Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

# SpineCloud: image analytics for predictive modeling of spine surgery outcomes

Tharindu De Silva
S. Swaroop Vedula
Alexander Perdomo-Pantoja
Rohan Vijayan
Sophia A. Doerr
Ali Uneri
Runze Han
Michael D. Ketcha
Richard L. Skolasky
Timothy Witham
Nicholas Theodore
Jeffrey H. Siewerdsen

SPIE.

Tharindu De Silva, S. Swaroop Vedula, Alexander Perdomo-Pantoja, Rohan Vijayan, Sophia A. Doerr, Ali Uneri, Runze Han, Michael D. Ketcha, Richard L. Skolasky, Timothy Witham, Nicholas Theodore, Jeffrey H. Siewerdsen, "SpineCloud: image analytics for predictive modeling of spine surgery outcomes," *J. Med. Imag.* **7**(3), 031502 (2020), doi: 10.1117/1.JMI.7.3.031502.

# SpineCloud: image analytics for predictive modeling of spine surgery outcomes

**Tharindu De Silva,[a,†] S. Swaroop Vedula,[b,†] Alexander Perdomo-Pantoja,[c] Rohan Vijayan,[a] Sophia A. Doerr,[a] Ali Uneri,[a] Runze Han,[a] Michael D. Ketcha,[a] Richard L. Skolasky,[d] Timothy Witham,[c] Nicholas Theodore,[c] and Jeffrey H. Siewerdsen[a,b,c,*]**

[a]Johns Hopkins University, Department of Biomedical Engineering, Baltimore, Maryland, United States

[b]Johns Hopkins University, Malone Center for Engineering in Healthcare, Baltimore, Maryland, United States

[c]Johns Hopkins University, School of Medicine, Department of Neurosurgery, Baltimore, Maryland, United States

[d]Johns Hopkins University, School of Medicine, Department of Orthopedic Surgery, Baltimore, Maryland, United States

## Abstract

**Purpose:** Data-intensive modeling could provide insight on the broad variability in outcomes in spine surgery. Previous studies were limited to analysis of demographic and clinical characteristics. We report an analytic framework called "SpineCloud" that incorporates quantitative features extracted from perioperative images to predict spine surgery outcome.

**Approach:** A retrospective study was conducted in which patient demographics, imaging, and outcome data were collected. Image features were automatically computed from perioperative CT. Postoperative 3- and 12-month functional and pain outcomes were analyzed in terms of improvement relative to the preoperative state. A boosted decision tree classifier was trained to predict outcome using demographic and image features as predictor variables. Predictions were computed based on SpineCloud and conventional demographic models, and features associated with poor outcome were identified from weighting terms evident in the boosted tree.

**Results:** Neither approach was predictive of 3- or 12-month outcomes based on preoperative data alone in the current, preliminary study. However, SpineCloud predictions incorporating image features obtained during and immediately following surgery (i.e., intraoperative and immediate postoperative images) exhibited significant improvement in area under the receiver operating characteristic (AUC): AUC = 0.72 ($CI_{95}$ = 0.59 to 0.83) at 3 months and AUC = 0.69 ($CI_{95}$ = 0.55 to 0.82) at 12 months.

**Conclusions:** Predictive modeling of lumbar spine surgery outcomes was improved by incorporation of image-based features compared to analysis based on conventional demographic data. The SpineCloud framework could improve understanding of factors underlying outcome variability and warrants further investigation and validation in a larger patient cohort.

 [DOI: 10.1117/1.JMI.7.3.031502]

**Keywords:** lumbar spine surgery; prediction models; machine learning; image analytics; outcome modeling; data analytics.

Paper 19187SSR received Jul. 25, 2019; accepted for publication Nov. 20, 2019; published online Feb. 18, 2020.

*Address all correspondence to Jeffrey H. Siewerdsen, E-mail: jeff.siewerdsen@jhu.edu

†Authors contributed equally

2329-4302/2020/$28.00 © 2020 SPIE

## 1 Introduction

Lumbar spine surgery suffers from high variability in patient outcomes,[1–9] heterogeneous treatment patterns, and a high frequency of undesirable outcomes, such as revision surgery,[8] failed back surgery,[7] and recurrent lumbar disk herniations.[10] This heterogeneity could be attributed to variable patient characteristics, surgeon preferences, hospital and surgeon volume, and complexity of the intervention.[11–13] However, current understanding of the contributing factors that determine the outcome variability remains ambiguous and underexplored.[14–16] In this context, models to predict patient outcomes provide a valuable resource for patient selection, treatment optimization, and rehabilitative pathway design.[11,17–19]

Previous studies that investigated variability in pain and functional outcomes were limited to analyzing patient demographic and preoperative clinical characteristics as the underlying factors.[11,12,19] While such factors [e.g., age, body mass index (BMI), and smoking status] partially contribute to the outcome variability, additional surgery-specific features could further illuminate and improve the utility of prediction models. Some studies have tended toward development of increasingly complex modeling techniques (e.g., deep neural networks[14,20]) to improve accuracy, particularly for scenarios in which the predictor variables arise from a limited number of available preoperative characteristics (i.e., demographic data). In spite of devising complex machine learning algorithms, model generalizability is challenging to achieve with a limited set of predictor variables that may not fully explain the variability in clinical outcomes. Even though advanced learning methods perform better under conditions of sparse data and/or limited input variables, increasingly complex models also introduce challenges to explainability in the predictions and the ability to identify actionable features.

A large number of images are routinely acquired for perioperative diagnostic and therapeutic evaluation purposes before, during, and after spine surgery. These images play a vital role in diagnosing spinal pathology, determining surgery indications, ensuring safe intervention via intraoperative guidance, and assessing the surgical product in postoperative and follow-up visits. Such images capture distinctive features related to anatomy, pathology, and the changes effected during surgery. In fact, qualitative radiological assessment has been an important basis for outcome measurement prior to (or complementary to) patient-reported outcomes (PROs). Thus, incorporating automatically derived image features in addition to previously investigated patient characteristics could improve the accuracy and utility of prediction models.

Consequently, several studies have evaluated the relationship between individual measures manually obtained from images and patient outcomes in lumbar, thoracic, and cervical spine surgery. Example measures include Modic changes,[21,22] signal intensity-based metrics,[23–25] and the lowest instrumented vertebra take-off angle.[26,27] In contrast to manually obtained features, recent advances facilitate the automatic extraction of several anatomical features from spine imaging.[28–31] Example image-based features include automatic labeling of vertebral levels, global spinal alignment, endplate angles, intervertebral space, number of levels treated, and measurements of surgical instrumentation. To our knowledge, there are no previous studies evaluating statistical models to predict patient outcomes in lumbar spine surgery that use automatically derived quantitative measurements from radiological images.

The objective of this study was to develop and validate a data-analytic framework referred to as "SpineCloud," which includes quantitative features automatically derived from spine imaging in addition to patient demographic and clinical characteristics to predict function and pain improvement after lumbar spine surgery. A learning algorithm was trained and validated to make predictions by combining disparate features based on demographics and image analytics. In this work, we used a boosted decision tree algorithm that inherently handles ordinal and continuous variables in a common predictor variable space. The prediction task was modeled as a binary classification to estimate functional improvement or nonimprovement at 3- and 12-month intervals after surgery.

