Variability in the utility of predictive models in predicting patient-reported outcomes following spine surgery for degenerative conditions: a systematic review

Nicholas Dietz, BS,¹ Mayur Sharma, MD, MCh,¹ Ahmad Alhourani, MD,¹
Beatrice Ugiliweneza, PhD, MSPH,¹ Dengzhi Wang, MS,¹ Miriam A. Nuño, PhD,²
Doniel Drazin, MD,³ and Maxwell Boakye, MD, MPH, MBA¹

¹Department of Neurosurgery, University of Louisville, Kentucky; ²Department of Public Health Sciences, Division of Biostatistics, University of California Davis, Davis, California; and ³Evergreen Hospital Neuroscience Institute, Kirkland, Washington

OBJECTIVE There is increasing emphasis on patient-reported outcomes (PROs) to quantitatively evaluate quality outcomes from degenerative spine surgery. However, accurate prediction of PROs is challenging due to heterogeneity in outcome measures, patient characteristics, treatment characteristics, and methodological characteristics. The purpose of this study was to evaluate the current landscape of independently validated predictive models for PROs in elective degenerative spinal surgery with respect to study design and model generation, training, accuracy, reliability, variance, and utility.

METHODS The authors analyzed the current predictive models in PROs by performing a search of the PubMed and Ovid databases using PRISMA guidelines and a PICOS (participants, intervention, comparison, outcomes, study design) model. They assessed the common outcomes and variables used across models as well as the study design and internal validation methods.

RESULTS A total of 7 articles met the inclusion criteria, including a total of 17 validated predictive models of PROs after adult degenerative spine surgery. National registry databases were used in 4 of the studies. Validation cohorts were used in 2 studies for model verification and 5 studies used other methods, including random sample bootstrapping techniques. Reported c-index values ranged from 0.47 to 0.79. Two studies report the area under the curve (0.71–0.83) and one reports a misclassification rate (9.9%). Several positive predictors, including high baseline pain intensity and disability, demonstrated high likelihood of favorable PROs.

CONCLUSIONS A limited but effective cohort of validated predictive models of spine surgical outcomes had proven good predictability for PROs. Instruments with predictive accuracy can enhance shared decision-making, improve rehabilitation, and inform best practices in the setting of heterogeneous patient characteristics and surgical factors.

<https://thejns.org/doi/abs/10.3171/2018.8.FOCUS18331>

KEYWORDS predictive models; patient reported outcomes; degeneration; spine surgery

THE shift toward value-based interventions to combat the rising costs of healthcare requires the adoption of quantitative quality measures.¹² In spine surgery, patient-reported outcomes (PROs) have come to the forefront as quantitative measures to evaluate pain, functional ability, and quality of life following spine surgery.^{21,34} While several spinal interventions have been shown to improve PROs through controlled clinical studies,^{26,27} individual outcomes can be heterogeneous. The ability

to predict individual outcomes can facilitate clinical decision-making in patient selection and managing patient expectations.

Ideal PRO tools are associated with high accuracy, reproducibility, and ease of use for wide applications in the clinical setting. Recently, there have been several efforts across multiple groups to develop predictive models for PROs in spine surgery. Incorporation of large patient registries that sample a more varied population improves

ABBREVIATIONS ASA = American Society of Anesthesiologists; AUC = area under the curve; JOABPEQ = Japanese Orthopaedic Association Back Pain Evaluation Questionnaire; LBOS = Low Back Outcome Scale; MCID = minimum clinically important difference; NRS = numeric rating scale; ODI = Oswestry Disability Index; PICOS = participants, intervention, comparison, outcomes, study design; PRO = patient-reported outcome; QOD = Quality Outcomes Database; QUADAS = Quality Assessment of Diagnostic Accuracy Studies; ROC = receiver operating characteristic; RTW = return to work; VAS = visual analog scale.

SUBMITTED June 29, 2018. **ACCEPTED** August 8, 2018.

INCLUDE WHEN CITING DOI: 10.3171/2018.8.FOCUS18331.

generalizability and validity of recent models. With the increased popularity of predictive models, a critical view of how they are derived and validated is needed as they will soon be incorporated in clinical workflow and decision-making. The present study represents the first of its kind to evaluate the landscape of predictive models in PROs for function and pain following elective spine surgery (cervical and lumbar). In this systematic review, we evaluated currently available predictive models with regard to accuracy, how they were created and validated, c-indices, and misclassification rates. With increased utilization and reliance on PROs, predictive models represent useful tools to guide future incorporation and measured applicability of their benefit to clinical practice.

Methods

Data Extraction

We framed the search around a PICOS (participants, intervention, comparison, outcomes, study design) model to define the population of interest used for predictive model creation and the nature of the studies performed to yield a comprehensive and reproducible topic search (see *Search Criteria*). We analyzed those articles that included only validated predictive model creation of PROs after spine surgery.

PICOS Outline

- Participants: operative adult patients ≥ 18 years of age undergoing spinal surgery, including spinal fusion, laminectomy, laminoplasty, and discectomy procedures.
- Intervention: cervical or lumbar spine surgery with at least 3 months of follow-up.
- Comparison: inclusive for control, nonsurgical management, and different surgical procedures.
- Outcomes: PROs, including metrics from functional, pain, and quality-of-life instruments, and certain related, indirect measures, such as return to work (RTW).
- Study design: inclusive of prospective and retrospective studies.

