargo BW, Bryce DA, Atluri S, Datta S, Hayek SM. Systematic review of diagnostic utility and therapeutic effectiveness of cervical facet joint interventions. *Pain Physician* 2009; 12:323-344.

30. Datta S, Lee M, Falco FJE, Bryce DA, Hayek SM. Systematic assessment of diagnostic accuracy and therapeutic utility of lumbar facet joint interventions. *Pain Physician* 2009; 12:437-460.

31. McKenzie-Brown AM, Shah RV, Sehgal N, Everett CR. A systematic review of sacroiliac joint interventions. *Pain Physician* 2005; 8:115-125.

32. Hansen HC, McKenzie-Brown AM, Cohen SP, Swicegood JR, Colson JD, Manchikanti L. Sacroiliac joint interventions: A systematic review. *Pain Physician* 2007; 10:165-184.

33. Rupert MP, Lee M, Manchikanti L, Datta S, Cohen SP. Evaluation of sacroiliac joint interventions: A systematic appraisal of the literature. *Pain Physician* 2009; 12:399-418.

34. Shah RV, Everett CR, McKenzie-Brown AM, Sehgal N. Discography as a diagnostic test for spinal pain: A systematic and narrative review. *Pain Physician* 2005; 8:187-209.

35. Buenaventura RM, Shah RV, Patel V, Benyamin RM, Singh V. Systematic review of discography as a diagnostic test for spinal pain: An update. *Pain Physician* 2007; 10:147-164.

36. Manchikanti L, Dunbar EE, Wargo BW, Shah RV, Derby R, Cohen SP. Systematic review of cervical discography as a diagnostic test for chronic spinal pain. *Pain Physician* 2009; 12:305-321.

37. Singh V, Manchikanti L, Shah RV, Dunbar EE, Glaser SE. Systematic review of thoracic discography as a diagnostic test for chronic spinal pain. *Pain Physician* 2008; 11:631-642.

38. Manchikanti L, Glaser S, Wolfer L, Derby R, Cohen SP. Systematic review of lumbar discography as a diagnostic test for chronic low back pain. *Pain Physician* 2009; 12:541-559.

39. Abdi S, Datta S, Lucas LF. Role of epidural steroids in the management of chronic spinal pain: A systematic review of effectiveness and complications. *Pain Physician* 2005; 8:127-143.

40. Abdi S, Datta S, Trescot AM, Schultz DM, Adlaka R, Atluri SL, Smith HS, Manchikanti L. Epidural steroids in the management of chronic spinal pain: A systematic review. *Pain Physician* 2007; 10:185-212.

41. Boswell MV, Singh V, Staats PS, Hirsch JA. Accuracy of precision diagnostic blocks in the diagnosis of chronic spinal pain of facet or zygapophysial joint origin. A systematic review. *Pain Physician* 2003; 6:449-456.

42. Agency for Healthcare Research and Quality, U.S. Department of Health and

Human Services; Prepared by Southern California Evidence-based Practice Center. Assessment of the Need to Update Comparative Effectiveness Reviews: Report of an Initial Rapid Program Assessment (2005–2009). September 22, 2009. http://effectivehealthcare.ahrq.gov/healthInfo.cfm?infotype=rr&ProcessID=65

43. Lind J. *A Treatise of the Scurvy in Three Parts. Containing an Inquiry into the Nature, Causes and Cure of that Disease, together with a Critical and Chronological View of What Has Been Published on the Subject*. A. Millar, London, 1753.

44. Glass GV. Primary, secondary, and meta-analysis of research. *Educ Res* 1976; 5:3-8.

45. Guyatt G, Cairns J, Churchill D, Haynes B, Hirsh J, Irvine J, Levine M, Nishikawa J, Sackett D, Brill-Edwards P, Gerstein H, Gibson J, Jaeschke R, Kerigan A, Neville A, Panju A, Detsky A. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA* 1992; 268:2420-2425.

46. Pearson K. Report on certain enteric fever inoculation statistics. *Br Med J* 1904; 3:1243-1246.

47. Egger M, Smith GD, Altman DG (eds). *Systematic Reviews in Health Care. Meta-Analysis in Context.* BMJ Publishing Group, London, 2001.

48. *Undertaking Systematic Reviews of Research on Effectiveness. CRDs Guidance for Carrying Out or Commissioning Reviews.* CRD Report Number 4 (2nd), CRD Centre for Reviews and Dissemination, University of York, York, UK. March 2001. www.york.ac.uk/inst/crd/report4.htm

49. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis. *J Clin Epidemiol* 1995; 48:167-171.

50. Mulrow C, Langhorne P, Grimshaw J. Integrating heterogeneous pieces of evidence in systematic reviews. *Ann Intern Med* 1997; 127:989-995.

51. Gotzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal anti-inflammatory drugs in rheumatoid arthritis. *Control Clin Trials* 1989; 10:31-56.

52. Wright RW, Brand RA, Dunn W, Spindler KP. How to write a systematic review. *Clin Orthop Relat Res* 2007; 455:23-29.

53. Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-Based Medicine: How to Practice and Teach EBM*. 2nd ed. Churchill Livingstone, Edinburgh, UK, 2000.

54. Crowther MA, Cook DJ. Trials and tribulations of systematic reviews and meta-analyses. *Hematology Am Soc Hematol Educ Program* 2007; 2007:493-497.

55. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. *N Engl J Med* 1987; 316:450-455.

56. Sacks HS, Reitman D, Pagano D, Kupelnick B. Meta-analysis: An update. *Mt Sinai J Med* 1996; 63:216-224.

57. Moher D, Tricco AC. Issues related to the conduct of systematic reviews: A focus on the nutrition field. *Am J Clin Nutr* 2008; 88:1191-1999.

58. Richardson WS, Wilson MS, Nishikawa J, Hayward RSA. The well-built clinical question: A key to evidence based decisions. *ACP J Club* 1995; A12-13.

59. American College of Occupational and Environmental Medicine Low Back Disorders Chapter. In: *Occupational Medicine Practice Guidelines: Evaluation and Management of Common Health Problems and Functional Recovery of Workers,* Second Edition. American College of Occupational and Environmental Medicine, Elk Grove Village, 2007.

60. American College of Occupational and Environmental Medicine. Chronic Pain Chapter (revised 2008). In: *Occupational Medicine Practice Guidelines: Evaluation and Management of Common Health Problems and Functional Recovery of Workers,* Second Edition. American College of Occupational and Environmental Medicine, Elk Grove Village, 2008.

61. Staal JB, de Bie RA, de Vet HC, Hildebrandt J, Nelemans P. Injection therapy for subacute and chronic low back pain: An updated Cochrane review. *Spine (Phila Pa 1976)* 2009; 34:49-59.

62. Taylor RS, Taylor R, Fritzell P. Balloon kyphoplasty and vertebroplasty for vertebral compression fractures: A comparative systematic review of efficacy and safety. *Spine (Phila Pa 1976)* 2006; 31:2747-2755.

63. Taylor RS, Fritzell P, Taylor RJ. Balloon kyphoplasty in the management of vertebral compression fractures: An updated systematic review and meta-analysis. *Eur Spine J* 2007; 16:1085-1100.

