Executive Summary

This technology assessment was commissioned by the Washington State Health Technology Assessment Program for use by the Health Technology Clinical Committee (HTCC). The HTCC uses evidence, primarily as assessed in this report to determine whether health technologies are safe, effective, and cost effective, and therefore should be covered by state programs that pay for health care.

This report evaluates relevant published research describing use of lumbar fusion and discography in patients with chronic low back pain and uncomplicated degenerative disc disease (DDD). ECRI Institute's technology assessment provides an independent, in-depth, formal evaluation of the strength of evidence for the safety and efficacy of lumbar fusion for the treatment of DDD associated with chronic low back pain. This report also evaluates the role of discography prior to lumbar fusion in this patient population. It is based on systematic review of the published, peer-reviewed scientific literature and methodological precepts described in Appendix C.

The word "uncomplicated" in the title of this report is intended to exclude the following conditions:

- Radiculopathy
- Functional neurologic deficits (motor weakness or EMG findings of radiculopathy)
- Spondylolisthesis (>Grade 1)
- Isthmic spondylolysis
- Primary neurogenic claudication associated with stenosis
- Fracture, tumor, infection, inflammatory disease
- Degenerative disease associated with significant deformity

Therefore, the conclusions of this report are not necessarily applicable to patients with any of the above conditions.

The degeneration of intervertebral discs is thought to be associated with altered biomechanics of adjacent vertebrae, musculature, and connective tissue, and can be associated with back pain and sciatica.(1) Discs are present between lower cervical (neck) vertebrae, thoracic (mid-back) vertebrae, and low back (lumbar) vertebrae. Discs at any level can degenerate and cause pain, but this most often occurs at cervical and lumbar levels, where there is the greatest amount of mobility. Patients with DDD in the absence of chronic low back pain would not be considered candidates for lumbar fusion.

The clinical presentation of low back pain may prompt the clinician to order diagnostic imaging. Since disc degeneration is not associated with pain in all individuals, imaging alone cannot be considered diagnostic. Both plain films and magnetic resonance imaging (MRI) may aid the clinician in confirming their diagnosis. Typical findings suggestive of discogenic pain include disc space collapse, endplate sclerosis, and vacuum disc phenomenon.(2) On a MRI, disc dehydration, high intensity zones, and endplate edema may also be evident.(2) However, there is currently no clear case definition for "discogenic back pain".

Low back pain was the most common cause of disability in persons younger than 45 in the U.S. in 2005.(3) It causes the most loss of productivity of any medical condition.(3) Only upper respiratory complaints cause people to miss more days of work annually.(3) In the United States, an estimated range of 8-56% of the population (the reason for this variation is unclear, but may be due to differences in diagnostic criteria or definition) experiences lower back pain every year, and the lifetime incidence rate is reportedly between 65% and 80%.(4) 2.4 million people are disabled because of low back pain, 1.2 million of them chronically.(3) Most patients improve within weeks; only 5-10% of people with low back pain develop chronic back pain.(4,5)

Chronic low back pain with DDD is typically managed conservatively for at least six months before surgery is considered. Rest is usually only recommended for the first couple days of onset.(3) A variety of conservative treatments can be tried, including back education, cognitive behavioral therapy, physical therapy, exercise, weight reduction, and alternative therapies (e.g., chiropractic manipulation), medications, and epidural injections.(2,6)

When conservative treatments fail after at least six months, spinal fusion may be considered. The goal of spinal fusion (also known as spinal arthrodesis) is to permanently immobilize the spinal column vertebrae surrounding the disc(s) that is (are) diagnosed as causing discogenic low back pain. Immobilizing the vertebrae is believed to reduce pain by limiting painful movement that may occur as degenerated discs subside. Spinal fusion is also used to treat other painful conditions, including spondylolisthesis (forward displacement of one of the lower lumbar vertebrae over the vertebra below it or on the sacrum), trauma resulting in spinal nerve compression, abnormal spinal curvatures (scoliosis or kyphosis), and vertebral instability caused by infections or tumors. Vertebral instability refers to a range of motion in the vertebrae that is greater than that of a normal range. Several surgical procedures may be used to achieve spinal fusion in patients with discogenic low back pain. They differ by surgical approach and instrumentation used. All methods may have advantages and disadvantages.

Approximately 300,000 people in the U.S. underwent lumbar spinal fusion surgery for any indication in 2001;(7) over 122,000 lumbar fusions were performed for degenerative conditions.(8) A retrospective cohort study of the Healthcare Cost and Utilization Project (HCUP) Nationwide Inpatient Sample (NIS) from 1988 through 2001 indicated that rates of lumbar fusion for degenerative conditions in the U.S. increased substantially during this period (220% increase from 1990 to 2001). The increase accelerated after 1996 following FDA approval of fusion cages; between 1996 and 2001, the number of lumbar fusions increased 113%, with the most rapid increase among patients of age 60 or above.(8)

A recent cross-sectional analysis of national Medicare data has revealed a substantial (close to 20-fold) range in regional rates of lumbar fusion (for any indication) among Medicare enrollees in 2002 and 2003.(9) The wide variation observed in this study may be due to scientific uncertainty regarding the evidence for lumbar fusion and a lack of consensus among surgeons on patient selection criteria and indications for lumbar fusion. Despite such uncertainty, the rates of lumbar fusion in the Medicare population increased nearly four-fold from 1992 to 2003.(9)The variation in regional rates of lumbar fusion among non-Medicare enrollees is unknown.

