Machine learning–based preoperative predictive analytics for lumbar spinal stenosis

Alessandro Siccoli, BMed,1 Marlies P. de Wispelaere, PDEng,2 Marc L. Schröder, MD, PhD,1 and Victor E. Staartjes, BMed1,3,4

1Department of Neurosurgery, Bergman Clinics, Amsterdam; 2Department of Clinical Informatics, Bergman Clinics, Amsterdam; 3Amsterdam UMC, Vrije Universiteit Amsterdam, Neurosurgery, Amsterdam Movement Sciences, Amsterdam, The Netherlands; and 4Department of Neurosurgery, Clinical Neuroscience Centre, University Hospital Zurich, University of Zurich, Switzerland

OBJECTIVE Patient-reported outcome measures (PROMs) following decompression surgery for lumbar spinal stenosis (LSS) demonstrate considerable heterogeneity. Individualized prediction tools can provide valuable insights for shared decision-making. The authors aim to evaluate the feasibility of predicting short- and long-term PROMs, reoperations, and perioperative parameters by machine learning (ML) methods.

METHODS Data were derived from a prospective registry. All patients had undergone single- or multilevel mini-open facet-sparing decompression for LSS. The prediction models were trained using various ML-based algorithms to predict the endpoints of interest. Models were selected by area under the receiver operating characteristic curve (AUC). The endpoints were dichotomized by minimum clinically important difference (MCID) and included 6-week and 12-month numeric rating scales for back pain (NRS-BP) and leg pain (NRS-LP) severity and the Oswestry Disability Index (ODI), as well as prolonged surgery (> 45 minutes), extended length of hospital stay (> 28 hours), and reoperations.

RESULTS A total of 635 patients were included. The average age was 62 ± 10 years, and 333 patients (52%) were male. At 6 weeks, MCID was seen in 63%, 76%, and 61% of patients for ODI, NRS-LP, and NRS-BP, respectively. At 12 months, 66%, 63%, and 51% of patients reported MCID; the observed accuracies were 62%, 74%, and 66%, with AUCs of 0.68, 0.72, and 0.79. Reoperations occurred in 60 patients (9.5%), of which 27 (4.3%) occurred at the index level. Overall and index-level reoperations were predicted with 69% and 63% accuracy, respectively, and with AUCs of 0.66 and 0.61. In 15%, a length of surgery greater than 45 minutes was observed and predicted with 78% accuracy and AUC of 0.54. Only 15% of patients were admitted to the hospital for longer than 28 hours. The developed ML-based model enabled prediction of extended hospital stay with an accuracy of 77% and AUC of 0.58.

CONCLUSIONS Preoperative prediction of a range of clinically relevant endpoints in decompression surgery for LSS using ML is feasible, and may enable enhanced informed patient consent and personalized shared decision-making. Access to individualized preoperative predictive analytics for outcome and treatment risks may represent a further step in the evolution of surgical care for patients with LSS.

https://thejns.org/doi/abs/10.3171/2019.2.FOCUS18723

KEYWORDS machine learning; outcome prediction; lumbar spinal stenosis; patient-reported outcome; decompression; laminectomy

With mounting life expectancy, the incidence of symptomatic lumbar spinal stenosis (LSS) has been steadily increasing.27 Although conservative treatment is recommended and effective in the majority of cases, many studies suggest an advantage of decompression surgery over nonsurgical treatments for selected patients, and many improvements in surgical techniques have been made over the past decades.21,22 However, indications for and outcomes after lumbar decompression surgery remain heterogeneous, particularly in those patients who present with concomitant low-grade spondylolisthesis.1,14,15,21,24

ABBREVIATIONS ASA = American Society of Anesthesiologists; AUC = area under the receiver operating characteristic curve; GLM = generalized linear model; LSS = lumbar spinal stenosis; MCID = minimum clinically important difference; ML = machine learning; NPV = negative predictive value; NRS = numeric rating scale; NRS-BP = NRS for back pain; NRS-LP = NRS for leg pain; ODI = Oswestry Disability Index; PPV = positive predictive value; PROMs = patient-reported outcome measures; XGBoost = extreme gradient boosting.

SUBMITTED December 28, 2018. ACCEPTED February 14, 2019.

INCLUDE WHEN CITING DOI: 10.3171/2019.2.FOCUS18723.
The combination of hard-to-predict outcomes and a strong disease-linked psychological component requires in-depth preoperative counseling. A range of factors influencing the outcome of surgery for LSS have been identified. However, it is not always feasible to integrate many single patient factors into one communicable risk for poor outcome. Prediction models represent an interface between these important and complex factors, and produce an output that can be benchmarked and is interpretable and clinically valuable for both patients and surgeons. Recently, interest has shifted toward machine learning (ML) algorithms in predictive modeling. While some ML algorithms have been around for decades, they have only recently gained major interest for predictive analytics in medicine. This increase in the use of ML is attributable to the advent of the “big data” era, but also to the development of new algorithms and steadily improving computational power. Various ML techniques such as gradient-boosting machines, decision trees, random forests, and neural networks have already proven an increased predictive ability compared to more traditional statistical modeling techniques in neurosurgical patient care.

