

Three-Month Leg Pain and Failure to Achieve the 1-Year ODI MCID After Unilateral Biportal Endoscopic Lumbar Interbody Fusion: A Landmark Comparison of Four Early Pain Measures

Xiaolin Zhang

The First Affiliated Hospital of Shandong Second Medical University, Weifang People's Hospital, Weifang, Shandong, China.

Hangkai Cui

The First Affiliated Hospital of Shandong Second Medical University, Weifang People's Hospital, Weifang, Shandong, China.

Zhichang Wang

The First Affiliated Hospital of Shandong Second Medical University, Weifang People's Hospital, Weifang, Shandong, China.

Haoyu Jiao

The First Affiliated Hospital of Shandong Second Medical University, Weifang People's Hospital, Weifang, Shandong, China.

Feng Li

fenglispine@163.com

The First Affiliated Hospital of Shandong Second Medical University, Weifang People's Hospital, Weifang, Shandong, China.

Xiaopeng Li

The First Affiliated Hospital of Shandong Second Medical University, Weifang People's Hospital, Weifang, Shandong, China.

Research Article

Keywords: unilateral biportal endoscopic lumbar interbody fusion, Oswestry Disability Index, postoperative leg pain, landmark analysis, restricted cubic spline, internal validation

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10678480/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background. Early postoperative pain may help identify patients whose functional recovery after unilateral biportal endoscopic lumbar interbody fusion (ULIF) is falling behind, but the most informative pain measure is uncertain. We compared four early pain measures and then examined the association, candidate threshold, and incremental information of the best-performing measure for failure to achieve the 1-year Oswestry Disability Index (ODI) minimum clinically important difference (MCID).

Methods. We retrospectively screened 282 surgery-system records of possible ULIF procedures performed from January 2023 through December 2024. After procedure confirmation and hierarchical eligibility review, 234 patients were clinically eligible; 189 had complete required pain and ODI data and formed the 3-month landmark cohort. Back- and leg-pain numerical rating scale (NRS) scores at 2 weeks and 3 months were compared using receiver operating characteristic analysis. The measure with the highest apparent area under the curve (AUC) was then examined with Firth logistic regression and restricted cubic splines. Fixed post-selection models underwent 1,000-sample conditional bootstrap validation, while a separate 10,000-sample bootstrap repeated marker and threshold selection.

Results. Of the 189 patients, 77 (40.7%) did not achieve the ODI MCID. Three-month leg pain had the highest apparent AUC (0.891; 95% CI, 0.850–0.933) and was selected in 9,999 of 10,000 marker-selection bootstraps. The probability of MCID non-achievement rose with higher leg-pain scores and followed an apparently nonlinear pattern. The selection-aware corrected AUC for the continuous score was 0.891, and the corrected balanced accuracy of the NRS ≥ 3 rule was 0.791. In the fixed post-selection models, concurrent 3-month ODI had a corrected AUC of 0.973. Adding the leg-pain spline increased corrected AUC by 0.0125 but moved the calibration slope farther from 1.

Conclusions. Three-month leg pain separated ODI-MCID achievers from non-achievers better than the other three early pain measures, although it added little beyond concurrent 3-month ODI. An NRS score of 3 or higher may be useful as a signal for closer reassessment, but not as a stand-alone prediction or treatment rule. External transportability remains unknown.

BACKGROUND

Degenerative lumbar disease is a major cause of disability, and lumbar fusion is increasingly performed as populations age and minimally invasive techniques become more widely available(1–3). Unilateral biportal endoscopic lumbar interbody fusion (ULIF) combines endoscopic decompression with interbody reconstruction and has shown favorable short-term clinical and radiographic results in selected patients(4–8). Even after technically successful surgery, however, the pace of recovery varies considerably between patients.

This variation is not fully captured by radiographs. Patient-reported outcomes therefore remain central to postoperative follow-up. The Oswestry Disability Index (ODI), first described by Fairbank et al. and later standardized for clinical and research use, is widely used to quantify disability related to low back disorders (9, 10). Our institution uses a validated simplified Chinese version(11). Interpreting ODI change through the minimum clinically important difference (MCID) can help distinguish improvement that is likely to be meaningful to patients(12, 13). Previous studies have examined factors associated with failure to achieve the ODI MCID after minimally invasive lumbar surgery(14), but a simple marker available during routine early follow-up would be clinically useful.

DiSilvestro et al. reported that pain measured 6–12 weeks after minimally invasive transforaminal lumbar interbody fusion was associated with 1-year ODI recovery(15). This finding raised a practical question for ULIF: does back pain or leg pain, and at which early follow-up visit, provide the clearest signal of later functional recovery? Because the procedure and assessment schedule differ from those in the earlier study, direct external validation of its threshold was not appropriate.

Recovery may also be shaped by age, comorbidity, functional dependence, and baseline disability (16–24). We therefore retained age, sex, BMI, and baseline ODI as routine clinical adjustment variables. The database also contained the four comorbidity domains used in the conventional mFI-5 together with a preoperative Barthel Index. Because the original ACS-NSQIP functional-status item was unavailable, we used these variables to construct a study-specific five-domain burden rather than labeling it a validated mFI-5.

The analysis was designed in two stages. First, we compared back- and leg-pain NRS scores at 2 weeks and 3 months. We then carried forward the measure with the highest discrimination to examine its continuous association with 1-year ODI-MCID non-achievement, explore a candidate integer threshold, and test whether it added information beyond a concurrent 3-month ODI landmark measure.

METHODS

Study Design and Cohort Construction

This single-center retrospective cohort used a 3-month landmark design and was reported according to STROBE(25). A surgery-system search identified 282 records as possible ULIF procedures performed between January 1, 2023, and December 31, 2024. Clinical data were reviewed through February 24, 2026. Eight records were not ULIF procedures, leaving 274 confirmed ULIF procedures for eligibility review.

Clinical exclusions were applied hierarchically so that each patient was counted once: non-degenerative indication (n=7), previous lumbar surgery (n=12), and multilevel surgery (n=21). This left 234 clinically eligible patients. Follow-up exclusions were then assigned according to the earliest required assessment that was missing: 2-week NRS (n=14), 3-month NRS and/or ODI (n=15), and 1-year NRS and/or ODI (n=16). The final complete-case cohort comprised 189 patients.

Postoperative assessments were assigned to prespecified windows: days 7–28 for the 2-week visit, days 60–120 for the 3-month landmark, and days 330–420 for the 1-year outcome. If more than one assessment fell within a window, the record closest to postoperative day 14, 90, or 365 was used; when two records were equally close, the earlier record was selected. Visit assignment was based on source-database dates. Exact postoperative days were not retained in the locked analytic dataset and therefore could not be independently rechecked from that file.

Surgical Procedure and Perioperative Variables

ULIF was performed under general anesthesia using separate endoscopic viewing and working portals through small paramedian incisions. Endoscopic decompression and discectomy were followed by endplate preparation, interbody cage insertion, and pedicle-screw fixation. The analytic dataset included age, sex, height, weight, body mass index (BMI), index level, American Society of Anesthesiologists class, operative time, estimated blood loss, and hospital length of stay. BMI was recalculated as weight in kilograms divided by height in meters squared and categorized as <25.0, 25.0–29.9, or ≥30.0 kg/m².

