The Spine Journal 19 (2019) 853–861

Clinical Study

Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar discectomy: feasibility of center-specific modeling

Victor E. Staartjes, BMed[a,b,e,*], Marlies P. de Wispelaere, MSc[c], William Peter Vandertop, MD, PhD[d], Marc L. Schröder, MD, PhD[a]

[a] *Department of Neurosurgery, Bergman Clinics Amsterdam, Rijksweg 69, 1411 GE Naarden, The Netherlands*
[b] *Amsterdam UMC, Vrije Universiteit Amsterdam, Neurosurgery, Amsterdam Movement Sciences, de Boelelaan 1117, 1081 HV Amsterdam, The Netherlands*
[c] *Department of Clinical Informatics, Bergman Clinics Amsterdam, Rijksweg 69, 1411 GE Naarden, The Netherlands*
[d] *Neurosurgical Center Amsterdam, Amsterdam University Medical Centers, de Boelelaan 1117, 1081 HV Amsterdam, The Netherlands*
[e] *Department of Neurosurgery, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Frauenklinikstrasse 10, 8091 Zurich, Switzerland*

Received 6 September 2018; revised 12 November 2018; accepted 12 November 2018

Abstract

BACKGROUND CONTEXT: There is considerable variability in patient-reported outcome measures following surgery for lumbar disc herniation. Individualized prediction tools that are derived from center- or even surgeon-specific data could provide valuable insights for shared decision-making.
PURPOSE: To evaluate the feasibility of deriving robust deep learning-based predictive analytics from single-center, single-surgeon data.
STUDY DESIGN: Derivation of predictive models from a prospective registry.
PATIENT SAMPLE: Patients who underwent single-level tubular microdiscectomy for lumbar disc herniation.
OUTCOME MEASURES: Numeric rating scales for leg and back pain severity and Oswestry Disability Index scores at 12 months postoperatively.
METHODS: Data were derived from a prospective registry. We trained deep neural network-based and logistic regression-based prediction models for patient-reported outcome measures. The primary endpoint was achievement of the minimum clinically important difference (MCID) in numeric rating scales and Oswestry Disability Index, defined as a 30% or greater improvement from baseline. Univariate predictors of MCID were also identified using conventional statistics.
RESULTS: A total of 422 patients were included (mean [SD] age: 48.5 [11.5] years; 207 [49%] female). After 1 year, 337 (80%), 219 (52%), and 337 (80%) patients reported a clinically relevant improvement in leg pain, back pain, and functional disability, respectively. The deep learning models predicted MCID with high area-under-the-curve of 0.87, 0.90, and 0.84, as well as accuracy of 85%, 87%, and 75%. The regression models provided inferior performance measures for each of the outcomes.
CONCLUSIONS: Our study demonstrates that generating personalized and robust deep learning-based analytics for outcome prediction is feasible even with limited amounts of center-specific data. With prospective validation, the ability to preoperatively and reliably inform patients about the likelihood of symptom improvement could prove useful in patient counselling and shared decision-making.

Keywords: Decision making; Disc herniation; Discectomy; Machine learning; Outcome measures; Sciatica.

FDA device/drug status: Not applicable.

Author disclosures: ***VES:*** Nothing to disclose. ***MPW:*** Nothing to disclose. ***WPV:*** Nothing to disclose. ***MLS:*** Nothing to disclose.

* Corresponding author. Department of Neurosurgery, c/o Bergman Clinics, Naarden, Rijksweg 69, 1411 GE Naarden, The Netherlands. Tel.: 0031 88 900 0500; fax: 0031 88 900 0568.
E-mail address: victor.staartjes@gmail.com (V.E. Staartjes).

https://doi.org/10.1016/j.spinee.2018.11.009
1529-9430/

Introduction

Although discectomy is an effective treatment for lumbar disc herniation (LDH) [1–7], there is still a subset (20%–30%) of patients who do not benefit from surgery. Few reliable radiological or clinical predictors of poor long-term outcomes following discectomy have been identified [5,8], and the integration of a range of known risk factors into a clinical decision is often impractical and quite speculative. Prediction tools for patient-reported outcomes can improve shared decision-making and patient counseling, when deciding to opt for surgery or not, and could possibly even improve cost-effectiveness of LDH treatment [9–13].

Machine learning (ML) techniques have gained interest for predictive analytics in medicine [14]. Various ML techniques such as gradient boosting machines, decision trees, random forests, and simple neural networks have provided increased predictive ability compared to more traditional modeling techniques [14–16]. Through the use of multiple layers of representation, instead of just a single one, deep neural networks have the ability to reduce complex problems and relationships into multiple simpler ones. Deep learning is a particularly effective ML method in medical image analysis, outcome prediction, genomic analysis, and drug discovery [17], but has not been widely applied yet in surgical spinal patient care [15].

