Prediction Tools in Spine Surgery: A Narrative Review

Martin Coia Jadresic[1)] and Joseph F Baker[1)2)]

1) Department of Orthopaedic Surgery, Waikato Hospital, Hamilton, New Zealand
2) Department of Surgery, University of Auckland, Auckland, New Zealand

Abstract:

There have been increasing reports on prediction models in spine surgery. Interest in prognostic tools or risk calculators can facilitate shared decision-making about treatment between patients and clinicians.

In recent years, there has been a steady increase in the number of models developed using varying methods. External validation is an essential component of prediction model testing to ensure the appropriate use of these models in populations outside of the developing center.

This narrative review aimed to provide an overview of the literature describing the development and validation of prediction models in the field of spine surgery.

Keywords:

spine, surgery, risk, prognosis, complications, validation

Spine Surg Relat Res 2025; 9(1): 1-10
dx.doi.org/10.22603/ssrr.2024-0189

Introduction

Prediction models in the field of surgery have two major applications: risk adjustment and prognostication. Risk adjustment is a measure of quality appraisal that enables individualization by casemix when comparing a clinician or a center's outcomes with a given standard[1)]. For example, the United States uses the Merit-Based Incentive Payment System to adjust clinician reimbursement based in part on quality outcomes[2)]. Prognostication relates to an individual's predicted risk when undergoing a surgical procedure and is an essential part of the process of informed consent[3)].

There has been a rapid expansion in the publication and dissemination of risk prediction models across numerous fields, including spinal surgery[4)]. A recent systematic review identified more than 30 prediction models for various outcomes following degenerative spine surgery alone[5)]. With the number and influence of these tools only likely to increase, critical appraisal is vital to assisting patients, clinicians, and policymakers in making informed choices about the use of these tools in clinical practice.

Prediction models exist for the risk of complications, reoperation, readmission, length of stay, and clinical prognosis following spine surgery[5)]. In this narrative review, we aimed to provide an overview of the development and validation of prediction models for the risk and adverse events (AEs) of spine surgery.

We selected models based on two recent systematic reviews and an unsystematic review of the literature using the terms "prediction model(s)" and "spine surgery"[5,6)]. We evaluated studies for completeness using the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis guideline and for bias using the prediction model risk of bias assessment tool, which are the preferred tools of the Cochrane collaboration[7-10)]. We searched PubMed, Medline, OVID, EMBASE, Cochrane, and Google Scholar for relevant articles. The adequacy of the articles for inclusion was determined by the first author with guidance of the senior author when in doubt.

Ethics

Ethical approval was not required for this literature review.

Study characteristics

We selected 9 model development and 7 associated external validation (EV) studies (Table 1, 2). Four models have been externally validated: SpineSage and the American College of Surgeons National Surgical Quality Improvement Program Risk Calculator (NSQIP) have both undergone in-

Corresponding author: Joseph F Baker, joseph.f.baker@gmail.com
Received: July 10, 2024, Accepted: September 11, 2024, Advance Publication: October 19, 2024

Table 1. Derivation Studies.

Author, Year	Model	Type	Sample Size	Validation	Age (mean)	Degenerative (%)	Complication rate (%)	Variables	EPV	Calibration measure	AUC	TRIPOD	Risk of bias (PROBAST)
Bilimoria, 2013	NSQIP	LR	1414006	Colon subset	58	Unknown	9	21	6044	Plot	0.72	25/30	Unclear
Lee, 2014	SpineSage	LR	1546	Cross-validation	49	67	23	23	23	None	0.76	25/30	High
McGirt, 2015	None	LR	1803	Cross-validation	56	91	6.6	46	3	χ>0.05	0.82	21/29	High
Ratliff, 2016	RAT	LR	279315	Training/ validation	47	97	14	19	2049	None	0.70	18/31	High
Scheer, 2017	None	ML	557	Training/ validation	58	100	27	45	3	None	0.89	19/29	High
Buchlak, 2017	Seattle Spine Score	LR	136	Training/ validation	63	100	26	12	3	χ=0.89	0.71	22/30	High
Kim, 2018	None	ANN	22629	Training/ validation	60	100	2.2*	12	41	None	0.59–0.71	17/30	High
Han, 2019	SpineAE	ML	1104233	Training/ validation	62	96	25	274	995	Plot α=0, β=1	0.70	21/29	Unclear
Broda, 2020	Universal Spine Surgery Score	LR	177928	Training/ validation	57	100	11	18	1101	None	0.75	21/30	High

EPV, events per variable; AUC, area under the receiver operating characteristic; TRIPOD, transparent reporting of a multivariable prediction model for individual prognosis or diagnosis; PROBAST, prediction model risk of bias assessment tool; LR, logistic regression; ML, machine learning; ANN, artificial neural network; α, Intercept; β, Calibration Slope; χ=Hosmer–Lemeshow Statistic; * Wound Complications Only.

Table 2. Validation Studies.