## 2 Materials and Methods

### 2.1 *Image and Data Analytics*

SpineCloud incorporates a variety of patient-specific variables, including automatically extracted image features detailed below in combination with conventional demographic, clinical,

and functional/pain outcomes. Automated analytics from radiological imaging is a core, novel constituent of SpineCloud. Image-derived analytics, illustrated in Fig. 1, were initiated with a machine-learning based vertebra annotation algorithm that automatically identified the centroid of each vertebra in computed tomography(CT) images.[32,33] Several features related to spine morphology were computed using the vertebral centroid as the input to algorithms. These included automatic quantification of the local and global curvature of the spine by extracting the vertebra endplate surfaces.[34] Spinal curvature measurements were extracted from digitally reconstructed radiographs after forward-projecting the vertebral endplate surfaces. Using these algorithms, vertebral endplate angle (EP), local curvature (LC, defined as the difference between two adjacent endplate angles), intervertebral distance (IVD), and lumbar lordosis (LL) were quantified for each CT image. The measurements were calculated using eight vertebra levels ranging from T11 to S1 for 23 measurements (i.e., 8 EP + 7 LC + 7 IVD + 1 LL) per image. In addition, an automatic segmentation algorithm based on continuous max-flow optimization[35] extracted vertebral bodies and any instrumentation present in the images. The number of levels treated and the length of the surgical construct were calculated from the segmentations for two features per surgery.

The quantitative features were calculated using images acquired at three time points before and after surgery [i.e., preoperative, 0-month (within 10 days) postsurgery, and 3-month postsurgery for a total of $23 \times 3 = 69$ features]. The changes in measurements were also calculated at 0- and 3-month time points relative to the preoperative state for a total of $23 \times 2 = 46$ features. Measurements were derived at each time point if the images were available within the relevant time period. If multiple images were available, the average of the (continuously quantitative) measurements computed from individual images was taken for that time point. A total of 117 (i.e., $69 + 46 + 2$) image analytic features were computed per surgical procedure for each patient.

In addition to image analytics, SpineCloud incorporated patient-specific demographic data extracted at baseline from medical charts. Such demographic/clinical features included age, sex, BMI, history of hypertension, diabetes, bone pathology (osteoporosis, osteopenia, or both), any prior spine surgery, and current or past exposure to smoking. Other relevant variables such as indication for surgery, procedure performed, and the vertebral levels targeted during surgery were extracted from the operative note. A physician and epidemiologist reviewed the medical charts to abstract functional outcomes at 3 and 12 months postsurgery using the modified

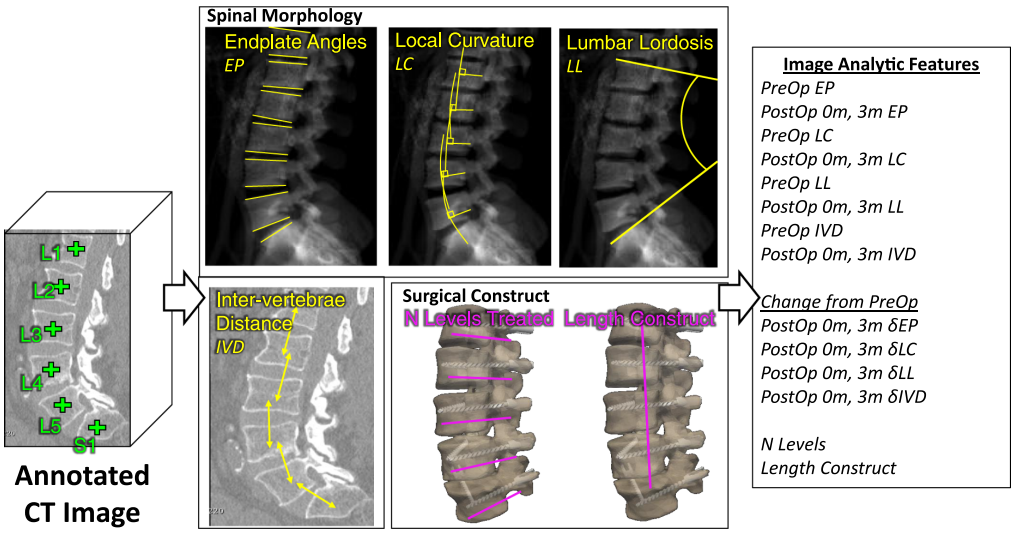


**Fig. 1** Calculation of automatic image analytics from spine CT at preoperative (PreOp), postoperative (PostOp) 0-month, and 3-month time points. Image analytic features included endplate angles (EP), LC, LL, IVD, number of levels, and the length of the construct.

Japanese Orthopedic Association (mJOA) and Nurick functional grading scales[36] and Likert pain scale.

## 2.2 *Prediction Modeling*

Prediction models in SpineCloud were developed using learning algorithms to predict improvements in the outcome using demographic and image-analytic predictor variables. Outcome improvement was a binary prediction on whether the patient improved at a given postoperative time point compared to the preoperative assessment.

A boosted decision tree ensemble classification method as shown in Fig. 2 was used to recursively partition the predictor variable space using a hierarchy of binary decision trees to best represent the variability in outcome improvement. A decision tree ($h_t$) combines multiple predictor variables ($x_n$) to classify the predicted binary output on whether the patient improved ($P$) or not-improved ($N$) an outcome after surgery. The boosting algorithm integrates multiple such decision trees $[h_t(x)]$ with different weights ($\alpha_t$) to construct the final classifier $[f(x)]$. Classification error $\in_t$ is minimized during training to align the predicted output $h_t(x_n)$ from input predictor variables ($x_n$) and the corresponding ground-truth outcome ($y_n$). Training was performed with a learning rate of 0.1 and a maximum of 30 partitions in a single decision tree. The models were trained and evaluated via leave-one-out cross validation. After the training was completed, the resulting decision trees were analyzed to understand which features contributed strongly to the outcome predictions. Feature importance was quantified within the classifier based on how frequently it was used to split the data and the contribution of the split in reducing the prediction error during training.[37]

Table 1 illustrates the prediction models developed and tested in this study with multiple variations on the features used as input to the model and the specific time points for which the outcome was predicted. Models for preoperative predictions were trained using demographics and patient characteristics (denoted as **D**) and with the addition of image analytic features (denoted as SpineCloud, **SC**). SpineCloud predictions using features derived from preoperative image data are denoted as $\mathbf{SC_{pre}}$, and with features derived from images acquired intraoperatively or immediately postoperative (within 10 days) are denoted as $\mathbf{SC_{0m}}$. Predictions made at 3 months after surgery used either a combination of demographics and outcomes at 3 months (denoted as $\mathbf{D} + \mathbf{O_{3m}}$) or a combination of demographics, image features, and outcomes at 3 months (denoted as $\mathbf{SC_{3m}} + \mathbf{O_{3m}}$). The 3-month outcomes were predicted using data at the preoperative and immediate postoperative (0 month) time points, and 12-month outcomes were predicted using data at the preoperative and postoperative 0- and 3-month time points.