Predictive analytics are often derived from multicenter data, with the advantage of large sample sizes and high generalizability. Such models are useful to clinicians around the world, but carry an inherent trade-off in center-specific prediction accuracy, as surgeons in one center will have different patient demographics and selection criteria, surgical technique, complication patterns, and outcomes, than in another [1,18–21]. With the advent of new techniques and big data-driven collection of patient data [22,23], it is conceivable that any center, or surgeon, would be able to train and apply a personalized prediction tool [20,21,23]. Therefore, the objective of this study was to evaluate the possibility of accurately predicting outcomes following lumbar discectomy, based on deep learning, applied to a limited cohort of prospectively collected data from a single center.

Methods

Patient population

From a prospective institutional registry of spinal interventions, we identified all patients who had undergone lumbar microdiscectomy between November 2013 and April 2018 at a single Dutch spine center. All patients underwent single-level tubular microdiscectomy by the senior author (MS) according to a validated protocol [2,24]. Patients were only considered for surgery with radiologically confirmed single-level LDH and failed conservative management for more than 8 weeks. Adult patients who had a complete baseline and 12-months' patient-reported outcomes (PROM) record, as well as ≥80% demographic data completion, were included in this study. This study was devised and reported according to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis statement [25]. All the patients included in the registry provided written informed consent. The prospective registry was authorized by the local institutional review board (Medical Research Ethics Committees United, Registration Number W16.065), and this study was carried out in accordance with the 2013 Declaration of Helsinki.

Data collection

Clinical and radiological baseline data were obtained at the first outpatient visit to the treating surgeon. Patients underwent magnetic resonance imaging and a full clinical workup. The 20 collected variables consisted of baseline PROM, as well as gender, age, height in centimeters (cm), weight in kilograms, body mass index, smoking status (active/inactive), alcohol consumption (regularly/not regularly), recreational drug use (yes/no), American Society of Anesthesiologists grade, prior discectomy at the index level (yes/no), index level, location of the herniation (left/right/midline/bilateral), classification as a far-lateral or sequestered herniation (yes/no), classification as a broad bulging disc without annular rupture compressing nerve roots (yes/no), presence of stenosis or spondylolisthesis at the index level (yes/no). Only clinically relevant stenosis and spondylolisthesis at the index level were captured.

For baseline PROM measurement, patients completed a standardized questionnaire including numeric rating scales (NRS) for back pain and leg pain severity, ranging from 0 to 10, and a validated Dutch version of the Oswestry Disability Index (ODI) to capture functional disability, ranging from 0 to 100, with higher values representing increasing severity [26]. All PROM data were collected using a validated web-based PROM assessment tool [27]. We defined clinical success as achievement of the minimum clinically important difference (MCID) threshold set by Ostelo et al. at the 1-year postoperative follow-up [28]. Thus, an improvement from baseline of greater or equal than 30% at the 12-months' follow-up represented clinical success in functional disability (ODI) or pain scores (NRS). The primary endpoint for this study was Achievement of MCID in leg pain at 1 year.

Statistical analysis

Continuous data are given as mean ± standard deviation, and categorical data as numbers and percentages. The effect of the preoperative variables on clinical success was

quantified using Welch's two-sample t tests and Pearson's χ^2 tests. Here, we controlled the false discovery rate using the correction described by Benjamini and Hochberg [29]. Intergroup differences and their 95% confidence intervals (CI) are provided.

Before modeling, missing demographic data, which were assumed to be missing at random, were imputed using predictive mean matching in a single imputation procedure [30]. As considerable class imbalance of around 80%/20% was present, we applied the synthetic minority oversampling technique to ensure robustness of our models [31]. For internal validation, data were randomly split into training, validation, and test sets in a 60%/20%/20% ratio. Deep feedforward artificial neural networks were built and trained in Keras with a TensorFlow (Google Brain Team, Mountainview, CA, USA) backend [32]. We implemented batch normalization, as well as random dropout to prevent overfitting, whenever necessary [33,34]. Three separate models for functional disability, back pain, and leg pain, were constructed. A range of models were trained by hyperparameter grid search, and continually evaluated on the validation set. The three most robust models, based on area under the receiver operating characteristics curve (AUC) analysis, were chosen as the final models, and subsequently evaluated on the test set. Performance measures were calculated from the resulting confusion matrices. We also trained logistic regression models on the same dataset. All analyses were carried out in *R* version 3.4.4 (The R Foundation for Statistical Computing, Vienna, Austria). A two-tailed $p \leq .05$ was considered statistically significant.

Results

During the study period, 2,695 patients underwent discectomy at our center, of whom 441 correctly returned all baseline and all 1-year PROM questionnaires. Of these, 19 had more than 20% missing baseline data, and were excluded. Finally, 422 patients were included in this study (Fig. 1). Detailed baseline characteristics are provided in Table 1.

Patient-reported outcome measures

At baseline, unadjusted NRS for leg pain severity averaged 7.3 ± 2.1, while NRS for back pain severity was at 5.3 ± 2.8, and ODI at 47.4 ± 18.5. During the first postoperative year, leg pain improved by a mean of 5.0 (95% CI: 5.3−4.7) to a 1-year measurement of 2.3 ± 2.8. Similarly, back pain on average improved by 2.0 (95% CI: 2.4−1.7) to a final value of 3.2 ± 2.8. Functional disability as measured by the ODI decreased by 31.3 points (95% CI: 33.5 −29.1) to 16.1 ± 17.0 points. In terms of MCID, 337 (80%) patients reported a relevant improvement in leg pain severity at 1 year, while 219 (52%) and 337 (80%) reported improvement in back pain and functional disability, respectively.

