European Spine Journal (2022) 31:2629–2638
https://doi.org/10.1007/s00586-022-07135-9

ORIGINAL ARTICLE

FUSE-ML: development and external validation of a clinical prediction model for mid-term outcomes after lumbar spinal fusion for degenerative disease

Victor E. Staartjes[1,2,3] · Vittorio Stumpo[1] · Luca Ricciardi[4] · Nicolai Maldaner[1] · Hubert A. J. Eversdijk[3] · Moira Vieli[1] · Olga Ciobanu-Caraus[1] · Antonino Raco[4] · Massimo Miscusi[4] · Andrea Perna[5,6] · Luca Proietti[5,6] · Giorgio Lofrese[7] · Michele Dughiero[7] · Francesco Cultrera[7] · Nicola Nicassio[7] · Seong Bae An[8] · Yoon Ha[8] · Aymeric Amelot[9,10] · Irene Alcobendas[11] · Jose M. Viñuela-Prieto[11] · Maria L. Gandía-González[11] · Pierre-Pascal Girod[12] · Sara Lener[12] · Nikolaus Kögl[12] · Anto Abramovic[12] · Nico Akhavan Safa[13] · Christoph J. Laux[13] · Mazda Farshad[13] · Dave O'Riordan[14] · Markus Loibl[15] · Anne F. Mannion[14] · Alba Scerrati[16] · Granit Molliqaj[17] · Enrico Tessitore[17] · Marc L. Schröder[3] · W. Peter Vandertop[2] · Martin N. Stienen[18] · Luca Regli[1] · Carlo Serra[1]

Received: 22 December 2021 / Accepted: 25 January 2022 / Published online: 21 February 2022

Abstract

Background Indications and outcomes in lumbar spinal fusion for degenerative disease are notoriously heterogenous. Selected subsets of patients show remarkable benefit. However, their objective identification is often difficult. Decision-making may be improved with reliable prediction of long-term outcomes for each individual patient, improving patient selection and avoiding ineffective procedures.

Communicated by SWITZERLAND.

✉ Victor E. Staartjes
victoregon.staartjes@usz.ch

1 Machine Intelligence in Clinical Neuroscience (MICN) Laboratory, Department of Neurosurgery, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Frauenklinikstrasse 10, 8091 Zurich, Switzerland

2 Amsterdam UMC, Neurosurgery, Amsterdam Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

3 Department of Neurosurgery, Bergman Clinics Amsterdam, Amsterdam, The Netherlands

4 Department of NESMOS, Azienda Ospedaliera Universitaria Sant'Andrea, Sapienza University, Rome, Italy

5 Department of Aging, Neurological, Orthopedic and Head-Neck Sciences, IRCCS A. Gemelli University Polyclinic Foundation, Rome, Italy

6 Department of Geriatrics and Orthopedics, Sacred Heart Catholic University, Rome, Italy

7 Neurosurgery Division, Department of Neurosciences, "M.Bufalini" Hospital, Cesena, Italy

8 Department of Neurosurgery, Spine and Spinal Cord Institute, Severance Hospital, College of Medicine, Yonsei University, Seoul, Korea

9 Department of Neurosurgery, La Pitié Salpétrière Hospital, Paris, France

10 Neurosurgical Spine Department, University Hospital of Tours, Tours, France

11 Department of Neurosurgery, Hospital Universitario La Paz, Madrid, Spain

12 Department of Neurosurgery, Medical University of Innsbruck, Innsbruck, Austria

13 University Spine Center, Balgrist University Hospital, University of Zurich, Zurich, Switzerland

14 Department of Teaching, Research and Development, Spine Center Division, Schulthess Klinik, Zurich, Switzerland

15 Department of Spine Surgery, Schulthess Klinik, Zurich, Switzerland

16 Department of Neurosurgery, Policlinico Universitario di Ferrara, Ferrara, Italy

17 Department of Neurosurgery, HUG Geneva University Hospital, Geneva, Switzerland

18 Department of Neurosurgery, Cantonal Hospital St. Gallen, St. Gallen, Switzerland

Methods Clinical prediction models for long-term functional impairment [Oswestry Disability Index (ODI) or Core Outcome Measures Index (COMI)], back pain, and leg pain after lumbar fusion for degenerative disease were developed. Achievement of the minimum clinically important difference at 12 months postoperatively was defined as a reduction from baseline of at least 15 points for ODI, 2.2 points for COMI, or 2 points for pain severity.
Results Models were developed and integrated into a web-app (https://neurosurgery.shinyapps.io/fuseml/) based on a multinational cohort [$N = 817$; 42.7% male; mean (SD) age: 61.19 (12.36) years]. At external validation [$N = 298$; 35.6% male; mean (SD) age: 59.73 (12.64) years], areas under the curves for functional impairment [0.67, 95% confidence interval (CI): 0.59–0.74], back pain (0.72, 95%CI: 0.64–0.79), and leg pain (0.64, 95%CI: 0.54–0.73) demonstrated moderate ability to identify patients who are likely to benefit from surgery. Models demonstrated fair calibration of the predicted probabilities.
Conclusions Outcomes after lumbar spinal fusion for degenerative disease remain difficult to predict. Although assistive clinical prediction models can help in quantifying potential benefits of surgery and the externally validated FUSE-ML tool may aid in individualized risk–benefit estimation, truly impacting clinical practice in the era of "personalized medicine" necessitates more robust tools in this patient population.