Author, Year	Country	Model	Independent	Design	Sample size	Age (mean)	Degenerative (%)	Outcomes	Predicted (%)	Observed (%)	Calibration	AUC	TRIPOD	Risk of bias (PROBAST)
Veeravagu, 2017	US	RAT	N	Prospective	4 5	62	88	69	27.3	29	Plot	0.67	24/29	High
Veeravagu, 2017	US	NSQIP	Y	Prospective	257	62	88	69	7.2	29	Plot	0.67	24/29	High
Wang, 2017	China	NSQIP	Y	Retrospective	242	79	100	106	14	44	χ=0.16	0.71	20/29	High*
Kasparek, 2018	Austria	Spine Sage	Y	Retrospective	273	61	83	44	15	16	None	0.71	18/29	High
Han, 2019	US	RAT	N	Retrospective	1104233	62	96	831487	?	25	α=−0.63 β=0.3	0.60	17/28	Unclear
McCarthy, 2020	US	NSQIP	Y	Retrospective	641	65	100	111	7	28	None	0.80	17/28	High
Fatemi, 2021	US	RAT/ Spine-AE	N	Prospective	276	64	89	76	Categorical	28	None	0.64	24/29	High

dependent external validation; the Risk Assessment Tool (RAT) and the Spine Adverse Event (SpineAE) Predictor have been validated by the developing or associated author groups. We identified the sample size and assessment of model performance as major sources of bias (Fig. 1, 2), among others.

Four studies used single or double internal registries, three utilized the NSQIP registry, and two used administrative claims databases. There was a wide range in the size of the derivation cohorts (136 - 14140066). The mean sample size of the validation cohorts was 324. All derivation and four validation studies belonged to the US cohorts. All but

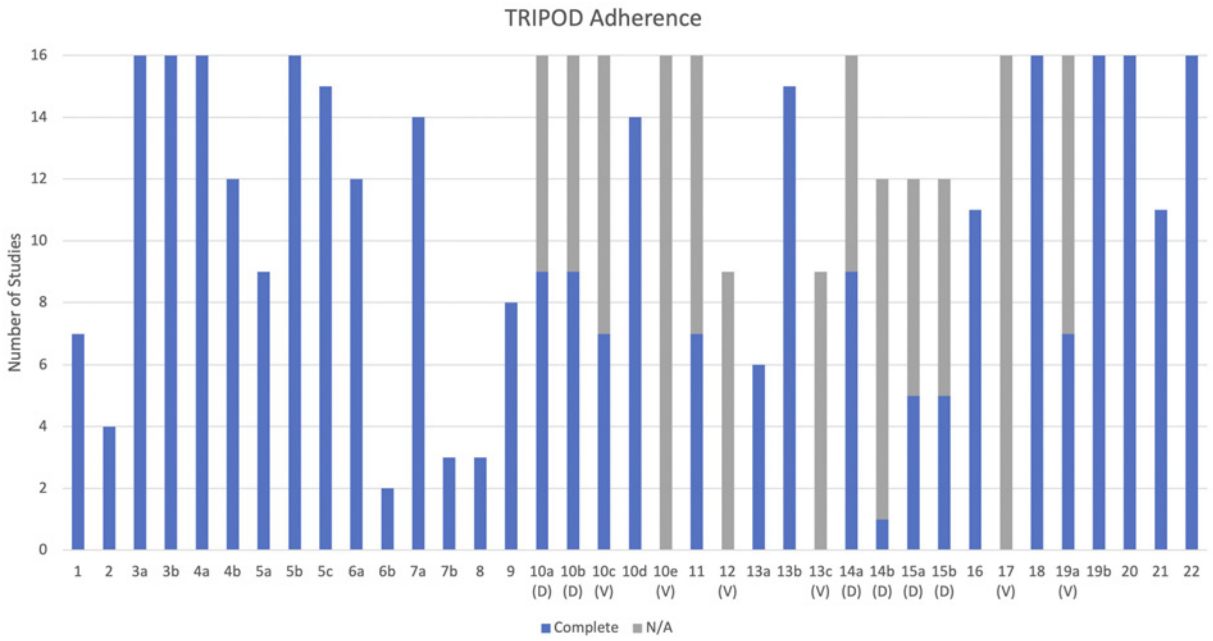

Figure 1. TRIPOD adherence, n=16. (D)=development studies only, (V)=validation studies only.

