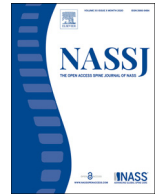Contents lists available at ScienceDirect

North American Spine Society Journal (NASSJ)

journal homepage: www.elsevier.com/locate/xnsj

Systematic Reviews /Meta-analyses

A scoping review of complication prediction models in spinal surgery: An analysis of model development, validation and impact ☆

Toros C. Canturk [a], Daniel Czikk [a], Eugene K. Wai [b], Philippe Phan [b], Alexandra Stratton [b], Wojtek Michalowski [c], Stephen Kingwell [b],*

[a] Faculty of Medicine, University of Ottawa, 451 Smyth Rd, Ottawa, ON K1H 8M5, Canada
[b] Division of Orthopaedic Surgery, The Ottawa Hospital, 1053 Carling Ave, Ottawa, ON K1Y 4E9, Canada
[c] Telfer School of Management, University of Ottawa, 55 Laurier Ave. E, Ottawa, ON K1N 6N5, Canada

ARTICLE INFO

ABSTRACT

Background: Predictive analytics are being used increasingly in the field of spinal surgery with the development of models to predict post-surgical complications. Predictive models should be valid, generalizable, and clinically useful. The purpose of this review was to identify existing post-surgical complication prediction models for spinal surgery and to determine if these models are being adequately investigated with internal/external validation, model updating and model impact studies.

Methods: This was a scoping review of studies pertaining to models for the prediction of post-surgical complication after spinal surgery published over 10 years (2010-2020). Qualitative data was extracted from the studies to include study classification, adherence to Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines and risk of bias (ROB) assessment using the Prediction model study Risk Of Bias Assessment Tool (PROBAST). Model evaluation was determined using area under the curve (AUC) when available. The Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement was used as a basis for the search methodology in four different databases.

Results: Thirty studies were included in the scoping review and 80% (24/30) included model development with or without internal validation. Twenty percent (6/30) were exclusively external validation studies and only one study included an impact analysis in addition to model development and internal validation. Two studies referenced the TRIPOD guidelines and there was a high ROB in 100% of the studies using the PROBAST tool.

Conclusions: The majority of post-surgical complication prediction models in spinal surgery have not undergone standardized model development and internal validation or adequate external validation and impact evaluation. As such there is uncertainty as to their validity, generalizability, and clinical utility. Future efforts should be made to use existing tools to ensure standardization in development and rigorous evaluation of prediction models in spinal surgery.

Introduction

Predictive analytics is gaining popularity in the field of spine surgery in line with patient safety and quality improvement initiatives, and this is certainly evident in surgical decision-making [1,2]. Given that electronic medical records (EMRs) are heavily invested in and can store data on surgical outcomes and complications, there has been a subsequent surge in prediction model development which can help inform surgical decision-making [3,4]. EMRs and large multicentre databases are often used for model development but the data may not be standardized, may have missing values and could be at risk of bias [5,6].

Prediction model development requires that model features have clinical relevance, are feasible to measure, and impact the outcomes of interest [7,8]. Logistic regression has traditionally been used for model development but there is significant interest in applying machine learning algorithms to model development. Regardless of the algorithm selected, the model must have a balance of accuracy, transparency, and generalizability [7–9]. The model must then be rigorously tested and

validated in different settings before it can be used on a larger scale [8–10].

Moons et al. described a framework for prediction model development and evaluation, which includes model development, internal validation, external validation, model updating, and model impact studies [8,10]. After development, the model should next undergo internal validation, which is the assessment of whether the predictive results accurately represent the studied population [8]. To ensure that the model is generalizable to distinct populations from those that developed the model, external validation must then occur [10]. The model should then be updated and evaluated with model impact studies to determine whether the model is being utilized correctly in clinical settings and is having its intended impact [10].

The current standard is that research groups presenting new models should be following the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines, which outline the necessary steps for reporting studies about the development, validation, and updating of prediction models [11]. With the increase in prediction model publications there has been an expected interest in analyzing and summarizing the quality and applicability of these models. The Prediction model study Risk Of Bias Assessment Tool (PROBAST) aims to assess the risk of bias and applicability of studies that develop, validate or update multivariable prognostic prediction models [12].

The objective of this review was to summarize existing complication prediction models in spinal surgery and to determine if the models are being adequately studied with external validation, model updating, and model impact studies. Prediction models developed for use in the pre-operative phase, and at the time of surgical or shared decision-making, for the prediction of post-operative complications in spinal surgery were the focus of this review. Secondary objectives of the review were to perform a risk of bias and applicability assessment using the PROBAST tool, to determine if the published research on prediction models in spinal surgery are adhering to the TRIPOD guidelines and to evaluate whether logistic regression or other machine learning algorithms are becoming the predominant approach to prediction model development.

Methods

Search strategy

The search methodology was completed in agreement with the Reporting Items of the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement [11]. Our search identified studies on prediction models, calculators, and algorithms for spinal surgery complications. The search strategy used Medical Subject Headings (MeSH) terms determined with the assistance of a medical librarian for each database to ensure consistency. The search was a convenience sample limited to English studies published between 2010 and the search date (July 2020) in Web of Science, Embase (Ovid), MEDLINE (Ovid), and Scopus.

