Lemel *et al. Art Int Surg.* 2025;5:82-102
DOI: 10.20517/ais.2024.42

Artificial Intelligence Surgery

Systematic Review

Open Access

Predictive models for patient-reported outcomes (PROs) in elective spine surgery: a systematic review

Hannah Lemel[1,2], David Shin[1,3], Seth Meade[1,3], Brittany Lapin[1], Thomas Mroz[1], Michael Steinmetz[1], Ghaith Habboub[1]

[1]Neurological Institute, Spine Research Laboratory, Cleveland Clinic, Cleveland, OH 44195, USA.
[2]School of Medicine, Case Western Reserve University, Cleveland, OH 44106, USA.
[3]Lerner College of Medicine, Cleveland Clinic, Cleveland, OH 44195, USA.

Correspondence to: Dr. Ghaith Habboub, Neurological Institute, Spine Research Laboratory, Cleveland Clinic, 9500 Euclid Ave, Cleveland, OH 44195, USA. E-mail: habboug@ccf.org

How to cite this article: Lemel H, Shin D, Meade S, Lapin B, Mroz T, Steinmetz M, Habboub G. Predictive models for patient-reported outcomes (PROs) in elective spine surgery: a systematic review. *Art Int Surg.* 2025;5:82-102. https://dx.doi.org/10.20517/ais.2024.42

Received: 15 Jun 2024 **First Decision:** 14 Jan 2025 **Revised:** 20 Jan 2025 **Accepted:** 23 Jan 2025 **Published:** 17 Feb 2025

Academic Editor: Andrew Gumbs **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Aim: A growing body of literature reports on prediction models for patient-reported outcomes of spine surgery, carrying broad implications for use in value-based care and decision making. This review assesses the performance and transparency of reporting of these models.

Methods: We queried four studies reporting the development and/or validation of prediction models for patient-reported outcome measures (PROMs) following elective spine surgery with performance metrics such as the area under the receiver operating curve (AUC) scores. Adherence to transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD-AI) guidelines was assessed. One representative model was selected from each study.

Results: Of 4,471 screened studies, 35 were included, with nine development, 24 development and evaluation, and two evaluation studies. Sixteen machine learning models and 19 traditional prediction models were represented. Oswestry disability index (ODI) and modified Japanese Orthopaedic Association (mJOA) scores were most commonly used. Among 29 categorical outcome prediction models, the median [interquartile range (IQR)] AUC was 0.79 [0.73, 0.84]. The median [IQR] AUC was 0.825 [0.76, 0.84] among machine learning models and 0.74

[0.71, 0.81] among traditional models. Adherence to TRIPOD-AI guidelines was inconsistent, with no studies commenting on healthcare inequalities in the sample population, model fairness, or disclosure of study protocols or registration.

Conclusion: We found considerable variation between studies, not only in chosen patient populations and outcome measures, but also in their manner of evaluation and reporting. Agreement about outcome definitions, more frequent external validation, and improved completeness of reporting may facilitate the effective use and interpretation of these models.

Keywords: Patient reported outcomes, predictive modeling, accuracy, machine learning

INTRODUCTION

Patient-reported outcome measures (PROMs) have been widely adopted across surgical subspecialties, including spine surgery, as the gold-standard method of assessing surgical success by incorporating the patient perspective. In comparison to traditional outcome measures such as reoperation and readmission rates, PROMs aim to capture nuanced aspects of the patient experience, including pain levels, functional improvement, and overall satisfaction with treatment. Accordingly, PROMs are becoming ubiquitous in care delivery and assessment and have been used in a variety of applications, ranging from use in value-based healthcare models and assessment of institution- and surgeon-level outcomes to clinical applications evaluating individual pre- and postoperative patient symptom status and prediction of surgical outcomes[1-5].

Implementing PROMs into standard clinical practice has presented several challenges. The collection of PROMs, whether in a research setting or as part of routine practice, requires significant resources in cost and time, leading to limited completeness in the collection of PROM data[1]. Interpretability also poses a significant limitation, as PROMs rely on subjective reporting of symptoms and their severity, whose perception necessarily varies from patient to patient for any single value and can vary widely over time with postoperative changes. A wide range of instruments are used in spine surgery alone, each with variable degrees of validation in psychometric properties including validity and reliability, complicating data comparison across individuals or groups[6,7]. Examples of commonly used PROMs in the evaluation of degenerative spine disease include the Oswestry disability index (ODI), modified Japanese Orthopaedic Association (mJOA) Scale, and patient reported outcomes measurement information system (PROMIS) scales[8]. Efforts have been made to identify the PROM instruments that are most appropriate and applicable to the assessment of spine surgery outcomes[6,7,9].

Given the importance of these measures to determine surgical success and their strong association with reimbursement and performance assessment models, many have built prediction models leveraging patient-reported outcome (PRO) as both outcomes and predictors; however, these models have classically been plagued by inter and intrapatient variability and often had only moderate quality performance. Machine learning has gained prominence as a set of approaches for developing such predictive models with potential improvements to predictive performance. These models highlight possible benefits to patient selection, decision making, and adverse event prevention that carry broad implications for patient-centered care. In the future, predictive models such as these may be used to help inform decision making between patients and physicians regarding the expected risks and benefits of surgery. Predictive models have already been developed to identify surgical indications, predict postoperative complications, and evaluate outpatient suitability for a range of spinal conditions[10]. In the context of these developments, it is important to assess trends in completeness of reporting for these predictive models, as well as to evaluate the risk of bias in handling challenges such as incomplete or missing data. Furthermore, in an age where machine learning is

being applied in several novel contexts, we aimed to summarize the impact of the rise of machine learning on predictive model performance and to identify avenues for future development of these models to improve our ability to identify candidates who have the best chances of attaining a benefit from surgery.

In this review, we summarize the existing predictive models for questionnaire-based PROMs outcomes of spine surgery, characterizing the PROMs commonly used for prediction as well as patterns in transparent reporting of model development and validation. We evaluate whether the application of machine learning in recent years has been accompanied by improvements in predictive performance for PROM outcomes.

METHODS

Search strategy

This systematic review was registered in the International Prospective Register of Systematic Reviews (PROSPERO ID: CRD42024536045) guidelines[11]. PubMed, Embase, Web of Science, and Scopus databases were queried to identify original publications describing prediction models for PROM outcomes following elective spine surgery published between January 1, 2010, and April 1, 2024. Queries were developed for each database with different domains of medical subject heading terms representing spine surgery, predictive models, and PROMs combined with "AND", and related words within each domain combined with "OR" [Supplementary Appendix 1]. Studies identified from these database searches were assessed for duplicates and then screened by two separate reviewers (HL, DS). The remaining full texts were assessed for eligibility. Disagreements over inclusion were resolved through consensus or through consultation with another author (SM). Additional searches were performed through a review of the referenced work of included articles.

Eligibility criteria

Studies were considered for inclusion if they described the development and/or validation of prediction models for PROM outcomes following elective spine surgery. Only studies that reported performance metrics for predictive models [e.g., the area under the receiver operating curve (AUC) for binary classification models] and assessed questionnaire-based PROMs (e.g., PROMIS, ODI, mJOA) as predicted outcomes were included. Studies were screened out if they included regression analyses without assessing their utility in a predictive model (e.g., studies that highlighted risk factors from multivariate analyses without commenting on or reporting model performance in the abstract). Where multiple studies met the above criteria and reported predictive models developed on identical or overlapping data, only the most recent such study was selected for inclusion.

Data extraction

Two separate authors (HL, DS) reviewed included studies and extracted details, including year and journal of publication, type of study (development and/or evaluation of a model), type of prediction model (machine learning *vs.* traditional approaches; specific models used), patient population, predicted outcomes, sample size, number of predictors, performance metrics assessed, handling of missing data, and adherence to items of the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD-AI) guidelines[12]. Disagreements were resolved first through consensus or by consultation with another author (SM).

To avoid overrepresentation of studies with multiple models in quantitative analyses, individual models were selected from those studies reporting multiple models (whether for multiple patient populations/subsets, multiple outcome definitions, model types) with the following method: the eligible model (predicting a questionnaire-based PROMs outcome) was selected for analysis that was presented by the authors as the best performing model, or if no individual model was designated in such a manner, the

model exhibiting greatest performance in discrimination was selected for categorical outcome prediction models, or the one explaining the greatest proportion of variation in continuous outcome prediction models.