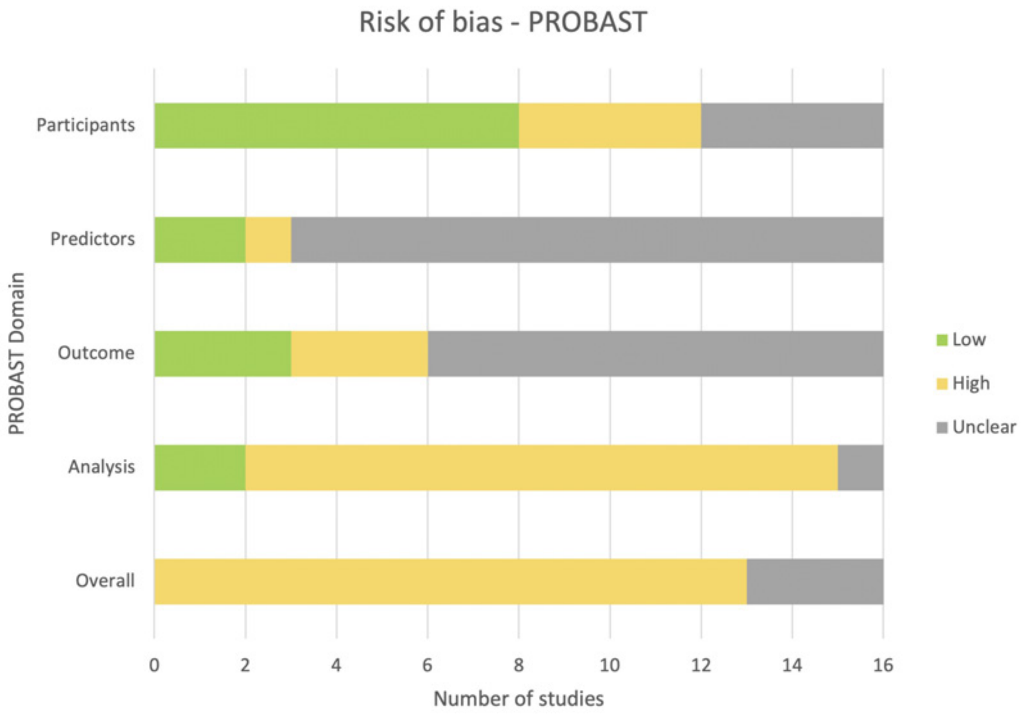

Figure 2. PROBAST domains, n=16.

one study cohort were primarily (>88%) degenerative cases.

Variable selection was largely *a priori* using a combination of previously published studies, expert opinion, and availability. Three studies modified their variable selection for the final model based on univariate analysis, a practice specifically identified as a source of bias in PROBAST. The variables or risk factors used in the model development are summarized in Table 3. The outcomes or complications analyzed are summarized in Table 4.

All nine development papers reported an appropriate detail for model building procedures; however, there was poor reporting of model iterations-no study provided detailed information of discarded models. Meanwhile, six models were derived using logistic regression (LR), two using machine learning (ML), and one using an artificial neural network (ANN).

To avoid overfitting, LR classically required at least 10 participants *with the outcome of interest* per candidate variable (events-per-variable [EPV]), although in the simulation studies, the regression models were robust only at EPVs>

Table 3. Risk Factors Including Patient, Comorbidity, and Surgery-specific Variables Considered in the Development of Each Model.

NSQIP	SpineSage	Risk Assessment Tool	Spine AE Predictor	McGirt	Scheer	Seattle Spine Score
Demographics Age group Sex Functional status Status: emergency or elective ASA class BMI class Laboratory markers Comorbidities: • Chronic steroid use • Ascites within 30 days preop. • Systemic sepsis within 48 h preop • Ventilator dependence • Disseminated cancer • Diabetes • Hypertension needing medical treatment • Previous cardiac event • CHF within 30 days preop • Dyspnoea • Current smoker within 1-year • COPD • Dialysis • Acute renal failure	Age Sex Smoking Alcohol abuse Drug use DM BMI Medical comorbidity: • MI • Non-MI cardiac disease • CHF • CVA • COPD • Asthma • HTN • RA • Preexisting neoplasm • Syncope or seizure • Anaemia • Bleeding disorder Previous spinal surgery Primary spinal diagnosis: • degenerative • trauma • neoplasm • other Level of surgery: • cervical • thoracic • lumbosacral Approach: • anterior • posterior • combined Surgical Invasiveness Index	Pulmonary disorders HTN Cardiac disorders other than HTN DM Neurologic disorders/deficit Hypercholesterolaemia Smoking Cancer/systemic malignancy Gastroesophageal disorder Alcohol/drug abuse Psychiatric disorder *all with multiple ICD-9-CM codes	Procedure details • Diagnosis • Instrumentation • Fusion • Fusion levels • BMP use Comorbidities: • Pulmonary • HTN • Cardiac disorders – not HTN • DM • Neurological disorder • Hypercholesterolaemia • Smoking • Cancer/systemic malignancy • Gastroesophageal disorder • Alcohol/drug abuse • Psychiatric disorder *all with multiple ICD-9-CM codes	Demographics: • Age • Race • Gender Employment Ambulatory status Smoking status Symptoms+duration ASA Patient comorbidities: • CAD • HTN • MI • A. fib • CHF • COPD • Arthritis • Diabetes • Osteoporosis Primary spinal diagnosis Motor deficits Surgery details: • Primary vs. revision • MIS vs. open • No. of levels involved Pre-op narcotic use No. of prior surgeries BMI	Demographics • Age • Sex BMI Number comorbidities CCI ASA Comorbidities: • Anaemia • Depression • Osteoporosis Surgical data: • Primary vs. revision • Single vs. staged • Rod diameter and material • UIV • LIV • Decompression details • Osteotomy details • Interbody fusion • Graft details: iliac crest, BMP • Number of levels fused	Sex Age BMI Comorbidities: • Smoking • HTN • Anxiety • Depression • Diabetes • Bipolar disorder • Parkinson's disease • Cancer • Anaemia

AE, adverse event; A. fib, atrial fibrillation; ASA, American Society of Anaesthesiologists; BMI, body mass index; BMP, bone morphogenetic protein; CCI, Charlson co-morbidity index; CHF, congestive heart failure; CVA, cerebrovascular accident; COPD, chronic obstructive pulmonary disease; DM, diabetes mellitus; HTN, hypertension; LIV, lower instrumented vertebra; MI, myocardial infarction; MIS, minimally invasive surgery; RA, rheumatoid arthritis; UIV, upper instrumented vertebra

20[11,12]. For example, the cohort reported by Scheer et al. had 148 complications, 45 initial variables, and 20 included in the final model (EPV=3) suggesting a lower than ideal number of complications per variable[13]. They reported the highest discrimination of any paper (area under the curve [AUC] =0.89) and high accuracy (87.6%), indicating significant

Table 4. Adverse Events and Outcomes Recorded in Each Tool.