64. Hulme PA, Krebs J, Ferguson SJ, Berlemann U. Vertebroplasty and kyphoplasty: A systematic review of 69 clinical studies. *Spine (Phila Pa 1976)* 2006; 31:1983-2001.

65. Felder-Puig R, Piso B, Guba B, Gartlehner G. Orthopade. Kyphoplasty and vertebroplasty for the management of osteoporotic vertebral compression fractures: A systematic review. *Orthopade* 2009; 38:606-615.

66. Gill JB, Kuper M, Chin PC, Zhang Y, Schutt R. Comparing pain reduction following kyphoplasty and vertebroplasty for osteoporotic vertebral compression fractures. *Pain Physician* 2007; 10:583-590.

67. Wolfer L, Derby R, Lee JE, Lee SH. Systematic review of lumbar provocation discography in asymptomatic subjects with a meta-analysis of false-positive rates. *Pain Physician* 2008; 11:513-538.

68. Conn A, Buenaventura R, Datta S, Abdi S, Diwan S. Systematic review of caudal epidural injections in the management of chronic low back pain. *Pain Physician* 2009; 12:109-135.

69. Parr AT, Diwan S, Abdi S. Lumbar interlaminar epidural injections in managing chronic low back and lower extremity pain: A systematic review. *Pain Physician* 2009; 12:163-188.

70. Benyamin RM, Singh V, Parr AT, Conn A, Diwan S, Abdi S. Systematic review of the effectiveness of cervical epidurals in the management of chronic neck pain. *Pain Physician* 2009; 12:137-157.

71. Buenaventura RM, Datta S, Abdi S, Smith HS. Systematic review of therapeutic lumbar transforaminal epidural steroid injections. *Pain Physician* 2009; 12:233-251.

72. Helm S, Hayek S, Benyamin RM, Manchikanti L. Systematic review of the effectiveness of thermal annular procedures in treating discogenic low back pain. *Pain Physician* 2009; 12:207-232.

73. Smith HS, Chopra P, Patel VB, Frey ME, Rastogi R. Systematic review on the role of sedation in diagnostic spinal interventional techniques. *Pain Physician* 2009; 12:195-206.

74. Frey ME, Manchikanti L, Benyamin RM, Schultz DM, Smith HS, Cohen SP. Spinal cord stimulation for patients with failed back surgery syndrome: A systematic review. *Pain Physician* 2009; 12:379-397.

75. Epter RS, Helm S, Hayek SM, Benyamin RM, Smith HS, Abdi S. Systematic review of percutaneous adhesiolysis and management of chronic low back pain

in post lumbar surgery syndrome. *Pain Physician* 2009; 12:361-378.

76. Patel VB, Manchikanti L, Singh V, Schultz DM, Hayek SM, Smith HS. Systematic review of intrathecal infusion systems for long-term management of chronic non-cancer pain. *Pain Physician* 2009; 12:345-360.

77. Hayek SM, Helm S, Benyamin RM, Singh V, Bryce DA, Smith HS. Effectiveness of spinal endoscopic adhesiolysis in post lumbar surgery syndrome: A systematic review. *Pain Physician* 2009; 12:419-435.

78. Hirsch JA, Singh V, Falco FJE, Benyamin RM, Manchikanti L. Automated percutaneous lumbar discectomy for the contained herniated lumbar disc: A systematic assessment of evidence. *Pain Physician* 2009; 12:601-620.

79. Manchikanti L, Boswell MV, Singh V, Derby R, Fellows B, Falco FJE, Datta S, Smith HS, Hirsch JA. Comprehensive review of neurophysiologic basis and diagnostic interventions in managing chronic spinal pain. *Pain Physician* 2009; 12:E71-E120.

80. Manchikanti L, Boswell MV, Datta S, Fellows B, Abdi S, Singh V, Benyamin RM, Falco FJE, Helm S, Hayek S, Smith HS. Comprehensive review of therapeutic interventions in managing chronic spinal pain. *Pain Physician* 2009: 12:E123-E198.

81. Datta S, Everett CR, Trescot AM, Schultz DM, Adlaka R, Abdi S, Atluri SL, Smith HS, Shah RV. An updated systematic review of diagnostic utility of selective nerve root blocks. *Pain Physician* 2007; 10:113-128.

82. Trescot AM, Chopra P, Abdi S, Datta S, Schultz DM. Systematic review of effectiveness and complications of adhesiolysis in the management of chronic spinal pain: An update. *Pain Physician* 2007; 10:129-146.

83. Boswell MV, Colson JD, Sehgal N, Dunbar E, Epter R. A systematic review of therapeutic facet joint interventions in chronic spinal pain. *Pain Physician* 2007; 10:229-253.

84. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC; Standards for Reporting of Diagnostic Accuracy. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med* 2003; 138:40-44.

85. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, Bouter LM, de Vet HC. The quality of diagnostic accuracy studies since the STARD statement: Has it improved? *Neurology* 2006; 67:792-797.

86. Light RJ, Pillemer DB. *Summing Up: The Science of Reviewing Research*. Harvard University Press, Cambridge, 1984.

87. Mulrow CD. The medical review article: State of the science. *Ann Intern Med* 1987; 106:485-488.

88. Higgins JPT, Green S. Introduction. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 (updated September 2006), Section 1. In: *The Cochrane Library*, Issue 4, 2006. John Wiley & Sons, Ltd., Chichester, UK.

89. Mulrow CD. Rationale for systematic reviews. *BMJ* 1994; 309:597-599.

90. Mulrow C, Cook D (eds). *Systematic Reviews: Synthesis of Best Evidence for Health Care Decisions*. ACP Press, Philadelphia, 1998.

91. Mulrow C. Rationale for systematic reviews. In: Chalmers I, Altman D (eds). *Systematic Reviews*, BMJ Books, London, 1995, pp 1-9.

92. Chalmers TC, Matta RJ, Smith H, Kunzler AM. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med* 1977; 297:1091-1096.

93. Juni, P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999; 282:1054-1060.

94. Shulz K, Chalmers I, Hayes R, Altman D. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995; 273:408-412.

95. Moher D, Pham B, Jones A, Cook D, Jadad A, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet* 1998 352:609-613.

96. Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database Syst Rev* 2007; (2):MR000012.

97. Petrie A, Bulman JS, Osborn JF. Further statistics in dentistry Part 8: Systematic reviews and meta-analyses. *Br Dent J* 2003; 194:73-78.

98. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Reitsma JB, Bossuyt PM, Bouter LM, de Vet HC. Quality of reporting of diagnostic accuracy studies. *Radiology* 2005; 235:347-353.

99. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, Bouter LM, de Vet HC. Reproducibility of the STARD checklist: An instrument to assess the quality of reporting of diagnostic accuracy studies. *BMC Med Res Methodol* 2006; 6:12.

100. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt pm. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999; 282:1061-1066.

101. Jaeschke R, Guyatt GH, Sackett DL. Users' guidelines to the medical literature, III: How to use an article about a diagnostic test, A: Are the results of the study valid? The Evidence-Based Medicine Working Group. *JAMA* 1994; 271:389-391.

