Web Appendix

Data Extraction

Each retrieved citation was reviewed by two independently working reviewers (A.R. and J.D.). Most articles were excluded on the basis of information provided by the title or abstract. Citations that appeared to be appropriate or those that could not be excluded unequivocally from the title and abstract were identified, and the corresponding full text reports were reviewed by the two reviewers. Any disagreement between them was resolved by reviewer consensus. From the included articles, the following data were extracted: patient demographics, diagnoses, type of surgery, reoperation risks, perioperative outcomes, adverse events, and predictive factors for outcome following LLIF surgery.

Study Quality

Articles selected for inclusion were classified by level of evidence. The method used for assessing the quality of evidence of individual studies as well as the overall quality of the body of evidence incorporates aspects of the rating scheme developed by the Oxford Centre for Evidence-based Medicine[1] and used with modification by *The Journal of Bone and Joint Surgery American Volume* (*J Bone Joint Surg Am*),[2] precepts outlined by the Grades of Recommendation Assessment, Development and Evaluation (GRADE) Working Group[3] and recommendations made by the Agency for Healthcare Research and Quality (AHRQ).[4] Each individual study was rated by two different investigators against preset criteria that resulted in an evidence rating (Level of Evidence I, II, III, or IV). Disagreements were resolved through discussion.

Determination of Overall Strength of Evidence

After individual article evaluation, the overall body of evidence with respect to each outcome is determined based on precepts outlined by the Grades of Recommendation Assessment, Development and Evaluation (GRADE) Working Group[3] and recommendations made by the Agency for Healthcare Research and Quality (AHRQ).[4] Qualitative analysis is performed considering the following AHRQ required and additional domains.[5] ►**Table 6** below provides an outline of the method used to determine the final SoE.

Risk of bias is evaluated during the individual study evaluation described above. After individual article review, the literature evidence was rated as "HIGH" initially if the majority of the articles are levels I or II. It is rated as "LOW" if the majority were level III or lower. This is the "baseline" strength of evidence, online supplementary "4a: Critical Appraisal for Articles on Therapy." The consistency, directness, precision, and subgroup effects are considered for potential "downgrading" the strength of the body of evidence (one or two levels depending on the degree and number of domain violations).

Criteria Evaluated for "Downgrading"

- *Consistency* refers to the degree of similarity in the effect sizes of different studies within an evidence base. If effect sizes indicate the same direction of effect and if the range of effect sizes is narrow, an evidence base was judged to be consistent. If meta-analyses were conducted, we evaluated the consistency with an "eye ball test." This test consists of a visual appraisal of the forest plots by two independent reviewers. Single study evidence bases were judged "consistency unknown (single study)" and downgraded.
- *Directness is* concerned with whether the evidence being assessed reflected a single, direct link between the interventions of interest and the ultimate health outcome; that is, a determination of whether the most clinically relevant outcome was measured or if a surrogate outcome was assessed. Directness also applies to indirect comparisons of treatment when head-to-head comparisons of interest could not made within individual studies.
- *Precision* of evidence pertains to the degree of certainty surrounding an estimate of effect for a specific outcome. This is based on whether the estimate of effect reached statistical significance and/or the inspection of confidence intervals around effect estimates. When there are only two subgroups, the overlap of the confidence intervals of the summary estimates of the two groups is considered. No overlap of the confidence intervals indicates statistical significance, but the confidence intervals can overlap to a small degree and the difference still is statistically significant.
- *Subgroup effects.* For evaluating subgroup effects (i.e., heterogeneity of treatment effects), we downgrade if the authors do not state *a priori* their plan to perform subgroup analyses and if there was no test for interaction.

Criteria Used for "Upgrading"

- Finally, if the strength of evidence is less than "HIGH," we "upgrade" the evidence if there is a dose–response association or a strong magnitude of effect.

The following four possible levels and their definition are reported:

- **High**: High confidence that the evidence reflects the true effect. Further research is very unlikely to change our confidence in the estimate of effect.
- **Moderate**: Moderate confidence that the evidence reflects the true effect. Further research may change our confidence in the estimate of effect and may change the estimate.
- **Low**: Low confidence that the evidence reflects the true effect. Further research is likely to change the confidence in the estimate of effect and likely to change the estimate.
- **Insufficient**: Evidence either is unavailable or does not permit a conclusion.