Keywords Predictive analytics · Outcome prediction · Machine learning · Spinal fusion · Neurosurgery · Clinical prediction model

Introduction

Degenerative disease of the lumbar spine, including chronic low back pain (CLBP), lumbar spinal stenosis, lumbar disc herniation, and degenerative lumbar spondylolisthesis, is part of the top-three causes of disability in Western societies and imposes significant direct and indirect socio-economic costs [1]. The gold standard treatment for these chronic degenerative diseases is multidisciplinary therapy including components of exercise therapy, cognitive behavioural therapy, and pharmacological therapy, although certain patients who are unresponsive to long-term conservative treatment may benefit from fusion [2, 3]. Nonetheless, with some reports showing no benefit compared with conservative treatment in a randomized population, patient selection is vitally important [4]. Various prognostic tests exist to attempt to identify subsets of patients that might truly benefit from surgery as a "last resort", but the validity of these tests is unclear [5, 6]. A relevant proportion of patients with intractable, conservative therapy-resistant lumbar degenerative disease does finally benefit from lumbar fusion surgery—the difficult question is how to identify these subsets reliably and how to avoid unnecessary, unsuccessful surgery [3].

Clinical prediction models can summarize a large number of factors into a single, potentially more accurate prediction of surgical risk or benefit, tailored to each individual patient [7–9]. The implementation of machine learning (ML) is increasing exponentially, although methodological rigour is only rarely upheld [8, 10]. Without thorough methodological foundations, development of clinical prediction models can very easily lead to pseudo-reliable predictions with seemingly high-performance measures due to issues such as data leakage, class imbalance, and overfitting [8, 11]. If clinical prediction models are not externally validated properly, real-world performance cannot be adequately estimated, and they should not to be applied in clinical practice [12, 13].

For patients with degenerative disease of the lumbar spine in whom spinal fusion surgery is considered, accurate prediction of long-term outcome in individual patients has been demonstrated to be extraordinarily difficult [5, 14]. The aim of the FUSE-ML consortium was to assemble a large multinational dataset of patients undergoing lumbar spinal fusion for degenerative disease in order to create robust clinical prediction models that take into account surgical variables and that are thoroughly developed and externally validated in a range of international centres.

Methods

Overview

A substantial multinational (7 countries), multicentre (11 centres) dataset (FUSE-ML) of patients who had undergone lumbar spinal fusion for degenerative disease was used to develop and externally validate a ML-based prediction tool for mid-term patient-reported outcomes. We then briefly compared the performance to that of the—to our knowledge—only other comparable, externally validated, clinical prediction model [14]. This study adheres to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis guidelines and is registered on ClinicalTrials.gov (NCT05161130) [7]. The use of patient data for research purposes was approved by each local institutional review board (IRBs), and patients provided informed consent or informed consent was waived, depending on the demands of the local IRB.

Inclusion and exclusion criteria

Patients with the following indications for thoracolumbar pedicle screw placement were considered for inclusion: degenerative pathologies (one or multiple of the following: spinal stenosis, spondylolisthesis, degenerative disc disease, disc herniation, failed back surgery syndrome (FBSS), radiculopathy, pseudarthrosis). Exclusion criteria were: surgery for—as the primary indication—infections, vertebral tumours, as well as traumatic and osteoporotic fractures or deformity surgery for scoliosis or kyphosis; moderate or severe scoliosis (Coronal Cobb's > 30°/Schwab classification sagittal modifier + or + +); surgery at more than 6 vertebral levels; missing endpoint data at 12 months; lack of informed consent; age < 18 years old.

Data collection

Each centre either extracted data retrospectively, from a prospective registry, or collected data in a prospective registry supplemented by retrospectively collected variables, with complete mid-term follow-up. The following data were collected: age, gender, surgical indication, index level(s), height, weight, BMI, smoking status, American Society of Anesthesiologists (ASA) Score, preoperative use of opioid pain medication, bronchial asthma as a comorbidity, prior thoracolumbar spine surgery, race/ethnicity, surgical approach, pedicle screw insertion and minimally invasive technique. PROMs included preoperative (baseline) and 12-month postoperative Oswestry Disability Index (ODI) (scaled from 0 to 100) or Core Outcome Measures Index (COMI) for multidimensional subjective functional impairment, numeric rating scale (NRS) for back pain severity, and NRS for leg pain severity [15, 16].

Primary endpoint definitions

Clinically relevant improvements in terms of functional impairment (ODI or COMI) and back/leg pain were dichotomized using the minimum clinically important difference (MCID) according to validated thresholds (Improvement from baseline to 12 months postoperatively of ≥ 15 points for ODI, ≥ 2.2 points for COMI, and ≥ 2 points for NRS pain severity) [17–19]. Thus, improvements from baseline that were greater than these validated thresholds were counted as achievement of MCID in the respective score.

Clinical prediction modelling

Numerical input variables were standardized using centring and scaling, and Yeo–Johnson transformation, and highly correlated variables (Pearson correlation coefficient ≥ 0.8) were filtered. A preoperative or postoperative ODI of ≤ 22 [20], COMI of ≤ 3.05 [21], or NRS pain severity of ≤ 3 [16] was considered as a probable "patient acceptable symptom state" (PASS) [22] based on established cut-offs. Patients with a preoperative PASS (minimal symptoms) in one of the three outcome dimensions were excluded from training for that respective dimension. Recursive feature elimination based on generalized linear models (GLMs) was carried out to identify the optimal, parsimonious set of inputs for each of the 3 models. Subsequently, GLMs were trained using Elastic Net Regularization using the Caret [23] library. During training, hyperparameters were tuned using fivefold cross-validation with 10 repeats, maximizing area under the curve (AUC). A *k*-nearest neighbour imputer was trained to impute missing data. The threshold for binary classification was selected based on the "closest-to-(0, 1)-criterion" and rounded. The models were then integrated into a web-app and underwent external validation. No recalibration was carried out. Quantile-based 95% confidence intervals (CIs) of the discrimination and calibration metrics were obtained from 1000 bootstrap resamples. Standardized model coefficients are reported to allow for explanation. [23] Finally, the models reported by Khor et al. [14] were reconstructed from the published coefficients, and external validation performance was compared. Notably, the Khor et al. model takes insurance status, which was not available within the FUSE-ML consortium. As has been done previously and due to the fact that virtually all inclusions in the FUSE-ML dataset stem from countries with either single-payer healthcare or compulsory health insurance, we adopted "*Medicare/Medicaid*" as the most appropriate choice for the entire cohort. [12] All analyses were carried out in *R* version 4.1.1.