Studies were designated as reporting development only if no methods of internal or external validation were reported beyond reporting performance on training data, evaluation only if an existing model was applied to new data, and both development and evaluation if model development and either internal or external development were reported. Methods of external validation were considered to include validation using an external dataset or a dataset consisting of test data from distinct centers from those used in training. Methods of internal validation were considered to include cross-validation (k-fold or leave-one-out), bootstrapping with resampling, temporal validation, or holdout test data from the same data sources as used in training. Studies were considered to report only evaluation on development data if performance metrics were only reported for the data points used in the training of models (i.e., without methods such as cross-validation, or bootstrapping).

While TRIPOD-AI may be used to guide study design and reporting in prediction model studies, its aims also include uses for the evaluation of published studies. Adherence to TRIPOD-AI items was recorded as "Yes", "No", or "Not applicable", with the last designation applied where an item was either not applicable to the study type as specified in the TRIPOD-AI guidelines or conditional on aspects of study design not applicable to the study or model. Acknowledgment or mention of following TRIPOD guidelines (either the prior 2015 TRIPOD guidelines or 2024 TRIPOD-AI guidelines) was also recorded[12,13].

Synthesis and analysis

Extracted data were tabulated for each included study. The distribution of predictive performance in discrimination, in the form of AUC, was assessed through median and interquartile range (IQR) and compared between models using machine learning *vs.* traditional methods for categorical prediction outcomes. These distributions were visualized with boxplots overlaid with scores from individual studies, designated as being derived from either external validation, internal validation, or training data. For studies reporting multiple scores for a predictive model, one representative score was selected for the model in the following order of preference by the method of evaluation: external validation, internal validation, and lastly, training data. Given the broad scope of models included in this review, with variations in patient populations and definitions of predicted outcomes, these distributions were presented to characterize broad trends, without further application of statistical tests to compare performance between models of the machine learning or traditional model categories. Adherence to TRIPOD-AI guidelines was represented with bar graphs denoting the proportion of models compliant with each item of the guidelines. In addition, TRIPOD-AI adherence was visualized with models grouped by their type (machine learning or traditional model), and whether the study mentioned adherence to either the TRIPOD 2015 or TRIPOD-AI 2024 guidelines.

For each study, a summary adherence score for the TRIPOD-AI guidelines was calculated as 100% times the number of items graded "Yes" divided by the sum of items graded "Yes" or "No" (i.e., items graded "Not applicable" were excluded). The distribution of these scores was represented through median and IQR.

Risk of bias was assessed through reporting of missing data handling (TRIPOD-AI item 11), effective sample size (using a threshold of 10 development data points for each predictor/level), and reporting of calibration.

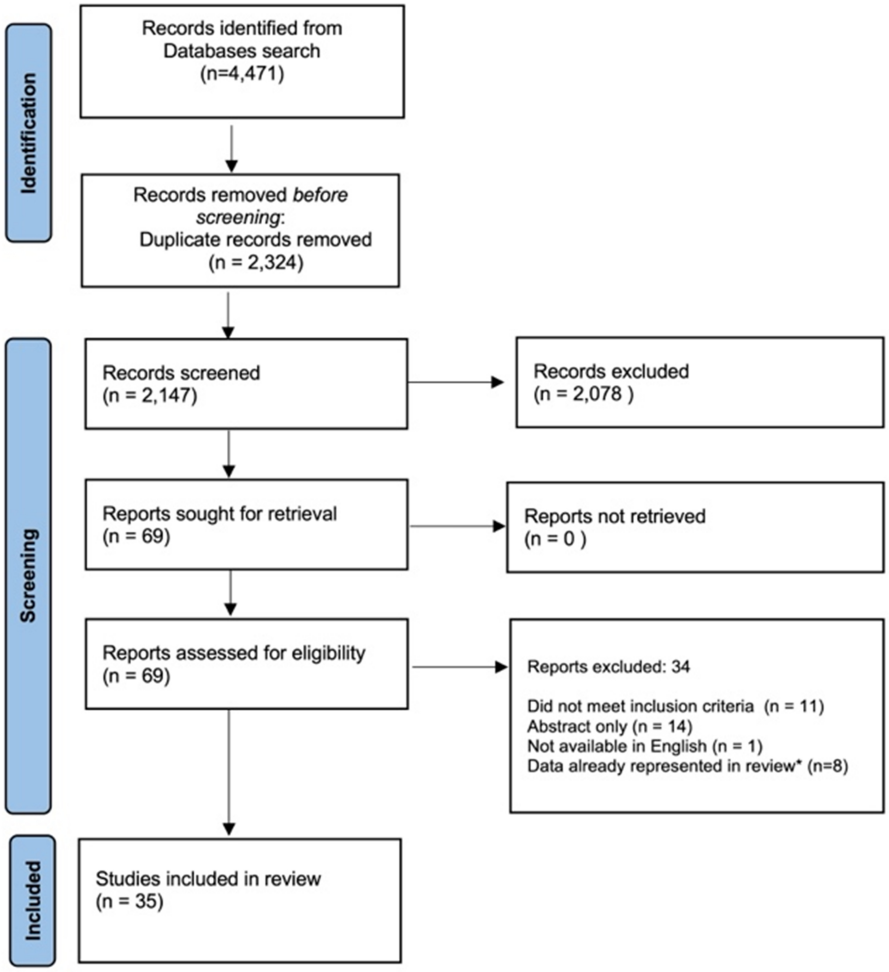

Figure 1. PRISMA flow diagram. *Studies reporting earlier versions of models trained on the same dataset were excluded, and the most recent publication evaluating model performance was included for review. PRISMA: Preferred reporting items for systematic reviews and meta-analyses.

Ethics, funding, conflicts of interest

Given the absence of any patient contact or study interventions for this systematic review, permission from the institutional review board was not required. This analysis was not supported by any funding, and the authors reported no additional disclosures or conflicts of interest.

RESULTS

Database search and study identification

Search queries identified 4,471 records from four databases, of which 2,324 were removed as duplicates. Screening by title and abstract of the remaining 2,147 records yielded 69 records for full-text screening, all of which were retrievable at the time of review. The application of selection criteria yielded 35 studies and their representative models. A flowchart depicting the search and review process is shown in Figure 1.

Study characteristics

The characteristics of the included studies and their selected representative models are shown in Table 1. Of the 35 included studies, nine reported only model development, 24 reported both model development and evaluation, and two reported only evaluation (external validation) of previously developed models. 16

Table 1. Characteristics of included studies

First author and year	Population	Outcome	Model type	Development and/or evaluation; method	Sample size	Predictors (N)	Performance metrics	Online calculator
Berg, 2024[24]	Lumbar discectomy[a]	Categorical: 1-year ODI improvement ≥ 22 pts[b]	ML: XGB[c]	DE; 5-fold regional cross-validation	22,707 (total), (internal-external cross-validation with 5-fold regional cross-validation)	25	AUC, PPV, NPV	https://huggingface.co/spaces/martingorosito/aidspine_hdsurgery_calculator
Carreon, 2024[29]	Elective lumbar spine surgery	Categorical: 1-year ODI improvement > 0 pts	T: logistic regression[d]	E; external dataset; temporal validation	8,105 (external dataset), 24,755 (temporal validation)	15	AUC	https://statcomp2.app.vumc.org/app_0
Pedersen, 2024[30]	Surgical decompression for LSS	Categorical: 1-year ODI improvement ≥ 14 pts	ML: MARS[c]	DE; independent test set (*N* = 228)	6,585 (total), 6,357 (development), 228 (testing)	7	AUC, Brier score, sensitivity, specificity, PPV, NPV	
Halicka, 2023[31]	Surgical decompression with or without fusion for LDH and/or LSS	Categorical: COMI improvement ≥ 2.2 pts, baseline to last available follow-up (collected 3, 12, 24 months)	T: logistic regression[d]	DE; temporal validation	4,307 (total), 2,691 (development),1,616 (testing)	34	AUC, Brier score, sensitivity, specificity, Nagelkerke R2	
Matsukura, 2023[32]	Surgery for cervical myelopathy due to OPLL	Categorical: mJOA recovery rate {2-year mJOA improvement/[(17 - preoperative JOA) * 100]} ≥ 52.8	T: logistic regression	D; evaluated on development data	395	6	AUC	
Rushton, 2023[33]	LSF	Categorical: 6-week ODI improvement ≥ 14.3 pts	T: logistic regression[c]	DE; temporal validation	1,200 (total), 600 (development), 600 (testing)	27	AUC	
Geere, 2023[34]	Surgical decompression with or without fusion for LDH and/or LSS[a]	Categorical: 1-year ODI ≤ 22 pts	T: thresholded linear regression[c]	DE; temporal validation	1,416 (total) 1,228 (development), 188 (testing)	18	AUC	https://spinepredictor.webflow.io/
Zhang, 2023[35]	Surgery for CSM[a]	Categorical: 2-year mJOA improvement ≥ 2 pts	ML: SVM[c]	DE; LOOCV	50	Unspecified	AUC, accuracy	
Chen, 2023[36]	Tubular microdiscectomy for LDH	Categorical: 1-year treatment improvement rate for lumbar spine JOA score {[(post-treatment score - pre-treatment score) ÷ (full score 29 - pre-treatment score)] × 100%} > 60%	T: logistic regression	DE; bootstrapping	273	5	AUC	https://fabinlin.shinyapps.io/DynNomapp/
Jaja, 2023[15]	Surgical decompression, cervical, with or without fusion[a]	Categorical: Dichotomized 2-year recovery trajectory of mJOA, derived by nonlinear GBTM	T: logistic regression	DE; bootstrapping	757	11	AUC sensitivity, specificity	