Patient-Reported Outcomes and Primary Endpoint

Back and leg pain were recorded separately on an 11-point NRS from 0 (no pain) to 10 (worst possible pain). All 189 patients had scores at baseline, 2 weeks, 3 months, and 1 year; no NRS values were imputed. Disability was assessed using the validated Simplified Chinese version of the Oswestry Disability Index (ODI), which has been previously translated and validated in Chinese-speaking populations(11). The ODI was originally developed by Fairbank and colleagues and has subsequently been widely used for assessing disability related to lumbar spinal disorders(9, 10). ODI totals were stored on a 0–100 scale, with higher values indicating greater disability. Item-level responses were not available in the locked dataset, so handling of unanswered sections and denominator adjustment could not be independently checked. ODI improvement was calculated as preoperative ODI minus 1-year ODI. Failure to achieve the MCID was defined as improvement <12.8 points(13). Because this threshold came from the literature rather than local validation, sensitivity analyses used thresholds of 10 and 15 points and also examined 1-year ODI continuously.

Failure to achieve the literature-derived ODI improvement threshold was the event in all binary models. Analyses were restricted to patients with pain and ODI data at the 3-month landmark and at 1 year. The results therefore apply to patients who remained under observation at 3 months and are not preoperative risk estimates.

Five-Domain Comorbidity and Functional-Dependence Burden

A five-domain deficit burden was constructed from diabetes mellitus, chronic obstructive pulmonary disease, congestive heart failure, treated hypertension, and preoperative functional dependence. These domains correspond to the conventional mFI-5(20, 24), but we did not treat the resulting measure as a validated mFI-5 because the original ACS-NSQIP functional-status category was unavailable. Functional status was assessed with the preoperative Barthel Index (26); a score <60 contributed one point as a study-defined indicator of substantial dependence in basic activities of daily living, and a score ≥60 contributed zero points. The other four domains each contributed one point, yielding a total of 0–5. Higher burden was defined as ≥2 deficits, consistent with thresholds commonly used in lumbar-fusion studies of the mFI-5(24).

Statistical Analysis

Continuous variables are summarized as mean ± standard deviation or median [interquartile range], and categorical variables as number (percentage). For Table 1, absolute standardized differences for continuous variables were calculated from group means using the square root of the average group variances, even when medians and quartiles were displayed. Binary differences used level-specific proportions and pooled Bernoulli variance. BMI and index-level categories were assessed separately. Table 2 is descriptive because the groups were defined by 1-year ODI change. Apparent AUCs for the four pain measures were compared with paired DeLong tests, with Holm adjustment across all six pairwise comparisons(27) (Supplementary Table S4).

Receiver operating characteristic curves were generated for back and leg pain at 2 weeks and 3 months. The measure with the highest AUC was carried forward. Integer cutoffs within that measure were assessed with the Youden index, with ties resolved in favor of the lower cutoff (28). Because both the marker and threshold were selected from the study data, these analyses were considered exploratory.

The best-performing measure, 3-month leg-pain NRS, was modeled continuously using Firth bias-reduced logistic regression(29) and a four-knot restricted cubic spline (30). Knots at 0, 2, 3, and 6 were chosen after inspection of the discrete score distribution and then fixed. A sensitivity model used the empirical 10th, 50th, and 90th percentiles (1, 2, and 5). Models adjusted for age, sex, BMI, higher five-domain deficit burden, and baseline ODI; these variables were selected on clinical grounds and from prior literature rather than univariable P values(15, 18-24). Ordinary maximum-likelihood likelihood-ratio tests were used only to describe functional form. Effect estimates and conditional internal validation came from the Firth models. Marginal probabilities were standardized over the observed covariate distribution. Delta-method confidence intervals used the Firth covariance matrix, were constrained to 0–1, and did not include uncertainty from marker or spline selection.

We compared six fixed 3-month landmark models: the clinical benchmark (age, sex, BMI, higher five-domain deficit burden, and baseline ODI); the benchmark plus the leg-pain spline; the benchmark plus NRS ≥3; the benchmark plus 3-month ODI; the benchmark plus 3-month ODI and NRS ≥3; and the benchmark plus 3-month ODI and the leg-pain spline. Each model was refitted in 1,000 patient-level bootstrap samples using Firth estimation to estimate model-fitting optimism(31). The selected marker, knot locations, and NRS ≥3 threshold remained fixed. These estimates therefore represent conditional internal validation of the fixed post-selection models rather than validation of the complete data-driven analysis pipeline. Calibration intercept and slope were estimated by Firth logistic recalibration of model-predicted log odds. Changes in AUC, Brier score, and calibration slope after adding leg pain to the 3-month ODI model were calculated within the same resamples. All fits converged.

A separate 10,000-sample bootstrap repeated the unadjusted marker-and-threshold selection procedure. In each sample, all four AUCs were recalculated, the best-performing marker was selected, and the integer Youden cutoff was re-estimated. Marker-level AUC ties followed the fixed order used in the analysis code: 2-week leg pain, 2-week back pain, 3-month leg pain, and 3-month back pain. Ties between Youden-optimal integer cutoffs favored the lower cutoff. The continuous marker and thresholded rule were evaluated separately. Corrected values were obtained by subtracting the mean difference between bootstrap-sample performance and performance when the bootstrap-selected choice was applied to the original sample. This procedure repeated marker and threshold selection only; spline knots and multivariable models were not reselected. Fixed random seeds were used for both bootstrap procedures to ensure reproducibility.

Sensitivity analyses used ODI-MCID thresholds of 10, 12.8, and 15 points, a baseline-adjusted continuous 1-year ODI outcome, and a three-knot spline. Concurrent 3-month ODI was treated as a same-domain landmark benchmark rather than as a biologically

independent predictor.

Potential effect modification was examined in a separate ordinary logistic model containing higher five-domain deficit burden, linear 3-month leg pain, and their interaction. This exploratory model was not adjusted for the full benchmark set. We did not calculate subgroup AUCs or perform decision-curve analysis.

No formal sample-size calculation was performed because the study used all clinically eligible patients with complete required follow-up measurements. The same 189 complete cases were used for the main analyses, without imputation. Model complexity was limited because 77 patients experienced the primary event.

RESULTS

Cohort Selection and Characteristics

The surgery-system search returned 282 records of possible ULIF procedures. Eight were not ULIF procedures, leaving 274 confirmed procedures. Clinical review then excluded 7 patients with a non-degenerative indication, 12 with previous lumbar surgery, and 21 who underwent multilevel surgery. Among the remaining 234 clinically eligible patients, 14 lacked the required 2-week NRS assessment, 15 lacked the 3-month NRS and/or ODI assessment, and 16 lacked the 1-year NRS and/or ODI assessment. The final complete-case cohort included 189 patients: 77 did not achieve the ODI MCID and 112 did (Fig. 1).

Figure 1.

The median age was 65.0 years [IQR, 58.0–69.0], 108 patients (57.1%) were women, and median BMI was 26.0 kg/m² [IQR, 23.8–28.0]. Thirty patients (15.9%) met the study-defined functional-dependence criterion, and 39 (20.6%) had at least two deficits (Table 1).

Patient-Reported Outcome Trajectories

Baseline ODI was nearly identical in non-achievers and achievers (51.8 ± 5.9 vs 52.5 ± 7.4). The groups separated during follow-up: ODI at 2 weeks was 46.6 ± 6.5 versus 38.5 ± 10.7 , at 3 months 48.3 ± 6.3 versus 33.0 ± 10.3 , and at 1 year 47.9 ± 6.7 versus 27.5 ± 8.1 . Mean 1-year improvement was 3.9 ± 3.3 in non-achievers and 25.1 ± 8.7 in achievers (Table 2 and Supplementary Figure S1). These trajectories are descriptive because group membership was defined by ODI improvement.

