Contents lists available at ScienceDirect

North American Spine Society Journal (NASSJ)

journal homepage: www.elsevier.com/locate/xnsj

Clinical Studies

PROPOSE. Development and validation of a prediction model for shared decision making for patients with lumbar spinal stenosis

Casper Friis Pedersen, MSSc, PhD[a,*], Mikkel Østerheden Andersen, MD[b], Leah Yacat Carreon, MD, MSc[b], Simon Toftgaard Skov, MD[c], Peter Doering, MD, PhD[d], Søren Eiskjær, MD[d]

[a] *Center for Spine Surgery and Research, Spinecenter of Southern Denmark, Lillebaelt Hospital, Oestre Hougvej 55, DK-5500, Middelfart, Denmark*
[b] *University of Southern Denmark, Center for Spine Surgery and Research, Spinecenter of Southern Denmark, Lillebaelt Hospital, Oestre Hougvej 55, DK-5500, Middelfart, Denmark*
[c] *Aarhus University, Elective Surgery Centre, Silkeborg Regional Hospital, Falkevej 3, DK-8600, Silkeborg, Denmark*
[d] *Department of Orthopedic Surgery, Aalborg University, Hobrovej 18-22, DK-9000, Aalborg, Denmark*

ARTICLE INFO

Keywords:
Lumbar decompression surgery
Lumbar spinal stenosis
Patient reported outcome
Prediction model
Shared decision-making

ABSTRACT

Background: Decompression for lumbar spinal stenosis (LSS) is the most frequently performed spine surgery in Denmark. According to the Danish spine registry DaneSpine, at 1 year after surgery, about 75% of patients experiences considerable pain relief and around 66% improvement in quality of life. However, 25% do not improve very much. We have developed a predictive decision support tool, PROPOSE. It is intended to be used in the clinical conversation between healthcare providers and LSS patients as a shared decision-making aid presenting pros and cons of surgical intervention. This study presents the development and evaluation of PROPOSE in a clinical setting.

Methods: For model development, 6.357 LSS patients enrolled in DaneSpine were identified. For model validation, predictor response and predicted outcome was collected via PROPOSE from 228 patients. Observed outcome at 1 year was retrieved from DaneSpine. All participants were treated at 3 Danish spine centers. The outcome measures presented are improvement in walking distance, the Oswestry Disability Index, EQ-5D-3L and leg/back pain on the Visual Analog Scale. Outcome variables were dichotomized into success (1) and failure (0). With the exception of walking distance, a success was defined as reaching minimal clinically important difference at 1-year follow-up. Models were trained using Multivariate Adaptive Regression Splines. Performance was assessed by inspecting confusion matrix, ROC curves and comparing GCV (generalized cross-validation) errors. Final performance of the models was evaluated on independent test data.

Results: The walking distance model demonstrated excellent performance with an AUC of 0.88 and a Brier score of 0.14. The VAS leg pain model had the lowest discriminatory performance with an AUC of 0.67 and a Brier score of 0.22.

Conclusions: PROPOSE works in a real-world clinical setting as a proof of concept and demonstrates acceptable performance. It may have the potential of aiding shared decision making.

Background

Surgical decompression for lumbar spinal stenosis (LSS) is the most frequently performed spine surgery in Denmark with 2,000 to 2,500 yearly operations, constituting 39% of all activity [1]. Treatment effectiveness of surgical interventions for LSS are measured using Patient Reported Outcome Measures (PROMs). Based on data from the Danish National Spine Registry (DaneSpine) as of 2021, at 1 year after

FDA device/drug status: Not applicable.

Author disclosures: ***CFP:*** Nothing to disclose. ***MA:*** Nothing to disclose. ***LYC:*** Scientific Advisory Board: Editorial Advisory Board, Spine (none), Editorial Advisory Board, The Spine Journal (none); Grants: IRSs Kursus- og Rejsepulje (none), TrygFonden (none), Region Syddanmark, ph.d.-puljen, SLBs Forskningsråd (none), Sygeforsikring "Danmark" donation, Sundhedsstyrelsen; eksp kræftkirurgi (none), SLBs Forskningsråd ProjektsØtte (none), SDU Faculty Scholarship (none), Johnson&Johnson (none), Cerapedics (none). ***STS:*** Nothing to disclose. ***PD:*** Nothing to disclose. ***SE:*** Nothing to disclose.

[*] Corresponding author. Center for Spine Surgery and Research, Spinecenter of Southern Denmark, Lillebaelt Hospital, Oestre Hougvej 55, DK-5500, Middelfart, Denmark.

E-mail address: casper.friis.pedersen@rsyd.dk (C.F. Pedersen).

https://doi.org/10.1016/j.xnsj.2024.100309
Received 14 November 2023; Received in revised form 5 January 2024; Accepted 5 January 2024
Available online 11 January 2024