Results

Patient cohort

Data from 1115 patients were provided by 11 participating centres in total. The development cohort was made up of 8 centres (817 patients, 42.7% male, age: 61.19 ± 12.36 years), while the remaining 3 centres were used for external validation (298 patients, 35.6% male, age: 59.73 ± 12.64 years). Achievement of MCID at 12-months was recorded in 761 (68.3%) patients for functional impairment, 862 (77.3%) patients for back pain severity, and 796 (71.4%) patients for leg pain severity. An overview of patient characteristics is provided in Table 1, and detailed patient characteristics including missingness and data per centre are shown in Supplementary Table 1. Overall, 3074 of 52′405 baseline data fields (5.9%) were incomplete.

Table 1 Summary of patient characteristics and outcome measures

Centre	Overall (pooled)	Development cohort	External validation cohort
N	1115	817	298
Male gender, *n* (%)	455 (40.8)	349 (42.7)	106 (35.6)
Age, mean (SD) [yrs.]	60.8 (12.5)	61.2 (12.4)	59.7 (12.6)
Height, mean (SD) [cm]	166.5 (9.8)	167.6 (9.6)	162.1 (9.4)
Weight, mean (SD) [kg]	73.5 (14.9)	74.7 (14.8)	69.1 (14.6)
Body mass index, mean (SD) [kg/m^2]	26.6 (4.61)	26.8 (4.9)	26.1 (3.9)
Smoking status, n (%)			
Active smoker	306 (27.4)	236 (29.0)	70 (24.1)
Ceased smoking	192 (17.2)	166 (20.4)	26 (9.0)
Never smoked	607 (54.4)	413 (50.7)	194 (66.9)
ASA score ≥ 3, *n* (%)	324 (29.1)	251 (31.4)	73 (24.5)
Opioid analgetic use, *n* (%)	364 (32.6)	314 (43.9)	50 (16.8)
Bronchial asthma, *n* (%)	63 (5.7)	51 (7.1)	12 (4.0)
Race/ethnicity, n (%)			
White	861 (77.2)	667 (93.0)	194 (65.5)
Black	30 (2.7)	29 (4.0)	1 (0.3)
Asian	106 (9.5)	6 (0.8)	100 (33.8)
Other	16 (1.4)	15 (2.1)	1 (0.3)
Prior thoracolumbar surgery, *n* (%)	257 (23.0)	204 (25.0)	53 (26.8)
Indication(s) for surgery, n (%)			
Spondylolisthesis	599 (53.7)	414 (50.7)	185 (62.1)
Lumbar disc herniation	202 (18.1)	139 (17.0)	63 (21.1)
Radiculopathy	323 (29.0)	230 (32.1)	93 (31.2)
Discogenic CLBP/DDD	457 (41.0)	337 (41.2)	120 (40.3)
FBSS	47 (4.2)	31 (4.3)	16 (5.4)
Lumbar spinal stenosis	618 (55.4)	429 (52.5)	189 (63.4)
Pseudarthrosis	56 (5.0)	55 (7.7)	1 (0.3)
Surgical index level(s), n (%)			
T12/L1	39 (3.5)	36 (4.4)	3 (1.0)
L1/L2	24 (2.2)	19 (2.3)	5 (1.7)
L2/L3	126 (11.3)	114 (14.0)	12 (4.0)
L3/L4	305 (27.4)	245 (30.0)	60 (20.1)
L4/L5	657 (58.9)	529 (64.7)	128 (64.6)
L5/S1	401 (36.0)	344 (42.1)	57 (28.8)
Surgical technique, n (%)			
TLIF	373 (33.5)	199 (27.8)	174 (58.4)
PLIF	449 (40.3)	325 (45.3)	124 (41.6)
ALIF	7 (0.6)	7 (1.0)	0 (0.0)
Lateral	73 (6.5)	73 (10.2)	1 (0.3)
Minimally invasive, *n* (%)	310 (27.8)	207 (25.3)	103 (34.6)
Pedicle screw insertion, *n* (%)	1081 (97.0)	783 (95.8)	298 (100.0)
Baseline patient-reported outcome			
Baseline ODI, mean (SD)	50.2 (17.9)	51.5 (17.5)	47.4 (18.6)
Baseline COMI, mean (SD)	7.5 (1.7)	7.5 (1.7)	–
Baseline back pain, mean (SD)	6.8 (2.3)	6.9 (2.3)	6.7 (2.4)
Baseline leg pain, mean (SD)	6.3 (2.8)	6.2 (2.8)	6.5 (2.7)
Baseline PASS[a] for function, *n* (%)	58 (5.2)	29 (3.8)	29 (9.7)
Baseline PASS[a] for back pain, *n* (%)	102 (9.1)	68 (8.4)	34 (11.4)
Baseline PASS[a] for leg pain, *n* (%)	192 (17.2)	152 (19.0)	40 (13.4)

Table 1 (continued)