102. Jaeschke R, Guyatt GH, Sackett DL. Users' guidelines to the medical literature, III: How to use an article about a diagnostic test, B: What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA* 1994; 271:703-707.

103. Greenhalgh T. How to read a paper: Papers that report diagnostic or screening tests. *BMJ* 1997; 315:540-543.

104. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: Getting better but still not good. *JAMA* 1995, 274:645-651.

105. Devillé WL, Buntinx F. Didactic Guidelines for Conducting Systematic Reviews of Studies Evaluating the Accuracy of Diagnostic Tests. In: Knottnerus A (ed). *The Evidence Base of Clinical Diagnosis.* BMJ Books, London, 2002, pp 145-165.

106. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests *BMJ* 2001; 323:157-162.

107. Nelemans PJ, Leiner T, de Vet HC, van Engelshoven JM. Peripheral arterial disease: Meta-analysis of the diagnostic performance of MR angiography. *Radiology* 2000; 217:105-114.

108. Devries SO, Hunink MGM, Polak JF. Summary receiver operating characteristic curves as a technique for meta-analysis of the diagnostic performance of duplex ultrasonography in peripheral arterial disease. *Acad Radiol* 1996; 3:361-369.

109. Carragee EJ. Validity of self-reported history in patients with acute back and neck pain after motor vehicle accident. *Spine J* 2008; 8:311-319.

110. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. *Can Med Assoc J* 1986; 134:587-594.

111. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* 1991; 11:88-94.

112. Kent DL, Larson EB. Disease, level of impact, and quality of research methods. Three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. *Invest Radiol* 1992; 27:245-254.

113. Rutjes AW, Reitsma JB, Di Nisio M, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 2006; 174:469-476.

114. Shunmugam M, Azuara-Blanco A. The quality of reporting of diagnostic accuracy studies in glaucoma using the Heidelberg retina tomograph. *Invest Ophthalmol Vis Sci* 2006; 47:2317-2323.

115. Rubinstein SM, van Tulder M. A best-evidence review of diagnostic procedures for neck and low-back pain. *Best Pract Res Clin Rheumatol* 2008; 22:471-482.

116. Hancock MJ, Maher CG, Latimer J, Spindler MF, McAuley JH, Laslett M, Bogduk N. Systematic review of tests to identify the disc, SIJ or facet joint as the source of low back pain. *Eur Spine J* 2007; 16:1539-1550.

117. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: A systematic review. *Ann Intern Med* 2004; 140:189-202.

118. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: Literature survey. *BMJ* 2006; 332:1127-1129.

119. Nordin M, Carragee EJ, Hogg-Johnson S, Weiner SS, Hurwitz EL, Peloso PM, Guzman J, Van Der Velde G, Carroll LJ, Holm LW, Cote P, Cassidy JD, Haldeman S. Assessment of neck pain and its associated disorders: Results of the Bone and Joint Decade 2000–2010 Task Force on Neck Pain and its Associated Disorders. *Spine (Phila Pa 1976)* 2008; 33:S101-S122.

120. Boswell MV, Trescot AM, Datta S, Schultz DM, Hansen HC, Abdi S, Sehgal N, Shah RV, Singh V, Benyamin RM, Patel VB, Buenaventura RM, Colson JD, Cordner HJ, Epter RS, Jasper JF, Dunbar EE, Atluri SL, Bowman RC, Deer TR, Swicegood JR, Staats PS, Smith HS, Burton AW, Kloth DS, Giordano J, Manchikanti L. Interventional techniques: Evidence-based practice guidelines in the management of chronic spinal pain. *Pain Physician* 2007; 10:7-111.

121. Bogduk N, McGuirk B. Causes and sources of chronic low back pain. In: *Medical Management of Acute and Chronic Low Back Pain. An Evidence-Based Approach: Pain Research and Clinical Management,* Vol. 13. Elsevier Science BV, Amsterdam, 2002, pp 115-126.

122. Bogduk N, McGuirk B. An algorithm for precision diagnosis. In: Bogduk N, McGuirk B, (eds). *Medical Management of Acute and Chronic Low Back Pain. An Evidence-Based Approach: Pain Research and Clinical Management*. Elsevier Science BV, Amsterdam, 2002; 13:177-186.

123. Deyo RA, Weinstein JN. Low back pain. *N Engl J Med* 2001; 344:363-370.

124. Deyo RA. Fads in the treatment of low back pain. *N Engl J Med* 1991; 325:1039-1040.

125. Deyo RA, Rainville J, Kent DL. What can the history and physical examination tell us about low back pain? *JAMA* 1992; 268:760-765.

126. van Tulder MWV, Koes BW, Bouter LM. Conservative treatment of acute and chronic nonspecific low back pain. A systematic review of randomized controlled trials of the most common interventions. *Spine (Phila Pa 1976)* 1997; 22:2128-2156.

127. Cohen SP, Argoff CE, Carragee EJ. Management of low back pain. *BMJ* 2008; 338:100-106.

128. Bogduk N. Low back pain. In: *Clinical Anatomy of Lumbar Spine and Sacrum*, 4th edition. Churchill Livingstone, New York, 2005, pp 183-216.

129. Nachemson AL. The natural course of low back pain. In: White A, Gordon SL (eds). *Symposium on Idiopathic Low Back Pain*. Mosby, St. Louis, 1982, pp 46-51.

130. Mooney V. Where is the pain coming from? *Spine (Phila Pa 1976)* 1987; 12:754-759.

131. Spitzer WO, LeBlanc FE, Dupuis M. Scientific approach to the assessment and management of activity-related spinal disorders: A monograph for clinicians. Report of Quebec Task Force on Spinal Disorders. *Spine (Phila Pa 1976)* 1987; 12:S1-S59.

132. Bogduk N, McGuirk B. Sources and causes of neck pain. In: *Management of Acute and Chronic Neck Pain. An Evidence-Based Approach*. Elsevier, 2006, pp 9-20.

133. Bogduk N, McGuirk B. Acute neck pain: Algorithm for acute neck pain. In: *Management of Acute and Chronic Neck Pain. An Evidence-Based Approach*. Elsevier, 2006, pp 69-77.

134. Bogduk N, Lord S. Cervical zygapophysial joint pain. *Neurosurg Q* 1998; 8:107-117.

135. Manchikanti L, Singh V, Pampati V, Damron K, Barnhill R, Beyer C, Cash K. Evaluation of the relative contributions of various structures in chronic low back pain. *Pain Physician* 2001; 4:308-316.

136. Manchukonda R, Manchikanti KN, Cash KA, Pampati V, Manchikanti L. Facet joint pain in chronic spinal pain: An evaluation of prevalence and false-positive rate of diagnostic blocks. *J Spinal Disord Tech* 2007; 20:539-545.

137. Manchikanti L, Manchukonda R, Pampati V, Damron KS, McManus CD. Prevalence of facet joint pain in chronic low back pain in postsurgical patients by controlled comparative local anesthetic blocks. *Arch Phys Med Rehabil* 2007; 88:449-455.