Search Criteria

We followed the PRISMA guidelines (2009) to construct the framework of the systematic review and conducted a search on May 22, 2018, using PubMed and Cochrane databases, limiting the search to articles published between 1980 and 2018.²⁴ Further, studies needed to report a standardized evaluation metric of accuracy with the inclusion of the area under the curve (AUC) or misclassification statistics. Additional articles were incorporated from the references of those articles identified in the searches. We used keyword and MeSH (Medical Subject Headings) terms for predictive outcomes to include the following terms with numbered iterations for the 2 databases as shown below, with the following eligibility criteria: follow-up for 3 months or longer, publication since 1980 in the English language.

1. PubMed: “Predictive model” AND “spine surgery” AND “PROs” (n = 21; 4 included). Ovid: “predictive models,” AND “spine surgery, spinal fusion, lumbar

vertebrae, cervical vertebrae” AND “patient reported outcome measures” (n = 0).

2. PubMed: (((predictive model OR predictive models) AND (spine surgery OR spinal surgery)) AND function (n = 238; 3 included, 3 redundant). Ovid: “predictive models,” AND “spine surgery, spinal fusion, lumbar vertebrae, cervical vertebrae” AND “function” (n = 1; 0 included).
3. PubMed: “Predictive model” AND “spine surgery” AND “pain” (n = 154; 5 included, 5 redundant). Ovid: “predictive models,” AND “spine surgery, spinal fusion, lumbar vertebrae, cervical vertebrae” AND “function” (n = 2; 1 included, 1 redundant).
4. PubMed: (((predictive OR prediction) AND (model OR models OR modeling OR tool*)) AND ((“patient reported outcome” OR PRO OR pain OR function OR Functional improvement OR Patient Satisfaction[Mesh] OR Treatment Outcome[Mesh] OR Quality of Life[Mesh] OR Patient Reported Outcome Measures[Mesh] OR Patient Outcome Assessment[Mesh] OR Patient Satisfaction OR Treatment Outcome OR Quality of Life OR Patient Reported Outcome Measures OR PROMs OR Patient Outcome Assessment))) AND (((spine OR spinal) AND surgery) OR (Spine AND Surgical Procedures, Operative) OR (Spine/Surgery) OR (Spinal Fusion AND Surgical Procedures, Operative) OR (Spinal Fusion/Surgery)) (n = 698; 7 included, 5 redundant).

Inclusion and Exclusion Criteria

In order to be included in the analysis, studies had to involve adult patients (age ≥ 18 years) who underwent elective spine surgery for a degenerative condition with at least a 90-day follow-up period and have a sample size of 50 patients or more. Randomized controlled trials and prospective and retrospective studies were eligible for inclusion as were studies involving discectomy, fusion, and decompression of cervical and/or lumbar vertebrae. Studies of spinal deformity, infection, or trauma; animal studies; studies involving pediatric patients; and studies without full text were excluded.

Data Evaluation

We used the QUADAS (Quality Assessment of Diagnostic Accuracy Studies) tool to evaluate risk bias and results applicability of the studies according to the 2003 guidelines as shown in Table 1. A total of 14 questions from the QUADAS survey were addressed for each study incorporated into the final analysis; these questions cover patient selection, index test, reference standard, and timing. The area under the receiver operating characteristic (ROC) curve (AUC/ROC) is a measure of accuracy for a model that reflects an overlap estimate of the plotted sensitivity and specificity.³⁵ The validation parameters were reported using c-index, c-statistic, or AUC in different studies; these all represent the same measure.

Results

Our search resulted in 1159 publications, 7 of which met the inclusion criteria based on the PRISMA guidelines (Fig. 1). A total of 17 prospectively created predictive

models were identified and evaluated that included PRO measures such as the most commonly assessed Oswestry Disability Index (ODI), numeric rating scale (NRS) for pain, and RTW, with follow-up ranging from 3 to 12 months. A summary of the relevant data is shown in Table 2. Several methods of validation were used across studies, including test cohort validation, in which a set of patient data from 15%–20% of the study population, not incorporated in training the model for linear regression, is used for data input to determine accuracy. Bootstrapping, another method of validation, used random data subsets from the study population for internal validity measures in two of the studies.^{5,13} None of the models had been validated externally.

The majority of the articles that fit our inclusion criteria described c-index analyses—concordance statistics used to assess the correlation of a binary outcome variable's goodness of fit included in a logistic regression model⁷—for the predictive models. Two papers (by Azimi et al.⁸ and Spratt et al.²⁹) reported AUC or classification percentages in their validation analyses. Reported c-index values ranged from 0.47 to 0.79. Two studies reported AUC (0.71–0.84) and misclassification rate (9.9%). Several demographic factors were routinely considered in creating the models, such as age and sex, as well as clinical factors such as BMI, American Society of Anesthesiologists (ASA) physical status classification score, smoking status, and previous spinal surgeries (Table 3). Coefficient inclusion was determined through a priori selection or Bayesian modeling coefficients related to outcome. The number of variables incorporated in each model varied significantly from 77 (39 clinical variables and 38 questionnaire items in the 2015 study by McGirt et al.²³) to 3 (age, present pain intensity, and the Roland-Morris function score in predicting the persistent postsurgical pain score in the study by Hegarty and Shorten¹³) (Tables 2 and 3).