Sundaramoorthy, 2023[16]	Conservative treatment or interlaminar sequestrectomy for low back pain	Continuous: 5-month ODI improvement	ML: deep neural network[c]	DE; 10-fold cross-validation	70	12	MAE	
Staartjes, 2022[14]	LSF for degenerative pathology	Categorical: 1-year ODI improvement ≥ 15 pts or 1-year COMI improvement ≥ 2.2 pts	ML: elastic net-regularized GLM	DE; holdout test set (3/11 centers)	1,115 (total), 730 (development), 269 (testing)	10	AUC, accuracy, sensitivity, specificity, PPV, NPV	https://neurosurgery.shinyapps.io/fuseml/
Dong, 2022[37]	LIF for degenerative lumbar spondylolisthesis[a]	Categorical: 2-week ODI improvement ≥ 60%	ML: SVM[c]	DE; 10-fold cross-validation	157	9	AUC, accuracy, precision, confusion matrix	
Pedersen, 2022[38]	Lumbar discectomy	Categorical: 1-year EQ-5D improvement ≥ 0.17 pts	ML: SVM[d]	DE; holdout test set (15%)	1,968 (total), 1,673 (development), 295 (testing)	16	AUC, accuracy, specificity, sensitivity, PPV, NPV	
Coric, 2022[39]	Lumbar total disc replacement	Categorical: 7-year ODI > 17.63 (mean in trial cohort)	T: logistic regression[c]	DE; holdout (nonrandomized from trial) test set ($N = 52$)	334 (total), 283 (development), 52 (testing)	5	AUC	
Purohit, 2022[40]	Lumbar spine surgery for degenerative disease	Categorical: 6-month mODI improvement > 0 pts	ML: RF[c]	DE; holdout test set (20%)	180 (total), 144 (development), 36 (testing)	24	AUC	http://134.209.148.167:5000*
Wirries, 2022[17]	Conservative treatment or microscopically or endoscopically assisted interlaminar or translaminar sequestrectomy for LDH	Continuous: 6-month ODI improvement	ML: decision tree[c]	DE; 10-fold cross-validation	123	19	MAE	
Khan, 2021[41]	Surgical decompression for DCM	Categorical: 1-year mJOA improvement ≥ 0 pts	ML: SVM[c]	DE; holdout test set (20%)	702 (total), 562 (development), 140 (testing)	Unspecified	AUC, accuracy, sensitivity, specificity	
Budiono, 2021[42]	L5/S1 ALIF for DDD	Categorical: ODI improvement ≥ 20 pts, at least 8 months postoperatively	T: logistic regression	D; evaluated on development data	68	4	AUC, sensitivity, specificity	
Werner, 2021[43]	Lumbar microdiscectomy, medium-risk group (34 pts > baseline ODI ≥ 29pts)[a]	Categorical: 12-month ODI ≥ 47	T: logistic regression	DE; holdout test set (30%)	3,796 (total), 2,772 (development), 1,024 (testing)	11	AUC	
Pilato, 2021[44]	Surgery for CSM	Categorical: postoperative (6-12 month) mJOA improvement ≥ 6 pts	T: logistic regression	D; evaluated on development data	76	9	AUC	
Karhade, 2021[45]	1- or 2-level posterior decompression for lumbar disc herniation or lumbar spinal stenosis	Categorical: 1-year PROMIS-PF improvement ≥ 2 pts	ML: elastic-net penalized logistic regression[c]	DE; holdout test set (30%)	759 (total), 532 (development), 227 (testing)	23	AUC, Brier score	https://sorg-apps.shinyapps.io/promis_pld_mcid/*
Berjano, 2021[25]	Lumbar arthrodesis	Categorical: 6-month ODI	ML: RF[c]	D; evaluated on	1,243	5	AUC, sensitivity,	

		improvement ≥ 12.7 pts		development data			specificity	
Zhang, 2021[46]	Surgical treatment for CSM[a]	Categorical: long-term follow-up (≥ 3 years) mJOA ≥ 16 pts	ML: SVM[c]	DE; holdout test set (*N* = 41)	151 (total), 110 (development), 41 (testing)	237	AUC, accuracy, precision, sensitivity, specificity	
Quddusi, 2020[47]	LSF	Categorical: 1-year ODI improvement ≥ 15 pts	T: logistic regression	E; external dataset	100	4	AUC, Brier score	https://becertain.org/spine-lumbar-fusion-outcomes-calculator
Ford, 2020[19]	Lumbar discectomy	Continuous: 6-month ODI improvement	T: linear regression	D; evaluated on development data	97	11	R2	
Rundell, 2020[20]	Laminectomy with fusion for LDH, spondylolisthesis, and stenosis[a]	Continuous: 1-year ODI	T: logistic regression	DE; bootstrapping	1,918	4	R2, overfitting-corrected c-index	https://statcomp2.app.vumc.org/lumbar12mby3m/
Staub, 2020[18]	Decompression surgery for LDH	Continuous: 1-year COMI	T: linear regression	DE; temporal validation	1,608 (total), 1,244 (development), 364 patients (testing)	15	R2, MAE, mean bias, RSME	https://linkup.kws.ch/prognostic
Siccoli, 2019[48]	Surgical decompression for LSS	Categorical: 6-week ODI improvement ≥ 30%	ML: best performing (unspecified) model of several tested[c]	DE; holdout test set (30%)	173 (total), 121 (development), 52 (testing)	15	AUC, accuracy	
Merali, 2019[49]	Surgical treatment of CSM	Categorical: 1-year mJOA improvement ≥ 2 pts	ML: RF[c]	DE; holdout test set (30%)	583 (total), 408 (development), 175 (testing)	108	AUC, accuracy sensitivity	
Staartjes, 2019[50]	Single-level tubular microdiscectomy for LDH	Categorical: 1-year ODI improvement ≥ 30%	ML: deep neural network[c]	DE; holdout test set (20%)	422 (total), 338 (development), 68 (testing)	20	AUC, accuracy	
De la Garza Ramos, 2019[51]	Surgery for moderate-to-severe cervical myelopathy	Categorical: 2-year mJOA ≥ 18 pts	T: logistic regression	D; evaluated on development data	251	12	AUC	
Rubery, 2019[52]	Lumbar discectomy[a]	Categorical: PROMIS PF improvement ≥ 3.75 pts, at least 40 days postoperatively	T: logistic regression	D; evaluated on development data	78	12	AUC	
Debnath, 2018[21]	Surgical repair of lumbar pars defect in young sporting individuals	Continuous: ODI improvement, at least 2 years postoperatively	T: logistic regression[c]	D; evaluated on development data	52	6	R2	
Nouri, 2015[53]	Decompression surgery for DCM[a]	Categorical: 1-year mJOA ≥ 16 pts	T: logistic regression	D; evaluated on development data	99	6	AUC	