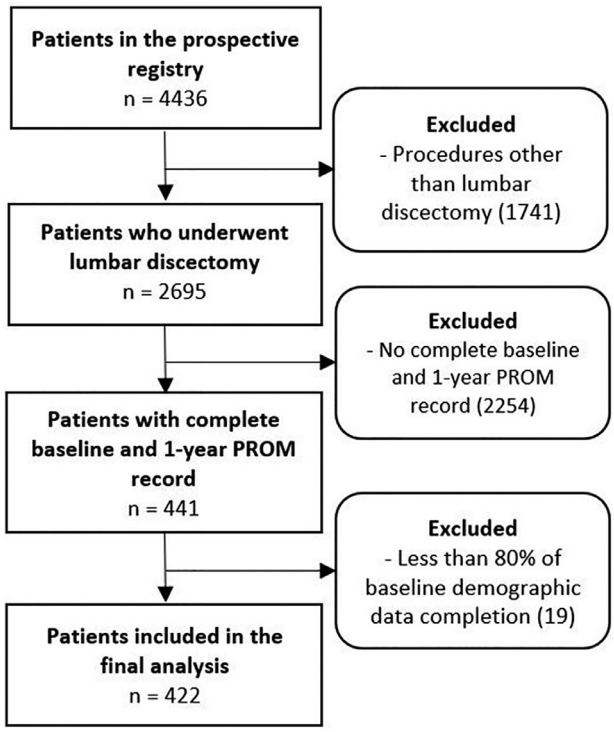

Fig. 1. Flowchart demonstrating the patient selection process for this study.
PROM, patient reported outcome measure.

Univariate predictors

The association of baseline variables with postoperative outcome was assessed using conventional statistical methods (Table 1). After correction for multiple testing, female gender, lower age, greater height, American Society of Anesthesiologists grade 1, absence of stenosis at the index level, absence of a bulging disc as the cause for symptoms, and greater baseline leg pain severity on the NRS scale were significantly associated with clinically relevant improvements in leg pain. Similarly, greater preoperative NRS back pain and ODI scores were related to improvement in back pain. Greater height and ODI scores were the only factors that were associated with achievement of MCID in functional disability.

Model development and validation

After extensive training and hyperparameter optimization, powerful deep learning and logistic regression models were arrived at. Performance, as measured on the test set and thus on data previously unknown to the model, was used for internal validation. Comparative model performance in terms of AUC on the test set for each of the three outcome measures is illustrated in Fig. 2.

Achievement of MCID in leg pain severity at 1 year, the primary endpoint, was predicted by the deep learning model with an AUC of 0.87, as well as an accuracy, sensitivity, and specificity of 85% (Table 2). In comparison, the

Table 1
Baseline patient characteristics. Values are provided for the overall cohort, as well as stratified by achievement of clinical success in the three outcome measures at 1 year. Clinical success was defined as a ≥30% improvement from baseline

