Health Policy Review


Laxmaiah Manchikanti, MD1, Sukdeb Datta, MD2, Richard Derby, MD3, Lee R. Wolfer, MD3, Ramsin M. Benyamin, MD4, and Joshua A. Hirsch, MD5

From: 1Pain Management Center of Paducah, Paducah, KY; 2Vanderbilt University Medical Center, Nashville, TN; 3Spinal Diagnostics & Treatment Center, Daly City, CA; 4Millennium Pain Center, Bloomington, IL; and 5Massachusetts General Hospital and Harvard Medical School, Boston, MA

Background: Clinical guidelines are defined as systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances. The clinical guideline industry has been erupting even faster than innovation in health care, constantly adding unhealthy perspectives with broad and complex mandates to health care interventions. Clinical guidelines are based on evidence-based medicine (EBM) and comparative effectiveness research (CER).

Multiple issues related to the development of clinical guidelines are based on conflicts of interest, controversies, and limitations of the guideline process. Recently, the American Pain Society (APS) developed and published multiple guidelines in managing low back pain resulting in multiple publications. However, these guidelines have been questioned regarding their development process, their implementation, and their impact on various specialties.

Objectives: To reassess the APS guidelines’ evidence synthesis for low back pain diagnostic interventions using the same methodology utilized by the APS authors. The interventions examined were diagnostic techniques for managing low back pain of facet joint origin, discogenic origin, and sacroiliac joint origin.

Methods: A literature search by two authors was carried out utilizing appropriate databases from 1966 through July 2008. Methodologic quality assessment was also performed by at least 2 authors utilizing the same criteria applied in APS guidelines. The guideline reassessment process included the evaluation of individual studies and systematic reviews and the translation into practice recommendations.

Results: Our reassessment of Chou et al’s evaluation, utilizing Chou et al’s criteria, showed good evidence for lumbar facet joint nerve blocks, fair evidence for lumbar provocation discography, and fair to poor evidence for sacroiliac joint blocks to diagnose sacroiliac joint pain. The reassessment illustrates that Chou et al have utilized multiple studies inappropriately and have excluded appropriate studies. Also, Chou et al failed to eliminate their bias in their study evaluations.

Conclusion: The reassessment, using appropriate methodology and including high quality studies, shows evidence that differs from published APS guidelines.

Key words: Guidelines, evidence-based medicine, comparative effectiveness research, systematic reviews, American Pain Society, interventional pain management, interventional techniques, low back pain, diagnostic interventions, lumbar facet joint nerve blocks, lumbar provocation discography, sacroiliac joint nerve blocks

Pain Physician 2010; 13:E141-E174
The Institute of Medicine (IOM) defines clinical guidelines as, “systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances” (1). Clinical guidelines are a constructive response to the reality that practicing physicians require assistance for assimilating and applying the exponentially expanding, often contradictory, body of medical knowledge (2). Clinical practice guidelines attempt to define practices that meet the needs of most patients under most circumstances. However, they do not attempt to supplant the independent judgment of clinicians in responding to particular clinical situations (3). In essence, guidelines enable the implementation of evidence-based medicine (EBM) or the findings of comparative effectiveness research (CER) in medical decision-making with the goal of encouraging effective care (4). Ideally, specific clinical recommendations contained within practice guidelines are systematically developed by expert panels who have access to available evidence, have an understanding of the clinical problem, and have clinical experience with the procedure being assessed as well as the relevant research methods and so are able to make considered, reasoned judgments. Further, these panels are expected to be objective and to produce recommendations that are unbiased, up-to-date, and free from conflicts of interest.

Part 1 of this critical review of the American Pain Society (APS) clinical practice guidelines for interventional techniques provides an assessment of diagnostic interventional techniques. Part 2 will concentrate on therapeutic interventions.

The pace of innovation in health care has been enormous, constantly adding to the broad and complex areas of health care interventions and systems (5,6). In addition, the demonstration of pervasive and persistent unexplained variability in clinical practice and high rates of perceived inappropriate care, combined with increased expenditures, have fueled a steady increase in demand for appropriate application of modalities that have clinical effectiveness (4-29). Consequently, it is essential to develop clinical guidelines, defined as a body of evidence regarding safety, effectiveness, appropriate indications, cost-effectiveness, and other attributes of medical care (5). However, if special interests twist their interpretations to drive an agenda, and develop guidelines based on personal biases and not on science and the best care for the patient, such guidelines have no relevance in clinical practice. Further, researchers, clinicians, professional organizations, and governments should recognize that the value of evidence is only as good as the type of evidence reviewed, the methodology utilized, the knowledge and experience of the reviewers, and many other factors, including bias, self-interest, and economics. A formal set of rules must complement medical training and common sense for clinicians to interpret the results of clinical research effectively (30-32). However, having knowledge of evidence-based practice tools (methodology) does not make one qualified to develop guidelines. Knowing the tools of evidence-based practice methodology is necessary, but not sufficient, for delivering the highest quality patient care. The clinical guidelines panel must incorporate not only the methodologists, but also the clinicians who actually practice medicine and are experts in the technique being reviewed.

Conflicts of interest in guideline development and inappropriate methodologies have been questioned (33,34). Sniderman and Furberg (2) described the conflicts, controversies, and limitations of the guideline process. Limitations of the guideline preparation process include governance and composition of the guideline committee, unanimity in guidelines, lack of independent review, and conflict of interest.

APS developed and published guidelines in managing low back pain resulting in multiple publications (35-42). Similarly, the American College of Occupational and Environmental Medicine (ACOEM) (43,44) and the Official Disability Guidelines (ODG) (45) have published their guidelines. There are numerous other guidelines available, including the ones from the American Society of Interventional Pain Physicians (ASIPP) which were updated in 2009 (4,46-51). ASIPP guidelines were developed starting with an extensive search and review of the literature, including developing systematic reviews and quality assessment of individual articles (52-73). In contrast, ACOEM guidelines for low back pain and chronic pain have been extensively criticized (74-76).

The controversial issues surrounding practice guidelines development and the evidence utilized in those guidelines are not limited to interventional pain management alone (77-79). The joint cardiovascular practice guidelines of the American College of Cardiology (ACC) and American Heart Association (AHA) have become important documents for guiding cardiology practice and establishing benchmarks for quality of care (79,80). However, evaluation of the scientific evidence underlying their clinical practice guidelines showed that the recommendations they issued are largely developed from lower levels of evidence or expert opin-
ion. Further, the proportion of recommendations for which there is no conclusive evidence also seems to be growing. It was noted that these findings highlight the need to improve the process of writing guidelines and to expand the evidence base from which clinical practice guidelines are derived (79).

**METHODS**

1.0 Types of Interventions Included
Common types of diagnostic interventional techniques in managing low back pain of facet joint origin, discogenic origin, and sacroiliac joint origin were included in this analysis.

2.0 Key Questions
Chou et al (35-37) utilized a key question format. The key questions in their analysis were evaluated, answered, and compared.

What is the diagnostic accuracy and what are the potential harms associated with invasive tests for identifying patients who may benefit from invasive procedures? How effective is prior use of these tests for selecting patients for invasive procedures in order to improve outcomes?

♦ Provocative discography
♦ Diagnostic selective nerve root blocks (not assessed)
♦ Diagnostic intraarticular facet joint blocks and medial branch blocks
♦ Diagnostic sacroiliac joint blocks

3.0 Literature Search
A comprehensive literature search was conducted which included the search of databases including PubMed and EMBASE from 1966 through July 2008 as utilized in the manuscripts. Search of Cochrane database, Clinical Trial Registry, systematic reviews, narrative reviews, and cross-references to the reviews published in the English language were also included.

The search was performed by at least 2 authors; their emphasis was chronic low back pain with a focus on diagnostic interventional techniques. The search strategy and MeSH terms utilized by Chou and Huffman (35) were followed.

4.0 Selection Criteria
Part 1 of this critical review focused on systematic reviews of diagnostic interventions and diagnostic accuracy studies. The population of interest was patients suffering with chronic low back pain. All the studies utilized by Chou and Huffman (35) that had appropriate management and outcome evaluations were analyzed along with other studies that were not included.

5.0 Methodologic Quality Assessment
Each study was evaluated by 2 physicians for stated criteria and disagreements were resolved by a third physician. If there was a conflict of interest with the reviewed manuscripts concerning authorship or any other conflict, the involved authors did not review the manuscripts for quality assessment or evidence synthesis. The methodologic quality assessment of various individual articles was based on the type of manuscript: systematic reviews of diagnostic accuracy studies and individual diagnostic accuracy studies.

5.1 Methodologic Quality Assessment of Systematic Reviews
Methodologic quality assessment of systematic reviews has been described by Agency for Healthcare Research and Quality (AHRQ) (81) and Oxman et al (82) which were adapted by Furlan et al (83). Table 1 illustrates the AHRQ criteria for systematic reviews. Chou et al (35-37) utilized the criteria developed by Oxman et al (82) and adapted by Furlan et al (83) as illustrated in Table 2. While both appear to be similar, there are significant differences between these tools; however, the basic assumptions of quality assessment criteria are the same for both. Thus, to satisfy the requirements by Chou and Huffman (35), the Oxman criteria were utilized.

5.2 Assessment of Accuracy of Invasive Diagnostic Tests
Chou and Huffman (35) assessed the quality of diagnostic accuracy studies using 9 criteria adapted from methods developed by the United States Preventive Services Task Force (USPSTF) (84) and on empiric studies (85,86) of sources of variation and bias in studies of diagnostic tests (Table 3). They also determined that studies which met at least 5 of the 9 criteria were considered higher-quality.

However, AHRQ has provided a quality rating assessment system for diagnostic interventions with 5 criteria (Table 4). These have been utilized in other studies (13,15,52-58,68)

For the present evaluation, criteria utilized by Chou and Huffman (35) as shown in Table 3 were utilized for the purposes of uniformity and simplicity. However, these have not been tested appropriately for interventional techniques in the past.
6.0 Analysis of Evidence

Chou and Huffman (35) utilized the method for grading the overall strength of evidence for an intervention as developed by USPSTF (84) as illustrated in Table 5. USPSTF also developed the grading system of research design utilized by ASIPP guidelines and multiple systematic reviews which continue to be quoted by numerous authorities as shown in Table 6. However, this approach has been criticized because it limits evaluating of internal validity (87).

Consequently, to avoid confusion for this analysis, the criteria utilized by Chou and Huffman (35) was adapted as illustrated in Table 5.