[a]Study included models for different patient populations; [b]Study included models for different outcomes (e.g., different PROMs); [c]Study included models of different types within the same category (ML, T); [d]Study included both ML and T models. To avoid overrepresentation of studies with multiple models in quantitative analyses, individual models were selected from those studies reporting multiple models (whether for multiple patient populations/subsets, multiple outcome definitions, model types) with the following method: the eligible model (predicting a questionnaire-based PROMs outcome) was selected for analysis that was presented by the authors as the best performing model, or if no individual model was designated in such a manner, the model exhibiting greatest performance in discrimination was selected for categorical outcome prediction models or the one explaining the greatest proportion of variation in continuous outcome prediction models. *Deprecated link at the time of publication. ODI: Oswestry disability index; ML: machine

learning; XGB: extreme gradient boosting; DE: development and evaluation; AUC: area under the curve; PPV: positive predictive value; NPV: negative predictive value; T: traditional model; E: evaluation; MARS: multivariate adaptive regression spline; LDH: lumbar disc herniation; LSS: lumbar spinal stenosis; COMI: core outcomes measures index; OPLL: ossification of the posterior longitudinal ligament; mJOA: modified Japanese Orthopaedic Association; D: development; LSF: lumbar spinal fusion; CSM: cervical spondylotic myelopathy; SVM: support vector machine; GBTM: group-based trajectory modeling; MAE: mean absolute error; GLM: generalized linear models; LIF: lumbar interbody fusion; EQ-5D: RF: random forest; DCM: degenerative cervical myelopathy; ALIF: anterior lumbar interbody fusion; DDD: degenerative disc disease; PROMIS-PF: Patient-reported outcomes measurement information system-physical function; PROMs: patient-reported outcome measures.

machine learning-based and 19 traditional prediction models were represented among the selected models from the included studies. Machine learning model types represented included neural network, random forest, decision tree, support vector machine, extreme gradient boosting, and multivariate adaptive regression splines, elastic net-regularized general linear model, and elastic-net-penalized logistic regression. Traditional models included multivariable logistic and linear regression. Of these models, six predicted a continuous outcome, and 29 predicted a binary categorical outcome.

Various PROMs were represented in the outcomes defined among selected models. Three models included the core outcome measures index (COMI), one included the EuroQol-5 dimensions (EQ-5D), 10 included the mJOA (or lumbar spine Japanese Orthopaedic Association) score, 20 included the ODI, and two included PROMIS-physical function. Of note, one model incorporated both ODI and COMI into the same predicted outcome[14]. There was considerable variation in the manner of outcome categorization, with studies choosing, for example, different thresholds for postoperative improvement or absolute score attainment, and different time points of postoperative assessment (including specified time points as well as baseline-to-last available follow-up). One study derived a dichotomized recovery trajectory outcome developed with nonlinear group-based trajectory modeling[15]. Models were developed for a range of variously defined patient populations. Nine models were developed for patients with cervical spine pathology, and 26 models were developed for patients with lumbar spine pathology. All models included patients undergoing surgery - however, two models also included conservatively treated patients in their defined patient populations[16,17].

Predictive performance

Binary categorical outcome prediction models included AUC as a measure of discrimination. Beyond this, considerable variation in the measures of predictive performance was reported, both for categorical and continuous outcome prediction models. Metrics reported included sensitivity, specificity, positive predictive value, negative predictive value, accuracy, Brier score, R^2, and mean absolute error (MAE), among others.

Categorical outcome prediction models reported median [IQR] AUC values of 0.79 [0.73, 0.84]. Of these 29 models, machine learning models reported median [IQR] AUC values of 0.825 [0.76, 0.84], while traditional models reported median [IQR] AUC values of 0.74 [0.71, 0.81]. Note that these distributions include values derived from external validation, internal validation, and performance on training data - for studies reporting multiple such scores, one representative score was selected in the following order of preference: external validation, internal validation, and training data. These distributions are shown by model category alongside the individual represented scores in Figure 2.

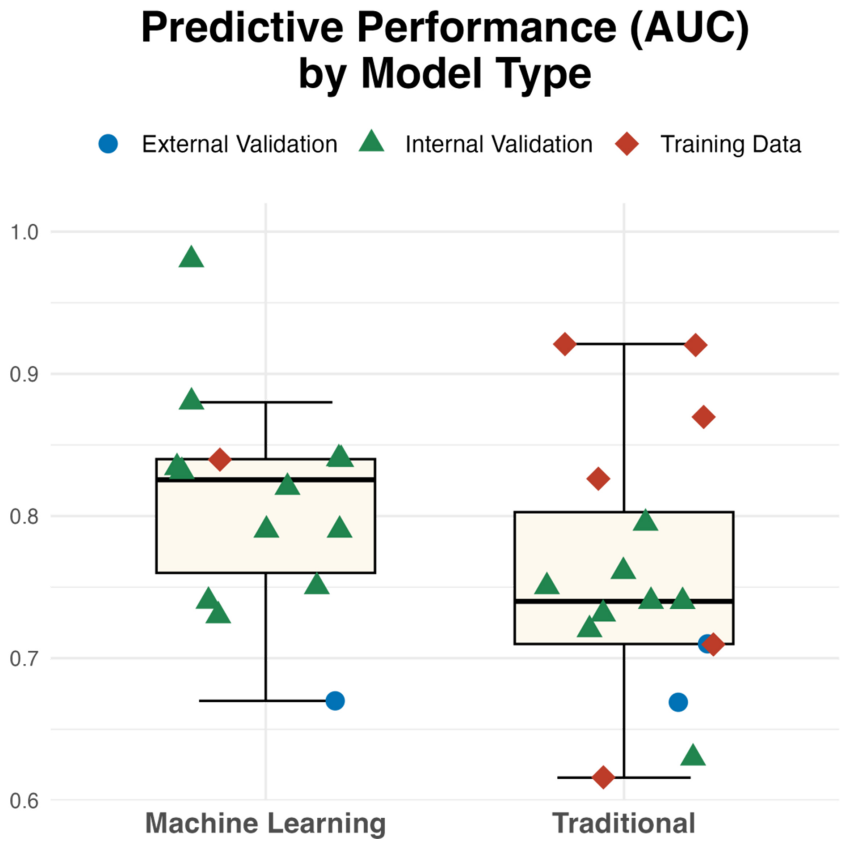

Figure 2. Predictive performance (AUC) of machine learning *vs.* traditional models. Boxplots show the distribution of discrimination scores (AUC) for categorical outcome prediction models, grouped by model type. Plots show median, IQR, and whiskers extending to smallest and largest values within 1.5 × IQR from the quartiles, overlaid with scores from each selected model from included studies. For studies reporting multiple scores for a predictive model, one representative score was selected for the model in the following order of preference by the method of evaluation: external validation, internal validation, and lastly, training data. Machine learning models (*N* = 14): AUC median [IQR]: 0.825 [0.76, 0.84]. Traditional models (*N* = 15): AUC median [IQR]: 0.74 [0.71, 0.81]. Total (*N* = 29): AUC median [IQR]: 0.79 [0.730, 0.84]. AUC: The area under the receiver operating curve, IQR: interquartile range.

The six continuous outcome prediction models included did not show consistent performance metric reporting. Three models reported MAE - 5.82[16], 8.68[17], 2.04[18], and four models reported R2 - 0.32[19], 0.247[20], 0.17[18], 0.809[21].

Methods of evaluation/validation

Of the 26 studies reporting model evaluation for selected models, three studies reported external validation, and 23 reported internal validation. Methods of external validation included validation with an external dataset, as well as validation with holdout test data from distinct centers from those used in training. Methods of internal validation included cross-validation (k-fold or leave-one-out), bootstrapping with resampling, temporal validation, or holdout test data from the same data sources as used in training. In nine studies, performance was only reported on development data.

TRIPOD-AI adherence

TRIPOD-AI item-level adherence for each item, along with a description of each item and notes on the manner of adherence grading, is shown in Table 2. Ten items showed 100% "Yes" grading (3a, 3b, 5a, 6c, 8a, 9b, 12e, 20b). When "Not applicable" was removed, nine additional items showed 100% "Yes" grading (8c, 9c, 12a, 12b, 12d, 12f, 23b), although six of these items were defined in a way that studies were likely to be

Table 2. TRIPOD-AI item-level adherence

Section/topic	Item	Development/ evaluation	Description	Notes on adherence assessment	Adherence
Title and abstract					
Title	1	D; E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted		Y: 21 (60.0%), N: 14 (40.0%), NA: 0 (0.0%)
Abstract	2	D; E	See TRIPOD + AI for abstracts checklist		Y: 15 (42.9%), N: 20 (57.1%), NA: 0 (0.0%)
Introduction					
Background	3a	D; E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models		Y: 35 (100.0%), N: 0 (0.0%), NA: 0 (0.0%)
	3b	D; E	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)		Y: 35 (100.0%), N: 0 (0.0%), NA: 0 (0.0%)
	3c	D; E	Describe any known health inequalities between sociodemographic groups		Y: 0 (0.0%), N: 35 (100.0%), NA: 0 (0.0%)
Objectives	4	D; E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)		Y: 34 (97.1%), N: 1 (2.9%), NA: 0 (0.0%)
Methods					
Data	5a	D; E	Describe the sources of data separately for the development and evaluation datasets (e.g., randomized trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data		Y: 35 (100.0%), N: 0 (0.0%), NA: 0 (0.0%)
	5b	D; E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up		Y: 30 (85.7%), N: 5 (14.3%), NA: 0 (0.0%)
Participants	6a	D; E	Specify key elements of the study setting (e.g., primary care, secondary care, general population), including the number and location of centers		Y: 34 (97.1%), N: 1 (2.9%), NA: 0 (0.0%)
	6b	D; E	Describe the eligibility criteria for study participants		Y: 34 (97.1%), N: 1 (2.9%), NA: 0 (0.0%)
	6c	D; E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant		Y: 35 (100.0%), N: 0 (0.0%), NA: 0 (0.0%)
Data preparation	7	D; E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups		Y: 30 (85.7%), N: 5 (14.3%), NA: 0 (0.0%)