Inclusion and exclusion criteria

The inclusion criteria were: (1) studies with expressed purpose of developing, validating, updating or determining impact of models, calculators or algorithms used for predicting postoperative complications after spinal surgery; (2) model features including preoperative variables or planned intraoperative variables (i.e., instrumentation plan, use of bone morphogenic protein); (3) studies with patients over the age of 18. The exclusion criteria were: (1) models that used intraoperative features that would not be known or anticipated preoperatively (i.e., blood loss, medications); (2) models that were not developed from any spinal surgery patients (i.e., frailty index); (3) studies with exclusively traumatic, infec-

tious or oncologic spinal diagnoses; (4) conference proceedings, poster presentations, letters to the editor, or abstracts; (5) non-English studies.

Data extraction and risk of bias and quality assessment

All identified studies through the database searches were uploaded to Covidence (2021 version), an online tool used for organization and data management of systematic and scoping reviews. Each study was independently reviewed by two of the authors (TC and DC), with conflicts resolved by the senior author (SK).

As per Moons et al., the extracted studies were grouped into tables based on whether they were model development and/or internal validation studies (Tables 1 and 2) or external validation, model updating and/or model impact studies (Table 3) [6–8,10]. Table 1 includes all studies that utilized logistic regression (LR) for model development and internal validation. Table 2 includes studies that used other machine learning algorithms for model development and internal validation.

A risk of bias and quality assessment was completed for each article using the Prediction study Risk of Bias Assessment Tool (PROBAST) [12]. PROBAST divides the risk of bias (ROB) into four domains: (1) Participants (i.e., study design); (2) Predictors (i.e., standardized and transparent predictors); (3) Outcomes (i.e., defined and measured the outcomes within a timely manner); and (4) Analysis (i.e., proper model development, whether lost data was handled appropriately). A set of questions are answered for each domain with either a 'yes', 'no', or unclear response, allowing a low, high, or unclear ROB to be assigned to each domain within an article. Risk of bias analysis is important as it provides an alternative metric to evaluate the article's limitations in study design and validation process. Additionally, ROB analysis allows readers to be able to draw stronger conclusions from a study's results, if it has low ROB [13].The tool also has a section for applicability where it is determined whether the prediction model applies to the research question of the scoping review [12]. The ROB assessment should be consistent amongst reviewers; however, the applicability can vary as it is dependent on the primary systematic review question.

Results

Our search identified 1910 studies that were inserted in COVIDENCE for duplication removal and subsequent review steps. Details of the review process including abstract removal for duplication, relevance and full text-review of eligibility are found in Fig. 1. The top three reasons for exclusion at this step were wrong outcomes (84.4%), wrong setting (4.8%) and wrong study design (3.8%). After this step, we manually added one more study that met our criteria after a hand search of references.

There were 30 studies that met the inclusion and exclusion criteria. Three studies (10%) were published between 2010-2015 and 90% were published between 2016-2020. There were 24 studies that included model development with or without internal validation and this represented the majority of the studies (80%) that were identified. All 24 of these studies describe distinct prediction model; however, some models were developed using the same databases and/or had minor variations in feature selection or predicted outcomes. Seventy-nine percent (19/24) of these studies only included internal validation with model development, 8% (2/24) only focused on model development and 13% (3/24) included either external validation or impact study in addition to internal validation and model development. Twenty of the 24 studies (83%) used logistic regression for model development while 17% (4/24) used other machine learning methods.

Table 1 summarizes the 20 model development and internal validation studies developed using LR. Table 2 includes the four prediction models that utilized ML algorithms during the model development. Table 3 summarizes the nine studies that included external validation, model updating or model impact analyses. Six studies were exclusively

Table 1
Studies that used logistic regression for model development and/or internal validation.