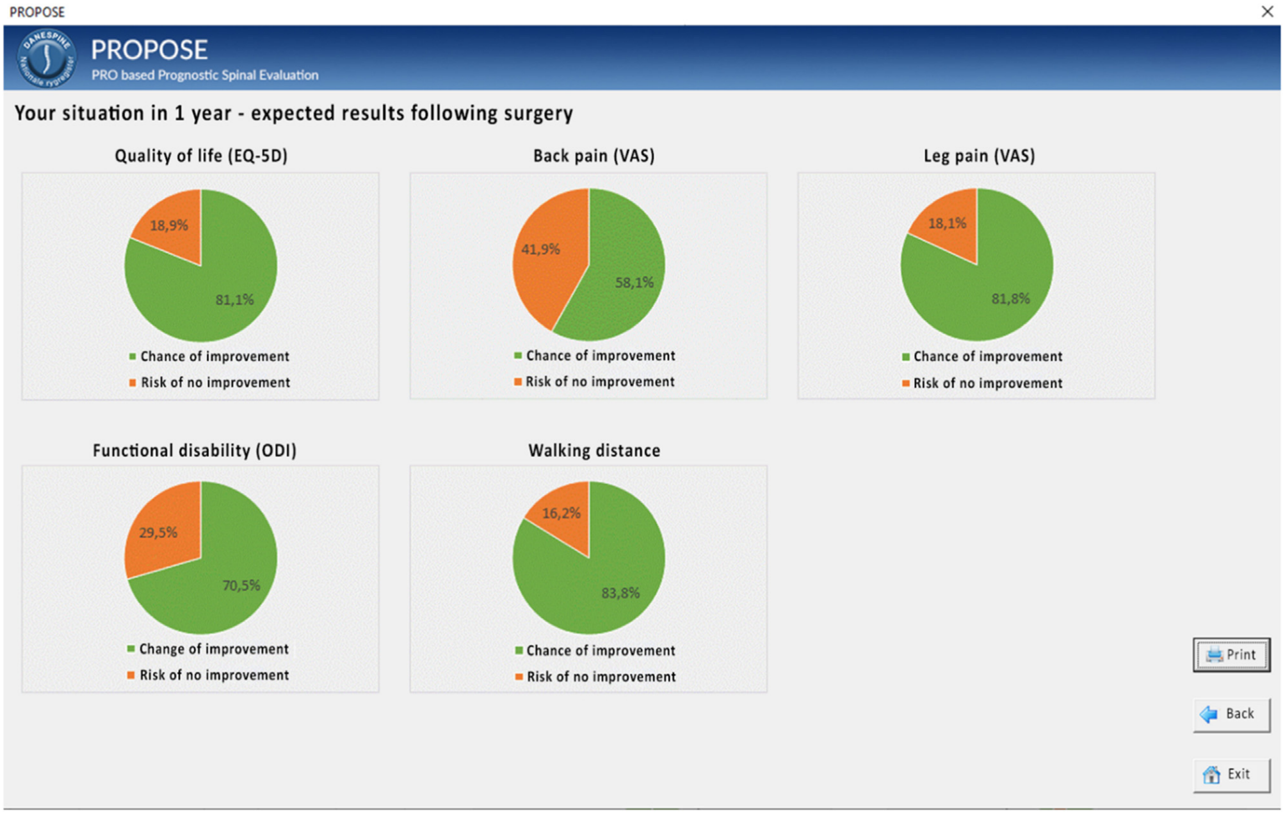

Fig. 1. Example case. PROPOSE results for a woman, 74 years, diagnosed with lumbar spinal stenosis. Leg pain for 3 - 12 months, back pain for 24 months or more, walking distance less than 100 meters, can only walk with cane or crutch, calculated life quality index of 0.159 (EQ-5D-3L), 90 and 60 leg/back pain on the Visual Analogue Scale. The green and orange proportion of the pie charts signifies the calculated probability of improvement and non-improvement respectively.

surgery, about 75% of patients experienced considerable pain relief and around 66% reported improvement in their quality of life. However, 25% of patients reported that they did not improve very much [2]. Due to several factors that may impact outcomes after spine surgery, identifying patients who will or will not have a successful result may be difficult.

DaneSpine collects patient-reported data before surgery and at 1 year postoperatively. The data consists of PROMs and demographic information. This allows for identification of relationships among preoperative factors and outcomes after surgery in order to make predictions at an individual level. We have developed a patient-oriented decision support tool, PROPOSE, based on data collected in DaneSpine. PROPOSE is intended to be used by the healthcare provider in the clinical conversation as a shared decision-making aid in accordance with patient preferences presenting pros and cons of the surgical intervention.

PROPOSE works in 2 stages. The individual patient's baseline factors are entered in an electronic questionnaire and the underlying predictive algorithms calculates dichotomized probabilities of success or failure based on the input. At the user's command the results are presented as pie charts with accompanying numbers indicating the proportions of successful outcome and failure (Fig. 1). Baseline values and predicted probabilities are saved on a secure server for future reference.

The purpose of the study is to present the development of PROPOSE and to evaluate its performance in a clinical setting.

Methods

Sample size and data source

For model development, 6,357 patients operated for lumbar spinal stenosis from 2010 to 2018 enrolled in DaneSpine were identified. To validate the models, predictor response and predicted outcome was collected via PROPOSE from January 15th 2019 to May 31st 2021 from 228 patients and one-year follow-up was retrieved from DaneSpine. All participants were treated at 3 Danish spine centers (Spine Centre of Southern Denmark, Elective Surgery Centre, Silkeborg and Dept. of Orthopedic Surgery, Aalborg University).

Predictors

A total of 72 predictors from the DaneSpine registry were evaluated. These included basic demographics and patient-reported data at baseline. The codebook can be found at the following link (in Danish): http://drks.ortopaedi.dk/wp-content/uploads/2019/12/Kodebog-l%C3%A6nd.pdf. As PROPOSE is intended to be used in a clinical setting where there is little time to enter the amount of information required to calculate compound scores such as the component scores of the 36-Item Short Form Survey (SF-36) [3] only individual questionnaire items and the EQ-5D-3L questionnaire [4] was chosen. The final models contained 7 predictors (Table 1).

