Aalborg Universitet

AALBORG UNIVERSITY

Applied Machine Learning for Spine Surgeons

Predicting Outcome for Patients Undergoing Treatment for Lumbar Disc Herniation Using PRO Data

Pedersen, Casper Friis; Andersen, Mikkel Østerheden; Carreon, Leah Yacat; Eiskjær, Søren

AO
SPINE

Original Article

Applied Machine Learning for Spine Surgeons: Predicting Outcome for Patients Undergoing Treatment for Lumbar Disc Herniation Using PRO Data

Casper Friis Pedersen, MSSc[1,2] , Mikkel Østerheden Andersen, MD[1,2] ,
Leah Yacat Carreon, MD, MSc[1,2] , and Søren Eiskjær, MD[3]

Abstract

Study Design: Retrospective/prospective study.

Objective: Models based on preoperative factors can predict patients' outcome at 1-year follow-up. This study measures the performance of several machine learning (ML) models and compares the results with conventional methods.

Methods: Inclusion criteria were patients who had lumbar disc herniation (LDH) surgery, identified in the Danish national registry for spine surgery. Initial training of models included 16 independent variables, including demographics and presurgical patient-reported measures. Patients were grouped by reaching minimal clinically important difference or not for EuroQol, Oswestry Disability Index, Visual Analog Scale (VAS) Leg, and VAS Back and by their ability to return to work at 1 year follow-up. Data were randomly split into training, validation, and test sets by 50%/35%/15%. Deep learning, decision trees, random forest, boosted trees, and support vector machines model were trained, and for comparison, multivariate adaptive regression splines (MARS) and logistic regression models were used. Model fit was evaluated by inspecting area under the curve curves and performance during validation.

Results: Seven models were arrived at. Classification errors were within $\pm$1% to 4% SD across validation folds. ML did not yield superior performance compared with conventional models. MARS and deep learning performed consistently well. Discrepancy was greatest among VAS Leg models.

Conclusions: Five predictive ML and 2 conventional models were developed, predicting improvement for LDH patients at the 1-year follow-up. We demonstrate that it is possible to build an ensemble of models with little effort as a starting point for further model optimization and selection.

Keywords

lumbar disc herniation, machine learning, predictive, deep learning, neural network, artificial intelligence, PRO, patient-reported outcomes

Introduction

For the past decade, various advanced techniques in predictive analytics commonly known as *machine learning* (ML) have gained interest in orthopaedics and medicine at large.[1] The effectiveness of ML compared with more traditional methods has been well demonstrated in solving classification problems, especially in medical image analysis, but as of yet has not been widely adopted by spine surgeons.[2,3] The increasing accumulation of health data leaves a gap between available data and actual data use. ML might leverage the use of large amounts of

[1] Lillebaelt Hospital, Middelfart, Denmark
[2] University of Southern Denmark, Odense, Denmark
[3] Aalborg University, Denmark

Corresponding Author:
Casper Friis Pedersen, Center for Spine Surgery and Research, Spinecenter of Southern Denmark, Lillebaelt Hospital, Oestre Hougvej 55, DK-5500 Middelfart, Denmark.
Email: casper.friis.pedersen@rsyd.dk

health data and accelerate the development of predictive, preventive, or personalized medicine.[4] However, the application of ML algorithms has traditionally required specific programming skills not readily available among medical professionals. The reliance on data scientists to explore health data may be an obstacle to widespread use of ML. The advent of modern visual code-free software platforms might put ML in the hands of surgeons.[5] The purpose of this study was to assess the feasibility of predicting outcomes following lumbar discectomy by comparing 5 ML methods using a modern visual data science software platform, RapidMiner Studio.[6] Results were compared with 2 conventional learning algorithms. To demonstrate differences and similarities, these various methods were applied to single-center retrospective registry data collected on patients following lumbar discectomy.

Methods

Patient Population and Data Source

Patients who had surgery for lumbar disc herniation (LDH) at a single center from 2010 to 2016 were identified in the Danish national registry for spine surgery (DaneSpine). Of 3216 patients identified, 1988 had complete baseline and 1-year follow-up patient-reported outcomes (PRO) data. In all, 20 patients had extreme body mass index (BMI) values (>80 kg/m^2) likely because of erroneous data entry and were excluded. A total of 1968 patients were included in this study.

Study Variables

Two primary health outcome variables were chosen for assessment: EuroQol (EQ-5D)[7] and the Oswestry Disability Index (ODI).[8] In addition, back and leg pain on the Visual Analog Scale (VAS; 0-100)[9] and patients' ability to return to work were included. In accordance with generally accepted practice when solving classification problems, outcome variables were binary coded as either success or nonsuccess.[10] With the exception of return to work, success was defined as achievement of minimal clinically important difference (MCID). MCID thresholds were arrived at using the anchor-based receiver operating characteristic curve (ROC) method.[11] For both EQ-5D and ODI, 1-year postoperative response to the Short Form-36 Health Transition Item (Item 2) that asks the question, "Compared to one year ago, how would you rate your health in general now?"[12] was used as the anchor. Possible answers are "much better," "somewhat better," "about the same," or "somewhat worse" or "much worse." Cutoff for success/non-success was set between patients who responded "somewhat better" versus "about the same." For VAS back and VAS leg pain, responses to the Global Assessment questions at 1 year postoperatively were used: "How is your back pain today compared with before the operation" and "How is your leg pain/sciatica today, compared with before the operation?"

Responses were "completely gone," "much better," "somewhat better," "unchanged," or "worse." Cutoff for success/nonsuccess was set between patients who responded "somewhat better" versus "unchanged."

In determining MCID thresholds, sensitivity and specificity were valued equally, and cutoff points were estimated from the coordinates of the ROC curves using the sum-of-squares approach. The smallest sum of squares of 1-sensitivity and 1-specificity identifies the point closest to the top-left corner in the ROC diagram space.[13] A wide variety of preoperative factors were selected as possible predictors, including gender, age, smoking status, level of pain, walking distance, and health-related measures (Table 1).

Statistical Analysis and Data Handling

During data preparation, distance-based outlier detection was applied to identify and remove extreme cases, which were assumed to be erroneous data entries.[14] Less than 1% were identified as outliers and removed case wise from the data set. The resulting data were randomly split into a training, validation, and test set by a 50%/35%/15% ratio. Class imbalances ranging from 60% to 78%/40% to 22% were present in the target outcome measures. Many ML classification algorithms are sensitive to imbalanced data and have poor accuracy for the infrequent class.[15] To ensure optimal class performance of the models, synthetic minority oversampling was applied to the training and validation data sets.[16] The test set (holdout data) was left untouched. All data preparation was done in RapidMiner (Figure 1).