The role of lumbar discography in selection of patients as surgical candidates is controversial. Discography is a diagnostic procedure in which contrast material is injected into the nucleus pulposus of a lumbar disc. The general intent is to determine whether the disc itself is the source of pain (i.e., a diagnosis of discogenic pain). This diagnosis has been used to justify the need for surgical intervention involving discectomy and lumbar fusion. Thus, discography may influence important decisions about the appropriateness of surgical intervention.

Discography yields two types of results: pain provocation (whether the patient's typical pain was reproduced by the injection), and morphology (whether the dye images an abnormal pattern in the disc, often based on CT scan). Controversy exists about the relative importance of these two test results. Some authors(3,10) assign much greater importance to pain provocation; for example, Bogduk (1996)(10) stated that "the morphology of the disc as revealed by discography is essentially irrelevant." By contrast, Buenaventura et al. (2007) cited disc morphology as the gold standard for discogenic pain, stating that "the imaging information is important since treating an anatomically normal disc, irrespective of its ability to cause pain, seems unethical."(11) Walsh (1990) proposed that a discography result should only be considered positive if the patient's typical pain was reproduced *and* the morphology was abnormal.(12) The extent of spread of the contrast material from the nucleus pulposus determines disc morphology. The Dallas Discogram Description categorizes several levels of disruption of the disc annulus, ranging from Grade 0 (normal) to Grade 5 (highest level of disruption).(13,14)

One major concern about discography is the rate of false positive results. Several authors have found that among people with no previous pain, the discography result can be positive.(15-22) Also, discography in lumbar discs has been reported to reproduce pain known to originate elsewhere in the body.(23) Various solutions have been proposed for these phenomena, including a more stringent definition of a positive test to require both typical pain provocation and abnormal morphology (Walsh definition),(12) the requirement that adjacent discs test negative,(24,25) and the avoidance of high pressure (≥22 pounds per square inch).(16) Carragee et al. (2006) found, however, that even when all of these conditions were met, the rate of false positives was still 25%.(26) Many have suggested that the origin of many false positives lies with the psychological status of some patients; a positive discography may be more likely in patients with psychological comorbidities who are predisposed to report pain.(17-20,22,23,27,28)

The analysis of evidence in this assessment is divided into two sections: Part I evaluates evidence comparing outcomes of lumbar fusion and nonsurgical treatments, while Part II evaluates evidence concerning the role of discography prior to lumbar fusion. We examined the evidence in the context of six clinical questions (three for Part I and three for Part II). Our strength of evidence ratings take into consideration not only the individual study quality for relevant outcomes, but also the quantity, consistency, and robustness of the evidence, in addition to the magnitude of observed effects. The instruments used to rate individual study quality appear in Appendix C, along with our system for rating the strength of evidence.

Part I – Lumbar fusion surgery and nonsurgical treatments for chronic lumbar back pain

1) Does lumbar fusion surgery reduce pain and improve functional status/quality of life more effectively than nonsurgical treatments?
2) What are the rates of adverse events (perioperative, long-term events, and reoperations) for lumbar fusion surgery and nonsurgical treatments?
3) What patient characteristics (i.e., workers' compensation population, patients with chronic pain, psychological distress, and age-groups) are associated with differences in the benefits and adverse events of lumbar fusion surgery?

Part II – Role of discography prior to lumbar fusion surgery

4) In patients being considered for lumbar fusion surgery, what is the reliability of discography?
   a. Test-retest reliability
   b. Inter-reader reliability
5) In patients undergoing lumbar fusion surgery, do the results of pre-surgical discography predict the degree of pain reduction or improvement in functional status/quality of life after lumbar fusion surgery?
6) In patients being considered for lumbar fusion surgery, do patients who receive discography that influences the treatment choice have better treatment outcomes than patients who do not receive discography?

Part I - Lumbar fusion surgery and nonsurgical treatments for chronic lumbar back pain

Overall, 30 articles reporting on 27 studies were included to address the clinical questions in Part I. Four randomized controlled trials (RCTs) that enrolled a total of 767 patients met the inclusion criteria for Key Question 1, which required a comparison of lumbar fusion to non-operative treatment in patients with DDD. These same RCTs also reported treatment complications and therefore also met the inclusion criteria for Key Question 2. In addition to the four RCTs described above, 23 studies with a total of 5,639 patients also met the inclusion criteria for Key Question 2. These studies were either case series of lumbar fusion or controlled studies (some randomized) that compared different lumbar fusion procedures. Data from one separate publication of one RCT (also included in Key Question 1 and 2) that enrolled 294 patients met the inclusion criteria for Key Question 3.