Table 1 Summary of outcomes in studies comparing LLIF/XLIF/DLIF with PLIF/TLIF for lumbar degenerative disease

Investigator (y) Study design CoE	Condition	Intervention	Perioperative outcomes[a]	Reoperation	Adverse events
Deluzio (2010) Retrospective cohort (using historical cohort from same institution) CoE: III	Degenerative spine conditions (details NR)	• XLIF from L1–L2 to L4–5 and MIS TLIF/transsacral fusion at L5-S1 • n = 109	LOS • XLIF: 1.2 d	NR	Study reported that there was a 76% decrease in the rate of residual events from the PLIF group to the XLIF group (details NR)
		• Open PLIF (historical cohort) • n = 102	LOS • PLIF: 3.2 d	NR	
Rodgers and Gerber (2010) Retrospective cohort (using historical cohort from same institution) CoE: III	Degenerative spine conditions (stenosis, spondylolisthesis, scoliosis, postlaminectomy)	• XLIF • n = 40	LOS • 1.3 d (p < 0.0001 vs. PLIF) Blood loss • 1.4 g (p < 0.0001 vs PLIF)	Reoperation risk • 5.0% (2/40) (compression fracture or NR) (p = NS vs. PLIF)	Overall complications: 7.5% (3/40) (p < 0.0001 vs. PLIF) • Implant fracture, replaced during surgery (n = 1) • Compression fracture, 4 wk post-op (n = 1) • Atrial fibrillation (n = 1) • Deep wound infections (n = 0) Mortality • Overall: 2.5% (1/40) (p < 0.0018 vs. PLIF) • 3 mo: (n = 0) • 6 mo: myocardial infarction (n = 1)
		• Open PLIF (historical cohort) • n = 20	LOS • 5.3 d EBL • 2.7 g	Reoperation risk • 15.0% (3/20) (deep wound infection or compression fracture)	Overall complications: 60.0% (12/20) • Deep wound infections (n = 3) • Superficial infection (n = 1) • Wound dehiscence (n = 1) • Pneumonia (n = 3) • Compression fracture (n = 1) • Intestinal obstruction (n = 1) • Postoperative ileus (n = 2) Mortality • Overall: 30% (6/20) • 3 mo: postinfection (n = 2), postpneumonia (n = 1) • 6 mo: (n = 1) • 12 mo: (n = 1) • 18 mo: (n = 1)

(Continued)

Table 1 (*Continued*)

Investigator (y) Study design CoE	Condition	Intervention	Perioperative outcomes[a]	Reoperation	Adverse events
Knight (2009) Retrospective cohort (using historical cohort from senior author's practice) CoE: III	Degenerative spine conditions (details NR)	• XLIF or DLIF • n = 58	LOS • 5 d (1–12) (p = NS vs. PLIF) EBL (mean) • 136 mL (p = 0.0000 vs. PLIF)	Reoperation risk • 1.7% (1/58) (loss of fixation at L2–L3 for acute subsidence)	Overall complications (mild or major): 22.4% (13/58) Approach-related: 15.5% (9/58) • Injured ipsilateral L4 nerve root, residual motor effects at 1 year (n = 2) • Irritation of LFCN resulting in meralgia paresthetica, significant paresthesia at 1 year (n = 6) • Significant psoas muscle spasm (n = 1) Other • Myocardial infarction (n = 1) • Acute dementia (n = 1) • Urinary retention (n = 1) • Implant bone interface failure (n = 1) • Psoas muscle spasm on surgical side (n = 5) • Vascular: 0% • Retrograde ejaculation: 0% • Mortality: 0% (0/58)
		• Open PLIF (historical cohort) • n = 40	LOS • 5 d EBL (mean) • 489 mL	NR	Overall complications (including wound infection and dural tears, details NR): 22.5% (9/40) Mortality: 2.5% (1/40)

Abbreviations: CoE, class of evidence; DLIF, direct lateral interbody fusion; EBL, estimated blood loss; f/u, follow-up; LFCN, lateral femoral cutaneous nerve; LLIF, lumbar lateral interbody fusion; LOS, length of hospital stay; MIS, minimally invasive surgical techniques; NR, not reported; NS, not significant; PLIF, posterior lateral interbody fusion; TLIF, transforaminal lumbar interbody fusion; XLIF, extreme lateral interbody fusion.
[a]Blood loss measured by average preoperative to postoperative hemoglobin change (Rodgers and Gerber, 2010).