Authors	Year	Calculator name	Study design	Number of cases	Registry	Diagnostic/ procedural classification	Complications measured	Study classification	Model evaluation	TRIPOD
McGirt et al. [51]	2015	None	Prospective	1,803	Single centre	Lumbar spine surgery	Overall complications	MD and IV	AUC: Development: 0.72 Validity: 0.82	no
Yilgor et al. [52]	2017	Global Alignment and Proportion (GAP) Score	Retrospective	222	Multicentre	Posterior fusion	Mechanical complications (proximal junctional kyphosis or failure, distal junctional kyphosis or failure, rod breakage, and implant-related complications)	MD and IV	AUC: Development: 0.88 Training: 0.92	no
Passias et al. [53]	2019	None	Retrospective	123	Multicentre	Cervical deformity surgery	Medical or surgical complications	MD and IV	AUC complications: Overall:0.79 Medical: 0.77 Surgical: 0.74	no
Bekelis et al. [54]	2014	None	Retrospective	13,660	NSQIP	Spinal surgery	Stroke, MI, death, infection, UTI, DVT, PE, return to OR, SSI	MD and IV	AUC: 0.65-0.95 (reported individually for each complication)	no
Lee et al. [55]	2014	SpineSage	Retrospective	1,476	Multicentre	Spinal surgery	Cardiac, pulmonary, GI, neurologic, hematologic, urologic; any medical complication and major medical complication	MD and IV	AUC: Any medical complication: 0.76 Major medical complications: 0.81	no
Klemencsics et al. [56]	2016	None	Prospective	1,030	Single centre	Lumbar de-compression, microdiscec-tomy, or instrumented fusion	SSI	MD and IV	AUROC: Development: 0.71 Validation :0.72	no
Belykh et al. [57]	2017	None	Retrospective	350	Single centre	Microdiscectomy	Recurrence of lumbar disc herniation	MD	0.60-0.99 (Accuracy of prediction rate for different models)	no
Han et al. [58]	2019	None	Retrospective	345,510 (MKS & MSM) and 760,724 (CMS)	MKS, MSM, CMS	Spinal surgery	Overall adverse event (AE) occurrence and types of AE occurrence during the 30-day postoperative follow-up.	MD and IV	AUC: Overall: 0.70 Specific:0.76	no
Fatima et al. [59]	2020	None	Retrospective	80,610	NSQIP	Lumbar degenerative spondylolis-thesis	Overall adverse events	MD and IV	AUC: 0.7	yes
Ratliff et al. [16]	2016	Spinal Risk Assessment Tool	Retrospective	279,135	Truven Health Analytics MarketScan Commercial Claims and Encounters and Medicare Supplement and Coordination of Benefits Databases	Spinal surgery	Major complications	MD and IV	AUC: 0.7	no

Table 1 (*continued*)

Authors	Year	Calculator name	Study design	Number of cases	Registry	Diagnostic/procedural classification	Complications measured	Study classification	Model evaluation	TRIPOD
Kim et al. [60]	2017	None	Cross-sectional database	22,629	ACS-NSQIP	Posterior lumbar spine fusion	Cardiac, wound complications, VTE, mortality	MD and IV	AUC: 0.59-0.70	no
Janssen et al. [61]	2019	None	Retrospective	898	Single centre	Instrumented thoracolumbar spine	SSI	MD and IV	AUC: 0.72	no
Kim et al. [62]	2018	None	Cross-sectional Database Study	4,073	ACS-NSQIP	ASD	Cardiac or wound complications, venous thromboembolism, mortality	MD and IV	AUC: 0.55-0.79	no
Li et al. [63]	2020	None	Retrospective	124	Single centre	ASD	Medical and surgical complications	MD	AUC: 0.82	no
Passias et al. [53]	2019	None	Prospective	117	Multicentre	Cervical deformity surgery	Distal junctional kyphosis	MD and IV	AUC: 0.87	no
Yagi et al. [64]	2020	PRISM	Retrospective	321	Multicentre	ASD	Mechanical failure	MD and IV	AUC: 0.81 (mechanical failure and risk grade correlation) AUROC: 0.96 (predictive model accuracy 92%)	no
Yagi et al. [28]	2019	None	Retrospective	151	Multicentre	ASD	Neurologic, implant related, SSI, other infection, cardiopulmonary, gastrointestinal	MD, IV and EV	AUC: Development: 0.82 Validation: 0.75	no
Yagi et al. [26]	2018	None	Retrospective	195	Multicentre	ASD	Major complications (all post-op complications recorded)	MD, IV and EV	AUROC: 0.96 (92% predictive model accuracy, 84% external validation accuracy)	no
Buchlak et al. [27]	2017	The Seattle spine score	Retrospective	136	Single centre	ASD	Cardiopulmonary, wound, infection, thrombotic, unplanned return to surgery, death (30 days)	MD, IV and Impact Study	AUROC: 0.71	no

MD: model development; IV: internal validation; EV: external validation, LR: logistic regression, ML: machine learning, AUC: area under the curve, AUROC: area under the receiver operating characteristic, MI: myocardial infarction, UTI: urinary tract infection, DVT: deep vein thrombosis, PE: pulmonary embolism, SSI: surgical site infection, GI: gastrointestinal, ASD: adult spinal deformity, NSQIP: National Surgical Quality Improvement Program®, ACS-NSQIP: American College of Surgeons National Surgical Quality Improvement Program®, MKS: Truven MarketScan Database, MSM: MarketScan Medicaid Database, CMS: Centers for Medicare and Medicaid Services Database.

Table 2
Studies that used machine learning for model development and/or internal validation.

| Authors	Year	Calculator name	Study design	# Of cases	Registry	Diagnostic/ procedural classification	Complications measured	Moon's classification	Model evaluation	TRIPOD
Scheer et al. [65]	2018	None	Retrospective	336	Multicentre	ASD	Pseudarthrosis at 2 years postoperatively	MD and IV	AUC Development: 0.97 Training: 0.94	No
Scheer et al. [66]	2017	None	Retrospective	557	Multicentre	ASD	Minor or major intraoperative and postoperative complications	MD and IV	AUROC: 0.89	No
Hopkins et al. [67]	2020	None	Retrospective	4046	Single centre	Posterior spine fusion	Surgical site infection	MD and IV	AUC: 0.78	No
Clark et al. [68]	2020	NZRISK-Neuro	Retrospective	18,375	New Zealand registry	Neuro or spinal surgery	Mortality	MD and IV	AUC: 0.9 (30-day), 0.91 (1 year) 0.91 (2 year)	Yes

MD: model development; IV: internal validation; EV: external validation, LR: logistic regression, ML: machine learning, AUC: area under the curve, AUROC: area under the receiver operating characteristic ASD: adult spinal deformity, NZRISK-Neuro: The New Zealand Neurosurgical Risk Tool.