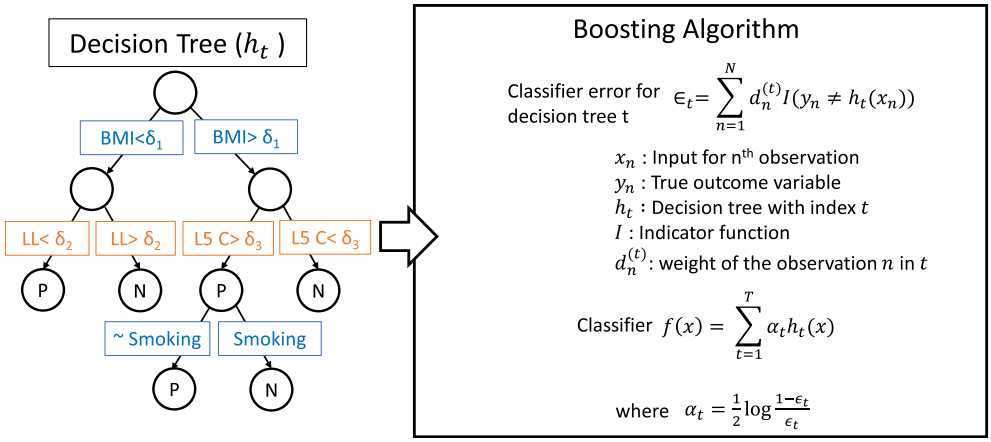


**Fig. 2** A boosted decision tree learning algorithm was used as the predictive model. Multiple binary decision trees were combined using the boosting algorithm to construct the classifier.

**Table 1** Summary of the predictive models developed and investigated in this study.

| | | Features included in a particular model | | | | |
|---|---|---|---|---|---|---|
| | | | Image analytics | | | Outcomes |
| Model | Notation | Demographic/ clinical | PreOp | Immediate PostOp (up to 10 days) | PostOp (up to 3 months) | 3 months |
| Demographic | **D** | ✓ | — | — | — | — |
| | $\mathbf{D} + \mathbf{O_{3m}}$ | ✓ | — | — | — | ✓ |
| SpineCloud | $\mathbf{SC_{pre}}$ | ✓ | ✓ | — | — | — |
| | $\mathbf{SC_{0m}}$ | ✓ | ✓ | ✓ | — | — |
| | $\mathbf{SC_{3m}} + \mathbf{O_{3m}}$ | ✓ | ✓ | ✓ | ✓ | ✓ |

Note: D, demographics; SC, SpineCloud; pre, preoperative; ✓ = feature was used in the analysis; — = feature was not used in the analysis; m, months.

## 2.3 *Materials*

The study involved a retrospective cohort with approval from the Institutional Review Board at Johns Hopkins Medical Institutions. Patient imaging data were retrieved from the hospital picture archiving and communication system. A total of 64 patients who underwent 84 lumbar spine surgeries with preoperative and postoperative CT imaging available were included in the cohort. Table 2 shows a descriptive summary of patients used in this analysis. The patient cohort had mean ($\pm$ standard deviation) age of 59.9 years ($\pm$11.8 years) and BMI of 28.8 ($\pm$6.7) with slight

**Table 2** Summary of clinical data set.

| Variable | Summary |
|---|---|
| Number of patients | 64 |
| Underwent one procedure | 46 |
| Underwent two procedures | 16 |
| Underwent three procedures | 2 |
| Sex; *N* (%)[a] | |
| Male | 28 (44%) |
| Female | 36 (56%) |
| Number of procedures | 84 |
| Age; mean (SD)[b] | 59.94 (11.80) |
| BMI; mean (SD)[b] | 28.83 (6.75) |
| Prior history; *N* (%)[b] | |
| Hypertension | 38 (45.24%) |
| Diabetes | 24 (28.57%) |
| Bone pathology | 9 (10.71%) |
| Smoking | 45 (53.57%) |
| Prior lumbar spine surgery | 64 (76.19%) |

Note: SD, standard deviation; BMI, body mass index.
[a]Denominator denotes number of patients.
[b]Denominator denotes number of procedures.

preponderance of women (56%). Approximately 28.1% (18/64) of the patients underwent more than one surgery during the study time period (March 2004 to July 2017), and 76.2% (64/84) of the surgeries had a record of a prior lumbar spine surgery. Improvement in physical function was observed at 3 months in 40.8% and 28.6% of the patients in terms of mJOA and Nurick scales, respectively, and at 12 months in 46.7% and 43.1% of the patients, respectively. Improvement in pain intensity was observed in 72.9% of the patients at 3 months and 77.6% of the patients at 12 months.

Prediction models were evaluated using estimates of accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), and receiver operating characteristic (ROC) area under the curve (AUC), along with corresponding 95% confidence intervals ($CI_{95}$). In computing these estimates, a positive sample constituted a surgery with an improved outcome relative to baseline, and a true positive was identified when the model correctly predicted the outcome improvement for a positive sample. Confidence intervals were computed using the Wilson method for accuracy, sensitivity, specificity, PPV, and NPV,[38] and using bootstrap (5000 iterations) for AUC. Prediction models using different feature subsets in SpineCloud were compared using bootstrap. Statistical significance was evaluated with respect to a $p$-value of 0.05 with no adjustment for multiple testing. MATLAB 2018b (The MathWorks, Inc., Natick, Massachusetts) was used to develop the prediction models, and R (Version 3.5.2; R Core Team, Vienna, Austria, 2018) was used for analysis and statistical testing.

## 3 Results

### 3.1 *Prediction Model Performance at Baseline and Immediately Following Surgery*

Figure 3 (Table 3 in Sec. 6) shows ROC curves comparing the different prediction models for mJOA functional outcome improvement at 3 and 12 months postsurgery. Conventional modeling based on patient demographic and clinical characteristics (**D**) without any image analytic features demonstrated an AUC of 0.49 ($CI_{95}$ = 0.36 to 0.63) for outcome prediction at 3 months and an AUC of 0.32 ($CI_{95}$ = 0.19 to 0.46) at 12 months. Thus, conventionally derived features (**D**) were not predictive of mJOA functional outcome improvement. The SpineCloud model with image analytics derived from preoperative imaging ($\mathbf{SC_{pre}}$) demonstrated a slight improvement in prediction performance with an AUC of 0.54 ($CI_{95}$ = 0.40 to 0.69) at 3 months and an AUC of 0.47 ($CI_{95}$ = 0.30 to 0.62) at 12 months. Outcomes prediction incorporating intraoperative and immediate postop imaging data ($\mathbf{SC_{0m}}$) exhibited markedly improved performance with an AUC of 0.71 ($CI_{95}$ = 0.59 to 0.82) for 3-month outcomes and AUC of 0.69 ($CI_{95}$ = 0.54 to 0.82) for 12-month outcomes. The improvements in AUC for $\mathbf{SC_{0m}}$ were statistically significant compared to both **D** and $\mathbf{SC_{pre}}$ for both 3- and 12-month analyses.