At 3 months, mean leg-pain NRS was 3.9 ± 1.7 in non-achievers and 1.5 ± 1.1 in achievers. Mean back-pain NRS was 4.5 ± 1.9 and 2.3 ± 1.6 , respectively (Fig. 2).

Comparison and Selection of Early Pain Measures

Three-month leg pain had the highest discrimination of the four postoperative pain measures, with an apparent AUC of 0.891 (95% CI, 0.850–0.933). This exceeded the 3-month back-pain AUC by 0.076 (Holm-adjusted $P = 0.002$), with larger differences from both 2-week measures (Holm-adjusted $P < 0.001$) (Table 3A and Fig. 3). Three-month leg pain was selected in 9,999 of 10,000 bootstrap samples.

Candidate Threshold and Association Shape

$\text{NRS} \geq 3$ was the most frequently selected integer cutoff. In the original cohort, sensitivity was 0.753, specificity 0.848, and balanced accuracy 0.801. After correction for marker and threshold selection, these values were 0.742, 0.840, and 0.791; corrected Youden J was 0.582 (Table 3B). Conditional on selecting 3-month leg pain, the cutoff was ≥ 2 in 20.36% of bootstrap samples, ≥ 3 in 79.55%, and ≥ 4 in 0.09% (Supplementary Table S2). The corrected AUC of the continuous score was 0.891.

For the adjusted model containing the fixed $\text{NRS} \geq 3$ indicator, conditional internal validation gave an optimism-corrected AUC of 0.829, a Brier score of 0.155, and a calibration slope of 0.937 (Table 4A). Because the threshold was held fixed during this bootstrap, these estimates do not include all threshold-selection uncertainty. The adjusted-model AUC (0.829), continuous-score

AUC (0.891), and binary-rule balanced accuracy (0.791) describe different analyses and should not be compared as if they were the same performance measure.

The probability of ODI-MCID non-achievement rose with increasing 3-month leg-pain NRS, and the spline suggested that the association was nonlinear. In a descriptive likelihood-ratio comparison, the four-knot spline differed from a linear term ($\chi^2=11.49$, 2 df; $P = 0.003$). Standardized event probabilities were close to zero at NRS 0 and were 41.7%, 60.7%, 74.3%, and 88.3% at scores of 2, 3, 4, and 5, respectively (Fig. 4 and Supplementary Table S3). All 27 patients with $NRS \geq 5$ were non-achievers, making estimates at the upper end of the scale imprecise. Adding the leg-pain spline to the clinical benchmark increased the conditionally corrected AUC from 0.556 to 0.910 (Table 4A).

Figure 4.

The interaction between 3-month leg pain and higher five-domain deficit burden was not clearly supported (OR, 0.87; 95% CI, 0.36–2.08; $P = 0.753$), although the estimate was imprecise (Supplementary Table S1).

Incremental Information and Sensitivity Analyses

Concurrent 3-month ODI was the strongest landmark measure in the fixed post-selection models, with a conditionally corrected AUC of 0.973, Brier score of 0.055, and calibration slope of 0.835. Adding $NRS \geq 3$ changed AUC by 0.0022 and Brier score by -0.0087, while calibration slope changed by -0.032. Adding the leg-pain spline changed AUC by 0.0125 and Brier score by -0.0135, with a -0.096 change in calibration slope. The added discrimination from leg pain was therefore small, and calibration did not improve (Table 4B).

Results were similar in the main sensitivity analyses. The four-knot model had a conditionally corrected AUC of 0.910, compared with 0.907 for the three-knot spline. Using ODI-MCID thresholds of 10, 12.8, and 15 points gave apparent AUCs of 0.900, 0.928, and 0.924, respectively; each model differed from the benchmark in descriptive likelihood-ratio testing ($P < 0.001$) (Supplementary Table S1). These estimates remain conditional on the fixed post-selection modeling choices and do not establish performance in an external cohort.

When 1-year ODI was analyzed continuously, the benchmark model explained 13.7% of the variance and the model including the leg-pain spline explained 37.1% (partial $F = 22.40$, 3 and 180 df; $P < 0.001$). This supported an association that was not dependent on dichotomizing the ODI outcome at the MCID threshold.

Thirty patients met the Barthel Index < 60 definition of functional dependence. Five-domain deficit scores were 0 in 82 patients, 1 in 68, 2 in 27, 3 in 11, and 4 in 1; 39 patients had at least two deficits.

DISCUSSION

Principal Findings

Of the four early pain measures, 3-month leg pain separated ODI-MCID achievers from non-achievers most clearly. The association also appeared nonlinear, with risk rising more sharply across the middle of the NRS range. $NRS \geq 3$ emerged most often as the candidate cutoff, but reducing the continuous score to a binary rule lost information. Importantly, concurrent 3-month ODI remained much more informative than leg pain, and adding leg pain produced only a modest gain in discrimination.

Comparison With Previous Research

Our findings point in the same direction as those of DiSilvestro et al. [15], who linked pain measured 6–12 weeks after minimally invasive transforaminal lumbar interbody fusion with 1-year ODI recovery. The weaker results at 2 weeks are clinically plausible: pain at that stage may still be dominated by the operation itself, limited activity, and short-term analgesic use. By 3 months, persistent leg pain is more likely to represent residual or recurrent radicular symptoms or a recovery course that has stalled. Back pain is less specific because it may arise from the surgical approach, paraspinal tissues, adjacent structures, or the underlying degenerative condition. These explanations remain tentative. Differences in procedure and timing also mean that our findings should not be presented as an external validation of the earlier threshold.

Clinical Interpretation

A leg-pain NRS score of 3 or higher at the 3-month visit may be a useful reason to look more closely at the patient's recovery rather than a treatment threshold in itself. The reassessment might focus on residual or recurrent radicular symptoms, neurological findings, rehabilitation progress, and whether imaging or specialist review is warranted. The interaction analysis did not provide a reliable basis for different cutoffs across deficit-burden groups. When a complete functional questionnaire is available, concurrent 3-month ODI remains more informative for subsequent ODI-MCID achievement.

The apparently high discrimination of the 3-month ODI model should also be interpreted in context. The landmark measure and the outcome come from the same disability scale, and baseline ODI was included in the model, so construct overlap and mathematical coupling are unavoidable. Calibration slopes of roughly 0.74–0.83 also show that predictions remained too extreme. In addition, the AUCs of 0.829, 0.910, and 0.985 are conditional internal-validation estimates for fixed post-selection models, not estimates from an independently validated model. Prospective multicenter data are needed before these models can be used for individual clinical decisions.

Strengths and Limitations

This study has several strengths. Rather than starting with a single preferred pain measure, we directly compared back and leg pain at two early follow-up visits. We then examined the best-performing measure as a continuous score, as a candidate binary cutoff, and in adjusted landmark models. The analysis also addressed nonlinearity and separation, used multiplicity-adjusted paired AUC comparisons and paired bootstrap resampling, and repeated the main findings across alternative ODI-MCID thresholds and spline specifications.