For each outcome measure, 5 popular ML models were trained: deep learning, decision trees, random forest, boosted trees, and support vector machine (SVM). For comparison, 2 conventional types of models—logistic regression (LR) and multivariate adaptive regression splines (MARS)—were trained. With the exception of MARS, all modeling was performed in RapidMiner Studio 9.3.001 using the "Auto-model" feature. Further tuning was done adjusting model hyperparameters during validation. Auto-model does not support full cross-validation. Instead, performance is evaluated for 7 disjoint subsets of the validation data. The largest and the highest performance are removed, and the average of the remaining 5 performances are reported. MARS models were built and trained using R version 3.5.3 and the CRAN package earth.[17,18] MARS models were tuned by manipulating model complexity (number of basic functions) and degree of interactions using the earth functions *nk* and *degree*. K-fold cross-validation was performed using the functions *nfold* (number of cross-validation folds) and *ncross* (number of cross-validations performed). In both software packages, model fit was evaluated during training by inspecting area under the curve (AUC) curves and the mean performance and SD of performance across validation folds. Final validation of all models was done by applying them to the test data set.

Table 1. Descriptive Statistics of Study Sample: Baseline Patient Characteristics.

Variable	Overall, n = 1988	Δ EQ-5D		Δ ODI		Δ VAS Leg	
		MCID, n = 1262	Non-MCID, n = 726	MCID, n = 1109	Non-MCID, n = 750	MCID, n = 1240	Non-MCID, n = 721
Demographic							
Gender, female, percentage (n)	47 (931)	47 (598)	46 (333)	51 (525)	46 (422)	47 (686)	47 (305)
Gender, male, percentage (n)	53 (1057)	53 (664)	54 (393)	49 (514)	54 (503)	53 (782)	53 (338)
Age, mean ± SD (years)	49.4 ± 14.1	49.6 ± 14.5	49.1 ± 13.4	50.1 ± 14.4	48.8 ± 14.0	49.3 ± 14.3	49.1 ± 13.8
Height, mean ± SD (cm)	175.3 ± 9.6	175.1 ± 9.6	175.5 ± 9.6	174.7 ± 9.8	175.3 ± 9.6	175.4 ± 9.6	174.9 ± 9.6
Weight, mean ± SD (kg)	81.9 ± 15.8	81.0 ± 15.7	83.4 ± 15.9	80.1 ± 15.9	83.2 ± 16.0	81.3 ± 15.8	83.1 ± 16.2
BMI, mean ± SD (kg/m²)	26.5 ± 4.3	26.3 ± 4.2	27.0 ± 4.4	26.1 ± 4.3	27.0 ± 4.3	26.3 ± 4.2	27.1 ± 4.5
Smoker, percentage (n)	30 (987)	30 (374)	30 (213)	29 (299)	30 (272)	29 (455)	31 (171)
Welfare recipient, percentage (n)	28 (557)	29 (367)	26 (190)	31 (327)	27 (247)	28 (407)	29 (189)
Sick-leave, no, percentage (n)	22 (311)	20 (178)	26 (133)	20 (144)	26 (167)	23 (241)	21 (96)
Sick leave, back related, percentage (n)	76 (1072)	79 (700)	72 (372)	79 (582)	71 (466)	75 (783)	76 (351)
Sick leave, other reason percentage (n)	2 (24)	1 (13)	2 (11)	1 (10)	3 (18)	2 (15)	3 (13)
Presurgical health/PROM status							
EQ-5D baseline, median (IQR)	0.260 (0.618)	0.101 (0.380)	0.656 (0.175)	0.101 (0.532)	0.587 (0.532)	0.159 (0.623)	0.475 (0.591)
VAS leg pain baseline, median (IQR)	74 (31)	79 (27)	62 (36)	80 (24)	62 (41)	77 (26)	57 (51)
VAS back pain baseline, median (IQR)	50 (50)	50 (51)	46 (48)	51 (53)	47 (47)	50 (51)	48 (49)
Walking distance, <100 m, percentage (n)	41 (814)	51 (638)	24 (176)	54 (561)	24 (225)	45 (663)	29 (185)
Walking distance, 100-500 m, percentage (n)	26 (522)	24 (296)	31 (226)	24 (247)	30 (275)	25 (363)	31 (200)
Walking distance, 0.5-1 km, percentage (n)	14 (275)	11 (141)	19 (134)	10 (107)	19 (174)	13 (189)	17 (108)
Walking distance, >1 km, percentage (n)	19 (368)	14 (181)	26 (187)	12 (120)	27 (245)	17 (249)	23 (144)
Pain does not prevent me walking, percentage (n)	18 (357)	14 (170)	26 (187)	9 (92)	29 (264)	16 (237)	23 (145)
Pain prevents me from walking >1 km, percentage (n)	24 (466)	20 (246)	31 (220)	19 (201)	30 (277)	22 (325)	28 (176)
Pain prevents me from walking >500 m, percentage (n)	24 (482)	23 (295)	26 (187)	24 (246)	25 (232)	23 (342)	27 (168)
Pain prevents me from walking >100 m, percentage (n)	23 (449)	28 (354)	13 (95)	31 (318)	12 (108)	26 (378)	15 (93)
I can only walk using a stick or crutches, percentage (n)	6 (119)	8 (100)	3 (19)	9 (94)	3 (26)	6 (93)	5 (33)
I am in bed most of the time, percentage (n)	5 (95)	7 (88)	1 (7)	8 (84)	1 (9)	6 (84)	2 (16)
Duration of leg pain, no leg pain, percentage (n)	1 (17)	1 (6)	1 (6)	0 (0)	2 (14)	0 (0)	3 (17)
Duration of leg pain, <3 months, percentage (n)	28 (546)	32 (404)	20 (142)	33 (346)	19 (180)	29 (429)	23 (151)
Duration of leg pain, 3-12 months, percentage (n)	50 (1001)	49 (620)	53 (381)	50 (516)	50 (463)	52 (761)	46 (296)
Duration of leg pain, 1-2 years, percentage (n)	11 (219)	9 (121)	13 (98)	9 (94)	14 (128)	11 (158)	12 (79)
Duration of leg pain, >2 years, percentage (n)	10 (204)	9 (110)	12 (94)	8 (81)	15 (139)	8 (119)	16 (100)
Duration of back pain, no back pain, percentage (n)	8 (153)	7 (90)	8 (63)	7 (72)	8 (73)	8 (116)	8 (49)
Duration of back pain <3 months, percentage (n)	16 (316)	19 (245)	10 (71)	21 (216)	10 (96)	17 (243)	14 (89)
Duration of back pain, 3-12 months, percentage (n)	40 (802)	43 (537)	37 (265)	42 (439)	37 (338)	43 (631)	34 (221)
Duration of back pain, 1-2 years, percentage (n)	11 (225)	10 (125)	14 (100)	10 (99)	13 (121)	11 (168)	11 (71)
Duration of back pain, >2 years, percentage (n)	25 (488)	21 (264)	31 (224)	20 (211)	32 (295)	21 (307)	33 (212)
I am already employed, percentage (n)	12 (230)	9 (108)	17 (122)	8 (82)	14 (130)	12 (170)	11 (72)
I expect to return to full-time work, percentage (n)	44 (861)	49 (605)	36 (256)	48 (499)	36 (328)	46 (678)	37 (232)
I expect to return to part-time work, percentage (n)	7 (141)	6 (77)	9 (64)	6 (65)	9 (83)	6 (84)	11 (67)
I expect to change work, percentage (n)	8 (161)	7 (89)	10 (72)	6 (66)	11 (100)	8 (116)	10 (61)
I expect to remain on sick leave, percentage (n)	3 (65)	2 (27)	5 (38)	2 (15)	6 (52)	2 (33)	5 (35)
I expect to remain retired, percentage (n)	26 (509)	27 (342)	23 (167)	30 (304)	24 (219)	26 (376)	26 (166)