Scoping review screening and extraction
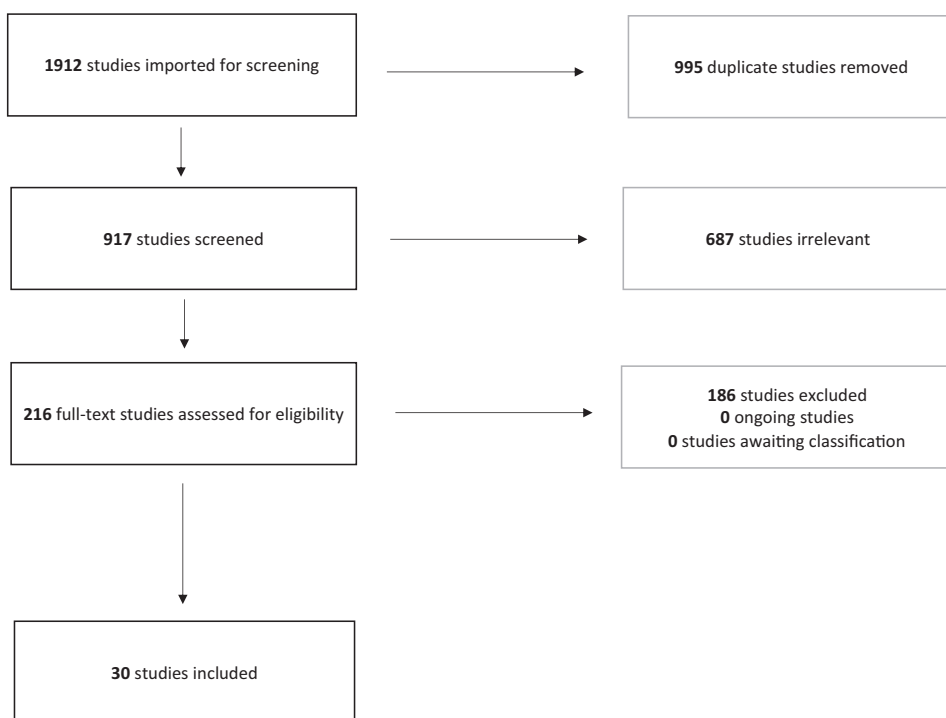
Fig. 1. Scoping review screening and extraction.

external validation studies and two of these studies evaluated the Spine sage tool [14] and ACS-NSQIP calculator [15] respectively and one study evaluated the spinal risk assessment tool [16]. While all nine studies included external validation, there was only one study that also conducted an impact study and another single study which conducted a model update in addition to the external validation.

PROBAST Results

The PROBAST assessment indicated high ROB for all 30 studies that we assessed. Seventy percent of studies (21/30) had low ROB for the participants domain as they met all the PROBAST criteria. PROBAST requires that appropriate data sources be used with appropriate inclusion/exclusion criteria, and this was generally the case.

The predictor, outcomes, and analysis components of the PROBAST assessment showed the highest ROB. For the predictor domain 57% (17/30) of the studies had high ROB. Due to the retrospective nature of many studies the predictor definitions and assessments were not clearly distinct from the outcome assessment, thus introducing bias. There was high ROB in 83% (25/30) of the studies for the outcome domain and 87% (26/30) for the analysis domain. As per the PROBAST assessment, high ROB was introduced by low numbers of participants with the outcome, mishandling of continuous and categorical predictors or missing data, predictor selection primarily based on univariate analysis and a lack of accounting of model overfitting or optimism. Finally, the PROBAST guidelines recommend downgrading low ROB studies to high ROB if they have a small data set and/or the model was developed without external validation [12].

6

Table 3
External validation, impact and model update studies that used either logistic regression or machine learning.