The limitations are also important. This was a retrospective single-center study, and case ascertainment depended on the procedure names and codes used in the surgery-system search. Forty-five of 234 clinically eligible patients (19.2%) were excluded because required follow-up data were incomplete, which may have introduced complete-case selection bias. Exact postoperative days and item-level ODI responses were not retained in the locked analytic dataset, so visit assignment and unanswered-section scoring could not be independently rechecked from that file. Marker, threshold, and spline choices were data driven, while the multivariable bootstrap held those choices fixed. The 3-month ODI benchmark is affected by construct overlap and mathematical coupling. High NRS scores were uncommon, the Barthel-based functional-dependence proxy and ≥ 2 -deficit threshold are not validated frailty definitions, and the interaction model was minimally adjusted with limited power. No external cohort was available. Finally, numerical reproduction of the analysis does not verify source-record accuracy, eligibility decisions, or transportability.

CONCLUSIONS

Among four early postoperative pain measures after ULIF, 3-month leg pain provided the clearest separation between patients who did and did not achieve the 1-year ODI MCID. Higher scores were associated with poorer functional recovery, and NRS ≥ 3 may be a practical signal for closer reassessment. Its added information beyond concurrent 3-month ODI was small, so the cutoff should not be used alone to predict outcome or guide treatment. External validation is needed before applying these findings to individual patients.

Abbreviations

ASA, American Society of Anesthesiologists; AUC, area under the receiver operating characteristic curve; BMI, body mass index; CHF, congestive heart failure; CI, confidence interval; COPD, chronic obstructive pulmonary disease; MCID, minimum clinically important difference; NPV, negative predictive value; NRS, numerical rating scale; ODI, Oswestry Disability Index; PPV, positive predictive value; RCS, restricted cubic spline; ROC, receiver operating characteristic; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology; ULIF, unilateral biportal endoscopic lumbar interbody fusion.

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and its later amendments. The Medical Research Ethics Committee of Weifang People's Hospital reviewed the investigators' qualifications, study protocol, application for waiver of informed consent, and related study materials, and approved the study on August 7, 2026 (approval no. KYLL20260807-1).

Consent for publication

Not applicable.

Clinical trial number:

Not applicable.

Competing interests

The authors declare that they have no competing interests.

ADDITIONAL FILES

Additional file 1

File format: PDF

Title: Supplementary tables and figure

Description: Supplementary Tables S1-S4 and Supplementary Figure S1, including sensitivity and functional-form analyses, bootstrap marker/threshold selection frequencies, score-level event distribution, paired DeLong comparisons, and longitudinal patient-reported outcome trajectories.

Funding

No external funding was received for this study.

Author Contribution

Authors' contributionsXZ drafted the manuscript. FL conceived and designed the study. HC and HJ collected and organized the clinical data. XL critically revised the manuscript. All authors contributed to the interpretation of the findings, read and approved the final manuscript, and agreed to be accountable for the work.

Acknowledgements

The authors thank the clinical and nursing teams involved in patient care and follow-up.

Data Availability

The datasets generated and/or analysed during the current study are not publicly available because they contain clinical data subject to patient privacy and institutional restrictions, but are available from the corresponding author on reasonable request, subject to institutional approval.

References

1. Global regional, national burden of low back pain. 1990–2020, its attributable risk factors, and projections to 2050: a systematic analysis of the Global Burden of Disease Study 2021. *Lancet Rheumatol.* 2023;5(6):e316–29.
2. Martin B, Mirza S, Spina N, Spiker W, Lawrence B, Brodke D. Trends in Lumbar Fusion Procedure Rates and Associated Hospital Costs for Degenerative Spinal Diseases in the United States, 2004–2015. *SPINE.* 2018;44:1.
3. Khan NR, Clark AJ, Lee SL, Venable GT, Rossi NB, Foley KT. Surgical Outcomes for Minimally Invasive vs Open Transforaminal Lumbar Interbody Fusion. *Neurosurgery.* 2015;77(6):847–74.
4. Heo DH, Son SK, Eum JH, Park CK. Fully endoscopic lumbar interbody fusion using a percutaneous unilateral biportal endoscopic technique: technical note and preliminary clinical results. *NeuroSurg Focus.* 2017;43(2).
5. Kim J-E, Choi D-J. Biportal Endoscopic Transforaminal Lumbar Interbody Fusion with Arthroscopy. *Clin Orthop Surg.* 2018;10(2).
6. Park M-K, Park S-A, Son S-K, Park W-W, Choi S-H. Clinical and radiological outcomes of unilateral biportal endoscopic lumbar interbody fusion (ULIF) compared with conventional posterior lumbar interbody fusion (PLIF): 1-year follow-up. *Neurosurg Rev.* 2019;42(3):753–61.
7. Liu G, Liu WX, Jin DJ, Yan PL, Yang ZC, Liu RP. Clinical outcomes of unilateral biportal endoscopic lumbar interbody fusion (ULIF) compared with conventional posterior lumbar interbody fusion (PLIF). *Spine J.* 2023;23(2):271–80.
8. Chen Y, Wu J, Li Q, Wu Y, Xu C, Chen R, et al. Clinical application of unilateral biportal endoscopic lumbar interbody fusion in lumbar degenerative diseases: a multi-center study. *Front Neurol.* 2025;16:1673110.
9. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. *Physiotherapy.* 1980;66(8):271–3.
10. Fairbank JC, Pynsent PB. The Oswestry Disability Index. *Spine (Phila Pa 1976).* 2000;25(22):2940–52. discussion 52.
11. Liu H, Tao H-R, Luo Z-J. Validation of the Simplified Chinese Version of the Oswestry Disability Index. *Spine.* 2009;34:1211–6. discussion 7.
12. Copay AG, Subach BR, Glassman SD, Polly DW, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J.* 2007;7(5):541–6.
13. Copay AG, Glassman SD, Subach BR, Berven S, Schuler TC, Carreon LY. Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and pain scales. *Spine J.* 2008;8(6):968–74.
14. Cha EDK, Lynch CP, Geoghegan CE, Jadcak CN, Mohan S, Singh K. Risk Factors for Failing to Reach a Minimal Clinically Important Difference Following Minimally Invasive Lumbar Decompression. *Int J Spine Surg.* 2022;16(1):51–61.
15. DiSilvestro KJ, Bay A, Kwas CT, Asada T, Hirase T, Zhang JS, et al. Early NRS Leg and Back Thresholds Predict Clinical Recovery After MIS Transforaminal Lumbar Interbody Fusion for Degenerative Spine Disease. *Spine.* 2025;50(9):628–35.
16. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, et al. Frailty in older adults: evidence for a phenotype. *J Gerontol Biol Sci Med Sci.* 2001;56(3):M146–56.
17. Clegg A, Young J, Iliffe S, Rikkert MO, Rockwood K. Frailty in elderly people. *Lancet.* 2013;381(9868):752–62.
18. Moskven E, Charest-Morin R, Flexman AM, Street JT. The measurements of frailty and their possible application to spinal conditions: a systematic review. *Spine J.* 2022;22(9):1451–71.
19. Chotai S, Gupta R, Pennings JS, Hymel AM, Archer KR, Zuckerman SL, et al. Frailty and Sarcopenia: Impact on Outcomes Following Elective Degenerative Lumbar Spine Surgery. *Spine.* 2022;47(20):1410–7.
20. Subramaniam S, Aalberg JJ, Soriano RP, Divino CM. New 5-Factor Modified Frailty Index Using American College of Surgeons NSQIP Data. *J Am Coll Surg.* 2018;226(2):173–e818.
21. Pierce KE, Naessig S, Kummer N, Larsen K, Ahmad W, Passfall L, et al. The Five-item Modified Frailty Index is Predictive of 30-day Postoperative Complications in Patients Undergoing Spine Surgery. *Spine.* 2021;46(14):939–43.
22. Subramanian T, Shahi P, Hirase T, Korsun M, Zhang J, Kim E, et al. High Modified 5 Factor Frailty Index is Associated With Worse PROMs and Decreased Return to Activities After 1 or 2 Level MI-TLIF. *Global Spine J.* 2025;15(6):2862–71.
23. Beauchamp-Chalifour P, Flexman AM, Street JT, Fisher CG, Ailon T, Dvorak MF, et al. The impact of frailty on patient-reported outcomes after elective thoracolumbar degenerative spine surgery. *J Neurosurgery-Spine.* 2021;35(5):607–15.