(continued)

Table 1. (continued)

Variable	Overall, n = 1988	Δ VAS Back		Return to work	
		MCID, n = 1468	Non-MCID, n = 643	Yes, n = 1026	No, n = 282
Demographic					
Gender, female, percentage (n)	47 (931)	47 (608)	46 (382)	42 (427)	52 (147)
Gender, male, percentage (n)	53 (1057)	53 (672)	54 (445)	58 (599)	48 (135)
Age, mean ± SD (years)	49.4 ± 14.1	49.7 ± 14.3	48.3 ± 13.8	43.8 ± 10.4	40.8 ± 10.2
Height, mean ± SD (cm)	175.3 ± 9.6	175.2 ±	175.3 ± 9.5	177.1 ± 9.5	175.7 ± 9.8
Weight, mean ± SD (kg)	81.9 ± 15.8	81.3 ± 16.0	82.8 ± 15.7	82.9 ± 15.8	85.0 ± 16.5
BMI, mean ± SD (kg/m^2)	26.5 ± 4.3	26.4 ± 4.3	26.9 ± 4.3	26.3 ± 4.2	27.5 ± 4.8
Smoker, percentage (n)	30 (987)	30 (379)	30 (247)	27 (274)	44 (121)
Welfare recipient, percentage (n)	28 (557)	30 (380)	25 (209)	1 (15)	4 (10)
Sick leave, no, percentage (n)	22 (311)	24 (219)	19 (117)	22 (199)	8 (21)
Sick leave, back related, percentage (n)	76 (1072)	75 (666)	78 (471)	77 (701)	90 (239)
Sick leave, other reason, percentage (n)	2 (24)	1 (9)	3 (19)	1 (10)	2 (5)
Presurgical health/PROM status					
EQ-5D baseline, median (IQR)	0.260 (0.618)	0.189 (0.623)	0.362 (0.634)	0.364 (0.635)	0.159 (0.590)
VAS leg pain baseline, median (IQR)	74 (31)	75 (27)	67 (37)	71 (33)	74 (29)
VAS back pain baseline, median (IQR)	50 (50)	61 (34)	22 (44)	46 (49)	62 (36)
Walking distance, <100 m, percentage (n)	41 (814)	36 (299)	41 (852)	37 (382)	35 (96)
Walking distance, 100-500 m, percentage (n)	26 (522)	28 (229)	26 (557)	23 (232)	34 (95)
Walking distance, 0.5-1 km, percentage (n)	14 (275)	15 (125)	14 (298)	15 (149)	17 (46)
Walking distance, >1 km, percentage (n)	19 (368)	21 (169)	19 (391)	25 (252)	14 (39)
Pain does not prevent me from walking, percentage (n)	18 (357)	17 (211)	21 (169)	25 (256)	14 (38)
Pain prevents me from walking >1 km, percentage (n)	24 (466)	24 (302)	24 (195)	25 (252)	27 (74)
Pain prevents me from walking >500 m, percentage (n)	24 (482)	24 (306)	25 (205)	21 (218)	31 (87)
Pain prevents me from walking >100 m, percentage (n)	23 (449)	24 (304)	21 (168)	19 (191)	20 (54)
I can only walk using a stick or crutches, percentage (n)	6 (119)	6 (81)	5 (44)	5 (47)	5 (14)
I am in bed most of the time, percentage (n)	5 (95)	5 (68)	4 (33)	5 (47)	3 (8)
Duration of leg pain, no leg pain, percentage (n)	1 (17)	1 (7)	1 (11)	1 (11)	1 (4)
Duration of leg pain, <3 months, percentage (n)	28 (546)	28 (352)	27 (224)	31 (313)	17 (48)
Duration of leg pain, 3-12 months, percentage (n)	50 (1001)	51 (658)	49 (402)	50 (512)	51 (141)
Duration of leg pain, 1-2 years, percentage (n)	11 (219)	10 (132)	13 (104)	10 (106)	15 (41)
Duration of leg pain, >2 years, percentage (n)	10 (204)	10 (130)	10 (86)	8 (79)	16 (44)
Duration of back pain, no back pain, percentage (n)	8 (153)	1 (21)	17 (141)	9 (87)	4 (12)
Duration of back pain, <3 months, percentage (n)	16 (316)	18 (225)	13 (107)	17 (173)	13 (36)
Duration of back pain, 3-12 months, percentage (n)	40 (802)	46 (583)	33 (267)	43 (442)	37 (101)
Duration of back pain, 1-2 years, percentage (n)	11 (225)	11 (144)	11 (93)	11 (111)	13 (37)
Duration of back pain, >2 years, percentage (n)	25 (488)	24 (305)	26 (217)	20 (207)	33 (91)
I am already employed, percentage (n)	12 (230)	11 (139)	13 (103)	19 (195)	4 (11)
I expect to return to full-time work, percentage (n)	44 (861)	45 (576)	41 (336)	68 (691)	40 (109)
I expect to return to part-time work, percentage (n)	7 (141)	7 (89)	8 (63)	6 (58)	14 (38)
I expect to change work, percentage (n)	8 (161)	7 (90)	11 (87)	6 (58)	27 (73)
I expect to remain on sick leave, percentage (n)	3 (65)	3 (33)	4 (35)	1 (7)	14 (38)
I expect to remain retired, percentage (n)	26 (509)	27 (343)	23 (192)	0 (0)	1 (4)