Outcome

The outcome measures presented to the patient are improvement in walking distance, the Oswestry Disability Index (ODI) [5], EQ-5D-3L and leg/back pain on the Visual Analog Scale (VAS) [6]. The outcome variables were dichotomized into 2 possible results, either success (1) or failure (0). With the exception of walking distance, a successful outcome was defined as reaching minimal clinically important difference (MCID) at 1-year follow-up. MCID thresholds were identified using anchor-based receiver operating characteristic curve (ROC) method [7].

The anchor used for ODI and EQ-5D-3L was the SF-36 health transition item 2 where patients are asked to rate their health compared with

Table 1
Predictors in the final predictive models and their encoding.

Predictors	Encoded
Quality of life (EQ-5D-3L)	-0.594 to 1.0
Preoperative VAS pain (legs)	0–100
Preoperative VAS pain (back)	0–100
Walking distance	
Walking distance, less than 100 m	1
Walking distance, 100–500 m	2
Walking distance, 0.5–1 km	3
Walking distance, > 1 km	4
Duration of pain in legs	
Duration of pain in legs, No pain	0
Duration of pain in legs, < 3 mo	1
Duration of pain in legs, 3–12 mo	2
Duration of pain in legs, 1–2 y	3
Duration of pain in legs, > 2 y	4
Duration of pain in back	
Duration of pain in back, No pain	0
Duration of pain in back, < 3 mo	1
Duration of pain in back, 3–12 mo	2
Duration of pain in back, 1–2 y	3
Duration of pain in back, > 2 y	4
Functional impairment, Walking (ODI section 4)	
Pain does not prevent me walking any distance	0
Pain prevents me from walking more than 1 km	1
Pain prevents me from walking more than 500 m	2
Pain prevents me from walking more than 100 m	3
I can only walk using a stick or crutches	4
I am in bed most of the time	5

Abbreviations: VAS, visual analogue pain scale; ODI, Oswestry Disability Index.

1 year ago with the following possible responses: "much better," "somewhat better," "about the same," or "somewhat worse," or "much worse". A cutoff defining success and failure was set between "somewhat better" and "about the same".

Anchors for VAS leg and back pain were the Global Assessment questions on leg pain/sciatica and back pain where patients are asked to compare their pain today in comparison with what they felt a year ago. Responses are "completely gone," "much better," "somewhat better," "unchanged," or "worse" [8]. Cutoff for success or failure was set between the responses "somewhat better" and "unchanged."

MCIDs were estimated by inspecting the coordinates of the ROC curves using the point closest to the top-left corner in the ROC curve as guideline. Cutoff points were established as 0.105 for EQ-5D-3L (sensitivity: 76.1; specificity: 64.5), 14 for ODI (sensitivity: 75.9; specificity: 74.5), 16 for VAS leg pain (sensitivity: 75.9; specificity: 74.5) and for back pain 14 (sensitivity: 71.3; specificity: 70.1).

Successful outcome for walking distance was defined as an improvement in the patient reported responses for walking distance from baseline to follow-up for the question: "How far can you walk at a normal pace?." Possible answers are "Less than 100 meters," "100 to 500 meters," "500 meters to 1 km," "More than one kilometer."

Missing data

Complete-case analysis was used during model development. Missingness was evenly distributed over predictors (22%) and outcome variables (40%) and assumed missing at random (MAR) [9]. Imputation of missing data was not attempted. The data collected to validate PROPOSE did not contain missing values.

Statistical analysis and methods

Data handling

The development data were screened for erroneous entries in accordance with the ranges and values given in the codebook. To identify unusual cases k-Nearest Neighbors (kNN) distance-based outlier detection was performed [10]. A few anomalies in BMI were found, with values in excess of 60, and replaced with blanks. Target variables representing the outcome were encoded according to the previously mentioned definition of success and failure. Minor class imbalances were found in favor of the positive class ranging from 56% to 66%. To prevent this from affecting the model's ability to discriminate correctly between classes synthetic minority oversampling (SMOTE) was applied [11]. The prepared data were then randomly split into a training and test set by a 70/30 % ratio. All data handling was done in RapidMiner Studio 9.1.0 [12].

Predictive modeling

A priority when selecting the predictive algorithm for PROPOSE was ease of implementation and use. Multivariate adaptive regression splines (MARS) is a transparent multivariate nonparametric machine learning (ML) algorithm capable of solving non-linear regression problems [13]. It is an extension of the stepwise linear regression and also handles classification problems. A further advantage of the algorithm is its built-in automatic feature selection. It is independent of predictors that do not add predictive performance to the model. MARS works by piecewise fitting together an ensemble of local linear functions thus adapting to nonlinearities [14].

The majority (80%) of the predictors consisted of categorical data with more than 2 levels. They were initially compared with the outcome variables by chi-square. To reduce computational modeling time, nonsignificant predictors were excluded from the analysis with a threshold of p≥.1. To identify multicollinearity, a mutual information matrix was produced [15].

A large number of models were trained for each outcome. Noninformative predictors were automatically detected by the MARS algorithm and removed in iterations. Predictors that contributed very little were also removed to reduce model complexity. Tuning of the models was done by manipulating the 2 parameters associated with MARS: maximal number of retained terms and the degree of the features that are added to the model. Degree of features limits the number of input variables that are considered for each piecewise function to reduce model complexity. Maximal number of terms was set at 5,000 and automatically reduced by MARS with the pruning option during training. To prevent overfitting and save computational time, degree of features was set at a maximum of 2.