The primary outcomes of interest addressing Key Question 1 are functional status measured by the Oswestry Disability Index (ODI), back pain measured by a visual analog scale (VAS), and quality of life measured by a previously validated instrument; the only instrument used to measure quality of life in the available evidence base was the short-form (SF)-36 questionnaire. The ODI is comprised of 10 questions on pain and pain-related disability in activities of daily life and social participation. Each question has six response alternatives, and the overall score ranges from 0 (no disability) to 100 (totally disabled or bedridden). The VAS for back pain is also scored from 0 (no pain) to 100 (worst pain imaginable). A recent study calculated the minimal clinically important difference for the ODI and VAS of back pain using linear regression analysis of score change compared to pre-treatment scores. The authors determined that the minimal clinically important difference for the ODI was 10, and for the VAS of back pain it was 18-19.(29) Accordingly, we used a difference of 10 for the ODI and a difference of 20 for the VAS as the minimal clinically important difference in our assessment of these outcomes (the FDA required an ODI change of 15). Although other estimates for clinically important change in ODI have ranged from 4 to 18.4 in other studies,(30-32) we consider the estimates in this study to be the best empirical estimates of clinically important change in ODI and VAS (for further discussion of the issues surrounding clinically important change thresholds, see Methods in the main text). The SF-36 is scored from 0 (worst health state) to 100 (best health state); we used a difference of 5 in the SF-36 as the minimal clinically important difference based on data from an earlier study that investigated this issue.(33)

A quality rating (and strength of evidence rating) was applied only to studies comparing lumbar fusion to non-operative therapy in Key Question 1. The remaining studies addressing Key Question 2 were not used to address comparative event rates of fusion and non-operative care; they were used only to provide additional data on adverse events and adverse event rates for lumbar fusion. Due to variability in the way complications are reported among different studies, lists of complications do not lend themselves to evidence ratings.

Our detailed assessments of the quality of the RCTs addressing Key Question 1 appear in Table 13 of Appendix D. The average quality of the studies was moderate due to several limitations, most notably lack of blinding of patients, providers, and outcome assessors (for the majority of outcomes) in all studies. This could lead to biased interpretation or reporting of outcomes, particularly of subjective outcomes; since placebo effects tend to be stronger with more invasive interventions, lack of blinding may be more likely to create bias favoring better outcomes with surgery. Two of the studies were further limited because more than 15% of patients did not receive their assigned treatment, either because they crossed over to the alternative treatment group or did not receive any of the trial treatments. Crossover to alternative treatments would tend to diminish a between-group difference in treatment outcome if it exists. Another potential limitation was differences between groups in additional treatments received during the trials (most trials did not record this information).

The average age of patients in all four RCTs was about 40-45 years, and the average age of patients in the additional 23 studies that addressed Key Question 2 ranged from 39 to 54 years, which is representative of the age at which most patients with degenerative disease undergo surgery in clinical practice. The proportion of patients receiving workers' compensation varied considerably (ranging from 21% to 94%) in the 12 studies that reported this information. Although the types of fusion procedures varied among different studies, all studies used fusion procedures that are currently employed in clinical practice.

Results and conclusions (Part I)

1. Does lumbar fusion surgery reduce pain and improve functional status/quality of life more effectively than nonsurgical treatments?

   ECRI Institute evidence assessments:

   We did not find sufficient evidence that lumbar fusion surgery is more effective to a clinically meaningful degree than nonsurgical treatments for any of the following patient populations, comparisons and outcomes:

   Fusion versus Intensive Exercise/Rehabilitation Plus CBT in Patients without Prior Back Surgery

   - Meta-analysis of postoperative changes in Oswestry disability scores from two moderate quality RCTs (n = 413 patients) revealed no clinically meaningful difference between fusion and intensive exercise/rehabilitation plus cognitive behavioral therapy (CBT) in patients without prior back surgery (95% CI 0.2 to 7.5, *a priori* 10-point difference defined as clinically meaningful), although the difference slightly favored fusion. Strength of evidence: Weak.
   - The evidence was insufficient to determine whether lumbar fusion provides a greater improvement in back pain (one moderate-quality RCT, n = 64 patients) or

quality of life (no acceptable evidence) compared to intensive exercise/rehabilitation plus CBT in patients without prior back surgery.

Fusion versus Intensive Exercise/Rehabilitation Plus CBT in Patients with Prior Back Surgery

- The evidence from one moderate-quality RCT (n = 60 patients) was insufficient to determine the relative benefits of lumbar fusion compared to intensive exercise/rehabilitation in patients with prior back surgery.

Fusion versus Non-intensive Physical Therapy in Patients without Prior Back Surgery

- The evidence from one moderate quality RCT (n = 294 patients) was insufficient to determine the relative benefits of lumbar fusion compared to conventional physical therapy in patients with or without prior back surgery.