Table 2 Predictive factors affecting outcomes following LLIF/XLIF/DLIF for lumbar degenerative disease

Investigator (y) Study design CoE	Potential predictive factors and outcomes evaluated	Results
Kepler (2012) Retrospective cohort CoE: III	Predictive factors • Demographic factors: age, sex, BMI • Surgical factors: none • Other factors: preoperative sagittal alignment at instrumented levels (degrees) Outcomes • Postoperative lumbar lordosis	Significant results (did not control for confounders) • Preoperative alignment was correlated significantly with postoperative lordosis ($p = 0.34$, $p = 0.003$) and correlated inversely with increase in lordosis ($p = -0.67$, $p < 0.001$). Other results • The following factors were not significantly correlated with change in segmental or global lordosis: age ($p = 0.09$, $p = 0.23$; $p = 0.12$, $p = 0.27$, respectively), sex ($p = 0.8$, $p = 0.9$, respectively; p NR), and BMI ($p = -0.1$, $p = 0.24$; $p = 0.13$, $p = 0.25$, respectively).
Isaacs (2010) Prospective cohort CoE: II	Predictive factors • Demographic factors: age, sex, BMI, comorbidities, severity of deformity • Surgical factors: inclusion of specific levels, number of levels treated, additional posterior decompression, type of fixation • Other factors: none Outcomes • Perioperative complications	Significant results (controlled for confounders) • Total number of levels operated per patient is the strongest independent predictor of complications ($p = 0.0004$): there is $\sim 59\%$ increase in the complication risk for each additional level treated (OR, 1.59; $p = 0.0105$). Other results • No other demographic or surgical factors were significant predictors when evaluated with the number of operative levels.
Rodgers and Cox (2010) Retrospective cohort CoE: III	Predictive factors • Demographic factors: BMI, age, sex, height and weight, smoking, comorbidities (including diabetes mellitus, coronary artery disease, chronic obstructive pulmonary disease, chronic steroid use) • Surgical factors: number of levels treated • Other factors: preoperative diagnosis Outcomes • Early complications (within first 3 months), including wound, nerve, cardiac, renal, GI, respiratory, vertebral body-related, and hardware-related	Significant results (controlled for confounders) • Preoperative diagnosis was the only variable found to significantly affect whether or not complications occurred ($p = 0.0075$). Higher complication risk reported in patients with diagnosis of DDD or recurrent disc herniation. Other results • No other demographic or surgical factors were significant predictors of the occurrence of a complication.

Abbreviations: BMI, body mass index; CoE, class of evidence; DDD, degenerative disc disease; DLIF, direct lateral interbody fusion; GI, gastrointestinal; LLIF, lumbar lateral interbody fusion; NR, not reported; OR, odds ratio; XLIF, extreme lateral interbody fusion.

Table 3a Summary of inclusion and exclusion criteria for studies comparing LLIF/XLIF/DLIF with PLIF or TLIF

	Inclusion	Exclusion
Patient	Adult patients (18 y or older) with • Lumbar (L1–L5) degenerative disc disease, with or without canal stenosis and with or without degenerative spondylolisthesis, or • Lumbar degenerative scoliosis	• Tumor/neoplasms • Trauma • Thoracic disc disease • Infection • Fractures
Intervention	• Lumbar lateral interbody fusion (LLIF) (also known as extreme lateral interbody fusion (XLIF), direct lateral interbody fusion (DLIF), and lateral transpsoas interbody fusion (LTIF)) • As stand-alone procedure or augmented with posterior pedicle screw fixations	
Comparison	• PLIF or TLIF as stand-alone procedures or augmented with posterior pedicle screw fixations (either open or percutaneous)	• ALIF
Outcome	• Differences in tissues dissection and trauma • Length of hospital stay • Fusion rates • Postoperative pain • Blood loss • Revision surgery • Complications including neural injuries, hip flexion weakness, hardware-related (including malposition, subsidence), infection, mortality, subsidence, adjacent level disease	
Study design	• Comparative studies • At least 10 patients per treatment group	• Case reports • Case series • Comparative study with less than 10 patients per treatment group • Animal, biomechanical, or cadaveric studies
Publication	• Peer-reviewed studies written in English	• Abstracts, editorials, letters • Duplicate publications of the same study, which do not report on different outcomes • Single reports from multicenter trials • White papers • Meeting abstracts, presentations or proceedings • Narrative reviews • Articles identified as preliminary reports when results are published in later versions