Variable	Overall	Leg pain			Back pain			Functional disability		
	N = 422	MCID N = 337	Non-MCID N = 85	p	MCID N = 219	Non-MCID N = 203	p	MCID N = 337	Non-MCID N = 85	p
Demographic										
Female gender, n (%)	207 (49)	154 (46)	53 (62)	0.049*	103 (47)	104 (51)	0.62	161 (48)	46 (54)	0.60
Age, mean ± SD [y]	48.5 ± 11.5	47.6 ± 11.5	52.0 ± 11.0	0.009*	49.2 ± 11.5	47.8 ± 11.5	0.49	48.5 ± 11.3	48.8 ± 12.5	0.93
Height, mean ± SD [cm]	177.6 ± 9.7	178.4 ± 9.4	174.4 ± 9.9	0.009*	178.4 ± 9.4	176.7 ± 9.9	0.27	178.3 ± 9.5	174.5 ± 9.7	0.015*
Weight, mean ± SD [kg]	80.6 ± 11.8	81.1 ± 12.1	78.6 ± 10.3	0.21	81.4 ± 12.8	79.7 ± 10.5	0.34	81.1 ± 12.0	78.8 ± 10.6	0.28
BMI, mean ± SD [kg/m^2]	25.6 ± 3.2	25.5 ± 3.1	25.9 ± 3.5	0.51	25.6 ± 3.4	25.6 ± 3.1	0.99	25.5 ± 3.1	25.9 ± 3.5	0.51
Active smoker, n (%)	137 (32)	106 (31)	31 (36)	0.62	68 (31)	69 (34)	0.74	101 (30)	36 (42)	0.17
Regular alcohol consumption, n (%)	221 (52)	177 (53)	44 (52)	0.99	112 (51)	109 (54)	0.79	175 (52)	46 (54)	0.92
Recreational drug use, n (%)	12 (3)	12 (4)	0 (0)	0.39	8 (4)	4 (2)	0.62	8 (2)	4 (5)	0.62
ASA grade 1, n (%)	246 (58)	209 (62)	37 (44)	0.018*	119 (54)	127 (63)	0.32	202 (60)	44 (52)	0.49
Disease-specific										
Prior discectomy, n (%)	47 (11)	38 (11)	9 (11)	0.99	20 (9)	27 (13)	0.49	33 (10)	14 (16)	0.33
Index level L5-S1, n (%)	205 (49)	162 (48)	43 (51)	0.89	112 (51)	93 (46)	0.56	160 (47)	45 (53)	0.62
Midline herniation, n (%)	22 (5)	15 (4)	7 (8)	0.51	13 (6)	9 (4)	0.78	16 (5)	6 (7)	0.72
Far-lateral herniation, n (%)	13 (3)	12 (4)	1 (1)	0.62	8 (4)	5 (2)	0.79	12 (4)	1 (1)	0.62
Spondylolisthesis, n (%)	22 (5)	15 (4)	7 (8)	0.51	14 (6)	8 (4)	0.60	16 (5)	6 (7)	0.72
Stenosis, n (%)	56 (13)	34 (10)	22 (26)	0.004*	29 (13)	27 (13)	0.99	40 (12)	16 (19)	0.34
Sequestered herniation, n (%)	79 (19)	60 (18)	19 (22)	0.62	41 (19)	38 (19)	0.99	62 (18)	17 (20)	0.93
Bulging disc, n (%)	19 (5)	9 (3)	10 (12)	0.009*	8 (4)	11 (5)	0.70	12 (4)	7 (8)	0.33
NRS leg pain, mean ± SD	7.3 ± 2.1	7.5 ± 1.8	6.3 ± 2.6	<0.001*	7.4 ± 2.0	7.1 ± 2.1	0.31	7.4 ± 2.1	6.8 ± 2.0	0.12
NRS back pain, mean ± SD	5.3 ± 2.8	5.1 ± 2.9	5.8 ± 2.4	0.17	6.3 ± 2.4	4.1 ± 2.8	<0.001*	5.1 ± 2.9	5.7 ± 2.4	0.27
ODI, mean ± SD	47.4 ± 18.5	47.8 ± 18.8	45.6 ± 17.0	0.55	50.0 ± 18.5	44.6 ± 18.0	0.018*	49.0 ± 18.4	41.2 ± 17.5	0.006*

MCID, minimum clinically important difference; SD, standard deviation; BMI, body mass index; ASA, American Society of Anesthesiologists; NRS, numeric rating scale; ODI, Oswestry Disability Inde.

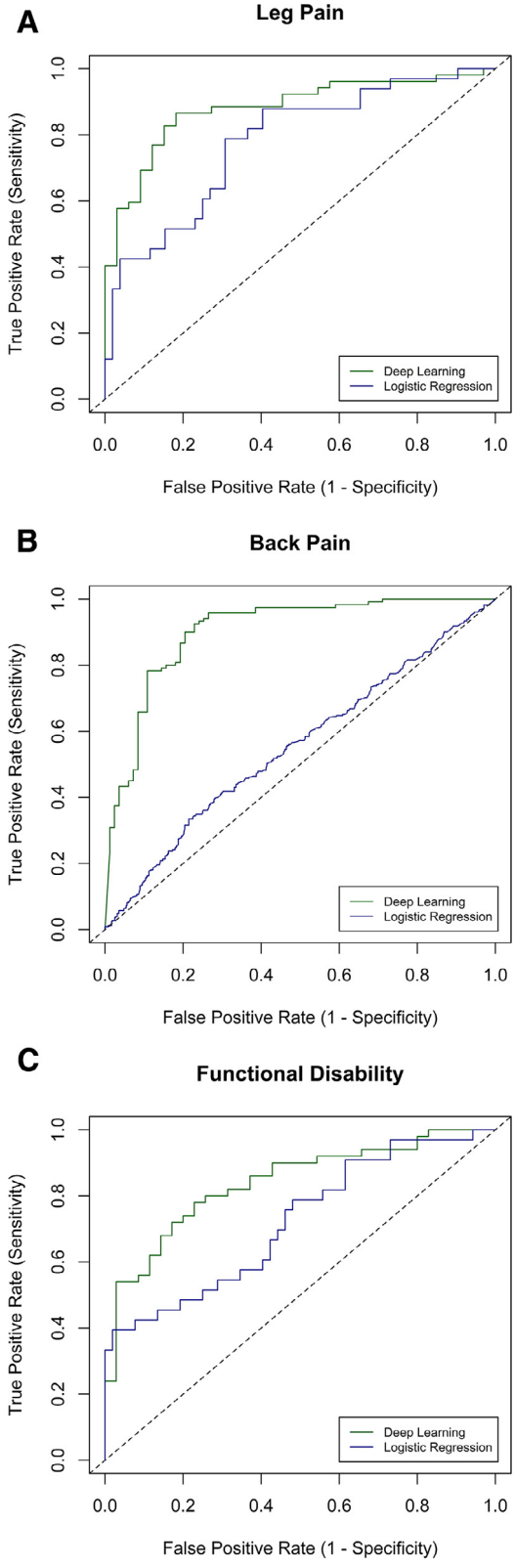

Fig. 2. Plots illustrating the comparative area under the receiver operating characteristics curve (AUC) performance of deep learning and logistic regression on the test set.