The four trials that met our inclusion criteria for this question differed in potentially important ways. Based upon independent assessment by two methodologists, we assumed that one difference that was likely to create variation in the effect size among trials was the intensity of non-operative therapy in the control groups. Three trials (Brox et al. 2003; Brox et al. 2006; Fairbank et al. 2005) used more intensive exercise/rehabilitation with cognitive behavioral strategies, while the remaining trial (Fritzell et al. 2001) used non-intensive physical therapy as the main component of an unstructured nonsurgical treatment program. The more intensive therapy seems more likely to benefit patients than the less intensive treatment (which patients had undergone without improvement prior to enrollment). If the amount of patient benefit from surgery is assumed to be the same in all studies, then one would expect a greater difference in patient benefit between patients treated surgically and patients treated with conventional physical therapy compared with patients treated surgically and patients treated with multidisciplinary and intensive exercise/rehabilitation. This is important to our analysis because the mean difference measures the difference between treatment and control groups. Therefore, the mean difference would vary depending on the control selected, causing heterogeneity (differences) in study findings. For this reason, the data from Fritzell et al. were not combined with data from the other three trials.

Another factor that might create heterogeneity among effect sizes is whether the patients had back surgery before enrolling in the studies in question. Patients with prior back surgery may be less likely to benefit from further surgery than patients who have never had back surgery. One of the three trials that used intensive exercise/rehabilitation (Brox et al. 2006) included only patients who had undergone prior surgery for disc herniation (most likely discectomy or laminectomy, as none of the patients had undergone prior lumbar fusion). The authors mentioned that "the prognosis after a second operation is generally considered poor compared with the prognosis in patients without previous surgery for disc herniation."(34) Of the remaining two trials, Brox et al. (2003) included no patients with prior back surgery, while Fairbank et al. (2005) had a small proportion of patients (8%) who had undergone prior laminectomy. Based upon the differences in the patient populations, we determined that the data from Brox et al. (2006) should not be combined with data from the remaining two trials.

Although the control therapies and patient characteristics were similar in the trials by Brox et al. (2003) and Fairbank et al. (2005), the two trials differed in the types of fusion performed and the length of followup. Brox et al. (2003) exclusively used posterolateral fusion (PLF) with pedicle screws, while Fairbank et al. (2005) used an unspecified variety of fusion procedures. Also, Brox et al. reported treatment outcomes at one year of followup, while Fairbank et al. reported treatment outcomes at two years of followup. However, we considered differences in the fusion procedure and length of followup less likely to create heterogeneity in effect sizes than the other factors described above. Therefore, we determined that combining the data from these two trials was appropriate.

The four RCTs were therefore analyzed in three separate groups: fusion versus intensive exercise/rehabilitation plus CBT – divided into patients without prior back surgery (Brox et al. 2003, Fairbank et al. 2005) and patients with prior back surgery (Brox et al. 2006) – and fusion versus non-intensive physical therapy (Fritzell et al. 2006).

Fusion versus Intensive Exercise/Rehabilitation Plus CBT in Patients without Prior Back Surgery

Two multicenter RCTs with a total of 413 patients compared intensive exercise/rehabilitation with cognitive behavioral therapy to fusion in patients who had not undergone back surgery before. Both studies reported the between-group difference in the pre-post change in ODI score (see Brox et al. 2003 and Fairbank et al. 2005 in Table 17, Appendix D). Both studies also reported the change scores adjusted for baseline values by analysis of covariance (ANCOVA); this is the best method for adjusting for imbalances in patient characteristics.(35) Thus, our analysis is based on the adjusted change scores.

As described above, these studies were considered suitable for a combined data analysis (meta-analysis), so the change score data were combined in a random effects meta-analysis. As shown in Figure 2, fusion led to a small but statistically significant increase in ODI change scores compared to intensive exercise/rehabilitation plus CBT; however, the upper 95% confidence limit (7.5) was below the minimum level that is considered clinically significant (ODI = 10). We therefore conclude that changes in ODI scores did not show a clinically meaningful difference between fusion and intensive exercise/rehabilitation plus CBT in patients without prior back surgery, although the difference slightly favored fusion (95% CI 0.2 to 7.5). Because the evidence base is of moderate quality and limited quantity, the strength of evidence supporting this conclusion is weak.

Only one of these studies (Brox et al. 2003) evaluated VAS back pain (Table 18, Appendix D). This study reported no statistically significant difference in change in VAS scores between patients undergoing fusion and patients undergoing intensive exercise/rehabilitation plus CBT. Because the 95% CI overlapped with zero and the boundary of minimum clinical significance, the evidence is insufficient to allow a conclusion for this outcome.

Although one of these studies measured quality of life using the SF-36 instrument, this outcome was excluded from analysis because <80% of patients completed the instrument.