138. Manchikanti L, Boswell MV, Singh V, Pampati V, Damron KS, Beyer CD. Prevalence of facet joint pain in chronic spinal pain of cervical, thoracic, and lumbar regions. *BMC Musculoskelet Disord* 2004; 5:15.

139. Manchikanti L, Singh V, Pampati V, Beyer CD, Damron KS. Evaluation of the prevalence of facet joint pain in chronic thoracic pain. *Pain Physician* 2002; 5:354-359.

140. Schwarzer AC, Aprill CN, Derby R, Fortin J, Kine G, Bogduk N. Clinical features of patients with pain stemming from the lumbar zygapophysial joints. Is the lumbar facet syndrome a clinical entity? *Spine (Phila Pa 1976)* 1994; 19:1132-1137.

141. Cohen SP, Larkin TM, Barna SA, Palmer WE, Hecht AC, Stojanovic MP. Lumbar discography: A comprehensive review of outcome studies, diagnostic accuracy, and principles. *Reg Anesth Pain Med* 2005; 30:163-183.

142. Schwarzer AC, Aprill CN, Derby R, Fortin J, Kine G, Bogduk N. The relative contributions of the disc and zygapophyseal joint in chronic low back pain. *Spine (Phila Pa 1976)* 1994; 19:801-806.

143. Irwin RW, Watson T, Minick RP, Ambrosius WT. Age, body mass index, and gender differences in sacroiliac joint pathology. *Am J Phys Med Rehabil* 2007; 86:37-44.

144. Manchikanti L, Boswell MV, Singh V, Benyamin RM, Fellows B, Abdi S, Buenaventura RM, Conn A, Datta S, Derby R, Falco FJE, Erhart S, Diwan S, Hayek SM, Helm S, Parr AT, Schultz DM, Smith HS, Wolfer LR, Hirsch JA. Comprehensive evidence-based guidelines for interventional techniques in the management of chronic spinal pain. *Pain Physician* 2009: 12:699-802.

145. Carragee EJ, Haldeman S, Hurwitz E. The pyrite standard: The Midas touch in the diagnosis of axial pain syndromes. *Spine J* 2007; 7:27-31.

146. Bogduk N. In defense of King et al: The validity of manual examination in assessing patients with neck pain. *Spine J* 2007; 7:749-752; author reply (Carragee EJ) 752-753.

147. O'Neill C, Owens D. Lumbar facet joint pain: Time to hit the reset button. *Spine J* 2009; 9:619-622.

148. Cohen SP, Stojanovic MP, Crooks M, Kim P, Schmidt RK, Shields CH, Croll S, Hurley RW. Lumbar zygapophysial (facet) joint radiofrequency denervation success as a function of pain relief during diagnostic medial branch blocks: A multicenter analysis. *Spine J* 2008; 8:498-504.

149. Cohen SP, Bajwa ZH, Kraemer JJ, Dragovich A, Williams KA, Stream J, Sireci A, McKnight G, Hurley RW. Factors predicting success and failure for cervical facet radiofrequency denervation: A multi-center analysis. *Reg Anesth Pain Med* 2007; 32:495-503.

150. Manchikanti L, Singh V. Are the results of a multicenter analysis of radiofrequency denervation success as a function of single diagnostic block reliable? *Spine J* 2009; 9:704-705.

151. Manchikanti L, Singh V. Diagnosis of facet joint pain and prediction of success and failure for cervical radiofrequency denervation. *Reg Anesth Pain Med* 2009; 34:81-82.

152. Carragee EJ, Tanner CM, Khurana S, Hayward C, Welsh J, Date E, Truong T, Rossi M, Hagle C. The rates of false-positive lumbar discography in select patients without low back symptoms. *Spine (Phila Pa 1976)* 2000; 25:1373-1380.

153. Carragee EJ, Chen Y, Tanner CM, Hayward C, Rossi M, Hagle C. Can discography cause long-term back symptoms in previously asymptomatic subjects? *Spine (Phila Pa 1976)* 2000; 25:1803-1808.

154. Carragee EJ, Alamin TF, Miller J, Grafe M. Provocative discography in volunteer subjects with mild persistent low back pain. *Spine J* 2002; 2:25-34.

155. de Graaf I, Prak A, Bierma-Zeinstra S, Thomas S, Peul W, Koes B. Diagnosis of lumbar spinal stenosis: A systematic review of the accuracy of diagnostic tests. *Spine (Phila Pa 1976)* 2006; 31:1168-1176.

156. Kent DL, Haynor DR, Larson EB, Deyo RA. Diagnosis of lumbar spinal stenosis in adults: A meta-analysis of the accuracy of CT, MR, and myelography. *AJR Am J Roentgenol* 1992; 158:1135-1144.

157. Devillé WL, van der Windt DA, Dzaferagi A, Bezemer PD, Bouter LM. The test of Lasègue: Systematic review of the accuracy in diagnosing herniated discs. *Spine (Phila Pa 1976)* 2000; 25:1140-1147.

158. Barnsley L, Lord S, Bogduk N. Comparative local anesthetic blocks in the diagnosis of cervical zygapophysial joints pain. *Pain* 1993; 55:99-106.

159. Lord SM, Barnsley L, Bogduk N. The utility of comparative local anesthetic blocks versus placebo-controlled blocks for the diagnosis of cervical zygapophysial joint pain. *Clin J Pain* 1995; 11:208-213.

160. Manchikanti L, Singh V, Pampati V. Are diagnostic lumbar medial branch blocks valid? Results of 2-year follow up. *Pain Physician* 2003; 6:147-153.

161. Pampati S, Cash KA, Manchikanti L. Accuracy of diagnostic lumbar facet joint nerve blocks: A 2-year follow-up of 152 patients diagnosed with controlled diagnostic blocks. *Pain Physician* 2009; 12:855-866.

162. Manchikanti L, Pampati V, Singh V, Beyer CD, Damron KS, Fellows B. Evaluation of the role of facet joints in persistent low back pain in obesity: A controlled, prospective, comparative evaluation. *Pain Physician* 2001; 4:266-272.

163. Manchikanti L, Pampati V, Damron KS, McManus CD, Jackson SD, Barnhill RC, Martin JC. A randomized, prospective, double-blind, placebo-controlled evaluation of the effect of sedation on diagnostic validity of cervical facet joint pain. *Pain Physician* 2004; 7:301-309.

164. Manchikanti L, Damron KS, Rivera J, McManus CD, Jackson SD, Barnhill RC, Martin JC. Evaluation of effect of sedation as a confounding factor in the diagnostic validity of lumbar facet joint pain: A prospective, randomized, double-blind, placebo-controlled evaluation. *Pain Physician* 2004; 7:411-417.

165. Manchikanti L, Pampati V, Damron KS, McManus CD, Jackson SD, Barnhill RC, Martin JC. The effect of sedation on diagnostic validity of facet joint nerve blocks: An evaluation to assess similarities in population with involvement in cervical and lumbar regions. *Pain Physician* 2006; 9:47-52.

166. Manchikanti L, Cash KA, Pampati V, Fellows B. Influence of psychological variables on the diagnosis of facet joint involvement in chronic spinal pain. *Pain Physician* 2008; 11:145-160.