The hope for these new techniques is that they will enable surgeons to accurately predict adverse events and outcomes that were previously near unpredictable, such as reoperations or a positive outcome, in order to identify those patients who will benefit most from surgical care. In addition, such tools may aid patients in making the best-informed decision for themselves. Therefore, the purpose of this study was to investigate whether comprehensive ML-based preoperative predictive analytics are feasible and robust in predicting a range of clinically relevant endpoints in patients undergoing decompression surgery for LSS.

Methods
Patient Population
In a retrospective analysis based on data from a prospective registry, we identified all patients who had undergone decompression surgery for LSS between November 2013 and November 2018 at a single Dutch spine center (Bergman Clinics). All patients underwent single- or multilevel decompression by the senior author (M.L.S.). Adult patients with ≥ 80% baseline data completion were included in this study. In patients with concomitant low-grade spondylolisthesis, the decision of whether to add a fusion procedure to decompression alone was based upon a validated decision-making protocol. This study was devised and reported according to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement. All patients included in the registry provided written informed consent. The prospective registry was authorized by the local IRB (Medical Research Ethics Committees United), and this study was performed in accordance with the 2013 Declaration of Helsinki.

Surgical Technique
While the patient was in the knee-elbow position, a 50-mm midline incision was made. Via a muscle-splitting approach, a mini-open retractor was inserted. The interspinous ligament was cut, and the spinous process was partially resected. A bilateral partial hemilaminectomy was then performed. The interspinous ligament was deliberately resected to contralaterally undercut the hypertrophic ligamentum flavum and osteophytes. Whenever possible, the facet joints were left untouched to preserve biomechanical integrity. Only if necessary, their hypertrophic medial part was partially resected. The lateral recesses and foramina were further opened until the nerve roots appeared to be fully released. Discectomy was performed only in cases of significant nerve root compression by a bulging disc at the index level.

Data Collection
Clinical and radiological baseline data were obtained at the first outpatient visit to the treating surgeon. Patients underwent MRI and a full clinical workup. The 15 collected variables consisted of baseline patient-reported outcome measures (PROMs), as well as sex, age, height (cm), weight (kg), BMI, smoking status (active/inactive), regular alcohol intake (yes/no), American Society of Anesthesiologists (ASA) grade, prior surgery at the index level, single- or multilevel decompression, index levels, presence of concomitant spondylolisthesis, and the predominant presenting symptom (neurogenic claudication or [mono] radiculopathy).

Outcome Measures
Patient-Reported Outcome
For PROM measurement, patients completed a standardized questionnaire including numeric rating scales (NRSs) for back pain (NRS-BP) and leg pain (NRS-LP) 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. All PROM data were collected using a validated web-based PROM assessment tool at baseline, 6 weeks, and 12 months after surgery. We defined clinical success as achievement of the minimum clinically important difference (MCID) threshold set by Ostelo et al. at the 6-week or 12-month postoperative follow-up. Thus, an improvement from baseline of ≥ 30% represented clinical success in functional disability (ODI) or pain severity scores (NRS).

Reoperation
Any reoperations were tracked. The analysis of reoperations was stratified into reoperations at the index level and overall reoperations.

Perioperative Parameters
Length of hospital stay in hours and length of surgery in minutes were obtained. We identified patients with an extended length of stay, defined as greater than 28 hours. This cutoff was chosen according to the observed distribution in our prospective registry, and leads to an approximate 15% prevalence of extended hospital stay. Similarly, a length of surgery greater than 45 minutes was considered a prolonged surgery.
Statistical Analysis

Continuous data are given as mean ± standard deviation, and categorical data as values and percentages. Missing baseline data, which were assumed to be missing at random, were imputed using single imputation by predictive mean matching. Data were randomly split into two sets in a 70%/30% ratio for training and testing, respectively. On the training set, the models were trained and continually validated using bootstrapping. We trialed the following 7 ML algorithms for each variable: random forests, extreme gradient boosting (XGBoost), Bayesian generalized linear models (GLMs), boosted trees, \( k \)-nearest neighbor, simple GLMs, as well as artificial neural networks with a single hidden layer. For each outcome, after hyperparameter tuning, the model with the best performance during bootstrapping was selected based on area under the receiver operating characteristic curve (AUC), which constitutes the primary endpoint of this study. The selected final model was then evaluated on the test set for internal validation. All baseline variables were provided to all models. Whenever relevant class imbalance was present, we applied conventional random upsampling or synthetic minority oversampling (SMOTE) to the training set to ensure robustness of our models. At internal validation, calibration—the degree to which predicted probabilities correspond to true, observed probabilities—was assessed using the Brier score. The Brier score ranges from 0 to 1, with values closer to 0 indicating greater calibration. AUC-based individual variable importance was obtained for each of the models. All analyses were performed in R (version 3.5.1, The R Foundation for Statistical Computing). The complete code is provided in the Online-Only Content.

Results

During the study period, 635 patients who underwent lumbar decompression were identified. All patients were included in at least one of the models. The average age was 62 ± 10 years, and 333 patients (52%) were male. Detailed baseline characteristics of the study population are shown in Table 1. An overview of the endpoints observed in the cohort is provided in Table 2.