A clear improvement in performance can be observed for leg pain (A) with AUC values of 0.87 and 0.78, back pain (B) with 0.90 and 0.55, and functional disability (C) with 0.84 and 0.72 for deep learning and logistic regression, respectively.

logistic regression model scored an AUC of 0.78, accuracy of 68%, sensitivity of 55%, and specificity of 77%.

In terms of functional disability as measured by ODI, the deep learning and logistic regression attained AUC values of 0.84 and 0.72, and accuracy of 75% and 59%, respectively. The greatest discrepancy in performance measures was observed in the models predicting back pain improvement. Here, deep learning achieved an AUC of 0.90 and accuracy of 87%, while the logistic regression model attained an AUC of 0.55 and accuracy of 58%. Results of the deep learning models applied to hypothetical patients are provided in Table 3. Fig. 3 illustrates correlation-based variable importance for each of the machine learning-based models [35].

Discussion

Clinical practice and research are increasingly shifting from surrogate markers of treatment success, such as radiological outcomes, towards value-based care with a focus on PROM. We demonstrate that derivation of an accurate personalized prediction tool is possible even from a small single-center cohort by applying advanced ML methods. The models appear to outperform logistic regression, which was used as a proxy for more conventional statistical models, by a relevant margin. The ability to preoperatively and reliably inform patients about the likelihood of symptom improvement could prove useful not only in patient counseling and shared decision-making, but also in improving cost-effectiveness of LDH treatment.

Clinical utility of prediction models

There is considerable variability in indications, patient selection, surgical techniques, and outcomes in elective surgery for LDH [2–5,8]. This variability contributes to a considerable lack of predictability in clinical outcome, often leading to unsatisfactory outcomes and reoperations for failed back surgery syndrome or recurrent LDH, both of which lead to substantial added health-care utilization and costs [36]. We identified factors univariately associated with 1-year PROM in our cohort. Most importantly, we found that patients with a worse preoperative status were more likely to improve. Even this intuitively convincing notion is not consistently supported as a reliable predictor in the literature [5,8]. Integrating the multitude of outcome predictors described in the literature into daily clinical practice is not always feasible. Subsets of patients who benefit more or less from surgery do, however, exist and are often hard to identify.

Clinical prediction models can provide valuable insights to physicians [9–13,16]. As opposed to informing patients about a generalized treatment success rate, based on historical published data, they derive likelihoods of improvement at an individual level, and provide results that are equally interpretable by patients and clinicians. If consistently used, they have the potential to lower the proportion of

Table 2
Performance measures of the deep learning and logistic regression models for clinical success. Final performance was assessed on the separate test set

Performance measure	Leg pain		Back pain		Functional disability	
	Deep learning	Regression	Deep learning	Regression	Deep learning	Regression
AUC	0.87	0.78	0.90	0.55	0.84	0.72
Accuracy	85%	68%	87%	58%	75%	59%
Sensitivity	85%	55%	96%	41%	86%	52%
Specificity	85%	77%	74%	70%	60%	75%
PPV (precision)	90%	60%	84%	49%	75%	57%
NPV	79%	73%	92%	63%	76%	71%
F1 score*	0.87	0.57	0.90	0.45	0.80	0.54

AUC, area under the curve; PPV, positive predictive value (precision); NPV, negative predictive value.
* The F1 score is a composite score, and represents the harmonic mean of precision and sensitivity.

patients undergoing surgery to no avail, subsequently avoiding complications and reoperations. Comprehensive personalized risk profiles generated using our clinical prediction tool may possibly even help hospitals gain insights into their own cost-effectiveness and ways to avoid unnecessary surgery, based upon their individualized predictive analytics.

Clinical prediction tools should however, not be used as absolute indicators of surgical success. Instead, they should be integrated within the day-to-day routine and adopted as an adjunct to surgeon experience and real-time interpersonal judgement. As such, they may enable more personalized informed decision-making. It is even conceivable that in some cases unfavorable patient characteristics identified by a prediction tool could be preoperatively modified to result in a higher likelihood of improvement.[13] Azimi et al. constructed the only previous ML model on outcomes after discectomy [16]. However, they used shallow neural networks instead, and predicted Macnab criteria with an AUC of 0.82. In addition, they reported that their logistic regression model performed worse than their ML model. In contrast to our study, Azimi et al. had access to a considerably smaller patient cohort, did not specify any measures taken against overfitting, and did not apply disease-specific PROM.