Outcome	8a	D; E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups		Y: 35 (100.0%), N: 0 (0.0%), NA: 0 (0.0%)
	8b	D; E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	NA if no subjective assessment of outcome described	Y: 0 (0.0%), N: 0 (0.0%), NA: 35 (100.0%)
	8c	D; E	Report any actions to blind assessment of the outcome to be predicted	NA if no outcome blinding described	Y: 5 (14.3%), N: 0 (0.0%), NA: 30 (85.7%)
Predictors	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building		Y: 31 (88.6%), N: 2 (5.7%), NA: 2 (5.7%)
	9b	D; E	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)		Y: 35 (100.0%), N: 0 (0.0%), NA: 0 (0.0%)
	9c	D; E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	NA if no subjective assessment of predictors described	Y: 2 (5.7%), N: 0 (0.0%), NA: 33 (94.3%)
Sample size	10	D; E	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation		Y: 5 (14.3%), N: 30 (85.7%), NA: 0 (0.0%)
Missing data	11	D; E	Describe how missing data were handled. Provide reasons for omitting any data		Y: 21 (60.0%), N: 14 (40.0%), NA: 0 (0.0%)
Analytical methods	12a	D	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements		Y: 33 (94.3%), N: 0 (0.0%), NA: 2 (5.7%)
	12b	D	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardization)		Y: 33 (94.3%), N: 0 (0.0%), NA: 2 (5.7%)
	12c	D	Specify the type of model, rationale2, all model-building steps, including any hyperparameter tuning, and method for internal validation		Y: 25 (71.4%), N: 8 (22.9%), NA: 2 (5.7%)
	12d	D; E	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations3	NA if no heterogeneity in estimates described	Y: 3 (8.6%), N: 0 (0.0%), NA: 32 (91.4%)
	12e	D; E	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models		Y: 35 (100.0%), N: 0 (0.0%), NA: 0 (0.0%)
	12f	E	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	NA if no model updating described	Y: 1 (2.9%), N: 0 (0.0%), NA: 34 (97.1%)
	12g	E	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object,		Y: 26 (74.3%), N: 1

			application programming interface)		(2.9%), NA: 8 (22.9%)
Class imbalance	13	D; E	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	NA if no class imbalance measures described	Y: 7 (20.0%), N: 0 (0.0%), NA: 28 (80.0%)
Fairness	14	D; E	Describe any approaches that were used to address model fairness and their rationale		Y: 0 (0.0%), N: 35 (100.0%), NA: 0 (0.0%)
Model output	15	D	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified		Y: 33 (94.3%), N: 0 (0.0%), NA: 2 (5.7%)
Training *vs.* evaluation	16	D; E	Identify any differences between the development and evaluation data in healthcare settings, eligibility criteria, outcome, and predictors	NA if evaluated through bootstrapping, leave-one-out cross-validation, development data	Y: 11 (31.4%), N: 11 (31.4%), NA: 13 (37.1%)
Ethical approval	17	D; E	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent		Y: 26 (74.3%), N: 9 (25.7%), NA: 0 (0.0%)
Open science					
Funding	18a	D; E	Give the source of funding and the role of the funders for the present study		Y: 31 (88.6%), N: 4 (11.4%), NA: 0 (0.0%)
Conflicts of interest	18b	D; E	Declare any conflicts of interest and financial disclosures for all authors		Y: 34 (97.1%), N: 1 (2.9%), NA: 0 (0.0%)
Protocol	18c	D; E	Indicate where the study protocol can be accessed or state that a protocol was not prepared		Y: 0 (0.0%), N: 35 (100.0%), NA: 0 (0.0%)
Registration	18d	D; E	Provide registration information for the study, including register name and registration number, or state that the study was not registered		Y: 0 (0.0%), N: 35 (100.0%), NA: 0 (0.0%)
Data sharing	18e	D; E	Provide details of the availability of the study data		Y: 7 (20.0%), N: 28 (80.0%), NA: 0 (0.0%)
Code sharing	18f	D; E	Provide details of the availability of the analytical code		Y: 3 (8.6%), N: 32 (91.4%), NA: 0 (0.0%)
Patient and public involvement					
Patient and public involvement	19	D; E	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement		Y: 1 (2.9%), N: 34 (97.1%), NA: 0 (0.0%)
Results					

Participants	20a	D; E	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful		Y: 34 (97.1%), N: 1 (2.9%), NA: 0 (0.0%)
	20b	D; E	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups		Y: 35 (100.0%), N: 0 (0.0%), NA: 0 (0.0%)
	20c	E	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome)	NA if evaluated through bootstrapping, leave-one-out cross-validation, development data	Y: 11 (31.4%), N: 11 (31.4%), NA: 13 (37.1%)
Model development	21	D; E	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)		Y: 32 (91.4%), N: 3 (8.6%), NA: 0 (0.0%)
Model specification	22	D	Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or re-use (e.g., freely available, proprietary)	Y if any included online calculators allowing for the use of the model	Y: 15 (42.9%), N: 18 (51.4%), NA: 2 (5.7%)
Model performance	23a	D; E	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation		Y: 29 (82.9%), N: 6 (17.1%), NA: 0 (0.0%)
	23b	D; E	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD Cluster for additional details	NA if no heterogeneity in model performance described	Y: 2 (5.7%), N: 0 (0.0%), NA: 33 (94.3%)
Model updating	24	E	Report the results from any model updating, including the updated model and subsequent performance	NA if no model updating described	Y: 0 (0.0%), N: 0 (0.0%), NA: 35 (100.0%)
Discussion					
Interpretation	25	D; E	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies		Y: 35 (100.0%), N: 0 (0.0%), NA: 0 (0.0%)
Limitations	26	D; E	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability		Y: 34 (97.1%), N: 1 (2.9%), NA: 0 (0.0%)
Usability of the model in the context of current care	27a	D	Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model		Y: 1 (2.9%), N: 32 (91.4%), NA: 2 (5.7%)
	27b	D	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users		Y: 2 (5.7%), N: 31 (88.6%), NA: 2 (5.7%)
	27c	D; E	Discuss any next steps for future research, with a specific view to the applicability and generalizability of the model		Y: 35 (100.0%), N: 0 (0.0%), NA: 0 (0.0%)

The complete checklist may be found at https://www.tripod-statement.org/wp-content/uploads/2019/12/TRIPODAI_checklist.pdf. TRIPOD-AI: Transparent reporting of a multivariable prediction model for

individual prognosis or diagnosis[12]; D: development, E: evaluation; Y: yes, N: no, NA: not applicable.

graded "Yes" for mentioning an applicable aspect of study methodology or otherwise "NA" - these adherence grading approaches are also detailed in Table 2. Two items showed 100% "Not applicable" grading (8b, 24), and four items showed 100% "No" grading (3c, 14, 18c, 18d).

For each study, an adherence score was calculated, shown in Supplementary Table 1, with a median [IQR] adherence score of 69% [64%, 73.5%]. Item-level TRIPOD-AI adherence is also visualized in bar graph form in Supplementary Figure 1. Item-level adherence is compared between studies with selected machine learning models and those with traditional models in Figure 3. Item-level adherence is compared between studies mentioning TRIPOD 2015 or TRIPOD-AI 2024 guideline adherence in Supplementary Figure 2.

Risk of bias

Fourteen of 35 studies did not report details of missing data handling. Ten of the 33 studies reporting development with or without internal/external validation showed a ratio of development sample size to number of predictors/levels below 10. Calibration was reported in 18 of 35 studies, and decision curve analysis in three studies. Twenty-two of 35 studies failed at least one of these criteria for risk of bias.