Patient-Reported Outcome

Short-Term Follow-Up

At 6 weeks postoperatively, a clinically meaningful improvement was seen in 63%, 76%, and 61% of patients for ODI, NRS-LP, and NRS-BP, respectively (Table 2). At internal validation (Table 3), the developed models were able to predict MCID in these variables with accuracies of 69%, 76%, and 85% and with AUC values of 0.75, 0.79, and 0.92, respectively (Fig. 1). This indicates fair to good discriminative ability, and the comparison of training and testing results indicates that no overfitting occurred. The Brier score indicates fair calibration of the predicted probabilities. The variables that were weighted most strongly by the models (Table 4) included baseline PROMs, concomitant spondylolisthesis, prior surgery, and smoking status.

| Parameter | Value |
|-----------|-------|
| Males, n (%) | 333 (52) |
| Mean age ± SD, yrs | 62 ± 10 |
| Prior surgery, n (%) | 59 (9.3) |
| Active smoker, n (%)* | 109 (43) |
| Regular alcohol intake, n (%)* | 228 (90) |
| Predominant presenting symptom, n (%) | Neurogenic claudication 502 (79) |
| Radiculopathy | 110 (17) |
| Concomitant spondylolisthesis, n (%) | 89 (14) |
| ASA score, n (%) | I 219 (39) |
| II | 334 (60) |
| III | 4 (0.7) |
| Mean height ± SD, cm | 173 ± 14 |
| Mean weight ± SD, kg | 81 ± 14 |
| Mean BMI ± SD, kg/m² | 26 ± 3.4 |
| Index level, n (%) | L1–2 6 (1.0) |
| L2–3 | 77 (12) |
| L3–4 | 276 (43) |
| L4–5 | 416 (66) |
| L5–S1 | 61 (9.6) |
| Multilevel decompression, n (%) | 178 (28) |
| Mean baseline PROMs ± SD | ODI 40 ± 16 |
| NRS-LP | 6.7 ± 2.6 |
| NRS-BP | 5.4 ± 3.1 |

* For these two variables data were available in 253 patients.

| Endpoint | No. Available (%) | Incidence (%) |
|----------|------------------|---------------|
| 6-wk PROMs* | ODI 173 (27) | 109 (63) |
| NRS-LP | 196 (31) | 149 (76) |
| NRS-BP | 196 (31) | 119 (61) |
| 12-mo PROMs* | ODI 111 (17) | 73 (66) |
| NRS-LP | 115 (18) | 73 (63) |
| NRS-BP | 115 (18) | 59 (51) |
| Reoperations | Overall 635 (100) | 60 (9.5) |
| At index level 635 (100) | 27 (4.3) |
| Periop parameters | Prolonged op, >45 mins 451 (71) | 68 (15) |
| Extended hospital stay, >28 hrs 633 (99) | 95 (15) |

* For PROMs, the rate of patients achieving an MCID of at least ≥ 30% improvement from baseline is provided.
Long-Term Follow-Up

At the 12-month follow-up, 66%, 63%, and 51% of patients reported experiencing a clinically meaningful improvement in ODI, NRS-LP, and NRS-BP, respectively (Table 2). At internal validation (Table 3), we observed accuracies of 62%, 74%, and 66%, as well as AUC values of 0.68, 0.72, and 0.79 for these long-term endpoints. Slight overfitting was observed for the 12-month ODI model, and all models demonstrated fair calibration. Baseline PROMs, as well as BMI, height, and weight, were found to have the highest variable importance (Table 4).

Reoperation

During the follow-up period, a total of 60 patients (9.5%) underwent reoperations, of which 27 (4.3%) were at the index level (Table 2). Overall, 27 patients (4.3%) underwent discectomy, 25 (3.9%) underwent additional decompression, 5 (0.8%) required fusion surgery, and 3 patients (0.5%) required reexploration. Of these reoperations, 15 (2.4%) discectomies were at the index level, as well as 7 (1.1%) decompression procedures for persistent symptoms, 2 (0.3%) fusion procedures for iatrogenic spondylolisthesis after decompression, and all 3 (0.5%) reexplorations for suspected dural defects. At internal validation (Table 3), overall and index-level reoperations were predicted with 69% and 63% accuracy, respectively, and with AUC values of 0.69 and 0.63. While slight overfitting was observed, the Brier score indicates excellent calibration of the predicted probabilities of reoperation. Age, sex, BMI, and baseline PROMs had the strongest influence on these models (Table 4).

Perioperative Parameters

Prolonged Surgery

In 15% of patients, a length of surgery greater than 45 minutes was observed (Table 2). The occurrence of a prolonged surgical procedure was predicted with an accuracy of 78%, but an AUC of 0.54. The model, based on the random forest algorithm, appears to possess a high sensitivity and positive predictive value (PPV) of 85% and 91%, respectively, but is rather unspecific (23%). This model would be fairly useful as a rule-out tool. Fair calibration was observed, and baseline leg pain severity, age, height, and high-lumbar procedures showed the highest variable importance (Table 4).

