PREDICTING COMPLICATIONS AFTER SPINAL SURGERY: SURGEONS' AIDED AND UNAIDED PREDICTIONS

Stephen Kingwell

Thesis submitted to the University of Ottawa in partial Fulfillment of the requirements for the Degree of Masters of Science in Health Systems

Telfer School of Management
University of Ottawa

© Stephen Kingwell, Ottawa, Canada, 2020

ABSTRACT

Despite the emergence of artificial intelligence (AI) and machine learning (ML) in medicine and the resultant interest in predictive analytics in surgery, there remains a paucity of research on the actual impact of prediction models and their effect on surgeons' risk assessment of post-surgical complications. This research evaluated how spinal surgeons predict post-surgical complications with and without additional information generated by a ML predictive model.

The study was conducted in two stages. In the preliminary stage an ML prediction model for post-surgical complications in spine surgery was developed. In the second stage, a survey instrument was developed, using patient vignettes, to determine how providing ML model support affected surgeons' predictions of post-surgical complications.

Results show that support provided by a ML prediction model improved surgeons' accuracy to correctly predict the presence or absence of a complication in patients undergoing spinal surgery from 49.1% to 54.8% ($p=0.024$).

It is clear that predicting post-surgical complications in patients undergoing spinal surgery is difficult, for models and experienced surgeons, but it is not surprising that additional information provided by the ML model prediction was beneficial overall. This is the first study in the spine surgery literature that has evaluated the impact of a ML prediction model on surgeon prediction accuracy of post-surgical complications.

ACKNOWLEDGEMENTS

I would like to thank the following people for their help and support with this research:

Wojtek Michalowski for his guidance, vision and infinite patience.

Szymon Wilk and Enea Parimbelli for their time and expertise.

Johanna Dobransky for her statistical assistance.

Dita Moravek for her support with editing and formatting.

DEDICATION

Thank you Sunita for your love and support, and to my children Rohan, Shaan and Anika, many years have passed since I embarked on this and I hope that it will serve as a lesson to try and finish what you start in life….no matter how long it may take.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACS	American college of surgeons
AI	Artificial intelligence
ASA	American Society of Anesthesiologists
AUC	Area under the curve
AUROC	Area under the receiver operating characteristics
BMI	Body mass index
CO	Carbon monoxide
CQI	Continuous quality improvement
DM	Diabetes mellitus
DS3	Decision for safer surgery model
HgbA1C	Hemoglobin A1C
ICU	Intensive care unit
ML	Machine learning
NSQIP	National surgical quality improvement program
P	Participant
Spinal RAT	Spinal risk assessment tool
SSI	Surgical site infection
TLIF	Transforaminal lumbar interbody fusion
TOH	The Ottawa Hospital
V	Vignette
VTE	Venous thromboembolism
XAI	Explainable artificial intelligence
XBT	Gradient boosted decision tree

Chapter 1: Introduction

1.1 Predicting Complications in Patients Undergoing Spinal Surgery

Surgical decision-making is a shared process between patients and surgeons and is ultimately made based on a balance of the expected risks and benefits, and influenced by patient values and surgeon experience. For truly informed patient decision-making it is essential that surgeons convey surgical risk that is personalized and accurate. It is expected that surgeons assess the potential risks and benefits of surgery based on their knowledge and experience but these generalizations do not answer many critical questions related to surgical risk prediction. Are surgeons good at personalizing and predicting surgical risk, are there important differences amongst surgeons and could predictive models help improve and standardize surgeons' risk assessments? The emergence of artificial intelligence (AI) and machine learning (ML) in medicine, coupled with increased utilization of large, multicenter databases, has generated interest in predictive analytics in medicine and specifically, developing prediction models to help with predicting post-surgical complications. However, it is not known how surgeons use results derived from these models and whether it improves their prediction accuracy. This research intends to fill this gap by studying how spinal surgeons predict post-surgical complications with and without additional information generated by a ML predictive model.

1.2 Spinal Surgery

Spinal surgery is most commonly performed for degenerative conditions of the spine. These conditions include spinal stenosis, disc herniations and deformities such as spondylolisthesis and scoliosis. In the United States, from 1992-2003, Medicare spending for inpatient back surgery doubled and spending for lumbar fusion increased more than 500% (1). In the United States, the total direct cost of treating low back pain is estimated at $100 billion (2). Direct medical costs of instrumented lumbar fusion in Ontario from 2002 to 2012 totaled $176 million (3). A decision to proceed with spinal surgery is made based on patient values and an understanding of the specific risks and benefits of the proposed procedure. As such, it is critical for patients to have an accurate appreciation of the accepted surgical risks as well as how their overall medical profile

and particular spinal diagnosis may alter these risks. It is essential to accurately track negative outcomes or complications in spinal surgery to truly understand and convey risks to patients.

1.2.1 Post-Surgical Complications

Prospective adverse event or complication collection is a critical part of continuous quality improvement (CQI) initiatives throughout healthcare (4, 5). The rate of adverse events in Canadian hospitals has previously been reported at 7.5% (6). There is significant variation in adverse event risk and subtypes between services in the hospital setting (7). In the context of surgery, adverse events can be defined as any deviation from normal post-operative care (8). Complications, as opposed to adverse events, imply a negative effect on patient's outcome (9). Rigorous collection of data on post-surgical complications can inform on the quality of care and direct change for quality improvement or identify existing strengths (10).

This research focuses on post-surgical complications after spinal surgery. Complication collection varies greatly between clinical reviewers, surgeons and discharge abstraction methods (11, 12). The significant variation between prospective and retrospective chart abstraction methods of complication collection has been demonstrated for an adult spinal surgery population. Street et al. identified 1850 postoperative adverse events in 942 patients undergoing major spinal surgery in comparison to 184 postoperative adverse events in 918 patients using traditional chart abstraction methods (9). This ten-fold increase in reported post-surgical complications in a spinal surgery population highlights the importance of data quality when being used by clinicians to frame a surgical risk discussion with patients. These concerns with data quality and a demand to better understand complications have led to the development of high quality, multinational prospective databases such as the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP) database.

1.2.2 American College of Surgeons National Surgical Quality Improvement Program

The ACS NSQIP database was developed to improve the quality of surgical care in the United States. A critical part of the initiative is the standardization of prospective surgical adverse event or complication collection throughout participating hospitals. The data are risk and case-mix

adjusted to allow for benchmarking and comparisons between centers (13). An increasing number of Canadian hospitals are participating in NSQIP as the importance of CQI and adverse event reporting disseminates. The Ottawa Hospital (TOH) has tracked 100% of spine surgery cases as part of the procedure-targeted program since January 2016 in accordance with NSQIP standards. Mortality and morbidity data are recorded and benchmarked. The database includes information on 30-day post-operative morbidity and mortality from participating hospitals and includes 135 variables. Specific complications that are recorded include cardiac complications, pneumonia, prolonged ventilation, venous thromboembolism, urinary tract infection, surgical site infection, sepsis, return to the operating room and readmissions. However, complications specific to spinal surgery such as dural tear, instrumentation failure and neurological deterioration are not captured within the NSQIP framework.

1.2.3 Predictive Analytics: Post-Surgical Complications

Post-surgical complications in spine surgery are dependent on patient factors as well as perioperative procedural considerations and postoperative events. Conventional outcome prediction, whereby a surgeon will offer their pre-operative risk assessment to the patient, is predicated on sound clinical knowledge, experience and the integration of clinical and pathological data. Considering that surgeons may have variable levels of experience, relevant knowledge and communication skills, there is an obvious role for an unbiased, standardized and data-driven surgical risk assessment that can potentially be achieved by way of predictive analytics.

Predictive analytics is the intersection of statistics and ML that develops models from data to predict future outcomes. Given the opportunity to improve quality and decrease costs there is significant interest in predictive analytics in health care (14). This is particularly relevant in spinal surgery where post-surgical complications can be devastating for patients and costly for the health care system (15).

Predictive models require external validation, model updating and an impact assessment (16). Although some spine specific risk prediction tools have been developed and are available as apps

or web-based calculators little is known about surgeon attitudes, use and understanding with respect to prediction models and how results generated by these models impact surgeons' decision-making or informed consent discussions (17-20). With increasing interest in model development it is critical to understand how the prediction generated by the models impacts surgeons as they synthesize available preoperative information and convey potential risks and anticipated benefits to patients.

Model adoption would certainly be influenced by factors such as surgeon understanding and trust in the model, as well as the perceived usefulness of the model. Although critical to understand a model's comprehensive impact, this research will not investigate how predictive models affect surgical decision-making or a specific model's perceived usefulness. These lines of investigation would be more appropriate for an established complication prediction model prior to workflow integration. This research aims to investigate more focused aspects of the model's impact. Specifically, how information provided by a prediction model impacts surgeons' pre-operative risk assessment and whether the provision of more explanatory information affects surgeons' perception of the model will be investigated.

1.2.4 Use of Machine Learning Models

To date, most of the models used in predicting the risk of complications after surgery were developed using statistical methods. Traditional statistical methods such as logistic regression have inherent limitations with heterogeneous distributions and nonlinear relationships among variables (21). Recently, there is growing interest in using ML models to supplement well-established methods. ML is a branch of AI in which the model "learns" patterns from data and may use these patterns for predicting future outcomes (14). ML models can update their predictions after providing new examples and are capable of making predictions without prior assumptions as required by most of the statistical models (22). There are two basic classes of ML models. In supervised learning, each data item (i.e. patient record) is labelled with the expected outcome (i.e. presence or absence of complication) and the role of the ML model is to learn how to predict this outcome. In unsupervised learning, such labelling is not available and the model is expected to detect if there are some commonalities among groups (clusters) of data

items (23). Supervised learning is the most common type of ML models used in medical research. Machine learning models have been used in the analysis of spinal imaging and spinal outcome research including complications (14).

1.3 Problem Statement and Rationale

A discussion of surgical outcomes and complications in the setting of spinal surgery is critical between patients and surgeons. A model that predicts the risk of post-surgical complications after spinal surgery could be invaluable in order to make an informed decision with respect to the risk/benefit profile of a proposed surgical intervention. As expected, research on prediction model development, including those derived from ML, has greatly expanded in surgical care. The problem, however, is that there remains a paucity of research on the actual impact of prediction models and the models' effect on surgeons' risk assessment for post-surgical complications. In short, more predictive models in surgery are being developed but it is not clear if the models outperform surgeons in prediction accuracy (this baseline is unknown), whether information provided by the model affects surgeon prediction accuracy and how surgeons perceive the predictions derived from these models.

1.4 Research Questions and Objectives

The purpose of this study is to determine how information provided by a ML prediction model affects spinal surgeons' predictions relating to the risk of a post-surgical complication. For this study the risk prediction problem will be considered as a binary classification (risk absent or present). The information provided will be the ML model's prediction of the presence or absence of a post-surgical complication. The prediction support is derived from the application of a ML prediction model that was developed from the TOH NSQIP database. Thus, the primary research question is:

> *How does the use of a ML prediction model impact surgeons' predictions of post-surgical complications in patients undergoing spinal surgery?*

The primary research question will be addressed by the following three research questions:

1) How accurate are spine surgeons at predicting the risk of a post-surgical complication as compared to a ML model's prediction?
2) How does support provided by a ML prediction model affect surgeons' prediction of the risk of a post-surgical complication in patients undergoing spinal surgery?
3) How does providing the explanation for the ML model's prediction affect the surgeon's perception of the model? Is the model prediction explanation presented in an understandable and useful way?

Answering the research questions requires the achievement of the following objectives:

1) The first objective is the development of ML prediction model. There are currently no ML complication prediction models in spinal surgery available for use and therefore one has to be developed for the purpose of this study.
2) The second objective is the development of a survey instrument for assessing surgeons' ability to predict the risk of post-surgical complications and to assess how the ML model affects surgeons' predictions. This will require the development of patient vignettes that summarize real patients that underwent spinal surgery at TOH. The vignettes will be representative of a broad range of clinical scenarios to adequately assess model performance and surgeon prediction accuracy.
3) The third objective requires the development of an explanation for the model's prediction. Typically, ML prediction models will provide output (complication present or absent) without an accompanying rationale or explanation. To investigate the surgeons' perception of the prediction model, the model will be presented with and without an accompanying explanation.

Table 1: Summary of research questions and objectives

Primary research question: How does the use of a ML prediction model impact surgeons' predictions of post-surgical complications in patients undergoing spinal surgery?				
		Objective required to answer research question		
Research questions		Development and internal validation of ML model for surgical risk prediction form TOH-NSQIP database	Development of survey instrument to assess surgeons' predictions of post-operative complications	Development of explanations for ML model's predictions
	How accurate are spine surgeons' at predicting the risk or post-operative complications compared to ML model?	√	√	X
	How does support provided by a ML prediction model affect surgeons' prediction of surgical risk?	√	√	X
	How does providing explanation of ML model support affect surgeons' perception of model?	√	√	√

Chapter 2: Literature Review

A literature review was undertaken to explore the landscape of existing research pertaining to the study, including: (1) complication prediction in spinal surgery; (2) use of ML for the development of prediction models in spinal surgery; (3) surgeons' perception of prediction models and comparison of prediction accuracy; and (4) impact of predictive models in surgery. The latter will include how surgeons use information provided by prediction models. Although complication prediction models using traditional statistical methods were not the focus of this research it was felt that this area should be included in order to assist with development of the TOH ML prediction model, including feature selection, and as a link to research on surgeons' use, perception and attitudes towards complication prediction. The literature review included an analysis of various existing methods of risk prediction in spinal surgery. Considering the primary area of interest, an analysis of how information provided by prediction models impacts surgeons' predictions, the review pertaining to ML models was expanded to include all surgical disciplines to ensure any existing research in this area was captured.

A search of PubMed was performed using the terms "machine learning or data mining" and "spine or spinal" and "surgery". This search yielded 107 references. We then added "complications" to the previous search terms and this narrowed the results to 21. Substituting adverse events for complications did not identify any new references. Revising the search to "surgery", "machine learning or data mining" and "complications" produced 231 results. The "similar article" PubMed function was used to identify related articles and article reference lists were perused for appropriate articles that were not initially identified. In order to search for articles pertaining to impact and usage of prediction models in surgery, the terms "surgery" and "prediction models" were combined with "impact" or "use/usage/utility and perceived usefulness" in PubMed. Ultimately, entering the phrase "how do surgeons use prediction models" resulted in 510 results and by using the similar article function, as well as searching reference lists, the most appropriate comparison and impact articles (topics 3 and 4 above) were identified.

2.1 Complication Prediction in Spinal Surgery

Surgical decision-making is made based on an assessment of the potential benefit and risk to a patient. Identification of risk factors in patients undergoing spinal surgery has long been an area of research interest to inform surgical decision-making. The initiation and growth of multicenter, prospective databases such as ACS NSQIP have led to a corresponding expansion of research on complications in surgery.

Schoefeld et al. reviewed 5887 patients that underwent spinal fusion to investigate the patient factors, comorbidities and surgical characteristics that increase complications and mortality in this patient population (13). Ten percent of patients sustained at least one complication and the mortality rate was 0.4%. Univariate and multivariate analysis identified increasing age, ASA (American Society of Anesthesiologists) classification >2, pulmonary and neurological conditions and a history of wound infection associated with an increased risk of developing complications. Surgical resident involvement, serum albumin level and procedure time were also associated with an increased risk of complications.

Bekelis et al. reviewed 13,660 patients that underwent spine surgery from 2005-2010 from the ACS NSQIP database and used multivariate regression analysis to develop a prediction model for adverse events (24). Specifically, the 30-day incidence of death, pulmonary embolus, deep vein thrombosis, myocardial infection, surgical site infection, urinary tract infection, stroke, pneumonia and re-operation were collected. A risk-factor based prediction model was developed; however, its utility in surgical decision-making is questionable and left unexplored.