Availability of prediction models for application

Ten studies provided online calculators for models developed and/or assessed. In addition to these studies, seven studies provided sufficient detail (e.g., formula/equation) of prediction models for outside use or reported availability of models for use.

DISCUSSION

We identified 35 studies reporting the development or evaluation of prediction models for questionnaire-based PROMs outcomes following elective spine surgery. This study reviews a growing body of literature reporting on prediction models, which continue to take on increasing importance, given implications for uses in reimbursement, physician evaluation, and patient-level decision making. In recent years, these predictive models have also more frequently taken advantage of machine learning methods, although traditional approaches continue to be used. We characterize patterns in the defined prediction tasks (patient population, outcome definition), models utilized, transparency of reporting, and risk of bias among these studies. Our study highlights a considerable degree of variability, not only in predictive performance but also in the completeness of prediction model reporting. These prediction models have been developed for a range of prediction tasks across variously defined patient populations and predicted outcomes, limiting the extent to which model performance may be compared head-to-head for a given task.

In addition to variation in the methodology of model design, the range of performance metrics available for these studies was variable beyond the reporting of discrimination in categorical outcome prediction models, with many studies missing calibration and/or decision curve analysis, and variable performance

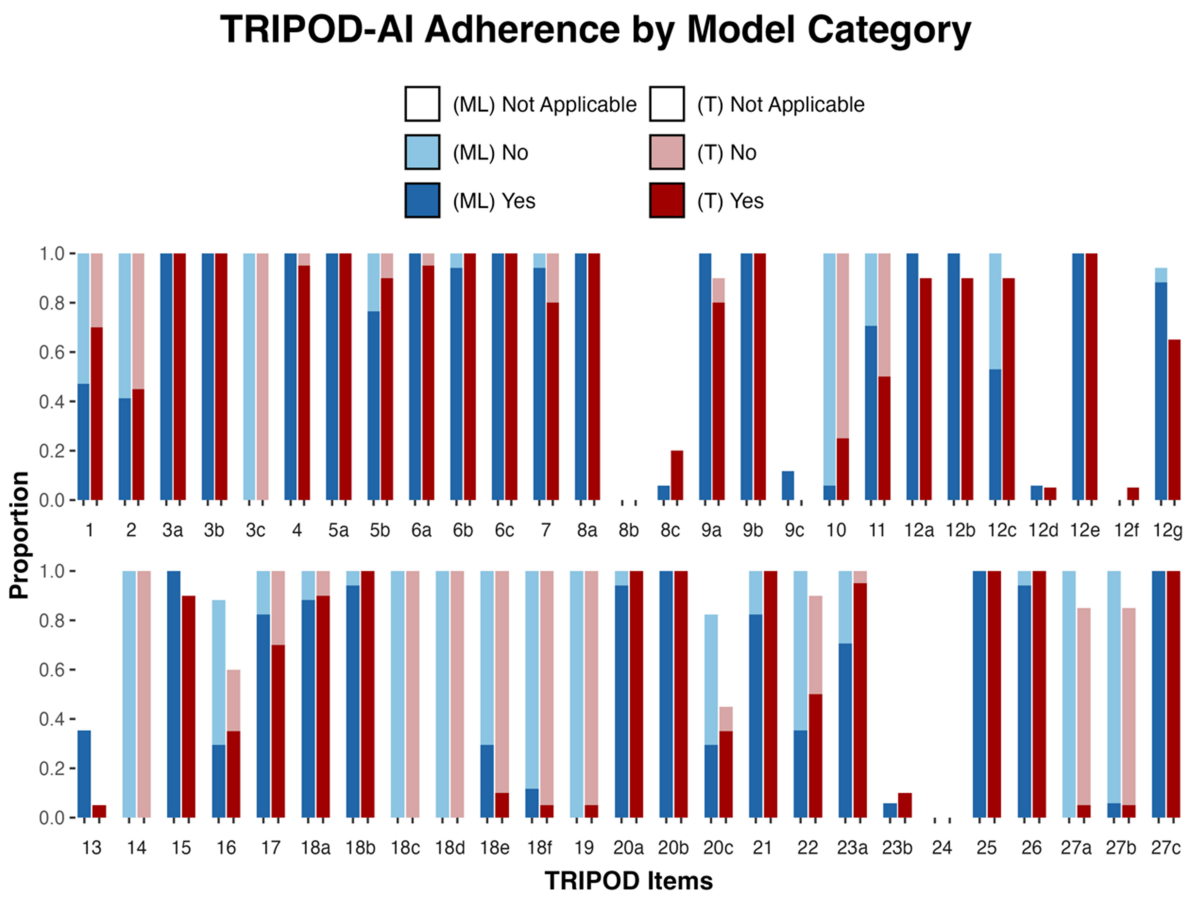

Figure 3. TRIPOD-AI adherence: machine learning *vs.* traditional models. ML: machine learning prediction model (*N* = 16); T: traditional statistical prediction model (*N* = 19); TRIPOD-AI: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis.

measures reported across continuous outcome prediction models. Among categorical outcome prediction models reporting AUC, we found the group of machine learning models to exhibit higher median (IQR) AUC values than the group of models developed with traditional methods. This comparison, however, should be understood in the context of each model varying in its prediction tasks (patient population, predicted outcome definition, and time point), sample size, number of predictors, and the setting of performance evaluation (external validation, internal validation, evaluation on training data). In addition, discrimination alone does not suffice to fully characterize the predictive performance of a model, and the meaning and utility of an individual score must be understood in the context of the assigned predictive task and the model applications under consideration. Nevertheless, our findings appear to indicate that both machine learning and traditional approaches can produce useful models, with considerable overlap between model types in the distribution of reported discriminative capacity, and AUC scores across both categories in ranges that have been characterized as suboptimal to outstanding in clinical applications[22].

Reported model performance should also be contextualized in terms of the completeness of reporting and the potential risk of bias. Particularly where models have yet to be externally or internally validated, predictive models carry the risk of overestimating performance through a number of potential pitfalls, including overfitting. This risk is heightened when the available sample size for the development of a model is low in proportion to the number of predictors incorporated into a given model - machine learning

models, in particular, carry the potential for overfitting through greater flexibility in modeling of predictor and outcome relationships that may represent patterns seen only in the training dataset and which may poorly generalize to other datasets. These concerns are illustrated by the variability we observed in predictive performance between models evaluated with training data, internal validation, and external validation. Although external validation was only reported for a few models, we noted a pattern of worse model performance in the examples of external validation with both machine learning and traditional methods, with scores for model discrimination falling at or below the 25th percentile for respective model types. Transparency of reporting in the development and evaluation of prediction models is also critical to maintaining interpretability, which may already be compromised in machine learning models due to the opacity of model decision-making processes[23]. This underscores the necessity for model transparency - ensuring that models are both powerful and interpretable enhances their reliability and applicability in clinical settings.

TRIPOD + AI is a recently updated set of guidelines for prediction model reporting[12]. This study is among the first to evaluate the transparency of predictive model reporting with these updated guidelines. While past studies may have relied on the 2015 TRIPOD guidelines available at the time, the newer guidelines are intended to serve as a more comprehensive tool. This review may provide insight into current model performance and reporting practices, as well as highlighting the updated guidelines to promote greater consistency in reporting at the study design and writing stage, where we believe it may yield the greatest utility. Taking these guidelines into consideration in the early stages of study design allows for thoughtful incorporation of measures to address problems of missing data, consider analyses of heterogeneity, and decide on an appropriate method of assessing model performance, whether that may include external validation or otherwise.

While the TRIPOD-AI guidelines include a checklist of items intended for use by authors in the development/evaluation and reporting of prediction models, the aims of the guidelines also include their use assessment of transparency and consistency of prediction model reporting[12]. We found four items (3c, 14, 18c, 18d) to be missing (graded "No") in every included study - these items encompassed description of known healthcare inequalities in the sample population, comments on model fairness, and disclosure of study protocols or registrations. Twenty percent of reported models included comments on the availability of the analytical code for the model (item 18f) and only one study included descriptions of patient and public involvement in the model development or testing process (item 19). Comparing machine learning models to traditional models, we found that sample size justification (item 10) and statements on data availability (item 18f) were more often missing among machine learning models than among traditional models. Overall, the adherence to the TRIPOD-AI guidelines was inconsistent and incomplete across studies, including among studies that had mentioned following either the prior TRIPOD 2015 or more recent TRIPOD-AI guidelines.