NSQIP	SpineSage	Risk assessment tool	Spine AE Predictor	McGirt	Scheer	Seattle Spine Score
30-day postoperative outcomes: • Mortality • Morbidity • Pneumonia • Cardiac • SSI • UTI • VTE • Renal failure	Cardiac Air embolism; arrest; arrhythmia; CHF; HTN; hypotension; MI; Inadequate fluid therapy; thermodysregulation; other Pulmonary ARDS; empyema; haemothorax; pleural effusion; hypoxia; pneumonia; pneumothorax; PE; respiratory arrest; other GI Ascites; colitis; GI bleeding; ileus; obstruction; pancreatitis; perforation; peritonitis; other Neurological CVA; cerebral hypoperfusion; delirium; diabetes insipidus; electrolyte change; meningitis; SAH/ ICH; seizure; withdrawal – alcohol/narcotic; other Hematologic Coagulopathy; DVT; OR haemorrhage>3 L; transfusion; other Urologic -Foley catheter-related trauma; renal insufficiency; urinary retention; UTI; other	All adverse events, including medical, within 30 days of surgery: • Wound infection • Pneumonia • Renal failure • MI • Pulmonary • Neurological • CHF • PE • DVT • Wound hematoma • Other wound complication • UTI • Cardiac dysrhythmia • Delerium *all relevant subheadings under the given ICD-9-CM code included	• Pulmonary complications • CHF • Neurological complications • Pneumonia • Cardiac dysrhythmia • Renal failure • MI • Wound infection • PE • DVT • Wound hematoma • Other wound complications • UTI • Delerium • Infection • Delerium *all relevant subheadings under the given ICD-9-CM code included	Perioperative course including morbidity and need for rehab Mortality	Intra- and perioperative complications within 6 weeks Classed as major and minor according to Glassman et al. (2007) • Cardiopulmonary • Electrolyte • Gastrointestinal • Implant • Infection • MSK • Neurological • Operative • Radiographic • Renal • Vascular • Wound • Other	Complication events within 30 days of surgery: • Cardiac event including MI • Pneumothorax • Pneumonia • Wound infection • Wound dehiscence • UTI • PE • VTE • Unplanned return to OR • Death

ARDS, adult respiratory distress syndrome; CHF, congestive heart failure; CVA, cerebrovascular accident; DVT, deep vein thrombosis; GI, gastrointestinal; HTN, hypertension; ICH, intracerebral haemorrhage; MI, myocardial infarction; MSK, musculoskeletal; PE, pulmonary embolus; OR, operating room; SAH, subarachnoid haemorrhage; SSI, surgical site infection; UTI, urinary tract infection; VTE, venous thromboembolism

overfitting of the model.

The inclusion of all "candidate" variables is critical-EPV is calculated using all the factors considered in the preliminary analysis, not just those included in the final model[14]. Overfitting can be ameliorated to some degree using techniques such as bootstrapping and cross-validation, but no model with a low EPV specifically addressed this.

Of the nine development studies, six internally validated their models by dividing the data into training and validation subsets. This random split-sample approach is common in predictive modeling but has been criticized as being inefficient as it does not use all available data for model development and is unlikely to demonstrate significant differences between the subsets considering the common data source[4]. Temporal or geographical split-sampling has been suggested to provide a more robust evaluation than sampling from contiguous time periods or regions[15,16]. The alternative approaches include cross-validation (validation between all possible combinations in a split sample) and bootstrapping, which can also be employed to ameliorate model optimism. Three studies used cross-validation, but only one reported the iterative performance of the model at each step.

A complete EV should present an assessment of calibration, discrimination, and clinical utility[17]. Briefly, calibration is a measure of agreement between predicted and observed outcome rates and is often visually reported, whereas discrimination reflects the likelihood of a model returning a higher risk estimate for participants who actually experience the outcome[4]. Assessment of how useful a risk calculator is in aiding clinical decision making can be inferred using decision curve analysis or relative utility[4]. Of these measures, discrimination is the most frequently reported in the orthopedic literature, usually in the form of an area under the receiver-operating characteristic (AUC)[18].

All studies specified the overall AUC of their model, although only seven reported confidence intervals. Eight studies conducted assessment of calibration. However, in 3 studies calibration was performed using the Hosmer-Lemeshow test statistic, which is sensitive to subgroup number and sample size and gives no information on the direction of any miscalibration[14]. Two studies provided appropriate measures of discrimination and calibration. No study included decision curve analysis or other assessment of clinical utility; thus, extrapolation of model usefulness to the clinical setting was not possible.

Validation studies are recommended to include at least 100 participants with outcomes to avoid biased estimates of model performance[14]. Of the six studies, three met this criterion for the primary outcome of any occurring complication; however, two of these three studies assessed secondary outcomes with significantly lower incidence, including one positive assessment of the NSQIP model's mortality prediction despite the very low incidence (n=2)[19]. Adjustments can be made for the analysis of rare outcomes, such as the substituting precision-recall (PR) metrics for the receiver-operating characteristic (ROC) curve; however, no studies reported actual PR metrics despite one including a PR plot[20,21].