7.0 Assessment of Integrity

USPSTF-defined evidence-based recommendation development with description of aims and processes to ensure integrity (88,89). The goals include transparency, accountability, consistency, and independence.

---

**Table 1. 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>
<tr>
<td></td>
<td>• Sufficiently comprehensive and rigorous with attention to possible publication biases</td>
</tr>
<tr>
<td></td>
<td>• Search restrictions justified (e.g., language or country of origin)</td>
</tr>
<tr>
<td>Search strategy</td>
<td>• Documentation of search terms and databases used</td>
</tr>
<tr>
<td></td>
<td>• Sufficiently detailed to reproduce study</td>
</tr>
<tr>
<td>Inclusion and exclusion criteria</td>
<td>• Selection methods specified and appropriate, with a priori criteria specified if possible</td>
</tr>
<tr>
<td>Interventions</td>
<td>• Intervention(s) clearly detailed for all study groups</td>
</tr>
<tr>
<td>Outcomes</td>
<td>• All potentially important harms and benefits considered</td>
</tr>
<tr>
<td></td>
<td>• Rigor and consistency of process</td>
</tr>
<tr>
<td></td>
<td>• Number and types of reviewers</td>
</tr>
<tr>
<td>Data extraction †</td>
<td>• Blinding of reviewers</td>
</tr>
<tr>
<td></td>
<td>• Measure of agreement or reproducibility</td>
</tr>
<tr>
<td></td>
<td>• Extraction of clearly defined interventions/exposures and outcomes for all relevant subjects and subgroups</td>
</tr>
<tr>
<td>Study quality and validity</td>
<td>• Assessment method specified and appropriate</td>
</tr>
<tr>
<td></td>
<td>• Method of incorporation specified and appropriate</td>
</tr>
<tr>
<td>Data synthesis and analysis</td>
<td>• Appropriate use of qualitative and/or quantitative synthesis, with consideration of the robustness of results and heterogeneity issues</td>
</tr>
<tr>
<td></td>
<td>• Presentation of key primary study elements sufficient for critical appraisal and replication</td>
</tr>
<tr>
<td>Results</td>
<td>• Narrative summary and/or quantitative summary statistic and measure of precision, as appropriate</td>
</tr>
<tr>
<td>Discussion</td>
<td>• Conclusions supported by results with possible biases and limitations taken into consideration</td>
</tr>
<tr>
<td>Funding or sponsorship</td>
<td>• Type and sources of support for study</td>
</tr>
</tbody>
</table>

* 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.
Table 2. Systematic reviews quality rating system.

<table>
<thead>
<tr>
<th>Criteria for Assessing Scientific Quality of Research Reviews*</th>
<th>Operationalization Of Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Were the search methods reported?</td>
<td></td>
</tr>
<tr>
<td>Were the search methods used to find evidence (original research) on the primary questions stated?</td>
<td></td>
</tr>
<tr>
<td>“Yes” if the review states the databases used, date of most recent searches, and some mention of search terms.</td>
<td></td>
</tr>
<tr>
<td>2. Was the search comprehensive?</td>
<td></td>
</tr>
<tr>
<td>Was the search for evidence reasonably comprehensive?</td>
<td></td>
</tr>
<tr>
<td>“Yes” if the review searches at least 2 databases and looks at other sources (such as reference lists, hand searches, queries experts).</td>
<td></td>
</tr>
<tr>
<td>3. Were the inclusion criteria reported?</td>
<td></td>
</tr>
<tr>
<td>Were the criteria used for deciding which studies to include in the overview reported?</td>
<td></td>
</tr>
<tr>
<td>4. Was selection bias avoided?</td>
<td></td>
</tr>
<tr>
<td>Was bias in the selection of studies avoided?</td>
<td></td>
</tr>
<tr>
<td>“Yes” if the review reports how many studies were identified by searches, numbers excluded, and gives appropriate reasons for excluding them (usually because of pre-defined inclusion/exclusion criteria).</td>
<td></td>
</tr>
<tr>
<td>5. Were the validity criteria reported?</td>
<td></td>
</tr>
<tr>
<td>Were the criteria used for assessing the validity of the included studies reported?</td>
<td></td>
</tr>
<tr>
<td>6. Was validity assessed appropriately?</td>
<td></td>
</tr>
<tr>
<td>Was the validity of all the studies referred to in the text assessed using appropriate criteria (either in selecting studies for inclusion or in analyzing the studies that are cited)?</td>
<td></td>
</tr>
<tr>
<td>“Yes” if the review reports validity assessment and did some type of analysis with it (e.g. sensitivity analysis of results according to quality ratings, excluded low quality studies, etc.)</td>
<td></td>
</tr>
<tr>
<td>7. Were the methods used to combine studies reported?</td>
<td></td>
</tr>
<tr>
<td>Were the methods used to combine the findings of the relevant studies (to reach a conclusion) reported?</td>
<td></td>
</tr>
<tr>
<td>“Yes” for studies that did qualitative analysis if there is some mention that quantitative analysis was not possible and reasons that it could not be done, or if “best evidence” or some other grading of evidence scheme used.</td>
<td></td>
</tr>
<tr>
<td>8. Were the findings combined appropriately?</td>
<td></td>
</tr>
<tr>
<td>Were the findings of the relevant studies combined appropriately relative to the primary question the overview addresses?</td>
<td></td>
</tr>
<tr>
<td>“Yes” if the review performs a test for heterogeneity before pooling, does appropriate subgroup testing, appropriate sensitivity analysis, or other such analysis.</td>
<td></td>
</tr>
<tr>
<td>9. Were the conclusions supported by the reported data?</td>
<td></td>
</tr>
<tr>
<td>Were the conclusions made by the author(s) supported by the data and/or analysis reported in the overview?</td>
<td></td>
</tr>
<tr>
<td>10. What was the overall scientific quality of the overview?</td>
<td>The score for Question 10, the overall scientific quality, should be based on your answers to the first nine questions. The following guidelines can be used to assist with deriving a summary score: If the &quot;Can't tell&quot; option is used one or more times on the preceding questions, a review is likely to have minor flaws at best and it is difficult to rule out major flaws (i.e. a score of 4 or lower). If the &quot;No&quot; option is used on Question 2, 4, 6 or 8, the review is likely to have major flaws (i.e. a score of 3 or less, depending on the number and degree of the flaws).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scoring: Each Question is scored as Yes, Partially/Can't tell or No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extensive Flaws</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

Table 3. Quality assessment of diagnostic study criteria utilized by Chou and Huffman (35).

1) Evaluated a consecutive series of patients or a random subset
2) Evaluated patients prospectively
3) Evaluated patients with a broad spectrum of symptoms
4) Adequately described the diagnostic test technique
5) Used current diagnostic techniques
6) Adequately described criteria for a positive diagnostic test
7) Used an appropriate definition for a positive diagnostic test
8) Performed statistical analysis on potential predictors or confounders of positive diagnostic tests, and
9) Performed testing blinded to patient symptoms and other clinical characteristics.

7.1 Transparency

Transparency is provided by standardized methodology described in methods.

7.2 Accountability

A conflict of interest policy, the process for prioritizing the literature, peer review of evidence synthesis and recommendations and updating of the recommendations consistent with current literature constitute accountability.

7.3 Consistency

Systematic reviews of the literature on effectiveness and harms utilizing use of outcome tables to assess the balance of benefits and harms with a defining evidence grid and descriptions in a standardized language.

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></td>
</tr>
<tr>
<td>• Independent, blind interpretation of test and reference</td>
<td></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>
</tbody>
</table>

TOTAL SCORE 100

Adapted and modified from West S et al. Systems to Rate the Strength of Scientific Evidence, Evidence Report, Technology Assessment No. 47. AHRQ Publication No. 02-E016 (81).

7.4 Independence

Finally, the evidence review process, voting process for members only, meeting attendance by invitation, and formalized communication among the stakeholders must be independent.

Table 5. Method for grading the overall strength of the evidence for an intervention.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>Evidence includes consistent results from well-designed, well-conducted studies in representative populations that directly assess effects on health outcomes (at least two consistent, higher-quality RCTs or studies of diagnostic test accuracy).</td>
</tr>
<tr>
<td>Fair</td>
<td>Evidence is sufficient to determine effects on health outcomes, but the strength of the evidence is limited by the number, quality, size, or consistency of included studies; generalizability to routine practice; or indirect nature of the evidence on health outcomes (at least one higher-quality trial or study of diagnostic test accuracy of sufficient sample size; or two or more higher-quality trials or studies of diagnostic test accuracy with some inconsistency; at least two consistent, lower-quality trials or studies of diagnostic test accuracy, or multiple consistent observational studies with no significant methodological flaws).</td>
</tr>
<tr>
<td>Poor</td>
<td>Evidence is insufficient to assess effects on health outcomes because of limited number or power of studies, large and unexplained inconsistency between higher-quality trials, important flaws in trial design or conduct, gaps in the chain of evidence, or lack of information on important health outcomes.</td>
</tr>
</tbody>
</table>

Critical Review of APS Clinical Practice Guidelines for Interventional Techniques

Table 6. Quality of evidence developed by USPSTF

<table>
<thead>
<tr>
<th>I</th>
<th>Evidence obtained from multiple properly conducted diagnostic accuracy studies.</th>
</tr>
</thead>
<tbody>
<tr>
<td>II-1</td>
<td>Evidence obtained from at least one properly conducted diagnostic accuracy study of adequate size.</td>
</tr>
<tr>
<td>II-2</td>
<td>Evidence obtained from at least one properly designed small diagnostic accuracy study.</td>
</tr>
<tr>
<td>II-3</td>
<td>Evidence obtained from diagnostic studies of uncertainty.</td>
</tr>
<tr>
<td>III</td>
<td>Opinions of respected authorities, based on clinical experience descriptive studies and case reports or reports of expert committees.</td>
</tr>
</tbody>
</table>

Adapted from the U.S. Preventive Services Task Force (USPSTF) (84).

RESULTS

The results are described below. The methodology utilized followed the sequence as described in the APS guidelines (35). The literature search extended through July 2008. The critical analysis in Part 1 included only diagnostic interventions.

1.0 Provocation Discography

Chou and Huffman (35) focused on several specific types of studies about provocative discography. Their literature search found 324 potentially relevant citations with consideration of multiple manuscripts (90-120).

Their summary indicated that in healthy, asymptomatic volunteers, positive responses to provocative discography were uncommon in several series of patients with fair evidence. Even then, based on all the other negative evidence and outdated guidelines from the Agency for Health Care Policy and Research (AHCPR) (34) and European COST Guidelines (108) they recommended against discography for the diagnosis of discogenic pain in patients with chronic low back pain.