Discussion

Predictive models for PRO in degenerative spine surgery quantitatively anticipate surgical outcomes and are most useful, with high degrees of accuracy, reproducibility, and external validation. We identified a total of 17 models from 7 articles that demonstrate largely good internal validation with c-index (0.47–0.79) and AUC (0.71–0.83) ranges. Moderate to significant heterogeneity was observed in variables considered, number of variables used in the models, selection of variables to include a priori, inclusion versus covariate identification through coefficient strength, model outcomes (e.g., ODI score, NRS score, visual analog scale (VAS) score, and RTW), training methodology to create binary versus continuous outcomes, and validation methodology to include separate cohorts versus bootstrapping. Key predictors considered across PRO models included age, sex, BMI, ASA score, smoking status, and previous spinal surgeries. Also notable were workers' compensation status considered in the 2017 study by Asher et al.⁵ and the 2017 study by McGirt et al.,²² depression in the 2017 study by Asher et al.,⁵ and preoperative ODI in the 2017 study by Asher et al.⁵ and the 2015 study by McGirt et al.²³

TABLE 1. QUADAS criteria for risk of bias for all 7 studies

QUADAS Question	Yes	No	Unclear
Was the spectrum of patients representative of the patients who will receive the test in practice?	7	0	0
Were selection criteria clearly described?	7	0	0
Is the reference standard likely to correctly classify the target condition?	5	2	0
Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	5	2	0
Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	5	2	0
Did patients receive the same reference standard regardless of the index test result?	2	5	0
Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	5	2	0
Was the execution of the index test described in sufficient detail to permit replication of the test?	2	5	0
Was the execution of the reference standard described in sufficient detail to permit its replication?	2	5	0
Were the index test results interpreted without knowledge of the results of the reference standard?	2	3	2
Were the reference standard results interpreted without knowledge of the results of the index test?	0	2	5
Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	7	0	0
Were uninterpretable/intermediate test results reported?	0	7	0
Were withdrawals from the study explained?	6	0	1

QUADAS criteria represented above with tally of ratings for all 7 studies included in the final analysis. The "reference group" was determined to be the validation cohort or bootstrapped validation technique for those studies that incorporated these methods. Bootstrapping for the predictive models produced variable index study and corresponding reference study based on the variables and patients used per trial. The bootstrap studies as well as those studies that only included regression validation methods were determined to be at higher risk of bias than those that used validation cohorts. QUADAS tool obtained with permission from Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J: The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Res Methodol* 3:25, 2003. <https://doi.org/10.1186/1471-2288-3-25>. © 2003 Whiting et al.; licensee BioMed Central Ltd.

Evaluation of Predictive Models

Evaluation of accuracy of models includes an analysis of the c-index or AUC/ROC values. These provide an estimate of the concordance of the goodness of fit that the variables match the logistic regression model. A value of 0.5 indicates 50% or random chance that a predictor outcome will be observed for a subject with the relevant out-

FIG. 1. Flow diagram showing the selection of articles for the systematic review. Data added to the PRISMA template (from Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group [2009]: Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med* 6[7]:e1000097) under the terms of the Creative Commons Attribution License.

come when compared with a subject without the outcome. A value of 1.0 indicates 100% precision that the model will predict that outcome. Generally, 0.7 and above indicates fair to good, and 0.9 indicates very good to excellent discrimination ability for c-index or AUC/ROC.^{30,36} We also consider the number of those in the sample size to determine if this enhances accuracy to identify covariates related to predictive outcome and determining the correct coefficient. While one proposed guideline has been to select a sample that has 10 outcome events per predictor variable (EPV), it has been reported that models that do not meet this threshold can be equally as effective in predictive discrimination.³² To date, no concrete guidelines

exist for sample size needed for predictive model creation. Additionally, we assessed the validation methods for the models to classify internal validity and test the potential generalizability of the model to new samples. Historically, proper external validation has limited the incorporation of predictive models in clinical practice.¹¹ Steyerberg and colleagues³⁰ analyzed predictive model generation for morbidity after acute myocardial infarction and have demonstrated bootstrapping to be the most effective measure of internal validation in which random sampling in the model is used for training and testing. Split-sample, a commonly used method, was found to undervalue true performance capability and have significant variance, but

TABLE 2. Summary of studies focusing on the predictive models for PROs in spine surgeries