Authors	Year	Calculator name	Study design	# Of cases	Registry	Diagnostic/ procedural classification	Complications measured	Moon's classification	Model evaluation	TRIPOD
Yagi et al. [28]	2019	None	Retrospective	151	Multicentre	ASD	Neurologic, implant related, SSI, other infection, cardiopulmonary, gastrointestinal	MD, IV and EV	AUC: Training: 0.82 Validation: 0.75	no
Sebastian et al. [29]	2019	ACS-NSQIP Surgical Risk Calculator	Retrospective	2808	ACS-NSQIP	Single-level posterior lumbar fusion	NSQIP 30-day complications	EV and performance assessment	C-statistic: 0.56-0.66	no
Yagi et al. [64]	2018	None	Retrospective	145	Multicentre	ASD	Proximal junctional failure	EV and model update	AUC: Training: 0.98 Testing: 1.00	no
Janssen et al. [31]	2018	Spine Sage	Retrospective	898	Single centre	Instrumented thoracolumbar spine cases	SSI	EV	AUC: 0.61	N/A
Kasparek et al. [32]	2018	Spine Sage	Retrospective	273	Single centre	Spinal surgery	Overall medical complications and major medical complications	EV	AUC: Overall complication: 0.71 Major complications: 0.85	N/A
Wang et al. [30]	2017	ACS-NSQIP Surgical Risk Calculator	Retrospective	242	Single centre	Lumbar laminectomy without fusion	Post-operative complications as per NSQIP	EV	AUC: All observed complications: 0.44 Predicted complications: 0.14 -See the paper for AUC of specific complications.	N/A
Veeravagu et al. [69]	2017	Spinal Risk Assessment Tool (RAT) and ACS NSQIP surgical risk calculator	2 cohorts: Retrospective and Prospective	200 (retrospective) 246 (prospective)	Single centre	Spine surgery	Cardiac, wound, thrombotic, pulmonary, urinary tract infection, radiculopathy, dysphagia, delirium, other	EV	AUC: Complications: 0.67 (for both SpinalRAT and ACS-NSQIP)	N/A
Yagi et al. [26]	2018	none	Retrospective	195	Multicentre	ASD	Major complications	MD, IV and EV	AUROC: 0.96	no
Buchlak et al. [27]	2017	The Seattle spine score	Retrospective	136	Single centre	ASD	30-day cardiopulmonary, wound, infection, thrombotic, unplanned return to surgery, and death	MD, IV and Impact Study	AUROC: 0.71	no

MD: model development; IV: internal validation; EV: external validation, LR: logistic regression, ML: machine learning, AUC: area under the curve, AUROC: area under the receiver operating characteristic, SSI: surgical site infection, ASD: adult spinal deformity, NSQIP: National Surgical Quality Improvement Program®, ACS-NSQIP: American College of Surgeons National Surgical Quality Improvement Program®.

Discussion

With the overarching goals of improving health care quality and safety, research using large data sets and predictive analytics continues to expand and evolve [2,3,6,17–24]. In particular, the field of spinal surgery has experienced an exponential increase in the development of prediction models [25]. Although more prediction models are being developed than ever before, this review confirms that only a small fraction of them undergo external validation, model updating and/or impact studies. This raises numerous concerns about their clinical utility, validity, and reliability in different populations.

As highlighted by Moons et al., it is vital for prediction models to undergo a multifactorial design process, which includes internal validation, external validation, and impact studies in order for them to be optimally tested for their respective niche [6–8,10]. This systematic review of complication prediction models for spinal surgery identified 30 articles and only 8 articles assessed specifically, or included, external validation. Moreover, there was only one model update [26] and one impact study [27]. Only two model development studies included external validation within their first publication based on the Moons criteria [26,28]. Additionally, only one study included an impact study with their model development and internal validation [27].

Given that most predictive models for spinal surgery complications don't undergo external validation, the large majority of models that are developed and internally validated are never tested on other patient populations, potentially limiting generalizability and clinical value of the model. External validation is important because different populations have unique, site-specific, unmeasurable features that may influence both patient complications and outcomes. This may lead to significant differences in model performance in various settings.

The NSQIP surgical risk calculator is a readily available online tool [29], but it was not an accurate and reliable predictor of post-surgical complications in an elderly Chinese population [30]. Similarly, the SpineSage tool [14] demonstrated poor predictive performance in a European population [31]. Conversely, Kasparek et al.'s external validation study [32] demonstrated similar area under the curve (AUC) values as the internal validation [14], both for medical and major medical complications (0.85 and 0.71 respectively).

External validation studies are critical to evaluate model performance in new settings and to determine whether re-calibration or other measures are warranted for local optimal performance [33,34]. Given the paucity of external validation studies and the variable model performance observed, this should give pause when applying the tools unreservedly and provide the impetus for further external validation research. Alternatively, other authors have advocated for site-specific prediction models, based on local data and features from EMRs, thus obviating the need for external validation [31].

It should be noted that 62% (13 of 20) of model development and internal validation studies included in our review were published between 2018-2020. This may limit the time for undertaking any external validation studies, however, it confirms the recent increase in spine surgery complication prediction model development.

The identification of only one impact study in this review is notable as impact studies have a critical role in real world model evaluation. In the context of a complication prediction model for surgery, the model may impact surgeon or patient decision making, treatment outcomes, patient satisfaction with the decision-making process, clinic logistics or patient flow and interfacing with electronic health records to name a few [35–38]. While adequate calibration, discrimination and classification of a prediction model may appear to indicate a successful clinical performance, multiple studies have shown that the real-life performance may not have any meaningful impacts in clinical decision making [39–41]. Furthermore, impact studies allow physicians to contextualize the true utility and effect of the developed model on patient care [42].