Broda et al. used the ACS NSQIP dataset from 2012-2016 and developed the Universal Spine Surgery (USS) score or Cervical and Lumbar Risk Prediction Model (CL-RPM) scores which demonstrated fair predictive accuracy and was comparable to the spinal risk assessment tool (RAT) and ACS NSQIP surgical risk calculator (25).

The generalizability of ACS NSQIP-derived spinal surgery risk factors has been questioned. NSQIP hospitals do represent a small proportion of all hospitals performing spinal surgery and typically these hospitals only sample 20-30% of the overall cases. Furthermore, ACS NSQIP is

not a specialty based database and, therefore, lacks many variables that could significantly influence outcomes and complications in spinal surgery (26).

McGirt et al. developed a prediction model for complications, return to work, hospital readmissions and functional improvement in patients undergoing low back surgery. Over 1800 patients from a single center database were analyzed using linear regression and Bayesian models. The model utilized clinical data and patient interview inputs. For prediction of complications, readmission and return to work the area under the receiver operating characteristic curve (AUROC) was 0.72-0.84. For predicting 12-month post-operative functional outcomes using the Oswestry Disability Index the R^2 was 0.51. The benefit of this model includes its ability to predict outcomes, complications and readmissions. This type of modelling can be used by patients, physicians, payers and hospital systems with decision support tools (2). The usability of the model in clinical practice was not assessed.

Marjoua et al. conducted a systematic review of orthopaedic and spinal research using the NSQIP database (27). Although the primary outcome was to determine the impact of the articles it should be noted that 114 articles were identified and 40 were spine-specific. Seventy-eight percent of the articles were identified as using adjusted statistical techniques for the model development and analysis. No mention was made of the use of data mining or machine learning techniques. Generally, spine-specific publications utilizing the NSQIP database have been limited to the assessment of isolated risk factors and complications associated with specific procedures (28). As such, there is an opportunity to develop a prediction model for complications in spine surgery using ML techniques applied to the ACS NSQIP database (29).

2.1.1 Risk Calculators in Spinal Surgery

ACS-NSQIP Surgical Risk Calculator

The American College of Surgeons National Surgical Quality Improvement Program morbidity and mortality calculator was developed for all surgical disciplines and is entered online. Its utility in spinal surgery has been questioned and it has demonstrated poor predictive performance in single-level posterior lumbar fusion (30).

SpinalRAT

The spinalRAT (Risk Assessment Tool) was developed from an administrative claims database of almost 280,000 patients and was compared with the Charleston Comorbidity Index and the ACS NSQIP surgical risk calculator (31). The surgical risk calculator and comorbidity index have not been specifically evaluated on spine surgical patients. The spinalRAT incorporates patient, procedure and diagnosis elements. The RAT requires 9 preoperative variables whereas the ACS NSQIP calculator has 21 preoperative and intraoperative factors. In the cohort studied by Veeravagu et al., the RAT and ACS NSQIP had comparable moderate efficacy in predicting risk although the ACS NSQIP risk calculator consistently underestimated the risk of complications (20).

SpineSage

Lee et al. developed a tool to predict complications after spinal surgery using a prospective registry (19). The single center database included outcomes and complications for 1476 patients. This model was designed to provide the absolute risk of complications based on the patients' comorbidities as well as the invasiveness of the surgery. The Surgical Invasiveness Index (SII), described by Mirza, was used in the predictive model (32). The model was used to predict the likelihood of any complication and any major complication after spinal surgery. The model was found to be a good measure and a website (SpineSage.com) was created for users of this model.

These risk calculators are available for use as online tools and smartphone apps but were not developed with ML.

2.2 Use of ML for the Development of Complication Prediction Models in Spinal Surgery

The parallel rise in big data and AI as well as continuous quality improvement initiatives in health care has provided a clear opportunity to develop predictive models in surgery.

Durand utilized the ACS NSQIP dataset from 2012-2015 to predict blood transfusion in patients undergoing spinal surgery for adult deformity. Classification tree and random forest ML techniques were utilized and 1029 patients were analyzed. The classification tree model utilized

hematocrit, weight and operative duration. The AUROC for the validation set was 0.79. The random forest model had an AUROC of 0.85 and used operative duration, surgical invasiveness, weight, age and hematocrit (33).

A large multicenter database was used to develop a preoperative predictive model for major intraoperative and postoperative complications in adults with spinal deformity. Forty five variables were used including demographic data, comorbidities, radiographic variables, surgical factors and baseline health measures. An ensemble of decision trees was used and the AUROC was calculated. The overall accuracy was 87% with an AUROC of 0.89 for predicting a major intraoperative or early postoperative complication. The development of a mobile application to assist as a point-of-care decision support tool was identified as an important practical application of the model (34). Similarly, the authors have developed a model using the same database to predict clinically significant proximal junctional kyphosis or proximal junctional failure. An ensemble of decision trees was used and seven preoperative variables were identified as important. These variables include age, lowest instrumented vertebrae, pre-operative sagittal vertical axis, upper instrumented vertebrae and type, preoperative pelvic tilt and pelvic incidence-lumbar lordosis (35).

The ACS-NSQIP database has been used by Kim et al. to evaluate the ability of ML models to predict surgical complications following posterior lumbar spine fusion surgery. In this study 22,629 patients were identified and pre-operative variables were used to develop predictive models using artificial neural networks and logistic regression. The variables included were sex, age, ethnicity, diabetes, smoking, steroid use, coagulopathy, functional status, ASA class, body mass index as well as pulmonary and cardiac comorbidities. The models developed were used to predict cardiac complications, wound complications, venous thromboembolism and mortality. While logistic regression is a common tool used for prediction in surgery, the purpose of the study was to assess predictive performance of the artificial neural networks. ASA was selected as a benchmark as it has previously been shown to predict complications in spine surgery. Both artificial neural networks and logistic regression outperformed ASA in predicting complications in this population and were felt to be comparable (22).

The ACS NSQIP database was used by Arvind et al. to predict complications in patients undergoing anterior cervical discectomy and fusion using ML. They identified 20,789 patients that underwent the procedure and compared ML models to ASA in predicting complications. They found that logistic regression and artificial neural networks outperformed ASA in predicting every complication but support vector machines and random forests were no better than chance at predicting any complications (36).

Han et al. extracted data from over 1 million patients from administrative databases and compared logistic regression and least absolute shrinkage and selection operator regularization method (LASSO) for predicting the risk of the six most commonly observed complications. Logistic regression showed higher AUROC's as compared to LASSO across all the adverse event prediction models. The AUROC for an adverse event was 0.7 and the highest value for any specific complication was 0.76 for pulmonary complication. The authors performed a simulation study and found that model performance increased from 0.67 to 0.7 with the number of patients increasing from 10,000 to 400,000 respectively but no improvement in accuracy with an increasing sample size greater than that (37)

2.3 Surgeons' Perception of Prediction Models and Comparison of Predictive Accuracy

As research on predictive model development has expanded, coinciding with interest in AI and emphasis on quality improvement, it is expected that there would be a similar expansion in research on how predictive models impact clinicians as well as how they are perceived and utilized. Model adoption by surgeons or physicians would be dependent on many factors including but not limited to their understanding of the model and whether the model demonstrated superior accuracy in predicting an outcome of interest.

For surgical risk prediction, little is known about surgeon attitudes and understanding with respect to prediction models and how surgeon prediction compares to model prediction (17, 18). The Decision for Safer Surgery (DS3) model was developed on data for 60,411 patients undergoing general and vascular surgery using logistic regression and had shown good agreement with experienced surgeons with respect to risk estimates (17). The DS3 model was

evaluated by 23 surgeons for its usefulness with a survey employing questions assessed using 5-point Likert scales and pertaining to: information technology, impact of the model on practitioners work, and impact of the model on the clinical care setting. Surgeons rated the model as very or moderately useful in 25% of cases. It was concluded that based on the survey, the DS3 was most useful for achieving particular tasks (patient engagement) or encouraging specific processes (18). The authors acknowledged the lack of validated measures to assess surgeons' perceptions of the model.

Sinuff et al. identified 12 observational studies in a systematic review that compared the ability of physicians and scoring systems to predict mortality in critically ill patients. The prediction tools included scoring systems, computer models and prediction rules. They concluded that ICU physicians discriminated between survivors and non-survivors more accurately than scoring systems and yet both had only moderate accuracy (38).

A ML model for predicting acute kidney injury was compared to ICU physicians' predictions in 250 patients. The ML acute kidney injury predictor had similar discriminative performance to physicians; however, physicians did overestimate the risk of acute kidney injury (39).

Dilaver et al. performed a systematic review and narrative synthesis of surgeons' perception of postoperative outcomes and risk. Their stated aim was to compare surgeons' "gut feeling" or perception of risk with available scoring systems and determine how accurately this correlated with patient outcomes. Twenty-seven studies were included and they found that surgeons consistently over-predicted mortality rates and were outperformed by existing risk scoring tools in six of seven studies with AUROC. They did note that surgeons' prediction of general morbidity was good, and equivalent or better than available risk prediction scores. There were no studies identified relating to spinal surgery (40).

The study by Samim et al. was not included in the previous systematic review and prospectively compared surgeons' risk assessment for major complications in patients undergoing hepato-pancreatico-biliary surgery to nine existing risk prediction models. Surgeons' assessment resulted in an AUROC of 0.71 for liver and 0.56 for pancreas surgery while the AUROC's for

existing models were 0.57-0.73 for liver and 0.51-0.57 for pancreas. They concluded that existing risk prediction models do not outperform surgeons' assessments (41).

2.4 Impact of Predictive Models in Surgery

The clinical implementation of predictive models in surgery requires bridging the research to practice gap. Models are used to assist physicians with their experience or intuition, and evidence based guidelines, to predict outcomes or assess risk (16). Physicians tend to approach clinical problems such as diagnosis and prediction in an intuitive and deductive manner while AI is inductive and analytical. Machine learning prediction models should be considered as a complimentary tool rather than a replacement for clinical reasoning (42).

Brennan et al. used the *MySurgeryRisk* tool and compared clinical judgment with the algorithm in the preoperative assessment of risk. This was a nonrandomized pilot study on 20 physicians to evaluate the usability and accuracy of the tool for preoperative risk assessment. The *MySurgeryRisk* tool is a validated ML model that is imbedded in a real-time, intelligent decision-support platform. They used the AUROC to compare the accuracy of physicians' risk assessment for six postoperative complications before and after using the tool for 150 cases. The *MySurgeryRisk* tool AUROC was 0.73-0.85 compared to physicians 0.47-0.69 except cardiovascular complications. After interaction with the model, physicians improved their risk assessment for acute kidney injury and ICU admission by 12% and 16% respectively. They noted that physicians rated the model as easy to use and useful (43).

There are no other studies identified in the literature review that evaluated the impact of prediction models on surgeons' assessment of risk.

The literature review has identified two spine-specific, accessible complication prediction models. These models are the spinalRAT and SpineSage. There are no studies that have compared the accuracy of these two models to surgeon prediction accuracy, nor are there any studies that have evaluated their usability, impact or surgeons' perception of these models. The review has also confirmed the rapidly growing interest in using ML to predict post-surgical

complications in surgery. There is a lack of research on the external validation of these models particularly in spinal surgery and no published model is available for use in clinical practice.

Importantly, there are no studies that have explored the impact of complication prediction models on spinal surgeons' own risk prediction. Only one study was identified and this pertained to the *MySurgeryRisk* tool for general surgery (43). It is critical to understand how spinal surgeons use information provided by prediction models given that there is a rapid expansion and interest in prediction model development without the corresponding research designed to understand how the models impact and affect the models' users. This research aims to fill this gap by investigating how information provided by a ML complication prediction model affects spinal surgeons' preoperative risk assessment.

Chapter 3: Research Methodology

3.1 Research Design

A surgeon's ability to accurately predict the risk of complications would greatly inform the surgical decision-making process with patients. Misleading or understating complication risk would likely have a negative effect on the decision-making process, patient expectations and ultimately patient satisfaction. Given the uncertainty and potential variability in a surgeon's risk prediction based on clinical and radiological factors, in addition to known complication rates, there is an understandable desire for a data-driven model that can produce predictions of post-surgical complications in the clinical setting. The literature review has confirmed the existence and emergence of ML models for predicting complications in spinal surgery without the concurrent research that would explore the models' impact on surgeons' risk assessment as well as surgeons' perceptions of complication prediction models.

This research was conducted in two stages. In the preliminary stage an ML prediction model for post-surgical complications in spine surgery was developed. In the second stage, a survey instrument was developed and used to answer the research questions pertaining to the study.

3.2 Preliminary Stage: Predicting Risk of Complication with ML Model

There is currently no ML model for patients undergoing spinal surgery that has been externally validated or that is available for this study. As such, the preliminary work for this research required the development of a model for predicting risk of complication for patients undergoing spinal surgery at TOH using TOH NSQIP data. The development of a complication prediction ML model for external use was not an objective of this research.

3.2.1 Study Population

The TOH NSQIP is a single center database of consecutive patients undergoing spinal surgery and these data were used to develop a model for predicting risk of complications up to 30 days

post-discharge including unplanned readmissions and emergency room visits. Unplanned readmissions and emergency room visits were captured and cross-referenced to ensure that all complications were accounted for. Use of this data was approved by the TOH and University of Ottawa Research Ethics Boards.

TOH utilizes the ACS NSQIP surgical quality improvement program to reduce costs, complications and death post-surgery. The hospital historically used the NSQIP essentials process and randomly sampled all surgeries for complications. However, since January 2016 TOH has converted to the procedure targeted NSQIP program which tracks 100% of cases including spinal surgery. The patient data includes demographic (i.e. age, gender), clinical findings (i.e. diagnosis, body mass index, comorbidities), laboratory test results, post-operative complications, and disposition details after discharge.

One thousand five hundred and fifty consecutive patients that underwent spinal surgery at TOH from 2016-2018 were included in the analysis. This includes both elective and urgent surgery, inpatient and outpatient surgery, and surgery for degenerative, traumatic, infectious and oncologic etiologies.

3.2.2 Data Set and Pre-processing

The dataset for the ML model development consists of de-identified data on 1550 patients. Feature selection was determined by the surgeon, computer scientist and health informatics expert. The primary requirements for feature selection were that the features had to be included in the TOH NSQIP database, they would include features that would be available at the time of surgical decision-making and they would be features that surgeons consider relevant to decision-making. As such the features selected were comparable to the NSQIP risk calculator, including demographic variables and comorbidities, but did not include features that would be determined at the time of surgery such as ASA class. Furthermore, our feature selection included diagnostic and procedural features specific to spine that are not considered by the NSQIP risk calculator. The SpineSage tool and to a lesser degree the spinalRAT consider procedural complexity that we cannot determine from our dataset (12, 19, 20). Other important factors such as diagnostic

imaging features were not included in our model as they were not available in the dataset and imaging features have not been considered previously in the NSQIP risk calculator, spinalRAT or SpineSage tool.

Thirty-day complications were separated into surgical and medical complications. These include superficial and deep surgical site infections (SSI) and unplanned return to the operating room related to the surgical procedure because of surgical complications. Medical complications include venous thromboembolism (VTE), pneumonia, urinary tract infection, renal failure, stroke, sepsis and cardiac arrest. Readmissions and unplanned emergency room visits within 30 days of the procedure were reviewed. Readmissions and emergency room visits were cross-referenced with specific complication outcomes to identify possible additional medical and surgical complications. Inclusion criteria included all elective and urgent spinal surgeries and patients were excluded if they underwent day surgery or surgery for intradural tumors.

3.2.3 Model Development and Internal Validation

Logistic regression and a number of ML models were developed and their performance compared. In the analysis a 10-fold cross validation repeated 5 times was used for model comparisons and the AUROC was used as a performance metric.