Authors & Year	No. of Cases	Pt Age (yrs)	Registry	Diagnosis/Spinal Procedure	Outcomes Measured	Follow-Up (mos)
Spratt et al., 2004	40	Median 59.8	None	Lumbar spinal stenosis/partial laminectomy/arthrectomy (or laminarthrectomy)	VAS, WDI, ODI, LBOS	12
Hegarty & Shorten, 2012	53	Range 18–65	None	Lumbar discectomy	Persistent post-surgical pain using VAS	3
McGirt et al., 2015	1803	Mean 55.92	QOD	Degenerative lumbar condition	ODI, RTW	12
McGirt et al., 2017	7618	Median 60	QOD	Stenosis, spondylolisthesis, disc herniation, revision surgery for disc herniation/lumbar decompression & fusion	ODI, EQ-5D, NRS–back pain, NRS–leg pain	12
Asher et al., 2017	4694	Median 52	QOD	Stenosis, spondylolisthesis, disc herniation, revision surgery for disc herniation/lumbar fusion	RTW	3
Khor et al., 2018	1965	Mean 61.3	Spine SCOAP	Stenosis, spondylolisthesis/lumbar fusion	NRS–back pain, NRS–leg pain, ODI	12
Azimi et al., 2018	133	Median 51.2	None	Lumbar disc herniation/lumbar discectomy	JOABPEQ, ODI	12

Pt = patient; SCOAP = Surgical Care and Outcomes Assessment Program; WDI = Waddell Disability Index.
Despite the retrospective nature of most studies, all data were collected in a prospective fashion.

may be considered in very large samples with EPV of 40 or greater.³⁰

Variability in Prediction Models and c-Indices (or AUC)

Khor and colleagues¹⁸ conducted one of the first state-wide analyses of PROs after lumbar fusion. PRO surveys were collected for ODI scores and NRS scores for back and leg pain 0–60 days preoperatively and at postsurgical months 2, 6, 12, 18, 24, 30, and 36.¹⁸ Model creation was derived from a randomly selected 85% of the cohort, with the remaining 15% used as a validation cohort. Logistic regression models were used to assess trend in 3 variables until 12 months postoperatively. Patient characteristics and select surgical data were related to the imputed PRO values across the preoperative and postoperative periods. Concordance statistics were found on the calibration to be 0.73–0.75 in discrimination and 0.66 for ODI, 0.79 for back pain, and 0.69 for leg pain in the validation cohorts. Improvements in outcomes were defined as a decrease of 15 points in ODI, and a 2-point decrease in the NRS in this study. At 12 months, all PRO measures improved significantly, with the greatest improvement in leg pain (76.5%), back pain (68.5%), and function (58%).¹⁸

In their 2015 article, McGirt and colleagues²³ analyzed data from the NeuroPoint Quality Outcomes Database (QOD), a prospective multicenter national registry dedicated to the formalized tracking of patients who undergo neurosurgical procedures,⁶ in a landmark study with a notably large sample size of 1803 patients who underwent surgery for degenerative lumbar conditions. Logistic regression and Bayesian models were used to create models for complications, readmissions, RTW, and need for inpatient rehabilitation using binary values for categorical variables. The AUC when compared to the validation model was 0.83 for RTW. A validation cohort, composed of 20% of the patient population, was used for internal validation of the models. Additionally, an ODI prediction model was created with a variance of 0.51, demonstrating that almost

half of the variance was not explained by the model. The same group conducted a later study, published in 2017,²² with the largest sample size to date of 7618 participants and demonstrated 12-month postoperative improvement in patients who underwent lumbar surgery using ODI, EQ-5D, and NRS for leg and back pain. In addition to age, smoking status, employment, baseline function, and pain scores, symptom duration and psychological distress were also significant factors related to disability, quality of life, and pain after surgery. Medians and interquartile ranges of the PROs, determined as ordinal dependent variables, were incorporated into 4 multivariable proportional-odds regression models and included patient characteristics and surgical factors. The c-index for the predictive value for the 12-month ODI model is 0.69, 0.69 for the EQ-5D, 0.67 for the 12-month NRS–back pain model, and 0.64 for the NRS–leg pain model.

Asher and colleagues⁵ report a predictive model for return to work, representing an indirect analysis of functionality and pain assessments related to patient quality of life. They found that 82% of patients successfully returned to work in 3 months, with preoperative work status as the main variable influencing likelihood of return to work. Unlike other models for PRO, work-related status accounted the most for the likelihood of a patient returning to work (33.3%). This emphasis on preoperative and other work-related factors, including education level, has been shown in similar models estimating RTW after spinal surgery.^{2,15} It was found that likelihood to return to work was significant reduced for female patients, African Americans versus white patients, those with a history of diabetes, and those with preoperative symptoms lasting longer than 3 months before degenerative spine surgery. The model incorporated linear continuous, binary, and categorical variables and was used to predict the percentage chance of return to work and was validated with a c-index of 0.71. Additionally, the model of McGirt et al. in their 2015 article²³ assessed RTW and showed baseline ODI and narcotic use to have larger negative coefficients

TABLE 3. Outcomes of studies focusing on the predictive models for PROs in spine surgeries