Extended Hospital Stay

Only 15% of patients were admitted to the hospital for longer than 28 hours (Table 2). The developed ML-based model enabled prediction of extended hospital stay with an accuracy of 77% and AUC of 0.58. Here, specificity was high at 87%, with a negative predictive value (NPV) of 86%, thus enabling its use as a rule-in tool. Calibration of predicted probabilities for extended hospital stay was fair, and baseline ODI and leg pain severity as well as age and BMI appeared to strongly influence the model’s decision-making process (Table 4).

### TABLE 3. Performance metrics of the ML-based prediction models on the training set and on unseen testing data, representing internal validation

| Metric | 6-wk PROMs | 12-mo PROMs | Reops | Periop Parameters |
|--------|------------|-------------|-------|-------------------|
|        | ODI        | NRS-LP      | NRS-BP | Overall           | At Index Level | Prolonged Op | Extended Hospital Stay |
| Training (bootstrapping) | | | | | | | |
| Accuracy | 0.67 | 0.78 | 0.81 | 0.71 | 0.73 | 0.68 | 0.81 | 0.78 | 0.81 | 0.81 |
| AUC | 0.71 | 0.75 | 0.87 | 0.73 | 0.72 | 0.75 | 0.63 | 0.70 | 0.61 | 0.77 |
| Sensitivity | 0.74 | 0.87 | 0.89 | 0.54 | 0.83 | 0.60 | 0.35 | 0.08 | 0.86 | 0.38 |
| Specificity | 0.55 | 0.46 | 0.71 | 0.79 | 0.55 | 0.77 | 0.85 | 0.96 | 0.29 | 0.87 |
| PPV (precision) | 0.73 | 0.84 | 0.79 | 0.80 | 0.76 | 0.72 | 0.19 | 0.32 | 0.93 | 0.32 |
| NPV | 0.56 | 0.53 | 0.84 | 0.50 | 0.66 | 0.66 | 0.93 | 0.81 | 0.17 | 0.90 |
| F1 score* | 0.74 | 0.86 | 0.84 | 0.79 | 0.80 | 0.65 | 0.25 | 0.13 | 0.89 | 0.34 |
| Testing (internal validation) | | | | | | | |
| Accuracy | 0.69 | 0.76 | 0.85 | 0.62 | 0.74 | 0.66 | 0.69 | 0.63 | 0.78 | 0.77 |
| AUC | 0.75 | 0.79 | 0.92 | 0.68 | 0.72 | 0.79 | 0.66 | 0.61 | 0.54 | 0.58 |
| Sensitivity | 0.80 | 0.82 | 0.86 | 0.57 | 0.90 | 0.56 | 0.32 | 0.50 | 0.85 | 0.27 |
| Specificity | 0.47 | 0.57 | 0.83 | 0.73 | 0.53 | 0.76 | 0.69 | 0.64 | 0.23 | 0.87 |
| PPV (precision) | 0.76 | 0.86 | 0.89 | 0.81 | 0.72 | 0.71 | 0.10 | 0.07 | 0.91 | 0.28 |
| NPV | 0.53 | 0.50 | 0.79 | 0.44 | 0.80 | 0.62 | 0.90 | 0.96 | 0.14 | 0.86 |
| F1 score* | 0.78 | 0.84 | 0.87 | 0.67 | 0.80 | 0.63 | 0.15 | 0.12 | 0.88 | 0.27 |
| Brier score | 0.23 | 0.16 | 0.11 | 0.20 | 0.20 | 0.18 | 0.09 | 0.05 | 0.13 | 0.13 |

The following algorithms were used: 1) 6-week PROMS: Bayesian GLM (ODI), neural network (NRS-LP), and boosted trees (NRS-BP); 2) 12-mo PROMS: Bayesian GLM (ODI), random forest (NRS-LP), and XGBoost machine (NRS-BP); 3) reoperations: XGBoost (Overall) and boosted trees (At Index Level); and 4) perioperative parameters: random forest (Prolonged Op) and XGBoost (Extended Hospital Stay).

* The F1 score is a composite performance measure and is defined as the harmonic mean of precision and sensitivity.
Discussion

In a prospective registry of 635 patients operated on for LSS, we investigated the feasibility of preoperative identification of patients who benefit most from surgery, as well as those who are likely to require reoperation or encounter prolonged surgery or hospital stay. Effective prediction models based on ML were produced and internally validated. The models were able to predict the endpoints of interest with employable performance at internal validation from easily accessible preoperative data. ML, through nonlinear integration of preoperative variables, may enable more accurate preoperative estimation of intra-, peri-, and postoperative events, and has the potential to enable access to information on previously unpredictable endpoints, such as reoperation or unsatisfactory clinical outcome, which will be valuable to both patients and physicians.