3.3 Second Stage: Surgeon Complication Prediction and Model Output Perception

The purpose of this stage of the research was to determine how spinal surgeons and trainees of different levels of training and experience, use information provided by a prediction model to predict the risk of post-surgical complications in patients undergoing spinal surgery. An online survey instrument was developed and it included patient vignettes describing a diverse population of patients undergoing spine surgery. Information about the patients was similar to that which a spine surgeon has available during the encounter when a decision about surgery is made.

3.3.1 Survey Participants

The participants formed a convenience sample of staff spinal surgeons, surgical fellows, and resident trainees with the Department of Surgery at TOH. All participant staff and fellows were spinal orthopedic surgeons or neurosurgeons and perform the majority of spinal surgery at TOH. Resident trainees were currently enrolled in orthopedic surgery or neurosurgery residency programs at the University of Ottawa and TOH.

3.3.2 Study Design

In order to answer the research questions, an online survey was developed with Qualtrix and it was conducted in two phases separated by two weeks (44):

Phase 1: To assess surgeons' unassisted, post-surgical complication prediction accuracy using clinical vignettes. The purpose of this phase is two-fold. Firstly, the literature review has demonstrated that there are no published studies that have determined spinal surgeon accuracy for predicting post-surgical complications. Secondly, this phase serves as baseline surgeon responses to determine the impact of information provided by the ML model on surgeon predictions in Phase 2.

Phase 2: To assess the impact of providing the model output (existence of post-surgical complication: present or absent) on surgeons' prediction of a post-surgical complication using clinical vignettes. The participants were informed about the ML model's performance in predicting complications. In this phase, the participants were also given a small number of vignettes with the model output and an explanation of the model's prediction. They were asked open-ended questions regarding their perception of the accompanying explanation.

The study design is summarized in Figure 1.

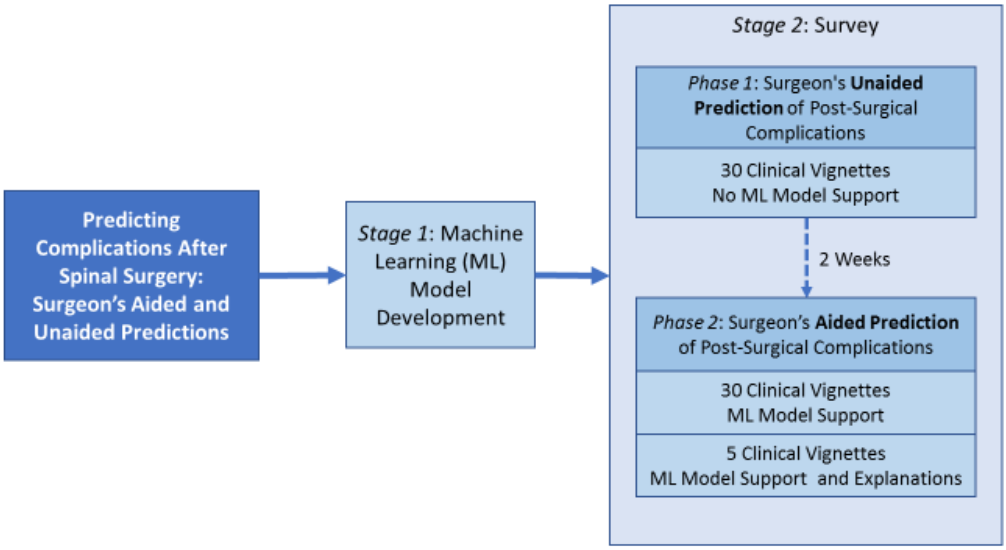


Figure 1: Summary of study design

To allow for a direct comparison of surgeons' predictions with and without information provided by the model, we used the same clinical vignettes in the two phases but the administration of the survey was separated by two weeks to minimize any recall or learning bias (44). The order with which the vignettes are presented remained unchanged but was different for each participant. The survey instrument included closed questions (agree, disagree, neutral), open-ended questions, and comment boxes.

The survey instrument was pre-tested on a small number of surgeons and trainees who did not participate in the study. Two clinical vignettes were used to allow for comments and feedback on format, instructions and ease of understanding. Comments were reviewed and changes were made as required prior to formal study initiation.

3.3.3 Clinical Vignettes

Thirty clinical vignettes were developed using patient data from the TOH NSQIP database. Each vignette consisted of simplified patient summaries, with descriptions of characteristics that could

impact a surgeon's prediction of the risk of complications. These characteristics include age, gender, diagnosis, surgical procedure, whether the procedure was planned or urgent, comorbidities, smoking history and body mass index (BMI). The description provided in the vignettes mimicked the overall picture a surgeon would have at the time of the surgical decision-making with a patient. The clinical vignettes were selected to reflect a variety of clinical contexts including elective and unplanned surgeries, diagnoses including degenerative, traumatic and oncologic conditions, and various types of procedures.

To minimize the risk that surgeons would state no complication as their default answer when predicting, and therefore, potentially outperform the model simply by knowing the expected prevalence of complications, we explicitly stated that the clinical vignettes describe patients who, as a cohort, had a higher than expected prevalence of complications. The goal of this was to ensure that surgeons would evaluate each case individually based on the features provided as the model does.

3.3.4 Survey Instrument

The online survey instrument included a description of the study, background information and rationale for the study. Each participant received an individual access link and was able to create a unique login and password. Before starting the survey, participants had to answer questions about his/her level of training and specialty. In the first phase of the survey, after being presented with a vignette, participants were asked to make a complication prediction as either present or absent.

Following completion of the 30 clinical vignettes, the first phase included questions pertaining to the surgeons' use of complication predictive tools and their opinion as to the patient's characteristics most likely to impact the risk of complications. As outlined in the literature review, given the paucity of research on surgeons' perceptions and use of predictive models, no questionnaires or survey instrument were available for adaptation in this study; therefore, they had to be developed for this purpose. The questions in the survey were:

Question 1: Do you use a surgical risk prediction tool in your practice? Check all that apply (Spine Sage, Spinal risk assessment tool, NSQIP risk calculator, other, none).
Question 2: I have a good understanding of how risk prediction tools work. (agree, neutral, disagree, not applicable).
Question 3: Factors most important in determining the risk of complications associated with spinal surgery. (Rank of age, BMI, comorbidities, diagnosis and procedure)

The second phase of the survey had the same purpose and background information as phase 1 in addition to information about the ML model's sensitivity, specificity, and AUROC. Each vignette also included prediction information produced by the ML model with some predictions being correct and some incorrect. Surgeons' perceptions were assessed with these questions at the conclusion of the survey:
Question 1: Prediction produced by the model helps spinal surgeons more accurately predict the presence or absence of a post-operative complication for a patient. (agree, neutral, disagree).
Question 2: Having a complication prediction generated by the model can improve shared decision making between patients and surgeons. (agree, neutral, disagree)
Question 3: Was the prediction generated by the model useful? Why or why not? Provide comments.

The last part of the Phase 2 survey included five repeated vignettes provided with additional information explaining the model's prediction. For example, the model not only provided the binary complication prediction (present or absent), but it also provided an explanation as to how the prediction was derived. This may have consisted of specific features such as age, BMI, diagnosis and comorbidities together with their values that drove the model's prediction. The purpose of this part of Phase 2 was to explore whether additional information provided by the ML model would affect surgeons' use and perception of the model, particularly if this new information helped the surgeon understand how the model derived its prediction. At the end, the participants were asked to answer these questions:
Question 1: I find having an explanation of the prediction result useful. (agree, neutral, disagree)

Question 2: Is the model prediction explanation presented in an understandable and useful way or should it be presented otherwise?
Question 3: If the model should be presented otherwise then please provide an example.
Each survey section had an open ended option for comments.

3.3.5 Analysis of Results

Considering the small sample size, results were analyzed using mostly descriptive statistics. Moreover, whenever appropriate we used non-parametric statistical tests that are applicable to smaller samples and do not require normal distribution of data.
In phase 1 we used Fleiss kappa statistics to assess inter-observer agreement between respondents in terms of predicting the risk of complications. This measure is a generalized version of Cohen's kappa that deals with more than two observers (45).

In phase 2 we checked the overall impact of the ML model support on participants' predictions of post-surgical complications using McNemar's test. McNemar is a specific test for dichotomous variables to assess consistency between two groups (46). To determine the impact of the support provided by the ML model prediction for each vignette on the surgeons' predictions we used signed-rank Wilcoxon test. We used this test to analyze responses provided by individual surgeons, so we were able to identify those who benefited from having additional information in terms of the ML model's predictions. Moreover, we also checked the impact within groups of participants (related to level of training) and over all participants. Cohen's kappa was used to determine intra-observer agreement between phases.

Chapter 4: Results

4.1 Preliminary Stage: Development of Machine Learning Model

One thousand five hundred and fifty consecutive patients that underwent spinal surgery at TOH from 2016-2018 were included in the analysis. This includes both elective and urgent surgery, inpatient and outpatient surgery, and surgery for degenerative, traumatic, infectious and oncologic etiologies. Patients undergoing outpatient surgery were excluded from model development given the low likelihood of complications in this group. Patients with intradural tumours were also excluded from the analysis as this procedure is not performed by orthopedic spine surgeons. This left us with 893 patients for model development. Feature selection was determined by the spine surgeon and supported by computer science and health informatics researchers and these are listed in the Table 2. Due to variable and overlapping hospital codes for procedures and diagnoses, all cases in the database were reviewed by the surgeon (SK) to create 4 mutually exclusive procedural categories and diagnostic categories.

Table 2: Patient features used for model development

Feature Name	**Domain of Values**
Age	Numerical
Urgent/Elective	Binary
Procedure*	Ordinal
Diagnosis**	Ordinal
Sex	Binary
BMI	Numerical
Diabetes	Binary
Hypertension	Binary
Renal Failure	Binary
Dialysis	Binary
Disseminated Cancer	Binary

Steroids Use	Binary
Bleeding Diathesis	Binary

*Procedures: anterior cervical, posterior cervical, thoracolumbar decompression +/- fusion <3 levels, thoracolumbar decompression +/- fusion >3 levels

**Diagnosis: degenerative, traumatic, infectious, and oncologic

A number of ML models were initially identified and their performance was assessed using a 10-fold cross validation that was repeated 5 times, so for each model 50 evaluations were obtained. Missing data were imputed with means for numerical data and medians for ordinal or binary features. The AUROC was used as the evaluation measure. As a result of this preliminary stage, logistic regression and gradient boosted decision tree (XBT) ML models were found to have the best performance with 0.631 and 0.596 AUC. The values of the AUROC's reflect the difficulty of the prediction problem. The XBT ML model was selected as it has been successfully applied and validated for predicting different clinical outcomes in different settings and would facilitate the development of the model output explanations in phase 2 (47, 48). Table 3 shows a comparison of ML models tested in this experiment.

Table 3: Comparison of ML algorithms

ML algorithm	**AUC**	**95% CI**
Adaboost	0.559	(0.533,0.586)
Decision tree with Gini index	0.505	(0.490, 0.520)
Decision tree with entropy	0.502	(0.487,0.516)
Logistic regression	0.631	(0.607,0.655)
Random forest	0.535	(0.514,0.557)
Gradient boosted decision tree	0.596	(0.571, 0.620)

4.1.1 Model Performance: Thirty Vignettes Selected for Survey

The XBT model's accuracy for the 30 clinical vignettes was 0.53 which is considered relatively poor but consistent with the result (0.60) obtained for the training set derived from the NSQIP

data. As such, the case-mix of difficulty for the 30 vignettes is likely representative, albeit slightly more difficult, than the whole cohort.

4.2 Second Stage: Participants Unassisted Predictions (Phase 1 Survey)

Twenty participants were approached to participate in the survey. Seventeen participants completed phase 1 of the survey and 14 completed phase 2 of the survey. The experience and specialty training of those completing both phases are summarized in Table 4.

Table 4: Experience and specialty training of survey participants completing Phase 1 and 2

Specialty			
Phase 1			
Training	Neurosurgery	Orthopedic Surgery	Total
Fellow	2	1	3
Resident	5	3	8
Staff	3	3	6
Total	10	7	17
Phase 2			
Fellow	2	1	3
Resident	3	3	6
Staff	2	3	5
Total	7	7	14

4.2.1 Participant Agreement: Phase 1

The agreement of the 17 participants' predictions of post-surgical complication (present or absent) for the 30 vignettes was determined using Fleiss kappa as the ratings are binary and there are multiple raters. The results are summarized in Table 5. Overall the Fleiss kappa demonstrates that there was fair agreement (kappa 0.21-0.40) between participants (49). That is, participants' predictions for each vignette demonstrated some agreement (better than random) but suggesting there is certainly no consensus for predicting the vignette outcomes overall. Surgeons trained in orthopedic surgery had better agreement than those trained in neurosurgery and residents had the best agreement when considering level of training.

Table 5: Participant agreement Phase 1 determined by Fleiss kappa

Category	**Fleiss kappa**
Overall	0.218
Neurosurgery	0.159
Orthopedic Surgery	0.275
Fellow	0.000
Resident	0.282
Staff	0.284

4.2.2 Participant Performance: Phase 1

The model performance (0.53) for the 30 vignettes may be considered poor, and therefore, not clinically useful by surgeons; however, the model actually outperformed the majority of participants. The model and prediction accuracy for the 30 vignettes, similar to flipping a coin, suggests that the predictions requested are inherently difficult. The prediction accuracy for phase 1 is described in Tables 6 and 7. In phase 1 only three residents and one fellow predicted post-surgical complications better than the model (0.57-0.63) while 11 participants predicted worse (0.37-0.50). Generally, residents and fellows predicted post-surgical complications better than staff surgeons for the 30 vignettes. As front line physicians that are more likely to initiate treatment of post-surgical medical complications residents may actually be more attuned to patients that are at higher risk despite less surgical experience. Staff surgeons may put more emphasis on surgical complications that are specific to spinal surgery and may not be captured by NSQIP unless there is a repeat operative intervention or readmission. However, the distribution of performance by specialty and experience may simply be a result of the small sample size.

Table 6: Prediction performance of participants in Phase 1

Participant	Level of Experience	Specialty	Accuracy
7	Resident	Orthopedic Surgery	0.63
11	Resident	Neurosurgery	0.60
6	Fellow	Orthopedic Surgery	0.57
8	Resident	Orthopedic Surgery	0.57
1	Staff	Orthopedic Surgery	0.53
4	Fellow	Neurosurgery	0.53
	Model		0.53
3	Resident	Neurosurgery	0.50
12	Staff	Orthopedic Surgery	0.50
9	Staff	Orthopedic Surgery	0.47
10	Resident	Neurosurgery	0.47
16	Staff	Neurosurgery	0.47
2	Resident	Neurosurgery	0.43
13	Resident	Neurosurgery	0.43
15	Staff	Neurosurgery	0.43
17	Staff	Neurosurgery	0.43
5	Fellow	Neurosurgery	0.40
14	Resident	Orthopedic Surgery	0.37

Table 7: Prediction performance based on specialty and level of experience

Category	Accuracy
Neurosurgery	0.47
Orthopedic Surgery	0.52
Fellow	0.50
Resident	0.50
Staff	0.47

One of the secondary objectives of this research was to provide a baseline reference for spinal surgeon complication prediction accuracy. In phase 1 of this study, surgeons' unaided prediction accuracy was 0.49. It is clear, however, that rather than a baseline reference of surgeon prediction performance, it is more relevant to have surgeon performance presented with model performance for specific models developed. As outlined in the literature review, the number of publications pertaining to ML-based prediction models in spinal surgery is increasing but without published surgeon accuracy there lacks a critical comparator (22, 33, 34, 36, 37). It is important that surgeon accuracy for predicting a gold standard be analyzed and compared to a model's accuracy for the same prediction and with similar information provided, as was done in this study. It is quite likely that depending on the prediction problem of interest, human and model performance will vary. For this study, the prediction of interest was the presence or absence of a post-surgical complication. However, this may not be the optimal prediction for surgeons or patients. A determination or prediction of high, medium and low risk patients may be of more value and may be more accurately predicted.