Authors & Year	No. of Cases	Binary or Continuous Outcome	No. of Predictive Models Evaluated	c-Index	AUC	Validation Cohort	Misclass Rate	Variance Compared to Validation	Variables Considered in Model or PPV Measured
Spratt et al., 2004	40	Binary	1	NA	NA	NA	9.9%	NA	Age, sex, non-demographic candidate predictors
Hegarty & Shorten, 2012	53	Continuous	1	0.658	NA	BS	NA	NA	Age, RMF, present pain intensity
McGirt et al., 2015	1803	Continuous	2	NA	0.83	20% of sample	NA	0.47	Age, BMI, no. of previous spine surgeries, no. of levels involved, preop narcotic use, smoking status, ambulatory, Sx duration >1 yr, ethnicity, ASA score, narcotic use in days, private insurance, employment status, ODI, SF-12, MSPQ, VAS-BP, VAS-LP, VAS-NP
McGirt et al., 2017	7618	Continuous	4	0.67–0.69	NA	NA	NA	NA	Age, sex, BMI, race, education level, h/o surgery, smoking, comorbidities, ASA score, Sx duration, indication for surgery, no. of surgical levels, h/o fusion, surgical approach, receipt of workers' comp, liability insurance, insurance, & ambulatory ability
Asher et al., 2017	4694	Continuous	1	0.71	NA	BS	NA	NA	Age; sex; race; smoking status; employment status; occupation; no. of spinal levels; ASA score; BMI; diabetes; osteoporosis; CAD; education; anxiety; depression; dominant Sx; motor deficit; Sx duration; principal spinal diagnosis; workers' comp; liability insurance for disability; insurance type; ambulatory status; lumbar fusion, anterior, posterior, or combined approach; & preop ODI, EQ-5D, & NRS-BP/LP scores
Khor et al., 2018	1965	Continuous	3	0.66–0.79	NA	15% of sample	NA	NA	Age, sex, race, insurance, ASA score, smoking status, diagnoses, prior surgery status, prescription opioid use, asthma, & baseline PRO scores
Azimi et al., 2018	133	Continuous	5	NA	0.714–0.836	NA	NA	NA	LBP, lumbar function, walking ability, social life function, mental health

BP = back pain; BS = bootstrapping; CAD = coronary artery disease; comp = compensation; h/o = history of; LBP = low-back pain; LP = leg pain; Misclass = misclassification; MSPQ = Modified Somatic Performance Index; NA = not available; NP = neck pain; PPV = positive predictive value; RMF = Roland-Morris function score; Sx = symptom(s).

affecting RTW. Employment status or other work-related factors were not reported in that model.

Hegarty and Shorten¹³ created a multivariate logistic regression model for persistent postsurgical pain after analysis from an evaluation at 3 months postoperatively. Persistent postoperative pain symptoms were determined if the patient had not improved by greater than 70% on the VAS from before surgery to after a 5-minute walking test was administered at 90 days after surgery. Six sur-

veys were also given preoperatively and postoperatively to determine clinical improvement. Of 53 total participants in the study, 20 (37.7%) experienced persistent postoperative pain symptoms determined by the McGill score, VAS, and present pain intensity. A total of 33 of 53 participants (62%) did not experience persistent postsurgical pain. A linear regression model was created from continuous variables and the c-index on internal validation by bootstrapping method was found to be 0.658. Akaike's information

criteria, Schwarz's Bayesian model, Cox and Snell's measure, and Nagelkerke's adjusted value were also incorporated to estimate the effect size. The final prediction model was published and included age, Roland-Morris function score, and present pain intensity to predict persistent postsurgical pain.

Azimi et al.⁸ created a predictive model from a sample of 133 patients who underwent lumbar disc herniation to determine minimum clinically important differences (MCIDs) in the ODI and Japanese Orthopaedic Association Back Pain Evaluation Questionnaire (JOABPEQ) after 12 months. A 13-point postoperative improvement on the ODI was defined as a success and was observed in 81.9% of cases, with a mean improvement of 19.6 scale points. A unique AUC cutoff assessment was used to identify the MCID in JOABPEQ score as opposed to using the traditional 20-point improvement on the scale to define surgical success.²⁸ Calculated MCID for JOABPEQ was a change of 19.1 for low-back pain, 21.3 points for the lumbar function subscale, 24.5 points on the walking ability subscale, 14.3 points for the social life function subscale, and 12.8 points on the mental health subscale.

Spratt and colleagues²⁹ investigated outcomes in patients with spinal stenosis, defining a successful outcome as an improvement in at least 3 of 4 PRO measures: VAS, Low Back Outcome Scale (LBOS), and reductions in claudication and leg pain. A total of 21 (58.3%) of 36 patients achieved successful improvement according to this classification. Two binary models were created, including a logistic regression model and a decision-tree method to create a chi-square automatic interaction detection (CHAID) model. The logistic regression model was not included in our analysis for lack of validation methods, however, the CHAID model was included, with a classification rate of 90.1% (29/32 outcomes)—without prospective predictive validation, however. Interestingly, the model identified predictive outcomes to be largely based on patient sex and vascular status through an aorta calcification score correlating with atherosclerotic disease.²⁹ Moreover, aorta calcification as a measure of vascular status was found to be more impactful on PRO after spine surgery than more commonly reported factors and comorbidities, such as classification of stenosis and number of levels operated.²⁹