Abbreviations: BMI, body mass index; EQ-5D; EuroQol; IQR, interquartile range; MCID, minimal clinically important difference; ODI, Oswestry Disability Index; PROM, Patient Reported Outcome Measures; VAS, Visual Analog Scale.
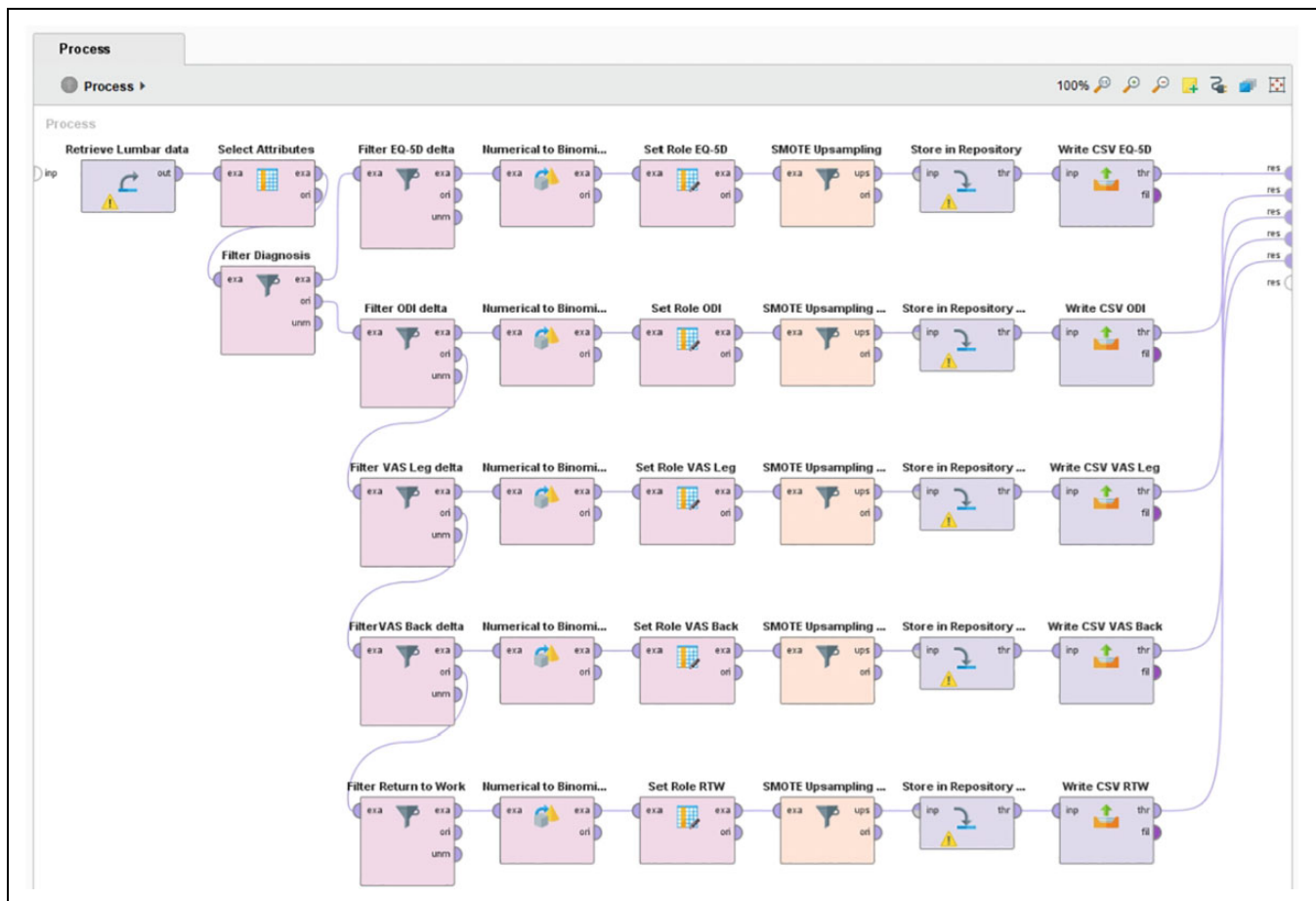
Figure 1. RapidMiner process for minority class upsampling of outcome measures using SMOTE.
Abbreviations: CSV, Comma-separated values; EQ-5D; EuroQol; ODI, Oswestry Disability Index; RTW, return to work; VAS, Visual Analog Scale.

Machine Learning

In this study, ML is simply regarded as the study on how computers can learn to solve problems without being explicitly programmed.[19] More formally, ML can be stated as methods that can automatically detect patterns in data. The uncovered patterns are used to predict future data or other outcomes of interest. ML is typically divided into 2 main types: supervised and unsupervised learners. In the supervised approach, the algorithms learn from labeled example data. In the unsupervised approach, no corresponding output variables are presented. Algorithms are left to their own to discover patterns in the data. The ML methods used in this study are supervised.[20]

Predictive Algorithms

LR is a highly popular method for solving classification problems and has been the de facto standard in many fields, including health care, for several decades. LR follows the same principles applied in general linear regression. However, because the outcome is binary, the mean of the regression must fall between 0 and 1. This is satisfied by using a logit model and assuming a binomial distribution.[21,22]

MARS is an advanced form of regression that extends the capabilities of standard linear regression by automatically modeling nonlinearities and interactions between variables. The MARS algorithm adapts to nonlinearities by piecewise fitting together smaller localized linear models that are defined pairwise.[23,24]

Deep learning belongs to a class of ML methods based on deep artificial neural networks (ANNs). An ANN is a simplistic representation of the functioning of a biological brain. The neural network consists of highly interconnected artificial neurons called nodes organized in several processing layers. Like synapses in a brain, the connections allow nodes to transmit signals to other nodes. Each node is initially assigned with a random numeric weight for each of its incoming connections. When a node receives signals from other nodes, the strength of the signals is adjusted by the associated weights. The resulting numbers are then summed and passed through a simple nonlinear function, which produces the output.[25-27]

Decision trees are models for classification and regression. A decision tree is structured like a flowchart resembling a tree. It learns by processing input from the top (root) and splitting data into increasingly smaller subsets following an if-then-else decision logic. Splits are made by decision nodes followed by

Table 2. EQ-5D Index Score Performance of the Final Machine Learning, MARS, and Logistic Regression Models Assessed on the Holdout Data Set.