Centre	Overall (pooled)	Development cohort	External validation cohort
12-month patient-reported outcome			
12-month ODI, mean (SD)	21.6 (16.5)	21.6 (16.7)	21.6 (16.1)
12-month COMI, mean (SD)	3.4 (2.9)	3.4 (2.9)	–
12-month back pain, mean (SD)	3.1 (2.4)	3.1 (2.4)	3.1 (2.4)
12-month leg pain, mean (SD)	2.5 (2.5)	2.5 (2.5)	2.5 (2.6)
12-month MCID[b] for function, *n* (%)	761 (68.3)	563 (74.4)	198 (66.4)
12-month MCID[b] for back pain, *n* (%)	862 (77.3)	640 (80.2)	222 (74.5)
12-month MCID[b] for leg pain, *n* (%)	796 (71.4)	564 (71.2)	232 (77.9)

SD standard deviation, *ASA* American society of anesthesiologists, *CLBP* chronic low back pain, *DDD* degenerative disc disease, *FBSS* failed back surgery syndrome, *TLIF* transforaminal lumbar interbody fusion, *PLIF* posterior lumbar interbody fusion, *ALIF* anterior lumbar interbody fusion, *ODI* Oswestry Disability Index, *COMI* Core Outcome Measures Index, *MCID* Minimum clinically important difference, *PASS* Patient acceptable symptom state

[a]PASS (Patient acceptable symptom state) was defined as a ODI of ≤ 22, COMI of ≤ 3.05, or a NRS of ≤ 3 for back and leg pain

[b]MCID (Minimum clinically important difference) was defined as a 15-point or greater improvement in ODI or a 2.2-point or greater improvement in COMI (function), or as a 2-point or greater improvement in NRS pain scores at 12 months compared to baseline, respectively

Table 2 Discrimination and calibration metrics of the machine learning-based prediction models for clinically relevant improvement

Metric	Models for improvement					
	Functional impairment [ODI/COMI] (MCID)		Back pain (MCID)		Leg pain (MCID)	
	Development	External validation	Development	External validation	Development	External validation
Model	Elastic net-regularized GLM		Elastic net-regularized GLM		Elastic net-regularized GLM	
Dichotomization cutoff	0.75		0.85		0.80	
No. observations	730	269	724	264	640	258
No. input variables	10		8		8	
Sampling	–	–	–	–	–	–
Discrimination						
AUC	0.75 (0.73–0.76)	0.67 (0.59–0.74)	0.71 (0.69–0.73)	0.72 (0.64–0.79)	0.72 (0.71–0.73)	0.64 (0.54–0.73)
Accuracy	0.70 (0.69–0.71)	0.61 (0.55–0.67)	0.68 (0.66–0.69)	0.70 (0.64–0.75)	0.74 (0.73–0.74)	0.71 (0.65–0.77)
Sensitivity	0.70 (0.68–0.72)	0.59 (0.52–0.66)	0.68 (0.67–0.69)	0.72 (0.65–0.77)	0.77 (0.76–0.78)	0.76 (0.71–0.82)
Specificity	0.70 (0.68–0.72)	0.66 (0.55–0.77)	0.63 (0.60–0.66)	0.64 (0.51–0.78)	0.58 (0.56–0.60)	0.42 (0.26–0.57)
PPV	0.88 (0.87–0.89)	0.81 (0.74–0.88)	0.91 (0.91–0.92)	0.90 (0.85–0.94)	0.90 (0.89–0.91)	0.88 (0.83–0.92)
NPV	0.43 (0.41–0.45)	0.39 (0.31–0.48)	0.26 (0.24–0.27)	0.34 (0.24–0.44)	0.34 (0.33–0.36)	0.23 (0.14–0.33)
F1 score	0.54 (0.52–0.55)	0.49 (0.41–0.58)	0.37 (0.34–0.39)	0.45 (0.34–0.54)	0.43 (0.42–0.45)	0.30 (0.19–0.41)
Calibration						
Intercept	0.00 (−0.05–0.06)	−0.07 (−0.36–0.22)	−0.00 (−0.07–0.07)	−0.38 (−0.70–0.06)	0.00 (−0.04–0.05)	0.14 (−0.22–0.51)
Slope	0.89 (0.84–0.95)	0.63 (0.34–0.93)	0.86 (0.77–0.94)	1.10 (0.62–1.57)	0.84 (0.79–0.89)	0.49 (0.12–0.86)

Metrics are provided with bootstrapped 95% confidence intervals based on 1000 samples with replacement. Reported development performance is the resampled cross-validation performance

MCID Minimum clinically important difference, *GLM* generalized linear model, *AUC* area under the curve, *PPV* positive predictive value, *NPV* negative predictive value

Table 3 Model coefficients of the fully trained models

Model coefficients (MCID)			
Variable	Function	Back pain	Leg pain
Model intercept	1.399	2.021	1.828
Male gender			0.214
Age	0.291		
Height	0.190		
ASA score ≥ 3	− 0.188		
Opioid analgetic use	− 0.156		
Prior thoracolumbar surgery		− 0.206	− 0.293
Indication(s) for surgery			
Lumbar disc herniation		0.157	
Radiculopathy	− 0.131	− 0.126	
Discogenic CLBP / DDD			− 0.238
Surgical index level(s)			
L4/L5			− 0.160
L5/S1		− 0.211	
Surgical technique			
TLIF	− 0.139	0.284	0.135
PLIF	0.169	0.271	0.299
Lateral	0.347	0.666	
Baseline patient-reported outcome			
Baseline ODI/COMI	1.026		
Baseline back pain	− 0.340	0.725	− 0.187
Baseline leg pain			0.812

Since centring and scaling were applied to the training data, the magnitude of the coefficients corresponds to variable importance

MCID minimum clinically important difference, *ASA* American society of anesthesiologists, *CLBP* chronic low back pain, *DDD* degenerative disc disease, *TLIF* transforaminal lumbar interbody fusion, *PLIF* Posterior lumbar interbody fusion, *ODI* Oswestry Disability Index, *COMI* Core Outcome Measures Index

Performance evaluation

Detailed model performance, including resampled development and external validation performance, is summarized in Table 2, and standardized model coefficients—enabling judgement of variable importance—are provided in Table 3. Calibration plots generated from the external validation cohort are shown in Fig. 1 including resampled training calibration, external validation calibration, and calibration from the Khor et al. model applied to the FUSE-ML external validation cohort. A detailed performance comparison with the Khor et al. model is available in Supplementary Table 2.