Personalized prediction models

Most outcome prediction models are derived from multicenter data. This leads to models that are widely used, highly generalizable, and with decent accuracy on external data, but at the cost of personalized accuracy. They usually also require elaborate and expensive prospective data collection with large sample sizes. This trade-off is attributable to the fact that surgeons in a particular center are exposed to different patient demographics, implement different diagnostics, selection criteria, guidelines, and surgical techniques, and achieve different complication patterns and clinical outcomes. Still, this is currently the preferred approach to predictive analytics. One reason is that some centers and clinicians would rather employ a validated, published multicenter model than to go through the efforts

Table 3
Results obtained after applying the deep learning models on five hypothetical patients

Patient characteristics	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
Age [y]	65	25	25	55	55
BMI [kg/m^2]	30	22	30	25	25
Gender	Female	Female	Male	Female	Male
ASA score	2	1	1	1	1
Smoking status	Nonsmoker	Nonsmoker	Active	Nonsmoker	Active
Prior discectomy at index level	No	No	No	Yes	Yes
Sequestered disc herniation	No	Yes	No	No	No
Preoperative status					
NRS leg pain	8	8	5	9	8
NRS back pain	3	3	3	3	8
Oswestry disability index	50	30	30	30	75
Predicted likelihood of achieving MCID*					
Leg pain [%]	60	74	68	99	83
Back pain [%]	4	0	0	0	100
Functional disability [%]	87	12	16	84	83

BMI, body mass index; ASA, American Society of Anesthesiologists; NRS, numeric rating scale; MCID, minimum clinically important difference.
* The MCID was defined as a ≥30% improvement from the baseline to the 1-year measurement.

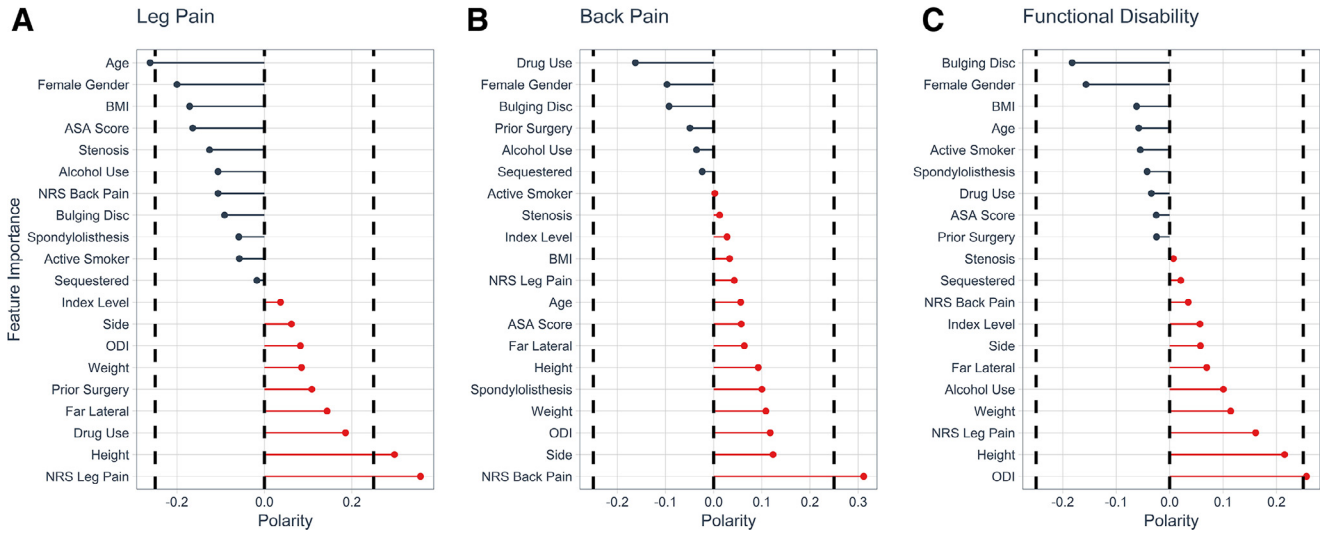

Fig. 3. Plots for leg pain (A), back pain (B), and functional disability (C) illustrating correlation-based variable importance in the machine learning-based models. Variable importance was obtained by correlation of input variables with the predicted labels on the training set.

ODI, Oswestry Disability Index; NRS, numeric rating scale; ASA, American Society of Anesthesiologists; BMI, body mass index.

of creating their own, personalized ML-based prediction tool from their own data. However, ML-based modeling is nowadays more accessible than ever. A range of industry-grade tools are freely and widely available, and any clinician with sufficient data, a computer, and the required background knowledge in biostatistics and data science could produce their own, individualized predictive analytics, as our study demonstrates.

On the other hand, individual centers or clinicians may also opt to train specific prediction tools based on data representative of their own patient cohort and clinical routine for optimal results. Retrospective data are often readily available. We demonstrate that center-specific modeling and internal validation based on advanced ML methods is feasible even with only a few hundred retrospectively collected patients. With the advent of the "big data" era, where large amounts of data and outcomes are automatically tracked, as well as the increase in computing power available to most clinicians, this type of personalized predictive modeling may soon become commonplace [22,23]. In the future, it may even become feasible for clinicians without any knowledge of ML to automatically obtain accurate, personalized prediction models tailored to their own data. Our findings could encourage surgeons to create personalized algorithms, tailored to their clinical practice, and to be used only by them.