1.1 Literature Search and Inclusion Criteria

For the present review, the computerized and manual search of literature yielded over 4,000 manuscripts, of which 600 were considered as potentially relevant; their abstracts were reviewed. After excluding abstracts not relevant to the present evaluation, and lack of full manuscripts, 130 full manuscripts were reviewed. Of these, 69 manuscripts were considered for inclusion related to diagnostic and outcome studies.

For this critical assessment, all the studies evaluating asymptomatic volunteers and symptomatic patients were considered. Discography, alone or in combination with other tests, must have been clearly described and performed according to the International Association for the Study of Pain (IASP) standards (121), International Spine Intervention Society (ISIS) standards (122), North American Spine Society (NASS) standards (123), or ASIPP guidelines (46). We also reassessed all the systematic reviews.

1.2 Methodologic Quality Assessment

1.2.1 Provocation Discography Studies

Chou and Huffman (35) included multiple studies without defined inclusion or exclusion criteria. They also failed to include 2 systematic reviews (56,76). Most systematic reviews have included inclusion criteria for reproduction of patients’ typical pain with disc stimulation.

Based on the present review, 19 studies performed discography under controlled conditions (92-101,105,111-118). However, of these, only 5 studies were performed with a controlled disc (101,111-113,118). Of these, 2 studies utilized the same patients (111,112). Thus, only 4 studies met inclusion criteria (101,111,113,118). Of the 7 studies included by Chou and Huffman (35) in their positive pain response to provocative discography (92-98), only one study (101) met inclusion criteria. Two studies analyzed other data (93,105).

Table 7 illustrates the quality rating of diagnostic discography trials. All the studies included by Chou and Huffman (35) and the studies included in other systematic reviews (56,58,76) were assessed. Inappropriate scoring was observed in at least 2 studies included by Chou and Huffman (35). This changes the importance of one study (101) with the score changing from 6/9 to 7/9. In addition, they claimed that the study by Manchikanti et al (101) was highly selective, because the patients had already undergone negative testing for facet joint pain, as well as an epidural steroid injection. However, such a high selection should actually increase the specificity of the study rather than reducing it (58,76).
Table 7. Quality rating of provocative discography diagnostic accuracy and outcome studies.

<table>
<thead>
<tr>
<th>Study</th>
<th>ASIPP</th>
<th>APS-AAPM</th>
<th>ASIPP</th>
<th>APS-AAPM</th>
<th>ASIPP</th>
<th>APS-AAPM</th>
<th>ASIPP</th>
<th>APS-AAPM</th>
<th>ASIPP</th>
<th>APS-AAPM</th>
<th>ASIPP</th>
<th>APS-AAPM</th>
<th>ASIPP</th>
<th>APS-AAPM</th>
<th>ASIPP</th>
<th>APS-AAPM</th>
<th>ASIPP</th>
<th>APS-AAPM</th>
<th>ASIPP</th>
<th>APS-AAPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consecutive series or random subset</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>NA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Prospective</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>NA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluates patients with a spectrum of symptoms</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>NA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Don't know</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adequate description of discography technique</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Use of current discography technique</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adequate description of criteria for positive test</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Appropriate definition for positive test</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical analysis of positive tests predictors</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Investigator not aware of clinical symptoms</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Included by Chou and Huffman (35) and Manchikanti et al (58)
+ Included by Manchikanti et al (58), but not Chou and Huffman (35)
* Included by Chou and Huffman (35), but not Manchikanti et al (58)
§ Used by Chou and Huffman (35) as predictors of positive response
# Used by Chou and Huffman (35) for outcome study
NA = Carragee et al (93) not rated for this review, not original data
NS = not scored by APS-AAPM review
Table 7 (cont.). Quality rating of provocative discography diagnostic accuracy and outcome studies.

<table>
<thead>
<tr>
<th>Study</th>
<th>Consecutive series or random subset</th>
<th>Prospective</th>
<th>Evaluates patients with a spectrum of symptoms</th>
<th>Adequate description of discography technique</th>
<th>Use of current discography technique</th>
<th>Adequate description of criteria for positive test</th>
<th>Appropriate definition for positive test</th>
<th>Statistical analysis of predictors for positive tests</th>
<th>Investigator not aware of clinical symptoms</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carragee et al 2006 (105)*</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>NS</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>6/9</td>
</tr>
<tr>
<td>Madan et al 2002 (107)+</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>NS</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>8/9</td>
</tr>
<tr>
<td>Derby et al 1999 (99) +</td>
<td>Yes</td>
<td>No</td>
<td>Don’t know</td>
<td>Yes</td>
<td>NS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>4/9</td>
</tr>
<tr>
<td>Derby et al 2005 +</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>NS</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>7/9</td>
</tr>
<tr>
<td>Schwarz et al 1995 (111) +</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>NS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>6/9</td>
</tr>
<tr>
<td>Cohen et al 2002 (118) +</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>NS</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>8/9</td>
</tr>
<tr>
<td>Manchikanti et al 2001 (113)+</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Don’t know</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>8/9</td>
</tr>
</tbody>
</table>

* included by Chou and Huffman (35) and Manchikanti et al (58)
+ Included by Manchikanti et al (58), but not Chou and Huffman (35)
♦ Included by Chou and Huffman (35), but not Manchikanti et al (58)
§ Used by Chou and Huffman (35) as predictors of positive response
# used by Chou and Huffman (35) for outcome study
NA – Carragee et al (93) not rated for this review, not original data
NS = not scored by APS-AAPM review
Cohen et al (118) evaluated the effect of needle insertion from the same side of the pain or the opposite side without any difference. This can be considered a validity study and was included by Manchikanti et al (58), however, was not included by Chou and Huffman (35).

1.2.2 Systematic Reviews of Lumbar Provocation Discography Studies

Chou and Huffman (35) identified 3 reviews as systematic (110,119,120). However, the search criteria missed 2 other systematic reviews, which were published in July 2008 (56,76). Further, the review by Cohen et al (119) was a narrative review. The second review by Willems et al (120) was a study of 435 consecutive discograms and a systematic review of the literature in relation to prophylactic antibiotic usage.

The quality rating of Buenaventura et al's (110) systematic review was inaccurate because they clearly avoided bias, whereas Chou and Huffman (35) assumed that they were unable to tell. They also assessed validity appropriately even though it was assumed as partial by Chou and Huffman (35), and finally they also reported the conclusions appropriately which Chou and Huffman (35) assumed that they had not reported. Thus, the score increased from 5/9 or 3/7 to 7/9, which increases the quality rating of this systematic review. Consequently, the conclusions made by Chou and Huffman (35) based on their assessment of systematic reviews were not of high quality or were not valid.

Further, systematic reviews performed by Wolfer et al (56) and Manchikanti et al (76) were not identified. These systematic reviews were of high quality as shown in Table 8. Surprisingly, utilizing the same European COST Guidelines, Rubinstein and van Tulder (124) evaluated systematic reviews performed by others very similar to the systematic review performed by Buenaventura et al (110).

### Table 8. Quality ratings of systematic reviews evaluating diagnostic discography systematic reviews.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Search Method</strong></td>
<td>ASIPP</td>
<td>APS-AAPM</td>
<td>ASIPP</td>
<td>APS-AAPM</td>
<td>ASIPP</td>
</tr>
<tr>
<td><strong>Comprehensive</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Partial</td>
<td>NRI</td>
<td>No</td>
</tr>
<tr>
<td><strong>Inclusion Criteria</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>NSR</td>
<td>Partial</td>
<td>NRI</td>
</tr>
<tr>
<td><strong>Validity Assessed</strong></td>
<td>Yes</td>
<td>Can't tell</td>
<td>Can't tell</td>
<td>NRI</td>
<td>Can't tell</td>
</tr>
<tr>
<td><strong>Validity Criteria</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>NSR</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Methods for Combining Studies</strong></td>
<td>Yes</td>
<td>Partial</td>
<td>NSR</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Appropriately Combined</strong></td>
<td>No</td>
<td>No</td>
<td>NSR</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Conclusions Supported</strong></td>
<td>No</td>
<td>No</td>
<td>NSR</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Overall Quality</strong></td>
<td>7/9</td>
<td>3/7</td>
<td>1/7</td>
<td>2/7</td>
<td>9/9</td>
</tr>
<tr>
<td><strong>Corrected Score</strong></td>
<td>7/9 (5/9)</td>
<td>(0/9)</td>
<td>(3/9)</td>
<td>9/9</td>
<td>7/9</td>
</tr>
</tbody>
</table>

*Study published in study period, but not identified by Chou and Huffman (35).

NSR = not a systematic review
NS = not scored by APS-AAPM review
NRI = not rated due to irrelevance
( ) corrected score by their own criteria
1.3 Standardization of Technique

The technique of lumbar discography is standardized by IASP criteria (121) and has been well studied (56,58,110,111,121-123,125,126). The definition of a positive discogram, per ISIS guidelines (122), is pain > 7/10, concordance, pressure < 50 psi above opening pressure, Grade III anular tear, and a painless control disc. ASIPP guidelines (46) have defined a positive discogram only if the target disc produces concordant pain with an intensity of at least 7 on a 10-point pain measurement scale or 70% of the highest reported pain (i.e., worst spontaneous pain of 7 = 7 x 70% = 5), and 2 adjacent discs with provocation discography do not produce any pain at all or only one disc in the case of L5/S1 with low volume and low pressure injection.

In an ideal situation, a gold standard or criterion is obtained by tissue confirmation of the presence or absence of a disease; however, surgical inspection of a degenerated disc cannot determine if discogenic pain is present or not. Thus, the greatest challenge concerning discography continues to be the gold standard problem. Four systematic reviews discussed these issues (56,58,109,110). However, the gold standard problem is not unique to discography. Knottnerus et al (127) stated that there are several methodological challenges that must be addressed in diagnostic accuracy studies. These include the gold standard problem, spectrum and selection bias, “soft” measures (subjective phenomena), observer variability and bias, complex relations, clinical impact, sample size, and the rapid progress of knowledge (127).

1.4 Validity

The face validity of discography is based on the premise that pressurizing a disc reproduces the physiological conditions that stress a disc until the nociceptive threshold is reached. Construct validity can be established by demonstrating a significant correlation between discography results and surgical outcomes. For a response to be considered positive, concordant pain must be reproduced; in order to meet validity standards, at least one adjacent disc must be painless upon injection.

The sensitivity and specificity of discographic pathology are 81% and 64%, using radiological imaging as the criterion standard. A recent meta-analysis of provocation discography in asymptomatic subjects found a false-positive rate of 6% when previously published data were re-analyzed based on IASP criteria (56).