Model performance was assessed by inspecting the resulting confusion matrix, ROC curves and comparing GCV (generalized cross-validation) errors. The final performance of the models was evaluated by applying them to the independent test data. The default threshold value of 0.5 was used for all outcomes. A predicted probability ≤0.5 corresponds to a failure (0) and if it is >0.5 a success (1). Threshold-moving to adjust for imbalanced classification was not performed. Model development was done in R version 3.5.3 using the CRAN package earth [16,17]. Earth is a free General Public Licensed implementation of the trademarked MARS algorithm.

Model performance assessment

The performance of the dichotomous classification models was evaluated by discriminative ability, agreement between observed and predicted outcome and overall probabilistic accuracy. Discrimination was measured by the area under the receiver operating characteristic curve (ROC) by plotting true positive rate (sensitivity) against false positive rate (1 - specificity) at all classification thresholds. The area under the curve (AUC) provides an aggregate metric of the models ability to rank random positives higher than random negatives, or degree of separation between classes. [18].

Agreement between observed and predicted outcome was assessed by calibration plots where predicted probabilities are plotted on the x-axis and observed outcome on the y-axis [19]. For binary outcomes, data must be converted into ratios by binning outcome and predictions into

Table 2
Descriptive statistics of the study samples as mean (SD) or proportions.

Characteristic	Validation cohort	Development data	p-value
Number of patients, (n)	228	6.357	
Age, years, mean (SD)	68.2 (10.2)	66.8 (11.4)	.030
Gender, females, n (%)	118 (51.8)	3.408 (53.6)	.470
Smoker, n (%)	52 (24.6)	1.224 (23.5)	.741
BMI, mean (SD)	28.6 (11.3)	27.8 (3.9)	.690
Comorbidity, n (%)	35 (17.2)	23.1 (1.221)	.053
Previous operated, n (%)	63 (31.0)	1.537 (29.2)	.571
No. of operated levels			
One level, n (%)	123 (54.2)	3.337 (53.5)	.629
Two levels, n (%)	82 (36.1)	2.155 (34.5)	
Three levels, n (%)	17 (7.5)	640 (10.3)	
Four levels, n (%)	5 (2.2)	102 (1.6)	
Five levels, n (%)	0 (0.0)	9 (0.1)	
Quality of life (EQ-5D-3L), mean (SD)	0.407 (0.297)	0.390 (0.313)	.634
Functional impairment (ODI), mean (SD)	40.9 (15.1)	42.3 (15.5)	.191
Preoperative VAS pain (legs), mean (SD)	65.6 (23.6)	64.0 (24.7)	.445
Preoperative VAS pain (back), mean (SD)	56.8 (26.4)	54.6 (28.1)	.360
Walking distance, less than 100 m, n (%)	57 (28.2)	1.866 (35.6)	.192
Walking distance, 100–500 m, n (%)	81 (40.1)	1.849 (35.3)	
Walking distance, 0.5–1 km, n (%)	34 (16.8)	797 (15.2)	
Walking distance, > 1 km, n (%)	30 (14.9)	728 (13.9)	
Duration of pain in legs, No pain, n (%)	3 (1.5)	120 (2.3)	.647
Duration of pain in legs, < 3 mo, n (%)	7 (3.5)	288 (5.5)	
Duration of pain in legs, 3–12 mo, n (%)	70 (34.8)	1.858 (35.4)	
Duration of pain in legs, 1–2 y, n (%)	54 (26.9)	1.289 (24.6)	
Duration of pain in legs, > 2 y, n (%)	67 (33.3)	1.691 (32.2)	
Duration of pain in back, No pain, n (%)	8 (4.0)	358 (6.9)	.237
Duration of pain in back, < 3 mo, n (%)	3 (1.5)	173 (3.3)	
Duration of pain in back, 3–12 mo, n (%)	45 (22.4)	1.217 (23.3)	
Duration of pain in back, 1–2 y, n (%)	41 (20.4)	926 (17.7)	
Duration of pain in back, > 2 y, n (%)	104 (51.7)	2.549 (48.8)	
Δ Quality of life (EQ-5D-3L), mean (SD)	0.287 (0.352)	0.269 (0.358)	.281
Δ Functional impairment (ODI), mean (SD)	18.0 (17.2)	16.1 (17.3)	.101
Δ VAS pain (legs), mean (SD)	33.3 (35.8)	29.0 (34.7)	.119
Δ VAS pain (back), mean (SD)	19.3 (33.0)	20.3 (32.0)	.649

Abbreviations: SD, standard deviation; ODI, Oswestry Disability Index; VAS, visual analogue pain scale.

equally sized subsets and calculating the average on both. Points on the 45-degree diagonal represents perfect calibration.

Overall probabilistic accuracy was measured by the Brier score. The Brier score is calculated as an average of the mean squared errors between predicted probabilities and observed values [20]. It quantifies the goodness of the predicted probabilities against outcome where a perfect model has a score of 0 and the worst possible model has a score of 1.

Application development

The decision support tool PROPOSE was coded in visual basic for application (VBA) in a Windows PC only environment. The application integrated with and relied on an external Excel file containing model specifics. Pie charts was chosen for graphical representation based on the results of a pilot-study with qualitative interviews including 20 patients.

Results

Study population

Baseline characteristics and outcome at 1-year of the study population are summarized in the second table (Table 2). Differences between development data and the validation cohort were small, the latter being slightly older.