It is difficult to compare ML prediction model performance in this study with other published studies that looked at complication predictions. Although the study by Kim et al. used the NSQIP database and as such would have comparable features and outcome collection methods, their study looked at specific complications including VTE, cardiac complications, wound complications and mortality as opposed to any complication. They reported that logistic regression and artificial neural networks outperformed ASA based on AUC in predicting complications but they did not specifically report the AUC values. Instead, they used a heatmap and the AUC values ranged from 0.37-0.71. Possible explanations for the difference in prediction accuracy between studies may include subtle variations in model development including feature selection or outcomes, specific characteristics of our patient population or overfitting of other published models. The study by Bekelis et al. may be an example of overfitting as they reported an AUC of 0.95 for post-operative stroke which had a 30 day risk of 0.05% in the cohort of 13,660 NSQIP patients (24). Certainly, it is difficult to understand and appreciate the specifics of other published prediction models given the lack of publications pertaining to external validation, model updating and impact analysis as outlined by Moons et al (16). The model developed by Scheer et al reported a prediction accuracy of 0.89 for predicting

major intraoperative or early postoperative complication in patients undergoing deformity surgery (34). Such a model could prove very useful for anticipating significant complications and informing patients preoperatively of the high surgical risk. The high accuracy reported for this model may be related to the more homogeneous cohort as well as the extensive database used including patient-specific features, quality of life data and radiographic features. Class imbalance would be less of an issue for this model as the complication rate for this cohort undergoing higher risk surgery exceeded 25%. Although this model appears to be a very valuable tool, without surgeons' unassisted prediction accuracy, the model accuracy reported lacks a comparator.

4.2.3 Vignette Difficulty: Phase 1

For each clinical vignette a difficulty value was determined. This value was calculated by first determining the average accuracy among all participants for the particular vignette and subtracting this value from 1. The most difficult vignettes have a value closer to 1 while vignettes that were easier to predict have a value closer to 0. Vignettes 10, 19 and 30 all had difficulty values of 0.94 while vignette 23 had a difficulty of 0.11. The most difficult vignettes were also incorrectly predicted by the ML model and also by all but one participant. By reviewing the details of these vignettes it is clear that they describe the cases where no complications occurred despite high risk features or where a complication occurred with low risk features. Certainly these are scenarios where surgeons would likely make the same predictions in future comparable cases using the same logic and would likely be correct more often than not. Fifteen participants correctly predicted the easiest vignette likely because of the overwhelming high risk features. A summary of these cases are presented in Table 8.

Table 8: Difficulty of selected vignettes in Phase 1

Vignette	Description	Complication	Prediction	Difficulty
V10	69 year old male undergoing unplanned (priority F) two-level thoracic decompression and instrumented fusion for thoracic stenosis. His BMI is 30 and he has diabetes mellitus and requires dialysis	absent	incorrect	0.94
V19	77 year old female undergoing elective lumbar decompression for radiculopathy. Her BMI is 23 and she has hypertension	present	incorrect	0.94
V30	53 year old female of prednisone undergoing urgent (priority E) thoracic decompression and instrumented fusion for thoracic myelopathy. Her BMI is 47 and she has hypertension.	absent	incorrect	0.94
V23	60 year old male smoker undergoing elective multilevel lumbar decompression and instrumented fusion (>3 levels) for lumbar stenosis. He has a BMI of 48 and hypertension	present	correct	0.11

4.2.4 Surgeons' Perceptions of Surgical Risk Factors and Use of Risk Prediction Tools

The participants were asked about their agreement with the statement, "I have a good understanding how surgical risk prediction tools work". Of the 17 participants only two (12%) reported that they had a good understanding of how risk prediction tools work. Six participants disagreed with this statement and nine were neutral. Nine participants reported that they do not use risk prediction tools. The SpineSage tool and NSQIP risk calculator were each used by four participants and the spinalRAT was used by one participant. Both participants that reported a good understanding of risk prediction tools used the SpineSage tool. Four out of six participants that disagreed with the above statement did not use risk prediction tools.

It is clear from the responses in phase 1 of the survey that the participants do not have a good understanding of risk prediction tools despite six fellowship trained spine surgeons' participation in this phase. There are no studies to reference as to whether this would be representative of

other surgical centers. Forty-seven percent of participants did report using tools and SpineSage and NSQIP calculator were the most commonly used. Surgeons without education or training in predictive analytics are not likely to use models if they feel they will not help their, or patients', decision-making and add time to an already busy practice. If surgeons feel the predictive tools are valuable and superior to current practice, they are likely to take time and effort to understand the models. Furthermore, if surgeons don't have a good understanding of the models they are not likely to expose their trainees or encourage learners to use prediction models. Although the models are available and well known, the lack of use and understanding may reflect the ambivalence with the current literature pertaining to prediction models in spinal surgery. Participant knowledge of risk prediction tools should not have affected the phase 1 results; however, phase 2 results could have differed if the participants had greater understanding or experience with risk prediction tools as they may have placed greater or less emphasis on the ML model support provided.

The participants were asked to rank five risk factors from most important to least important and these responses are summarized in Table 9. The ranking of risk factors was used to understand and analyze the participants' prediction of post-surgical complications for the specific vignettes. Risk factors for complications in spinal surgery are well known (13, 24) but how surgeons rank or weigh their relative importance is not. From this small sample, there is no clear pattern that develops. Surgical procedure was most commonly rated as the most important risk factor (47%) but it was also rated the least important by 17% of participants. Diagnosis and BMI were most frequently ranked the second most important risk factor but were very infrequently ranked as the most important risk factor. How surgeons weigh risk factors was analyzed in greater depth as part of the assessment of vignette difficulty in phase 2.

Table 9: Participant ranking frequency of most important (1) to least important (5) risk factor for predicting post-surgical complication

	Frequency Risk Factor is Reported as Most Important (1) to Least Important (5)				
Risk Factor	Most Important	Second Most Important	Middle	Second Least Important	Least Important
Surgical procedure	8	2	0	4	3
Comorbidity	4	3	3	2	5
Age	2	2	9	2	2
Diagnosis	2	5	1	5	4
BMI	1	5	4	4	3

4.3 Second Stage: Phase 2 Survey Results

Out of the 17 participants who completed phase 1 of the survey, 14 completed phase 2. All seven female participants completed both phase 1 and 2.

4.3.1 Reliability Measures: Phase 2

The inter-rater reliability of the participants improved from phase 1 to 2, as determined using Fleiss kappa, from 0.218 to 0.339. The poor reliability in phase 1 may be partially reflective of the inherent difficulty in predicting post-surgical complications and the improved reliability in phase 2 may be related to the ML model support provided. Intra-rater reliability between the phases was moderate at 0.481 and 79% of participants had moderate or substantial agreement between phases. This suggests that most participants were consistent with their predictions even with the ML support. The two phases were conducted two weeks apart and therefore, vignette recall was likely minimal and insignificant regardless as participants were never given the "gold standard" outcome of the vignettes. The intra-rater reliability results do support the quality of the vignettes as they were sufficiently descriptive to allow for moderate agreement considering the additional ML support. In other words, if these vignette descriptions lacked sufficient detail, the intra-rater reliability would be expected to be poor.

4.3.2 Participant Prediction Accuracy: Comparing Phase 1 and 2

In phase 2 of the survey, the participants received support with the ML model's prediction as to the presence or absence of a post-surgical complication. A comparison of the surgeons' overall accuracy for phase 1 and 2 is depicted in Figure 2. Overall, surgeon prediction accuracy improved from an average of 49.1% (206 correct responses) in phase 1 to an average of 54.8% (230 correct responses) in phase 2. This difference is statistically significant using McNemar's test ($p=0.024$). Specifically, 8 participants demonstrated improved complication prediction accuracy with ML model support, 2 had equivalent prediction accuracy between phase 1 and 2, while 4 had worse prediction accuracy with the addition of ML model support. The prediction accuracy of surgeons in phase 2 is comparable but slightly greater than the model performance (0.53) for the 30 cases selected for the survey. The statistically significant improvement in prediction accuracy for the 14 participants using the ML model suggests that despite a model with mediocre prediction accuracy, the ML model can improve prediction performance. These findings are important as the additional information derived from a model may cause surgeons to reflect on their own outcome and complication prediction abilities. It is clear that predicting post-surgical complications in patients undergoing spinal surgery is difficult, for models and experienced surgeons, but it is not surprising that additional information provided by the ML model was beneficial overall. This is the first study in the spine surgery literature that has evaluated the impact of a ML prediction model on surgeon complication prediction accuracy. In general surgery, the *MySurgeryRisk* tool was found to improve physician risk assessment for certain types of postoperative complications but this tool is already embedded in a decision-support platform (43).

It is important to examine the participants with a decrease in prediction performance in phase 2. Two orthopedic surgery residents (P7 and P8) were in this group but they had above average prediction performance in phase 1 (0.63 and 0.57) and their drop in prediction accuracy may reflect a move toward the mean of the group. The prediction accuracy drop for P7 was 0.63 to 0.53 and for P8 was 0.57 to 0.47. This may have occurred because they were anchored by the model and moved towards the model's performance or they may have simply performed better than would be expected in phase 1. Of greater concern would be that two staff surgeons (P9 and

P16) predicted below average in phase 1 (both 0.47) and their prediction performance worsened in phase 2 to 0.37 and 0.43 respectively. Given that these are more experienced surgeons possible explanations for poor prediction accuracy in both phases include a false perception that post-surgical complications are more likely to occur in patients with greater risk factors or that complications in this population are rare. The latter consideration should have been mitigated by informing the participants that there was a greater prevalence of complications in the 30 vignettes than normal. The deterioration in performance in phase 2 could be explained by experienced staff surgeons recognizing that the ML model performance of 0.53 reported would be considered poor and, considering their confidence in their own prediction ability, actively questioning the ML model support more than other participants to their detriment.

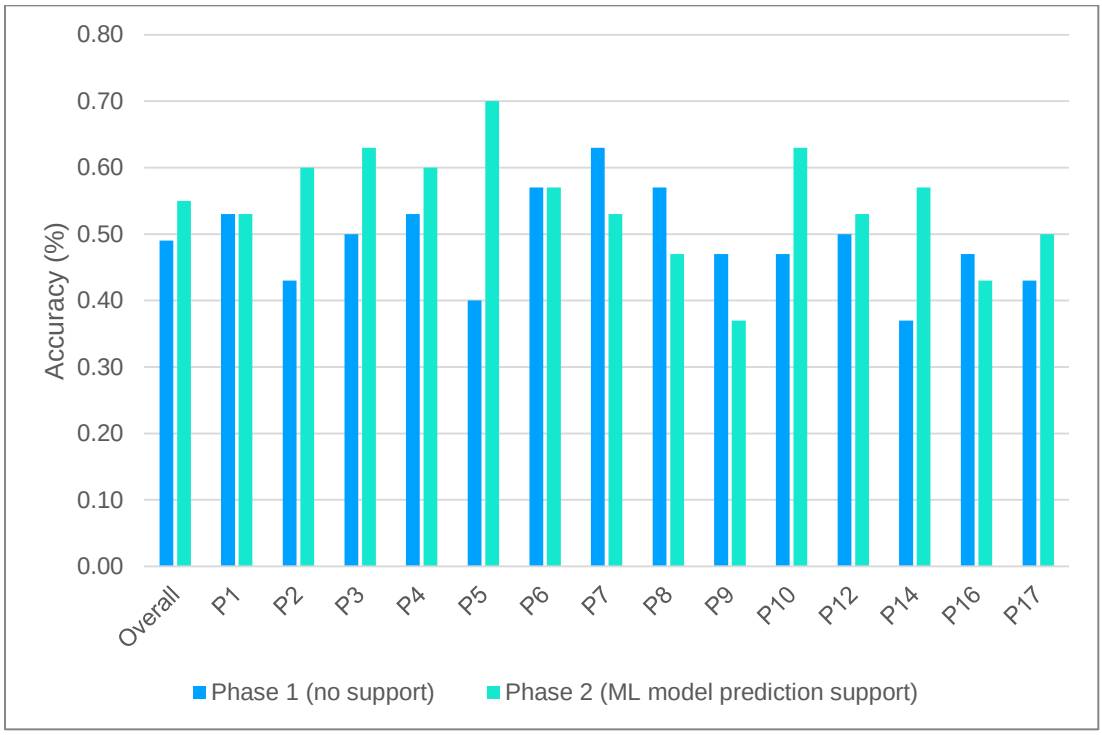


Figure 2: Complication prediction accuracy of surgeons in phase 1 and 2

4.3.3 Participant Prediction Accuracy: Correct and Incorrect ML Model Predictions

The effect of ML model support on prediction accuracy was analyzed separately when the model provided correct and incorrect predictions. In this study there were 16 vignettes with correct

model predictions and 14 vignettes with incorrect model predictions. These results are summarized in Figures 3 and 4 respectively. Figure 5 depicts the differences in prediction accuracy between phase 1 and 2 for all vignettes and when separated into correct and incorrect ML model predictions. Positive values are indicative of improvement in prediction accuracy with ML model support.