24. Weaver DJ, Malik AT, Jain N, Yu E, Kim J, Khan SN. The Modified 5-Item Frailty Index: A Concise and Useful Tool for Assessing the Impact of Frailty on Postoperative Morbidity Following Elective Posterior Lumbar Fusions. *World Neurosurg.* 2019;124:e626–32.
25. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet.* 2007;370(9596):1453–7.
26. Mahoney FI, Barthel DW. FUNCTIONAL EVALUATION: THE BARTHEL INDEX. *Md State Med J.* 1965;14:61–5.
27. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44(3):837–45.
28. Youden WJ. Index for rating diagnostic tests. *Cancer.* 1950;3(1):32–5.
29. Alam TF, Rahman MS, Bari W. On estimation for accelerated failure time models with small or rare event survival data. *BMC Med Res Methodol.* 2022;22(1):169.
30. Harrell FE Jr., Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. *J Natl Cancer Inst.* 1988;80(15):1198–202.
31. Steyerberg EW, Harrell FE Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol.* 2001;54(8):774–81.

Tables

Table 1. Baseline and perioperative characteristics of the analytic cohort.

Variable	Overall (n=189)	ODI-MCID non-achievement (n=77)	ODI-MCID achievement (n=112)	Absolute standardized difference
Age, years	65.0 [58.0, 69.0]	65.0 [57.0, 70.0]	64.0 [58.0, 69.0]	0.002
Female sex	108 (57.1%)	43 (55.8%)	65 (58.0%)	0.044
BMI, kg/m ²	26.0 [23.8, 28.0]	26.1 [23.4, 28.0]	25.8 [23.9, 27.7]	0.087
BMI category, kg/m ²				
<25	78 (41.3%)	31 (40.3%)	47 (42.0%)	0.035
25–29.9	88 (46.6%)	34 (44.2%)	54 (48.2%)	0.081
≥30	23 (12.2%)	12 (15.6%)	11 (9.8%)	0.174
Higher five-domain deficit burden (≥2 deficits)	39 (20.6%)	23 (29.9%)	16 (14.3%)	0.383
Diabetes mellitus	38 (20.1%)	22 (28.6%)	16 (14.3%)	0.354
Hypertension	66 (34.9%)	34 (44.2%)	32 (28.6%)	0.328
COPD	6 (3.2%)	2 (2.6%)	4 (3.6%)	0.056
Congestive heart failure	19 (10.1%)	11 (14.3%)	8 (7.1%)	0.232
Functional dependence (Barthel Index <60)	30 (15.9%)	14 (18.2%)	16 (14.3%)	0.106
ASA class	2.0 [2.0, 2.0]	2.0 [2.0, 2.0]	2.0 [2.0, 2.0]	0.102
Length of stay, days	7.0 [5.0, 10.0]	7.0 [5.0, 10.0]	7.0 [6.0, 10.0]	0.076
Operative time, h	5.0 [4.3, 6.0]	5.0 [4.4, 5.9]	5.1 [4.3, 6.0]	0.045
Estimated blood loss, mL	100 [50, 100]	100 [50, 100]	100 [50, 100]	0.028
Index level				
L2/3	4 (2.1%)	0 (0.0%)	4 (3.6%)	0.272
L3/4	15 (7.9%)	4 (5.2%)	11 (9.8%)	0.176
L4/5	159 (84.1%)	69 (89.6%)	90 (80.4%)	0.261
L5/S1	11 (5.8%)	4 (5.2%)	7 (6.2%)	0.045

Values are median [interquartile range] or n (%). Continuous standardized differences used group means and pooled variances; binary differences used level-specific proportions and pooled Bernoulli variances. BMI and index-level categories were assessed separately. Values were descriptive and not used for covariate selection. Functional dependence was defined as preoperative Barthel Index <60. The five-domain deficit burden combined this variable with diabetes mellitus, COPD, congestive heart failure, and treated hypertension; higher burden was defined as ≥2 deficits. This study-defined burden is not a validated mFI-5 score. COPD, chronic obstructive pulmonary disease; ASA, American Society of Anesthesiologists.

Table 2. Patient-reported outcome trajectories in the analytic cohort.

Outcome	Time point	Overall (n=189)	ODI-MCID non-achievement (n=77)	ODI-MCID achievement (n=112)
NRS back pain	Preoperative	5.6 ± 2.4	5.7 ± 2.3	5.5 ± 2.5
NRS back pain	2 weeks	2.8 ± 1.5	3.3 ± 1.6	2.5 ± 1.4
NRS back pain	3 months	3.2 ± 2.0	4.5 ± 1.9	2.3 ± 1.6
NRS back pain	1 year	2.2 ± 1.9	3.8 ± 1.8	1.2 ± 1.0
NRS back pain	Improvement at 1 year	3.3 ± 2.6	1.8 ± 1.9	4.3 ± 2.5
NRS leg pain	Preoperative	5.4 ± 2.4	5.6 ± 2.4	5.3 ± 2.5
NRS leg pain	2 weeks	2.9 ± 1.5	3.3 ± 1.6	2.6 ± 1.4
NRS leg pain	3 months	2.5 ± 1.8	3.9 ± 1.7	1.5 ± 1.1
NRS leg pain	1 year	1.8 ± 1.9	3.3 ± 1.8	0.8 ± 1.0
NRS leg pain	Improvement at 1 year	3.6 ± 2.6	2.4 ± 2.0	4.5 ± 2.6
ODI	Preoperative	52.2 ± 6.9	51.8 ± 5.9	52.5 ± 7.4
ODI	2 weeks	41.8 ± 10.0	46.6 ± 6.5	38.5 ± 10.7
ODI	3 months	39.2 ± 11.7	48.3 ± 6.3	33.0 ± 10.3
ODI	1 year	35.8 ± 12.6	47.9 ± 6.7	27.5 ± 8.1
ODI	Improvement at 1 year	16.4 ± 12.5	3.9 ± 3.3	25.1 ± 8.7

Values are presented as mean ± standard deviation for descriptive purposes. Back- and leg-pain NRS values were available at all four time points for all 189 patients; no NRS values were imputed.

The table is descriptive. Outcome groups were defined by 1-year ODI improvement; repeated time-specific P values were therefore removed to avoid circular and multiplicity-heavy inference.

Improvement is defined as the preoperative score minus the 1-year score; positive values indicate improvement. ODI, Oswestry Disability Index; NRS, numerical rating scale.

Table 3A. Apparent ROC performance and post hoc marker-selection frequency in the analytic cohort.