Performance measures

An assessment of model performance is given as various metrics for the model testing set and the PROPOSE validation cohort in the third and fourth tables (Table 3, Table 4). As indicated by the Brier score the mean squared errors between predicted probabilities and observed values ranged from 0.16 to 0.22 in the model test data, and 0.14 to 0.22 in the validation set. Score rankings were consistent between test and validation data. The best accuracy of probabilities was demonstrated by the walking distance model. The VAS leg pain model had the lowest degree of accuracy.

Agreement between observed and predicted outcome (calibration) by PROPOSE is illustrated in the second figure (Fig. 2). The models predicting EQ-5D, back pain and walking distance demonstrated fair concordance on a group level. Models for ODI and leg pain were less concordant.

The discriminative ability of PROPOSE is also demonstrated for all 5 outcomes (Fig. 3). For EQ-5D, ODI, back pain and walking distance the AUC level is good to excellent ranging from 0.76 to 0.88. The performance of the leg pain model is less convincing with an AUC level of 0.67 and a lower confidence interval value of 0.57 (Fig. 3).

Discussion

Systematic data collection of PRO-based registry data enables the development of predictive prognostic models that can support shared decision making and possibly align expectations in the preoperative discussion of treatment options between patients and surgeons. Variability in clinical outcome following LSS surgery often makes it a difficult task for the surgeon to provide the patient with precise information on what to expect. Often, the surgeon must rely on clinical experience or knowledge on estimated average success rates in the literature. Predictive models based on existing cases makes it possible to estimate the most likely outcome on an individual case level. If implemented in an easy to interpret application understood by both patients and surgeons, such models

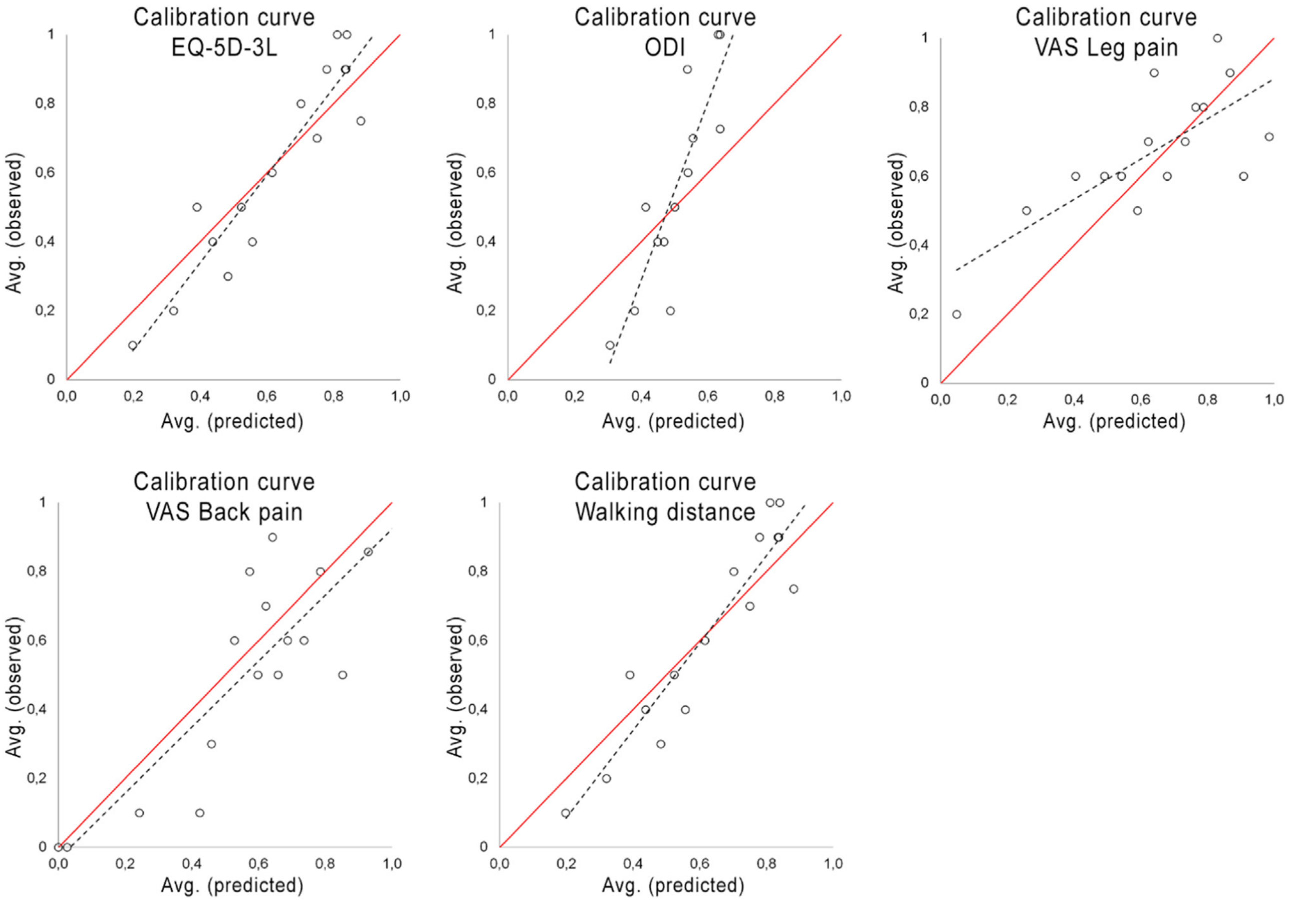

Fig. 2. Calibration curves for PROPOSE.