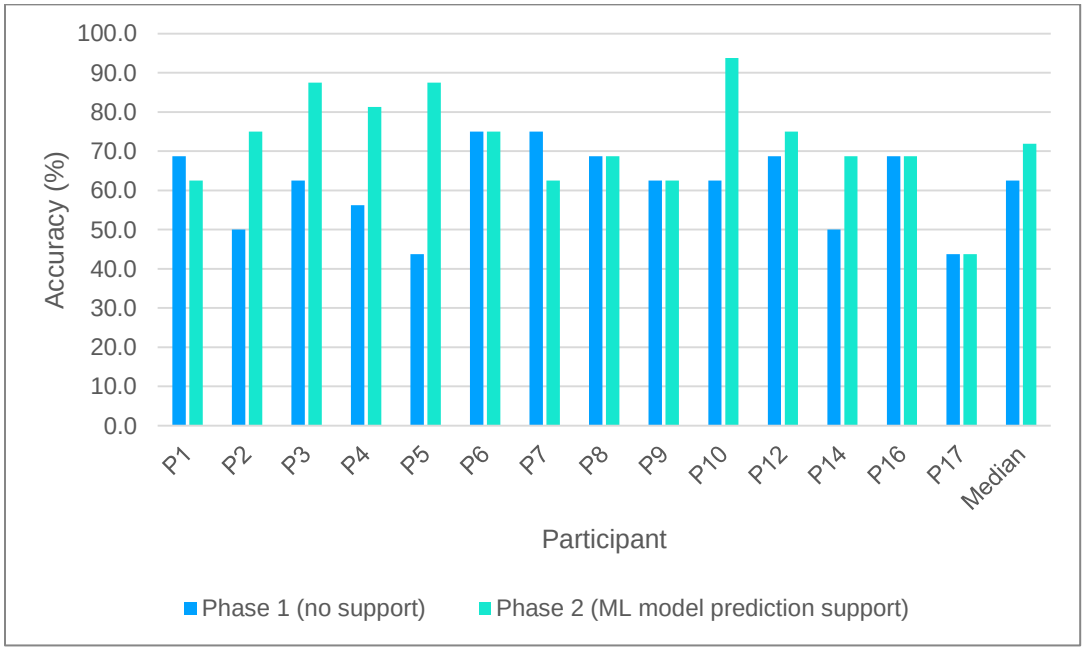


Figure 3: Surgeon prediction accuracy for vignettes with correct ML prediction

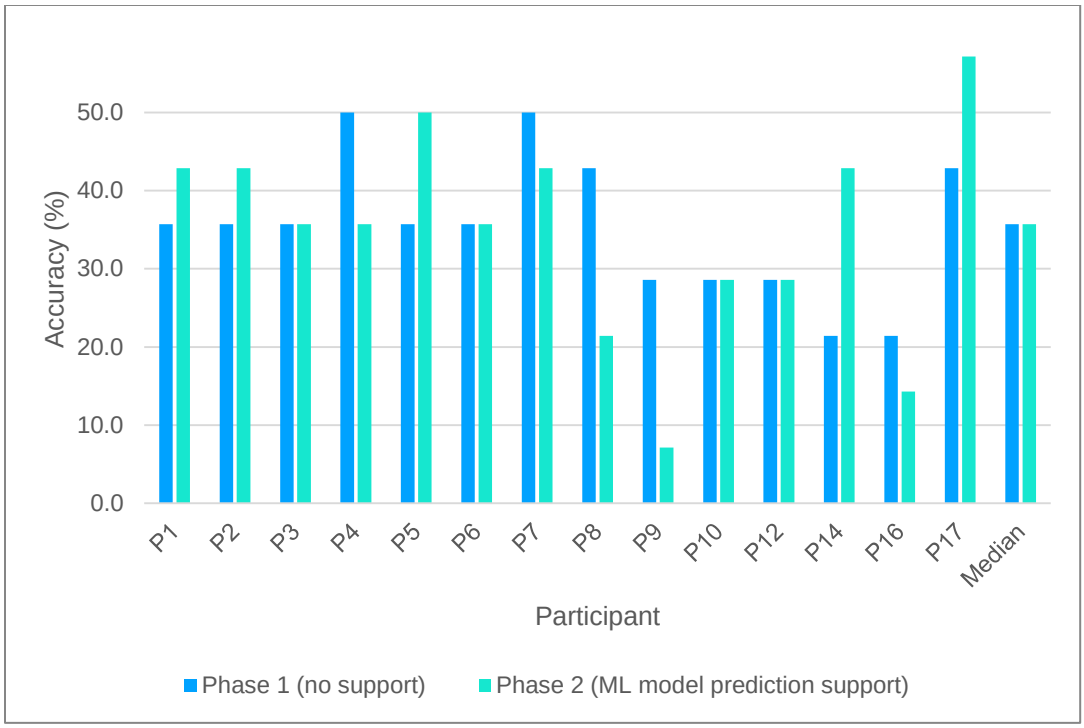


Figure 4: Surgeon prediction accuracy for vignettes with incorrect ML prediction

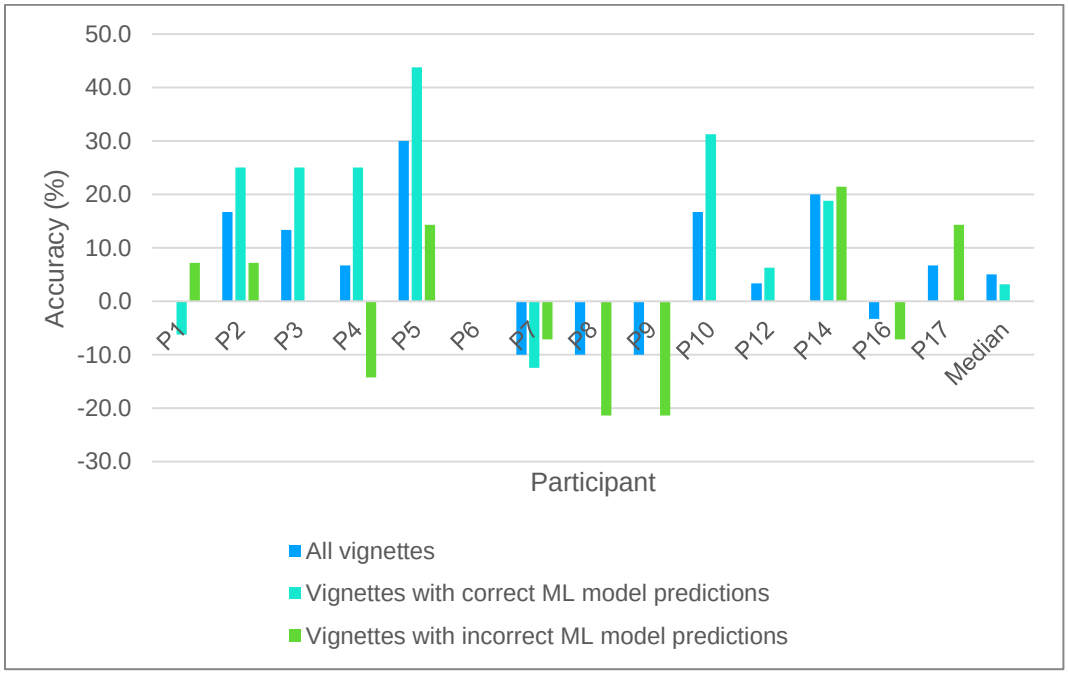


Figure 5: Difference in participant accuracy between phase 2 and 1

The effect of ML model support between phase 2 and 1 overall and with correct and incorrect predictions is summarized in Table 10.

Table 10: Effect of ML model support on prediction accuracy between phase 2 and 1

	Percentage of participants and prediction performance from phase 1 to 2		
	Improved	Worsened	No change
Vignettes with correct ML prediction	50%	14%	36%
Vignettes with incorrect ML prediction	36%	36%	29%
All vignettes	57%	29%	14%

Most participants benefitted from ML model predictions and this was most notable for correct predictions. The ML model support was most helpful when the ML model predictions were correct. Prediction accuracy improved in 50% or remained unchanged for 36% of participants when provided with correct predictions in phase 2. These results are not surprising as most participants were likely able to use the correct ML model predictions to reinforce their correct predictions or question, and on occasion alter their incorrect predictions. Only 2 participants had a drop in their prediction accuracy in this setting and one of these participants was a resident (P7) who had the highest accuracy (0.63) in phase 1. The resident's drop in performance with correct predictions may simply be an expected move toward the mean accuracy of the group. Participants 9 and 16 were two staff surgeons who predicted worse in phase 2. They did not benefit overall from correct predictions and, again, this may have been due to devaluing the ML model support based on the reported model accuracy of 0.53.

When incorrect predictions were made in phase 2, the effect was more varied as five participants had improvement, five had a deterioration and four remained unchanged in their accuracy overall. Four of the participants (P2, P5, P14 and P17) that demonstrated improved prediction accuracy in phase 2 when given incorrect ML model predictions had poor prediction accuracy in phase 1 (0.43 or less). All of these participants had improved accuracy in phase 2, including with correct ML model predictions, and again, this may reflect a move toward the mean of the

group. This is a more likely explanation for their improvement as opposed to correctly determining when the ML model provided correct and incorrect support.

Participants 8 and 9 had the greatest drop in prediction performance when given incorrect ML model support. Once again, given P8 (resident) performed well in phase 1 but did not change with correct ML support and worsened significantly with incorrect ML support, this may be an expected move to the mean accuracy of the group. Participant 9 is a staff surgeon whose prediction performance dropped between phase 1 and 2 from 0.47 to 0.37. This can be fully accounted for by the performance on vignettes with incorrect ML model support, and therefore this staff surgeon appeared to overvalue the ML model support.

4.3.4 Participant Prediction Accuracy: Level of Training and Specialty

The participant prediction accuracy results based on level of training are depicted in Figure 6. Resident and fellow participants benefitted most from the ML model support as the median accuracy for their group improved between phases and in particular for correct ML model predictions. This finding would be expected if the ML model support is more helpful for less experienced surgeons. These surgeons in training may have recognized their limitations with respect to predicting complications and deferred to the ML model. The ML model support did not affect staff surgeons' predictions notably. There was no difference in median accuracy between phases for staff overall and with correct or incorrect ML model support. Staff surgeons may feel that their ability to accurately predict post-surgical complications is superior to the ML model, particularly if they recognized that the model accuracy (0.53) is considered poor, although the model was equivalent to one staff and superior to four staff in phase 1.

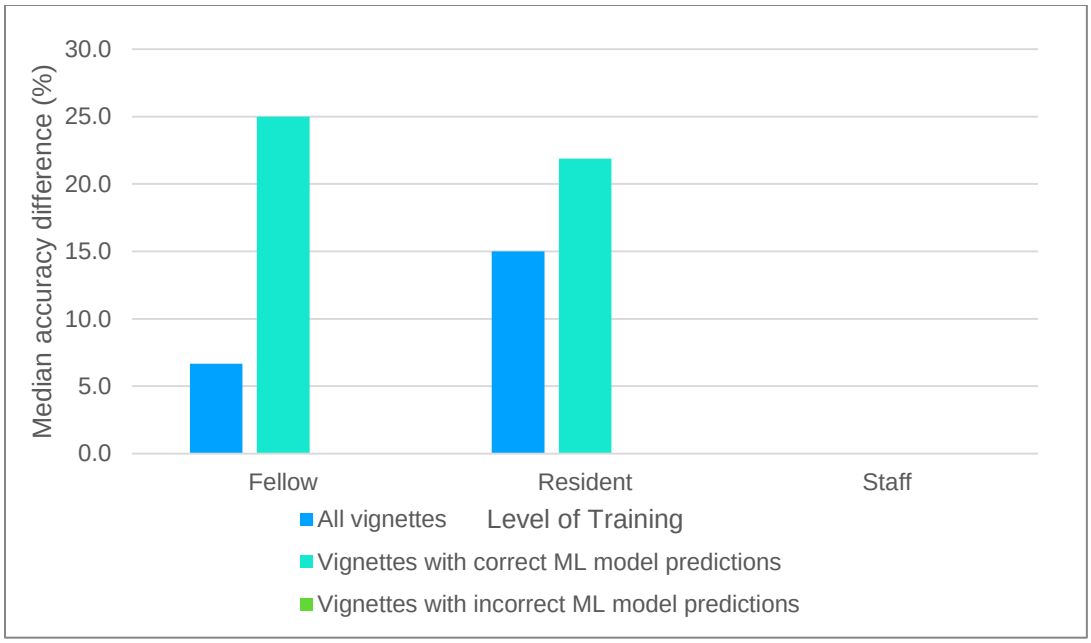


Figure 6: Median difference in accuracy between phase 2 and 1 and by level of training

Figure 7 shows the median difference in accuracy between phase 2 and 1 based on specialty. As a specialty group, the median prediction accuracy of neurosurgical participants was positively affected by the ML model support and most notably, with correct predictions. This may reflect a training or education effect whereby prediction models have greater support within the neurosurgery community. As described previously, given the relative poor prediction performance of neurosurgical participants in phase 1, these results may reflect an expected movement toward the mean of the group.

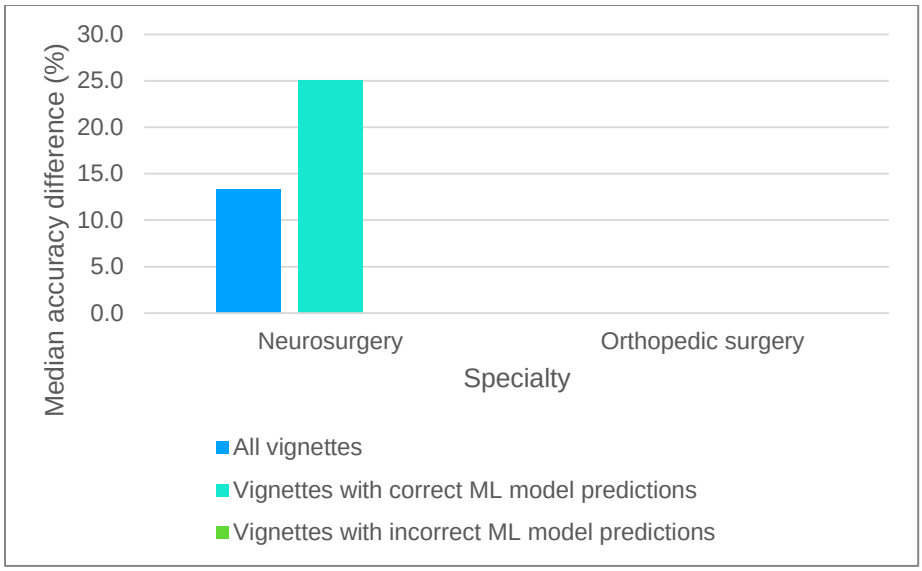


Figure 7: Median difference in accuracy between phase 2 and 1for specialty

4.3.5 Analysis of Vignette Difficulty with ML Model Support

Complication prediction accuracy across vignettes, for correct and incorrect ML predictions, is depicted in figures 8 and 9.

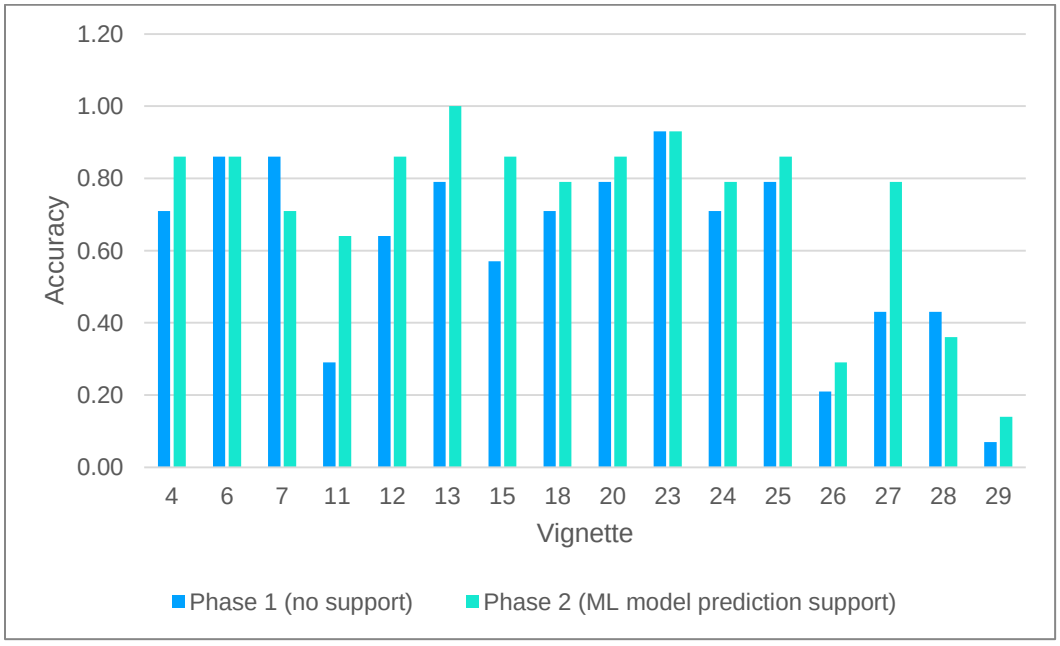


Figure 8: Complication prediction accuracy across vignettes with correct predictions

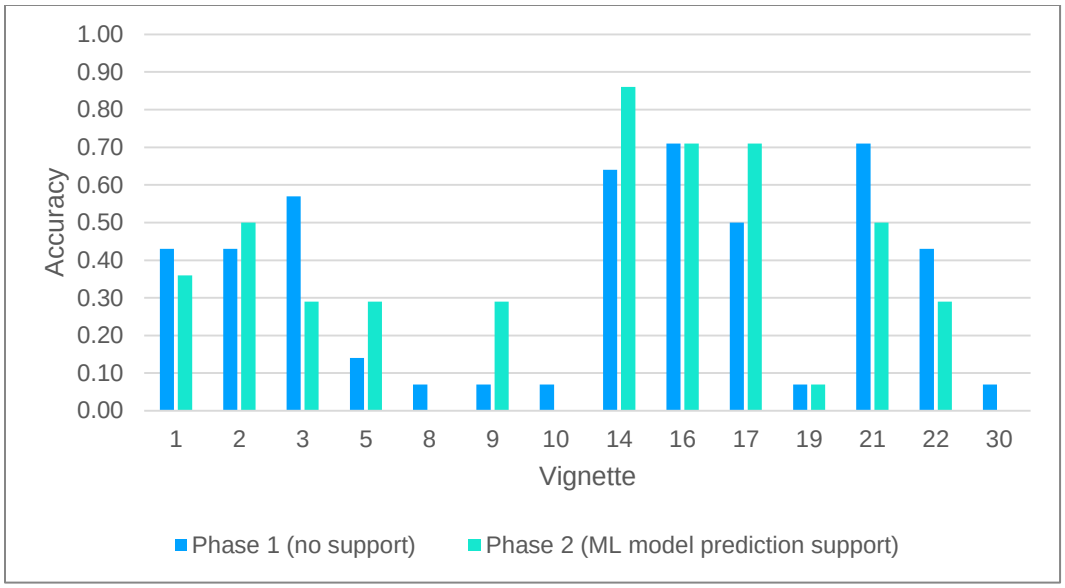


Figure 9: Complication prediction accuracy across vignettes with incorrect predictions

Despite ML support, four vignettes (V8, V10, V19, V30) remained very difficult (accuracy <0.10) to predict and three of these vignettes were incorrectly predicted by all the participants in phase 2. Three of these cases are described in table 8 and all of these vignettes were incorrectly predicted by the ML model. Participants had difficulty correctly predicting complications for vignette 9 in phase 1; however, their accuracy improved (0.22) with the ML model support. These cases highlight the difficulty with complication prediction as they all had considerable high or low risk features which dictated the prediction. Given that these vignettes were incorrectly predicted by the ML model and all participants, these are likely cases where experts would make the same prediction in similar future scenarios based on the same logic. These are cases that are considered very high or low risk and although incorrect predictions were made in these cases, the predictions would more often be correct in future similar scenarios. These four cases are outliers in the sense that vignettes with significant agreement between participants and the ML model would typically be the easiest cases to predict.