Early pain measure	AUC	95% CI	Observed threshold	Sensitivity	Specificity	PPV	NPV	Accuracy	Youden J	Marker selected in 10,000 marker-and-threshold selection bootstraps
2-week leg-pain NRS	0.594	0.516-0.673	≥2	100.0%	27.7%	48.7%	100.0%	57.1%	0.277	0 (0.00%)
2-week back-pain NRS	0.632	0.556-0.708	≥2	100.0%	27.7%	48.7%	100.0%	57.1%	0.277	0 (0.00%)
3-month leg-pain NRS	0.891	0.850-0.933	≥3	75.3%	84.8%	77.3%	83.3%	81.0%	0.601	9,999 (99.99%)
3-month back-pain NRS	0.816	0.758-0.874	≥4	71.4%	75.9%	67.1%	79.4%	74.1%	0.473	1 (0.01%)

AUC confidence intervals were calculated using the DeLong method. Integer thresholds maximized the Youden index; when thresholds were tied, the smaller integer threshold was selected. Threshold positivity was defined as predicting 1-year ODI-MCid non-achievement.

Marker selection frequencies are from 10,000 ordinary patient-level bootstrap samples repeating the marker-and-threshold selection workflow, in which all four AUCs and the maximum-Youden integer threshold were reselected.

The 3-month leg-pain AUC exceeded the 3-month back-pain AUC by 0.076 (Holm-adjusted $P=0.002$) and both 2-week measures by larger margins (Holm-adjusted $P<0.001$). Supplementary Table S4 reports all six paired comparisons. PPV, positive predictive value; NPV, negative predictive value.

Table 3B. Selection-aware performance of the selected continuous marker and its thresholded binary rule.

Metric	Apparent	Mean optimism	Corrected
Selected continuous marker AUC	0.8915	0.0001	0.8914
Thresholded-rule sensitivity	0.753	0.012	0.742
Thresholded-rule specificity	0.848	0.008	0.840
Thresholded-rule balanced accuracy	0.8007	0.0096	0.7911
Thresholded-rule PPV	0.773	0.008	0.765
Thresholded-rule NPV	0.833	0.007	0.826
Thresholded-rule accuracy	0.810	0.009	0.800
Thresholded-rule Youden J	0.601	0.019	0.582

The continuous-marker AUC and thresholded-rule metrics are distinct. For a binary rule, the corresponding two-point ROC AUC is numerically equal to balanced accuracy=(sensitivity+specificity)/2; balanced accuracy is reported in the table to avoid confusion with the continuous-marker AUC. Corrected estimates equal apparent performance minus mean bootstrap-training-to-original-sample-test optimism. The selected-marker AUC optimism was only 0.0001 because 3-month leg pain was selected in 9,999 of 10,000 bootstrap samples.

Table 4A. Conditional bootstrap performance of six fixed post-selection 3-month landmark models.

Landmark model	Apparent AUC	Corrected AUC	Corrected Brier	Corrected calibration intercept	Corrected calibration slope	Successful bootstraps
Clinical benchmark	0.611	0.556	0.246	-0.121	0.675	1000/1000
Benchmark + leg-pain RCS	0.928	0.910	0.126	0.026	0.893	1000/1000
Benchmark + leg-pain NRS ≥ 3	0.851	0.829	0.155	-0.019	0.937	1000/1000
Benchmark + 3-month ODI	0.980	0.973	0.055	0.001	0.835	1000/1000
Benchmark + 3-month ODI + NRS ≥ 3	0.982	0.975	0.047	0.009	0.803	1000/1000
Benchmark + 3-month ODI + leg-pain RCS	0.993	0.985	0.042	-0.032	0.739	1000/1000

The event was failure to achieve the 1-year ODI MCID. All models used Firth logistic regression. The benchmark included age, sex, BMI, higher five-domain deficit burden, and baseline ODI. Three-month leg pain was selected after comparing four candidate markers; RCS denotes its fixed four-knot restricted cubic spline.

Corrected values used the same 1,000 patient-level bootstrap samples and are conditional on the selected marker, fixed knot locations, and fixed data-derived NRS ≥ 3 threshold.

AUC, area under the receiver operating characteristic curve; BMI, body mass index; ODI, Oswestry Disability Index; NRS, numerical rating scale; RCS, restricted cubic spline. These are internal-validation estimates and do not establish transportability.

Table 4B. Conditional paired bootstrap incremental performance beyond the fixed 3-month ODI landmark model.

Added fixed leg-pain term	Conditional corrected Δ AUC	Conditional corrected Δ Brier	Conditional corrected Δ calibration slope
NRS ≥ 3	0.0022	-0.0087	-0.032
Four-knot RCS	0.0125	-0.0135	-0.096

Differences are full minus the 3-month-ODI model and use the same fixed-model resamples. Positive Δ AUC and negative Δ Brier favor the added term. Paired corrected differences were calculated within shared bootstrap resamples and therefore do not necessarily equal the arithmetic difference between separately summarized corrected estimates in Table 4A. Estimates are conditional on the fixed post-selection marker, threshold, and spline choices.