NSQIP

The NSQIP Risk Calculator is an LR model derived from a retrospective cohort of 1,414,006 surgical procedures across multiple specialties from 393 hospitals in the United States[22]. Internal validation was performed on a subcohort of colorectal surgical patients for whom a model had previously been derived. The AUC was 0.944 for mortality and 0.816 for morbidity. Calibration was visually reported using the Hosmer-Lemeshow-type calibration curves and calibration plots, and good performance was observed. The calculator is freely available online at https://riskcalculator.facs.org/RiskCalculator/

Wang et al. assessed the performance of the NSQIP calculator in a cohort of 242 patients aged >60 years undergoing exclusive single-level laminectomies at a single center[19]. They found that NSQIP had poor discrimination, poor calibration-in-the-large (Table 2), and poor overall performance (Brier score=0.321) for predicting any complication. The authors concluded that these findings were likely because of the older cohort. Performance analysis of the NSQIP for complication subtype and death was also conducted despite the very low incidence of these outcomes.

Veeravagu et al. found similar overall discrimination for the NSQIP calculator (ROC=0.67) to Wang and systematic risk underestimation (Table 2). However, their study included only 69 patients who had complications. Validation studies should ideally include at least 100 participants with outcomes to avoid biased estimates of model performance[23].

McCarthy et al retrospectively analyzed 641 patients who underwent either cervical or lumbar primary arthrodesis at a single center in the USA[24]. They reported excellent discrimination for NSQIP with an AUC of 0.801. However, the gross calibration was poor, with a predicted event rate of 6.9% and observed rate of 27.5%.

In all validation studies, the NSQIP calculator exhibited poor calibration for spine surgery. A systematic deviation of observed risk from predicted risk indicates the absence of an important predictor from the model[25]. The calculator only allows a single current procedural terminology code to be entered. Thus, procedure complexity and invasiveness may not be appropriately represented, particularly for multilevel or dual-approach surgeries.

Broda et al. addressed these deficiencies by deriving a model using only the subset of the NSQIP cohort that underwent spine surgery for degenerative diseases of the cervical or lumbar spine (n=177, 928). Univariate analysis was employed to find significant predictors, which were then included in the multivariate LR model[26]. This approach, though common, ignores important confounding relationships between predictors in the final model and omits factors that may significantly influence outcome in the multivariate analysis[27].

External validity was evaluated via comparison with the extant NSQIP calculator; however, EV requires testing against cohort independent from that used for development[15,28]. The complication rate was 11.1%, consistent with the low overall morbidity rates in the NSQIP cohort. A final AUC of 0.75 was reported for any complication, with good agreement between the predicted and observed complication rates within the 12 risk groups analyzed.

Kim et al. derived yet another model from the NSQIP cohort of spine surgeries (n=22629) using ANNs[29]. The outcomes analyzed were cardiac and wound complications, venous thromboembolism (VTE), and mortality. The EPV was 2.8 for mortality and 41 for wound complications. The authors employed stepwise LR to select variables for the final model based on coefficient magnitude and significance in the initial regression model. This approach is counterintuitive as ANNs are designed to "learn" associations between many different variables without influence of a prespecified hypothesis[30]. After the split-sample training and validation, the models were tested and compared on a further split sample. LR exhibited a better AUC than ANN for VTE (0.588 vs. 0.567), wound complications (0.61 vs. 0.6), and mortality (0.7 vs. 0.68). The calibration of the model was not assessed.

SpineSage

The SpineSage calculator was derived from a cohort of 1476 adult patients who underwent spine surgery at one of two tertiary centers in the USA[31]. Variables were selected based on previous published analyses and depending on whether they were known confounders or had strong univariate association. Broad definitions of postoperative complications were used, and the definition of a major complication was more restrictive than that used in NSQIP (cardiac arrest, myocardial infarction, acute respiratory distress syndrome, postop hypoxia, pulmonary embolism, bowel perforation, and/or meningitis). The authors employed the 50:50 cross-validation method for internal validation. SpineSage is freely available online at https://depts.washington.edu/spinersk/

The AUC was 0.76 for any medical complication and 0.81 for major complications. Unfortunately, the number of patients with major complications was not reported; thus, the EPV for major complication is unknown. The authors described calibration using the Hosmer-Lemeshow test; however, no statistic was reported, and no calibration plot was included.

The final model was disseminated at the time of publication as an online calculator (SpineSage.com). The calculated risks were stratified using the surgical invasiveness index (SII), a validated tool to measure the invasiveness of spinal procedures based on the number of vertebrae decompressed, instrumented, anterior, and/or posterior to the pedicles[32]. In the development study, SII>25 was the greatest risk factor in the univariate analysis (odds ratio: 6.95, 95% confidence interval: 3.43-10.3, P<0.001), suggesting that granular assessment of surgical factors is critical in the development of an accurate model.

Kasparek et al. investigated the performance of the SpineSage tool for major and all medical complications at a single center in Vienna (n=273)[33]. The study was underpowered with 44 occurrences of the main outcome (any medical complication)[25]. Nine patients had major complications. The model showed similar discrimination to the derivation study (Table 1, 2) with reasonable gross calibration. However, the authors chose to assess calibration on the basis of median risk in arbitrarily defined risk groups, rather than reporting calibration curve or intercept, which was criticized due to the possibility of false assumptions of risk profile within these arbitrary groups[34,35].