Table 11 describes the 6 vignettes (3 positive and 3 negative) where the ML model support had the greatest impact on prediction accuracy.

Table 11: Prediction results for the vignettes where ML model support had greatest effect

Vignette	Complication	ML prediction	Surgeon mean accuracy change	Vignette description
27	present	correct	improved 0.43-0.79	74 year old female undergoing elective lumbar decompression and TLIF for lumbar radiculopathy. Her BMI is 37 and she has a past medical history of hypertension
11	present	correct	improved 0.29-0.64	82 year old female undergoing elective lumbar decompression for spinal stenosis. She has a BMI of 33 and hypertension.
15	absent	correct	improved 0.57-0.86	60 year old female undergoing elective revision lumbar decompression and instrumented fusion for pseudarthrosis. Her BMI is 23.
3	present	incorrect	worsened 0.57-0.29	58 year old female smoker undergoing elective lumbar decompression and TLIF for spondylolisthesis. She has a BMI of 20.
21	present	incorrect	worsened 0.71-0.50	68 year old female undergoing urgent posterior cervicothoracic decompression and instrumented fusion for cervical spine fracture. She has a BMI of 29 and her past medical history includes diabetes mellitus and hypertension.
7	present	correct	worsened 0.86-0.71	87 year old male undergoing non-elective cervical decompression and instrumented fusion for stenosis with a BMI of 31 and hypertension.

The three vignettes where the ML model had the greatest positive effect were all cases where the ML model support was correct and participant accuracy was below the median in phase 1. These were all vignettes that could be considered to have balanced features that did not strongly favor a specific complication outcome but that were more often than not incorrectly predicted by participants. For V11, participants predicted very poorly in phase 1 (0.29) likely due to the fact it was a low-risk surgical procedure but with the ML model support, the participants may have reconsidered the patient's age and BMI, which led to more participants correctly predicting a post-operative complication (0.64) in phase 2. It is likely that the correct ML model support in

these cases was sufficient to change the participant prediction or lead the participants to view the case differently.

It would be expected that incorrect ML model support would have the greatest negative effect on prediction performance and this was the case for V3 and V21. These were both vignettes where prediction accuracy was better than the median for vignettes with incorrect ML model support and therefore, where surgeons performed well in phase 1. The incorrect prediction was deleterious on performance for these vignettes. It should be noted, however, that prediction performance decreased in phase 2 in only seven (50%) vignettes with incorrect ML model support and the median decrease in accuracy for the vignettes was 0.07.

For V7, performance worsened in phase 2 despite correct ML model support. Participants performed well in phase 1 (0.86) as this was the second easiest vignette to predict and the best explanation is that the vignette risk assessment changed in phase 2 for a few participants and the correct ML support was ignored by those surgeons.

There were also a number of vignettes where the surgeons differed greatly from the ML model predictions. An analysis of vignettes where surgeon accuracy was at or below 0.30 for vignettes with correct model predictions and accuracy at or above 0.70 for vignettes with incorrect model predictions is found in Table 12. These are vignettes where the surgeons performed well and the ML model did not and vice versa.

Table 12: Analysis of vignettes where surgeon predictions differed greatly from ML model

Vignette	Complication	ML model	Surgeon mean accuracy phase 1 and 2	Vignette description
26	absent	correct	0.21-0.29	56 year old male on prednisone undergoing elective anterior cervical discectomy and fusion for myelopathy. His BMI is 41 and he has hypertension.
29	absent	correct	0.07-0.14	61 year old female on prednisone undergoing elective lumbar decompression and TLIF for lumbar radiculopathy. Her BMI is 33 and she has a known bleeding diathesis.
14	present	incorrect	0.64-0.86	68 year old female undergoing elective lumbar decompression for spinal stenosis. She has a BMI of 47 and hypertension.
16	present	incorrect	0.71-0.71	67 year old male undergoing elective revision lumbar decompression and instrumented fusion for lumbar radiculopathy. His BMI is 43 and he has a past medical history of hypertension.
17	absent	incorrect	0.50-0.71	45 year old male smoker undergoing unplanned non-elective anterior cervical discectomy and fusion for myelopathy. His BMI is 30.
21	present	incorrect	0.71-0.50	68 year old female undergoing urgent posterior cervicothoracic decompression and instrumented fusion for cervical spine fracture. She has a BMI of 29 and her past medical history includes diabetes mellitus and hypertension.

The vignettes summarized in Table 12 exemplify the challenges in understanding how surgeons predict post-operative complications. In six vignettes (20% of the survey) the surgeon prediction performance differed significantly from the ML model prediction. For V26, the ML model correctly predicted the absence of a complication yet the surgeons' accuracy was below 0.30. This is likely due to surgeons' consideration of prednisone use and a BMI of 41 as high risk features. It is interesting to note that most participants ranked surgical procedure as the most

important risk factor and BMI the least important in the phase 1 post-survey questions yet in this case a high BMI in the setting of a low risk procedure seemed to lead to the erroneous prediction of a complication. Similarly, for V29, the ML model correctly predicted no complication yet surgeon mean accuracy was 0.07-0.14 between phase 1 and 2. The features that surgeons likely considered high risk were the history of a bleeding diathesis, prednisone use and BMI of 33. Despite the correct predictions by the ML model, participant accuracy only improved marginally in phase 2.

There were four vignettes where the ML model incorrectly predicted the presence or absence of a complication and surgeons performed well. Vignette 14 describes a 68 year old patient undergoing a low risk procedure but with a BMI of 47. Participants were clearly influenced by the BMI and correctly predicted the presence of a post-surgical complication in phase 1. Despite the incorrect prediction by the ML model participant accuracy paradoxically improved in phase 2 and this was likely due to the participants valuing the BMI more than the model support. Once again, although the participants reported surgical procedure as more important than BMI in phase 1, it is clear that a threshold for BMI was reached that influenced participants to predict correctly the presence of a post-operative complication. Vignette 16 is similar in that it describes a 67 year old patient undergoing a higher risk procedure (revision and fusion) but again with a high BMI of 43. The participants correctly predicted the presence of a complication despite incorrect ML model prediction and this may have been due to the procedure complexity and BMI. The procedural complexity is difficult to capture for the ML model given the sheer number and heterogeneity of hospital procedural codes that need to be amalgamated during model development. Despite incorrect ML model prediction, participant accuracy did not change between phase 1 and 2. As with V14, despite incorrect ML model prediction participant accuracy improved from 0.50-0.71 between phases for V17. It is difficult to determine why this would occur. For an unknown reason, participant confidence increased that there would be no complication on their assessment in phase 2. In this case, the patient was a smoker and the surgery unplanned but there were no other high risk features and the procedure was low risk. For V21, the incorrect ML model prediction worsened participant accuracy between phases (0.71-0.50). This patient did have a complication and participants predicted well in phase 1 based on the fact it was urgent surgery for a traumatic fracture and the patient had diabetes mellitus. For

this vignette, the ML model prediction was clearly influential and perhaps resulted in the participants over-emphasizing the lower risk features in phase 2.

These six vignettes exemplify the inherent difficulty in predicting complications for models and experts. There are clearly features that overly influence experts and lead to erroneous predictions. Many features are actually spectrums but are dichotomized for the purpose of model development and presented as such in the vignettes. For example, if the data allowed for greater granularity features such as "diabetes" could be presented based on HgbA1C or "well controlled and poorly controlled". Similarly, prednisone use may be considered as "long-term use or short-term use". BMI was reported as a continuous variable but it is unclear at which value the expert begins to consider it high risk for complications. Is this done in conjunction with other comorbidities or is there a value at which BMI outweighs all other features?

The last four vignettes typify the concerns that clinicians may have with ML prediction models and the "black box" effect. Participants predicted well in these vignettes despite incorrect ML model predictions. It would be concerning for clinicians that the model was incorrect and there is no accompanying explanation that could be evaluated or rationalized. However, it is reassuring to see that in only one of these cases (V21) the model had a negative effect on prediction accuracy in phase 2 and the median decrease in prediction accuracy for vignettes with an incorrect ML model prediction was only 0.07. Furthermore, despite these four cases the model performed better than the participants overall. Finally, it should be noted that there were two cases (V26, V29) where the model correctly predicted the outcome and surgeons performed poorly, and for the other ten vignettes with incorrect ML model predictions, participants generally performed poorly as well.

4.3.6 Surgeons' Perceptions of the ML Prediction Model after Phase 2

After completing phase 2 of the survey, surgeons were asked whether they agreed that the ML model would help with predicting post-surgical complications, improve shared decision making with patients and if the model was useful. Despite, or perhaps because of, the participants' naivety with respect to surgical risk prediction tools, participants responded positively when

questioned about the potential benefits of the prediction model. Seventy one and 86% of participants answered that the prediction model could help surgeons improve complication prediction accuracy and help with shared-decision making respectively. Participant comments did highlight previously identified limitations of human and model predictions including over-simplification of features including comorbidities or other risk factors. The results are summarized in Table 13.

Table 13: Summary of surgeons' perceptions of the ML model after phase 2

	Number of participants and agreement		
Statement	Agree	No opinion	Disagree
Model prediction helps spinal surgeons more accurately predict the absence or presence of a post-surgical complication	10	2	2
A complication prediction generated by the model can improve shared decision making	12	0	2
I find the predictions generated by the model useful	9	2	3

After these questions, participants were asked to make comments and seven participants included comments. When appropriate, grammar was corrected for clarity. These are found below:
"*It will be useful to have a system to predict risk of complication. It will ease the surgeon job when discussing with the patient and it will encourage patient to correct modified risks (smoking, BMI...). It will provide common language among treating team (surgical staff, nurses, residents and other sub-specialty like anesthesia) to identify high risk groups of patients. However, there will be some challenges as not all the risks are the same for example: patient with DM type 2 for 10 years differ from DM type 1 for 5 years, whether DM is under tight control or not, smoking (how many cigarettes per day and how many years and type of smoking {concentration of CO and nicotine}), there different types of bleeding diathesis and there is different stages of chronic renal diseases... a lot of variables among specific risk, which will make accurate risk estimation difficult which need the physician to study each case individually.*"

This comment addresses many of the same issues that have been discussed in the study. This participant recognizes the model's potential value for surgeons, patients and the health care team for risk assessment and shared decision-making. The model could also provide value as a teaching tool to show patients how addressing modifiable risk factors can lower their surgical risk. This participant also recognizes the limitations of the model. A risk factor such as smoking or a medical comorbidity loses predictive value when it is not given a relative weight or severity. Model risk estimates will be affected by the granularity of the features used when developing the ML model.

"*The model has excellent potential but with the brief vignettes, I erred on the side of expecting a complication. With more information, I would be more likely to trust the prediction."*

This participant has also acknowledged the potential of the model and does state that more information could improve their trust in the model. This comment may be referencing the "black box" effect of the model and the place for an accompanying explanation of the model's prediction.

"Model can improve but needs to be based on individual surgeon's results as opposed to aggregate results. (It) would be great if there was a running model based on one own's results.

This participant's comments suggest that the ultimate clinical prediction model in this setting would be surgeon specific and change over time. This is very perceptive and quite accurate as some surgeons may actually treat patients with certain comorbidities better than other surgeons. This differential risk may be related to pre-operative planning, surgical technique or experience. Not surprisingly, some surgeons may be better than others at certain procedures for similar reasons and this would, therefore, alter the risk profile for patients. This could be addressed by including the surgeon specific information as one of the features of the model. Increased capacity of electronic health records may allow for real time model adaptation over time and this would certainly be welcomed.

"There is no clear way to decide about the complication risk based on patient characteristics beside models based on database, so those tools can be of big help."

This participant recognizes the potential benefit of the prediction model and the difficulty with predicting risk based on patient characteristics.

"*Model prediction seems to be spot on. I differ very occasionally from the model when it comes to high BMI. BMI greater than 40 is concerning for me, and is a predictor to me, anecdotally, for complications. But this may not necessarily be true. The predictor model outcome is dichotomous..present or absent. Perhaps a grade or scale of complication presence may be generated. These grades could be absent, low likelihood of complication, moderate likelihood of complication, present. This may be difficult to do at this stage.*"

This participant provided feedback on the model based on the survey. This comment does address the issue of feature thresholds for predicting risk. For continuous variables such as age and BMI there is likely a threshold where surgeons will change their risk prediction. This participant also addresses the model output. Given that there is a class imbalance for complications (complication rate is low) and surgeons may feel predicting post-surgical complications as dichotomous outcomes is not clinically relevant it may be more appropriate to separate model predictions into low, medium and high risk. This prediction outcome may be more intuitive to surgeons and patients.

"*It is hard to use a prediction system that I don't understand. What is the model based on? What is the accuracy of the model? Without this the model does not change my prediction."*

This comment is certainly appropriate. Surgeons are unlikely to use a prediction model that is not understood. Again, this does highlight the "black box" effect. A brief model description was provided to the participants in addition to the model's performance; however, interpretation does require some statistical knowledge. Once again, an accompanying explanation of the model's prediction could certainly be helpful.

"AI will help with avoiding your oversight (sic). It may not replace your experience and expertise but it may prove to help reveal aspects of your patient's health that you may have otherwise."

This participant is acknowledging the complimentary benefit of predictive analytics and AI. The goal is not to replace surgeon assessment and judgment but rather assist with another tool.

In summary, the comments provided highlighted and reinforced areas that have been discussed in the study. These include: the great potential of predictive models for patients and surgeons, the shortcomings of limiting features' granularity, the importance of determining the most clinically intuitive prediction outcome and that the value of the model is to assist but not replace surgeon experience and judgment.