Fusion versus Intensive Exercise/Rehabilitation Plus CBT in Patients with Prior Back Surgery

One RCT (Brox et al. 2006) with 60 patients studied the efficacy of exercise/rehabilitation plus cognitive behavioral therapy to fusion in patients who had previously undergone back surgery. This study reported the between-group difference in the pre-post change in ODI score, after statistically adjusting for baseline between-group differences in gender and treatment expectations (see data in Table 17, Appendix D). However, the results were inconclusive because the 95% CI overlapped with zero (not statistically significant) as well as the boundary of clinical significance (ODI = -10), meaning the true difference (if one exists) could favor either treatment. Thus, the evidence is insufficient for a conclusion regarding the relative benefit of fusion versus intensive exercise/rehabilitation plus CBT in patients with prior back surgery.

This same study reported no statistically significant difference in change in VAS scores between patients undergoing fusion and patients undergoing intensive exercise/rehabilitation plus CBT (Table 18, Appendix D). The results of a single moderate quality study are insufficient to allow a conclusion for this outcome.

Fusion versus Non-intensive Physical Therapy in Patients without Prior Back Surgery

One RCT (Fritzell et al. 2001) with 294 patients addressed this comparison; however, a minority of patients (18.7%, evenly distributed between groups) had prior discectomy. This study reported ODI pre-post change scores for each comparison group (see data in Table 17, Appendix D). A significantly larger improvement in ODI was observed in the fusion group compared to the physical therapy group (11.6 vs 2.8, p = 0.015); group changers were included in the analysis of difference (although not in their tabled data). However, although the difference in change is statistically significant, the mean difference in change between groups (ODI = 8.8) is below the level of clinical significance (ODI = 10). Because this is a single trial of moderate quality, the evidence is insufficient to allow a conclusion for this comparison.

This same study reported a statistically significant difference in the change in VAS score favoring fusion when compared to non-intensive physical therapy (Table 18, Appendix D). However, the mean difference between groups (16.7) did not exceed the boundary of minimum clinical significance for VAS back pain (difference = 20). Because this study did not include group changers in their tabled data, we cannot be certain of the difference if group changers had been included. In any event, because this is a single study of moderate quality without a large effect (see Appendix C, Strength of Evidence Algorithm, Decision Point 10 for definition of large effect), the evidence is inconclusive for this outcome.

2. What are the rates of adverse events (perioperative, long-term events, and reoperations) for lumbar fusion surgery and nonsurgical treatments?

   - Lumbar fusion leads to higher rates of both early and late adverse events compared to non-intensive physical therapy or intensive exercise/rehabilitation plus CBT.

   - None of the four RCTs comparing fusion to non-intensive physical therapy or intensive exercise/rehabilitation plus CBT reported any adverse events occurring in patients who only received non-operative care. Most of the reported adverse events for patients in the surgical group could not have occurred in patients who did not undergo surgery (e.g., surgical complications).

   - Categories of adverse events most frequently reported in fusion studies include reoperation (18/27 studies), infection (14/27 studies), various device-related complications (13/27 studies), neurologic complications (12/27 studies), thrombosis (11/27 studies), bleeding/vascular complications (10/27 studies), and dural injury (10/27 studies).

- The ranges of rates of the most frequently reported complications in fusion studies were: reoperation (0% to 46.1%), infection (0% to 9%), device-related complications (0% to 17.8%), neurologic complications (0.7% to 25.8%), thrombosis (0% to 4%), bleeding/vascular complications (0% to 12.8%), and dural injury (0.5% to 29%).

All four RCTs with 767 patients that met our inclusion criteria for Key Question 1 compared adverse event rates for lumbar fusion surgery and nonsurgical treatments. None of the trials reported the rate of total adverse events (from intraoperative to last followup). Instead, they generally divided complication rates by time of occurrence.

Two trials (Brox et al. 2003, Fritzell et al. 2001) separately reported "early" (usually meaning perioperative) and "late" complications (which either occur at a later time or are persistent or permanent). Fritzell et al. defined early as within the first two weeks post-treatment, while Brox et al. did not report the cutoff time for early complications (although it likely did not exceed one month). Another trial by Brox et al. (2006) appeared not to report all early complications; the authors stated that "early complications included two wound infections among the 23 operated patients", but no other early complications are mentioned. Thus, we cannot be certain that these were the only early complications. However, the authors stated that no late complications occurred. The remaining trial (Fairbank et al. 2005) divided adverse events into intraoperative (during surgery) and post-operative (any time after surgery) categories, which is a somewhat different division than early and late. The only postoperative complications mentioned were need for reoperation; we cannot be certain that there were no late complications that did not require reoperation.

All trials calculated adverse event rates on a per protocol basis, meaning only patients who actually received surgery were included in calculations of surgical adverse events. This is the most conservative approach for analysis of adverse events; calculations on an intent-to-treat basis would underestimate the surgical complication rate, as some patients assigned to surgery never received it.