Performance metrics	Logistic regression	MARS	Deep learning	Decision tree	Random forest	Boosted trees	SVM
AUC	0.84	0.84	0.81	0.77	0.81	0.80	0.89
Accuracy	79%	79%	78%	78%	76%	75%	77%
Sensitivity	70%	70%	68%	67%	66%	65%	67%
Specificity	84%	84%	85%	85%	84%	81%	84%
PPV	83%	82%	80%	79%	78%	79%	79%
NPV	71%	72%	76%	76%	74%	69%	75%
F1 score	0.83	0.83	0.82	0.82	0.80	0.80	0.82
MCC	0.54	0.54	0.54	0.53	0.50	0.47	0.53

Abbreviations: AUC, area under the curve; EQ-5D; EuroQol; F1 score, measure for harmonic mean of precision and sensitivity; MARS, multivariate adaptive regression splines; MCC, Matthews correlation coefficient (a balanced measure sensitive to true and false positives and negatives); NPV, negative predictive value; PPV, positive predictive value (precision); SVM, support vector machine.

another decision node or a leaf (prediction). Splits below nodes are called branches. The decision tree algorithm evaluates all possible splits in each step and multiple thresholds in order to make the most homogeneous subsets.[28]

Random forest is an ensemble method where the training model consists of a multitude of ordinary decision trees. The final model is determined by majority vote when solving classification problems and mean value when tasked with regression. The basic idea is that merged together, the predictions of several trees should be closer to the true value than any single tree.[29,30]

Boosted trees are ensembles of very simple decision trees referred to as weak learners. The model is trained by gradually improving estimations by applying trees in a sequence, where each new tree is optimized for predicting the residuals (errors) of the preceding tree. Optimization is done by an algorithm that favors incorrectly classified predictions by assigning them a higher weight, which forces subsequent trees to adapt to the examples that were incorrectly classified by the previous trees. The idea is that many simple models when combined perform better than 1 complex model.[31]

SVMs are commonly used to classify a data set into 2 classes. It achieves this by finding the line that best separates data points. Support vectors are the data points closest to the dividing line. When a data set is nonlinearly separable, SVM transforms data into a 3D dimension. The dividing line now resembles a plane. Data will be mapped into increasingly higher dimensions until a plane succeeds in segregating the classes.[32,33]

Results

Outcome Measures

MCID cutoff points for the chosen dichotomous outcome measures (target variables) were established as follows: EQ-5D = 0.17; ODI = 18; VAS Back = 10; VAS Leg = 17.

Predictors

During training and validation, feature selection was reduced to the following independent preoperative variables: employment status, BMI, sick leave status, EQ-5D, pain duration (back and leg), VAS pain level (back and leg), walking distance, walking impairment caused by pain, and self-reported expectation to return to work after surgery.

Predictive Modeling

Following training, validation, and optimization, 7 different models were arrived at for each of the 5 selected outcome measures. Classification errors for all models were within $\pm 1\%$ to 4% SD across validation folds. Model performance from the final holdout data set is compared in Tables 2 to 6 and illustrated in Figure 2. Evaluation was done using performance metrics from the resulting confusion matrices, including the Matthews correlation coefficient.[34]

Improvement in EQ-3D at 1 year (positive predictive value [PPV]) was predicted with an average accuracy of 80% (median = 79%; range = 5%). The mean AUC value was 0.82 (median = 0.81; range = 0.12). Nonimprovement (negative predictive value [NPV]) was predicted with an average accuracy of 73% (median = 74%; range = 7%). LR and MARS models performed on par with the best performing ML models.

Improvement in ODI at 1 year (PPV) was predicted by the models with an average accuracy of 69% (median = 71%; range = 12%). The mean AUC value was 0.75 (median = 0.76; range = 0.09). Nonimprovement (NPV) was predicted with an average accuracy of 69% (median = 71%; range = 12%). MARS performed on par with the best performing ML models. LR performed poorly but better than the worst performing ML models.

Improvement in VAS Leg at 1 year (PPV) was predicted by the models with an average accuracy of 67% (median = 66%; range = 12%). The mean AUC value was 0.73 (median = 0.74; range = 0.12). Nonimprovement (NPV) was predicted with an average accuracy of 67% (median = 68%; range = 12%). MARS performed on par with the best ML models. LR did not perform well.

Improvement in VAS Back at 1 year (PPV) was predicted by the models with an average accuracy of 82% (median = 79%; range = 14%). The mean AUC value was 0.81 (median = 0.82;

Table 3. ODI Index Score Performance of the Final Machine Learning, MARS, and Logistic Regression Models Assessed on the Holdout Data Set.

Performance metrics	Logistic regression	MARS	Deep learning	Decision tree	Random forest	Boosted trees	SVM
AUC	0.74	0.79	0.76	0.70	0.71	0.77	0.77
Accuracy	69%	72%	70%	67%	67%	72%	68%
Sensitivity	67%	68%	68%	65%	62%	70%	66%
Specificity	70%	75%	72%	68%	71%	73%	70%
PPV	71%	71%	72%	62%	62%	74%	71%
NPV	65%	72%	67%	71%	72%	69%	66%
F1 score	0.71	0.73	0.72	0.69	0.66	0.74	0.71
MCC	0.37	0.43	0.40	0.33	0.34	0.43	0.37

Abbreviations: AUC, area under the curve; F1 score, measure for harmonic mean of precision and sensitivity; MARS, multivariate adaptive regression splines; MCC, Matthews correlation coefficient (a balanced measure sensitive to true and false positives and negatives); NPV, negative predictive value; ODI, Oswestry Disability Index; PPV, positive predictive value (precision); SVM, support vector machine.

Table 4. VAS Leg Performance of the Final Machine Learning, MARS, and Logistic Regression Models Assessed on the Holdout Data Set.