The variability in indications, choice of surgical technique, and outcomes is notorious in elective spine surgery for LSS.\textsuperscript{21,22,27} This variability, combined with the psychological influencing factors and beliefs inherent to degenerative pathologies of the lumbar spine, makes accurate clinical predictions exceedingly hard.\textsuperscript{4} This lack of predictability for both the clinical outcome and adverse events—such as, for example, iatrogenic spondylolisthesis requiring fusion after decompression alone—means that correctly informing our patients on a case-to-case basis is complicated.\textsuperscript{3,16,36} It is often not possible to combine the large amount of known predictors and risk factors described in the literature into a single outcome prediction with which to inform our patient. Moreover, whenever a patient inquires about the odds of specific risks and benefits, often the best we can do is provide a carefully selected estimate from the literature. Subsets of patients who benefit more or less from surgery do, however, exist, and are often hard to identify. Vast clinical experience certainly enables accurate judging of a patient’s individual chances of successful outcome after surgery. However, it is not inconceivable that ML-based prediction tools (artificial intelligence) will soon overtake the performance of this “natural intelligence,” as has been the case in multiple areas of clinical neurosurgery already.\textsuperscript{31}

Such prediction tools, whether based on ML or on more traditional statistical modeling methods, also have distinct advantages other than potentially increased precision. As opposed to providing the patients with a generalized treatment success rate, based on historical published data from the peer-reviewed literature, they estimate the likelihood of a given endpoint at the individual patient level. Furthermore, their output is equally and easily interpretable by patient and physician alike. The outputs of well-validated models could even be used to compare the results of surgeons or hospitals based on their individual case complexity. If used consistently, through enhanced patient selection, they have the potential to lower the proportion of patients undergoing surgery to no avail, and possibly even to avoid subsequent complications and reoperations such as iatrogenic spondylolisthesis requiring fusion. The comprehensive personalized risk profiles generated using highly accurate clinical prediction tools may even help hospitals gain insights into cost-effectiveness and ways to avoid unnecessary surgery through comparison of predicted and

![AUC analysis of the ML models for prediction of 6-week (A) and 12-month (B) PROMs. Values were obtained from prediction on the testing (holdout) set, representing internal validation. Early outcome was predicted with AUC values of 0.75, 0.79, and 0.92 for functional disability, leg pain severity, and back pain severity, respectively. Similarly, AUC values of 0.68, 0.72, and 0.79 were observed for 12-month outcomes.](image)
achieved endpoints. Knowledge on the likelihood of an extended hospital stay can guide postoperative planning, and it may even become possible to predict and prevent costly readmissions in patients discharged too early.²,¹⁸–²⁰ Lastly, it has been suggested that clinical prediction tools could guide the preoperative optimization process by targeting and modifying specific patient factors—such as weight loss—to maximize the chances of clinical success.³⁸

However, the outputs of clinical prediction tools may never be used as absolute “red or green lights” for surgery or for surgical success. Instead, they should complement, not replace, the vast clinical experience that neurosurgeons have, and be adopted as an adjunct to interpersonal judgment in daily clinical routine. In this way, they may enable enhanced and more individualized shared decision-making. For example, a specific patient who has suffered from mild symptomatic LSS for 3 years may be unsure as to whether he or she should undergo surgery. Tailored predictive analytics may suggest that the patient, while having a very high chance at a minimum 30% symptomatic improvement, is also prone to reoperation at the index level, and to an extended length of hospital stay. In combination with counseling by an experienced surgeon, this would allow the patient to make a somewhat more informed decision as to whether or not to undergo surgery.

Although there has been a vast increase in the use of prediction models in spine surgery, relatively little has been done in decompression surgery for LSS, and even less so using ML methods. Using conventional logistic regression, Held et al. were able to predict MCID in pain and functional disability after decompression, with AUC values of 0.62 and 0.70, respectively.¹⁷ Azimi et al. used shallow neural networks to predict treatment satisfaction in patients with LSS, with an AUC of 0.80, which was better than with their logistic regression model.¹ Kalagara et al.¹⁸ were able to preoperatively predict 30-day readmissions following lumbar laminectomy using gradient boosting with an AUC of 0.69. Our results compare favorably with these works. In addition, these studies, in combination with analyses performed in other areas of spinal and cranial neurosurgery, corroborate the notion that ML models can be more precise in their predictions than more traditional statistical modeling techniques.¹,¹⁸,²⁰,²⁵,²⁹,³²,³⁴,³⁷

Application of ML algorithms to predictive analytics is not without drawbacks. First, many ML algorithms—notably deep learning—³⁴,³⁷—are often criticized for being “black box” models. That is, the user can observe what inputs are provided to the model and what the predicted output is, but the internal decision-making process of the model is unknown. Although we can visualize which variables are generally preferred by the model, as we have done using AUC-based variable importance in the current analysis, it is thus hard for the physician to explain to a patient why a particular prediction came together. In short, ML algorithms enable enhanced precision at the cost of reduced interpretability. In contrast, logistic regression provides