4.4 Analysis of Prediction Accuracy for Vignettes with Additional Explanation of ML Model Prediction

The last part of the survey involved vignettes that were the same as presented earlier but now accompanied with an additional explanation about predictions made by the ML model (some of these predictions were correct and some were wrong). Fourteen participants completed the predictions for these five vignettes in phase 1, phase 2 (ML model support) and phase 2 with ML model support and explanation. The results are depicted in figure 9. The addition of the model prediction explanation improved accuracy for three participants and worsened accuracy for three participants compared to the phase 2 scenario without the model explanation. Overall, providing an explanation alongside the ML model prediction did not affect the prediction accuracy for the five vignettes at the completion of phase 2. However, given the limited number of vignettes with an explanation, general conclusions cannot be drawn. The overwhelming majority of participants did find the explanation useful and explainable AI (XAI) warrants further research.

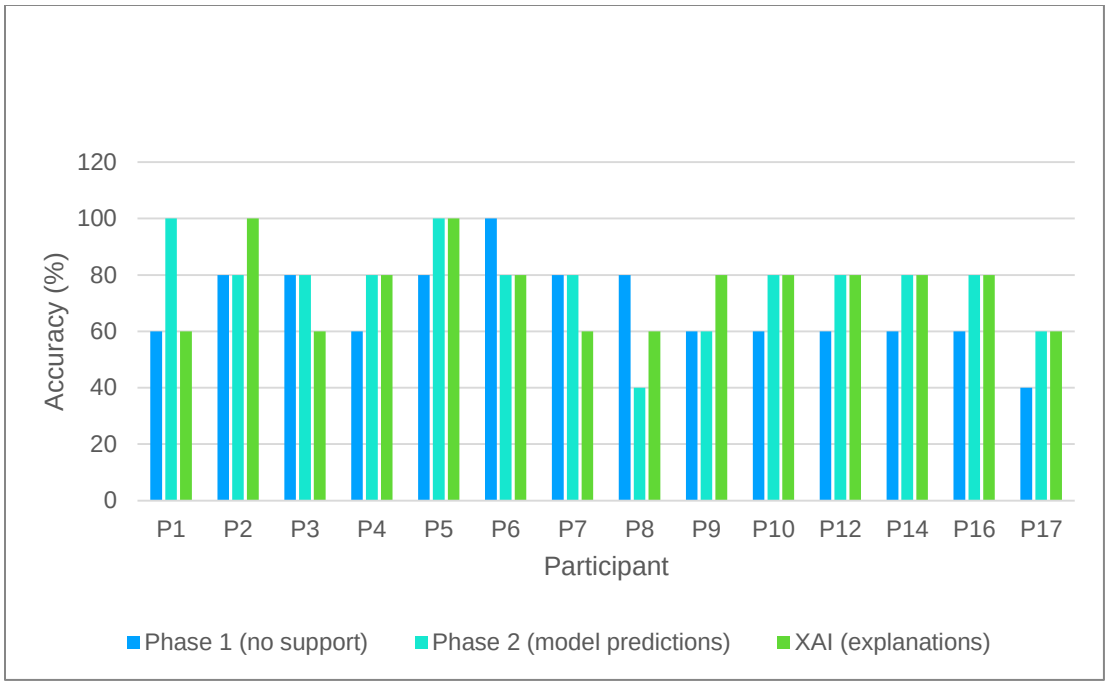


Figure 10: Comparison of surgeon prediction accuracy for 5 vignettes without model support, with model support and with model support and explanation

Twelve of 14 participants agreed that having an explanation of the prediction is useful. Ten participants thought that the way the explanation was presented was useful while four thought it should be presented differently. It is not surprising that most participants thought having an explanation was useful as the perception of prediction models as "black boxes" is certainly a potential barrier to uptake in the clinical setting.

5. Conclusions

Health care quality and safety have become a societal and health care system priority. Surgeons recognize the potential value of predictive analytics to improve care and ML has emerged as a popular approach to predictive model development. Research on the impact of predictive models is nominal and has certainly not kept pace with ML model development. This research aimed to fill the knowledge gap between predictive model development and model impact. The primary research question was "*How does the use of ML prediction models impact surgeons' predictions of post-surgical complications in patients undergoing spinal surgery?*" To answer this question preliminary work was performed to develop a ML prediction model for post-surgical complications in patients undergoing spinal surgery and the model performance was 0.596. Following this preliminary work, a survey instrument was created with 30 clinical vignettes and administered to a convenience sample of spinal surgeons and trainees. By surveying the participants without and with support provided by the ML prediction model, and separated by 2 weeks, the effect of the ML model on surgeons' predictions of post-operative complications was determined.

In this study, support provided by a ML prediction model improved surgeons' accuracy to correctly predict the presence or absence of a complication in patients undergoing spinal surgery from 49.1% to 54.8% ($p=0.024$). The prediction accuracy of surgeons in phase 2 was comparable but slightly greater than the ML model performance (0.53) for the 30 cases selected for the survey. The statistically significant improvement in prediction accuracy for the 14 participants using the ML model suggests that despite a model with mediocre prediction accuracy, the ML model can improve prediction performance in surgeons most likely because having some additional prediction information "nudges" surgeons to reflect about their own reasoning. This finding is important as a prediction model with an AUROC of 0.60 may be discounted as inadequate or potentially harmful and yet in this study, surgeons were outperformed by this model and benefitted from its support. Correct ML model predictions were most beneficial for participants and it is reassuring that although incorrect ML model predictions were not as helpful the overall effect of these predictions was relatively neutral. These findings are important as the study results may cause surgeons to reflect on their own outcome and

complication prediction abilities and make them more receptive to ML models. Predictive models are unlikely to replace surgeon judgement but can function to provide additional information and work in synergy. It is clear that predicting post-surgical complications in patients undergoing spinal surgery is difficult, for models and experienced surgeons, but it is not surprising that additional information provided by the ML model was beneficial overall. This is the first study in the spine surgery literature that has evaluated the impact of a ML prediction model on surgeon complication prediction accuracy.

In this research, the ML model may have been developed without using data that could have improved its performance but the surgeon was also predicting without knowing such features as direct patient contact, procedural considerations and imaging findings. As such, it is possible that model and surgeon complication prediction accuracy could improve with a more comprehensive set of features than those used in the study but the magnitude of improvement is uncertain. It remains that predicting complications after spinal surgery is difficult for ML models and experts.

From the detailed analysis of the vignettes, it becomes more evident as to why it is difficult to predict the presence or absence of complications in patients undergoing spinal surgery. Considering the 30 vignettes, four vignettes were nearly impossible to predict correctly in either phase and in six others, the performance of participants and the ML model were significantly different. The latter fact highlights the potentially numerous unexplained factors affecting participant and model performance. Despite these divergent predictions for the six vignettes it was reassuring that surgeon prediction accuracy decreased in only one of six vignettes after phase 2 and the magnitude was modest.

A secondary objective of this research was to provide a baseline reference for spinal surgeon complication prediction accuracy. In phase 1 of this study surgeons' unaided prediction accuracy was 0.49. In other words, correctly predicting the presence or absence of a post-surgical complication in this cohort of 30 vignettes was analogous to flipping a coin. This may come as a surprise to experienced surgeons but it certainly reflects the inherent difficulty for this prediction task. There may be other ways to frame the prediction (high, medium, low risk) that

would prove to be more accurate and, therefore, clinically useful. It is clear that surgeon prediction accuracy will be dependent on the prediction task and research on prediction models that are presented without surgeon prediction accuracy lack important contextual information critical for model evaluation. This study design may serve as a valuable tool for future studies on ML prediction models.

Surgeons may be hesitant to trust ML prediction models as they are perceived as a "black box". For example, a model may provide a prediction but this is done without an explanation. As such, ML model uptake, even if it provides clinically useful information, may be delayed. This study did not demonstrate an impact on prediction performance when the ML model support was provided with an explanation but the sample size was too small for definitive conclusions. The overwhelming majority of participants did find the explanation useful and XAI warrants further research.

5.1 Limitations

There are limitations associated with the study. The participants represent a convenience sample drawn from a single medical center. The generalizability of the study results may be limited. Furthermore, the depth of the survey, including the number of vignettes and associated questions, had to be balanced with the time required to complete the survey to ensure satisfactory response rate. This was particularly relevant for the final 5 vignettes of phase 2 that were developed to investigate the ML model accompanied by explanations. The value of ML model explanations needs to be studied in a more robust manner given the positive response by participants.

The clinical vignettes were developed from actual patient data at TOH. The vignettes included limited pertinent data, which was also used by the ML model, but did not include all patient data including imaging. A greater number and granularity of features could certainly improve both surgeon and model performance for a specific prediction task. It is possible that some surgeons were involved with the cases and as such would know the actual output but the descriptions were left sufficiently broad with respect to diagnosis and surgical procedure that surgeons were

unlikely to recall the cases with certainty. The prediction accuracy demonstrated in this study would suggest that knowledge of actual case outcome was not a significant issue.

The percentage of patients with a complication in the overall dataset is small. This class imbalance creates challenges in model development but also when surveying surgeons. To address this issue, the participants were informed that the set of 30 vignettes had a greater than expected prevalence of patients that had a post-surgical complication. Although this was felt to be necessary to avoid surgeons predicting "no complication" based on the known low complication rate it may have changed the surgeons' approach to the prediction task. Furthermore, predicting the "presence or absence" of a post-surgical complication may not be intuitive for surgeons and may be better framed in a more clinically appropriate way such as levels of risk.

Finally, there are complications that were not considered as they are not captured by TOH NSQIP. For example, dural tear and neurological injury are complications related to the procedure that are not necessarily captured unless there is an unplanned return visit to the operating room related to the procedure. While the types of included complications were explicitly stated in the survey, and were generally medical based on NSQIP data capture, it is possible that responding surgeons could have forgotten what they were over the course of analyzing the vignettes and completing the survey.

5.2. Contributions

This study has highlighted many critical considerations, particularly with respect to model features and prediction task, that are required to develop a clinically relevant ML prediction model. The results have reinforced the importance of database variable selection and quality as they pertain to predictive model feature granularity and accuracy. As prediction model development continues to expand there should be a renewed effort to ensure optimal database quality and granularity. Computer scientists, health informatics experts and physicians should explore ways in which models can produce an outcome or prediction of interest that is consistent with clinically relevant prediction tasks. In this study, the prediction of interest was

complication “yes or no” which was the logical prediction based on the database outcomes available yet this may not be the most clinically relevant prediction. Clinicians and patients may prefer ordinal scales of risk such as high, medium or low or absolute risk. Further research should be done to determine the most clinically relevant prediction tasks and ways to translate or optimize binary outcome data.

With the rapid expansion of AI and ML in health care and the potential to use large databases for predicting outcomes there is an understandable focus amongst spine surgeons and researchers to develop and publish models that can predict outcomes and complications. Importantly, the interest in model development has not been matched to this point with research on model impact and surgeons’ perceptions of predictive models. However, with published predictive models of varying prediction accuracies it is critically important to understand how surgeons will use or synthesize the output of the models in determining a patient’s risk with spinal surgery. This study has advanced the knowledge on how information provided by ML prediction models impact surgeons’ predictions regarding the risk of complications in patients undergoing spinal surgery. This is the first study in the spine surgery literature that has evaluated the impact of a ML prediction model on surgeon complication prediction accuracy. It is clear that surgeon prediction accuracy will be dependent on the prediction task and research on prediction models that are presented without surgeon prediction accuracy lack important contextual information critical for model evaluation. This study design may serve as a valuable tool for future studies on ML prediction models.

.

REFERENCES

1. Weinstein JN, Lurie JD, Olson PR, Bronner KK, Fisher ES. United States' trends and regional variations in lumbar spine surgery: 1992-2003. Spine. 2006;31(23):2707-14.
2. 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. Neurosurgical focus. 2015;39(6):E13.
3. Xu Y, Yen D, Whitehead M, Xu J, Johnson AP. Use of instrumented lumbar spinal surgery for degenerative conditions: trends and costs over time in Ontario, Canada. Can J Surg. 2019;62(6):393-401.
4. Fisher ES, McClellan MB, Safran DG. Building the path to accountable care. The New England journal of medicine. 2011;365(26):2445-7.
5. Beaulé PE, Roffey DM, Poitras S. Continuous quality improvement in orthopedic surgery: changes and implications with health system funding reform. Canadian Journal of Surgery. 2016;59(3):149-50.
6. Baker GR, Norton PG, Flintoft V, Blais R, Brown A, Cox J, et al. The Canadian Adverse Events Study: the incidence of adverse events among hospital patients in Canada. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2004;170(11):1678-86.
7. Forster AJ, Worthington JR, Hawken S, Bourke M, Rubens F, Shojania K, et al. Using prospective clinical surveillance to identify adverse events in hospital. BMJ quality & safety. 2011;20(9):756-63.
8. Dindo D, Demartines N, Clavien PA. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. Annals of surgery. 2004;240(2):205-13.
9. Street JT, Lenehan BJ, DiPaola CP, Boyd MD, Kwon BK, Paquette SJ, et al. Morbidity and mortality of major adult spinal surgery. A prospective cohort analysis of 942 consecutive patients. The spine journal : official journal of the North American Spine Society. 2012;12(1):22-34.
10. Montroy J, Breau RH, Cnossen S, Witiuk K, Binette A, Ferrier T, et al. Change in Adverse Events After Enrollment in the National Surgical Quality Improvement Program: A Systematic Review and Meta-Analysis. PloS one. 2016;11(1):e0146254.

11. Chen BP, Garland K, Roffey DM, Poitras S, Dervin G, Lapner P, et al. Can Surgeons Adequately Capture Adverse Events Using the Spinal Adverse Events Severity System (SAVES) and OrthoSAVES? Clinical orthopaedics and related research. 2017;475(1):253-60.
12. Sebastian AS, Polites SF, Glasgow AE, Habermann EB, Cima RR, Kakar S. Current Quality Measurement Tools Are Insufficient to Assess Complications in Orthopedic Surgery. The Journal of hand surgery. 2017;42(1):10-5.e1.
13. Schoenfeld AJ, Carey PA, Cleveland AW, 3rd, Bader JO, Bono CM. Patient factors, comorbidities, and surgical characteristics that increase mortality and complication risk after spinal arthrodesis: a prognostic study based on 5,887 patients. The spine journal : official journal of the North American Spine Society. 2013;13(10):1171-9.
14. Galbusera F, Casaroli G, Bassani T. Artificial intelligence and machine learning in spine research. JOR Spine. 2019;2(1):e1044-e.
15. Alluri R, Kang HP, Bouz G, Wang J, Hah RJ. The True Effect of a Lumbar Dural Tear on Complications and Cost. Spine. 2020;45(3):E155-e62.
16. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart (British Cardiac Society). 2012;98(9):691-8.
17. Glasgow RE, Hawn MT, Hosokawa PW, Henderson WG, Min SJ, Richman JS, et al. Comparison of prospective risk estimates for postoperative complications: human vs computer model. Journal of the American College of Surgeons. 2014;218(2):237-45.e1-4.
18. Norton WE, Hosokawa PW, Henderson WG, Volckmann ET, Pell J, Tomeh MG, et al. Acceptability of the decision support for safer surgery tool. American journal of surgery. 2015;209(6):977-84.
19. Lee MJ, Cizik AM, Hamilton D, Chapman JR. Predicting medical complications after spine surgery: a validated model using a prospective surgical registry. The spine journal : official journal of the North American Spine Society. 2014;14(2):291-9.
20. Veeravagu A, Li A, Swinney C, Tian L, Moraff A, Azad TD, et al. Predicting complication risk in spine surgery: a prospective analysis of a novel risk assessment tool. Journal of neurosurgery Spine. 2017;27(1):81-91.
21. Wang G, Lam KM, Deng Z, Choi KS. Prediction of mortality after radical cystectomy for bladder cancer by machine learning techniques. Computers in biology and medicine. 2015;63:124-32.