Nevertheless, TRIPOD items are not interchangeable, and the implications for incomplete reporting may vary considerably between items. For instance, TRIPOD item 11 involves the description of missing data handling - a complete case analysis may meet or fail completeness of this item not based on measures taken to address missing data, but based on whether such measures or the absence thereof were described appropriately. As a tool designed around transparency in reporting rather than on robustness of study design, this focus on the adequacy of reporting may be appropriate, but for this reason, it is important to consider the definitions of these items when interpreting item-level adherence, which may not capture the full scope of a given problem. In the case of missing data, completeness of PROMs data presents a particular challenge, with unavailable data for many patients and the associated risk of excluding patients without

complete data, resulting in the underrepresentation of these patients in predictive model development and evaluation[1].

Some TRIPOD items were also found in this review to be difficult to assess from the perspective of a reader or reviewer, as they were conditional on knowledge that may only have been available to authors during the study design or analysis phase. In addition, many TRIPOD items include several components (e.g., item 2 for abstracts), and some of these items may have been better suited for consultation during the writing of a publication rather than as tools for post-hoc evaluation of a publication, particularly given that the most recent guidelines were not available to most authors of the studies included in this review. While we assessed an adherence score for the studies reviewed for the purposes of aggregate analysis, we found the utility of the TRIPOD-AI tool in the assessment of individual studies to lie primarily in identifying particular gaps in reporting to contextualize findings and enable further evaluation, rather than in the reduction of reporting in a study to a single score.

One of the paramount challenges for comparative assessment of prediction model performance lies in the variability of outcome reporting, which impedes the ability to compare aggregated data and appropriately assess the clinical utility of one model over another. Even for similarly defined populations of surgical patients, the use of a variety of PROMs exacerbates this challenge. Where the same PROMs are used, studies may vary in their incorporation into outcome definitions - for instance, studies may define treatment success by different time points or thresholds for minimal clinically important differences (MCIDs). One study might define MCID as achieving a 22-point improvement in ODI scores, whereas another might set a threshold of 12.7 points[24,25]. These variations underscore an opportunity for discussion on the standardization of such PROM-based definitions of treatment success. While such standardization would face considerable logistical obstacles, it would enhance interpretability and enable more effective comparative assessments, not only within the prediction model literature but also across the broader body of literature on surgical outcomes.

Several included studies had not yet reported validation beyond performance on training data, and we found a paucity of reported external validation of models, mirroring trends observed in other systematic reviews[26]. External validation, in particular, will remain critical for assessing the generalizability of models to different countries, healthcare systems, and institutional methods of predictor measurement[27]. Methods of incorporating external validation may include applying existing models to new datasets - alternatively, external validation may be integrated alongside the development of new models to be reported together. In both cases, collaboration between institutions, as well as utilization of publicly available datasets where appropriate, will help to promote more widespread external validation.

Limitations

This review has several limitations. Firstly, our broad inclusion criteria encompassed a broad range of models - varying in model type, patient population, and predicted outcome, among many other characteristics. This heterogeneity limits the appropriate comparisons that can be made between models. In addition, we chose to incorporate only the best-performing eligible model from each included study for our analysis - while this approach avoids the overrepresentation of individual datasets or studies in quantitative comparisons, it also reduces the breadth of models represented. This study was also limited in scope to those analyses that included prediction models of questionnaire-based outcomes, with the understanding that prediction models may improve the utility of PROMs. This focus, however, excludes studies reporting prediction models for other types of outcomes, such as visual analog scales for pain, or quantified rates of adverse events. In addition, our review excludes studies that assess machine learning model performance for

non-prediction tasks - for example, a recently published model that assesses vertebral body fractures on CT imaging[28]. Studies meeting inclusion criteria may have been missed in our search strategy, although we believe the use of separate queries of four databases may have mitigated some of this risk. Studies included in our review may also have been subject to publication bias, with more poorly performing models less likely than better-performing models to be published. This may have resulted in overestimates of performance in the distribution of performance metrics across our pool of included studies. This study may best be regarded as a review of the current landscape of prediction models and their reporting, rather than as definitive aggregate evidence in favor of selecting one model type or another for a given prediction task.

In conclusion, this study highlights the current state of predictive modeling for PROM-based outcomes of elective spine surgery, with a focus on the variation between machine learning and traditional regression models. We noted considerable variation in patterns of model development, not only in chosen patient populations and outcome measures but also in their manner of evaluation. We found 22/35 studies lacking in at least one domain in risk of bias assessment (missing data handling, sample size-predictor ratio, calibration), and noted median [IQR] adherence to TRIPAD-AI guideline components of 69% [64%, 73.5%]. Future studies stand to benefit from referring to guidelines such as these in order to improve transparency and completeness of model reporting. There remains a considerable opportunity for improved consistency in reporting as well as in possible standardization of outcome definitions to enable meaningful model comparisons. Future efforts should also be directed toward external validation of existing models to enable better characterization of model generalizability. Such developments may facilitate the application of this growing body of literature toward effective clinical uses.

DECLARATIONS

Authors' contributions

Made substantial contributions to the conception and design of the study and performed data analysis and interpretation: Lemel H, Shin D, Meade S

Provided supervision, manuscript revision: Habboub G, Lapin B, Mroz T, Steinmetz M

Availability of data and materials

Not Applicable.

Financial support and sponsorship

None.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Falavigna A, Dozza DC, Teles AR, et al. Current status of worldwide use of patient-reported outcome measures (PROMs) in spine care. *World Neurosurg.* 2017;108:328-35. DOI
2. Finkelstein JA, Schwartz CE. Patient-reported outcomes in spine surgery: past, current, and future directions. *J Neurosurg Spine.* 2019;31:155-64. DOI PubMed
3. Lapin B, Li Y, Davin S, et al. Comparison of stratification techniques for optimal management of patients with chronic low back pain in spine clinics. *Spine J.* 2023;23:1334-44. DOI
4. Lee TJ, Thomas AA, Grandhi NR, et al. Cost-effectiveness applications of patient-reported outcome measures (PROMs) in spine surgery. *Clin Spine Surg.* 2020;33:140-5. DOI
5. Pronk Y, Pilot P, Brinkman JM, van Heerwaarden RJ, van der Weegen W. Response rate and costs for automated patient-reported outcomes collection alone compared to combined automated and manual collection. *J Patient Rep Outcomes.* 2019;3:31. DOI PubMed PMC
6. Beighley A, Zhang A, Huang B, et al. Patient-reported outcome measures in spine surgery: a systematic review. *J Craniovertebr Junction Spine.* 2022;13:378-89. DOI PubMed PMC
7. Guzman JZ, Cutler HS, Connolly J, et al. Patient-reported outcome instruments in spine surgery. *Spine.* 2016;41:429-37. DOI
8. Ravishankar P, Winkleman R, Rabah N, Steinmetz M, Mroz T. Analysis of patient-reported outcomes measures used in lumbar fusion surgery research for degenerative spondylolisthesis. *Clin Spine Surg.* 2022;35:287-94. DOI PubMed
9. Teles AR, Khoshhal KI, Falavigna A. Why and how should we measure outcomes in spine surgery? *J Taibah Univ Med Sci.* 2016;11:91-7. DOI
10. Wellington IJ, Karsmarski OP, Murphy KV, Shuman ME, Ng MK, Antonacci CL. The use of machine learning for predicting candidates for outpatient spine surgery: a review. *J Spine Surg.* 2023;9:323-30. DOI PubMed PMC
11. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* 2021;372:n71. DOI PubMed PMC
12. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. DOI PubMed PMC
13. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162:W1-73. DOI PubMed
14. Staartjes VE, Stumpo V, Ricciardi L, et al. FUSE-ML: development and external validation of a clinical prediction model for mid-term outcomes after lumbar spinal fusion for degenerative disease. *Eur Spine J.* 2022;31:2629-38. DOI
15. Jaja BNR, Witiw CD, Harrington EM, et al. Analysis of recovery trajectories in degenerative cervical myelopathy to facilitate improved patient counseling and individualized treatment recommendations. *J Neurosurg Spine.* 2023;38:644-52. DOI
16. Sundaramoorthy K, Al Ansari MS, Koteswari S, Kumari M. Artificial intelligence and machine learning-driven decision-making in spinal disease treatment. *J Theor Appl Inf Technol.* 2023;101:8388-406. http://www.jatit.org/volumes/Vol101No24/38Vol101No24.pdf . (accessed 2025-02-13)
17. Wirries A, Geiger F, Hammad A, et al. AI prediction of neuropathic pain after lumbar disc herniation-machine learning reveals influencing factors. *Biomedicines.* 2022;10:1319. DOI PubMed PMC
18. Staub LP, Aghayev E, Skrivankova V, Lord SJ, Haschtmann D, Mannion AF. Development and temporal validation of a prognostic model for 1-year clinical outcome after decompression surgery for lumbar disc herniation. *Eur Spine J.* 2020;29:1742-51. DOI PubMed
19. Ford JJ, Kaddour O, Page P, Richards MC, McMeeken JM, Hahne AJ. A multivariate prognostic model for pain and activity limitation in people undergoing lumbar discectomy. *Br J Neurosurg.* 2020;34:381-7. DOI PubMed
20. Rundell SD, Pennings JS, Nian H, et al. Adding 3-month patient data improves prognostic models of 12-month disability, pain, and satisfaction after specific lumbar spine surgical procedures: development and validation of a prediction model. *Spine J.* 2020;20:600-13. DOI
21. Debnath UK, Scammell BE, Freeman BJC, McConnell JR. Predictive factors for the outcome of surgical treatment of lumbar spondylolysis in young sporting individuals. *Global Spine J.* 2018;8:121-8. DOI PubMed PMC
22. Çorbacıoğlu ŞK, Aksel G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: a guide to interpreting the area under the curve value. *Turk J Emerg Med.* 2023;23:195-8. DOI PubMed PMC
23. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Min Knowl Discov.* 2020;10:e1379. DOI
24. Berg B, Gorosito MA, Fjeld O, et al. Machine learning models for predicting disability and pain following lumbar disc herniation surgery. *JAMA Netw Open.* 2024;7:e2355024. DOI PubMed PMC
25. Berjano P, Langella F, Ventriglia L, et al. The influence of baseline clinical status and surgical strategy on early good to excellent result in spinal lumbar arthrodesis: a machine learning approach. *J Pers Med.* 2021;11:1377. DOI
26. Ogink PT, Groot OQ, Karhade AV, et al. Wide range of applications for machine-learning prediction models in orthopedic surgical outcome: a systematic review. *Acta Orthop.* 2021;92:526-31. DOI PubMed PMC
27. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol.* 2014;14:40. DOI PubMed PMC
28. Polzer C, Yilmaz E, Meyer C, et al. AI-based automated detection and stability analysis of traumatic vertebral body fractures on