### TABLE 4. AUC-derived variable importance for each of the ML-based prediction models

| Variable                                      | 6-wk PROMs | 12-mo PROMs | Reops | Periop Parameters |
|-----------------------------------------------|------------|-------------|-------|-------------------|
|                                               | Overall    | At Index Level | Prolonged Op | Extended Hospital Stay |
| Sex                                           | 33.69      | 29.75       | 6.13   | 45.53*            | 37.13     | 8.35   | 6.90 |
| Age                                           | 12.89      | 0           | 2.14   | 60.03*            | 4.51      | 9.36*  | 27.44*       | 92.51*      | 53.63*     | 60.07* |
| BMI                                           | 17.04      | 26.16       | 100.00*| 49.52*            | 28.56*    | 30.64* | 7.74*         | 56.34       | 17.61      | 48.79* |
| Height                                        | 52.11*     | 61.60*      | 12.19  | 38.99             | 17.05*    | 17.54* | 48.75*        | 75.67*      | 59.23*     | 23.58* |
| Weight                                        | 46.97*     | 6.96        | 22.19  | 40.57*            | 5.79      | 2.39   | 3.21          | 97.02*      | 4.35       | 27.50* |
| Prior surgery                                 | 0          | 29.33       | 94.30* | 12.38             | 1.61      | 3.55   | 0             | 15.55       | 1.60       | 5.76   |
| Active smoker                                 | 43.44*     | 9.49        | 1.17   | 6.38              | 2.31      | 0      | 0             | 99.69*      | 83.82*     | 0.71   |
| Regular alcohol intake                        | 5.14       | 13.08       | 23.39  | 1.27              | 0         | 0      | 0             | 26.09       | 32.76      | 22.53* |
| Predominant presenting symptom                | 7.75       | 1.90        | 19.54  | 14.25             | 11.85*    | 0      | 0             | 48.05       | 0         | 4.15   |
| Concomitant spondylolisthesis                 | 12.74      | 46.62*      | 39.79* | 3.68              | 11.08*    | 0      | 1.98          | 6.23        | 8.23       | 0.63   |
| ASA score                                     | 19.34      | 8.02        | 0.59   | 7.83              | 2.56      | 0      | 3.07          | 48.82       | 10.23      | 1.60   |
| L1–2                                         | 9.75       | 0.21        | 0      | 2.13              | 0         | 0      | 0             | 49.34       | 0         |
| L2–3                                         | 5.53       | 50.42*      | 31.75  | 19.13             | 0.09      | 0      | 0             | 43.05       | 52.20*     | 2.42   |
| L3–4                                         | 6.67       | 48.52*      | 21.62  | 11.10             | 7.24      | 0      | 0             | 17.88       | 24.76      | 4.86   |
| L4–5                                         | 10.59      | 33.97       | 12.26  | 4.54              | 1.04      | 0      | 0             | 39.08       | 38.08      | 3.30   |
| L5–S1                                        | 8.67       | 1.06        | 8.62   | 5.41              | 3.14      | 0      | 0             | 13.10       | 16.98      | 0      |
| Multilevel decompression                     | 54.87*     | 41.35       | 31.87  | 9.50              | 5.63      | 0      | 5.74          | 38.43       | 42.94      | 6.75   |
| Baseline ODI                                  | 100.00*    | 100.00*     | 54.71* | 58.03*            | 67.54*    | 12.96* | 0             | 30.10       | 49.29      | 100.00* |
| Baseline NRS-LP                               | 1.23       | 7.17        | 31.94* | 100.00*           | 31.03*    | 8.05   | 100.00*       | 22.80       | 100.00*    | 24.84* |
| Baseline NRS-BP                               | 53.26*     | 30.38       | 18.10  | 31.18             | 100.00*   | 100.00*| 3.78          | 100.00*     | 8.18       | 19.19 |

* The five variables with the highest variable importance per model are marked with an asterisk.
easily interpretable odds ratios and confidence intervals. However, recent developments such as local interpretable model-agnostic explanations (LIME) are challenging this notion. Second, some ML architectures are prone to overfitting (Fig. 2), which is a problem in logistic regression, where overfitting can be easily combated by shrinking the regression coefficients. Overfitting occurs when an algorithm adjusts too closely to the training data set, and subsequently demonstrates poor performance on new data. This occurs because the model starts memorizing specific patients in the training set, without actually learning generalizable relationships between the inputs, which would allow it to make precise forecasts on new, unseen patient data. Training in \( k \)-fold cross-validation or bootstrapping as well as internal validation enable detection and quantification of overfitting. In addition, some ML algorithms integrate specific regularization mechanisms to reduce overfitting.

**Limitations**

Although we obtained all data from a prospective registry, selection bias cannot effectively be ruled out. Furthermore, our methodology was not prospectively decided upon. The findings of our variable importance analysis may be biased by confounders. Our prediction tool does not include measures of quality of life and objective functional impairment, nor did we have access to data on anxiety and depression, severity of stenosis, comorbidities, or neurological deficits, which are relevant data points in the management of patients with LSS. ML algorithms rely on large amounts of training data. Thus, although a comparatively large number of patients were included, and although we employed advanced data augmentation techniques, this study has to be seen as a pilot study demonstrating the feasibility of predicting these endpoints with a relatively small amount of data. Further supervised training with a larger sample, or semi-supervised training that integrates unlabeled data, is likely to improve the robustness and accuracy of the models. This would allow reducing the amount of required input variables and construction of a simpler model. Furthermore, we used MCID to dichotomize clinical outcome. However, the choice of other, equally viable options such as minimum detectable change or AUC-anchored methods could lead to different findings. Single-center data were used to derive our models. Although we have internally validated the models, this could possibly limit the generalizability of our prediction tools to different cohorts with varying patient characteristics, different clinical and surgical decision-making protocols, and surgeons with a different level of experience. Models trained on data with enough variance from a single setting, such as ours, can often be applied with considerable success at other centers. Nonetheless, external validation of existing models is highly desirable. In the future, however, individual centers may also opt to train their own, center-specific models. Our study shows that this is feasible.