Risk assessment tool

The RAT was derived from a large cohort of patients (n= 279, 315) using an administrative claims database (MarketScan)[36]. The PROBAST guidelines suggest that development studies have a higher potential for bias when participant data are collected from existing sources, such as registries, as data are often collected for a purpose other than development (i.e., administrative claims). Historically, there have been important issues with the use of claims data to judge clinical outcomes[37-39]. The authors cited a study they previously published claiming improved sensitivity for capturing AE using longitudinal data captured in MarketScan for 30 days; however, in that study, the 30-day outcomes were simply compared with the complication rates at an earlier time point, not with retrospectively/prospectively collected data based on electronic medical records[23]. The AUCs were presented for the cohort as a whole and for each surgical subgroup. No assessment of calibration was performed, but good calibration is expected in a cohort of this size.

Veeravagu et al. conducted a prospective validation of the RAT (*n=246)*, which showed good calibration-in-the-large but poor discrimination (Table 2)[40]. A calibration curve was reported to show a good fit up to a predicted/actual risk of 40% with progressive overestimation of risk thereafter. The complication incidence was 73. Calibration-in-the-large was very poor for complication subtype; however, the event rate for these complications was extremely low (2-26). The AUC for the NSQIP calculator in this population was 0.67 with systematic risk underestimation in each of the groups analyzed.

For their analysis of accuracy of risk prediction, Veeravagu et al. reclassified the risk predictions into risk groups, namely, low, medium, and high, with each having equal tertiles with the same/similar number of patients. As previously discussed, this reclassification can result in loss of information and risks bias; when reclassification is performed, the ideal method is to reclassify into at least 10 subgroups[14].

Spine AE predictor

The same research group that developed the RAT created another model (SpineAE Predictor) in 2019 using similar methods, with the addition of the Medicare and Medicaid databases[41]. In this new cohort, the previous RAT model exhibited an AUC of 0.6 and poor calibration (intercept −0.63, slope 0.301). Spine AE Predictor is freely available at https://spineaepredictor.shinyapps.io/app-1/

The SpineAE Predictor was prospectively validated by Fatemi et al. at a single institution (n=276). However, the RAT model was also used for some aspects of the analysis, with conflicting information in the study regarding whether the risk thresholds and absolute risk were calculated using

RAT or SpineAE[20]. The authors reported an AUC of 0.64, with no assessment of calibration. The bulk of the analysis was based on the sensitivity/specificity of the tool according to the risk thresholds. The calculator had a sensitivity of 0.38 for high-risk patients (probability of AE>0.278). The accuracy of the calculator ranged from 0.5 to 0.69, depending on the subgroup, with no overall accuracy reported. The authors reported their results as a successful EV.

Notably, the same research group that developed the RAT and the SpineAE Predictor also contributed to and performed the subsequent validation, thereby increasing the risk of bias[42]. Ideally, EV should be conducted by independent actors[18].

McGirt

McGirt et al. used a single-center patient database (*n*=1803) to derive an LR model for complications, readmission, 12-month Oswestry Disability Index, and a composite outcome of any unplanned occurrence[43]. Varying sample sizes from 750 to 1200 were used in model development "depending on outcome of interest." Patients with missing information were excluded from all analyses. Surgical procedure variables were included types of procedure: primary or revision; and, minimally invasive or open. No explicit definition of a complication was given, but the complication rate was 6.6%. A total of 46 model covariates were listed with yielding an EPV of 2.6. For postoperative complications, model performance was assessed using AUC (0.82) and the Hosmer-Lemeshow test for calibration. No value was reported for the latter, but it was stated as >0.05. The model coefficients were reported, but there was no intercept. We found this development study to be at a high risk of bias in each of the four PROBAST domains, with the most relevant point being the degree of overfitting. The model has yet to undergo EV.

Scheer

Scheer et al. used the ML algorithm to derive a model for complications on a multicenter adult spinal deformity database (*n*=557)[13]. The model had an extremely low EPV (3)[44]. It is unclear whether the model returns a binary categorical outcome or an absolute risk, but the former is inferred, considering the absence of information on predicted and observed risks. The AUC and overall "accuracy" were reported, without a supporting confusion matrix detailing metrics, such as sensitivity and specificity. No calibration assessment was conducted. The authors acknowledged that the limitation of this type of model is the lack of transparency-the specific influence of the covariates is unknown, limiting the potential application for both risk adjustment and prognostication.

Seattle spine score

Buchlak et al. derived an LR model (Seattle Spine Score [SSS]) using a retrospective cohort of adult spinal deformity patients (*n*=136)[45]. The authors included variables that were deemed clinically relevant or achieved univariate significance level of 0.2 or less. The low EPV in this study is compounded by evidence suggesting that ML models exhibit high optimism even when the EPV is >200[44]. The SSS can be accessed in the following: https://safetyinspinesurgery.com/publications/seattle-spine-score/

Clinical utility was assessed by comparing predictive performance with and without access with the model based on 100 de-identified cases. Performance was marginally improved with access to the model (61% vs. 50% accuracy), although it appears the cases used in testing were pulled from the same cohort from which the model was developed.

Discussion

In 2014, a systematic review of EV studies reported that 54% of the studies did not mention missing data and that 67% did not evaluate calibration. The authors concluded that the vast majority of EV studies were poorly designed and suggested that this may explain the lack of uptake of predictive models in clinical practice[46]. A more recent systematic review of the general orthopedic literature obtained similar findings and recommended multicenter collaboration to increase the inadequate sample sizes in the current EV studies[18]. Publication bias of negative EV studies is also a concern. Despite the existence of robust, explicit guidance in the areas of model development and evaluation, there is poor awareness of these guidelines - none of the studies we analyzed mentioned an accepted framework.