Prediction of functional impairment

At external validation, the FUSE-ML prediction model for clinical success in terms of functional impairment (ODI/COMI) achieved an AUC of 0.67 (95% CI: 0.59–0.74), sensitivity of 0.59 (95% CI: 0.52–0.66) and specificity of 0.66 (95% CI: 0.55–0.77). In terms of calibration, we measured a calibration intercept of − 0.07 (95% CI: − 0.36–0.22) and a calibration slope of 0.63 (95% CI: 0.34–0.93). When studying the standardized model coefficients, it was clear that predictions were mostly driven by greater baseline ODI/COMI scores, age, and lower back pain severity preoperatively, and application of a lateral surgical approach. The Khor et al. model achieved an AUC of 0.71 (95% CI: 0.64–0.77) on the same external validation cohort.

Prediction of back pain severity

Prediction of clinical success in terms of back pain severity in the external validation dataset was achieved with an AUC of 0.72 (95% CI: 0.64–0.79), sensitivity of 0.72 (95% CI: 0.65–0.77) and specificity of 0.64 (95% CI: 0.51–0.78). The calibration intercept was − 0.38 (95% CI: − 0.70–0.06) and slope, 1.10 (95% CI: 0.62–1.57). Higher baseline back pain and a lateral surgical approach were assigned the highest importance by the model. The Khor et al. model demonstrated an AUC of 0.73 (95% CI: 0.65–0.79) at external validation.

Prediction of leg pain severity

At external validation, long-term leg pain severity was predicted with an AUC of 0.64 (95% CI: 0.54–0.73), sensitivity of 0.76 (95% CI: 0.71–0.82) and specificity of 0.42 (95% CI: 0.26–0.57). The calibration intercept was 0.14 (95% CI: − 0.22–0.51) and calibration slope, 0.49 (95% CI: − 0.12–0.86). Looking at model coefficients, it appeared that greater baseline leg pain, a posterior surgical approach, and the absence of prior thoracolumbar surgery contributed most to the predictions of leg pain. The Khor et al. model had a corresponding AUC of 0.63 (95% CI: 0.54–0.71).

Model deployment

The prediction model was integrated into a freely available, web-based application accessible at https://neurosurgery.shinyapps.io/fuseml/.

Discussion

The rationale of the FUSE-ML study was to develop and thoroughly externally validate clinical prediction models for 12-month MCID in ODI/COMI, back pain, and leg pain in patients undergoing lumbar fusion for degenerative disease of the lumbar spine. Using data from 11 centres in 7 countries, a web-app was generated. After thorough external validation, we found that the fully trained clinical prediction models demonstrated only moderate ability to