The major issue with personalized prediction tools is overfitting. Overfitting occurs when a model adjusts too closely to training data, and subsequently demonstrates poor performance on new data. This is because, during overfitting, the model starts memorizing specific training observations, without actually extracting generalizable relationships between variables. Due to the nature of personalized, center-specific prediction tools, external validation is not always the most useful method for estimation of out-of-sample error for these models, as this would not correspond to the intended use of such individualized predictive models, which is to be applied to future patients from the same center. Instead, prospective internal validation would be the method of choice, since this represents the actual clinical use of such tools. Internal validation enables quantification of overfitting. In addition, some ML methods integrate specific tools to reduce overfitting. For deep learning, we applied the dropout technique, which has been shown to effectively minimize overfitting [33].

Ultimately, a prediction tool can only be as good as the inputs provided to it. While we included a range of clinical and demographic baseline data, we did not include neurologic deficits, duration of symptoms, or baseline anxiety and depression scores, which may affect the likelihood of improvement after surgery. Furthermore, intra- and perioperative parameters, such as complications, could improve model performance. However, including data that are not available preoperatively would defeat the purpose of a prediction tool for use in clinical shared decision-making. Rather, separate models incorporating these intra- and perioperative data for a refined prediction at discharge would be valuable.

Limitations

Although all data was obtained from a prospective registry, selection bias cannot effectively be ruled out, and the methodology was not prospectively decided upon. The data that was used for training the models in our study was derived from a prospective single-center registry. Our model may only be accurate in patients undergoing tubular microdiscectomy. The findings of our univariate analysis of factors associated with outcomes may be biased by confounders. Furthermore, comorbidities such as diabetes and

cardiovascular disease were not available as inputs. The decision of good vs. bad outcomes was based on MCID, a commonly used metric [28]. However, other metrics such as substantial clinical benefit (SCB) or even patient satisfaction would certainly result in different analytics. Our models do not include measures of quality of life and objective functional impairment. Lastly, further supervised training with a larger sample, or semisupervised training that integrates unlabeled data are likely to improve robustness and accuracy [37]. This would allow reducing the amount of required input variables and construction of a simpler model more suitable for clinical practice.

Conclusions

We have created a preoperative clinical prediction model for patients' individual likelihood of improvement in leg pain, back pain, and functional disability after surgery for lumbar disc herniation. Clinical prediction tools have the potential to improve personalized shared decision-making. Our study demonstrates that constructing and internally validating a center- or even surgeon-specific model for outcome prediction is feasible with relatively small amounts of data by applying deep learning. In the "big data" era, this personalized approach may become a standard for predictive analytics.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.1016/j.spinee.2018.11.009.

References

[1] Gibson JNA, Waddell G. Surgical interventions for lumbar disc prolapse. Cochrane Database Syst Rev 2007:CD001350. https://doi.org/10.1002/14651858.CD001350.pub4.

[2] Arts MP, Brand R, van den Akker ME, Koes BW, Bartels RHMA, Peul WC, et al. Tubular diskectomy vs conventional microdiskectomy for sciatica: a randomized controlled trial. JAMA 2009;302:149–58. https://doi.org/10.1001/jama.2009.972.

[3] Weinstein JN, Tosteson TD, Lurie JD, Tosteson ANA, Hanscom B, Skinner JS, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT): a randomized trial. JAMA 2006;296:2441–50. https://doi.org/10.1001/jama.296.20.2441.

[4] Weinstein JN, Lurie JD, Tosteson TD, Skinner JS, Hanscom B, Tosteson ANA, et al. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT) observational cohort. JAMA 2006;296:2451–9. https://doi.org/10.1001/jama.296.20.2451.

[5] Arts MP, Brand R, Koes BW, Peul WC. Effect modifiers of outcome of surgery in patients with herniated disc related sciatica? A subgroup analysis of a randomised clinical trial. J Neurol Neurosurg Psychiatr 2010;81:1265–74. https://doi.org/10.1136/jnnp.2009.192906.

[6] Atlas SJ, Keller RB, Wu YA, Deyo RA, Singer DE. Long-term outcomes of surgical and nonsurgical management of sciatica secondary to a lumbar disc herniation: 10 year results from the maine lumbar spine study. Spine 2005;30:927–35.

[7] Konstantinou K, Dunn KM. Sciatica: review of epidemiological studies and prevalence estimates. Spine 2008;33:2464–72. https://doi.org/10.1097/BRS.0b013e318183a4a2.

[8] Koerner JD, Glaser J, Radcliff K. Which variables are associated with patient-reported outcomes after discectomy? Review of SPORT disc herniation studies. Clin Orthop Relat Res 2015;473:2000–6. https://doi.org/10.1007/s11999-014-3671-1.

[9] Khor S, Lavallee D, Cizik AM, Bellabarba C, Chapman JR, Howe CR, et al. Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery. JAMA Surg 2018. https://doi.org/10.1001/jamasurg.2018.0072.

[10] McGirt MJ, Sivaganesan A, Asher AL, Devin CJ. Prediction model for outcome after low-back surgery: individualized likelihood of complication, hospital readmission, return to work, and 12-month improvement in functional disability. Neurosurg Focus 2015;39:E13. https://doi.org/10.3171/2015.8.FOCUS15338.