Furthermore, the focus on discrimination as the primary measure of model performance should be highlighted, particularly in the context of small EV studies where ROC may be particularly misleading[21]. Decision curve analysis was conspicuously lacking from all the studies we analyzed, presenting a further barrier to face validity of the models. This may be due to poor awareness of the technique in the orthopedic field or the perceived difficulty in the reporting and interpretation of decision curves, as the analysis itself is prescriptive and easy to undertake, with freely disseminated step-by-step instructions available for all major statistical packages published online[47,48].

In this review, only models that published a web-based tool have undergone EV. Although many of the LR development studies published model coefficients, none reported an intercept, whereas the ML/ANN models were completely opaque. Even if these metrics are reported, validation requires their incorporation into a statistical package and subsequent generation of risk scores, a step that is obviated by an online tool.

Significant heterogeneity was observed in the definition of a complication. Some models focused on a composite outcome of "any complication," whereas others attempted to predict specific complications. The latter may be more suited to risk stratification-a patient may question why their specific risk of a urinary tract infection or deep vein thrombosis is relevant when more salient risks, such as paralysis and admission to the intensive care unit, are discussed in the clinic. Standardized, composite outcome measures may improve the generalizability and comparison of future models[49]. Furthermore, composite outcome measures of clinically relevant complications allow relatively small cohorts to make statistically valid conclusions. For example, the SSI rate after instrumented fusion ranges from 2.4% to 8.5%-an EV study examining this complication will require a minimum sample size of 1176-4166 to make robust conclusions regarding predictive ability[50].

It is perhaps surprising that such a large database as the NSQIP does not result in the generation of a reliable RAT. However, as pointed out by others, the database covers various surgical procedures rather than just spinal; thus, the influence of variables may be either enhanced or diminished by the inclusion of nonspinal procedures. The approach by Broda et al. on this regard may allow a more refined analysis, but as aforementioned, an improvement of the methodology used is needed[26]. Furthermore, validation of the NSQIP may yield disappointing results when performed in centers outside the USA where healthcare resource and funding models are often strikingly different.

Focusing only on spinal surgical procedures improves the accuracy of model development, as seen in SpineSage. In comparison with SpineSage, which uses the SII to allow a risk scale based on the selected variables, NSQIP does not provide a nuanced assessment of the spinal procedure[32]. However, it is notable that this calculator does not include comorbid conditions, such as fibromyalgia, for which there is clear evidence of a high complication rate[51]. The exclusion of conditions in this manner likely relates to the clinicians' inherent knowledge and experience that they predispose to complications; surgery is avoided, and the condition therefore fails to appear as a variable in any data analysis. This potential pitfall supports the need to continuously add prospectively to datasets allowing interim reanalysis lest previous findings are forgotten and omitted from the risk assessment.

Considering the shortcomings of current EV studies for complications following spine surgery, it is unsurprising that clinical impact studies of any of the described models are absent. Clinicians should exercise caution before adopting routine use of these tools into their daily practice and consider conducting their own validation studies because regional variation in patient demographics and healthcare resources may result in different performances. Future efforts in this field should adhere to established guidelines in the development and validation of these tools and encourage continued assessment of their accuracy across multiple systems.

Conflicts of Interest: JFB: received consultation fees from Fisher & Paykel; teaching honorarium from Medtronic

Sources of Funding: Nil relevant

Author Contributions: MCJ completed literature review and extraction of data, initial drafting; JFB developed concept and methodology, assisted in data interpretation, final drafting

Ethical Approval: Not applicable-this a review article

Informed Consent: Not applicable-no patients involved

References

1. Iezzoni LI. Risk adjustment for measuring health care outcomes. Chicago: Health Administration Press; 1997; 471-516
2. MIPS Explore Measures - QPP. Accessed September 14, 2021. https://qpp.cms.gov/mips/explore-measures
3. Powell JM, Rai A, Foy M, et al. The "three-legged stool": a system for spinal informed consent. Bone Joint J. 2016;98-B(11):1427-30.
4. Moons KG, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.
5. Lubelski D, Hersh A, Azad TD, et al. Prediction models in degenerative spine surgery: a systematic review. Glob Spine J. 2021;11(1_suppl):79S-88S.
6. Romiyo P, Ding K, Dejam D, et al. Systematic review and evaluation of predictive modeling algorithms in spinal surgeries. J Neurol Sci. 2021;420:117184.
7. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD). Ann Intern Med. 2015;162(10):735-6.
8. Wolff RF, Moons KG, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51-8.
9. Abstracts of the 25th Cochrane Colloquium. In: *Abstracts of the 25th Cochrane Colloquium*. Wiley; 2018. doi:10.1002/14651858.CD201801
10. Abstracts accepted for the 26th Cochrane Colloquium. *Cochrane Database Syst Rev*. Published online January 31, 2020. doi:10.1002/14651858.CD201901
11. Courvoisier DS, Combescure C, Agoritsas T, et al. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol. 2011;64(9):993-1000.
12. van Smeden M, de Groot JA, Moons KG, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16(1):163.
13. Scheer JK, Smith JS, Schwab F, et al. Development of a preoperative predictive model for major complications following adult spinal deformity surgery. J Neurosurg Spine. 2017;26(6):736-43.
14. Moons KG, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med. 2019;170(1):W1-33.
15. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19(4):453-73.
16. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338(7708):b605.
17. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925-31.
18. Groot OQ, Bindels BJ, Ogink PT, et al. Availability and reporting

quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. Acta Orthop. 2021;92(4):385-93.