Kappen et al. highlight several potential reasons for why impact studies are difficult to undertake [42]. The ideal way to measure a model's clinical impact on real-life decision making, change in practice, and health outcomes is via large scale cluster-randomized studies. However, these studies have substantial cost and effort and may be impractical given growth of prediction models [10,43–45]. While significant efforts should be given to the model development, equal importance should be given to all levels of model validation, impact and updating in order for these models to be clinically valuable.

The TRIPOD guidelines provide a comprehensive checklist for prediction model development and they are separated into 'Model Development' and 'Model Validation' [46]. There are numerous benefits in following TRIPOD guidelines including standardization of the model development process, transparency, and ensuring high quality research. Transparency provides the foundation for external validation studies, can help to avoid duplication, and can potentially improve clinical uptake. Even though the TRIPOD guidelines were initially published in 2015, the vast majority of the models that were developed after did not follow the TRIPOD-recommended steps. Only two out of 21 studies in this review pertaining to model development and published after 2015 referenced the use of TRIPOD guidelines during their development.

PROBAST is a ROB assessment tool specifically developed for prediction models which differentiates it from other ROB approaches [12,47,48]. Flaws in study design, conduction, and/or analysis in the prediction model development should be identified in order to have a clear understanding and appreciation of the model's predictive performance [12]. Our PROBAST assessment indicated high ROB for 100% of the studies. Specifically, there was high ROB in 30%, 57%, 83% and 87% of the studies for the participant, predictor, outcomes, and analysis domains. All studies had a high ROB in at least one domain of outcomes or analysis and the PROBAST tool recommends downgrading model development studies from low to high ROB if external validation is not included as part of the study.

Similar observations on ROB were reported by other studies including, White et al.'s PROBAST analysis in their systematic review, which found high ROB in all 13 of the studies that they analyzed [49]. Additionally, Vemena et al.'s study on 556 clinical prediction models (CPMs) found 529 (95%) of the CPMs had high, 20 (4%) had low,7(1%) had unclear ROB classification [50].

The primary objective of this review was to evaluate whether prediction models for postoperative spine surgery complications are being adequately studied with external validation, model updating, and model impact studies. Other published systematic reviews pertaining to postoperative complications in orthopedic surgery[25] and complications or outcomes in spinal surgery [49] have been descriptive and comprehensive but have not analyzed the existing research in this manner. This review has confirmed the increase in published complication prediction models in spinal surgery, but the studies are heavily weighted towards model development and internal validation. The lack of external validation and model impact studies represent significant potential limitations in their generalizability and clinical utility.

There are limitations to this scoping review. The goal of the review was to identify all published complication prediction models in spinal surgery that a spine surgeon might want to apply to their own practice and to evaluate the existing research as it pertains to model development, validation, updating and impact. It became apparent that many studies reported on risk factors or multivariate predictors of various outcomes and these studies were excluded unless the expressed purpose of the study was to develop a prediction model. As such, it is possible some relevant models were erroneously excluded during the process, but it is unlikely that these models were at a stage where they were internally or externally validated or operational in a clinical practice. Similarly, it was felt that prediction models reported in abstracts, conference proceedings, letters or otherwise would not be able to undergo the evaluation utilized in this scoping review. There are some important limitations to the PROBAST assessment. A comprehensive external assessment of ROB is aided with tools such as PROBAST, but the assessment is generally rendered more difficult by a lack of explicit information rather than overt

signs of bias. It is certainly important for all researchers publishing on prediction models to be aware of TRIPOD guidelines and the PROBAST assessment to ensure consistency and transparency in the reporting of prediction models.

Conclusions

The objective of this scoping review was to determine whether existing complication prediction models in spinal surgery have undergone adequate internal and external validation, model updating and evaluated with model impact studies. The majority of studies identified in this review pertained exclusively to model development and internal validation. A small number of external validation studies were conducted and only one impact study was identified. All studies had high ROB as determined using the PROBAST tool and very few studies referenced the TRIPOD guidelines. While complication prediction models in spinal surgery may be useful adjuncts to surgical decision-making questions will remain as to their validity, generalizability, and clinical utility unless the models undergo appropriate validation, updating and impact analysis. Users of prediction models in spinal surgery should be aware of their current, inherent limitations. Going forward, researchers with an interest in predictive analytics and the development of prediction models for spinal surgery should familiarize themselves with the TRIPOD guidelines and PROBAST tool in the earliest stages of model planning and focus additional research on the impact of prediction models.

Funding Disclosure

Competing Interests

The authors have no competing interests to disclose.

Acknowledgements

The authors have no individuals to acknowledge.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.xnsj.2022.100142.

References

[1] Elwyn G, et al. Implementing shared decision making in the NHS. BMJ 2010;341:971–2.

[2] Chang M, Canseco JA, Nicholson KJ, Patel N, Vaccaro AR. The Role of Machine Learning in Spine Surgery: The Future Is Now. Front. Surg. 2020;0:54.

[3] AK J, et al. Use of electronic health records in U.S. hospitals. N. Engl. J. Med. 2009;360:1628–38.

[4] Mahmoudi E, et al. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. BMJ 2020;369.

[5] Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J. Am. Med. Informatics Assoc. 2017;24:198–208.

[6] Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:1317–20.

