Journal of Clinical Neuroscience 43 (2017) 247–255

Contents lists available at ScienceDirect

Journal of Clinical Neuroscience

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

Tools and techniques

The Seattle spine score: Predicting 30-day complication risk in adult spinal deformity surgery

Quinlan D. Buchlak [a,*], Vijay Yanamadala [a], Jean-Christophe Leveque [a], Alicia Edwards [a], Kellen Nold [a], Rajiv Sethi [a,b]

[a] Neuroscience Institute, Virginia Mason Medical Center, Seattle, WA, USA
[b] Department of Health Services, University of Washington, Seattle, WA, USA

ARTICLE INFO

Article history:
Received 5 February 2017
Accepted 8 June 2017

Keywords:
Risk stratification
Surgical complications
Spine surgery
Predictive clinical decision support

ABSTRACT

Background: Complication rates in complex spine surgery range from 25% to 80% in published studies. Numerous studies have shown that surgeons are not able to accurately predict whether patients are likely to face post-operative complications, in part due to biases based on individual experience. The purpose of this study was to develop and evaluate a predictive risk model and decision support system that could accurately predict the likelihood of 30-day postoperative complications in complex spine surgery based on routinely measured preoperative variables.

Methods: Preoperative and postoperative data were collected for 136 patients by reviewing medical records. Logistic regression analysis (LRA) was applied to develop the predictive algorithm based on patient demographic parameters, including age, gender, and co-morbidities, including obesity, diabetes, hypertension and anemia. We additionally compared the performance of the predictive model to a spine surgeon's ability to predict patient complications using signal detection theory statistics representing sensitivity and response bias (A′ and B″ respectively). We developed a decision support system tool, based on the LRA predictive algorithm, that was able to provide a numeric probabilistic likelihood statistic representing an individual patient's risk of developing a complication within the first 30 days after surgery.

Results: The predictive model was significant ($\chi^2 = 16.242$, $p < 0.05$), showed good fit, and was calibrated by using area under the receiver operating characteristics curve analysis (AUROC = 0.712, $p < 0.01$). The model yielded a predictive accuracy of 75.0%. It was validated by splitting the data set, comparing subset models, and testing them with unknown data. Validation also involved comparing the classification of cases by experts with the classification of cases by the model. The model significantly improved the classification accuracy of physicians involved in the delivery of complex spine surgical care.

Conclusions: The application of technology and data-driven tools to advanced surgical practice has the potential to improve decision making quality, service quality and patient safety.

1. Introduction

The population of patients with spinal deformity requiring surgical treatment is growing [1,2]. With the move towards value-based care, surgical care for these patients is being rewarded for higher quality with controlled cost [2–4]. Efforts to improve the accuracy of surgical decision-making and to develop data-driven risk stratification methods are likely to improve patient safety and outcomes, and thereby increase the overall quality and value of spine surgery care.

Complex spine surgery, defined as a procedure involving six or more levels of spinal fusion, is a high-risk undertaking with high complication rates. Complication rates range from 10 percent up to 80 percent [5–12], and are often associated with increased hospital stay, cost and long-term morbidity [9,10,12]. These complications occur as a result of a complex web of social, physiological and environmental factors [13].

Preoperative assessment of complication risk in complex spine surgery is often based on broad prevalence rates and retrospective percentage statistics. The development of debiasing strategies in high-risk medical decision making has the potential to increase service quality and patient safety. Debiasing involves moving away from intuitive processing towards processing that is more analytical, evidence-based and system-supported [14]. Robust predictive

* Corresponding author at: Virginia Mason Medical Center Neuroscience Institute, 1201 Terry Ave, Seattle, WA 98101, USA.
E-mail address: quinlanbuchlak@gmail.com (Q.D. Buchlak).

http://dx.doi.org/10.1016/j.jocn.2017.06.012
0967-5868/

models are one method to improve risk assessment and achieve gains in service quality. While work on predictive modeling in spine surgery is progressing [15–20], the application of data-driven methods for accurately and reliably predicting surgical risk and patient complications in spine surgery is rare.

The purpose of this study was to generate and calibrate a statistical model to predict the risk of 30-day complications associated with complex spine surgery. The utility of the model was maximized by focusing on preoperative variables that were readily available and easily measurable. We hypothesized that a statistical model developed using preoperative patient characteristics would accurately predict the likelihood of 30-day complications. We used the predictive model to develop a decision support system (DSS) with a quantified output representing the risk of complication within 30 days of surgery. We performed an evaluation experiment to assess the utility of this statistical model-driven DSS in helping physicians involved in the delivery of complex spine surgical care to identify patients who were at higher or lower risk of postoperative complications. We hypothesized that the additional information provided by the DSS would increase the capability of physicians to accurately predict whether patients would or would not go on to experience postoperative complications.