167. Manchikanti L, Pampati V, Fellows B, Rivera JJ, Damron KS, Beyer CD, Cash KA. Influence of psychological factors on the ability to diagnose chronic low back pain of facet joint origin. *Pain Physician* 2001; 4:349-357.

168. Manchikanti L, Manchikanti K, Cash KA, Singh V, Giordano J. Age-related prevalence of facet joint involvement in chronic neck and low back pain. *Pain Physician* 2008; 11:67-75.

169. Manchikanti L, Manchikanti K, Pampati V, Brandon D, Giordano J. The prevalence of facet joint-related chronic neck pain in postsurgical and non-postsurgical patients: A comparative evaluation. *Pain Pract* 2008; 8:5-10.

170. Yin W, Bogduk N. The nature of neck pain in a private pain clinic in the United States. *Pain Med* 2008; 9:196-203.

171. Manchikanti L, Singh V, Fellows B, Pampati V. Evaluation of influence of gender, occupational injury, and smoking on chronic low back pain of facet joint origin: A subgroup analysis. *Pain Physician* 2002; 5:30-35.

172. Manchikanti L, Singh V, Falco FJE, Cash KA, Pampati V. Effectiveness of thoracic medial branch blocks in managing chronic pain: A preliminary report of a randomized, double-blind controlled trial; Clinical trial NCT00355706. *Pain Physician* 2008; 11:491-504.

173. Manchikanti L, Singh V, Falco FJ, Cash KA, Fellows B. Cervical medial branch blocks for chronic cervical facet joint pain: A randomized double-blind, controlled trial with one-year follow-up. *Spine (Phila Pa 1976)* 2008; 33:1813-1820.

174. Manchikanti L, Singh V, Falco FJ, Cash KA, Pampati V. Lumbar facet joint nerve blocks in managing chronic facet joint pain: One-year follow-up of a randomized, double-blind controlled trial: Clinical Trial NCT00355914. *Pain Physician* 2008; 11:121-132.

175. Manchikanti L, Damron KS, Cash KA,

Manchukonda R, Pampati V. Therapeutic cervical medial branch blocks in managing chronic neck pain: A preliminary report of a randomized, double-blind, controlled trial: Clinical Trial NCT0033272. *Pain Physician* 2006; 9:333-346.

176. Nath S, Nath CA, Pettersson K. Percutaneous lumbar zygapophysial (facet) joint neurotomy using radiofrequency current, in the management of chronic low back pain: A randomized double blind trial. *Spine (Phila Pa 1976)* 2008; 33:1291-1298.

177. Lord S, Barnsley L, Wallis B, McDonald G, Bogduk N. Percutaneous radiofrequency neurotomy for chronic cervical zygapophyseal-joint pain. *N Engl J Med* 1996; 335:1721-1726.

178. Pham Dang C, Lelong A, Guilley J, Nguyen JM, Volteau C, Venet G, Perrier C, Lejus C, Blanloeil Y. Effect on neurostimulation of injectates used for perineural space expansion before placement of a stimulating catheter: Normal saline versus dextrose 5% in water. *Reg Anesth Pain Med* 2009; 34:398-403.

179. Tsui BCH, Wagner, A, Finucane B. Electrophysiologic effect of injectates on peripheral nerve stimulation. *Reg Anesth Pain Med* 2004; 29:189-193.

180. Pham Dang C, Guilley J, Dernis L, Langlois C, Lambert C, Nguyen JM, Pinaud M. Is there any need for expanding the perineural space before catheter placement in continuous femoral nerve blocks? *Reg Anesth Pain Med* 2006; 31:393-400.

181. Pham Dang C, Kick O, Collet T, Gouinn F, Pinaud M. Continuous peripheral nerve blocks with stimulating catheters. *Reg Anesth Pain Med* 2003; 28:83-88.

182. Tsui BC, Kropelin B. The electrophysiologic effect of dextrose 5% in water on single-shot peripheral nerve stimulation. *Anesth Analg* 2005; 100:1837-1839.

183. Levin JH. Prospective, double-blind, randomized placebo-controlled trials in interventional spine: What the highest quality literature tells us. *Spine J* 2009; 9:690-703.

184. Manchikanti L, Cash KA, McManus CD, Pampati V, Smith HS. Preliminary results of randomized, equivalence trial of fluoroscopic caudal epidural injections in managing chronic low back pain: Part 1. Discogenic pain without disc herniation or radiculitis. *Pain Physician* 2008; 11:785-800.

185. Manchikanti L, Singh V, Cash KA, Pampati V, Damron KS, Boswell MV. Preliminary results of randomized, equivalence trial of fluoroscopic caudal epidural injections in managing chronic low back pain: Part 2. Disc herniation and radiculitis. *Pain Physician* 2008; 11:801-815.

186. Manchikanti L, Singh V, Cash KA, Pampati V, Datta S. Preliminary results of randomized, equivalence trial of fluoroscopic caudal epidural injections in managing chronic low back pain: Part 3. Post surgery syndrome. *Pain Physician* 2008; 11:817-831.

187. Manchikanti L, Cash KA, McManus CD, Pampati V, Abdi S. Preliminary results of randomized, equivalence trial of fluoroscopic caudal epidural injections in managing chronic low back pain: Part 4. Spinal stenosis. *Pain Physician* 2008; 11:833-848.

188. Petticrew M, Song F, Wilson P, Wright K. Quality-assessed review of health care interventions and the Database of Abstracts of Reviews of Effectiveness (DARE). *Int J Technol Assess Health Care* 1999; 15:671-678.

189. Jadad AR, Cook DJ, Jones A, Klassen TP, Tugwell P, Moher M, Moher D. Methodology and reports of systematic reviews and meta-analyses: A comparison of Cochrane reviews with articles published in paper-based journals. *JAMA* 1998; 280:278-280.

190. Stengel D, Bauwens K, Rademacher G, Mutze S, Ekkernkamp A. Association between compliance with methodological standards of diagnostic research and reported test accuracy: Meta-analysis of focused assessment of US for trauma. *Radiology* 2005; 236:102-111.

191. Whiting P, Rutjes A, Reitsma J, Bossuyt P, Kleijnen J. The Development of QUADAS: A tools for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003; 3:25.

192. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. The selection of diagnostic tests. In: *Clinical Epidemiology, A Basic Science for Clinical Medicine*, 2nd ed. Little, Brown and Company, Boston, 1991, pp 47-57.

193. Vamvakas EC. Meta-analyses of studies of the diagnostic accuracy of laboratory tests. A review of the concepts and methods. *Arch Pathol Lab Med* 1998; 122:675-686.

194. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. *J Clin Epidemiol* 1995; 48:119-130.

195. Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. *Stat Methods Med Res* 1998; 7:337-353.

196. Hui SL, Zhou XH. Evaluation of diagnostic tests without gold standards. *Stat Methods Med Res* 1998; 7:354-370.

197. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* 2007; 11:ix-51.

198. Oxman AD. Systematic reviews: Checklists for review articles. *BMJ* 1994; 309:648-651.

199. McAlister FA, Clark HD, van Walraven C, Straus SE, Lawson FM, Moher D, Mulrow CD. The medical review article revisited: Has the science improved? *Ann Intern Med* 1999; 131:947-951.

200. Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. Evidence-Based Medicine Working Group. *JAMA* 1994; 272:1367-1371.

201. Altman DG. Confidence intervals in research evaluation. *Ann Intern Med* 1991; 116:A28.

202. West S, King V, Carey TS, Lohr KN, McKoy N, Sutton SF, Lux L. *Systems to Rate the Strength of Scientific Evidence*, Evidence Report, Technology Assessment No. 47. AHRQ Publication No. 02-E016. Rockville, MD: Agency for Healthcare Research and Quality, 2002. www.thecre.com/pdf/ahrq-system-strength.pdf

203. Proposed Evaluation Tools for COMPUS [www.ccohta.ca/compus/compus_pdfCOMPUS_Evaluation_Methodology_draft_e.pdf]. Canadian Coordinating Office for Health Technology Assessment, Ottawa, November 29, 2005.

204. Lavis JN, Davies HTO, Oxman AD, Denis JL, Golden-Biddle K, Ferlie E. Towards systematic reviews that inform healthcare management and policy making. *Journal of Health Services Research and Policy* 2005; 10:35-48.

205. Glenton C, Underland V, Kho M, Pennick V, Oxman AD. Summaries of findings, descriptions of interventions, and information about adverse effects would make reviews more informative. *J Clin Epidemiol* 2006; 59:770-778.

206. Glasziou P, Oxman AD, Higgins J. Summary of Findings Tables within Cochrane Reviews: Draft Specification for Rev Man 5.0. December 2004. *Obtaining a Consensus on the Content and Methods of a Summary of Findings Table for Cochrane Reviews. Report to the Cochrane Collaboration Steering*

Group, 2005.

207. Oxman AD, Schünemann HJ, Fretheim A. Improving the use of research evidence in guideline development: 8. Synthesis and presentation of evidence. *Health Res Policy Syst* 2006; 4:20.

208. Auperin A, Pignon JP, Poynard T. Review article: Critical review of meta-analyses of randomized clinical trials in hepatogastroenterology. *Alimentary Pharmacol Ther* 1997, 11:215-225.

209. Khan KS, Ter Riet G, Glanville J, Sowden AJ, Kleijnen J. *Undertaking Systematic Reviews of Research on Effectiveness. CRD's Guidance for Carrying Out or Commissioning Reviews*. CRD Centre for Reviews and Dissemination, University of York, York, UK, 2000.

210. Barnes DE, Bero LA. Why review articles on the health effects of passive smoking reach different conclusions. *JAMA* 1998; 279:1566-1570.

211. Oxman AD, Guyatt GH, Singer J, Goldsmith CH, Hutchison BG, Milner RA, Streiner DL. Agreement among reviewers of review articles. *J Clin Epidemiol* 1991; 44:91-98.

212. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. Quality of reporting of met-analyses. *Lancet* 1999; 354:1896-1900.

213. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *JAMA* 1992; 268:240-248.

214. Higgins JPT, Green S (eds). Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 (updated September 2006). In: *The Cochrane Library*, Issue 4, 2006, John Wiley & Sons, Ltd. Chichester, UK.

215. Higgins JPT, Green S (eds). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.0.1 (updated September 2008). The Cochrane Collaboration, 2008. www.cochrane-handbook.org.

216. Glasziou P, Irwig L, Bain C, Colditz G. *Systematic Reviews in Health Care. A Practical Guide*. University Press, Cambridge, 2001.

217. Higgins JPT, Green S (eds). Formulating the problem. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 (updated September 2006), Section 4. In: *The Cochrane Library*, Issue 4, 2006. John Wiley & Sons, Ltd., Chichester, UK.

218. Jackson GB. Methods for integrative reviews. *Rev Educ Res* 1980; 50:438-460.

219. Cooper HM. The problem formulation stage. In: Cooper HM (ed). *Integrating Research. A Guide for Literature Reviews*. Sage Publications, Newbury Park, 1984, pp 19-37.

220. Hedges LV. Statistical considerations. In: Cooper H, Hedges LV (eds). *The Handbook of Research Synthesis*. Russell Sage Foundation, New York, 1994; 29-38.

221. Knottnerus JA. *The Evidence Base of Clinical Diagnosis*. BMJ Books, London, 2002.

222. Bossuyt PM. Interpreting diagnostic test accuracy studies. *Semin Hematol* 2008; 45:189-195.

223. Zhou X-H, Obuchowski N, McClish D. *Statistical Methods in Diagnostic Medicine*. Hoboken, NJ: J Wiley; 2002.

224. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. *Ann Intern Med* 1997; 127:380-387.

225. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: Assessing new tests against existing diagnostic pathways. *BMJ* 2006; 332:1089-1092.

226. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? *Ann Intern Med* 2006; 144:850-855.

227. Thornbury JR. Eugene W. Caldwell Lecture. Clinical efficacy of diagnostic imaging: Love it or leave it. *AJR Am J Roentgenol* 1994; 162:1-8.

228. Glanville J. Stage II — Conducting the review, Phase 3 — Identification of research. In: *Undertaking Systematic Reviews of Research on Effectiveness.* CRDs guidance for carrying out or commissioning reviews. CRD Report Number 4 (2nd), CRD Centre for Reviews and Dissemination, University of York, York, UK. March 2001. www.york.ac.uk/inst/crd/report4.htm

229. Goodman C. *Step 3: Formulate Plan for Literature Search and Step 4: Conduct Literature Search and Retrieval.* Swedish Council on Technology Assessment in Health Care, 1993.

230. Clarke M, Oxman A. Section 5. Locating and selecting studies. In: *Cochrane Reviewers' Handbook*. 4.1 ed. Cochrane Collaboration, Oxford, 2000.

231. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. In: Mulrow C, Cook D (eds). *Systematic Reviews. Synthesis of Best Evidence for Health Care Decisions.* ACP Press, Philadelphia, 1998, pp. 67-79.

232. Higgins JPT, Green S (eds). Locating and selecting studies. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 (updated September 2006); Section 5. In: *The Cochrane Library*, Issue 4, 2006. John Wiley & Sons, Ltd., Chichester, UK.

233. Royle P, Waugh N. Literature searching for clinical and cost effectiveness studies used in health technology assessment reports carried out for the National Institute for Clinical Excellence appraisal system. *Health Technology Assessment* 2003, 7:www.ncchta.org/execsumm/summ734.htm.

234. Manchikanti L, Singh V, Derby R, Helm S, Trescot AM, Staats PS, Prager JP, Hirsch JA. Review of occupational medicine practice guidelines for interventional pain management and potential implications. *Pain Physician* 2008; 11:271-289.

235. Manchikanti L, Singh V, Helm S, Trescot AM, Hirsch JA. A critical appraisal of 2007 American College of Occupational and Environmental Medicine (ACOEM) practice guidelines for interventional pain management: An independent review utilizing AGREE, AMA, IOM, and other criteria. *Pain Physician* 2008; 11:291-310.