[11] McGirt MJ, Bydon M, Archer KR, Devin CJ, Chotai S, Parker SL, et al. An analysis from the Quality Outcomes Database, Part 1. Disability, quality of life, and pain outcomes following lumbar spine surgery: predicting likely individual patient outcomes for shared decision-making. J Neurosurg: Spine 2017;27:357–69. https://doi.org/10.3171/2016.11.SPINE16526.

[12] Asher AL, Devin CJ, Archer KR, Chotai S, Parker SL, Bydon M, et al. An analysis from the Quality Outcomes Database, Part 2. Predictive model for return to work after elective surgery for lumbar degenerative disease. J Neurosurg: Spine 2017;27:370–81. https://doi.org/10.3171/2016.8.SPINE16527.

[13] Steinmetz MP, Mroz T. Value of adding predictive clinical decision tools to spine surgery. JAMA Surg 2018. https://doi.org/10.1001/jamasurg.2018.0078.

[14] Senders JT, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman ML, et al. Natural and artificial intelligence in neurosurgery: a systematic review. Neurosurgery 2017. https://doi.org/10.1093/neuros/nyx384.

[15] Azimi P, Mohammadi HR, Benzel EC, Shahzadi S, Azhari S, Montazeri A. Artificial neural networks in neurosurgery. J Neurol Neurosurg Psychiatry 2015;86:251–6. https://doi.org/10.1136/jnnp-2014-307807.

[16] Azimi P, Benzel EC, Shahzadi S, Azhari S, Mohammadi HR. The prediction of successful surgery outcome in lumbar disc herniation based on artificial neural networks. J Neurosurg Sci 2016;60:173–7.

[17] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.

[18] Malik AT, Panni UY, Mirza MU, Tetlay M, Noordin S. The impact of surgeon volume on patient outcome in spine surgery: a systematic review. Eur Spine J 2018;27:530–42. https://doi.org/10.1007/s00586-017-5447-2.

[19] Shriver MF, Xie JJ, Tye EY, Rosenbaum BP, Kshettry VR, Benzel EC, et al. Lumbar microdiscectomy complication rates: a systematic review and meta-analysis. Neurosurg Focus 2015;39:E6. https://doi.org/10.3171/2015.7.FOCUS15281.

[20] Ng K, Sun J, Hu J, Wang F. Personalized predictive modeling and risk factor identification using patient similarity. AMIA Jt Summits Transl Sci Proc 2015;2015:132–6.

[21] Snyderman R. Personalized health care: from theory to practice. Biotechnol J 2012;7:973–9. https://doi.org/10.1002/biot.201100297.

[22] Oravec CS, Motiwala M, Reed K, Kondziolka D, Barker FG, Michael LM, et al. Big data research in neurosurgery: a critical look at this popular new study design. Neurosurgery 2018;82:728–46. https://doi.org/10.1093/neuros/nyx328.

[23] Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 2016;375:1216–9. https://doi.org/10.1056/NEJMp1606181.

[24] Staartjes VE, de Wispelaere MP, Miedema J, Schröder ML. Recurrent lumbar disc herniation after tubular microdiscectomy: analysis of learning curve progression. World Neurosurg 2017. https://doi.org/10.1016/j.wneu.2017.07.121.

[25] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594.

[26] Van Hooff ML, Spruit M, Fairbank JCT, van Limbeek J, Jacobs WCH. The Oswestry Disability Index (version 2.1a): validation of a Dutch language version. Spine 2015;40:E83–90. https://doi.org/10.1097/BRS.0000000000000683.

[27] Schröder ML, de Wispelaere MP, Staartjes VE. Are patient-reported outcome measures biased by method of follow-up? Evaluating paper-based and digital follow-up after lumbar fusion surgery. Spine J 2018. https://doi.org/10.1016/j.spinee.2018.05.002.

[28] Ostelo RWJG, Deyo RA, Stratford P, Waddell G, Croft P, Von Korff M, et al. Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change. Spine 2008;33:90–4. https://doi.org/10.1097/BRS.0b013e31815e3a10.

[29] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodological) 1995;57:289–300. https://doi.org/10.2307/2346101.

[30] van Buuren S. Groothuis-Oudshoorn CGM. Mice: multivariate imputation by chained equations in R. J Stat Softw 2011: 45. https://doi.org/10.18637/jss.v045.i03.

[31] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. JAIR 2002;16:321–57. https://doi.org/10.1613/jair.953.

[32] Chollet. Keras: deep learning library for Theano and TensorFlow URL: Https://KerasIo/K, 2015;7:8.

[33] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.

[34] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. ArXiv:150203167 [Cs] 2015.

[35] Hall M.A. Correlation-based feature selection for machine learning. 1999.

[36] Ambrossi GLG, McGirt MJ, Sciubba DM, Witham TF, Wolinsky J-P, Gokaslan ZL, et al. Recurrent lumbar disc herniation after single-level lumbar discectomy: incidence and health care cost analysis. Neurosurgery 2009;65:574–8 discussion 578. https://doi.org/10.1227/01.NEU.0000350224.36213.F9.

[37] Erhan D, Bengio Y, Courville A, Manzagol P-A, Vincent P, Bengio S. Why does unsupervised pre-training help deep learning? J Mach Learn Res 2010;11:625–60.