Current Landscape and Future Directions

Pain related to spinal structures accounts for the largest subset of those experiencing chronic pain, estimated at 54%–80%.²⁰ The lifetime incidence of neck pain has also been shown to be as high as 66%.⁹ The incidence of cervical spine procedures increased from 2002 to 2009, with a total of 1,323,979 procedures being performed between those years according to Oglesby et al.,²⁶ and Katz¹⁶ reported an annual 298,000 lumbar fusion procedures. Given the prevalence of pathology requiring cervical and lumbar spinal procedures, predictive models that guide shared decision-making on projected outcomes are likely widely applicable in spine surgery. PROs, as measures of surgical success, are critical, given the incidence of persistent postsurgical pain, reported to be present in 10%–50% of patients after any surgery.¹⁷

The creation of validated predictive models of spine

surgical outcomes has shown sufficiently good predictability of PROs at 3–12 months following surgery according to the internal validation methods of the c-index and AUC. Instruments with predictive accuracy can enhance shared decision-making, improve rehabilitation, and inform best practices in the setting of patient characteristics and surgical factors considered for elective adult degenerative surgery of the lumbar spine. Those models with good accuracy and simplicity of use may, therefore, prove cost-effective and reliable in clinical settings. Questions remain, however, regarding the degree to which these models can be clinically incorporated and how well they may perform under external validation techniques. Risk-adjusted predictions of surgical improvement project anticipated scores from PRO measures. We acknowledge that successful surgical outcome is not limited to the consideration of single variables or outcomes in isolation, but must be evaluated in context to improve patient-centered discussions regarding surgery and shared decision-making.

Future directions include an analysis of cervical conditions and surgical interventions to demonstrate applicability and success as demonstrated for degenerative lumbar conditions. Hermansen et al.¹⁴ analyzed anterior cervical discectomy and fusion (ACDF) procedures with 10–13 years of follow-up and showed that initial high neck-related pain intensity, nonsmoking status at the time of surgery, and male sex were preoperative factors predictive of good outcome. Also, improvements in pain intensity surpassed those of neck-specific disability, which was associated with psychosocial characteristics.¹⁴ Functional outcomes in surgery for degenerative cervical myelopathy show that nonsmoking status, fewer comorbidities, shorter symptom duration, absence of gait impairment, and younger age have also been related to minimal impairment scores after cervical spinal surgery.^{3,31}

Synthesis and Suggested Guidelines

A limited, but growing set of internally validated predictive models related to PRO measures is available to aid in postsurgical projections and shared decision-making. The models demonstrate marginal heterogeneity in moderately fair accuracy with c-index and AUC between 0.49–0.83. Differing approaches of internal validation are seen across studies, including split-cohort validation of certain percentages or bootstrapping that may also over- or undervalue c-index or AUC values. Consistency of internal validation would benefit model comparison and guide choice of variables and sample selection in future studies. Specifically, we propose that models be trained with bootstrapping internal validation methods as suggested by Steyerberg and colleagues.³⁰ Validation cohorts with sufficiently large sample sizes may also be used to determine predictive accuracy in a new sample separate from the derivation cohort.^{1,33} As shown in the 2015 study by McGirt et al.,²³ even the inclusion of 77 variables may not account for more than 51% of the variance in the model and may risk “over-fitting,” suggesting too many variable coefficients that compete. While more variables may be needed to account for this variance, the nature of the continuous ODI scale also contributes to the inherent variance observed. Indeed, such an extensive model might also be cumbersome.

some in practice. Studies demonstrate a convergence of variables that seem to be important in predicting PROs such as age, sex, BMI, ASA score, smoking status, and previous spinal surgeries. Future studies should report the coefficients to identify those factors that are most influential to the outcome predicted and report c-index, AUC, misclassification index for binary models, and variance. Outcome variables that are continuous should also be used to describe outcome with greatest specificity. When constructing models, follow-up data should also match current data on the variable considered, as ODI, for example, is best approximated at 12 months after surgery, yielding improved long-term individual patient accuracy when compared to 3-month projections.⁴ Once validated internally, we propose that models be used in a new sample to determine external validity with respect to the extent of their accuracy, success, and generalizability, as well as their clinical usefulness. Of all the studies identified, we found that only the 2015 study by McGirt et al.²³ included patient narcotic use in the model. Narcotics could certainly play a role in the perceived postoperative improvement and recovery such that changes in preoperative baseline scores and pain thresholds might confound results in assessment of PRO.²⁴ Future studies of PROs should include analysis of or screening for narcotic use to improve validity. Another issue may be the breadth of surgical indications included in the models, potentially limiting specificity. For example, McGirt and colleagues²² included 7618 patients in the creation of 4 predictive models, and these patients underwent fusion, discectomy, or laminectomy for a variety of indications such as stenosis, herniation, or revision surgery. Outcomes and healing are known to differ based on type of surgery and indication for treatment.^{10,19} We recommend confining analysis to cohorts with a single surgical indication or including patients undergoing only one type of spine procedure to optimize specificity of models to predict PRO most relevant to patient condition. Future studies may investigate the benefit of cohort selectivity in external validation methods.