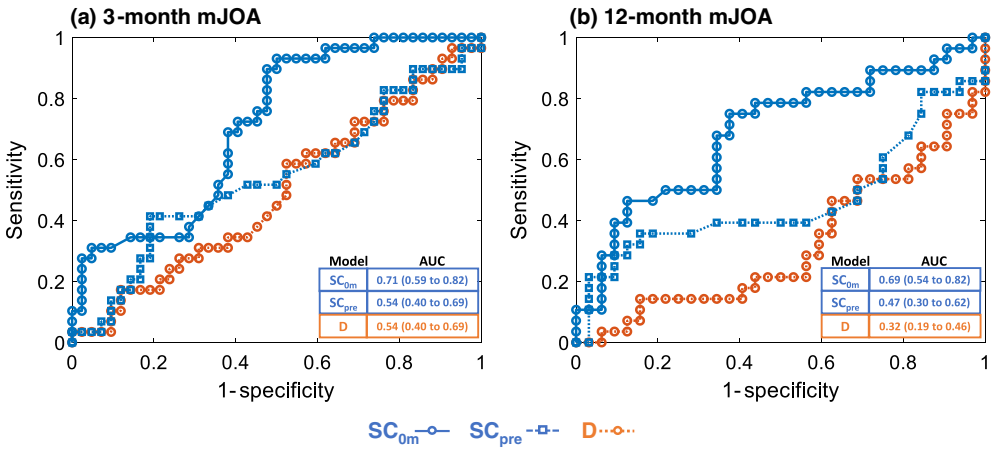


**Fig. 3** ROC curve for SpineCloud prediction of physical function (mJOA) at (a) 3 months and (b) 12 months after surgery.

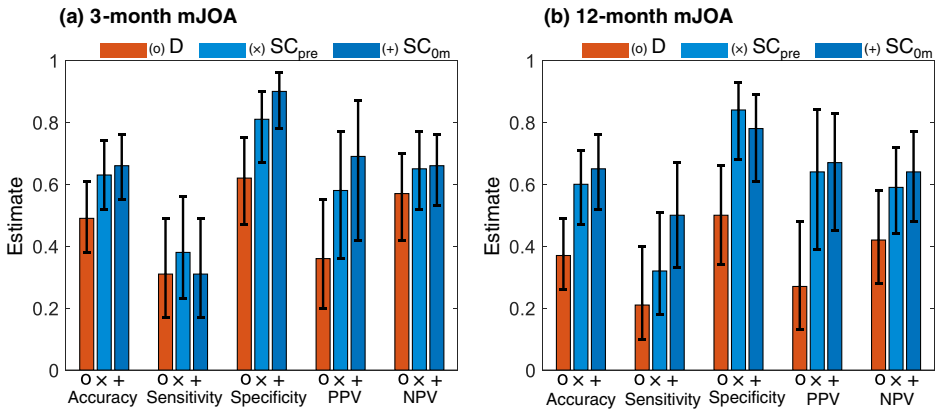


**Fig. 4** Performance measures for SpineCloud prediction of physical function (mJOA) at (a) 3 months and (b) 12 months after lumbar spine surgery.

As shown in Fig. 4 (Table 3 in Sec. 6), prediction model performance assessed in terms of precision, specificity, accuracy, PPV, and NPV showed similar patterns, with $\mathbf{SC_{0m}}$ exhibiting higher performance compared to $\mathbf{D}$ and $\mathbf{SC_{pre}}$. In addition, $\mathbf{SC_{pre}}$ moderately improved the performance compared to $\mathbf{D}$. The $\mathbf{SC_{pre}}$ and $\mathbf{SC_{0m}}$ models both exhibited higher specificity in prediction (0.81 and 0.90, respectively) than the $\mathbf{D}$ model (0.62) at 3 months. However, all three models exhibited relatively low sensitivity to predict outcomes at both 3 and 12 months (0.21 for $\mathbf{D}$, 0.34 for $\mathbf{SC_{pre}}$, and 0.34 for $\mathbf{SC_{0m}}$).

## 3.2 *Prediction Model Performance at 3 Months Postsurgery*

As shown in Fig. 5 (Table 4 in Sec. 6), incorporating information available at 3 months after surgery significantly improved the performance of predicting the 12-month outcome, compared to $\mathbf{D}$, $\mathbf{SC_{pre}}$, and $\mathbf{SC_{0m}}$. These models included 3-month outcomes in addition to patient demographic data and image analytics available at 3 months. Predicting 12-month outcome using the $\mathbf{D} + \mathbf{O_{3m}}$ model yielded an AUC of 0.79 ($CI_{95} = 0.65$ to 0.91), and the $\mathbf{SC_{3m}} + \mathbf{O_{3m}}$ model yielded an AUC of 0.82 ($CI_{95} = 0.70$ to 0.93). Thus, image analytics improved the prediction performance beyond that obtained from 3-month outcome data alone. Establishing the statistical significance of the improvement requires further investigation with a larger data set for 12-month outcomes. Models incorporating postoperative image analytics and outcomes available at 3 months after surgery also demonstrated improved performance, including accuracy, sensitivity, specificity, PPV, and NPV (Fig. 6).

Figure 7 illustrates the importance of various features for predictions in the $\mathbf{D}$ and $\mathbf{SC_{0m}}$ models for 12-month mJOA outcome. In $\mathbf{SC_{0m}}$, features derived from images had a strong influence in the decision trees, compared to traditionally used demographic features. While

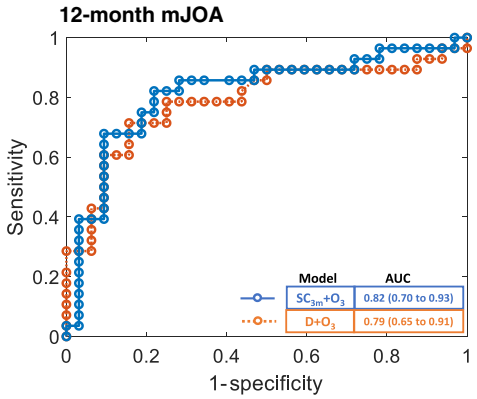


**Fig. 5** ROC curve for SpineCloud including 3-month outcomes to predict physical function (mJOA) at 12 months after lumbar spine surgery.

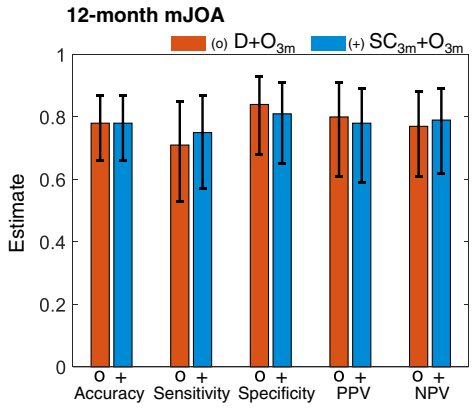


**Fig. 6** Performance measures for SpineCloud including 3-month outcomes to predict physical function (mJOA scale) at 12 months after lumbar spine surgery.