Figures

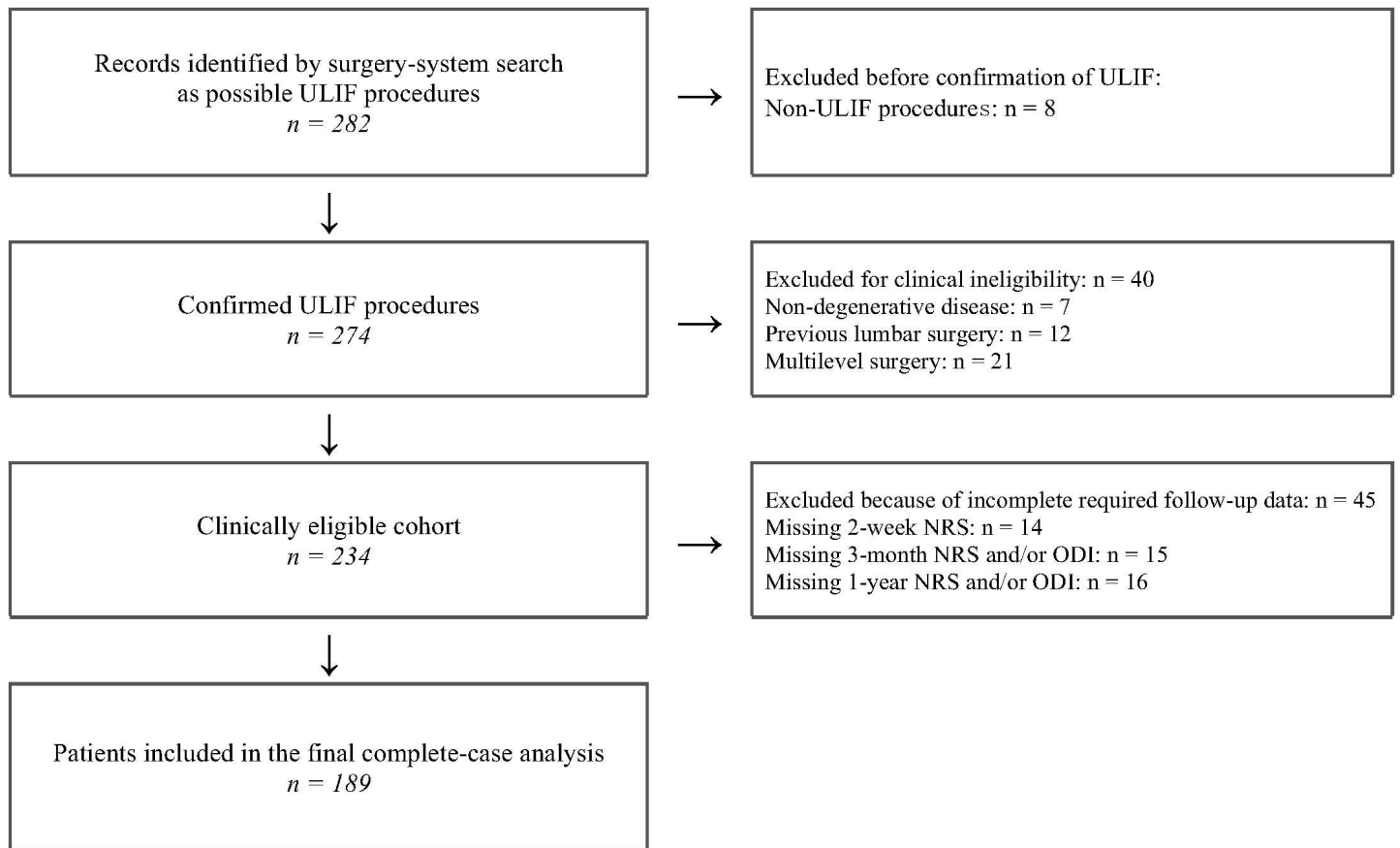


Figure 1

Flow diagram of patient selection. A total of 282 records identified by the surgery-system search as possible ULIF procedures were screened. After exclusion of non-ULIF procedures and clinically ineligible cases, 234 patients constituted the clinically eligible cohort. A further 45 patients were excluded because of incomplete required follow-up data, leaving 189 patients in the final complete-case analytic cohort. ULIF, unilateral biportal endoscopic lumbar interbody fusion; NRS, numerical rating scale; ODI, Oswestry Disability Index.

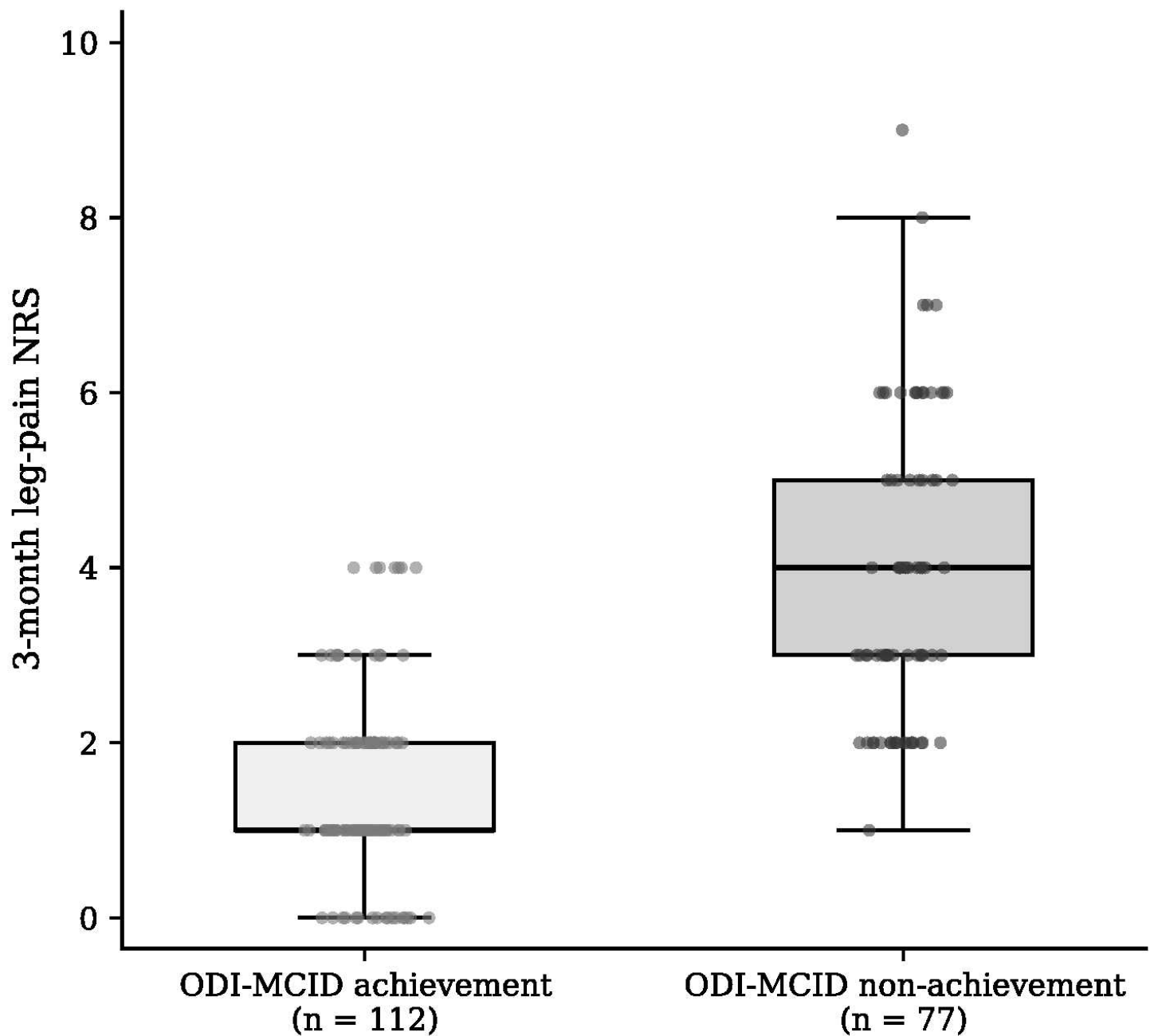


Figure 2

Distribution of 3-month leg-pain NRS scores according to 1-year ODI-MCID status. Individual points represent patients. Boxes indicate the interquartile range, center lines indicate the median, and whiskers extend to 1.5 times the interquartile range. The ODI-MCID achievement and non-achievement groups comprised 112 and 77 patients, respectively. ODI, Oswestry Disability Index; MCID, minimum clinically important difference; NRS, numerical rating scale.