1.5 Outcomes Assessment

Carragee et al (105) used fusion results as the criterion standard in demonstrating the lack of validity of lumbar provocation discography; however, there is sparse evidence that either fusion or disc replacement is an effective treatment for discogenic low back pain (119,128-133). In spite of the widespread use or misuse of lumbar discography as a presurgical screening tool, few studies have evaluated its effect on surgical outcomes. The relative lack of controlled studies is further compounded by significant variability in outcomes and the controversy surrounding spinal arthrodesis and disc prosthesis procedures for discogenic low back pain (133,134). The surgical outcomes for the treatment of IDD are widely acknowledged to be inferior than for radiculopathy, with reported success rates ranging from less than 50% to greater than 80% (129-131). The randomized studies comparing fusion outcomes to conservative treatment demonstrated mixed results (129,132,133). In addition, of multiple published studies evaluating disc replacement outcomes (133), none directly compared outcomes between patients whose selection was contingent on discography results and those who underwent disc replacement based solely on clinical and radiological findings. The presence of concomitant pain sources in most patients with discogenic pain, along with inconsistent clinical outcomes, even with a technically successful surgery, are factors that must be considered when evaluating the predictive value of discography for surgical outcomes.

Health Technology Assessment of Spinal Fusion and Discography for Chronic Low Back Pain Secondary to Uncomplicated Lumbar Degenerative Disc Disease (134) concluded the following: 1) the evidence is insufficient to permit conclusions about the reliability of discography, to predict fusion outcomes, and to permit conclusions about the influence of discography on fusion outcomes in patients with chronic low back pain and uncomplicated lumbar degenerative disc disease. Two studies (135,136) were used to assess the test-retest and inter-reader reliability of discography. Three studies (106,137,138) were used to evaluate the ability of presurgical discography to predict outcomes based on fusion. The results of this analysis revealed that 2 of the 3 studies argued favorably for discographic screening. In the Willems et al study (106), no differences were found in fusion success rates between patients who had a positive discogram(s) adjacent to the fused levels and those who did not. In the Colhoun et al (138) study, 89% of those with provoked pain experienced a posi-
tive fusion outcome, which favorably compared to the 52% success rate in those whose discograms revealed morphological abnormalities but no pain provocation.

Chou and Huffman (35) utilized only 2 studies for outcomes assessment Carragee et al (105) and Madan et al (107). However, neither of these studies were utilized by Health Technology Assessment. Consequently, none of the 3 studies (106,137,138) utilized by Health Technology Assessment (134) were utilized by Chou and Huffman (35). However, the analyses by Chou and Huffman (35) and Health Technology Assessment (134) did not utilize the study by Derby et al (99), which was the only study utilizing manometry as a determining factor in discography interpretations. In addition, Essex et al (139) and Madan et al (107) failed to duplicate the results of Colhoun et al (138). Further, Cohen et al (119,128) also found no pooled differences in fusion outcomes between studies that use discography and those that did not. As a consequence, the lack of strong evidence for the use of fusion to treat degenerative disc disease as well as methodological flaws in component studies, make data interpretation exceptionally difficult (138,140). Therefore, utilizing fusion outcomes as the criterion standard for evaluating the accuracy of lumbar provocation discography is not only unrealistic, but also unscientific.

Further, Carragee et al’s (105) report of the low predictive value of discography was based on a study comparing surgical outcomes in a control isthmic spondylolisthesis group versus a discogenic pain group (both with a single level low pressure positive discogram). In the control group, 72% (23/32) of patients had a highly effective success rate as opposed to 27% (8/30) in the discogenic pain group. Even though they used surgical intervention as the gold standard, they mistakenly compared 2 completely different low back populations. It is common knowledge that state-of-the-art surgical interventions for painful single level ischemic spondylolisthesis are superior to those for discogenic pain (105,140-142). In contrast, outcomes for the surgical treatment of chronic axial discogenic low back pain are not only variable, but inferior (93,141-144). Thus, Carragee et al’s (105) outcomes are within the range of expected results.

1.6 Magnetic Resonance Imaging (MRI) Correlation

Lei et al (145) concluded that magnetic resonance imaging (MRI) is an excellent tool for assessing disc morphology, but should be used in conjunction with discography for planning surgical treatment. They concluded a new MRI classification of disc degeneration, found to have good intra- and inter-observer agreement, with discography. The sensitivity and specificity of MRI in predicting a painful disc was 94% and 77%, which favorably compared to endplate signal changes and high intensity zones, which were found to have sensitivities of 32% and 27%, respectively. O’Neill et al (146) in evaluating the accuracy of MRIs in diagnosing discogenic pain concluded that MRI parameters correlate both with each other and discography findings, and that a nuclear signal is the most important MRI characteristic to consider.

1.7 Inclusion of Controversial Studies

Systematic reviews (46,48,56,58,76,109,110) have evolved. Refinements have been made in their execution. Despite this, multiple studies were excluded from the analysis. Carragee et al (93-97,105,116,117) performed multiple studies related to discography in subjects with or without low back pain. None of the studies were performed utilizing IASP criteria. Derby et al also published multiple studies (98-100,114). However, in the APS guidelines, only one study - Derby et al (98) was included.

Chou and Huffman (35) also included Walsh et al (92), which also does not meet inclusion criteria. Further, Chou et al (35-37) also included other studies evaluating predictors of positive pain responses (101-104). Of these, 3 studies have not been included in other systematic reviews (102-104) because they were unrelated to the accuracy of provocation discography.

1.8 Exclusion of High-Quality Studies

Chou and Huffman (35) failed to identify 2 high quality systematic reviews. They also excluded 2 high quality diagnostic accuracy studies (111,113).

In 1995, Schwarzer et al (111) sought to identify historical or physical exam features associated with discogenic pain and determine prevalence (56,76) by performing discography utilizing IASP criteria on 92 patients with chronic low back pain. In addition to concordant pain reproduction at a disc containing a grade 3 or 4 radial fissure, a negative control disc had to be present for a disc to be deemed a pain generator. Overall, 36 patients, or 39%, satisfied the criteria for a positive discogram. The 95% confidence limits for this proportion were 29% to 49%. The authors concluded that a diagnosis of painful IDD can only be made with discography.

The study by Manchikanti et al (113) evaluated the relative contributions of potential pain generators in...
120 patients with chronic non-radicular low back pain. All patients initially underwent controlled comparative diagnostic facet joint nerve blocks with lidocaine and bupivacaine. In patients with negative medial branch blocks, sacroiliac joint injections were performed in those patients with tenderness overlying the joint and positive provocative maneuvers. In subjects in whom the facet and SI joints were ruled out as causative factors, provocation discography was performed in accordance with IASP criteria. Overall, the prevalence of discogenic pain was estimated to be 26% (95% confidence interval [CI], 18%, 34%).

1.9 False-Positive Rates

Chou and Huffman (35) considered lumbar provocation discography as unreliable because of false-positive rates. A series of systematic reviews (56,58,76) and published studies investigated the potential for false-positive results of discography (92,94-97,114,116). The issue of inordinately high false-positive rates date back to the Holt study (90), which was performed on prisoners. That study was done with outdated techniques and noxious, irritating contrast dye (147). It also did not consider pain response as a criterion for a positive discogram.

In an attempt to determine the effect establishing pressure thresholds has on the rate of false-positives, Carragee et al re-analyzed previously published data (94-96) according to low pressure criteria. They (93) reported a false-positive rate of 25% (17/69 patients), which was not statistically significantly different from the 27% positive rate (14/52) in their comparison cohort of patients with presumed chronic discogenic pain. This exploratory post-hoc analysis was performed on 5 prior experimental groups (no pain, no low back pain [n = 10]; chronic pain [n = 10]; somatization disorder [n = 4]; post-discectomy [n = 20]; and mild persistent backache [n = 25]). Low pressure positive was defined as ≤ 22 psi above opening pressure, which is higher than the standard set by ISIS/IASP of ≤ 15 psi a.o. (106). The individual groups were found to have the following false-positive responses: pain free 0/10, chronic pain 3/10, somatization disorder 2/4, post-discectomy 5/20, and “benign” backache, 7/25 patients.

There are significant shortcomings in Carragee et al’s (93) re-analysis. The pain-free group had a 0% false-positive rate. The chronic pain group included 10 chronic pain patients who were disabled volunteers with failed cervical fusions, on regular medications (including opioids), with markedly abnormal psychometric scores, and active worker’s compensation litigation. Using high-pressure provocation (pressure ≤ 100 psi a.o.), Carragee et al (93) reported a false-positive rate of 40%; however, because of the small numbers, the 95% confidence level ranged between 10% and 70%. If one substitutes the ISIS/IASP (121,122) standard of ≤ 15 psi a.o., the false-positive rate decreases to 10% per patient (1/10) (95% CI, 0% – 33%) and 8.3% per disc (1/12) (95% CI, 0% – 27%) (56). Furthermore, Carragee et al (93) included 4 patients with somatization disorder in this analysis who might arguably be removed from consideration. However, a prospective study by Manchikanti et al (101) found no difference in the rate of positive discograms between patients with and without somatization disorder.

Lastly, Carragee et al (94) included 25 patients with a history of persistent, low intensity back pain. Thirty-six percent (n = 9) of these subjects were deemed false-positive in the original protocol analysis, which declined only slightly to 28% (n = 7) in the re-analysis. Yet, the contention that these patients represent false-positive responses is contestable. An alternative explanation is that these individuals were in a more quiescent phase of their illness, or simply were more stoic. This argument is bolstered by the original 36% false-positive rate, which is similar to the 39% prevalence rate of discogenic pain reported by Schwarzer et al (111). In summary, Carragee et al’s (93) post-hoc analysis of select populations with low pressure positive discograms is subject to different interpretations. When more stringent criteria are applied, the false-positive rate in individuals without confounding factors is very low.