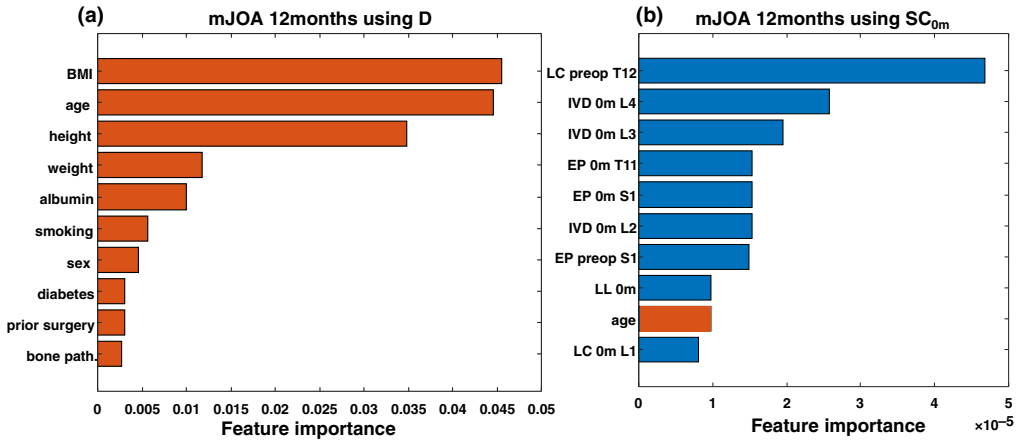


**Fig. 7** Comparison of importance metrics for the 10 most frequently used features in 12-month mJOA outcome prediction using (a) **D** and (b) $\mathbf{SC_{0m}}$. Image-analytic-based features are shown in blue, and patient demographic based features are shown in orange. The strong majority of important features in $\mathbf{SC_{0m}}$ are seen to be image-analytic-based features.

specific features selected from the learning algorithms could be spurious due to the limited number of training examples in this study, the availability of relevant features correlating with outcome variability facilitates the construction of more accurate and generalized models. The propensity of image-based features (viz., 9 out of the top 10 features) indicates that quantitative image analytics has strong influence on the prediction.

Finally, Tables 5–8 summarize estimates of model performance measures to predict the improvement in function on the Nurick scale and the improvement in pain intensity.

## 4 Discussion

A clinical outcomes prediction framework referred to as SpineCloud was shown to incorporate analytical features derived automatically from patient images and thereby improve prediction performance, compared to conventional analysis based on demographic and clinical characteristics alone. The results demonstrated that information extracted from perioperative images of patients undergoing surgery can provide additional relevant, quantitative, and potentially actionable predictive variables to improve performance in modeling postsurgery outcomes. The overall performance of the prediction models is similar to previous studies that developed models predominantly using patient preoperative characteristics. Khor et al.[11] presented the closest relevant study in which models are developed for prediction of pain and function improvement in lumbar spine surgery, reporting performance in terms of a concordance statistic (=0.66 to 0.79). Other studies have demonstrated superior performance in more targeted procedures and patient

cohorts undergoing lumbar discectomy[14] and decompression surgery for lumbar spinal stenosis.[20] While it is challenging to make reliable direct comparison of absolute performance metrics among models involving a limited number of subjects, the improvements gained by image analytics as reported in this study could translate to other works by further improving those models as well—that is, incorporating image analytics in previously reported models based on clinical/demographic data.

The image analytics automatically computed in this study included measurements related to spine curvature, morphology, and surgical construct. The prediction models could benefit further from other analytical measures stemming from previously developed or existing algorithms for surgical planning, spine image registration, and segmentation. For example, automatic planning of pedicle screw trajectories[39,40] and algorithms for detecting the delivered screw trajectory[30] permit the computation of geometric deviation between the delivered screw and the optimal plan. Image registration algorithms that solve for the geometric relationship between preoperative and intraoperative imaging[41] allow computation of relevant morphological metrics from intraoperative imaging. Automatic vertebra and surrounding structure segmentation[35] allows the computation of metrics related to bone density, intervertebral disk space, and texture-based, radiomic-type measurements from spine images. Moreover, while the primary imaging modality in this study was CT, automatic computation of image analytics can be extended to other commonly acquired modalities such as magnetic resonance imaging and radiographs as well as other anatomical sites, such as the cervical and thoracic spine.

Prediction models can be helpful in estimating the likely outcome of the patient both prior to surgery for patient selection and after surgery for guiding postoperative rehabilitative care. Intraoperative and immediate postoperative image analytics, in addition to guiding the rehabilitative pathway, could identify distinct measurements as an immediate quality check during surgery and inform the need to revise the surgical construct based on outcome prediction. Such models could also inform the development of patient-specific intraoperative image guidance and verification solutions driven by predicted outcomes for a given patient. The SpineCloud models presented in this work were constructed using image analytics derived from preoperative diagnostic images as well as immediate-postoperative and postoperative images in addition to patient demographic features. Our experiments demonstrated larger improvements in prediction performance when immediate-postoperative image analytics ($\mathbf{SC_{0m}}$) were included in the model. These analytics captured quantitative measurements of the surgical product (e.g., orientation of instrumentation) and the change effected during surgery (e.g., change in spinal curvature). It is intuitive that the quantitation of the change effected by surgery could have a large impact on patient outcome. Moreover, when predicting 12-month outcomes, incorporation of 3-month outcomes as predictors in the model substantially improved the performance, and these findings are consistent with previous studies investigating the incorporation of early PROs when predicting 12-month outcomes.[42] Thus, the prediction accuracy of the models improved when incorporating more information along the time course after surgery. However, making accurate predictions at earlier stages (i.e., preoperative and intraoperative) will certainly improve the utility of such predictions by informing actionable surgical decisions prior to or during surgery. Therefore, improving the performance of the preoperative models (specifically, by increasing the amount of data used to build the models and/or by implementing effective learning methods) will advance the clinical utility to stages of patient selection and surgical planning.

SpineCloud uses more predictor variables than conventional demographics-based methods in model generation, providing an inherent advantage to search for models that align well with training data. However, having more predictor variables would not necessarily improve model performance and could lead to overfitting due to the lack of generalizability. In addition, a boosted decision tree algorithm builds a parsimonious model and eliminates redundant features that are extraneous to explain variability in training data. Thus, the improvements in SpineCloud compared to demographic-based models with different number of predictor variables as input to the models show the utility of including additional relevant variables to explain outcome variability.

The machine learning algorithm used in this study was a boosted decision tree classifier, and the prediction task was modeled as a binary classification to differentiate patients with improved or nonimproved outcomes. While the binarization simplified the analysis to be feasible with a limited number of patients, the models can in theory be constructed to directly predict numerical

outcome scales such as pain Likert scale (e.g., range: 0 to 10), mJOA scale (e.g., range: 0 to 18), or more comprehensive scales such as Patient-Reported Outcomes Measurement Information System (PROMIS). In the current study, prediction of multiclass outcomes (cf., binary improvement or nonimprovement) is challenged by the limited number of data points (∼60 to 71) and high-dimensional feature space (∼127 features). Recognizing the importance of a more continuous outcome prediction, future work aims to increase the amount of data incorporated in the learning model to directly predict the relevant outcome variables. Moreover, limited data in the current study precluded model evaluation using more rigorous validation methods in lieu of leave-one-out cross validation. Possible overfitting resulting from such limitations may be evident in some of the low AUC findings (some <0.5). The limitations of the current validation approach would induce similar bias across all the tested models, and in this setting, we observed improvement in prediction performance when image analytic features were included in some of the models. More rigorous validation in a larger patient cohort is required to provide confirmatory evidence of the statistical and clinical significance of such improvements. Furthermore, extending the SpineCloud framework to a larger number of patients would improve the model generalizability to derive important predictor variables and provide recommendations to help shape the process of shared decision-making in clinical practice.