2. Method

2.1. Predictive modeling and DSS development

2.1.1. Participants and data collection

This retrospective predictive modeling study included a total of 136 consecutive spine deformity patients. Inclusion criteria were as follows. Patients were (1) at least 18 years of age, (2) diagnosed with adult spinal deformity with a coronal lumbar or thoracic curve greater than 40 degrees and/or significant sagittal plane imbalance with SVA greater than 10 cm and LL-PI mismatch of 20 degrees or greater, and (3) treated with a spinal fusion procedure involving six or more vertebral levels. All patients underwent a complex spine procedure involving a posterior approach. A subset of cases had a secondary minimally invasive lateral approach for anterior fusion. Surgeries were performed at a single high-volume institution in the United States with a large multistate referral pattern for adult spinal deformity cases. Data was collated for all cases based on queries of the institution's data warehouse and abstraction of electronic medical records (EMR). Abstraction was conducted by two trained abstractors. A random sample of 15% of cases was selected to assess the accuracy of chart abstraction. Inter-rater data concordance was 100%.

2.1.2. Predictive model development

The primary outcome of interest was the occurrence of a complication event within 30 days of surgery. A complication event was defined as a patient experiencing one or more of the following: cardiac event including myocardial infarction, pneumothorax, pneumonia, wound infection, wound dehiscence, urinary tract infection, pulmonary embolism, thromboembolism, unplanned return to surgery and death. The presence of any complication was coded "1" and absence was coded "0". Multiple complications were not additive.

Univariate and multivariate logistic regression analyses (LRA) were conducted to predict the probability of a postoperative complication event. The odds ratio of each risk factor was indicated. A set of theoretically and clinically relevant potential predictive variables was devised based upon the expertise and recommendations of five senior surgeons. Potential predictive variables needed to be captured adequately within the EMR for inclusion in this study. Potential predictors included age, gender, BMI, a history of smoking, and a preoperative diagnosis of hypertension, anxiety, depression, diabetes, bipolar disorder, Parkinson's disease, cancer, and anemia.

Summary statistics were calculated, including frequency and percentage statistics for categorical variables and means and standard deviations for continuous variables. In assessing the magnitude of associations, we calculated odds ratios and 95% confidence intervals. For the multivariate LRA, we included variables that (1) were clinically relevant or (2) achieved a univariate significance level of 0.2 or less, in line with the methods of other predictive modeling researchers [21].

To achieve sufficient power in multivariate LRA, the model must be based on a sample size that is at least 10 times the number of predictors [22]. In this case, the sample size was sufficient to substantially exceed this minimum benchmark. The final model contained seven predictor variables (BMI, age, gender, smoking status, and a preoperative diagnosis of diabetes, hypertension, or anemia) and was based on 136 cases.

Multivariate LRA models were considered significant if they achieved a p value less than 0.05. To calibrate the models and establish an indicator of their performance, discrimination between high- and low-risk patients was assessed using the area under the receiver operating curve (AUROC). The model was developed in line with predictive model development guidelines [13,23–25]. Analyses were conducted using SPSS (SPSS, Chicago, IL).

Three multivariate models were developed and their relative quality was assessed. Model calibration measures how closely actual outcomes align with those predicted by the model. Calibration was measured using the Hosmer-Lemeshow Chi-square statistic [21,26]. Model quality was assessed by reviewing the (1) model's chi-square statistic, (2) percentage of correct predictions, and (3) Nagelkerke's pseudo-R^2. The model that demonstrated the best fit and the highest percentage of correct predictions was selected for subsequent validation and experimental evaluation.

To validate the model, we divided our dataset into five distinct sets. In line with the process articulated by Assman, Cullen & Schulte (2002) [21], combinations of four of these five sets were used for generating the model and training the algorithms. The final set was used for testing the performance of the models on unknown data. This validation process was conducted for every possible 4-part, 1-part combination [21]. This internal validation process showed that the performance of the model was robust. Results in each of the subsets did not differ substantially from the model derived from the full data set.

A predictive algorithm was developed using the beta coefficients and the constant of the model based on the full dataset.

2.1.3. Decision support system development

A DSS was developed to enable the application of the predictive algorithm created. This DSS is an interactive system that applies the exponentiated regression equation, weighting each predictive variable independently. The algorithm was mathematically converted to yield a quantified probability score [24]. The DSS allows for calculation of risk in an individual patient by inputting the value for each of the seven predictor variables. The output of the DSS is a single global percentage statistic, which suggests the likelihood of complications occurring within 30 days for each individual case. Fig. 1 shows the design of the DSS dashboard that was developed. This design adheres to DSS-development guidelines [27,28].