[7] Lee Y, Bang H, Kim DJ. How to Establish Clinical Prediction Models. Endocrinol. Metab. 2016;31:38–44.

[8] Moons KGM, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker 2012;98:683–90.

[9] Abbott D. Applied Predictive Analytics : Principles and Techniques for the Professional Data Analyst 2014;453.

[10] KG M, et al. Risk prediction models: II. External validation, model updating, and impact assessment 2012;98:691–8.

[11] Liberati A, et al. The PRISMA statement for reporting systematic reviews and meta–analyses of studies that evaluate health care interventions: explanation and elaboration. J. Clin. Epidemiol. 2009;62:e1–e34.

[12] Wolff RF, et al. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann. Intern. Med. 2019;170:51–8.

[13] Higgins JPT, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011;343.

[14] Lee MJ, Cizik AM, Hamilton D, Chapman JR. Predicting medical complications after spine surgery: a validated model using a prospective surgical registry. Spine J 2014;14:291–9.

[15] Bilimoria KY, 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;217 833-842.e3.

[16] Ratliff JK, et al. Predicting occurrence of spine surgery complications using big data modeling of an administrative claims database. J. Bone Jt. Surg. - Am. 2016;98:824–34.

[17] Bernstein DN, et al. Patient-reported outcomes use during orthopaedic surgery clinic visits improves the patient experience. Musculoskeletal Care 2019;17:120–5.

[18] Hwang EJ, et al. Development and Validation of a Deep Learning–Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. JAMA Netw. Open 2019;2 e191095–e191095.

[19] Wang X, et al. ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases 2017:2097–106.

[20] Li Z, et al. Thoracic Disease Identification and Localization With Limited Supervision 2018:8290–9.

[21] Singh R, et al. Deep learning in chest radiography: Detection of findings and presence of change. PLoS One 2018;13:e0204155.

[22] Nam, J. G. et al. Development and Validation of Deep Learning–based Automatic Detection Algorithm for Malignant Pulmonary Nodules on Chest Radiographs. https://doi.org/10.1148/radiol.2018180237 290, 218–228 (2018).

[23] Geras, K. J. et al. High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks. (2017).

[24] Wu N, et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Trans. Med. Imaging 2020;39:1184–94.

[25] Merrill RK, Ibrahim JM, Machi AS, Raphael JS. Analysis and Review of Automated Risk Calculators Used to Predict Postoperative Complications After Orthopedic Surgery. Curr. Rev. Musculoskelet. Med. 2020;13:298.

[26] Yagi M, et al. Predictive model for major complications 2 years after corrective spine surgery for adult spinal deformity. Eur. Spine J. 2018 281 2018;28:180–7.

[27] Buchlak QD, et al. The Seattle spine score: Predicting 30-day complication risk in adult spinal deformity surgery. J. Clin. Neurosci. 2017;43:247–55.

[28] Yagi M, et al. Surgical risk stratification based on preoperative risk factors in adult spinal deformity. Spine J 2019;19:816–26.

[29] Sebastian A, et al. Assessing the Performance of National Surgical Quality Improvement Program Surgical Risk Calculator in Elective Spine Surgery: Insights from Patients Undergoing Single-Level Posterior Lumbar Fusion. World Neurosurg 2019;126:e323–9.

[30] Wang X, Hu Y, Zhao B, Su Y. Predictive validity of the ACS-NSQIP surgical risk calculator in geriatric patients undergoing lumbar surgery. Medicine (Baltimore) 2017;96.

[31] Janssen, D. M. C. C. et al. External validation of a prediction model for surgical site infection after thoracolumbar spine surgery in a Western European cohort. 13, 1–6 (2018).

[32] Kasparek, M. F. et al. Predicting medical complications in spine surgery: evaluation of a novel online risk calculator. 27, 2449–2456 (2018).

[33] Kappen, T. H. et al. Adaptation of Clinical Prediction Models for Application in Local Settings: http://dx.doi.org/10.1177/0272989×12439755 32, (2012).

[34] Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J. Clin. Epidemiol. 2008;61:76–86.

[35] Schaafsma JD, van der Graaf Y, Rinkel GJE, Buskens E. Decision analysis to complete diagnostic research by closing the gap between test characteristics and cost-effectiveness. J. Clin. Epidemiol. 2009;62:1248–52.

[36] Koffijberg H, van Zaane B, Moons KG. From accuracy to patient outcome and cost-effectiveness evaluations of diagnostic tests and biomarkers: an exemplary modelling study. BMC Med. Res. Methodol. 2013 2013;131(13):1–12.

[37] Vickers, A. J. & Elkin, E. B. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models: http://dx.doi.org/10.1177/0272989×06295361 26, 565–574 (2016).

[38] Hunink MGM, et al. Decision Making in Health and Medicine: Integrating Evidence and Values. Decis. Mak. Heal. Med. 2014. doi:10.1017/CBO9781139506779.

[39] di Ruffano LF, et al. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ 2012;344.

[40] Peters SAE, et al. Multiple imputation of missing repeated outcome measurements did not add to linear mixed-effects models. J. Clin. Epidemiol. 2012;65: 686–695.