While more data, in principle, could alleviate these problems in any learning algorithm, the decision trees resulting from the boosted decision tree classifier helped qualitative interpretation of the importance of predictors (which can be hidden in other predictive models). For example, Fig. 7 illustrates the relative importance of predictor variables used in decision trees for predicting mJOA at 12 months using **D** and $\mathbf{SC_{0m}}$. Clearly, multiple image analytic measures align well with the patterns in outcome variability in $\mathbf{SC_{0m}}$ and contribute to more accurate predictions relative to the predictors in **D**. Such explainability may not be necessary to achieve high levels of prediction/classification performance, but it carries enormous clinical value in guiding the shared decision-making process between surgeons and patients. Model explainability is important for extending the utility of the learning method beyond predictions *per se* to improve the understanding of simple associations among variables and, more importantly, the causal relationships underlying variations in surgical outcomes.

This study involved a retrospective cohort, and its findings are susceptible to potential biases induced in such design. While all the patient demographic and outcome data were derived from retrospective analysis of patient charts and surgeon notes, outcome measurements were limited to scales such as mJOA and Nurick that were attainable via retrospective review. A prospective study design, on the other hand, could benefit from more reliable PRO scales, such as PROMIS,[43] Oswestry Disability Index, or short form 36.[44] Moreover, the clinical findings of this study could be limited by using data arising from a single surgeon at one institution and selecting patients based on the availability of preoperative and postoperative imaging. Limited and missing data presented another drawback in the retrospective study. In the current analysis, mJOA functional outcome was accessible in 71/84 of the procedures at 3 months and 60/84 of the procedures at 12 months. Similarly, image analytics were computed only if an image was available within the relevant time period. Model training was performed excluding procedures without outcome data and any predictors without a valid measurement. A prospective study that records patient outcome at controlled time points and protocols for consistent image acquisition would further strengthen the findings in this paper.

While statistically significant improvements in predicting outcomes were observed for mJOA, predicting the outcome improvements in Nurick scale and pain intensity is not conclusive and requires further investigation with a larger patient cohort. However, possible improvements were observed in predicting Nurick at 3 months using $\mathbf{SC_{0m}}$ compared to $\mathbf{SC_{pre}}$ and **D**. A summary of Nurick predictions can be found in Sec. 6 (Tables 5 and 6). On the other hand, when predicting pain intensity at 3 months, there was no evidence of difference in AUC for algorithms using **D**, $\mathbf{SC_{pre}}$, or $\mathbf{SC_{0m}}$. The summary of findings is detailed in Sec. 6 (Tables 7 and 8).

This study provides preliminary evidence supporting the hypothesis that SpineCloud image analytics are predictive of outcomes in lumbar spine surgery. Overall, image analytic features helped in boosting the prediction performance compared to conventional modeling using patient demographic data alone. When predicting outcomes prior to or immediately after surgery, SpineCloud reached performance up to AUC = 0.71. Such performance is consistent with

similar studies that assessed lumbar spine surgery outcomes in population-based studies.[11] Although the predictive power of such models is far from perfect, it is helpful to understand the extent and relevant importance of these variables in explaining outcomes. In fact, the limited performance of such models using relevant predictor variables is a reflection of the uncertainty associated with clinical decision-making. Thus, incorporating additional features derived from images (or other means) to further improve predictive performance will contribute to explaining outcome variability and mitigating the uncertainty associated with surgical decision-making. Since the pilot study included a fairly limited number of patients, these findings warrant further investigation in larger cohort studies to better understand the extent and clinical significance of the improvements. A larger dataset could also open the model construction to more advanced, robust learning algorithms, including deep learning methods.

## 5 Conclusion

SpineCloud analytics uses high-level image features combined with patient demographics as a foundation for machine learning-based predictive models. Initial studies demonstrated improved prediction of surgical outcome compared to analysis based on demographics alone, providing a framework within which features automatically derived from image data could guide patient selection, surgical planning, and rehabilitative care.

## 6 Appendix

Tables 3 to 6 show estimates and 95% confidence intervals of model performance in predicting physical function in mJOA and Nurick scales and pain outcomes in Likert scale.

**Table 3** Estimates of performance for SpineCloud predicting physical function at 3 and 12 months after lumbar spine surgery—mJOA scale.

| Model | Prediction time point | Accuracy | Sensitivity | Specificity | PPV | NPV | AUC |
|---|---|---|---|---|---|---|---|
| **D** | 3 months | 0.49 (0.38 to 0.61) | 0.31 (0.17 to 0.49) | 0.62 (0.47 to 0.75) | 0.36 (0.20 to 0.55) | 0.57 (0.42 to 0.70) | 0.49 (0.36 to 0.63) |
| **$SC_{pre}$** | 3 months | 0.63 (0.52 to 0.74) | 0.38 (0.23 to 0.56) | 0.81 (0.67 to 0.90) | 0.58 (0.36 to 0.77) | 0.65 (0.52 to 0.77) | 0.54 (0.40 to 0.69) |
| **$SC_{0m}$** | 3 months | 0.66 (0.55 to 0.76) | 0.31 (0.17 to 0.49) | 0.90 (0.78 to 0.96) | 0.69 (0.42 to 0.87) | 0.66 (0.53 to 0.76) | 0.71 (0.59 to 0.82) |
| **D** | 12 months | 0.37 (0.26 to 0.49) | 0.21 (0.10 to 0.40) | 0.50 (0.34 to 0.66) | 0.27 (0.13 to 0.48) | 0.42 (0.28 to 0.58) | 0.32 (0.19 to 0.46) |
| **$SC_{pre}$** | 12 months | 0.60 (0.47 to 0.71) | 0.32 (0.18 to 0.51) | 0.84 (0.68 to 0.93) | 0.64 (0.39 to 0.84) | 0.59 (0.44 to 0.72) | 0.47 (0.30 to 0.62) |
| **$SC_{0m}$** | 12 months | 0.65 (0.52 to 0.76) | 0.50 (0.33 to 0.67) | 0.78 (0.61 to 0.89) | 0.67 (0.45 to 0.83) | 0.64 (0.48 to 0.77) | 0.69 (0.54 to 0.82) |

**Table 4** Estimates of performance for SpineCloud with 3-month outcomes predicting physical function at 12 months after lumbar spine surgery—mJOA scale.

| Model | Prediction time point | Accuracy | Sensitivity | Specificity | PPV | NPV | AUC |
|---|---|---|---|---|---|---|---|
| **$D + O_{3m}$** | 12 months | 0.78 (0.66 to 0.87) | 0.71 (0.53 to 0.85) | 0.84 (0.68 to 0.93) | 0.80 (0.61 to 0.91) | 0.77 (0.61 to 0.88) | 0.79 (0.65 to 0.91) |
| **$SC_{3m} + O_{3m}$** | 12 months | 0.78 (0.66 to 0.87) | 0.75 (0.57 to 0.87) | 0.81 (0.65 to 0.91) | 0.78 (0.59 to 0.89) | 0.79 (0.62 to 0.89) | 0.82 (0.70 to 0.93) |