computed tomography. *Eur J Radiol.* 2024;173:111364. DOI

29. Carreon LY, Nian H, Archer KR, Andersen MØ, Hansen KH, Glassman SD. Performance of the streamlined quality outcomes database web-based calculator: internal and external validation. *Spine J.* 2024;24:662-9. DOI PubMed
30. Pedersen CF, Andersen MØ, Carreon LY, Skov ST, Doering P, Eiskjær S. PROPOSE. Development and validation of a prediction model for shared decision making for patients with lumbar spinal stenosis. *N Am Spine Soc J.* 2024;17:100309. DOI PubMed PMC
31. Halicka M, Wilby M, Duarte R, Brown C. Predicting patient-reported outcomes following lumbar spine surgery: development and external validation of multivariable prediction models. *BMC Musculoskelet Disord.* 2023;24:333. DOI PubMed PMC
32. Matsukura Y, Egawa S, Inose H, et al. Preoperative symptom duration influences neurological recovery and patient-reported outcome measures after surgical treatment of cervical ossification of the posterior longitudinal ligament. *Spine.* 2023;48:1259-65. DOI
33. Rushton AB, Jadhakhan F, Verra ML, et al. Predictors of poor outcome following lumbar spinal fusion surgery: a prospective observational study to derive two clinical prediction rules using British Spine Registry data. *Eur Spine J.* 2023;32:2303-18. DOI
34. Geere JH, Hunter PR, Swamy GN, Cook AJ, Rai AS. Development and temporal validation of clinical prediction models for 1-year disability and pain after lumbar decompressive surgery. The Norwich Lumbar Surgery Predictor (development version). *Eur Spine J.* 2023;32:4210-9. DOI PubMed
35. Zhang JK, Jayasekera D, Javeed S, et al. Diffusion basis spectrum imaging predicts long-term clinical outcomes following surgery in cervical spondylotic myelopathy. *Spine J.* 2023;23:504-12. DOI PubMed PMC
36. Chen X, Lin F, Xu X, Chen C, Wang R. Development, validation, and visualization of a web-based nomogram to predict the effect of tubular microdiscectomy for lumbar disc herniation. *Front Surg.* 2023;10:1024302. DOI PubMed PMC
37. Dong S, Zhu Y, Yang H, et al. Evaluation of the predictors for unfavorable clinical outcomes of degenerative lumbar spondylolisthesis after lumbar interbody fusion using machine learning. *Front Public Health.* 2022;10:835938. DOI PubMed PMC
38. Pedersen CF, Andersen MØ, Carreon LY, Eiskjær S. Applied machine learning for spine surgeons: predicting outcome for patients undergoing treatment for lumbar disc herniation using PRO data. *Global Spine J.* 2022;12:866-76. DOI PubMed PMC
39. Coric D, Zigler J, Derman P, Braxton E, Situ A, Patel L. Predictors of long-term clinical outcomes in adult patients after lumbar total disc replacement: development and validation of a prediction model. *J Neurosurg Spine.* 2022;36:399-407. DOI
40. Purohit G, Choudhary M, Sinha VD. Use of artificial intelligence for the development of predictive model to help in decision-making for patients with degenerative lumbar spine disease. *Asian J Neurosurg.* 2022;17:274-9. DOI PubMed PMC
41. Khan O, Badhiwala JH, Akbar MA, Fehlings MG. Prediction of worse functional status after surgery for degenerative cervical myelopathy: a machine learning approach. *Neurosurgery.* 2021;88:584-91. DOI PubMed
42. Budiono GR, McCaffrey MH, Parr WCH, et al. Development of a multivariate prediction model for successful oswestry disability index changes in L5/S1 anterior lumbar interbody fusion for degenerative disc disease. *World Neurosurg.* 2021;148:e1-9. DOI
43. Werner DAT, Grotle M, Småstuen MC, et al. A prognostic model for failure and worsening after lumbar microdiscectomy: a multicenter study from the Norwegian Registry for Spine Surgery. *Acta Neurochir.* 2021;163:2567-80. DOI PubMed PMC
44. Pilato F, Calandrelli R, Distefano M, Tamburrelli FC. Multidimensional assessment of cervical spondylotic myelopathy patients. Usefulness of a comprehensive score system. *Neurol Sci.* 2021;42:1507-14. DOI PubMed PMC
45. Karhade AV, Fogel HA, Cha TD, et al. Development of prediction models for clinically meaningful improvement in PROMIS scores after lumbar decompression. *Spine J.* 2021;21:397-404. DOI
46. Zhang MZ, Ou-Yang HQ, Jiang L, et al. Optimal machine learning methods for radiomic prediction models: clinical application for preoperative T_2^*-weighted images of cervical spondylotic myelopathy. *JOR Spine.* 2021;4:e1178. DOI PubMed PMC
47. Quddusi A, Eversdijk HAJ, Klukowska AM, et al. External validation of a prediction model for pain and functional outcome after elective lumbar spinal fusion. *Eur Spine J.* 2020;29:374-83. DOI
48. Siccoli A, de Wispelaere MP, Schröder ML, Staartjes VE. Machine learning-based preoperative predictive analytics for lumbar spinal stenosis. *Neurosurg Focus.* 2019;46:E5. DOI PubMed
49. Merali ZG, Witiw CD, Badhiwala JH, Wilson JR, Fehlings MG. Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy. *PLoS One.* 2019;14:e0215133. DOI PubMed PMC
50. Staartjes VE, de Wispelaere MP, Vandertop WP, Schröder ML. Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar discectomy: feasibility of center-specific modeling. *Spine J.* 2019;19:853-61. DOI PubMed
51. De la Garza Ramos R, Nouri A, Nakhla J, et al. Predictors of return to normal neurological function after surgery for moderate and severe degenerative cervical myelopathy: an analysis of a global AOSpine cohort of patients. *Neurosurgery.* 2019;85:E917-23. DOI
52. Rubery PT, Houck J, Mesfin A, Molinari R, Papuga MO. Preoperative patient reported outcomes measurement information system scores assist in predicting early postoperative success in lumbar discectomy. *Spine.* 2019;44:325-33. DOI PubMed
53. Nouri A, Tetreault L, Côté P, Zamorano JJ, Dalzell K, Fehlings MG. Does magnetic resonance imaging improve the predictive performance of a validated clinical prediction rule developed to evaluate surgical outcome in patients with degenerative cervical myelopathy? *Spine.* 2015;40:1092-100. DOI PubMed