**Conclusions**

Using data from a prospective registry, we trained and internally validated preoperative predictive analytics for patient-reported outcome, reoperations, prolonged surgery, and extended length of hospital stay. Prediction of clinically relevant endpoints in decompression surgery for LSS using machine learning is feasible, and may enable enhanced informed patient consent and personalized shared decision-making. Access to individualized preoperative predictive analytics for outcome and treatment risks may represent a further step in the evolution of surgical care for patients with LSS.

**References**

1. Azimi P, Benzel EC, Shahzadi S, Azhari S, Mohammadi HR: Use of artificial neural networks to predict surgical satisfaction in patients with lumbar spinal canal stenosis: clinical article. J Neurosurg Spine 20:300–305, 2014
2. Bernatz JT, Anderson PA: Thirty-day readmission rates in spine surgery: systematic review and meta-analysis. Neurosurg Focus 39(4):E7, 2015
3. Blumenthal C, Curran J, Benzel EC, Potter R, Magge SN, Harrington JF Jr, et al: Radiographic predictors of delayed
instability following decompression without fusion for degenerative grade I lumbar spondylolisthesis. J Neurosurg Spine 18:340–346, 2013

4. Burgstaller JM, Wertli MM, Steurer J, Kessels AGH, Held U, Gramke HF: The influence of pre- and postoperative fear avoidance beliefs on postoperative pain and disability in patients with lumbar spinal stenosis: analysis of the Lumbar Spinal Outcome Study (LSOS) data. Spine (Phila Pa 1976) 42:E425–E432, 2017

5. Chan AK, Bisson EF, Bydon M, Glassman SD, Foley KT, Potts EA, et al: Laminectomy alone versus fusion for grade I lumbar spondylolisthesis in 426 patients from the prospective Quality Outcomes Database. J Neurosurg Spine 30:234–241, 2018

6. Choma TJ, Mroz TE, Goldstein CL, Arnold P, Shamji MF: Emerging techniques in degenerative thoracolumbar surgery. Neurosurgery 80(3S):S55–S60, 2017

7. Cobo Soriano J, Sendino Revuelta M, Fabregate Fuente M, Cimarro Diaz I, Martinez Ureña F, Deglane Meneses R: Predictors of outcome after decompressive lumbar surgery and instrumented posterior-lateral fusion. Eur Spine J 19:1841–1848, 2010

8. 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 350:g7594, 2015

9. Copay AG, Glassman SD, Subach BR, Berven S, Schuler TC, Carreon LY: Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study Questionnaire Short Form 36, and pain scales. Spine 36(6):E3–E9, 2014

10. Dietz N, Sharma M, Alhourani A, Ugliuwenza B, Wang D, Nuño MA, et al: Variability in the utility of predictive models in predicting patient-reported outcomes following spine surgery for degenerative conditions: a systematic review. Neurosurg Focus 45(5):E10, 2018

11. Elsayed G, McClugage SG, Erwood MS, Davis MC, Dupépé EB, Szerlip P, et al: Association between payer status and patient-reported outcomes in adult patients with lumbar spinal stenosis treated with decompression surgery. J Neurosurg Spine 30:198–210, 2018

12. Försth P, Ólafsson G, Carlsson T, Frost A, Borgström F, Fritzell P, et al: A randomized, controlled trial of fusion surgery for lumbar spinal stenosis. N Engl J Med 374:1413–1423, 2016

13. Ghogawala Z, Benzel EC, Amin-Hanjani S, Barker FG II, Harrington JF, Magne SN, et al: Prospective outcomes evaluation after decompression with or without instrumented fusion for lumbar stenosis and degenerative grade I spondylolisthesis. J Neurosurg Spine 1:267–272, 2004

14. Ghogawala Z, D'Amato J, Butler WE, Dai F, Terrin N, Magne SN, et al: Laminectomy plus fusion versus laminectomy alone for lumbar spondylolisthesis. N Engl J Med 374:1424–1434, 2016

15. Ghogawala Z, Shaffrey CI, Asher AL, Heary RF, Logvinenko T, Malhotra NR, et al: The efficacy of lumbar discectomy and single-level fusion for spondylolisthesis: results from the NeuroPoint-SD registry: clinical article. J Neurosurg Spine 19:555–563, 2013

16. Guha D, Heary RF, Shamji MF: Iatrogenic spondylolisthesis following laminectomy for degenerative lumbar stenosis: systematic review and current concepts. Neurosurg Focus 39(4):E9, 2015

17. Held U, Burgstaller JM, Wertli MM, Pichieri G, Winkelhofer S, Brunner F, et al: Prognostic function to estimate the probability of meaningful clinical improvement after surgery—results of a prospective multicenter observational cohort study on patients with lumbar spinal stenosis. PLoS One 13:e0207126, 2018

18. Kalágara S, Eltorai AEM, Durand WM, DePasse JM, Daniels AH: Machine learning modeling for predicting hospital readmission following lumbar laminectomy. J Neurosurg Spine 18:49–59, 2018