Performance metrics	Logistic regression	MARS	Deep learning	Decision tree	Random forest	Boosted trees	SVM
AUC	0.65	0.74	0.67	0.75	0.75	0.74	0.78
Accuracy	64%	71%	69%	67%	65%	67%	67%
Sensitivity	43%	51%	48%	46%	45%	45%	45%
Specificity	80%	82%	85%	85%	84%	83%	83%
PPV	66%	74%	69%	64%	62%	66%	68%
NPV	60%	62%	70%	72%	72%	68%	65%
F1 score	0.57	0.55	0.76	0.73	0.72	0.74	0.75
MCC	0.25	0.35	0.35	0.33	0.32	0.31	0.31

Abbreviations: AUC, area under the curve; F1 score, measure for harmonic mean of precision and sensitivity; MARS, multivariate adaptive regression splines; MCC, Matthews correlation coefficient (a balanced measure sensitive to true and false positives and negatives); NPV, negative predictive value; PPV, positive predictive value (precision); SVM, support vector machine; VAS, Visual Analog Scale.

Table 5. VAS Back Performance of the Final Machine Learning, MARS, and Logistic Regression Models Assessed on the Holdout Data Set.

Performance metrics	Logistic regression	MARS	Deep learning	Decision tree	Random forest	Boosted trees	SVM
AUC	0.78	0.83	0.82	0.80	0.82	0.80	0.82
Accuracy	72%	75%	74%	75%	75%	74%	74%
Sensitivity	64%	84%	69%	73%	79%	67%	70%
Specificity	77%	70%	77%	76%	73%	79%	76%
PPV	79%	88%	79%	84%	90%	76%	79%
NPV	61%	63%	66%	62%	54%	70%	66%
F1 score	0.78	0.78	0.78	0.80	80%	0.78	0.78
MCC	0.41	0.52	0.46	0.47	0.47	0.46	0.45

Abbreviations: AUC, area under the curve; F1 score, measure for harmonic mean of precision and sensitivity; MARS, multivariate adaptive regression splines; MCC, Matthews correlation coefficient (a balanced measure sensitive to true and false positives and negatives); NPV, negative predictive value; PPV, positive predictive value (precision); SVM, support vector machine; VAS, Visual Analog Scale.

range = 0.05). Nonimprovement (NPV) was predicted with an average accuracy of 63% (median = 63%; range = 16%). MARS performed on par with the best performing ML models. The performance of the LR model was inferior.

Whether patients were successfully able to return to work at 1 year (PPV) was predicted by the models with an average accuracy of 89% (median = 90%; range = 7%). The mean AUC value was 0.84 (median = 0.84; range = 0.06). Nonsuccess (NPV) was predicted with an average accuracy of 68% (median = 68%; range = 18%). MARS performed on par with

the best ML model. The LR model performed slightly worse than the least successful ML models.

Discussion

The ML methods applied in this study did not yield overall superior performance compared with the conventional methods. Performance of MARS was on par with the best performing ML method, deep learning. MARS and deep learning models performed consistently well across outcome measures.

Table 6. Return to Work Performance of the Final Machine Learning, MARS, and Logistic Regression Models Assessed on the Holdout Data Set.

Performance metrics	Logistic regression	MARS	Deep learning	Decision tree	Random forest	Boosted trees	SVM
AUC	0.81	0.86	0.85	0.81	0.87	0.84	0.84
Accuracy	86%	86%	87%	86%	84%	85%	85%
Sensitivity	61%	61%	63%	62%	55%	59%	59%
Specificity	92%	93%	93%	91%	94%	92%	92%
PPV	91%	90%	90%	92%	85%	89%	89%
NPV	63%	72%	71%	59%	77%	68%	68%
F1 score	0.91	0.91	0.92	0.91	0.89	0.91	0.91
MCC	0.53	0.57	0.59	0.51	0.55	0.54	0.54

Abbreviations: AUC, area under the curve; F1 score, measure for harmonic mean of precision and sensitivity; MARS, multivariate adaptive regression splines; MCC, Matthews correlation coefficient (a balanced measure sensitive to true and false positives and negatives); NPV, negative predictive value; PPV, positive predictive value (precision); SVM, support vector machine.

Mixed results were observed across outcome measures for both LR and other ML models. Discrepancy in performance measures was greatest among the models predicting leg pain improvement. In some cases, ML methods were slightly outperformed by the logistic regression models. One possible explanation for the results may be attributed to the univariate correlations found during model training. This suggests that outcome is linearly related to the severity of patients' health status preoperatively. That is, patients who are worse off tend to improve the most. This is consistent with the findings of Staartjes et al.[2] Previous studies indicate that ML primarily has performance advantages when data exhibit strong interactions between predictors and nonlinearities.[35] The absence of these qualities might explain the failure of ML methods to prove superior in this study.

Transparency Versus Explainability

Parallel to the empirical success of ML methods in recent years, there is a rising concern about their lack of transparency.[36] Complex models such as deep learning are in essence black boxes once they have been trained. Their inner workings are, although accessible, beyond human comprehension and interpretation. Complex models are often more accurate, but less interpretable and vice versa.[37] Ethically, the problem of explainability is especially important in a clinical setting where patient's lives and well-being depend on decision-making.[38] It has been well documented that ML can lead to unforeseen bias and discrimination inherited by the algorithms from either human prejudices or artefacts in the training data.[39] Considerable efforts to bridge accuracy and explainability in ML have already been done but remains in its infancy.[40] From an epistemological point of view, the problem of explainability underlines the basic requirement of science to be able to describe the cause and effect of any given system and align the inputs with any given output. Explanatory and predictive accuracy have different qualities and may be viewed as 2 dimensions that all models possess.[41] This suggests that depending on the purpose, trade-offs between transparency and accuracy should be carefully evaluated in model selection. In short, choosing the simpler model might be preferable.

A Priori Model Selection

Several studies have suggested that simple models often perform just as well as more advanced models.[42-44] In a recent systematic review including 71 studies (Christodoulou et al[45]), the authors found no evidence of superior performance of ML over LR. However, they did not investigate which factors might explain this and recommend that future research should focus on identifying which algorithms are optimal for different types of prediction problems. The above-mentioned findings are in line with the theoretical work of Wolpert,[46] which states that averaged across all possible problems, all learning algorithms will perform equally well because of their inherent inductive bias. According to this theorem, there are no objective a priori reasons to favor any algorithm over any others.[46] In conclusion, despite lacking evidence of the superiority of ML, we suggest that clinical prediction should always explore and compare multiple models.