Table 3
Performance of predicted outcome measures – model test data.

Metrics	EQ-5D-3L	ODI	VAS Leg pain	VAS back pain	Walking distance
Number of patients, (n)	1.035	954	1.054	1.047	1.082
AUC, (CI)	80.4 (77.7; 83.0)	74.3 (71.2; 77.4)	70.6 (67.4; 73.8)	75.7 (72.7; 78.7)	81.9 (79.3; 84.5)
Sensitivity, %	69.0	62.7	65.2	69.5	70.0
Specificity, %	78.2	70.3	84.3	79.6	89.0
Classification error, % (CI)	27.8	32.4	32.6	28.3	25.0
Accuracy, %	72.2	67.6	67.4	71.7	75.0
Precision, %	85.9	53.5	97.0	92.5	95.0
Recall, %	69.0	62.7	65.2	69.5	70.0%
F-score	0.77	0.58	0.78	0.79	0.81
MCC	0.45	0.32	0.32	0.41	0.51
Youden's J	0.47	0.33	0.49	0.49	0.59
Cohen's kappe	0.45	0.22	0.58	0.55	0.53
Brier score	0.20	0.21	0.22	0.20	0.16
PPV, %	69.0	62.7	65.2	69.5	70.2
NPV, %	78.2	70.3	84.3	79.6	88.8

Abbreviations: AUC, area under the curve; CI, confidence interval; MCC, Matthews correlation coefficient; PPV, positive predicted value; NPV, negative predicted value.

could supplement the knowledge and experience of the surgeon and potentially improve the decision on treatment.

We have developed and validated a decision support tool PROPOSE which incorporates ML models predicting likelihood of improvement in EQ-5D-3L, ODI, VAS leg, VAS back and walking distance at 12 months postoperatively. To our knowledge, this is the first shared decision-making support tool of its kind in Denmark. PROPOSE was tested by surgeons in a real-world clinical setting. Predictive performance was on par with our findings during model development and testing. The Walking distance model demonstrated excellent discriminatory performance with an AUC of 0.88 and a Brier score of 0.14.

The lowest discriminatory performance was seen with the VAS leg pain model with an AUC of 0.67 CI (57.2; 76.1) and a Brier score of 0.22. The difficulty of predicting leg pain outcome may be related to variability in individual day-to-day activities. Leg pain is often aggravated by standing, walking or lumbar extension. Therefore, VAS leg scores reported at baseline could probably differ considerably depending on recent physical activities.

A deciding factor when choosing the predictive ML algorithm MARS was intelligibility and ease of implementation during the development of PROPOSE. More advanced ML methods such as deep learning neural networks, support vector machines, and random forests are well suited

Table 4
Performance of predicted outcome measures – PROPOSE validation.

Metrics	EQ-5D-3L	ODI	VAS leg pain	VAS back pain	Walking distance
Number of patients, (n)	158	133	157	157	144
AUC, (CI)	81.3 (74.2; 88.4)	79.0 (71.2; 86.7)	66.6 (57.2; 76.1)	75.9 (21.0; 35.1)	88.3 (82.1; 94.6)
Sensitivity, %	77.6	69.3	72.4	84.2	80.0
Specificity, %	73.3	67.2	57.7	60.5	78.0
Classification error, % (CI)	24.0 (17.0; 31.0)	23.0 (24.0; 39.0)	32.5 (25.2; 39.8)	28.0 (21.0; 35.1)	20. (14.0; 27.0)
Accuracy, %	75.9	68.4	67.5	72.0	79.0
Precision, %	82.6	73.2	77.6	66.7	90.0
Recall, %	77.6	69.3	72.4	84.2	80%
F-score	0.80	0.71	0.75	0.74	0.85
MCC	0.50	0.36	0.29	0.46	0.53
Youden's J	0.51	0.37	0.30	0.45	0.57
Cohen's kappe	0.48	0.39	0.48	0.41	0.58
Brier score	0.17	0.21	0.22	0.19	0.14
PPV, %	77.5	69.3	72.4	84.2	79.8
NPV, %	73.3	67.2	57.7	60.5	77.5

Abbreviations: AUC, area under the curve; CI, confidence interval; MCC, Matthews correlation coefficient; PPV, positive predicted value; NPV, negative predicted value.