**Table 5** Estimates of performance for SpineCloud with 3-month outcomes predicting physical function at 12 months after lumbar spine surgery—Nurick scale.

| Model | Prediction time point | Accuracy | Sensitivity | Specificity | PPV | NPV | AUC |
|---|---|---|---|---|---|---|---|
| **D** | 3 months | 0.64 (0.53 to 0.74) | 0.20 (0.08 to 0.42) | 0.82 (0.69 to 0.90) | 0.31 (0.13 to 0.58) | 0.72 (0.59 to 0.82) | 0.54 (0.38 to 0.69) |
| $\mathbf{SC_{pre}}$ | 3 months | 0.66 (0.54 to 0.76) | 0.15 (0.05 to 0.36) | 0.86 (0.74 to 0.93) | 0.30 (0.11 to 0.60) | 0.72 (0.59 to 0.81) | 0.47 (0.32 to 0.63) |
| $\mathbf{SC_{0m}}$ | 3 months | 0.69 (0.57 to 0.78) | 0.15 (0.05 to 0.36) | 0.90 (0.79 to 0.96) | 0.38 (0.14 to 0.69) | 0.73 (0.60 to 0.82) | 0.63 (0.47 to 0.78) |
| **D** | 12 months | 0.57 (0.44 to 0.69) | 0.36 (0.20 to 0.55) | 0.73 (0.56 to 0.85) | 0.50 (0.29 to 0.71) | 0.60 (0.45 to 0.74) | 0.67 (0.52 to 0.80) |
| $\mathbf{SC_{pre}}$ | 12 months | 0.62 (0.49 to 0.73) | 0.24 (0.11 to 0.43) | 0.91 (0.76 to 0.97) | 0.67 (0.35 to 0.88) | 0.61 (0.47 to 0.74) | 0.59 (0.44 to 0.73) |
| $\mathbf{SC_{0m}}$ | 12 months | 0.55 (0.42 to 0.67) | 0.24 (0.11 to 0.43) | 0.79 (0.62 to 0.89) | 0.46 (0.23 to 0.71) | 0.58 (0.43 to 0.71) | 0.63 (0.48 to 0.77) |

**Table 6** Estimates of performance for SpineCloud with 3-month outcomes predicting physical function at 12 months after lumbar spine surgery—Nurick scale.

| Model | Prediction time point | Accuracy | Sensitivity | Specificity | PPV | NPV | AUC |
|---|---|---|---|---|---|---|---|
| $\mathbf{D + O_{3m}}$ | 12 months | 0.84 (0.73 to 0.92) | 0.72 (0.52 to 0.86) | 0.94 (0.80 to 0.98) | 0.90 (0.70 to 0.97) | 0.82 (0.67 to 0.91) | 0.92 (0.83 to 0.98) |
| $\mathbf{SC_{3m} + O_{3m}}$ | 12 months | 0.88 (0.77 to 0.94) | 0.76 (0.57 to 0.89) | 0.97 (0.85 to 1.00) | 0.95 (0.76 to 1.00) | 0.84 (0.70 to 0.93) | 0.94 (0.87 to 0.99) |

**Table 7** Estimates of performance for SpineCloud predicting pain intensity at 3 and 12 months after lumbar spine surgery.

| Model | Prediction time point | Accuracy | Sensitivity | Specificity | PPV | NPV | AUC |
|---|---|---|---|---|---|---|---|
| **D** | 3 months | 0.68 (0.55 to 0.78) | 0.79 (0.65 to 0.89) | 0.38 (0.18 to 0.61) | 0.77 (0.63 to 0.87) | 0.40 (0.20 to 0.64) | 0.66 (0.51 to 0.81) |
| $\mathbf{SC_{pre}}$ | 3 months | 0.73 (0.60 to 0.83) | 0.95 (0.85 to 0.99) | 0.12 (0.03 to 0.36) | 0.75 (0.62 to 0.84) | 0.50 (0.15 to 0.85) | 0.62 (0.46 to 0.76) |
| $\mathbf{SC_{0m}}$ | 3 months | 0.75 (0.62 to 0.84) | 0.95 (0.85 to 0.99) | 0.19 (0.07 to 0.43) | 0.76 (0.63 to 0.85) | 0.60 (0.23 to 0.88) | 0.68 (0.53 to 0.82) |
| **D** | 12 months | 0.72 (0.60 to 0.82) | 0.89 (0.77 to 0.95) | 0.15 (0.04 to 0.42) | 0.78 (0.65 to 0.88) | 0.29 (0.08 to 0.64) | 0.55 (0.40 to 0.70) |
| $\mathbf{SC_{pre}}$ | 12 months | 0.71 (0.58 to 0.81) | 0.89 (0.77 to 0.95) | 0.08 (0.00 to 0.33) | 0.77 (0.64 to 0.86) | 0.17 (0.01 to 0.56) | 0.33 (0.18 to 0.50) |
| $\mathbf{SC_{0m}}$ | 12 months | 0.69 (0.56 to 0.79) | 0.89 (0.77 to 0.95) | 0.00 (0.00 to 0.23) | 0.75 (0.62 to 0.85) | 0.00 (0.00 to 0.43) | 0.25 (0.10 to 0.41) |

**Table 8** Estimates of performance for SpineCloud with 3-month outcomes predicting pain intensity at 12 months after lumbar spine surgery.

| Model | Prediction time point | Accuracy | Sensitivity | Specificity | PPV | NPV | AUC |
|---|---|---|---|---|---|---|---|
| $\mathbf{D} + \mathbf{O_{3m}}$ | 12 months | 0.79 (0.67 to 0.88) | 0.87 (0.74 to 0.94) | 0.54 (0.29 to 0.77) | 0.87 (0.74 to 0.94) | 0.54 (0.29 to 0.77) | 0.73 (0.55 to 0.88) |
| $\mathbf{SC_{3m}} + \mathbf{O_{3m}}$ | 12 months | 0.76 (0.63 to 0.85) | 0.89 (0.77 to 0.95) | 0.31 (0.13 to 0.58) | 0.82 (0.69 to 0.90) | 0.44 (0.19 to 0.73) | 0.69 (0.49 to 0.86) |

## Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.

## Acknowledgments

This work was supported by funding from NIH R01-EB-017226 and the Malone Center for Engineering in Healthcare, Johns Hopkins University.