22. Kim JS, Merrill RK, Arvind V, Kaji D, Pasik SD, Nwachukwu CC, et al. Examining the Ability of Artificial Neural Networks Machine Learning Models to Accurately Predict Complications Following Posterior Lumbar Spine Fusion. Spine. 2017.
23. Senders JT, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman MLD, et al. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. World neurosurgery. 2018;109:476-86.e1.
24. 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. The spine journal : official journal of the North American Spine Society. 2014;14(7):1247-55.
25. Broda A, Sanford Z, Turcotte J, Patton C. Development of a Risk Prediction Model with Improved Clinical Utility in Elective Cervical and Lumbar Spine Surgery. Spine. 2019.
26. Wang MC. Calculating risks: the power and pitfalls of registry data. The spine journal : official journal of the North American Spine Society. 2013;13(10):1180-2.
27. Marjoua Y, Xiao R, Waites C, Yang BW, Harris MB, Schoenfeld AJ. A systematic review of spinal research conducted using the National Surgical Quality Improvement Program. The spine journal : official journal of the North American Spine Society. 2017;17(1):88-95.
28. Buerba RA, Fu MC, Gruskay JA, Long WD, 3rd, Grauer JN. Obese Class III patients at significantly greater risk of multiple complications after lumbar surgery: an analysis of 10,387 patients in the ACS NSQIP database. The spine journal : official journal of the North American Spine Society. 2014;14(9):2008-18.
29. Paulino Pereira NR, Janssen SJ, van Dijk E, Harris MB, Hornicek FJ, Ferrone ML, et al. Development of a Prognostic Survival Algorithm for Patients with Metastatic Spine Disease. The Journal of bone and joint surgery American volume. 2016;98(21):1767-76.
30. Sebastian A, Goyal A, Alvi MA, Wahood W, Elminawy M, Habermann EB, 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 neurosurgery. 2019;126:e323-e9.
31. Ratliff JK, Balise R, Veeravagu A, Cole TS, Cheng I, Olshen RA, et al. Predicting Occurrence of Spine Surgery Complications Using "Big Data" Modeling of an Administrative Claims Database. The Journal of bone and joint surgery American volume. 2016;98(10):824-34.
32. Mirza SK, Deyo RA, Heagerty PJ, Konodi MA, Lee LA, Turner JA, 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; discussion 62.

33. Durand WM, DePasse JM, Daniels AH. Predictive Modeling for Blood Transfusion Following Adult Spinal Deformity Surgery: A Tree-Based Machine Learning Approach. Spine. 2017.
34. Scheer JK, Smith JS, Schwab F, Lafage V, Shaffrey CI, Bess S, et al. Development of a preoperative predictive model for major complications following adult spinal deformity surgery. Journal of neurosurgery Spine. 2017;26(6):736-43.
35. Scheer JK, Osorio JA, Smith JS, Schwab F, Lafage V, Hart RA, et al. Development of Validated Computer-based Preoperative Predictive Model for Proximal Junction Failure (PJF) or Clinically Significant PJK With 86% Accuracy Based on 510 ASD Patients With 2-year Follow-up. Spine. 2016;41(22):E1328-e35.
36. Arvind V, Kim JS, Oermann EK, Kaji D, Cho SK. Predicting Surgical Complications in Adult Patients Undergoing Anterior Cervical Discectomy and Fusion Using Machine Learning. Neurospine. 2018;15(4):329-37.
37. Han SS, Azad TD, Suarez PA, Ratliff JK. A machine learning approach for predictive models of adverse events following spine surgery. The spine journal : official journal of the North American Spine Society. 2019;19(11):1772-81.
38. Sinuff T, Adhikari NK, Cook DJ, Schunemann HJ, Griffith LE, Rocker G, et al. Mortality predictions in the intensive care unit: comparing physicians with scoring systems. Critical care medicine. 2006;34(3):878-85.
39. Flechet M, Falini S, Bonetti C, Guiza F, Schetz M, Van den Berghe G, et al. Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: a prospective evaluation of the AKIpredictor. Critical care (London, England). 2019;23(1):282.
40. Dilaver NM, Gwilym BL, Preece R, Twine CP, Bosanquet DC. Systematic review and narrative synthesis of surgeons' perception of postoperative outcomes and risk. BJS open. 2020;4(1):16-26.
41. Samim M, Mungroop TH, AbuHilal M, Isfordink CJ, Molenaar QI, van der Poel MJ, et al. Surgeons' assessment versus risk models for predicting complications of hepato-pancreato-biliary surgery (HPB-RISC): a multicenter prospective cohort study. HPB : the official journal of the International Hepato Pancreato Biliary Association. 2018;20(9):809-14.
42. Pelaccia T, Forestier G, Wemmert C. Deconstructing the diagnostic reasoning of human versus artificial intelligence. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2019;191(48):E1332-e5.

43. Brennan M, Puri S, Ozrazgat-Baslanti T, Feng Z, Ruppert M, Hashemighouchani H, et al. Comparing clinical judgment with the MySurgeryRisk algorithm for preoperative risk assessment: A pilot usability study. Surgery. 2019;165(5):1035-45.
44. Phan P, Mezghani N, Nault ML, Aubin CE, Parent S, de Guise J, et al. A decision tree can increase accuracy when assessing curve types according to Lenke classification of adolescent idiopathic scoliosis. Spine. 2010;35(10):1054-9.
45. Fleiss JL. Statistical Methods for Rates and Proportions.
46. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ open. 2016;6(11):e012799.
47. Goto T, Camargo CA, Jr., Faridi MK, Freishtat RJ, Hasegawa K. Machine Learning-Based Prediction of Clinical Outcomes for Children During Emergency Department Triage. JAMA Netw Open. 2019;2(1):e186937-e.
48. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O, written on behalf of AMEB-DCTCG. Predictive analytics with gradient boosting in clinical medicine. Annals of translational medicine. 2019;7(7):152-.
49. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. biometrics Biometrics. 1977;33(1):159-74.

Appendix A

Phase 1: Study and survey description

TELFER

VOTRE LIEN AVEC CE QUI COMPTE — CONNECTS YOU TO WHAT MATTERS

Phase 1: Study and Survey Description

The purpose of this survey is to determine how spinal surgeons, of different levels of training and experience, use information provided by a prediction model for surgical decision-making.

The survey is divided into two phases. In Phase 1 you will be asked to provide your prediction as to the presence or absence of a post- operative complication in patients undergoing spinal surgery. Phase 2 will be administered at a later date. During Phase 2 you will be asked to provide your own predictions when supported with information derived from a prediction model that was developed from a spine surgical database using a machine learning method (artificial intelligence or AI method).

The survey responses will be automatically saved and it does not have to be completed in a single session. However, the same computer needs to be used for each session to ensure the survey responses are restored.

→

Background

- Decision making in spinal surgery requires an understanding of a proposed procedure's complications and benefits by both the patient and surgeon.
- Surgical risk prediction tools utilize known covariates to predict the risk of a complication over a specific time period.
- For the purpose of this survey, a complication is defined as any of the following:
  - *Surgical*: surgical site infection (superficial, deep, or wound dehiscence), or unplanned return to the operating room related to the index spinal procedure.
  - *Medical*: pneumonia, pulmonary embolism, renal insufficiency, renal failure, urinary tract infection, stroke, deep vein thrombosis, sepsis or septic shock, or cardiac arrest.

Instructions

- The following survey consists of 30 clinical vignettes and will take approximately 15 minutes to complete.
- For each vignette you will be asked to predict the absence or presence of a complication within the first 30 days following surgery.
- The vignettes have a greater prevalence of complications than the original database. This has been done to ensure that you will predict the presence or absence of a complication based on the specific features of the vignette as opposed to the expected or known complication rate in a population.

Appendix B

Phase 1: Sample survey vignette

Vignette 1

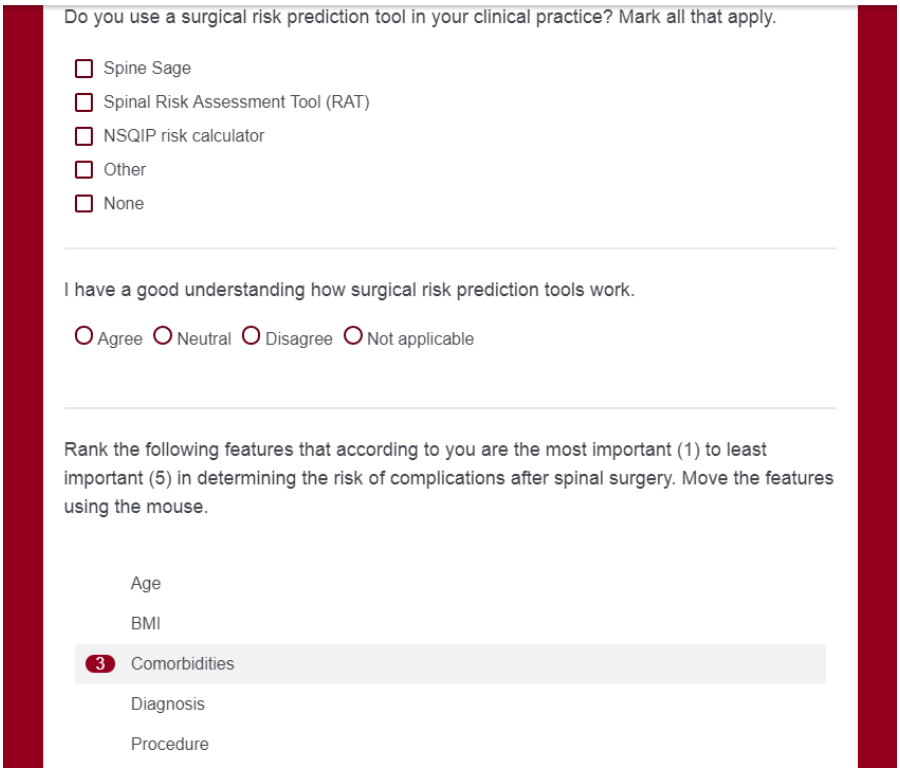
59 year old male undergoing elective thoracolumbar decompression and instrumented fusion/TLIF for flatback. He has a BMI of 25 and past medical history includes diabetes mellitus and hypertension.

Please select your own prediction of a complication

- ○ Absent
- ○ Present

Appendix C

Questions pertaining to surgeons' use of complication prediction models



Appendix D

Phase 2: Study and survey description

Phase 2: Study and Survey Description

The purpose of this phase of the survey is to determine how spinal surgeons, of different levels of training and experience, use information provided by a prediction model for surgical decision-making. The prediction model was developed from local spine surgical data using machine learning methods (AI methods) to predict the absence or presence of complications in patients undergoing spinal surgery.

Explainable AI will be explored to determine how a surgeon's confidence in the model's prediction is impacted when they are provided with additional information about specific features and their values that were used by the model to predict surgical risk.

The prediction model's sensitivity, specificity and area under the receiver operating curve (AUROC) will be provided.

The survey responses will be automatically saved and it does not have to be completed in a single session. However, the same computer needs to be used for each session to ensure the survey responses are restored.

Background

- Decision making in spinal surgery requires an understanding of a proposed procedure's complications and benefits by both the patient and surgeon.
- Surgical risk prediction tools utilize known covariates to predict the risk of a complication over a specific time period.
- For the purpose of this survey, a complication is defined as any of the following:
  - *Surgical*: surgical site infection (superficial, deep, or wound dehiscence), or unplanned return to the operating room related to the index spinal procedure.
  - *Medical*: pneumonia, pulmonary embolism, renal insufficiency, renal failure, urinary tract infection, stroke, deep vein thrombosis, sepsis or septic shock, or cardiac arrest.
- We developed a prediction model using the Ottawa Hospital NSQIP spinal surgery database and machine learning algorithms that belong to a broader class of AI methods. The model does not predict a percentage risk of a complication but rather the presence or absence of a complication in an individual patient. The model's AUROC is 0.63 and the sensitivity and specificity are 0.66 and 0.53, respectively.

Appendix E

Phase 2: Sample survey vignette

Vignette 1

59 year old male undergoing elective thoracolumbar decompression and instrumented fusion/TLIF for flatback. He has a BMI of 25 and past medical history includes diabetes mellitus and hypertension.

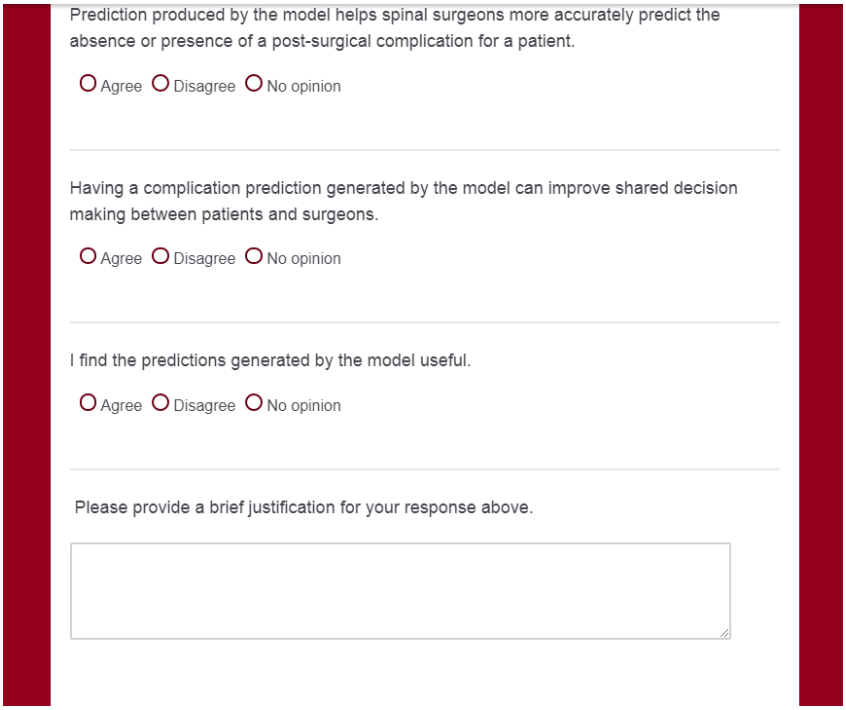
Model prediction of a complication: **absent**

Please select your own prediction of a complication

- ○ Absent
- ○ Present

Appendix F

Surgeons' perceptions of ML prediction model after Phase 2



Appendix G

Phase 2: Introduction to vignettes with model prediction and explanation

Explainable Artificial Intelligence

A possible barrier to the clinical uptake of prediction models, developed using machine learning, is their lack of transparency or explainability. To the clinician the model may appear as a "black box" with respect to the primary features, and their values, that most significantly influenced the model's prediction.

Instructions

The next 5 clinical vignettes will be provided with automatically generated rules that explain the complication predictions that were established for specific patients (i.e., explainable artificial intelligence).

Appendix H

Phase 2: Sample survey vignette with ML model explanation

Vignette 1

48 year old female undergoing unplanned (priority F) anterior cervical discectomy and fusion for myelopathy. Her BMI is 35 and she has hypertension.

Model prediction of a complication: **present**

Explanation of model prediction: **surgery is urgent and patient's BMI is above 29.6**

Please select your own prediction of a complication

- ○ Absent
- ○ Present

Appendix I

Surgeons' perceptions of ML model explanation

I find having an explanation of the prediction result useful.

○ Agree ○ Disagree ○ No opinion

Please provide a brief justification for your response above.

Is the model prediction explanation presented in an understandable and useful way or should it be presented otherwise?

○ Yes, it should be presented as is ○ No, it should be presented otherwise

If you think the model prediction should be presented otherwise, then please give an example.