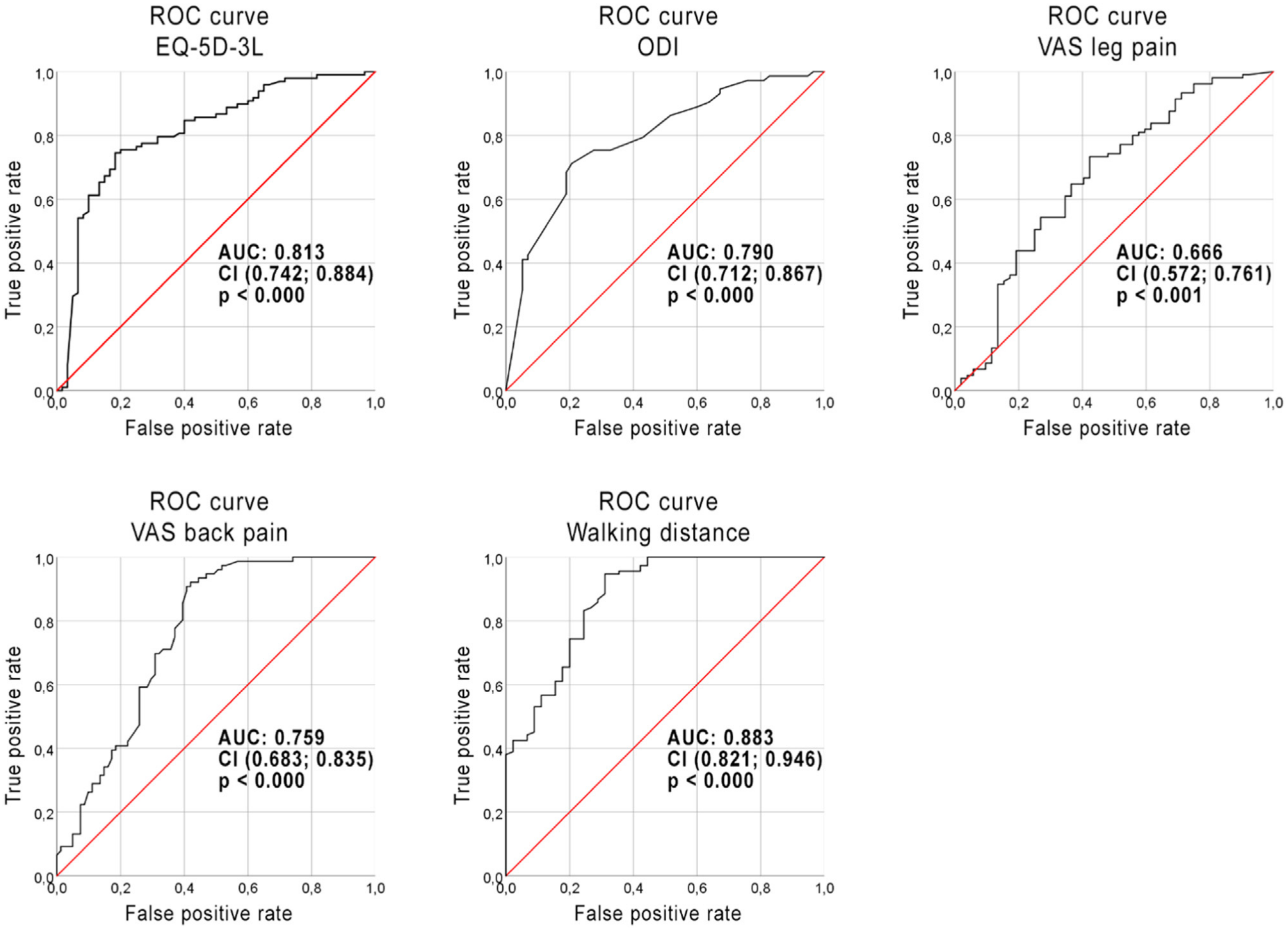

Fig. 3. ROC curves for PROPOSE.

for handling high dimensional data, less vulnerable to missingness and are claimed to outperform traditional algorithms [21,22]. This may not always be the case. A systematic review by Christodoulou et al. [23] of 71 studies on clinical prediction found no evidence of superior performance of advanced ML algorithms over logistic regression (LR). In a previous study conducted on DaneSpine data we reached the same conclusion after comparing the predictive abilities of LR and MARS to several advanced algorithms, with MARS performing slightly better than LR in most cases [24]. It is not inconceivable that alternative algorithms including advanced ML techniques could enhance future predictions depending on added variables. However, one of the shortcomings of many advanced ML algorithms are their lack of transparency. It is often not possible to comprehend their complex inner workings, commonly referred to as "black boxes" [25]. This makes it difficult for surgeons to communicate to the patient what caused a particular prediction.

Limitations

Although PROPOSE performed on par with results from model development tests, the study is limited by the small sample size used to validate the tool. Furthermore, validation should preferably be done on an external data source unrelated to model development (eg, other spine centers). For this reason, PROPOSE may not generalize well to patients outside the spine centers participating in this study.

Missingness in the development data could be a source of selection bias. Although the data validity of the DaneSpine registry appears to be largely unaffected by loss of follow-up [26], there is still a subset of patients with incomplete data at baseline.

In this study MCID was used to dichotomize outcome. Alternative methods such as minimal clinically important improvement (MCII) [27] could lead to different results. Finally, in determining cut-off values for MCIDs we deemed sensitivity and specificity as equally important. This might not reflect patient's preferences. Ideally, threshold setting should take both prevalence and the risk associated with misclassification into account [28].

Besides, decision curve analysis the arguably most important part of the performance evaluation of the predictive model is lacking [29]. Hence, the true net benefit of PROPOSE is unknown.

Lastly, any benefit of the predictive model PROPOSE will depend on the users. If performance expectancy (the degree to which an individual believes that using a new technology will help them attain gains in job performance) or usability is non-satisfactory or the use of a predictive model like PROPOSE is not advocated by the leadership in all probability it will not be used in the clinical setting [30].

Future research

Introducing new variables to model development may enhance predictive performance, for example, educational level which is usually associated with better health outcome. Carefully tuning the default binary classification threshold of 0.5 could yield better balanced prediction of classes [31]. Handling missing data by multiple imputation [31] thus increasing available data might produce slightly better results. The clinical utility of tools like PROPOSE cannot be evaluated by their predictive performance alone. Therefore, it would be beneficial to quantify the net benefits of PROPOSE before being introduced into clinical practice. This can only be achieved by decision curve analysis [32]. For ease of maintenance and future implementation it would be desirable to rewrite PROPOSE into a platform independent web-application.