## References

1. P. Försth et al., "A randomized, controlled trial of fusion surgery for lumbar spinal stenosis," *N. Engl. J. Med.* **374**, 1413–1423 (2016).
2. M. D. Alvin et al., "Spine surgeon treatment variability: the impact on costs," *Global Spine J.* **8**, 498–506 (2018).
3. M. Raad et al., "US regional variations in rates, outcomes, and costs of spinal arthrodesis for lumbar spinal stenosis in working adults aged 40–65 years," *J. Neurosurg. Spine* **30**(1), 83–90 (2018).
4. T. D. Azad et al., "Geographic variation in the surgical management of lumbar spondylolisthesis: characterizing practice patterns and outcomes," *Spine J.* **18**, 2232–2238 (2018).
5. R. A. Deyo, "Trends, major medical complications, and charges associated with surgery for lumbar spinal stenosis in older adults," *JAMA* **303**, 1259 (2010).
6. A. Desai et al., "Variation in outcomes across centers after surgery for lumbar stenosis and degenerative spondylolisthesis in the spine patient outcomes research trial," *Spine* **38**(8), 678–691 (2013).
7. Z. Baber and M. Erdek, "Failed back surgery syndrome: current perspectives," *J. Pain Res.* **9**, 979–987 (2016).
8. F.-C. Kao et al., "Short-term and long-term revision rates after lumbar spine discectomy versus laminectomy: a population-based cohort study," *BMJ Open* **8**, e021028 (2018).
9. R. L. Skolasky et al., "United States hospital admissions for lumbar spinal stenosis: racial and ethnic differences, 2000 through 2007," *Spine J.* **12**, S55 (2012).
10. D. Drazin et al., "Treatment of recurrent disc herniation: a systematic review," *Cureus* **8**(5), e622 (2016).
11. S. Khor et al., "Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery," *JAMA Surg.* **153**, 634 (2018).
12. F. B. Christensen, "Lumbar spinal fusion. Outcome in relation to surgical methods, choice of implant and postoperative rehabilitation," *Acta Orthop. Scand. Suppl.* **75**, 2–43 (2004).
13. P. Farjoodi, R. L. Skolasky, and L. H. Riley, "The effects of hospital and surgeon volume on postoperative complications after lumbar spine surgery," *Spine* **36**, 2069–2075 (2011).
14. V. E. Staartjes et al., "Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar diskectomy: feasibility of center-specific modeling," *Spine J.* **19**(5), 853–861 (2019).
15. P. Fritzell et al., "2001 Volvo award winner in clinical studies: lumbar fusion *versus* nonsurgical treatment for chronic low back pain: a multicenter randomized controlled trial

from the Swedish Lumbar Spine Study Group," *Spine* **26**, 2521–2532; discussion 2532–4 (2001).
16. Z. Ghogawala et al., "Laminectomy plus fusion versus laminectomy alone for lumbar spondylolisthesis," *N. Engl. J. Med.* **374**, 1424–1434 (2016).
17. N. Dietz 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).
18. M. J. McGirt 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* **27**, 357–369 (2017).
19. M. P. Steinmetz and T. Mroz, "Value of adding predictive clinical decision tools to spine surgery," *JAMA Surg.* **153**, 643 (2018).
20. K. F. Spratt et al., "A predictive model for outcome after conservative decompression surgery for lumbar spinal stenosis," *Eur. Spine J.* **13**, 14–21 (2004).
21. A. F. Laustsen and R. Bech-Azeddine, "Do Modic changes have an impact on clinical outcome in lumbar spine surgery? A systematic literature review," *Eur. Spine J.* **25**, 3735–3745 (2016).
22. T. A. Mattei et al., "The 'Lumbar Fusion Outcome Score' (LUFOS): a new practical and surgically oriented grading system for preoperative prediction of surgical outcomes after lumbar spinal fusion in patients with degenerative disc disease and refractory chronic axial low back pain," *Neurosurg. Rev.* **40**, 67–81 (2017).
23. J.-K. Kim et al., "Clinical outcomes and prognostic factors in patients with myelopathy caused by thoracic ossification of the ligamentum flavum," *Neurospine* **15**, 269–276 (2018).
24. W.-C. Lin et al., "The impact of preoperative magnetic resonance images on outcome of cemented vertebrae," *Eur. Spine J.* **19**, 1899–1906 (2010).
25. N. E. Epstein, "High cord signals on magnetic resonance and other factors predict poor outcomes of cervical spine surgery: a review," *Surg. Neurol. Int.* **9**, 13 (2018).
26. H. Koller, O. Meier, and W. Hitzl, "Criteria for successful correction of thoracolumbar/lumbar curves in AIS patients: results of risk model calculations using target outcomes and failure analysis," *Eur. Spine J.* **23**, 2658–2671 (2014).
27. H. Koller et al., "Factors influencing radiographic and clinical outcomes in adult scoliosis surgery: a study of 448 European patients," *Eur. Spine J.* **25**, 532–548 (2016).
28. S.-F. L. Lo et al., "Automatic localization of target vertebrae in spine surgery," *Spine* **40**, E476–E483 (2015).
29. T. De Silva et al., "Utility of the LevelCheck algorithm for decision support in vertebral localization," *Spine* **41**, E1249–E1256 (2016).
30. A. Uneri et al., "Known-component 3D-2D registration for quality assurance of spine surgery pedicle screw placement," *Phys. Med. Biol.* **60**, 8007–8024 (2015).
31. T. De Silva et al., "3D-2D image registration for target localization in spine surgery: investigation of similarity metrics providing robustness to content mismatch," *Phys. Med. Biol.* **61**, 3009–3025 (2016).
32. M. Schwier et al., "Automated spine and vertebrae detection in CT images using object-based image analysis," *Int. J. Numer. Method Biomed. Eng.* **29**, 938–963 (2013).
33. J. H. Siewerdsen et al., "Automatic vertebrae localization in spine CT: a deep-learning approach for image guidance and surgical data science," *Proc. SPIE* **10951**, 109510S (2019).
34. J. H. Siewerdsen et al., "Automatic analysis of global spinal alignment from spine CT images," *Proc. SPIE* **10951**, 1095104 (2019).
35. T. De Silva et al., "Registration of MRI to intraoperative radiographs for target localization in spinal interventions," *Phys. Med. Biol.* **62**, 684–701 (2017).
36. K. K. Revanappa and V. Rajshekhar, "Comparison of Nurick grading system and modified Japanese Orthopaedic Association scoring system in evaluation of patients with cervical spondylotic myelopathy," *Eur. Spine J.* **20**(9), 1545–1551 (2011).
37. H. Ishwaran, "Variable importance in binary regression trees and forests," *Electron. J. Stat.* **1**, 519–537 (2007).

38. A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, New Jersey (2013).
39. J. Goerres et al., "Spinal pedicle screw planning using deformable atlas registration," *Phys. Med. Biol.* **62**, 2871–2891 (2017).
40. RC. Vijayan et al., "Automatic trajectory and instrument planning for robot-assisted spine surgery," *Proc. SPIE* **10951**, 1095102 (2019).
41. M. D. D. Ketcha et al., "Multi-stage 3D-2D registration for correction of anatomical deformation in image-guided spine surgery," *Phys. Med. Biol.* **62**(11), 4604–4622 (2017).
42. R. L. Skolasky et al., "Does reduction in sciatica symptoms precede improvement in disability and physical health among those treated surgically for intervertebral disc herniation? Analysis of temporal patterns in data from the spine patient outcomes research trial," *Spine J.* **18**, 1318–1324 (2018).
43. D. Cella et al., "The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks:
2005–2008," *J. Clin. Epidemiol.* **63**, 1179–1194 (2010).
44. L. Lins and F. M. Carvalho, "SF-36 total score as a single measure of health-related quality of life: scoping review," *SAGE Open Med.* **4**, 2050312116671725 (2016).

Biographies of the authors are not available.