Table 3b Summary of inclusion and exclusion criteria for studies investigating factors affecting patient outcome after LLIF/XLIF/DLIF surgery

	Inclusion	Exclusion
Patient	Adult patients (18 years or older) with one of the following conditions who have received lumbar lateral interbody fusion (LLIF) (also known as extreme lateral interbody fusion (XLIF), direct lateral interbody fusion (DLIF), and lateral transpsoas interbody fusion (LTIF)), as a stand-alone procedure or augmented with posterior pedicle screw fixation: • Lumbar (L1-L5) degenerative disc disease, with or without canal stenosis and with or without degenerative spondylolisthesis, or • Lumbar degenerative scoliosis	• Tumor/neoplasms • Trauma • Thoracic disc disease • Infection • Fractures
Predictive factors	• Degree of scoliosis • Degree of spondylolisthesis • Degree and type of canal stenosis • Demographic factors including sex, age, obesity	
Outcome	• Postoperative pain • Neurological improvement • Sagittal and coronal balance • Pelvic parameters	
Study design	• Studies assessing factors affecting patient outcome following LLIF	• Studies with less than 20 patients • Case reports • Animal, biomechanical, or cadaveric studies

Table 4a Class of evidence summary table for included studies comparing LLIF/XLIF/DLIF with PLIF or TLIF

Methodological principle	Deluzio (2010)	Rodgers and Gerber (2010)	Knight (2009)
Study design			
Randomized controlled trial			
Prospective cohort study			
Retrospective cohort study	√	√	√
Case-series			
Random sequence generation[a]			
Statement of concealed allocation[a]			
Intention to treat[a]			
Independent or blind assessment[b]			
Co-interventions applied equally			
Complete follow-up of ≥ 80%			
Adequate sample size		√	√
Controlling for possible confounding[c]			
Evidence level	III	III	III

Note: Blank cells indicate that the criterion was either not met or that it could not be determined
[a]Applies only to randomized controlled trials only.
[b]Applies to assessment of the primary study outcome(s).
[c]Groups must be comparable on baseline characteristics or evidence of control for confounding presented.

Table 4b Definition of class of evidence for articles on therapy

		Studies of therapy	
Class	Risk of bias	Study design	Criteria
I	**Low risk** Study adheres to commonly held tenets of high quality design, execution, and avoidance of bias	Good quality RCT	• Random sequence generation • Allocation concealment • Intent-to-treat analysis • Blind or independent assessment for important outcomes • Cointerventions applied equally • F/U rate of 80%+ • Adequate sample size
II	**Moderately low risk** Study has potential for some bias; study does not meet all criteria for class I, but deficiencies not likely to invalidate results or introduce significant bias	Moderate or poor quality RCT	• Violation of one of the criteria for good quality RCT
		Good quality cohort	• Blind or independent assessment in a prospective study, or use of reliable data[a] in a retrospective study • Co-interventions applied equally • F/U rate of 80%+ • Adequate sample size • Controlling for possible confounding[b]
III	**Moderately high risk** Study has significant flaws in design and/or execution that increase potential for bias that may invalidate study results	Moderate or poor quality cohort	• Violation of any of the criteria for good quality cohort
		Case-control	• Any case-control design
IV	**High risk** Study has significant potential for bias; lack of comparison group precludes direct assessment of important outcomes	Case series	• Any case series design

Abbreviations: F/U, follow-up; RCT, randomized controlled trial.
[a]Outcome assessment is independent of healthcare personnel judgment. Reliable data are data such as mortality or reoperation.
[b]Authors must provide a description of robust baseline characteristics, and control for those that are unequally distributed between treatment groups.