Study Strengths and Limitations

Strengths of the listed studies include large data sets from large registries such as the NeuroPoint QOD (or N²QOD)^{5,22,23} and statewide, population-based¹⁸ methods of analysis. Limitations include heterogeneity of variables in the generation of predictive models. Differences in ordinal versus continuous variables, for example, affect c-index results, and heterogeneity in variable definitions and implementation limit effective comparison across studies. Spratt and colleagues²⁹ developed a decision-tree approach using a nominal-variable predictive model as opposed to a continuous-variable predictive model. Current models would be strengthened with convergence of items used for prediction. For example, the use of aorta calcification uniquely targets vascular status, which may affect healing, and predicts PROs more accurately than other commonly used variables. Future predictive models may incorporate this metric or explore the economic benefit of achieving this information with a more easily obtained surrogate metric such as smoking status or age. Standards of choosing variables used in the training models also dif-

fered across studies, as McGirt and colleagues²³ acknowledge choosing certain variables a priori from clinical interpretation known to be related to outcome, while most other metrics were determined from Bayesian and linear regression to determine correlative strength of outcome. Additionally, varying definitions of successful outcomes used in model creation compromise standardization of PRO metrics and models for prediction. Variability in length of follow-up may inhibit generalizability of results. For example, 2 of the predictive models were generated from 3-month follow-up data, and outcomes were pain and functional assessments.^{5,13} It has been shown from a sample of 593 patients that 3-month postoperative measures are not predictive of 12-month outcomes; 11.5% of patients who achieved MCID at 3 months after elective lumbar spine surgery did not maintain the MCID at 12 months.²⁶ Similarly, 10.5% of patients who did not meet the threshold MCID at 3 months were able to achieve MCID at 12 months. Overlapping patient cohorts used in different studies (such as McGirt and colleagues' 2015²³ and 2017²² publications) may have an impact on the overall interpretation of pooled data. However, since the number of publications included in our study was small ($n = 7$), pooled analysis was not used; therefore, overlapping of patient cohorts had no effect on our results. Strategies such as using a validated instrument (QUADAS or GRADE [Grading of Recommendations Assessment, Development, and Evaluation]) may be used to mitigate such effects in meta-analysis. Finally, these models fail to incorporate the possibility of adverse events and limit the inherent accuracy of accounting for such cases.

Conclusions

A limited but effective cohort of validated predictive models of spine surgical outcomes have proven good predictability for PROs. Instruments with predictive accuracy can enhance shared decision-making, improve rehabilitation, and inform best practices in setting of patient characteristics and surgical factors. Current models would benefit from external validation methods, convergence of variable types, and standardization of metrics incorporated.

References

- Altman DG, Vergouwe Y, Royston P, Moons KG: Prognosis and prognostic research: validating a prognostic model. **BMJ** 338:b605, 2009
- Anderson PA, Schwaegler PE, Cizek D, Levenson G: Work status as a predictor of surgical outcome of discogenic low back pain. **Spine (Phila Pa 1976)** 31:2510–2515, 2006
- Anderson PA, Subach BR, Riew KD: Predictors of outcome after anterior cervical discectomy and fusion: a multivariate analysis. **Spine (Phila Pa 1976)** 34:161–166, 2009
- Asher AL, Chotai S, Devin CJ, Speroff T, Harrell FE Jr, Nian H, et al: Inadequacy of 3-month Oswestry Disability Index outcome for assessing individual longer-term patient experience after lumbar spine surgery. **J Neurosurg Spine** 25:170–180, 2016
- Asher AL, Devin CJ, Archer KR, Chotai S, Parker SL, Bydon M, et al: An analysis from the Quality Outcomes Database, Part 2. Predictive model for return to work after elective surgery for lumbar degenerative disease. **J Neurosurg Spine** 27:370–381, 2017