Overall Early Adverse Events

The results for overall early adverse events appear in Table 19, Appendix D. Despite variation in types of fusion and nonsurgical therapies used in these studies, the four trials had one factor in common; none of them identified any adverse event resulting from nonsurgical treatment (intensive exercise/rehabilitation plus CBT in three trials, non-intensive physical therapy in one trial). The three trials that reported overall intraoperative or early adverse event rates found similar rates (range 12.7% to 18%) despite differences in the time period observed (intraoperative to one month). The differences between early adverse events in the surgical versus physical therapy groups was statistically significant in all three of these trials. The reported early adverse events in the surgical groups included bleeding, thrombosis, wound infection (deep and superficial), neurological (pain, sympathetic cord damage) complications, device-related (problems with screws or implants) complications, reoperations for various causes, and others (dural tears, peritoneal tears). A complete list of reported early complications and their occurrence rates in these trials appears in Table 21, Appendix D (note: some complications in this table may not be early; most studies did not report time cutoffs for the complications). Most of these complications could not have occurred in the absence of surgery.

Overall Late Adverse Events

Overall late adverse event rates showed more variation among studies, ranging from 0% to 7.4% (Table 19, Appendix D). A number of factors might account for this variation. It could have resulted from differences in the length of followup; the two trials with only one-year followup reported no late events, while the two trials with two-year followup reported that 6.2% and 7.4% of patients who underwent fusion had late events (in both trials, the difference in event rates between surgical and nonsurgical patients was statistically significant). The size of the trials may also have influenced these differences, as the two trials with one-year followup were also much smaller than the other trials, and therefore less likely to detect less common adverse events. A third factor is that the authors of these trials may have had different definitions of what constitutes an adverse event. Reported late adverse events most frequently included reoperations for various problems (mostly infections and pseudoarthroses) and continuing pain at the donor site from bone graft harvesting. Specific causes of reoperations and other late complications and their rates are listed in Table 22, Appendix D. Again, these events could not have occurred in the absence of surgery.

We examined additional studies of lumbar fusion that lacked a non-operative control group to determine whether these studies report adverse events not reported in the four RCTs described above, and also to determine if the adverse event rates differed from those reported in the RCTs. We selected studies with at least 100 patients total that received any type of lumbar fusion procedure and met all of our other inclusion criteria.

Twenty-three studies with a total of 5,639 enrolled patients met our criteria for this question. Fourteen of these studies were prospective studies; of these 14, six were randomized trials comparing different fusion procedures (a comparison not addressed in this report). The remaining studies were retrospective. Some studies focused only on specific adverse events such as need for reoperation, while others reported all adverse events that occurred during the course of the study. Only eight studies reported any type of overall adverse event rates (operative, postoperative, total, etc.), and the studies varied considerably in the manner in which these events were summarized (Table 20, Appendix D). Because a patient may experience more than one adverse event, we could not calculate the percent of patients experiencing any adverse event when studies only reported rates for specific adverse events. These studies also showed considerable variation in the types of fusion procedures performed, which may contribute to variation in the types of adverse events that occurred in different studies.

A concise summary of reported ranges of specific adverse event rates appears in Table 4. These ranges combine data from the four RCTs described earlier with data from the 23 additional studies. In this table, we do not attempt to separate early from late events, as several studies did not report the specific time of occurrence for each event. Categories of adverse events most frequently reported in fusion studies include reoperation (18/27 studies), infection (14/27 studies), neurologic complications (12/27 studies), thrombosis (11/27 studies), bleeding/vascular complications (10/27 studies), and dural injury (10/27 studies). Death related to surgery was relatively rare, occurring only in 4/27 studies with a maximum reported rate of 2% (we assumed no deaths related to surgery occurred in the other 23 studies). Certain adverse events showed substantial variation in reported rates: these include reoperation (0% to 46.1%), dural injury (0.5% to 29%), neurologic complications (0.7% to 25.8%), and device-related complications (0% to 17.8%). Reported rates in the four RCTs comparing fusion to non-operative care were either at the low end (0% for death) or within the indicated ranges but below the maximum reported rate.

An important issue with reoperation is whether the reoperation was due to problems related to the initial surgery. Of the 17 studies that reported reoperations, 13 reported the specific causes for reoperation. In these 13 studies, the percentage of reoperations that could be definitely determined to be directly related to the initial surgery ranged from 61% to 100% (in five studies, 100% of reoperations were directly related to the initial surgery, and in three studies more than 90% of reoperations were directly related to the initial surgery). The possibility exists that some reoperations that could not be definitely attributed to the initial surgery (e.g., reoperation at another level) were nevertheless related to the initial surgery, so these estimates are conservative.

Complete information on the rates of all adverse events reported in these studies is summarized in Tables 23 and 24, Appendix D.

3. What patient characteristics (i.e., workers' compensation population, patients with chronic pain, psychological distress, and age-groups) are associated with differences in the benefits and adverse events of lumbar fusion surgery?

   ECRI Institute Evidence Assessment:

   - The evidence from one moderate-quality RCT (n = 294 patients) is insufficient to determine what patient characteristics are associated with differences in the benefits and adverse events of lumbar fusion surgery.