Not all studies have found high false-positive rates in asymptomatic volunteers. Walsh et al (92) sought to replicate Holt’s work (90), but attempted to remediate some of the shortcomings by including in their criteria pain intensity ratings, concordance, and observed pain behaviors. Although discograms were morphologically abnormal in 5 of the 10 subjects, none elicited concordant pain. Derby et al (98) also performed 3 or more discograms in 13 volunteers with no low back pain history. Although 44% of injected discs elicited pain, most required high pressures to reach the nociceptive threshold, and even then, were only mildly painful. The authors concluded that if one takes into consideration pain intensity and the amount of pressure needed to provoke symptoms, the false-positive rate is less than 10%. Wolfer et al (56) conducted a meta-analysis on all complete data sets obtained from lumbar discography studies done in subjects asymptomatic for low back pain. Using ISIS/IASP standards, the pooled analysis of 75 patients and 116 discs revealed a false-positive rate...
Pain Physician: May/June 2010; 13:E141-174

Table 9. Summary of false-positive rates (%) per patient and per disc for experimental studies in subjects asymptomatic of low back pain.*†

<table>
<thead>
<tr>
<th>STUDY</th>
<th>Walsh et al (92)/Carragee et al (96)</th>
<th>Derby et al (99)</th>
<th>ISIS/IASP (122)</th>
<th>Low pressure &lt; 22 psi a.o. (Carragee)</th>
<th>Low pressure ≤ 15 psi a.o. (Derby)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%FP /pt</td>
<td>%FP /disc</td>
<td>%FP /pt</td>
<td>%FP /disc</td>
<td>%FP /pt</td>
</tr>
<tr>
<td>Walsh et al (92): Asymptomatic volunteers (95% CI)</td>
<td>0 ( - )</td>
<td>0 ( - )</td>
<td>0 ( - )</td>
<td>0 ( - )</td>
<td>0 ( - )</td>
</tr>
<tr>
<td>Carragee et al (97) Iliac crest (95% CI)</td>
<td>50 (5 – 95%)</td>
<td>28.6 (2 – 56%)</td>
<td>37.5 (0 – 81%)</td>
<td>21.4 (0 – 46%)</td>
<td>12.5 (0 – 42%)</td>
</tr>
<tr>
<td>Carragee et al (96): pain-free (cs-good) (95% CI)</td>
<td>10 (0 – 33%)</td>
<td>10 (0 – 33%)</td>
<td>10 (0 – 33%)</td>
<td>10 (0 – 33%)</td>
<td>10 (0 – 33%)</td>
</tr>
<tr>
<td>Carragee et al (96): chronic pain (cs-failed) (95% CI)</td>
<td>40 (3 – 77%)</td>
<td>58.3 (26 – 91%)</td>
<td>30 (0 – 65%)</td>
<td>33.3 (2 – 65%)</td>
<td>20 (0 – 50%)</td>
</tr>
<tr>
<td>Carragee et al (114): Somatization disorder (95% CI)</td>
<td>75 (0 – 100%)</td>
<td>44.4 (4 – 85%)</td>
<td>50 (0 – 100%)</td>
<td>22.2 (0 – 56%)</td>
<td>50 (0 – 100%)</td>
</tr>
<tr>
<td>Derby et al (114): Asymptomatic volunteers (95% CI)</td>
<td>0 ( - )</td>
<td>0 ( - )</td>
<td>0 ( - )</td>
<td>0 ( - )</td>
<td>0 ( - )</td>
</tr>
<tr>
<td>Carragee et al (94) mild backache (95% CI)</td>
<td>36 (16 – 56%)</td>
<td>37.5 (20 – 55%)</td>
<td>36 (16 – 56%)</td>
<td>31.3 (14 – 48%)</td>
<td>20 (3 – 37%)</td>
</tr>
<tr>
<td>Carragee et al (95): Post-discectomy (95% CI)</td>
<td>35 (12 – 58%)</td>
<td>24.2 (9 – 40%)</td>
<td>35 (12 – 58%)</td>
<td>24.2 (9 – 40%)</td>
<td>25 (4 – 46%)</td>
</tr>
</tbody>
</table>

*ISIS = International Spine Intervention Society; IASP = International Association for the Study of Pain; a = no control disc; b = control disc ≤ 6/10; c = painless control disc; FP = false-positive; pt = patient; cs-good=cervical spine surgery, good outcome; cs-failed=cervical spine surgery, poorest outcome; CI = confidence interval

† Studies by Holt (90) and Massie and Steven (91) are not included since pain and pressure were not reported in the published study.


of 9.3% (95% CI, 3%–16%) per patient and 6.0% (95% CI, 2%–10%) per disc (Table 9).

Thus, it is demonstrated that lumbar discography performed in accordance with accepted guidelines is associated with a low false-positive rate of 3% in subjects without confounding factors, 0% in the pain-free group, 10% in the low pressure positive chronic pain group, 15% in prior discectomy, and 12.5% in patients with residual pain after iliac crest bone harvesting – ranging from 0% to 12.5% with a total false-positive rate of 9.3% (95% CI, 3%, 16%) in contrast to the high false-positive rates of 40% to 83% described by Carragee et al (96) and adapted by Chou and Huffman (35).

1.10 Results of Reassessment of Provocation Discography

Chou and Huffman (35), based on the evidence, concluded that there were high false-positive rates and
also that there was no value for provocation discography. They based their recommendations on outdated AHCPR guidelines and European COST guidelines (108), recommending against discography for diagnosing discogenic pain in patients with chronic low back pain. Based on the USPSTF criteria, the indicated evidence has been determined to be Level II-2 for lumbar discography, with a prevalence of discogenic pain of 26% and internal disc disruption of 39% with false-positive rates of 9.3%.

It is inappropriate to utilize an unproven treatment to question the diagnostic accuracy of a test, in this case, provocation discography.

Considering the inclusion of inappropriate studies and the exclusion of high quality studies; and using inappropriate methodology for assessing the quality of systematic reviews and individual studies, Chou and Huffman (35) have erred in their analysis, conclusion, and recommendations. The evidence after reassessment is “fair” based on criteria utilized by Chou and Huffman (35).

2.0 Diagnostic Intraarticular Facet Joint Block and Medial Branch Block

Chou and Huffman (35) recommends against diagnostic intraarticular facet joint blocks and medial branch blocks based on European COST guidelines (108). Their search included only one study by Birkenmaier et al (148) despite the vast literature available about evaluating of diagnostic facet joint blocks (46,48,54). The basis was that as in other invasive diagnostic procedures for low back pain, no reliable reference standard for facet joint pain was available to estimate the diagnostic accuracy of intraarticular facet joint blocks and medial branch blocks. Thus, they concluded that it was unknown whether the decreased positive response rate was due to fewer false-positives, fewer true positives, or some combination. Further, they also utilized the basic indication for facet joint blocks as a contraindication for utilizing the lack of correlation of facet joint pain diagnosed by controlled diagnostic blocks as not correlating well with findings on imaging studies.

2.1 Literature Search and Inclusion Criteria

The literature search included all studies published on diagnosing lumbar facet joint pain in patients who had chronic pain for more than 3 months. However, only the studies of controlled diagnostic blocks, either placebo or comparative local anesthetic blocks utilizing fluoroscopy, were included. The criterion standard for diagnosing lumbar facet joint pain was at least 80% pain relief for the duration of the local anesthetic and the ability to perform previously painful movements.

A computerized and manual search of the literature yielded over 1,700 manuscripts with 6 systematic reviews (54,124,149-152) and multiple other manuscripts (113,153-191).

2.2 Methodologic Quality Assessment

2.2.1 Studies of Diagnostic Lumbar Facet Joint Nerve Blocks

Chou and Huffman (35), from 46 potentially relevant citations, identified only one lower-quality trial. It was not included in any of the systematic reviews that evaluated clinical outcomes in patients selected for percutaneous facet joint cryodenervation based on a positive uncontrolled medial branch block versus pericapsular block (148,187). This study, published as 2 reports, failed to meet inclusion criteria because of uncontrolled blocks comparing pericapsular blocks in evaluating cryodenervation – a technique with lack of literature. Further, in this study they utilized generally higher volumes than that utilized in other studies, specifically, concentrated bupivacaine 0.5% with volumes higher than 1 mL resulting in an improvement in the patients’ specific low back pain of 50% or more for at least 3 hours. Surprisingly, even at this low quality, the results were superior with a single uncontrolled medial branch block compared to a pericapsular block. Even then, Chou and Huffman (35) reached negative conclusions based on the results of this study.

Our assessment found 7 studies (113,154-156,158,161,162) meeting the inclusion criteria of 80% pain relief and the ability to perform previously painful movements with controlled diagnostic blocks of lumbar facet joint nerves, published prior to Chou et al’s (35) search criteria.

Seven studies were excluded since they evaluated only a single block (173-177,188,189), 14 studies were excluded since the inclusion criteria was pain relief of less than 80% (156,157,160,163,169,172,179-184), 3 studies (166-168) were excluded as these were a subgroup analyses of other studies, 3 studies (170,171,190) evaluated the validity of diagnosis, 2 studies (185,191) evaluated the effect of sedation, 2 studies (165,166) evaluated the role of psychological factors, and one study (186) was excluded because it studied opioid exposure.

The methodologic quality assessment is shown in Table 10. Since the methodologic quality assessment
Table 10. Quality rating of intraarticular facet joint block and medial branch blocks: diagnostic accuracy and outcome studies.

<table>
<thead>
<tr>
<th>Study</th>
<th>Methodologic Quality Assessment Criteria</th>
<th>Participants</th>
<th>Prevalence</th>
<th>False-Positive Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manchikanti et al 2002 (159)</td>
<td>8/9 NS</td>
<td>120</td>
<td>40% (95% CI; 31%–49%)</td>
<td>30% (95% CI; 20%–40%)</td>
</tr>
<tr>
<td>Manchikanti et al 2004 (155)</td>
<td>8/9 NS</td>
<td>397</td>
<td>31% (95% CI; 27%–36%)</td>
<td>27% (95% CI; 22%–32%)</td>
</tr>
<tr>
<td>Manchukonda et al 2007 (154)</td>
<td>7/9 NS</td>
<td>303</td>
<td>27% (95% CI; 22%–33%)</td>
<td>45% (95% CI; 36%–53%)</td>
</tr>
<tr>
<td>Schwarzer et al 1995 (161)</td>
<td>8/9 NS</td>
<td>63</td>
<td>37% (95% CI; 25%–49%)</td>
<td>NA</td>
</tr>
<tr>
<td>Manchikanti et al 2001 (113)</td>
<td>8/9 NS</td>
<td>120</td>
<td>40% (95% CI; 31%–49%)</td>
<td>47% (95% CI; 35%–59%)</td>
</tr>
<tr>
<td>Manchikanti et al 2003 (158)</td>
<td>8/9 NS</td>
<td>300</td>
<td>I. 21% (95% CI; 14%–27%) II. 41% (95% CI; 33%–49%)</td>
<td>I. 17% (95% CI; 10%–24%) II. 27% (95% CI; 18%–36%)</td>
</tr>
<tr>
<td>Manchikanti et al 2007 (162)</td>
<td>8/9 NS</td>
<td>117</td>
<td>16% (95% CI; 9%–23%)</td>
<td>49% (95% CI; 39%–59%)</td>
</tr>
<tr>
<td>Overall</td>
<td>1,420</td>
<td>31% (95% CI; 28%–33%)</td>
<td>30%* (95% CI; 27%–33%)</td>
<td></td>
</tr>
</tbody>
</table>

* Included by Datta et al (54), but not Chou and Huffman (35)
NS = not scored by APS-AAPM review

Table 11. Data of prevalence with controlled diagnostic blocks and false-positive rates in lumbar region.