19. Kanaan SF, Yeh HW, Waitman RL, Burton DC, Arnold PM, Sharma NK: Predicting discharge placement and health care needs after lumbar spine laminectomy. J Allied Health 43:88–97, 2014

20. Karhade AV, Oginik P, Thio Q, Broekman M, Cha T, Gormley WB, et al: Development of machine learning algorithms for prediction of discharge disposition after elective inpatient surgery for lumbar degenerative disc disorders. Neurosurg Focus 45(5):E6, 2018

21. Lønne G, Fritzell P, Hägg Ö, Nordvall D, Gerdhem P, Lagerbäck T, et al: Lumbar spinal stenosis: comparison of surgical practice variation and clinical outcome in three national spine registries. Spine J 19:41–49, 2019

22. Lønne G, Schoenfeld AJ, Cha TD, Nygaard ØP, Zwart JAH, Solberg T: Variation in selection criteria and approaches to surgery for lumbar spinal stenosis among patients treated in Boston and Norway. Clin Neurosurg Neuroplast 156:173–182, 2017

23. Machado GC, Ferreira PH, Yoo RI, Harris IA, Pinheiro MB, Koes BW, et al: Surgical options for lumbar spinal stenosis. Cochrane Database Syst Rev 11:CD012421, 2016

24. Mommaneni PV, Whitmore RG, Curran JN, Ziewacz JE, Wadwha R, Shaffrey CI, et al: Cost-effectiveness of lumbar discectomy and single-level fusion for spondylolisthesis: experience with the NeuroPoint-SD registry. Neurosurg Focus 36(6):E3, 2014

25. Obermeyer Z, Emanuel EJ: Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 375:1216–1219, 2016

26. 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 (Phila Pa 1976) 33:90–94, 2008

27. Raad M, Donaldson CJ, El Daafrawy MH, Sciubba DM, Riley LH, Neuman BJ, et al: Trends in isolated lumbar spinal stenosis surgery among working US adults aged 40–64 years, 2010–2014. J Neurosurg Spine 29:169–175, 2018

28. Ribeiro MT, Singh S, Gaestrin C: “Why should I trust you?” Explaining the predictions of any classifier. arXiv:1602.04938, 2016

29. Scheer JK, Oh T, Smith JS, Shaffrey CI, Daniels AH, Sciubba DM, et al: Development of a validated computer-based preoperative predictive model for pseudarthrosis with 91% accuracy in 336 adult spinal deformity patients. Neurosurg Focus 45(5):E11, 2018

30. 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 19:65–70, 2019

31. Senders JT, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman ML, et al: Natural and artificial intelligence in neurosurgery: a systematic review. Neurosurgery 83:181–192, 2018

32. Senders JT, Zaki MM, Karhade AV, Chang B, Gormley WB, Broekman ML, et al: An introduction and overview of machine learning in neurosurgical care. Acta Neurochir (Wien) 160:29–38, 2018

33. Sigmundsson FG, Jönsson B, Strömqvist B: Outcome of decompression with and without fusion in spinal stenosis with degenerative spondylolisthesis in relation to preoperative pain pattern: a register study of 1,624 patients. Spine J 15:638–646, 2015

34. 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 [epub ahead of print], 2018

35. Staartjes VE, Schröder ML: Class imbalance in machine learning for neurosurgical outcome prediction: are our models valid? J Neurosurg Spine 29:611–612, 2018 (Letter)

36. Staartjes VE, Schröder ML: Effectiveness of a decision-making protocol for the surgical treatment of lumbar stenosis with grade 1 degenerative spondylolisthesis. World Neurosurg 110:e355–e361, 2018

37. Staartjes VE, Serra C, Muscas G, Malduner N, Akeret K, van Niftrik CHB, et al: Utility of deep neural networks in predicting gross-total resection after transsphenoidal surgery for pituitary adenoma: a pilot study. Neurosurg Focus 45(5):E12, 2018

38. Steinmetz MP, Mroz T: Value of adding predictive clinical decision tools to spine surgery. JAMA Surg 153:643, 2018

39. van Buuren S, Groothuis-Oudshoorn CGM: mice: multivariate imputation by chained equations in R. J Stat Softw 45(3):1–67, 2011

Disclosures
The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

Author Contributions
Conception and design: Staartjes, Schröder. Acquisition of data: all authors. Analysis and interpretation of data: Staartjes, Siccoli, Schröder. Drafting the article: Staartjes, Siccoli, Schröder. Critically revising the article: Staartjes, de Wispelaere, Schröder. Reviewed submitted version of manuscript: Staartjes, de Wispelaere, Schröder. Approved the final version of the manuscript on behalf of all authors: Staartjes. Statistical analysis: Staartjes, Siccoli. Administrative/technical/material support: Staartjes, de Wispelaere, Schröder. Study supervision: Schröder.

Supplemental Information
Online-Only Content
Supplemental material is available online.
R code for the ML modeling process. The code was executed in R (version 3.5.1, The R Foundation for Statistical Computing) on a machine running Windows 10 (Microsoft Corp.).
R code: https://thejns.org/doi/suppl/10.3171/2019.2.FOCUS18723.

Correspondence
Victor E. Staartjes: c/o Bergman Clinics, Naarden, The Netherlands. victor.staartjes@gmail.com.