Limitations

The registry used in this single-center study has previously been demonstrated to be unaffected by loss of follow-up.[47] Still, some degree of selection bias cannot be ruled out. No attempt to evaluate missingness of the initial data set (n = 3216) was made, and it was assumed to be missing at random. Consequently, data were not imputed because the final study sample was relatively large. Application of an imputation technique—for example, multiple imputation by chained equations,[48] would have made a larger source of information available to the models, possibly leading to better results. ML techniques appear to require far more data per variable to achieve stability compared with conventional methods such as LR.[49] Comorbidities were not factored in. Neither were surgical methods or complications. Furthermore, to counter overfitting and reduce redundancy, the feature selection was limited to a small subset. Although this strategy could help improve the robustness of the models, it is possible that more predictors could have resulted in a higher degree of accuracy. Finally, the decision in this study to value sensitivity and specificity equally in determining MCIDs is somewhat arbitrary. Prevalence, severity of the condition, and possible adverse effects from treatment should ideally all be
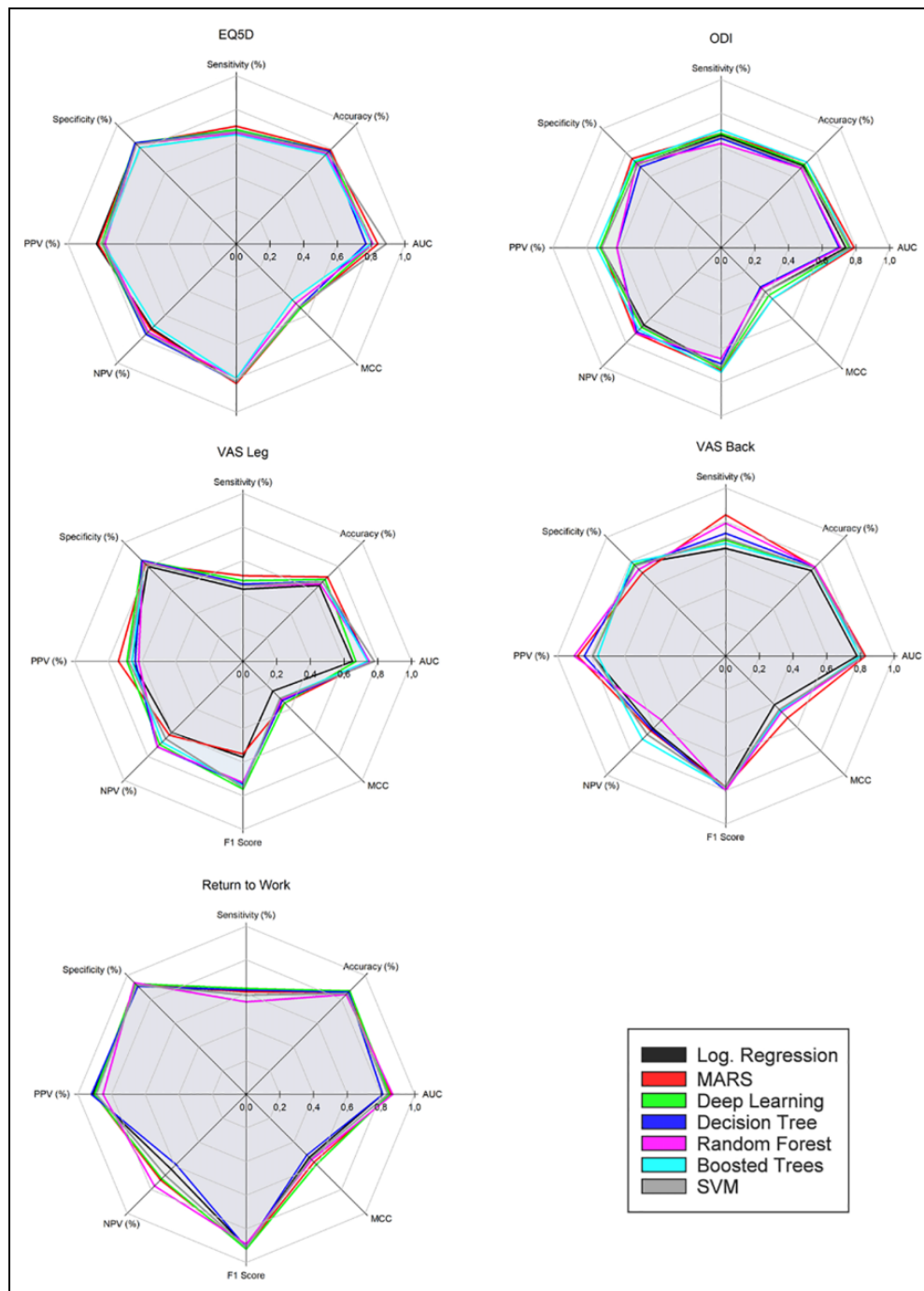
Figure 2. Performance metrics for models: EQ5D, ODI, VAS Leg, VAS Back, Return to Work.
Abbreviations: AUC, area under the curve; EQ-5D; EuroQol; MARS, multivariate adaptive regression splines; MCC, Matthews correlation coefficient; NPV, negative predictive value; ODI, Oswestry Disability Index; PPV, positive predictive value; SVM, support vector machine; VAS, Visual Analog Scale.

considered when deciding an acceptable trade-off between true positives and false positives.[50]

Conclusions

We have developed 5 different ML models and 2 conventional models across 5 outcome measures, predicting improvement for patients following surgery for LDH at 1 year after surgery: EQ-5D, ODI, VAS Leg, VAS Back, and Return to Work. The study demonstrates that it is possible to build and train an ensemble of different predictive ML models with little effort and no programming skills, as a starting point for model comparison and further optimization and development. Modern code-free software like RapidMiner may encourage the use

of PRO data for predictive purposes by surgeons and other medical professionals by eliminating the need for programming skills.