Table 4c Class of evidence summary table for included studies investigating factors affecting patient outcome after LLIF/XLIF/DLIF surgery

Methodological principle	Kepler (2012)	Isaacs (2010)	Rodgers and Cox (2010)
Study design			
Randomized controlled trial			
Prospective cohort study		√	
Retrospective cohort study	√		√
Case control			
Case-series			
Patients at similar point in the course of their disease or treatment	√	√	√
Patients followed long enough for outcomes to occur	√	√	√
Complete follow-up of $>$ 80%			
Controlling for extraneous prognostic factors[a]		√	√
Evidence level	III	II	III

[a]Authors must consider other factors that might influence patient outcomes.

Table 4d Criteria for class of evidence (CoE) for prognostic studies

		Studies	
Class	Risk of bias	Study design	Criteria
I	**Low risk** Study adheres to commonly held tenets of high quality design, execution, and avoidance of bias	Good quality cohort	• Prospective design • Patients at similar point in the course of their disease or treatment • F/U rate of $\geq$ 80% • Patients followed long enough for outcomes to occur • Accounting for other prognostic factors
II	**Moderately low risk** Study has potential for some bias; does not meet all criteria for class I but deficiencies not likely to invalidate results or introduce significant bias	Moderate quality cohort	• Prospective design, with violation of one of the other criteria for good quality cohort study • Retrospective design, meeting all the rest of the criteria in class I
III	**Moderately high risk** Study has flaws in design and/or execution that increase potential for bias that may invalidate study results	Poor quality cohort Good quality case-control or cross-sectional study	• Prospective design with violation of 2 or more criteria for good quality cohort, or • Retrospective design with violation of 1 or more criteria for good quality cohort • A good case-control study • A good cross-sectional study
IV	**High risk** Study has significant potential for bias; does not include design features geared toward minimizing bias and/or does not have a comparison group	Poor quality case-control or cross-sectional Case series	• Other than a good case-control study • Other than a good cross-sectional study • Any case series design

Table 5 Evidence summary

Baseline quality: High = majority of articles level I/II; low = majority of articles level III/IV *Upgrade:* Large magnitude of effect (1 or 2 classes); dose response gradient (1 class) *Downgrade:* Inconsistency of results (1 or 2 classes); indirectness of evidence (1 or 2 classes); imprecision of effect estimates (1 or 2 classes)					
	Strength of evidence	**Conclusions/ comments**	**Baseline**	**Upgrade (classes)**	**Downgrade (classes)**
In adult patients, what is the comparative effectiveness of LLIF/XLIF/DLIF surgery compared with PLIF or TLIF surgery?					
LLIF/XLIF/DLIF versus PLIF/TLIF	Insufficient	None of the studies reported the comparative effectiveness of radiographic or patient-reported outcomes.	Insufficient		
In adult patients, what is the comparative safety of LLIF/XLIF/DLIF surgery compared with PLIF or TLIF surgery?					
LLIF/XLIF/DLIF versus PLIF/TLIF	Low	Overall, the evidence on the comparative safety of LLIF compared with PLIF is low. The LLIF treatment group had less estimated blood loss and a lower mortality risk than the PLIF treatment group. However, results for other outcomes were inconsistent. Two studies reported a shorter length of hospital stay for the LLIF group, yet one study reported the same length of hospital stay for both treatment groups. One study reported a significantly lower complication risk for the LLIF group, but another study reported approximately the same risk for both treatment groups. And only one study reported the reoperation risk for both treatment groups.	Low	No	No
In adult patients, are there any factors affecting patient outcome after LLIF/XLIF/DLIF surgery?					
LLIF/XLIF/DLIF	Insufficient	Overall, the evidence that factors predict patient outcome after LLIF surgery is insufficient. The 3 studies examined predictive factors for different outcomes. Two studies performed a multivariate analysis to control for confounders: 1 study found that number of levels treated was a significant predictor of perioperative complications and 1 study found that preoperative diagnosis was a significant predictor of early complications. The third study found that preoperative sagittal alignment was a significant predictor of postoperative lumbar lordosis, but did not control for confounders in the analysis. All 3 studies found that age, sex, and BMI were not predictors of outcome after LLIF.	Low	No	Yes (consistency unknown: one study for each type of outcome analyzed)

Abbreviations: DLIF, direct lateral interbody fusion; LLIF, lumbar lateral interbody fusion; PLIF, posterior lateral interbody fusion; TLIF, transforaminal lumbar interbody fusion; XLIF, extreme lateral interbody fusion.