19. Wang X, Hu Y, Zhao B, et al. Predictive validity of the ACS-NSQIP surgical risk calculator in geriatric patients undergoing lumbar surgery. Medicine. 2017;96(43):e8416.

20. Fatemi P, Zhang Y, Han SS, et al. External validation of a predictive model of adverse events following spine surgery. Spine J. 2022;22(1):104-12.

21. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE. 2015;10(3):e0118432.

22. Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. J Am Coll Surg. 2013;217(5):833-42.

23. Veeravagu A, Cole TS, Azad TD, et al. Improved capture of adverse events after spinal surgery procedures with a longitudinal administrative database. J Neurosurg Spine. 2015;23(3):374-82.

24. McCarthy MH, Singh P, Nayak R, et al. Can the American College of Surgeons risk calculator predict 30-day complications after spine surgery? Spine. 2020;45(9):621-8.

25. Vergouwe Y, Steyerberg EW, Eijkemans MJ, et al. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58(5):475-83.

26. Broda A, Sanford Z, Turcotte J, et al. Development of a risk prediction model with improved clinical utility in elective cervical and lumbar spine surgery. Spine. 2020;45(9):E542-51.

27. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996;49(8):907-16.

28. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8.

29. Kim JS, Merrill RK, Arvind V, et al. Examining the ability of artificial neural networks machine learning models to accurately predict complications following posterior lumbar spine fusion. Spine. 2018;43(12):853-60.

30. Forsström JJ, Dalton KJ. Artificial neural networks for decision support in clinical medicine. Ann Med. 1995;27(5):509-17.

31. Lee MJ, Konodi MA, Cizik AM, et al. Risk factors for medical complication after spine surgery: a multivariate analysis of 1,591 patients. Spine J. 2012;12(3):197-206.

32. Mirza SK, Deyo RA, Heagerty PJ, et al. Development of an index to characterize the “invasiveness” of spine surgery: validation by comparison to blood loss and operative time. Spine. 2008;33(24):2651-61.

33. Kasparek MF, Boettner F, Rienmueller A, et al. Predicting medical complications in spine surgery: evaluation of a novel online risk calculator. Eur Spine J. 2018;27(10):2449-56.

34. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Med Res Methodol. 2012;12:21. DOI.org/10.1186/1471-2288-12-21

35. Shahsavari S, Naderi M, Abbasi M. Letter to the Editor concerning “Predicting medical complications in spine surgery: evaluation of a novel online risk calculator” by M. F. Kasparek et al. (Eur Spine J: DOI 10.1007/s00586-018-5707-9). Eur Spine J. 2018;27(11):2885-6.

36. Ratliff JK, Balise R, Veeravagu A, et al. Predicting occurrence of spine surgery complications using “big data” modeling of an administrative claims database. J Bone Joint Surg Am. 2016;98(10):824-34.

37. Fisher ES, Whaley FS, Krushat WM, et al. The accuracy of Medicare’s hospital claims data: progress has been made, but problems remain. Am J Public Health. 1992;82(2):243-8.

38. Guimarães PO, Krishnamoorthy A, Kaltenbach LA, et al. Accuracy of medical claims for identifying cardiovascular and bleeding events after myocardial infarction: a secondary analysis of the TRANSLATE-ACS study. JAMA Cardiol. 2017;2(7):750-7.

39. Rudrapatna VA, Glicksberg BS, Avila P, et al. Accuracy of medical billing data against the electronic health record in the measurement of colorectal cancer screening rates. BMJ Open Qual. 2020;9(1):e000856.

40. Veeravagu A, Li A, Swinney C, et al. Predicting complication risk in spine surgery: a prospective analysis of a novel risk assessment tool. J Neurosurg Spine. 2017;27(1):81-91.

41. Han SS, Azad TD, Suarez PA, et al. A machine learning approach for predictive models of adverse events following spine surgery. Spine J. 2019;19(11):1772-81.

42. Siontis GC, Tzoulaki I, Castaldi PJ, et al. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015;68(1):25-34.

43. McGirt MJ, Sivaganesan A, Asher AL, et al. Prediction model for outcome after low-back surgery: individualized likelihood of complication, hospital readmission, return to work, and 12-month improvement in functional disability. Neurosurg Focus. 2015;39(6):E13.

44. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014;14:137.

45. Buchlak QD, Yanamadala V, Leveque JC, et al. The Seattle spine score: predicting 30-day complication risk in adult spinal deformity surgery. J Clin Neurosci. 2017;43:247-55.

46. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40.

47. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18.

48. Vickers AJ, Holland F. Decision curve analysis to evaluate the clinical benefit of prediction models. Spine J. 2021;21(10):1643-8.

49. Clavien PA, Vetter D, Staiger RD, et al. The Comprehensive Complication Index (CCI®): added value and clinical perspectives 3 years “down the line”. Ann Surg. 2017;265(6):1045-50.

50. Schimmel JJ, Horsting PP, de Kleuver M, et al. Risk factors for deep surgical site infections after spinal fusion. Eur Spine J. 2010;19(10):1711-9.

51. Donnally CJ, 3rd, Vakharia RM, Rush AJ, 3rd, et al. Fibromyalgia as a predictor of increased postoperative complications, readmission rates, and hospital costs in patients undergoing posterior lumbar spine fusion. Spine. 2019;44(4):E233-8.