6. Asher AL, Speroff T, Dittus RS, Parker SL, Davies JM, Selden N, et al: The National Neurosurgery Quality and Outcomes Database (N²QOD): a collaborative North American outcomes registry to advance value-based spine care. **Spine (Phila Pa 1976)** **39** (22 Suppl 1):S106–S116, 2014
7. Austin PC, Steyerberg EW: Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. **BMC Med Res Methodol** **12**:82, 2012
8. Azimi P, Yazdanian T, Benzel EC: Determination of minimally clinically important differences for JOABPEQ measure after discectomy in patients with lumbar disc herniation. **J Spine Surg** **4**:102–108, 2018
9. Cassidy JD, Carroll LJ, Côté P: The Saskatchewan health and back pain survey. The prevalence of low back pain and related disability in Saskatchewan adults. **Spine (Phila Pa 1976)** **23**:1860–1867, 1998
10. Chang W, Yuwen P, Zhu Y, Wei N, Feng C, Zhang Y, et al: Effectiveness of decompression alone versus decompression plus fusion for lumbar spinal stenosis: a systematic review and meta-analysis. **Arch Orthop Trauma Surg** **137**:637–650, 2017
11. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al: External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. **BMC Med Res Methodol** **14**:40, 2014
12. Graves C, Meyer S, Knightly J, Glassman S: Quality in spine surgery. **Neurosurgery** **82**:136–141, 2018
13. Hegarty D, Shorten G: Multivariate prognostic modeling of persistent pain following lumbar discectomy. **Pain Physician** **15**:421–434, 2012
14. Hermansen A, Hedlund R, Vavruch L, Peolsson A: Positive predictive factors and subgroup analysis of clinically relevant improvement after anterior cervical decompression and fusion for cervical disc disease: a 10- to 13-year follow-up of a prospective randomized study: clinical article. **J Neurosurg Spine** **19**:403–411, 2013
15. Hodges SD, Humphreys SC, Eck JC, Covington LA, Harrom H: Predicting factors of successful recovery from lumbar spine surgery among workers' compensation patients. **J Am Osteopath Assoc** **101**:78–83, 2001
16. Katz JN: Lumbar disc disorders and low-back pain: socioeconomic factors and consequences. **J Bone Joint Surg Am** **88** (Suppl 2):21–24, 2006
17. Kehlet H, Jensen TS, Woolf CJ: Persistent postsurgical pain: risk factors and prevention. **Lancet** **367**:1618–1625, 2006
18. Khor S, Lavalley D, Cizik AM, Bellabarba C, Chapman JR, Howe CR, et al: Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery. **JAMA Surg** **153**:634–642, 2018
19. Liang HF, Liu SH, Chen ZX, Fei QM: Decompression plus fusion versus decompression alone for degenerative lumbar spondylolisthesis: a systematic review and meta-analysis. **Eur Spine J** **26**:3084–3095, 2017
20. Manchikanti L, Boswell MV, Singh V, Derby R, Fellows B, Falco FJ, et al: Comprehensive review of neurophysiologic basis and diagnostic interventions in managing chronic spinal pain. **Pain Physician** **12**:E71–E120, 2009
21. McCormick JD, Werner BC, Shimer AL: Patient-reported outcome measures in spine surgery. **J Am Acad Orthop Surg** **21**:99–107, 2013
22. McGirt MJ, Bydon M, Archer KR, Devin CJ, Chotai S, Parker SL, et al: An analysis from the Quality Outcomes Database, Part 1. Disability, quality of life, and pain outcomes following lumbar spine surgery: predicting likely individual patient outcomes for shared decision-making. **J Neurosurg Spine** **27**:357–369, 2017
23. McGirt MJ, Sivaganesan A, Asher AL, Devin CJ: Prediction model for outcome after low-back surgery: individualized likelihood of complication, hospital readmission, return to work, and 12-month improvement in functional disability. **Neurosurg Focus** **39**(6):E13, 2015
24. Moher D, Liberati A, Tetzlaff J, Altman DG: Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. **J Clin Epidemiol** **62**:1006–1012, 2009
25. Morris BJ, Sciascia AD, Jacobs CA, Edwards TB: Preoperative opioid use associated with worse outcomes after anatomic shoulder arthroplasty. **J Shoulder Elbow Surg** **25**:619–623, 2016
26. Oglesby M, Fineberg SJ, Patel AA, Pelton MA, Singh K: Epidemiological trends in cervical spine surgery for degenerative diseases between 2002 and 2009. **Spine (Phila Pa 1976)** **38**:1226–1232, 2013
27. Parker SL, Asher AL, Godil SS, Devin CJ, McGirt MJ: Patient-reported outcomes 3 months after spine surgery: is it an accurate predictor of 12-month outcome in real-world registry platforms? **Neurosurg Focus** **39**(6):E17, 2015
28. Parker SL, Godil SS, Shau DN, Mendenhall SK, McGirt MJ: Assessment of the minimum clinically important difference in pain, disability, and quality of life after anterior cervical discectomy and fusion: clinical article. **J Neurosurg Spine** **18**:154–160, 2013
29. Spratt KF, Keller TS, Szpalski M, Vandeputte K, Gunzburg R: A predictive model for outcome after conservative decompression surgery for lumbar spinal stenosis. **Eur Spine J** **13**:14–21, 2004
30. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD: Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. **J Clin Epidemiol** **54**:774–781, 2001
31. Swets JA: Measuring the accuracy of diagnostic systems. **Science** **240**:1285–1293, 1988
32. Tetreault L, Kopjar B, Côté P, Arnold P, Fehlings MG: A clinical prediction rule for functional outcomes in patients undergoing surgery for degenerative cervical myelopathy: analysis of an international prospective multicenter data set of 757 subjects. **J Bone Joint Surg Am** **97**:2038–2046, 2015
33. Vittinghoff E, McCulloch CE: Relaxing the rule of ten events per variable in logistic and Cox regression. **Am J Epidemiol** **165**:710–718, 2007
34. Waljee AK, Higgins PD, Singal AG: A primer on predictive models. **Clin Transl Gastroenterol** **5**:e44, 2014
35. Wrdring T, Smith SM: Patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs). **Health Serv Insights** **6**:61–68, 2013
36. Zweig MH, Campbell G: Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. **Clin Chem** **39**:561–577, 1993

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

Acquisition of data: Dietz. Analysis and interpretation of data: Dietz. Drafting the article: Dietz, Sharma, Alhourani. Critically revising the article: Dietz, Sharma, Alhourani, Ugiliweneza, Nuño. Reviewed submitted version of manuscript: Sharma, Ugiliweneza. Statistical analysis: Wang. Administrative/technical/material support: Boakye. Study supervision: Boakye, Dietz, Drazin.

Correspondence

Maxwell Boakye: School of Medicine, University of Louisville, KY. maxwell.boakye@ulpl.org.