Declaration of Conflicting Interests

Funding

ORCID iD

Casper Friis Pedersen, MSSc 🔟 https://orcid.org/0000-0002-1426-4096
Mikkel Østerheden Andersen, MD 🔟 https://orcid.org/0000-0001-8478-8218
Leah Yacat Carreon, MD, MSc 🔟 https://orcid.org/0000-0002-7685-9036
Søren Eiskjær, MD 🔟 https://orcid.org/0000-0002-6673-0116

References

1. Kim JT. Application of machine and deep learning algorithms in intelligent clinical decision support systems in healthcare. *J Health Med Inform*. 2018;9:1-6. doi:10.4172/2157-7420.1000321

2. 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;19:853-861. doi:10.1016/j.spinee.2018.11.009

3. Cabitza F, Locoro A, Banfi G. Machine learning in orthopedics: a literature review. *Front Bioeng Biotechnol*. 2018;6:75. doi:10.3389/fbioe.2018.00075

4. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018;319:1317-1318. doi:10.1001/jama.2017.18391

5. Van S, Zhang Z, Schmitz M, et al. Scalable predictive analysis in critically ill patients using a visual open data analysis platform. *PLoS One*. 2016;11:e0145791. doi:10.1371/journal.pone.0145791

6. RapidMiner. RapidMiner Studio. *Accessed October 12*, 2020. https://rapidminer.com/products/studio/

7. EuroQol Group. EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy*. 1990;16:199-208.

8. Fairbank JCT, Pynsent PB. The Oswestry disability index. *Spine (Phila Pa 1976)*. 2000;25:2940-2953. doi:10.1097/00007632-200011150-00017

9. 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

10. Greene WH.Functional form and structural change. In: *Econometric Analysis*. 5th ed. Prentice Hall; 2003:116-147.

11. de Vet HCW, Ostelo RWJG, Terwee CB, et al. Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Qual Life Res*. 2007;16:131-142. doi:10.1007/s11136-006-9109-9

12. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (Sf-36): I. Conceptual framework and item selection. *Med Care*. 1992;30:473-483. doi:10.1097/00005650-199206000-00002

13. 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:e114468. doi:10.1371/journal.pone.0114468

14. Ramaswamy S, Rastogi R, Shim K, Ramaswamy S, Rastogi R, Shim K.Efficient algorithms for mining outliers from large data sets. Accessed October 12, 2020. https://webdocs.cs.ualberta.ca/∼zaiane/pub/check/ramaswamy.pdf

15. Drummond C, Holte RC. Severe class imbalance: why better algorithms aren't the answer. Paper presented at: Machine Learning: ECML 2005, 16th European Conference on Machine Learning; October 3-7, 2005; Porto, Portugal. doi:10.1007/11564096_52

16. Fernandez A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *J Artif Intell Res*. 2018;61:863-905. doi:10.1613/jair.1.11192

17. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2020.

18. Milborrow S. earth: Multivariate adaptive regression splines. Accessed October 12, 2020. https://cran.r-project.org/package=earth

19. Koza JR, Bennett FH III, Andre D, Keane MA. Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In: Gero JS, Sudweeks F, eds. *Artificial Intelligence in Design '96*. Springer; 1996:151-170. doi:10.1007/978-94-009-0279-4_9

20. Murphy KP. *Machine Learning: A Probabilistic Perspective*. MIT Press; 2012.

21. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2nd ed. John Wiley & Sons; 2000. doi:10.1002/0471722146

22. Agresti A. *Categorical Data Analysis*. Wiley-Interscience; 2013. Accessed October 14, 2020. https://www.wiley.com/en-us/Categorical+Data+Analysis%2C+3rd+Edition-p-9780470463635

23. Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer; 2009.

24. Friedman JH. Multivariate adaptive regression splines. *Ann Stat*. 1990;199:1-67. doi:10.1214/aos/1176347963

25. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436-444. doi:10.1038/nature14539

26. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2016. Accessed October 14, 2020. https://www.deeplearningbook.org/

27. Rojas R. *Neural Networks: A Systematic Introduction*. Springer-Verlag; 1996.

28. Venkatasubramaniam A, Wolfson J, Mitchell N, Barnes T, Jaka M, French S. Decision trees in epidemiological research. *Emerg Themes Epidemiol*. 2017;14:11. doi:10.1186/s12982-017-0064-4

29. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32. doi:10.1023/A:1010933404324

30. Kuhn M, Johnson K. *Applied Predictive Modeling*. New York, NY: Springer; 2013. doi:10.1007/978-1-4614-6849-3

31. Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Front Neurorobot*. 2013;7:21. doi:10.3389/fnbot.2013.00021

32. Zhang Y. Support vector machine classification algorithm and its application. In: Lu C, Wang L, Yang A, eds. *Information Computing and Applications. ICICA 2012. Communications in Computer and Information Science*. Vol 308. Springer; 2012:179-186. doi:10.1007/978-3-642-34041-3_27

33. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20:273-297. doi:10.1007/BF00994018

34. Tharwat A. Classification assessment methods. *Applied Computing and Informatics*. https://www.researchgate.net/publication/327148996_Classification_Assessment_Methods_a_detailed_tutorial.

35. Steyerberg EW, van der Ploeg T, Van Calster B. Risk prediction with machine learning and regression methods. *Biom J*. 2014;56:601-606. doi:10.1002/bimj.201300297

36. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. *Big Data Soc*. 2016;3. doi:10.1177/2053951716679679

37. Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. *Digit Signal Process*. 2018;73:1-15. doi:10.1016/j.dsp.2017.10.011

38. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare. Paper presented at: 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD'15; August 10-13, 2015; New York, NY. doi:10.1145/2783258.2788613

39. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366:447-453. doi:10.1126/science.aax2342

40. DARPA. Broad Agency announcement: explainable artificial intelligence (XAI). DARPA-BAA-16-53. Published August 10, 2016. Accessed October 12, 2020. https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

41. Shmueli G. To explain or to predict? *Stat Sci*. 2010;25:289-310. doi:10.1214/10-STS330

42. Hand DJ. Classifier technology and the illusion of progress. *Stat Sci*. 2006;21:1-14. doi:10.1214/088342306000000060

43. Nusinovici S, Tham YC, Chak Yan MY, et al. Logistic regression was as good as machine learning for predicting major chronic diseases. *J Clin Epidemiol*. 2020;122:56-69. doi:10.1016/j.jclinepi.2020.03.002

44. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115-118. doi:10.1038/nature21056

45. 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

46. Wolpert DH. The lack of a priori distinctions between learning algorithms. *Neural Comput*. 1996;8:1341-1390. doi:10.1162/neco.1996.8.7.1341

47. 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;25:282-286. doi:10.1007/s00586-015-4127-3

48. Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv Methodol*. 2001;27:85-96.

49. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol*. 2014;14:137. doi:10.1186/1471-2288-14-137

50. 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