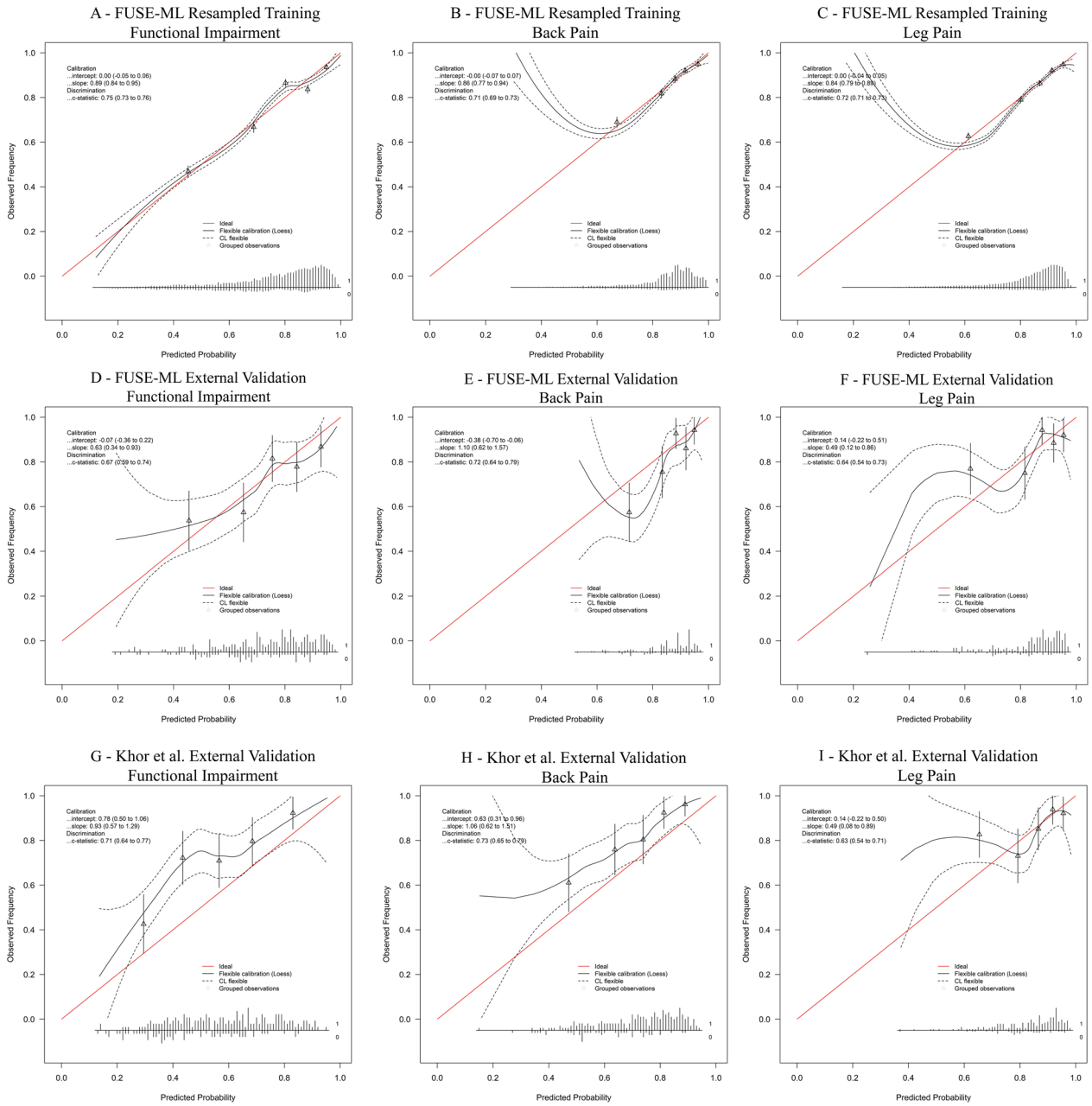

Fig. 1 Calibration curves of the three clinical prediction models for function, back pain, and leg pain on the resampled development cohort (**a**–**c**), cross-validation performance), the external validation cohort (**d**–**f**), FUSE-ML models at external validation), as well as those generated from the performance of the Khor et al. [14] prediction model applied to the FUSE-ML external validation cohort (**g**–**i**). The predicted probabilities for functional impairment (ODI/COMI) are distributed into five equally sized groups and contrasted with the actually observed frequencies of functional impairment. Calibration intercept and slope are calculated. A perfectly calibrated model has a calibration intercept of 0 and slope of 1. Metrics are provided with bootstrapped 95% confidence intervals

dichotomize patients who did and those who did not benefit from lumbar fusion surgery (discrimination performance). Calibration performance—the reliability of the predicted probabilities—was fair. Generally, our models performed comparably well to those published previously by Khor et al. although our models appeared to require only around half of the inputs to achieve the same performance, which streamlines implementation.

Our findings, coupled with those reported in the literature for patients with degenerative disease of the lumbar spine, demonstrate that the accurate prediction of long-term postoperative PROMs in this patient population remains remarkably difficult, and that clinical prediction models should only

have a minor role in clinical decision-making. It is well-known that even expert surgeons can overestimate the benefits and underestimate complications of certain procedures [24]. Clinical outcomes in degenerative disease of the lumbar spine and spinal fusion—and in particular CLBP, FBSS, and low-grade spondylolisthesis—are known as distinctly difficult to anticipate, and few independent predictors with a sufficiently large effect size are known [5, 14, 25]. Taking the example of discogenic CLBP, all recent randomized studies show that fusion surgery—overall—does not produce significantly better results than conservative treatment [4]. While surgery may not provide a benefit compared to conservative treatment for CLBP in the general patient population, there are subsets of patients that will truly benefit [5, 6]. Rigorous patient selection is the key to success in degenerative spine surgery.

In theory, clinical prediction models can provide valuable insights, since they enable calculation of individualized likelihoods of improvements or complications for each patient—as opposed to informing patients about a generalized treatment success rate that is based on historical data in the literature [26]. The hopes of being able to predict the effects of fusion surgery more robustly by generating "objective" risk–benefit profiles for each individual patient have not been fulfilled to date [26]. Janssen et al. [27] achieved an externally validated AUC of 0.68 for prediction of MCID in the predominant pain complaint using a nomogram. Apart from this nomogram, to our best knowledge, the only other externally validated prediction tools that predict pain and functional outcomes for this population are the prediction models of Khor et al. [14]. The latter was developed using the data of 1965 adult lumbar fusion surgery patients collected from a registry of fifteen Washington state hospitals. This model has recently been externally validated at a single Dutch centre, demonstrating AUCs of 0.71–0.83, sensitivities of 0.64–1.00, and specificities of 0.38–0.65, with fair calibration. [12] This analysis demonstrated that the discrimination and calibration performance generalized relatively well to a new population, although this level of performance unfortunately still would not allow any reliable decision support in actual clinical practice. FUSE-ML is largely based on the same inputs as those used in the Khor et al. [14] tool, although we attempted to improve upon the predictions by introducing surgical variables. In our extensive, multinational external validation study, the FUSE-ML models demonstrated only moderate discrimination and calibration, both of which appeared similar to the performance of the Khor et al. models when applied to our external validation dataset. Still, judging by these performance measures, these models would likely not be very helpful in clinical practice. The discrimination and calibration performance of expert surgeons has not been established as yet for lumbar fusion in degenerative disease. As long as these metrics remain unknown and as long as comparative or randomized studies do not demonstrate superiority of a decision-making approach integrating machine learning, these supportive tools ought to be used only adjunctively and with great caution in this patient population.

Even with the considerable amount of development data available to us for FUSE-ML, and the application of, e.g. regularization techniques, outcomes after lumbar spinal fusion remained difficult to predict with high reliability. One likely contributing factor is the input data: while we included a wide range of relevant socio-demographic, disease-specific, and surgical variables, the addition of imaging data for radiomic analysis and the inclusion of psychological factors could potentially improve predictions. The rationale behind the current approach was to only include few simple, preoperatively and easily available variables, with the intention of keeping prediction tools simple, accessible, and quick to use. This goal was also achieved: we demonstrated that our models generalized to an external validation dataset as approximately equally well as previously published, robust models did—although the FUSE-ML models appeared to enable the same level of performance with only around half of the inputs required. [14] More parsimonious models, rather than more complex models that require hard-to-collect inputs, are more prone to overfitting and may not be interpretable at all ("black box") [28, 29].