3. Experimental evaluation

3.1. Aims and hypotheses

An experiment was conducted to assess whether the output of the DSS improved the predictive accuracy of expert physicians

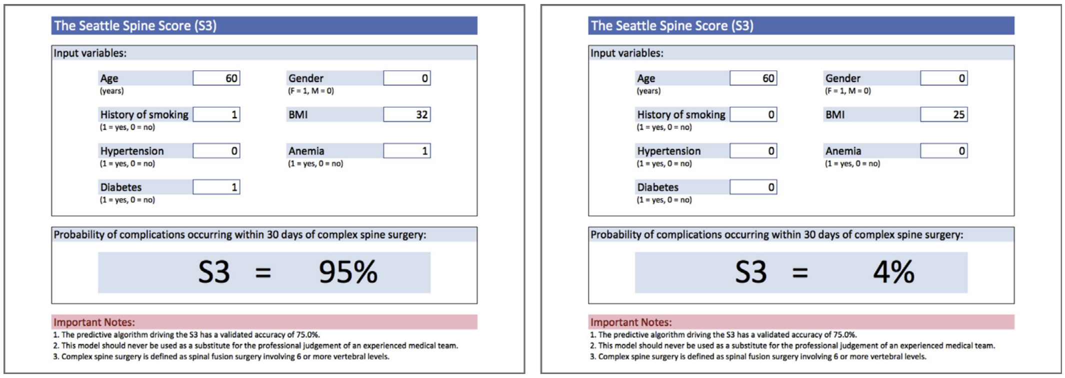

Fig. 1. Two examples of the DSS interface. One displaying an example of a hypothetical high-risk test case. The other displaying an example of a hypothetical low-risk test case.

involved in the delivery of complex spine care. The purpose of this study was to determine the effect of the DSS output (the risk metric) on the ability of physicians to accurately decide whether patients would or would not experience surgical complications as a result of complex spine surgery. The experiment aimed to contribute to building an understanding of whether the output of predictive risk calculators could assist in improving surgical decision making. It involved collecting and analyzing decision-making data from a sample of senior physicians directly involved in the delivery of complex spine surgical care. Hypotheses guiding the design and implementation of this experimental evaluation study were as follows.

1. Physicians will be more able to correctly identify whether or not patients will go on to experience postoperative complications when they are presented with preoperative patient information along with the probabilistic risk metric than when they are presented with just preoperative patient information alone.
2. The predictions of the model-driven DSS will be more accurate than the predictions of expert physicians when they do not have access to the risk metric.

3.2. Participants

Eight senior physicians involved in the delivery of complex spine surgical care participated in this study. Participants included orthopaedic surgeons, neurosurgeons, anaesthesiologists, and physiatrists.

3.3. Design

A within-subjects experiment was conducted to gather data on how the risk model affected the quality of physician decision making. De-identified data was used to create a list of 100 random cases. Data for each case included the seven relevant DSS model input variables. The list consisted of data from 26 surgical cases that experienced postoperative complications within 30 days of surgery and 74 surgical cases that did not experience complications. This list was used in three test conditions. The experimenter sat with participants to mitigate the risk of them using additional resources to aid in the decision making process.

3.3.1. Condition X: Surgeon only

In this condition, each physician was asked to predict whether each case would result in complicated outcomes based only on the seven clinical variables presented. Their prediction for each case was recorded as a yes or a no response.

3.3.2. Condition Y: Surgeon and DSS

In this condition, the list of cases presented to physicians contained the DSS risk metric for each case. Physicians were informed of the accuracy of the risk metric. For each case, physicians were asked to decide whether or not postoperative complications were likely. Again, their prediction for each case was recorded as a yes or a no response.

3.3.3. Model only condition

The list was analyzed using the DSS. The resulting risk metric was recorded for each case. A DSS probability estimate greater than 0.5 suggested that complications would occur. An estimate less than 0.5 suggested that complications would not occur.

The presentation of stimulus conditions to participants was balanced to control for practice and memory effects. Half of the participant pool was presented with condition X first, followed by condition Y. The other half of the participant pool was presented with condition Y first, followed by condition X. Assignment to these conditions was random. The participant pool provided data for a total of 1600 trials (800 trials in the X-Y condition and 800 trials in the Y-X condition). A power analysis demonstrated that this sample size was sufficient to allow for significance testing.

3.4. Analysis

Classification performance was compared to observed patient outcomes. Statistical significance testing was conducted to identify performance differences between conditions. A-prime (A') statistics were calculated to assess group discrimination sensitivity. An A' of one indicates perfect performance. This means that participants are able to discriminate between the two patient groups (complications vs. no-complications) accurately every time. An A' close to zero indicates that participants are not able to distinguish the signal from the noise. As predictive accuracy improves, A' moves closer to one. Response bias was measured using B''. A B'' of negative one indicates an extreme bias in favour of yes responses. A B'' of zero indicates no bias. A B'' of positive one indicates an extreme bias in favour of no responses [27].

