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

From: Laxmaiah Manchikanti, MD1, Richard Derby, MD2, Lee Wolfer, MD2, Vijay Singh, MD3, Sukdeb Datta, MD4, and Joshua A. Hirsch, MD5

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

www.painphysicianjournal.com
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, reproducitive 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 (47,48).

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 choose 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 (AHORQ) 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 es-
sential 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): reducing 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
litxature 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 re-
views and meta-analysis of randomized trials, a dan-
gerous discrepancy exists between experts and evi-
dence 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 informa-
tion. 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 with-
out disc herniation or radiculitis. Precision diagnostic
blocks have changed this substantially. Nachemson
(129) reported that in only 15% of cases could a path-
oanatomical 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 diagnos-
tic 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 dem-
onstrated 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) evalu-
ated 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 diagno-
sis of sacroiliac joint pain. The validity of diagnostic
interventional techniques with variable accuracy and
reliability has been described in multiple studies yield-
ing mixed results (26-40,67-79,81-83,120,144).

Even though extensive criticism has been fo-
cused 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 diagno-
tic 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 scor-
ing 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 per-
formance 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 rais-
ing test had a sensitivity of 0.9 with pooled specificity
of 0.88. Finally, they concluded that the diagnostic ac-
curacy of the straight leg raising test is limited by its
low specificity. Even then, the reliability on these di-
agnostic tests is enormous in the practice of medicine
and guidelines.

In contrast to tests with low accuracy, diagnostic
interventional techniques have been shown to pres-
ent with significant evidence of accuracy. Diagnostic
spinal interventional techniques have presented con-
cept validity, content validity, face validity, and con-
struct validity. Multiple studies also have evaluated
the false-positive rates of the diagnostic interventio-
al 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 not 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
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 1. Comparison of traditional and systematic reviews.**

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

Guidelines illustrating the evaluation of systematic reviews of diagnostic accuracy (12) and meta-analysis (14) delineate 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.

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

<table>
<thead>
<tr>
<th>DOMAIN</th>
<th>ELEMENTS*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study question</td>
<td>• Question clearly specified and appropriate</td>
</tr>
</tbody>
</table>
| 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.

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, 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 provocative 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 methodol-
ogy 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.

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

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

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

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

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 with-in 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.

Table 5. The QUADAS tool.

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

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 agreement 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)

---

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.

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. 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. West et al 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. 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.

Table 8 illustrates panel ratings of available evidence supporting guideline statements developed by AHRQ (formerly AHCPR). However, the document publishing these ratings is considered as outdated. The United States Preventive Services Task Force (USPSTF) describes level of evidence for therapy rather than diagnosis. These have been modified to incorporate diagnostic studies in multiple systematic reviews performed (Table 9).
<table>
<thead>
<tr>
<th>Phase in Review Process</th>
<th>Key Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Definition of the review objectives</td>
<td>To identify the review question:</td>
</tr>
<tr>
<td></td>
<td>State the patient group and define presenting condition(s), previous test results, and health care setting.</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Define the target condition to be diagnosed and reference standards to be used.</td>
</tr>
<tr>
<td>2. Study identification and selection</td>
<td>Search several electronic databases.</td>
</tr>
<tr>
<td></td>
<td>Use a search strategy built around terms for the index test, target condition, and possibly patient characteristics.</td>
</tr>
<tr>
<td></td>
<td>Do not use restrictive methodological search filters.</td>
</tr>
<tr>
<td>3. Quality assessment</td>
<td>Identify biases for which the included studies are at risk.</td>
</tr>
<tr>
<td></td>
<td>Use the QUADAS checklist as a tool for identifying many common deficiencies.</td>
</tr>
<tr>
<td></td>
<td>Comment on the adequacy of each aspect of study design. Do not use summary quality scores.</td>
</tr>
<tr>
<td>4. Data extraction, analysis, and presentation</td>
<td>Extract paired estimates of test sensitivity and specificity from each study overall and, if available, for patient subgroups.</td>
</tr>
<tr>
<td></td>
<td>Plot studies in ROC space to identify the location, variability, and correlations.</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Compute average values of sensitivity and specificity when the data combined share a common threshold.</td>
</tr>
<tr>
<td></td>
<td>Use summary ROC curves to describe test performance and to compare tests without restricting to particular thresholds.</td>
</tr>
<tr>
<td></td>
<td>Obtain estimates of summary likelihood ratios from average values of sensitivity and specificity and not through separate pooling of likelihood ratios.</td>
</tr>
<tr>
<td></td>
<td>Global tests for heterogeneity before data synthesis or tests for publication bias are typically not useful.</td>
</tr>
<tr>
<td></td>
<td>Meta-analyze and present studies that compare tests by using randomized or within-patient designs separately from the results of indirect comparisons.</td>
</tr>
<tr>
<td>5. Interpretation</td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Consider complementing the interpretation with decision modeling by using results of the review.</td>
</tr>
</tbody>
</table>

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

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-
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.

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


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.

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

continued on page 951
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.

---

Fig. 2. Progress through the stages of a systematic 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).

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.

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

References


Moher D, Tricco AC. Issues related to the conduct of systematic reviews: A focus on the nutrition field. Am J Clin Nutr 2008; 88:1191-1199.


Hulme PA, Krebs J, Ferguson SJ, Berlemann U. Vertebroplasty and kypho-

...


175. Manchikanti L, Damron KS, Cash KA,


www.painphysicianjournal.com
The Evidence Base of Cochrane Database Syst Rev


240. Armon C, Argoff CE, Samuels J, Backonja MM; Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. As-
Systematic Reviews and Meta-Analyses of Diagnostic Accuracy Studies

Spin - Assessing Interventions - 2017


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.


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.


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