Note: All AHRQ "required" and "additional" domains[a] are assessed. Only those that influence the baseline grade are listed in table.

Baseline strength: Risk of bias (including control of confounding) is accounted for in the individual article evaluations. High = majority of articles level I/II; low = majority of articles level III/IV.

Downgrade: Inconsistency[b] of results (1 or 2); indirectness of evidence (1 or 2); imprecision of effect estimates (1 or 2); sub-group analyses not stated a priori and no test for interaction (2).

Upgrade: Large magnitude of effect (1 or 2); Dose response gradient (1).

[a]Required domains: risk of bias, consistency, directness, precision. Plausible confounding that would decrease observed effect is accounted for in our baseline risk of bias assessment through individual article evaluation. Additional domains: dose–response, strength of association, publication bias.

[b]Single study = "consistency unknown."

Table 6 Methodology outline for determining overall strength of evidence (SoE)

All AHRQ "required" and "additional" domains[a] are assessed. Only those that influence the baseline grade are listed in table
Baseline strength: Risk of bias (including control of confounding) is accounted for in the individual article evaluations
High = majority of articles level I/II; low = majority of articles level III/IV
Downgrade: Inconsistency[b] of results (1 or 2); Indirectness of evidence (1 or 2); Imprecision of effect estimates (1 or 2); Subgroup analyses not stated *a priori* and no test for interaction (2)
Upgrade: Large magnitude of effect (1 or 2); Dose response gradient (1)

Outcome	Strength of evidence	Conclusions and comments	Baseline	Downgrade	Upgrade
Outcome	High	Summary of findings	High Level I/II studies	No consistent, direct, and precise estimates	No
Outcome	Moderate	Summary of findings	Low Level III studies	No consistent, direct, and precise estimates	Yes Large effect
Outcome	Low	Summary of findings	High Level I/II studies	Yes (2) Inconsistent Indirect	No

[a]Required domains: risk of bias, consistency, directness, precision. Plausible confounding that would decrease observed effect is accounted for in our baseline risk of bias assessment through individual article evaluation. Additional domains: dose–response, strength of association, publication bias.
[b]Single study = "consistency unknown."

Table 7 List of articles excluded at full-text review

Article	Reason for exclusion
Studies comparing LLIF/XLIF/DLIF with PLIF or TLIF	
Tormenti MJ, Maserati MB, Bonfield CM, Okonkwo DO, Kanter AS. Complications and radiographic correction in adult scoliosis following combined transpsoas extreme lateral interbody fusion and posterior pedicle screw instrumentation. Neurosurgical focus 2010;28(3):E7	Number of patients < 10/treatment group
Studies investigating factors affecting patient outcome after LLIF surgery	
Johnson RD, Valore A, Villaminar A, Comisso M, Balsano M. Pelvic parameters of sagittal balance in extreme lateral interbody fusion for degenerative lumbar disc disease. Journal of clinical neuroscience: official journal of the Neurosurgical Society of Australasia 2013;20(4):576–581	No analysis of predictive factors
Smith JS, Shaffrey CI, Glassman SD, et al. Risk-benefit assessment of surgery for adult scoliosis: an analysis based on patient age. Spine 2011;36(10):817–824	XLIF surgery not included in review

References

1 OCEBM Levels of Evidence Working Group. The Oxford 2011 Levels of Evidence.   Available at: http://www.cebm.net/index.aspx?o=5653. Accessed December 11, 2012

2 Wright JG, Swiontkowski MF, Heckman JD. Introducing levels of evidence to the journal. J Bone Joint Surg Am 2003;85-A(1): 1–3

3 Atkins D, Best D, Briss PA, et al; GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ 2004; 328(7454):1490

4 West S, King V, Carey TS, et al. Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center, Contract No. 290–97–0011). Rockville, MD: Agency for Healthcare Research and Quality; 2002

5 Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions—agency for healthcare research and quality and the effective health-care program. J Clin Epidemiol 2010;63(5):513–523