236. Manchikanti L, Singh V, Derby R, Schultz DM, Benyamin RM, Prager JP, Hirsch JA. Reassessment of evidence synthesis of occupational medicine practice guidelines for interventional pain management. *Pain Physician* 2008; 11:393-482.

237. Dennison PL. *Official Disability Guidelines*. 13th Ed. Work Loss Data Institute, 2008.

238. Staal JB, de Bie R, de Vet HC, Hildebrandt J, Nelemans P. Injection therapy for subacute and chronic low-back pain. *Cochrane Database Syst Rev* 2008; 3:CD001824.

239. Peloso PMJ, Gross A, Haines T, Trinh K, Goldsmith CH, Burnie SJ, Cervical Overview Group. Medicinal and injection therapies for mechanical neck disorders. *Cochrane Database Syst Rev* 2007; 3:CD000319.

240. Armon C, Argoff CE, Samuels J, Backonja MM; Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. As-

sessment: Use of epidural steroid injections to treat radicular lumbosacral pain: Report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. *Neurology* 2007; 68:723-729.

241. Manchikanti L, Jasper J, Singh V. Cochrane Review by Nelemans et al. *Spine (Phila Pa 1976)* 2001; 26:2641-2643.

242. Chou R, Loeser JD, Owens DK, Rosenquist RW, Atlas SJ, Baisden J, Carragee EJ, Grabois M, Murphy DR, Resnick DK, Stanos SP, Shaffer WO, Wall EM; American Pain Society Low Back Pain Guideline Panel. Interventional therapies, surgery, and interdisciplinary rehabilitation for low back pain: An evidence-based clinical practice guideline from the American Pain Society. *Spine (Phila Pa 1976)* 2009; 34:1066-1077.

243. Carragee EJ, Hurwitz EL, Cheng I, Carroll LJ, Nordin M, Guzman J, Peloso P, Holm LW, Côté P, Hogg-Johnson S, van der Velde G, Cassidy JD, Haldeman S; Bone and Joint Decade 2000–2010 Task Force on Neck Pain and Its Associated Disorders. Treatment of neck pain: Injections and surgical interventions: Results of the Bone and Joint Decade 2000–2010 Task Force on Neck Pain and Its Associated Disorders. *Spine (Phila Pa 1976)* 2008; 33:S153-S169.

244. Côté P, van der Velde G, Cassidy JD, Carroll LJ, Hogg-Johnson S, Holm LW, Carragee EJ, Haldeman S, Nordin M, Hurwitz EL, Guzman J, Peloso PM; Bone and Joint Decade 2000–2010 Task Force on Neck Pain and Its Associated Disorders. The burden and determinants of neck pain in workers. Results of the Bone and Joint Decade 2000–2010 Task Force on Neck Pain and Its Associated Disorders. *Spine (Phila Pa 1976)* 2008; 33:S60-S74.

245. 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; 1:447-458.

246. Deville´ WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. *J Clin Epidemiol* 2000; 53:65-69.

247. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: Reducing the number needed to read. *J Am Med Inform Assoc* 2002; 9:653-658.

248. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: Analytical survey. *BMJ* 2004; 328:1040.

249. Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. *J Clin Epidemiol* 2005; 58:444-449.

250. Leeflang MM, Scholten RJ, Rutjes AW, Reitsma JB, Bossuyt PM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. *J Clin Epidemiol* 2006; 59:234-240.

251. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *BMJ* 1994; 309:1286-1291.

252. Lefebvre C, Eisinga A, McDonald S, Paula N. Enhancing access to reports of randomized trials published worldwide – the contribution of EMBASE records to the Cochrane Central Register of Controlled Trials (CENTRAL) in The Cochrane Library. *Emerg Themes Epidemiol* 2008; 5:13.

253. McDonald S, Taylor L, Adams C. Searching the right database. A comparison of four databases for psychiatry journals. *Health Libr Rev* 1999; 16:151-156.

254. Turp JC, Schulte JM, Antes G. Nearly half of dental randomized controlled trials published in German are not included in MEDLINE. *Eur J Oral Sci* 2002; 110:405-411.

255. Odaka T, Nakayama A, Akazawa K, Sakamoto M, Kinukawa N, Kamakura T, Nishioka Y, Itasaka H, Watanabe Y, Nose Y. The effect of a multiple literature database search — a numerical evaluation in the domain of Japanese life science. *J Med Syst* 1992; 16:177-181.

256. Smith BJ, Darzins PJ, Quinn M, Heller RF. Modern methods of searching the medical literature. *Med J Aust* 1992; 157:603-611.

257. Rovers JP, Janosik JE, Souney PF. Crossover comparison of drug information online database vendors: Dialog and MEDLARS. *Ann Pharmacother* 1993; 27:634-639.

258. Ramos-Remus C, Suarez-Almazor M, Dorgan M, Gomez-Vargas A, Russell AS. Performance of online biomedical databases in rheumatology. *J Rheumatol* 1994; 21:1912-1921.

259. Royle P, Bain L, Waugh N. Systematic reviews of epidemiology in diabetes: Finding the evidence. *BMC Med Res Methodol* 2005; 5:2.

260. Cook DJ, Guyatt GH, Ryan G, Clifton J, Buckingham L, Willan A, McIlroy W, Oxman AD. Should unpublished data be included in meta-analyses? Current convictions and controversies. *JAMA* 1993; 269:2749-2753.

261. Hopewell S, Clark M, Lefebvre C, Scherer R. Handsearching still a valuable element of the systematic review. *Evid Based Dent* 2008; 9:85.

262. Khan KS, Kleijnen J. Stage II — Conducting the review, Phase 4 — Selection of studies. In: *Undertaking Systematic Rev iews of Research on Effectiveness.* CRDs guidance for carrying out or commissioning reviews. CRD Report Number 4 (2nd), CRD Centre for Reviews and Dissemination, University of York, York, UK, March 2001. www.york.ac.uk/inst/crd/report4.htm

263. Khan KS, ter Riet G, Popay J, Nixon J, Kleijnen J. Stage II — Conducting the review, Phase 5 — Study quality assessment. In: *Undertaking Systematic Reviews of Research on Effectiveness.* CRDs guidance for carrying out or commissioning reviews. CRD Report Number 4 (2nd), CRD Centre for Reviews and Dissemination, University of York, York, UK, March 2001. www.york.ac.uk/inst/crd/report4.htm

264. Higgins JPT, Green S (eds). Assessment of study quality. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 (updated September 2006); Section 6. In: *The Cochrane Library*, Issue 4, 2006. John Wiley & Sons, Ltd., Chichester, UK.

265. Oxman AD, Stachenko SJ. Meta-analysis in primary care: Theory and Practice. In: Tudiver F, Bass MJ, Dunn EV, Norton PG (eds). *Assessing Interventions: Traditional and Innovative Research Methods for Primary Care*. Sage Publications, Newbury Park, 1992, pp 191-207.

266. Slavin B. Best evidence synthesis: An intelligent alternative to meta-analysis. *J Clin Epidemiol* 1995; 48:9-18.

267. Goodman C. Step 2: Specify inclusion criteria for studies. Swedish Council on Technology Assessment in Health Care, 1993.