<table>
<thead>
<tr>
<th>Study</th>
<th>Methodologic Quality Assessment Criteria</th>
<th>Participants</th>
<th>Prevalence</th>
<th>False-Positive Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manchikanti et al 2002 (159)</td>
<td>8/9 NS</td>
<td>120</td>
<td>40% (95% CI; 31%–49%)</td>
<td>30% (95% CI; 20%–40%)</td>
</tr>
<tr>
<td>Manchikanti et al 2004 (155)</td>
<td>8/9 NS</td>
<td>397</td>
<td>31% (95% CI; 27%–36%)</td>
<td>27% (95% CI; 22%–32%)</td>
</tr>
<tr>
<td>Manchukonda et al 2007 (154)</td>
<td>7/9 NS</td>
<td>303</td>
<td>27% (95% CI; 22%–33%)</td>
<td>45% (95% CI; 36%–53%)</td>
</tr>
<tr>
<td>Schwarzer et al 1995 (161)</td>
<td>8/9 NS</td>
<td>63</td>
<td>37% (95% CI; 25%–49%)</td>
<td>NA</td>
</tr>
<tr>
<td>Manchikanti et al 2001 (113)</td>
<td>8/9 NS</td>
<td>120</td>
<td>40% (95% CI; 31%–49%)</td>
<td>47% (95% CI; 35%–59%)</td>
</tr>
<tr>
<td>Manchikanti et al 2003 (158)</td>
<td>8/9 NS</td>
<td>300</td>
<td>I. 21% (95% CI; 14%–27%) II. 41% (95% CI; 33%–49%)</td>
<td>I. 17% (95% CI; 10%–24%) II. 27% (95% CI; 18%–36%)</td>
</tr>
<tr>
<td>Manchikanti et al 2007 (162)</td>
<td>8/9 NS</td>
<td>117</td>
<td>16% (95% CI; 9%–23%)</td>
<td>49% (95% CI; 39%–59%)</td>
</tr>
<tr>
<td>Overall</td>
<td>1,420</td>
<td>31% (95% CI; 28%–33%)</td>
<td>30%* (95% CI; 27%–33%)</td>
<td></td>
</tr>
</tbody>
</table>

CI = confidence interval
NA = not available
# Schwarzer et al (161) not included
criteria carried out by Chou and Huffman (35) included only one study (148,187), there were no values to be compared. The data of prevalence and false-positive rates of these studies is included in Table 11. Thus, based on strict criteria, the prevalence has been shown to be 21% to 40% with an overall rate of 31% in a heterogeneous population with chronic low back pain with CIs ranging from 14% to 53%, along with false-positive rates of 17% to 19% with CIs ranging from 10% to 59% with an overall false-positive rate of 30%.

2.2.2 Systematic Reviews of Diagnostic Lumbar Facet Joint Nerve Blocks

Our literature search yielded 7 systematic reviews (54,76,124,149-151,152). Of these, Chou and Huffman (35) identified 2 systematic reviews (149,151). They criticized that neither review included any study evaluating the use of intraarticular facet joint blocks or medial branch blocks to select patients for procedures intended to treat presumed facet joint pain; and whether doing so improves clinical outcomes compared to relying on other methods to select patients’ procedures.

Datta et al (54) was excluded as it was published after the search criteria. Hancock et al (152), Rubinstein and van Tulder (124), Manchikanti et al (76) were considered for inclusion. However, Hancock et al (152) evaluated non-invasive tests, whereas Rubinstein and van Tulder (124) utilized data from Sehgal et al (151).

Thus, assessment of systematic reviews was performed for 3 reviews concerning diagnostic accuracy studies (76,151,152). Table 12 illustrates the results.

2.3 Exclusion of Appropriate Studies

As illustrated in this methodologic quality assessment, Chou and Huffman (35) excluded all high quality individual diagnostic accuracy studies and systematic reviews from consideration and evidence synthesis. Even with substantial tightening of the criteria, in this analysis, we were able to include 7 diagnostic accuracy studies (54,76,124,149-151,152), which met criteria based on Chou et al’s (35) recommendations.

2.5 Inclusion of Inappropriate Studies

Chou and Huffman (35) also inappropriately included Birkenmaier’s study (148,187), which is not only of poor quality, but has not met any inclusion criteria for a diagnostic accuracy study. It essentially compared 2 uncontrolled procedures. They utilized higher concentrations of bupivacaine 0.5% and volumes higher than 1 mL. They also utilized cryodenervation, which has not been proven to be an effective procedure; there is no substantial literature available on cryodenervation. Even so, Birkenmaier et al (148,187) showed that patients who had been selected by medial branch blocks had better pain relief than did patients who had been diagnosed using pericapsular blocks with statistical significance noted at 6 weeks and...
3 months. They concluded that the results suggest that uncontrolled medial branch blocks are superior to percapsular blocks when selecting patients for facet joint cryodenervation. They also concluded that both blocks work and further commented that if serial controlled blocks cannot be used, lumbar facet joint pain remains a diagnostic dilemma. A multitude of these factors were not taken into consideration by Chou and Huffman (35). Further, percapsular blocks may be misleading as the injection can be made into the joint itself. They also might leak into the epidural or subarachnoid space, producing a false-positive result of relief.

2.5 Reassessment of the Evidence of Diagnostic Lumbar Facet Joint Nerve Blocks

Based on the present review, there is strong evidence for diagnostic lumbar facet joint nerve blocks. This has been confirmed by a recent review by Datta et al (54), as well as an independent review by Rubinstein and van Tulder (124).

Consequently, the assessment by Chou and Huffman (35) regarding diagnosing lumbar facet joint pain was not performed accurately, leading to inaccurate conclusions and recommendations.

Based on reassessment, utilizing Chou et al’s (35) criteria, the evidence for diagnostic lumbar facet joint nerve blocks is “good.”

3.0 Diagnostic Sacroiliac Joint Blocks

Chou and Huffman (35) described diagnostic sacroiliac joint blocks and concluded that there are no studies on how using sacroiliac joint blocks to evaluate patients for sacroiliac joint pain affects the choice of therapy and clinical outcomes compared to the use of non-invasive methods alone. They described that other guidelines do not address diagnostic sacroiliac joint blocks even though these have been addressed in multiple guidelines in the past (192,193). They opined that the rates of positive intraarticular sacroiliac joint blocks range from 2% to 27%, depending in part on the population evaluated and method of block used – namely controlled or uncontrolled, fluoroscopically guided or not (194-196). They also hypothesized that as in other invasive diagnostic procedures for low back pain, no reliable reference standard for sacroiliac pain is available for estimating the diagnostic accuracy of sacroiliac joint blocks. Further, they claim the main reason why sacroiliac joint blocks are used as a negative correlate is that results of sacroiliac joint blocks may not correlate well with findings on imaging studies. They focused only on phase 4 evidence of diagnostic research hierarchy studies but none were available. They also failed to find any relevant studies for diagnostic sacroiliac joint blocks.

Contrary to the assertions of Chou and Huffman (35), there is substantial evidence for sacroiliac joint blocks, even though this is not as robust as for the diagnosis of lumbar facet joint pain (68,76,124,152,194).

3.1 Literature Search and Inclusion Criteria

Our literature search yielded over 2,000 manuscripts resulting in review of 82 full manuscripts with multiple manuscripts considered for inclusion. Of these, 8 systematic reviews were identified (68,76,124,152,194,197-199), along with 16 individual manuscripts that were considered for inclusion (113,189,200-213). The following studies were excluded for failure to include patients with comparative blocks, unresolveable technical flaws, or study of aspects other than diagnostic accuracy (189,202,203,206-213).

3.2 Methodologic Quality Assessment Criteria

3.2.1 Diagnostic Sacroiliac Joint Injection Studies

Inclusion criteria for diagnostic sacroiliac joint interventions was chronic low back pain and/or lower extremity pain for greater than 3 months, controlled diagnostic blocks with a criterion standard of 80% or greater pain relief for the duration of action of local anesthetic; increased ability to perform previously painful movements. Of the 15 manuscripts available for inclusion, in only 2 studies was 80% relief utilized as the criterion standard (113,204). In contrast, Maigne et al (200) utilized 75% pain relief; Irwin et al (201) utilized 70% relief; and finally, van der Wurff et al (205) utilized 50% relief as the criterion standard. Due to lack of availability of studies, the inclusion criteria were modified to include 70% relief as the criterion standard. Thus, 4 of the 5 studies were included in the methodologic quality assessment (Table 13). Table 14 illustrates the prevalence and false-positive rate.

3.2.2 Systematic Reviews of Diagnostic Sacroiliac Injections

Of the 8 systematic reviews available (68,76,124,152,194,197-199), only one (194) was considered by Chou and Huffman (35) for evidence synthesis. One systematic review was published in 2009 (68), another systematic review (124) based its synthesis on a previous systematic review (194). Hancock et al (152), in a systematic review of tests to identify sacroiliac joint
pain, evaluated a multitude of clinical tests, but not the value of diagnostic blocks. Of the other 3 systematic reviews, Song et al (197) evaluated the diagnostic value of scintigraphy in assessing sacroiliitis in ankylosing spondylitis, Simpson and Gemmell (199) evaluated the accuracy of spinal orthopaedic tests, and Szadek et al (198) evaluated the diagnostic validity of criteria for sacroiliac joint pain. Thus, 2 systematic reviews met inclusion criteria (Table 15) (76,194). The reassessment of the rating quality yielded different results for Hansen et al (194) with high quality systematic review. Similar results were obtained for Manchikanti et al (76) with high quality assessment.

### 3.3 Exclusion of Appropriate Studies

Based on the above evaluation, Chou and Huffman (35) excluded appropriate studies to evaluate diagnostic accuracy. Prevalence appears to range from 10% to 26.6% with CIs ranging from 0% to 39% as shown in Table 14, with false-positive rates ranging from 0% to 53.8% with CIs of 0% to 64%. Of interest, Laslett et al (206) showed false-positive rates of 0%. However, this is the only study which has shown 0% false-positive rates with dual blocks.