[41] Kappen TH, Peelen LM. Prediction models: The right tool for the right problem. Curr. Opin. Anaesthesiol. 2016;29:717–26.

[42] Kappen TH, et al. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagnostic Progn. Res. 2018 2018;21(2):1–11.

[43] Reilly BM, Evans AT. Translating clinical research into clinical practice: Impact of using prediction rules to make decisions. Ann. Intern. Med. 2006;144:201–9.

[44] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:1432–5.

[45] Steyerberg EW, et al. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. PLOS Med 2013;10:e1001381.

[46] KG M, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann. Intern. Med. 2015;162:W1–W73.

[47] Luchini C, Stubbs B, Solmi M, Veronese N. Assessing the quality of studies in meta-analyses: Advantages and limitations of the Newcastle Ottawa Scale 2017;5:80–4. http://www.wjgnet.com/.

[48] Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J. Epidemiol. Community Heal. 1998;52:377–84.

[49] White HJ, et al. Predicting Patient-Centered Outcomes from Spine Surgery Using Risk Assessment Tools: a Systematic Review. *Curr. Rev. Mus*culoskelet. Med. 2020 2020;133(13):247–63.

[50] Venema E, et al. Large-scale validation of the Prediction model Risk Of Bias ASsessment Tool (PROBAST) using a short form: high risk of bias models show poorer discrimination. medRxiv 2021 2021.01.20.21250183. doi:10.1101/2021.01.20.21250183.

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

[52] Yilgor C, et al. Global Alignment and Proportion (GAP) Score: Development and Validation of a New Method of Analyzing Spinopelvic Alignment to Predict Mechanical Complications After Adult Spinal Deformity Surgery. Spine J 2017;17:S155–6.

[53] Passias PG, et al. Predicting the occurrence of complications following corrective cervical deformity surgery: Analysis of a prospective multicenter database using predictive analytics. J. Clin. Neurosci. 2019;59:155–61.

[54] Bekelis K, Desai A, Bakhoum SF, Missios S. A predictive model of complications after spine surgery: The National Surgical Quality Improvement Program (NSQIP) 2005-2010. Spine J 2014;14:1247–55.

[55] Lee MJ, Cizik AM, Hamilton D, Chapman JR. Predicting medical complications after spine surgery: a validated model using a prospective surgical registry. Spine J 2014;14:291–9.

[56] Klemencsics I, et al. Risk factors for surgical site infection in elective routine degenerative lumbar surgeries. Spine J 2016;16:1377–83.

[57] Belykh E, et al. Preoperative estimation of disc herniation recurrence after microdiscectomy: predictive value of a multivariate model based on radiographic parameters. Spine J 2017;17:390–400.

[58] Han SS, Azad TD, Suarez PA, Ratliff JK. A machine learning approach for predictive models of adverse events following spine surgery. Spine J 2019;19:1772–81.

[59] Fatima N, et al. Development and Validation of Machine Learning Algorithms for Predicting Adverse Events After Surgery for Lumbar Degenerative Spondylolisthesis. World Neurosurg 2020;140:627–41.

[60] Kim JS, et al. Examining the Ability of Artificial Neural Networks Machine Learning Models to Accurately Predict Complications Following Posterior Lumbar Spine Fusion. Spine (Phila. Pa. 1976) 2018;43:853.

[61] Janssen DMC, van Kuijk SMJ, d'Aumerie B, Willems P. A prediction model of surgical site infection after instrumented thoracolumbar spine surgery in adults. Eur. Spine J. 2019;28:775–82.

[62] Kim JS, et al. Predicting Surgical Complications in Patients Undergoing Elective Adult Spinal Deformity Procedures Using Machine Learning. Spine Deform 2018;6:762–70.

[63] Li XJ, et al. Surgeon-specific risk stratification model for early complications after complex adult spinal deformity surgery. Spine Deform. 2020 2020;81(8):97–104.

[64] Yagi M, et al. The patient demographics, radiographic index and surgical invasiveness for mechanical failure (PRISM) model established for adult spinal deformity surgery. *Sci. Rep*orts 2020 2020;101(10):1–10.

[65] Scheer JK, et al. Development of a validated computer-based preoperative predictive model for pseudarthrosis with 91% accuracy in 336 adult spinal deformity patients. Neurosurg. Focus 2018;45:E11.

[66] Scheer JK, et al. Development of a preoperative predictive model for major complications following adult spinal deformity surgery. J. Neurosurg. Spine 2017;26:736–743.

[67] Hopkins BS, et al. Using artificial intelligence (AI) to predict postoperative surgical site infection: A retrospective cohort of 4046 posterior spinal fusions. Clin. Neurol. Neurosurg. 2020;192:105718.

[68] Clark S, et al. Development and Validation of a Multivariate Prediction Model of Perioperative Mortality in Neurosurgery: The New Zealand Neurosurgical Risk Tool (NZRISK-NEURO). Neurosurgery 2020;87:E313–20.

[69] Veeravagu A, et al. Predicting complication risk in spine surgery: a prospective analysis of a novel risk assessment tool. J. Neurosurg. Spine 2017;27:81–91.