One RCT (Hagg et al. 2003) with 294 patients met the inclusion criteria for this question. This was another publication derived from the Swedish Lumbar Spine Study originally described in Fritzell et al. (2001). The efficacy and safety findings of Fritzell et al. were discussed under Key Questions 1 and 2. In their subsequent publication, Hagg et al. presented data concerning prognostic factors that were not included in Fritzell et al. Hagg et al. conducted a multivariate analysis to identify factors that predicted various outcomes of treatment in the surgical and nonsurgical (non-intensive physical therapy) patient groups. The main outcome measures in their analysis included change of disability (measured as ≥50% reduction of the ODI score), patient global assessment of treatment effect (improvement/no improvement), and work status at followup. Stepwise, forward multiple logistic regression analyses were performed within each treatment group, with the outcomes as dependent variables.

As shown in Table 25 (Appendix D), only one patient characteristic (neurotic personality) showed a statistically significant association with change in disability in the surgical group; patients with neurotic personalities were less likely to show improvement in the ODI score. No patient characteristic was significantly associated with improvement in ODI score in the nonsurgical group.

The study also identified patient characteristics significantly associated with the patient global assessment (improved or not improved). In the surgical group, neurotic personality was again associated with poor outcome (less likely to be improved), while disc height <50% was significantly associated with improvement. In the nonsurgical group, one patient characteristic (depressive symptoms) was significantly associated with poor outcome. No other factors were significantly associated with patient global assessment in either group.

Certain patient characteristics were significantly associated with work status at followup in both groups. Among surgical patients, older age and longer period of current sick leave were significant predictors of not working at followup. Among nonsurgical patients, only longer period of current sick leave was significantly associated with not working at followup. No positive predictors of working at followup were identified for either patient group.

The following variables did not show significant associations with any of the three outcomes at followup: pain (multiple measures), clinical findings (multiple measures), sociodemographics (disability pension, workers' compensation, unemployment, heavy job, comorbidity, smoking, prior surgery, gender, or marital status), other psychological measures (pain behavior, personality disorders), or radiographic indicators.

Although not specifically stated in the text of the study, it appears that patients who changed treatment groups after enrollment were not included in the analyses described above. The effect this might have on the observed associations is unknown.

Although multicenter, this was a single study of moderate quality; furthermore, none of the observed associations were large effects (see Appendix C, Strength of Evidence Algorithm, Decision Point 10 for definition of large effect). Therefore, the evidence is insufficient to allow a conclusion regarding patient characteristics associated with differences in the benefits and adverse events of lumbar fusion surgery.

Part II – Role of discography prior to lumbar fusion surgery

Overall, six studies were included to address the clinical questions in Part II.

Results and Conclusions – Part II

4. In patients being considered for lumbar fusion surgery, what is the reliability of discography?
   a. Test-retest reliability
   b. Inter-reader reliability

   ECRI Institute Evidence Assessment:

   - The evidence was insufficient to permit conclusions about the reliability of discography for patients with chronic low back pain and uncomplicated lumbar degenerative disc disease.

Two studies met the inclusion criteria for this Key Question[1].(36,37) Agorastides (2002)(36) reported data on both test-retest reliability and inter-rater reliability (133 discs in 72 patients), whereas Milette (1999)(37) only reported data on inter-rater reliability (132 discs in 45 patients).

Both studies investigated at least one specific type of reliability: whether a given discogram is judged to have the same morphology grade by the same reader at different times (i.e., test-retest) or by different readers (i.e., inter-rater). Notably, neither study performed two discography exams on the same disc to determine whether the results were consistent between discography injections. Also, neither study investigated the reliability of patients' reports of pain provocation or similarity to their typical pain. These types of reliability represent additional potential sources of variability in discography examinations that have not been assessed in patients with chronic low back pain and uncomplicated lumbar degenerative disc disease.

[1] After finding only two studies, we removed the date requirement (that studies must have been published in 1990 or later), but when we examined earlier studies, none of them met the other inclusion criteria.

We rated the quality of both studies as moderate (quality scores of 7.1 and 7.9). Both studies used consecutive enrollment, reported data on all or almost all enrolled patients, and the discograms were read without consultation of prior discograms or other clinical information about the patient. However, both were retrospective studies that did not report the funding source, and also the Agorastides study did not report whether patient inclusion/exclusion criteria were applied consistently to all patients.

For test-retest reliability, the Agorastides study observed good reliability (values for kappa ranging from 0.80 to 0.85 for the three raters),[2] but because it was a single moderate-quality study at a single center, we deemed this evidence limited quantity to permit conclusions. For inter-rater reliability, neither study observed large reliability (values for kappa ranging from 0.66 to 0.77), and neither study was multicenter. These factors, considered together with the moderate quality and limited quantity, mean that the evidence base was insufficient to permit conclusions.

5. In patients undergoing lumbar fusion surgery, do the results of pre-surgical discography predict the degree of pain reduction or improvement in functional status/quality of life after lumbar fusion surgery?