Chou and Huffman (35) failed to identify multiple high quality systematic reviews. These systematic re-

---

### Table 13. Quality rating of diagnostic accuracy studies of sacroiliac joint pain.

<table>
<thead>
<tr>
<th>Study</th>
<th>Assessment Criteria</th>
<th># of Subjects</th>
<th>Prevalence Estimates</th>
<th>False-Positive Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manchikanti et al (113)</td>
<td>8/9 NS</td>
<td>20</td>
<td>10% (95% CI, 0% - 23%)</td>
<td>22% (95% CI, 3% - 42%)</td>
</tr>
<tr>
<td>Maigne et al (200)</td>
<td>8/9 NS</td>
<td>54</td>
<td>18.5% (95% CI, 8% - 29%)</td>
<td>20% (95% CI, 8% - 33%)</td>
</tr>
<tr>
<td>Irwin et al (201)</td>
<td>7/9 NS</td>
<td>158</td>
<td>26.6% (95% CI, 20% - 34%)</td>
<td>53.8% (95% CI, 43% - 64%)</td>
</tr>
<tr>
<td>Laslett et al (204)</td>
<td>9/9 NS</td>
<td>43/48</td>
<td>25.6% (95% CI, 12% - 39%)</td>
<td>0%</td>
</tr>
</tbody>
</table>

CI = confidence interval
views provide value for diagnosing sacroiliac joint pain using controlled diagnostic blocks. They also provide value for non-invasive testing. Chou and Huffman (35) also excluded Rubinstein and van Tulder (124), which was a best evidence review of diagnostic procedures for low back pain that focused on previously published systematic reviews (194). Chou and Huffman (35) rated it as low quality. Ironically, Rubinstein and van Tulder (124) performed their best evidence synthesis based on European COST guidelines (108).

3.4 Results of Reassessment of Diagnostic Sacroiliac Joint Injections

Based on this critical reassessment, it appears that Chou and Huffman (35) did not follow the appropriate principles of diagnostic accuracy studies or the determination of a criterion standard. As a result, their conclusions are inaccurate.

The reassessment shows moderate evidence for the diagnostic accuracy of sacroiliac joint blocks in diagnosing sacroiliac joint pain, meeting specific criteria. This has been echoed by multiple systematic reviews in the past. The independent review by Rubinstein and van Tulder (124), utilizing European COST guidelines (108), also concluded that there was moderate evidence for diagnostic facet joint blocks based on the systematic review (194), which was considered as of low quality by Chou and Huffman (35).

Based on reassessment, utilizing Chou et al’s criteria, the evidence is “fair to poor.”

### 4.0 Assessment of Integrity

USPSTF has defined evidence-based recommendation development with a description of aims and processes to ensure integrity (88,89). The goals include transparency, accountability, consistency, and independence.

Based on the original document (35), the guidelines appear to have been sponsored by APS and American Academy of Pain Medicine (AAPM); however, only opioid guidelines state that this is the case. In addition, all the tables show that the evidence synthesis was for APS and AAPM. Thus, the relationship of AAPM with this guideline development has not been divulged. Multiple authors who have withdrawn from this guideline synthesis have not been mentioned or disclosed. Lack of clinical expertise will add substantial issues to raise questions of credibility. Of the 4 authors of the principle manuscript of interventional techniques, only 2 of them appear to be pain specialists and at least 2 of them have withdrawn their support of the guidelines (36). The large document describing the entire guidelines shows only doctors Chou and Huffman as the authors. Huffman is not shown as an author in any of the other manuscripts (36,37).

Further, Chou is an employee of the United States government as the Scientific Director of the Oregon Evidence-based Practice Center, which is funded by the AHRQ, and is also Lead Investigator for the Center’s support for the USPSTF. The data shows that a grant

<table>
<thead>
<tr>
<th><strong>Table 15. Quality rating of systematic reviews of sacroiliac joint pain.</strong></th>
<th><strong>Manchikanti et al 2008 (76)</strong></th>
<th><strong>Hansen et al 2007 (194)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ASIPP</strong></td>
<td><strong>APS-AAPM</strong></td>
<td><strong>ASIPP</strong></td>
</tr>
<tr>
<td>Search Method</td>
<td>Yes</td>
<td>NS</td>
</tr>
<tr>
<td>Comprehensive</td>
<td>Yes</td>
<td>NS</td>
</tr>
<tr>
<td>Inclusion Criteria</td>
<td>Yes</td>
<td>NS</td>
</tr>
<tr>
<td>Bias Avoided</td>
<td>Yes</td>
<td>NS</td>
</tr>
<tr>
<td>Validity Criteria</td>
<td>Yes</td>
<td>NS</td>
</tr>
<tr>
<td>Validity Assessed</td>
<td>Yes</td>
<td>NS</td>
</tr>
<tr>
<td>Methods for Combining Studies</td>
<td>No</td>
<td>NS</td>
</tr>
<tr>
<td>Appropriately Combined</td>
<td>No</td>
<td>NS</td>
</tr>
<tr>
<td>Conclusions Supported</td>
<td>Yes</td>
<td>NS</td>
</tr>
<tr>
<td>Overall Quality</td>
<td>7/9</td>
<td>NS</td>
</tr>
<tr>
<td>Corrected score</td>
<td>7/9</td>
<td>NS</td>
</tr>
</tbody>
</table>

* Included by Chou and Huffman (35) and in the present review
** Study published in study period, but not identified by Chou and Huffman (35).
NS = not scored by APS-AAPM review
() corrected score by their own criteria
was provided by APS to prepare the guidelines. It appears that there were also other societies involved. The data is not available with regards to the amount of the grant and if the salary was also provided during this work. Further, the conflicts of interest of other authors has not been disclosed. The conflicts of interest are not only related to economic interests, but also academic interests and personal biases, etc. Thus, the transparency has not been provided.

With regards to the accountability, Chou et al admitted to several potential limitations even though these are a few and far between and have not been popularized (35,36). In this extensive review, they have included only the randomized controlled trials and also provided substantial criticism to other systematic reviews, which included observational studies. Further issues of accountability include exclusion of the opinions of the members and non-publication of the withdrawn authors.

There was no consistency as the evaluation methodology changed based on authorship of the individual articles and systematic reviews exerting significant bias.

Finally, while editors retained the independence, authors of this reassessment concluded that the participants seemed to have no control on the decision-making process and the input was not related to the clinical issues discussed. Thus, at least 2 participants for interventional techniques have withdrawn their names from inclusion of these reviews.

**Discussion**

This critical assessment attempts to review APS guidelines developed by Chou and Huffman (35). In this analysis, the same principles described by Chou and Huffman (35) were utilized. However, the reevaluation provided different results. Based on this reevaluation, utilizing the criteria described by Chou and Huffman (35), the evidence was fair for lumbar provocation discography, good for lumbar diagnostic facet joint nerve blocks, whereas the evidence was considered fair to poor for diagnostic sacroiliac joint blocks.

The development process of the guidelines by APS appears to be superior to the ACOEM guideline process and others. However, there were deficiencies and inappropriate evaluation in almost all areas; inappropriate studies were included and appropriate studies were excluded. The major impact is based on the utilization of outdated guidelines, inappropriate application of evidence assessment criteria, methodologic quality assessment without weighted values, and lack of conflict management.

The differences in rating strength for diagnosing discogenic pain by provocation discography and facet joint pain by diagnostic facet joint nerve blocks is identified with fair grading based on Chou et al’s utilization of the criteria and strong evidence by ASIPP utilization of USPSTF criteria. The evidence for diagnosis of sacroiliac joint pain by diagnostic sacroiliac joint blocks is fair to poor by the present evaluation based on Chou et al’s criteria and Level II-2 by USPSTF criteria in prior evaluations. Thus, the discrepancy is extremely important for Medicare or insurance coverage or non-coverage of these techniques.

**Lumbar Provocation Discography**

Lumbar provocation discography is a procedure that is used to characterize the pathoanatomy and architecture of the intervertebral disc and to determine if the intervertebral disc is a source of chronic low back pain (46,48,58). Discography is an invasive diagnostic test that should only be applied to those chronic low back pain patients in whom one suspects a discogenic etiology. Basic and clinical studies have overwhelmingly illustrated the nerve supply of the disc and pathomorphologic correlates (111,213-244). However, specific neurobiological events involved in the provocation of pain with discography have not been illustrated. Yet, present knowledge of provocation discography is based on sound anatomic, histopathological, radiological, and biomechanical evidence to identify symptomatic and pathological intervertebral discs (28,46,93). Consequently, the rationale is well established for lumbar provocation discography (28,46,56,58,76,93), just as clinical and radiological examinations might demonstrate a favorable correlation with discography or disc-related pain (48,56,58,111,125,152,189-252).

The accuracy of discography was evaluated by examining cadaver lumbar discs, and by comparing myelograms, CTs, and MRIs. Examinations of cadaver lumbar discs typically confirmed the presence of annular tears and disc degeneration, as revealed by discograms (253-257). High intra-observer agreement has been demonstrated in assessing discographic morphology, i.e., based on Adam’s classification, specifically with exact reproduction of pain (107,253,258). In comparison with other evaluations, CT discography was reported to be more accurate than myelography (259-269) and plain CT (270,271). Regarding MRIs, some have identified advantages of discography with pain provocation when MRIs were normal or equivocal (271-275). Wolfer et al (56) in a recent meta-analysis of provocation discogra-
The basic deficiency of Chou and Huffman (35) was their failure to recognize that discography must not be performed in asymptomatic volunteers or patients with mild low back pain. They also utilized outdated guidelines from AHCPR (34) and European COST guidelines (108). They included multiple studies without defining inclusion or exclusion criteria and failed to include 2 systematic reviews. The methodologic quality assessment of the studies revealed inappropriate quality criteria (98,101). The methodologic quality assessment of systematic reviews was also inappropriate. They based conclusions on multiple controversial studies which failed to meet criteria in the other systematic reviews, none of the studies were performed utilizing IASP criteria, and they excluded high quality studies (111,113). Finally, they based their evaluation on flawed outcome studies to evaluate the gold standard (105,107). It is surprising that the Health Technology Assessment of Washington State (134) used different studies (106,137,138).