Still, even generally—in other patient populations—there is little to no high-quality evidence that clinical prediction models have any measurable clinical impact in their current state. A simulation analysis by Joshi et al. [30] found that only if applied on a population scale, prediction models in adult spinal deformity may overall decrease healthcare costs by better redirection of resources. Prospective clinical studies evaluating the real-world impact of integrating decision support tools into practice are currently not available. All of the above indicates a need for improving the methods, performance, and in silico/in vivo validation of clinical prediction models. However, caution must be exercised: the publication of clinical prediction models has increased exponentially over the past few years, as a result of the equally exponential access to computing power and "big data". [8] Exactly because it has become relatively easy to generate prediction models, many of these publications fall into common methodological ML "traps", which can catch out reviewers of expert medical journals. An important notion is the fact that it is relatively easy to generate prediction models with seemingly high-performance measures if certain concepts are disregarded—such as class imbalance, data leakage, adequate resampling, and proper validation, among others [8, 11]. Furthermore, the vast majority of published models have not undergone external validation and would very likely perform considerably worse in external validation studies [10, 13]. A recent review by Lubelski et al.

highlighted the vast methodological deficits in the spinal prediction modelling literature [10]. Lastly, the hopes that ML may help improve predictive performance compared to "traditional statistical modelling" have not been fulfilled, as a systematic analysis by Christodoulou et al. concludes [31]. ML certainly has advantages when analysing highly dimensional data, imaging data, or in natural language processing and time series analysis, but for "simple" tabulated clinical data as is the case with most prediction models, the advantages of ML over, e.g. "traditional" generalized linear models likely, do not outweigh their drawbacks [8, 31].

We do not recommend the use of clinical prediction models—even those with very high-performance metrics—as absolute "red light" or "green light" indicators, but advocate carefully balancing all available clinical data against patient wishes and expectations as well as clinical expertise. There is a need for improved clinical prediction models in spinal fusion for degenerative disease of the lumbar spine, and development will require major international collaborative efforts to collect larger amounts of data and to enable thorough validation of developed models. The FUSE-ML collaborators will continue investigating approaches to improving patient selection in this population.

Strengths and limitations

Our study used data from 11 centres in different countries, with unified variable definitions. The models have been made available as a web-based tool. Different degenerative spinal diseases were included. Consequently, our models may perform better for more common pathologies, whereas performance may be limited for the less prevalent ones. Conversely, this heterogeneity in training data may equip the models for the heterogenous presentations of spinal degenerative disease. We also directly compare the performance of our models to the current "benchmark" model in spinal fusion surgery and demonstrate approximate equivalence of our performance at external validation, as well as fair calibration of our models.

Our data consisted of a mix of retrospectively and prospectively collected data from institutional registries. Many definitions of MCID—and, in the same vein, of PASS—exist, and their choice determines the interpretation of generated predictions [15]. We chose a MCID based on robust MCID studies[17–19], and we excluded patients unlikely to improve by determining a minimally symptomatic state (PASS) based on thresholds from analyses that were anchored to patient-rated symptom satisfaction [16, 20, 21]. Our prediction tool does not include measures of quality of life and psychological factors, which may improve performance. Learning techniques rely on large amounts of development data and often improve their performance linearly with an increasing number of training samples. Thus, although we included a relatively large cohort of patients, further training with a larger sample is likely to improve the performance and generalization of the models. We excluded patients under the age of 18 and those with spinal deformity. Our models may not necessarily generalize when extrapolating to these patients.

Conclusions

With the great heterogeneity of outcomes after lumbar spinal fusion for degenerative disease and the countless physical and psychological factors that may modulate the effects of procedures, identifying those patients most likely to benefit from surgical treatment in an objective fashion remains difficult. Although assistive clinical prediction models can help in quantifying potential benefits of surgery and the externally validated FUSE-ML tool (https://neurosurgery.shinyapps.io/fuseml) may aid in individualized risk–benefit estimation, truly impacting clinical practice in the era of "personalized medicine" will necessitate improvements in reliability of clinical prediction models in this patient population. When thoroughly externally validated, current approaches based on tabulated clinical data fail to break the performance barrier required to prevent ineffective surgery or to allow meaningful decisions that are at least partially informed by such clinical prediction models.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00586-022-07135-9.

Acknowledgements We thank the patients whose anonymized data were used for this research.

Funding Open access funding provided by University of Zurich. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations

Conflict of interest The authors declare that the article and its content were composed in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ravindra VM, Senglaub SS, Rattani A et al (2018) Degenerative lumbar spine disease: estimating global incidence and worldwide volume. Glob Spine J 8:784–794. https://doi.org/10.1177/2192568218770769
2. Manchikanti L, Abdi S, Atluri S et al (2013) An update of comprehensive evidence-based guidelines for interventional techniques in chronic spinal pain. Part II: guidance and recommendations. Pain Physician 16:S49-283
3. Bono CM, Lee CK (2004) Critical analysis of trends in fusion for degenerative disc disease over the past 20 years: influence of technique on fusion rate and clinical outcome. Spine 29:455–463. https://doi.org/10.1097/01.brs.0000090825.94611.28
4. Mannion AF, Brox J-I, Fairbank JC (2016) Consensus at last! long-term results of all randomized controlled trials show that fusion is no better than non-operative care in improving pain and disability in chronic low back pain. Spine J Off J North Am Spine Soc 16:588–590. https://doi.org/10.1016/j.spinee.2015.12.001
5. Staartjes VE, Vergroesen P-PA, Zeilstra DJ, Schröder ML (2018) Identifying subsets of patients with single-level degenerative disc disease for lumbar fusion: the value of prognostic tests in surgical decision making. Spine J 18:558–566. https://doi.org/10.1016/j.spinee.2017.08.242
6. Willems P (2013) Decision making in surgical treatment of chronic low back pain: the performance of prognostic tests to select patients for lumbar spinal fusion. Acta Orthop Suppl 84:1–35. https://doi.org/10.3109/17453674.2012.753565
7. Collins GS, Reitsma JB, Altman DG, Moons KGM (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 350:g7594
8. Kernbach JM, Staartjes VE (2020) Machine learning-based clinical prediction modeling—A practical guide for clinicians. http://arxiv.org/abs/200615069 Cs Stat
9. Steyerberg EW (2008) Clinical prediction models: a practical approach to development, validation, and updating. Springer, Berlin
10. Lubelski D, Hersh A, Azad TD et al (2021) Prediction models in degenerative spine surgery: a systematic review. Glob Spine J 11:79S-88S. https://doi.org/10.1177/2192568220959037
11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357. https://doi.org/10.1613/jair.953
12. Quddusi A, Eversdijk HAJ, Klukowska AM et al (2019) External validation of a prediction model for pain and functional outcome after elective lumbar spinal fusion. Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc. https://doi.org/10.1007/s00586-019-06189-6
13. Collins GS, de Groot JA, Dutton S et al (2014) External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 14:40. https://doi.org/10.1186/1471-2288-14-40
14. Khor S, Lavallee D, Cizik AM et al (2018) Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery. JAMA Surg 153:634–642. https://doi.org/10.1001/jamasurg.2018.0072
15. Ostelo RWJG, Deyo RA, Stratford P et al (2008) Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine 33:90–94. https://doi.org/10.1097/BRS.0b013e31815e3a10
16. Fekete TF, Haschtmann D, Kleinstück FS et al (2016) What level of pain are patients happy to live with after surgery for lumbar degenerative disorders? Spine J Off J North Am Spine Soc 16:S12-18. https://doi.org/10.1016/j.spinee.2016.01.180
17. Mannion AF, Porchet F, Kleinstück FS et al (2009) The quality of spine surgery from the patient's perspective: part 2. minimal clinically important difference for improvement and deterioration as measured with the core outcome measures index. Eur Spine J 18:374–379. https://doi.org/10.1007/s00586-009-0931-y
18. Farrar JT, Young JP, LaMoreaux L et al (2001) Clinical importance of changes in chronic pain intensity measured on an 11-point numerical pain rating scale. Pain 94:149–158. https://doi.org/10.1016/S0304-3959(01)00349-9
19. Parker SL, Adogwa O, Paul AR et al (2011) Utility of minimum clinically important difference in assessing pain, disability, and health state after transforaminal lumbar interbody fusion for degenerative lumbar spondylolisthesis. J Neurosurg Spine 14:598–604. https://doi.org/10.3171/2010.12.SPINE10472
20. van Hooff ML, Mannion AF, Staub LP et al (2016) Determination of the Oswestry disability index score equivalent to a "satisfactory symptom state" in patients undergoing surgery for degenerative disorders of the lumbar spine—a Spine Tango registry-based study. Spine J 16:1221–1230. https://doi.org/10.1016/j.spinee.2016.06.010
21. Genevay S, Marty M, Courvoisier DS et al (2014) Validity of the French version of the core outcome measures index for low back pain patients: a prospective cohort study. Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc 23:2097–2104. https://doi.org/10.1007/s00586-014-3325-8
22. Tubach F, Dougados M, Falissard B et al (2006) Feeling good rather than feeling better matters more to patients. Arthritis Care Res 55:526–530. https://doi.org/10.1002/art.22110
23. Kuhn M (2008) Building Predictive Models in *R* Using the caret Package. J Stat Softw 28:1–26. https://doi.org/10.18637/jss.v028.i05
24. Sacks GD, Dawes AJ, Ettner SL et al (2016) surgeon perception of risk and benefit in the decision to operate. Ann Surg 264:896–903. https://doi.org/10.1097/SLA.0000000000001784
25. Alentado VJ, Caldwell S, Gould HP et al (2017) Independent predictors of a clinically significant improvement after lumbar fusion surgery. Spine J Off J North Am Spine Soc 17:236–243. https://doi.org/10.1016/j.spinee.2016.09.011
26. Steinmetz MP, Mroz T (2018) Value of adding predictive clinical decision tools to spine surgery. JAMA Surg. https://doi.org/10.1001/jamasurg.2018.0078
27. Janssen ERC, Punt IM, van Kuijk SMJ et al (2020) Development and validation of a prediction tool for pain reduction in adult patients undergoing elective lumbar spinal fusion: a multicentre cohort study. Eur Spine J Off Publ Eur Spine Soc Eur Spinal Deform Soc Eur Sect Cerv Spine Res Soc 29:1909–1916. https://doi.org/10.1007/s00586-020-06473-w
28. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206. https://doi.org/10.1038/s42256-019-0048-x
29. Ariew R (1976) Ockham's razor: a historical and philosophical analysis of Ockham's principle of parsimony. Dissertation, PhD Thesis
30. Joshi RS, Serra-Burriel M, Pellise F et al (2020) 15. Use of predictive machine learning models at the population level has the potential to save cost by directing economic resources to those likely to improve most: a simulation analysis stratified by risk in largest combined US/European ASD registry. Spine J 20:S8. https://doi.org/10.1016/j.spinee.2020.05.118
31. Christodoulou E, Ma J, Collins GS et al (2019) A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 110:12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004