Conclusion

We have developed a clinical prediction tool to estimate individualized likely improvements in quality of life, leg and back pain, functional disability and walking distance following surgery for lumbar spinal stenosis. PROPOSE works in a real-world clinical setting as a proof of concept and demonstrates acceptable performance. If thoroughly externally validated on a sufficient large sample, it may have the potential of aiding shared decision making between patients and surgeons when discussing treatment options.

Declarations of competing interests

One or more of the authors declare financial or professional relationships on ICMJE-NASSJ disclosure forms.

Funding

There are no financial conflicts of interest to disclose.

References

[1] Andersen M, Nielsen M, Bech-Azeddine R, Helmig P ES. Danish society of spinal surgery. DaneSpine. Yearly report 2021. 2022.

[2] Dane Spine. The Danish national database for spinal surgery. Middelfart, Denmark: Data extraction; 2022.

[3] Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (Sf-36): I. conceptual framework and item selection. Med Care 1992;30:473–83. doi:10.1097/00005650-199206000-00002.

[4] EuroQol GroupEuroQol–a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.

[5] Fairbank JCT, Pynsent PB. The Oswestry Disability Index. Spine (Phila Pa 1976) 2000;25:2940–53. doi:10.1097/00007632-200011150-00017.

[6] Price DD, McGrath PA, Rafii A, Buckingham B. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain 1983;17:45–56. doi:10.1016/0304-3959(83)90126-4.

[7] Froud R, Abel G. Using ROC curves to choose minimally important change thresholds when sensitivity and specificity are valued equally: the forgotten lesson of pythagoras. Theoretical considerations and an example application of change in health status. PLoS One 2014;9:1–11. doi:10.1371/journal.pone.0114468.

[8] Parai C, Hägg O, Lind B, Brisby H. The value of patient global assessment in lumbar spine surgery: an evaluation based on more than 90,000 patients. Eur Spine J 2018;27:554–63. doi:10.1007/S00586-017-5331-0.

[9] Rubin DB. Inference and missing data. Biometrika 1976;63:581–92. doi:10.1093/BIOMET/63.3.581.

[10] Ramaswamy S, Rastogi R, Shim K, Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. In: Proc. 2000 ACM SIGMOD Int. Conf. Manag. data - SIGMOD '00. New York, New York, USA: ACM Press; 2000. p. 427–38. vol. 29. doi:10.1145/342009.335437.

[11] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority oversampling technique. J Artif Intell Res 2002;16:321–57. doi:10.1613/jair.953.

[12] Mierswa, Ingo, Klinkenberg R. Rapid miner, RapidMiner Studio 9.1.0 2019. https://rapidminer.com/products/studio/. Accessed November 11, 2022.

[13] Friedman JH. Multivariate adaptive regression splines. Ann Stat 1990;199:1–67. doi:10.1214/aos/1176347963.

[14] Hastie T, Tibshirani R, Friedman JH, Jerome H. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2009.

[15] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27:1226–38. doi:10.1109/TPAMI.2005.159.

[16] Core Team R. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019 https://www.r-project.org/ Accessed 11 November 2022.

[17] Milborrow S. earth: multivariate adaptive regression splines 2019. https://cran.r-project.org/package=earth.

[18] Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27:861–74. doi:10.1016/J.PATREC.2005.10.010.

[19] Hilden J, Habbema JDF, Bjerregaard B. The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. Methods Inf Med 1978;17(4):227–37.

[20] Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med 1997;16:965–80 16:9<965::aid-sim509>3.0.co;2-o. doi:10.1002/(sici)1097-0258(19970515).

[21] Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:1–2. doi:10.1038/S41746-018-0029-1.

[22] Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319:1317. doi:10.1001/jama.2017.18391.

[23] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12–22. doi:10.1016/j.jclinepi.2019.02.004.

[24] 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. Glob Spine J 2022;12:866–76. doi:10.1177/2192568220967643.

[25] Adler P, Falk C, Friedler SA, et al. Auditing black-box models for indirect influence. Knowl Inf Syst 2018;54:95–122. doi:10.1007/s10115-017-1116-3.

[26] Højmark K, Støttrup C, Carreon L, Andersen MO. Patient-reported outcome measures unbiased by loss of follow-up. Single-center study based on DaneSpine, the Danish spine surgery registry. Eur Spine J 2016;25:282–6. doi:10.1007/s00586-015-4127-3.

[27] Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003;41:582–92. doi:10.1097/01.MLR.0000062554.74615.4C.

[28] Smits N. A note on Youden's J and its cost ratio. BMC Med Res Methodol 2010;10:89. doi:10.1186/1471-2288-10-89.

[29] Jaderesic M, Baker FB. Predicting complications of spine surgery: external validation of three models. Spine J 2022;22:1801–10.

[30] Eiskjær S, Pedersen CF, Skov ST, Andersen MØ. Usability and performance expectancy govern spine surgeons' use of a clinical decision support system for shared decision-making on the choice of treatment of common lumbar degenerative disorders. Front. Digit. Health 2023;5:1225540. doi:10.3389/fdgth.2023.1225540.

[31] Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from imbalanced data sets. Cham, Switzerland: Springer Int Publ; 2018.

[32] Rombach I, Gray AM, Jenkinson C, Murray DW, Rivero-Arias O. Multiple imputation for patient reported outcome measures in randomised controlled trials: Advantages and disadvantages of imputing at the item, subscale or composite score level. BMC Med Res Methodol 2018;18:87. doi:10.1186/s12874-018-0542-6.