   ECRI Institute Evidence Assessment:

   - Because of low quality and heterogeneous results from three studies (n = 330 patients), the evidence was insufficient to permit conclusions about the use of discography to predict fusion outcomes in patients with chronic low back pain and uncomplicated lumbar degenerative disc disease.

This question involves a comparison in surgical outcomes between those who had a positive discography before surgery and those who had a negative discography before surgery. Three studies met the inclusion criteria.(39-41) Willems (2007)(39) included 82 patients, Gill (1992)(40) included 53 patients, and Colhoun (1988)(41) included 195 patients.

Importantly, the three studies each used a different definition of a "positive" discography test:

- Willems (2007)(39) categorized two groups of patients based on *typical pain provocation* in *adjacent-disc(s)*: 1) patients whose adjacent lumbar disc(s) provoked typical pain on discography (N = 22); and 2) patients whose adjacent lumbar disc(s) did not provoke typical pain (or no pain) on discography (N = 60).
- Gill (1992)(40) categorized three groups of patients based on the *morphology* of the *suspected disc*: 1) annular tear beyond the periphery (N = 20); 2) annular tear and contrast extension to the periphery, but not beyond (N = 19); and 3) small annular tear that did not extend to the periphery (N = 14).
- Colhoun (1988)(41) categorized four groups of patients based on both *typical pain provocation and morphology* of the *suspected disc:* 1) typical pain provocation and abnormal morphology (N = 137); no pain provocation and abnormal morphology (N = 25); 3) neither pain provocation nor abnormal morphology (N = 6); and 4) total disc

[2] Kappa measures chance-corrected agreement. 0 represents chance, and 1 represents perfect agreement. The standard interpretation of kappa values is that Below 0.0 is Poor agreement; 0.00-0.20 is Slight agreement; 0.21-0.40 is Fair agreement; 0.41-0.60 is Moderate agreement; 0.61-0.80 is Substantial agreement; 0.81-1.00 is Almost Perfect agreement.(38)

resorption of contrast material thus morphology not assessable and pain provocation not reported (N = 27).

Also, the three studies assessed different surgical outcomes:

- Willems (2007)(39) reported mean VAS pain scores at followup as well as the percentage of patients who experienced at least 30% pain relief (at a mean followup 6.7 years)
- Gill (1992)(40) reported a composite outcome involving the percentage of patients showing "improvement on functional testing and pain report", which was based on three items (Oswestry Pain Questionnaire, VAS, and pain drawing) (at a mean followup of 3 years)
- Colhoun (1988)(41) reported a composite outcome involving the percentage of patients who were considered a "success", which was defined as meeting all three conditions: 1) complete relief or significant subjective improvement in symptoms; 2) resumption of work and/or normal duties; 3) no intake of analgesics (at a mean followup 3.6 years).

Furthermore, the three studies reported qualitatively different results (the data appear in Table 36 of Appendix E):

- Willems (2007)(39) found evidence of no statistical difference in VAS pain scores at followup between the two groups, suggesting that discography results do not predict surgical outcomes.
- Gill (1992)(40) did not enroll enough patients to determine whether their data demonstrated a difference or no difference, leaving open the question of whether discography results predict surgical outcomes.
- Colhoun (1988)(41) found evidence of a difference in success rates, suggesting that discography results do predict surgical outcomes. Specifically, "success" was found to be more likely among patients with positive pain provocation and abnormal morphology (88%) than for other groups (52% to 85%).

We rated the quality of all three studies as low (with scores ranging from 4.1 to 4.3). All three were retrospective, non-randomized, unblinded studies. Only one of the three studies (Willems) reported baseline data to assess comparability of patient groups at baseline or attempted to enhance comparability using statistical methods.

Given the low quality, the different definitions of a positive discography, the different outcomes examined, and the qualitatively different results reported, we drew no conclusions about whether discography results predict surgical outcomes.

6. In patients being considered for lumbar fusion surgery, do patients who receive discography that influences the treatment choice have better treatment outcomes than patients who do not receive discography?

   ECRI Institute Evidence Assessment:

   - No evidence of acceptable quality was available to address this question; thus, the evidence was insufficient to permit conclusions about the influence of discography on fusion outcomes in patients with chronic low back pain and uncomplicated lumbar degenerative disc disease.

This question involves comparison of treatment outcomes between patients who had received discography before treatment and patients who had not received discography before treatment. Only one study met the inclusion criteria. Madan (2002)(42) retrospectively compared the surgical outcomes of two groups of patients at a single center: 32 patients who were seen between January 1998 and January 1999 and had a positive discography result; and 2) 41 patients who were seen prior to 1998 and had not received discography. All patients underwent the same surgical procedure (instrumented PLIF with posterolateral fusion).

Our quality assessment indicated that the study was very low quality (score 3.4), therefore we excluded the study from further consideration. The primary factors influencing this quality rating were a retrospective, non-concurrent, non-randomized, unblinded design in which the groups were not well-matched at baseline and authors had not attempted statistical methods that may have enhanced group comparability. Due to the lack of evidence of sufficient quality, we drew no conclusions about whether performing discography influences surgical outcomes.