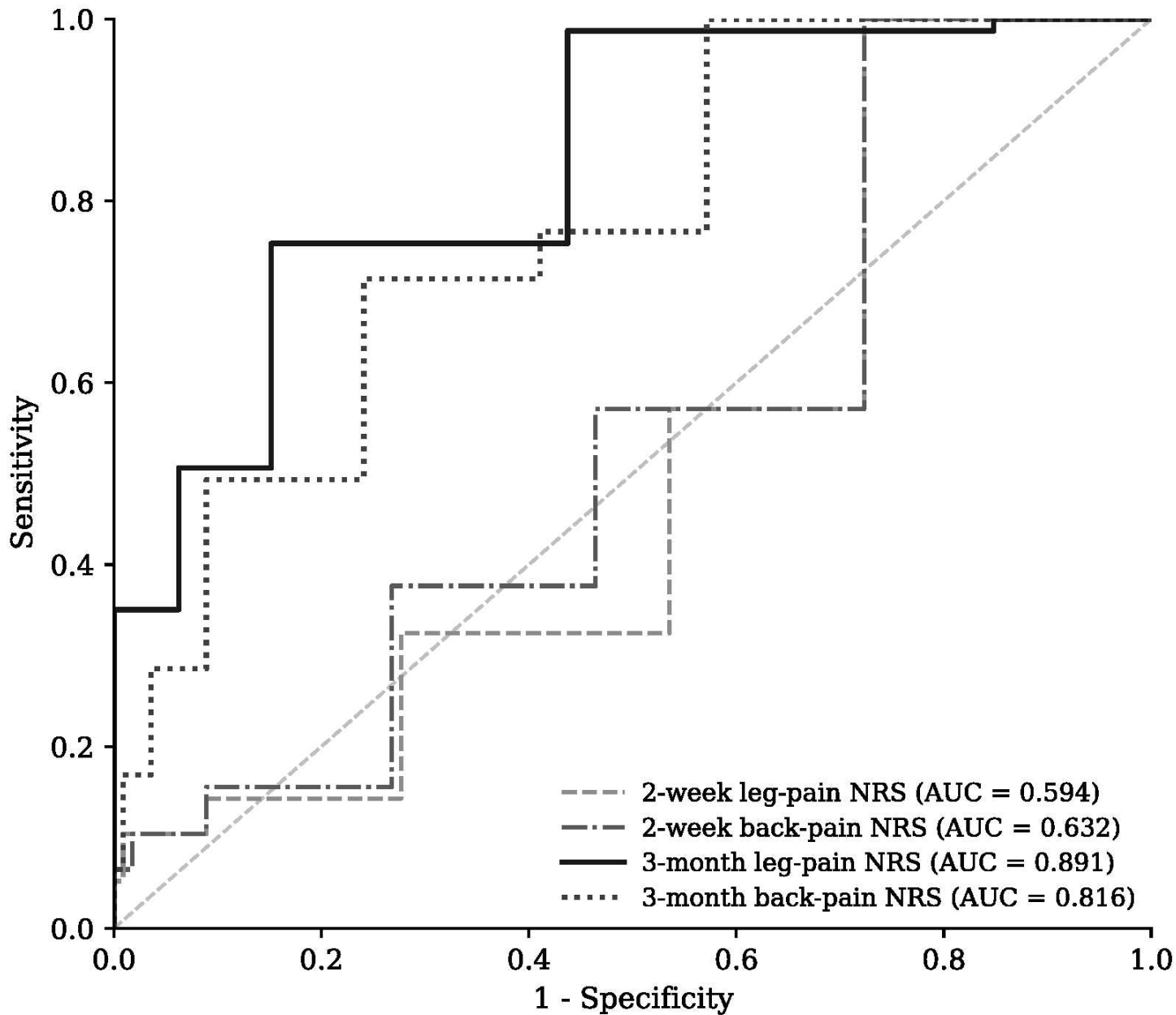


Figure 3

ROC curves of four early postoperative pain measures for 1-year ODI-MCID non-achievement. Stepwise receiver operating characteristic curves are shown for 2-week leg-pain NRS, 2-week back-pain NRS, 3-month leg-pain NRS, and 3-month back-pain NRS. The corresponding areas under the curve were 0.594, 0.632, 0.891, and 0.816, respectively. The diagonal dashed line represents discrimination no better than chance. AUC, area under the curve; ODI, Oswestry Disability Index; MCID, minimum clinically important difference; NRS, numerical rating scale.

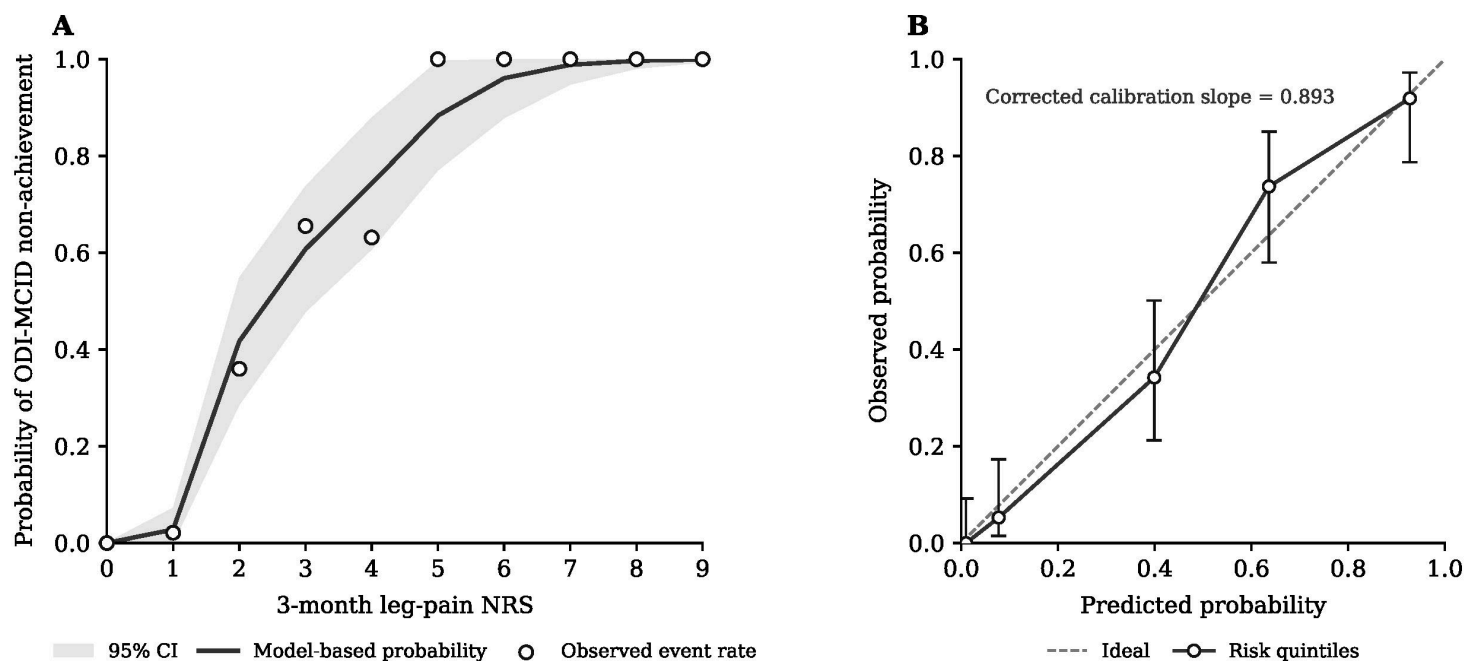


Figure 4

Association of 3-month leg-pain NRS with 1-year ODI-MCID non-achievement and calibration of the selected-marker model. (A) Standardized marginal probabilities of 1-year ODI-MCID non-achievement across 3-month leg-pain NRS scores estimated using the four-knot restricted cubic spline model. The solid line represents model-based probabilities, the shaded area represents 95% confidence intervals, and open circles indicate observed event rates at each integer NRS score. (B) Calibration of the selected-marker model across apparent risk quintiles. Points represent observed event probabilities within each quintile, vertical error bars indicate Wilson 95% confidence intervals, and the dashed diagonal line represents ideal calibration. The optimism-corrected calibration slope was 0.893. ODI, Oswestry Disability Index; MCID, minimum clinically important difference; NRS, numerical rating scale.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1Supplementarymaterial2.pdf](#)