268. Cooper H, Ribble RG. Influences on the outcome of literature searches for integrative research reviews. *Knowledge* 1989; 10:179-201.

269. Oxman A, Guyatt G. The science of reviewing research. *Ann NY Acad Sci* 1993; 703:125-134.

270. Eysenck HJ. Meta-analysis and its problems. *BMJ* 1994; 309:789-792.

271. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. *BMJ* 2002; 324:669-671.

272. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol* 2005; 58:882-893.

273. Whiting P, Harbord R, Kleijnen J. No role for quality scores in systematic reviews of diagnostic accuracy studies. *BMC Med Res Methodol* 2005; 5:19.

274. Jadad AR, Moher D, Klassen TP. Guides for reading and interpreting systematic reviews: II. How do the authors find the studies and assess their quality? *Arch Pediatr Adolesc Med* 1998; 152:812-817.

275. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: Current issues and future directions. *Int J Technol Assess Health Care* 1996; 12:195-208.

276. Greenland S. Quality scores are useless and potentially misleading. *Am J Epidemiol* 1994; 140:300-301.

277. Topfer LA, Parada A, Menon D, Noorani H, Perras C, Serra-Prat M. Comparison of literature searches on quality and costs for health technology assessment using the MEDLINE and EMBASE databases. *Int J Technol Assess Health Care* 1999; 15:297-303.

278. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, Lijmer JG; Standards for Reporting of Diagnostic Accuracy. The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Ann Inter Med* 2003; 138:W1-W12.

279. Feinstein AR. *Clinical Epidemiology: The Architecture of Clinical Research*. WB Saunders Co., Philadelphia, 1985.

280. Higgins JPT, Green S (eds). Collecting data. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 (updated September 2006); Section 7. In: *The Cochrane Library*, Issue 4, 2006. John Wiley & Sons, Ltd.Chichester, UK.

281. Deeks JJ, Higgins JPT, Altman DG (eds). Analysing and presenting results. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]; Section 8. In: *The Cochrane Library*, Issue 4, 2006. John Wiley & Sons, Ltd. Chichester, UK.

282. Deeks J, Khan KS, Song F, Popay J, Nixon J, Kleijnen J. Stage II – Conducting the review, Phase 7 – Data synthesis. In: *Undertaking Systematic Reviews of Research on Effectiveness*. CRDs guidance for carrying out or commissioning reviews. CRD Report Number 4 (2nd), CRD Centre for Reviews and Dissemination, University of York, York, UK. March 2001. www.york.ac.uk/inst/crd/report4.htm

283. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: Methodologic primer. *AJR Am J Roentgenol*. 2006; 187:271-281.

284. Marcus SH, Grover PL, Revicki DA. The method of information synthesis and its use in the assessment of health care technology. *Int J Technol Assess Health Care* 1987; 3:497-508.

285. Khan KS, ter Riet G, Kleijnen J. Stage III — Reporting and dissemination, Phase 8 — The report and recommendations. In: *Undertaking Systematic Reviews of Research on Effectiveness*. CRDs guidance for carrying out or commissioning reviews. CRD Report Number 4 (2nd), CRD Centre for Reviews and Dissemination, University of York, York, UK. March 2001. www.york.ac.uk/inst/crd/report4.htm

286. How to use the evidence: Assessment and application of scientific evidence. National Health and Medical Research Council, Canberra, Commonwealth of Australia, 2000, pp 1-84.

287. Bigos SJ, Boyer OR, Braen GR, Brown K, Deyo R, Haldeman S, Hart JL, Johnson EW, Keller R, Kido D, Liang MH, Nelson RM, Nordin M, Owen BD, Pope MH, Schwartz RK, Stewart DH, Susman J, Triano JJ, Tripp LC, Turk DC, Watts C, Weinstein JN. Acute low back problems in adults. Clinical Practice Guideline No.14, AHCPR Publication No. 95-0642. Rockville, Maryland. U.S.A., Agency for Healthcare Policy and Research, Public Health Service, U.S., Department of Health and Human Services, December, 1994, pp 1-60.

288. Berg AO, Allan JD. Introducing the third U.S. Preventive Services Task Force. *Am J Prev Med* 2001; 20:S3-S4.

289. Guyatt G, Gutterman D, Baumann MH, Addrizzo-Harris D, Hylek EM, Phillips B, Raskob G, Lewis SZ, Schünemann H. Grading strength of recommendations and quality of evidence in clinical guidelines. Report from an American College of Chest Physicians task force. *Chest* 2006; 129:174-181.

290. Deeks JJ, Higgins JPT, Altman DG (eds). Interpreting results. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]; Section 9. In: *The Cochrane Library*, Issue 4, 2006. John Wiley & Sons, Ltd., Chichester, UK.

291. Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D. How quickly do systematic reviews go out of date? A survival analysis. *Ann Intern Med* 2007; 147:224-233.

292. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB. Meta-analysis of observational studies in epidemiology: A proposal for reporting. Meta-analysis of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 2000; 283:2008-2012.

293. Moher D, Schulz KF, Altman D, for the CONSORT Group. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 2001; 285:1987-1991.

294. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ; CONSORT Group. Reporting of noninferiority and equivalence randomized trials: An extension of the CONSORT statement. *JAMA* 2006; 295:1152-1160.

295. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P; CONSORT Group. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: Explanation and elaboration. *Ann Intern Med* 2008; 148:295-309.

296. Taddio A, Pain T, Fassos FF, Boon H, Ilersich AL, Einarson TR. Quality of nonstructured and structured abstracts of original research articles in the *British Medical Journal*, the *Canadian Medical Association Journal* and the *Journal of the American Medical Association*. *CMAJ* 1994; 150:1611-1615.

297. Haynes RB, Mulrow CD, Huth EJ, Altman DG, Gardner MJ. More informative abstracts revisited. *Ann Intern Med* 1990; 113:69-76.

298. Hartley J, Sydes M, Blurton A. Obtaining information accurately and quickly: Are structured abstracts more efficient? *Journal of Information Science* 1996; 22:349-356.

299. World Medical Association declaration of Helsinki. Recommendations guiding physicians in biomedical research involving human subjects. *JAMA* 1997; 277:925-926.

300. Savulescu J, Chalmers I, Blunt J. Are research ethics committees behaving unethically? Some suggestions for improving performance and accountabil-

ity. *BMJ* 1996; 313:1390-1393.

301. McAuley L, Moher D, Tugwell P. The influence of grey literature on meta-analysis. MSc Thesis, University of Ottawa, 1999.

302. Moher D, Pham B, Klassen TP, Schulz KF, Berlin JA, Jadad AR, Liberati A; International Cochrane Colloquium. Does the language of publication of reports of randomized trials influence the estimates of intervention effectiveness reported in meta-analyses? [conference presentation]. In: 6th Cochrane Colloquium, 1998, Baltimore. Providence, New England Cochrane Center Providence Office.

303. Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. *Lancet* 1997; 350: 326-329.

304. Khan KS, Daya S, Collins JA, Walter S. Empirical evidence of bias in infertility research: Overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. *Fertil Steril* 1996; 65:939-945.

305. *Annals of Internal Medicine.* Information for authors. Available at www.annals.org.