One of the major flaws of Chou and Huffman (35) is the exclusion of the only 2 studies (111,113) performed according to IASP criteria that used a control disc. Schwarzer et al (111) evaluated the prevalence of internal disc derangement, whereas Manchikanti et al (113) evaluated discogenic pain with results yielding a 39% prevalence of internal disc disruption and 26% prevalence of discogenic pain. The exclusion of both of these studies questions the integrity of the evaluation provided by Chou and Huffman (35). Further, they also excluded multiple studies by Derby et al (99,100,114). Carragee et al (93) analyzed and reanalyzed their data on multiple occasions providing poor quality studies which were included by Chou et al. The reanalysis of Carragee et al’s (93) data, along with the inclusion of Derby et al’s (98-100,114) data provides a different picture as shown by Wolfer et al (56) who showed low false-positive rates.

It appears that there is general confusion regarding placebo control, not just in therapeutic trials, but also that placebo-controlled neural blockades are not viable, even though they have been misinterpreted (276-283). It is a common practice in interventional pain management, especially by those with a lack of understanding and bias, to report any local anesthetic injection as a placebo. These interpretations are inaccurate. Further, even the differences between various types of placebo injections have been demonstrated, as well as injections into various structures. The experimental and clinical findings from investigation of the electrophysiologic effects of 0.9% sodium chloride solution and dextrose 5% in water solution illustrate the potential inaccuracy created by 0.9% sodium chloride solution versus 5% dextrose (280,281). In addition, the effect of sodium chloride solution when injected into either the disc, the facet joint, or paraspinal muscles have shown to be variable. Indahl et al (282,283) in their study of the electromyographic response of the porcine multifidus musculature after nerve stimulation (283) and interaction between the porcine lumbar intervertebral disc, zygapophysial joint, and paraspinal muscles (282), showed that stimulation of the disc and the facet joint capsule produced contractions in the multifidus fascicles (283). In addition, they demonstrated that the introduction of lidocaine into the facet joint resulted in a significantly reduced electromyographic response with the most drastic reduction seen when stimulating the facet joint capsule. They (282) also showed that the introduction of physiologic saline into the zygapophysial joint reduced the stimulation pathway from the intervertebral disc to the paraspinal musculature. Consequently, they hypothesized that the paraspinal muscle activation caused by nerve stimulation in the annulus fibrosis of the lumbar intervertebral disc could be altered by saline injection into a zygapophysial joint. Thus, it is essential to rule out zygapophysial joint pain prior to discography.

Consequently, the reassessment resulted in “fair” evidence for lumbar provocation discography utilizing the criteria established by Chou and Huffman (35).

**Diagnostic Lumbar Facet Joint Nerve Blocks**

Chou and Huffman (35) recommends against diagnostic intraarticular facet joint blocks and medial branch blocks based on European COST guidelines (108) and the inclusion of a single study by Birkenmaier et al (148) which was poorly performed. Chou and Huffman (35) decided to demonstrate the accuracy of diagnostic facet joint nerve blocks. Manchikanti et al (46) summarized the rationale for using facet joint blocks for diagnosis which is based on the fact that lumbar facet joints have a nerve supply and are capable of causing pain (284-294), and that they have been shown to be a source of pain in patients using diagnostic techniques of known reliability (46,48,112,150-175). Finally, the value, validity, and clinical effectiveness of diagnostic facet joint
nerve blocks has also been demonstrated by applying therapeutic modalities based on a diagnosis utilizing controlled comparative local anesthetic blocks (46,190,245,295-298). The face validity and construct validity of facet joint blocks has been evaluated repeatedly and has been affirmed (153,286,287,299-305). Based on the systematic review by Datta et al (54), a prevalence of 21% to 40%, with an overall rate of 31%, was determined in heterogeneous populations with chronic low back pain, with false-positive rates of 17% to 19% and an overall prevalence of 30% with single blocks. The deficiencies of Chou and Huffman (35) is that even though there were 7 high quality studies (113,154,156,158,161,162) meeting inclusion criteria performed according to modified IASP criteria of 80% pain relief, they chose not to utilize them based on one poorly performed study (148) assessing outcomes. They applied the same philosophy to all diagnostic interventions for which there was no gold standard and attempted to establish a gold standard by utilizing unproven techniques. Further, they ignored long-term follow-up criteria as a gold standard. They also exhibited substantial bias when evaluating systematic reviews. Ironically, Rubinstein and van Tulder (124), utilizing the same criteria they quote as the basis – namely, European COST guidelines (108) – have provided strong evidence for the diagnosis of lumbar facet joint pain utilizing the same systematic review by Boswell et al (150) and Sehgal et al (151). They also ignored the systematic review by Hancock et al (152) which essentially showed the value of diagnostic blocks and a lack of value for all other investigations. Further, the 7 studies utilized a total of 1,320 patients to study diagnostic accuracy. In addition, there have been multiple publications evaluating confounding factors (164,191), along with a systematic review (64) evaluating sedation. As a result, the influence of psychological factors, opioid exposure, and age have been evaluated appropriately (165-168).

The literature has been replete with studies illustrating a lack of correlation between clinical evaluation, radiological findings, and nerve conduction studies. Rubinstein and van Tulder (124) commented that it is quite remarkable that while named orthopedic tests of the low back are often illustrated in orthopedic textbooks, there is little evidence to support their diagnostic accuracy, and therefore their use in clinical practice. Consistent with clinical experience, many studies have demonstrated that physical examination serves primarily to confirm suspicions raised during the history.

Thus, based on this reassessment, the evidence for diagnostic lumbar facet joint nerve blocks is “good.”

Diagnosis of Sacroiliac Joint Pain

Sacroiliac joint blocks are recommended for diagnosing sacroiliac joint pain because diagnosing sacroiliac joint pain cannot be done with non-invasive tests. It has been described that sacroiliac joint blocks are the evaluation of choice to provide an appropriate diagnosis. Multiple studies have evaluated the value and validity of sacroiliac joint blocks in diagnosing chronic sacroiliac joint pain. Consequently, the face validity, as well as construct validity, has been established, at least initially, by determining prevalence and false-positive rates in specific population groups. It also has been recognized that the sacroiliac joint is one of the structures which can leak through the joint capsules, resulting in false-positive results; this can also cause false-negative results due to faulty needle placement, intravascular injection, and, finally, the inability of the local anesthetic to reach the painful portion of the joint due to loculations. In the analysis of sacroiliac joint nerve blocks, Chou and Huffman (35), utilizing Hansen et al’s (194) study, which they rated as low quality, reached inappropriate conclusions. Further, they also excluded multiple appropriate studies meeting the inclusion criteria. Ironically, Rubinstein and van Tulder (124), in a best evidence review of diagnostic procedures for low back pain that focused on a previously published systematic review (194) which was published earlier than the one used by Hansen et al, concluded that there was moderate evidence for diagnostic sacroiliac joint blocks based on European COST Guidelines.

Hancock et al (152), in their systematic review evaluating a battery of tests to identify the sacroiliac joint as one of the sources of low back pain, showed that a combination of sacroiliac joint pain provocative maneuvers appear to be useful in pinpointing the sacroiliac joint as the principal source of symptoms in patients with pain below the 5th lumbar vertebra. However, they also concluded that although a positive bone scan has high specificity, it is associated with a very low sensitivity, which means that the majority of patients with sacroiliac joint pain will not be accurately identified. Szadek et al (198) evaluated the diagnostic validity of the IASP criteria for sacroiliac joint pain in a meta-analysis showing that the thigh thrust test, the compression test, and 3 more positive stressing tests contain sufficient discriminative power for diagnosing sacroiliac joint pain. Even so, they also included, acknowledging the lack of a gold standard for
sacroiliac joint pain, that the diagnostic validity of tests for sacroiliac joint pain should be regarded with caution. Song et al (197), in a systematic review evaluating the diagnostic value of scintography in assessing sacroiliitis and ankylosing spondylitis, concluded that scintography is at best of limited value in establishing a diagnosis of ankylosing spondylitis.

Manchikanti et al (76), in the reassessment of ACOEM practice guidelines, concluded there was moderate evidence for sacroiliac joint blocks in the diagnosis of sacroiliac joint pain. Finally, in a more recent evidence synthesis, which was published after the publication of Chou and Huffman (35), Rupert et al (68), utilizing 50% relief as the criterion standard with ability to perform previously painful maneuvers, showed moderate evidence with the prevalence of sacroiliac joint pain ranging between 10% and 38% using a double block paradigm in the study population; estimated false-positive rates of single, uncontrolled, sacroiliac joint injections were 20% to 54%.

Thus, based on reassessment, utilizing Chou et al’s criteria, the evidence is “fair to poor.”

Assessment of Integrity
As discussed in the results section, the goals of integrity include transparency, accountability, consistency, and independence. Thus, none of these criteria were met in their evaluation. The integrity seems to have not been applied uniformly because of the lack of clear information with a multitude of conflicts of interest, financial, as well as academic.

Transparency is lacking in many aspects.

The accountability has not been provided appropriately as conflict of interest policies with the prioritization of the literature peer review of evidence synthesis and recommendations and updating of the recommendations consistent with the current literature were not appropriately utilized.

There was no consistency as literature was reviewed inappropriately with inclusion and exclusion criteria based on convenience rather than a standardized approach.

Finally, the independence was breached as at least 2 of the involved experts in interventional techniques withdrew their names from publication and this was not publicized.

CONCLUSION
The reassessment of any and all guidelines is a crucial part of modern medicine. Chou et al have performed an extensive analysis; however, due to inappropriate evaluations, the inclusion of inappropriate studies, and the exclusion of appropriate studies. Their evidence synthesis reached conclusions that appear to be not of as high a quality as they are illustrated to be.

ACKNOWLEDGMENTS
The authors wish to thank Sekar Edem for his assistance in the literature search; Tom Prigge for manuscript review, and Tonie M. Hatton and Diane E. Neihoff, transcriptionists, for their assistance in preparation of this manuscript. We would like to thank the editorial board of Pain Physician for review and criticism in improving the manuscript.

DISCLOSURES
Dr. Hirsch is a consultant for Cardinal Healthcare. He is a minor shareholder in Medtronic and Cardinal Healthcare. He serves on the Steering Committee for KAVIAR trial (volunteer position), and on the Data and Safety Monitoring Board (DSMB): CEEP trial (volunteer position). Dr. Datta receives research support from Sucampo Pharmaceuticals and an honorarium from Smith and Nephew. Dr. Derby has stock options with Laurimed and Medtronics.

REFERENCES
7. Manchikanti L. Evidence-based medicine, systematic reviews, and guide-


65. Frey ME, Manchikanti L, Benyamin RM, Schultz DM, Smith HS, Cohen SP. Spi-
Critical Review of APS Clinical Practice Guidelines for Interventional Techniques


Critical Review of APS Clinical Practice Guidelines for Interventional Techniques