Retrospective evaluation was conducted in accordance with Pick (2008). Retrospective evaluation involved comparing the statistical model predictions with actual outcomes in the retrospective dataset. Each case was assessed by the DSS and a risk classification determination was made to assess the accuracy of the predictive model on the set of cases presented.

3.5. Ethics review and approval

Ethics approval for this predictive modeling and experimental evaluation study was granted by the institution's IRB (IRB file number: IRB15133).

4. Results

The mean age of patients was 63.2 years (range 20.0–85.1, SD = 11.2). Mean BMI was 28.5 (range 17.1–47.0, SD = 6.1). Most patients (73.5%) were female, 46.3% had a history of smoking, 55.1% had hypertension, 8.1% had diabetes, and 3.7% had preoperative anemia. Complications occurring within 30 days of surgery were evident in 25.7% of cases.

4.1. Predictive modeling

Univariate LRA indicated that age, BMI, gender, smoking status and preoperative diagnoses of anemia, diabetes and hypertension were valid predictor variables to be included in the multivariate model. Univariate LRA resulted in the exclusion of the following variables from the multivariate model: Preoperative diagnoses of depression, anxiety, bipolar, Parkinson's disease and cancer.

The multivariate LRA model was significant ($\chi^2 = 16.242$, $p < 0.05$) and demonstrated a predictive accuracy of 75% (Table 1). The ability of our model to discriminate between those who experienced complication from those who did not was measured using area under the receiver-operating characteristics (AUROC) curve analysis, with an AUROC curve statistic of 1.0 indicating perfect discrimination and 0.5 representing chance. The AUROC statistic obtained by means of the model algorithm was 0.712 ($p < 0.01$), indicating a good level of discriminative functionality (Fig. 2). The risk estimates generated by our model showed very good agreement with the observed incidence of complications (Hosmer-Lemeshow $\chi^2 = 3.692$, $p = 0.884$; p should be greater than 0.05), further demonstrating the ability of the model to discriminate between cases that did and did not go on to experience complications. A classification plot is presented in Fig. 3. This provides detailed insight into how well the predictive model classified complicated (1) and uncomplicated (0) cases.

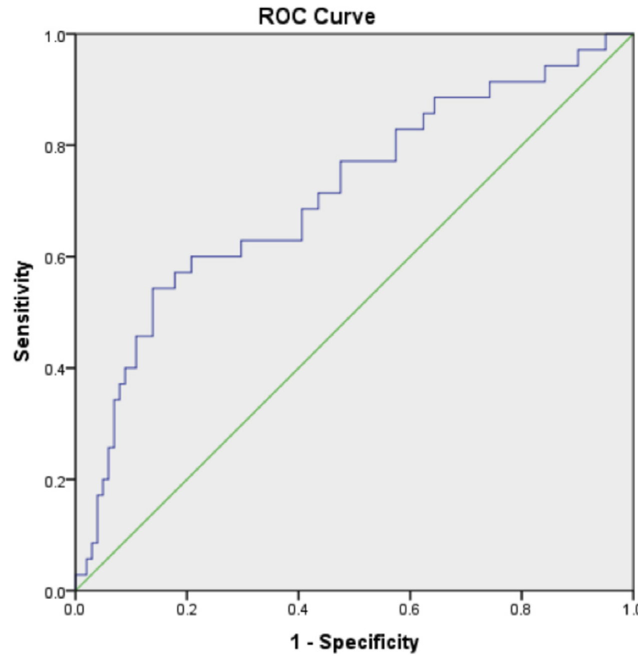

Fig. 2. The receiver operating characteristics curve for the multivariate logistic regression model.

5. Experimental evaluation

5.1. Retrospective evaluation

When the 100 cases were analyzed by the DSS, it demonstrated a 76% accuracy rate, which was significantly better than chance (50%; $\chi^2 = 21.825$, $p < 0.01$). A 0.5 threshold was used as the DSS decision making criterion. The DSS was significantly more accurate than participants who were exposed only to the case list that included patient data alone, and were not exposed to the risk metric at all ($\chi^2 = 21.825$, $p < 0.01$). Participants demonstrated an accuracy of 50%.

5.2. The X-Y Condition: Blinded cases (X), then cases with the risk metric (Y)

The four participants in this condition were first presented with the list of 100 cases that did not include the DSS risk metric. These participants were subsequently presented with the list of 100 cases that included patient data along with the DSS risk metric. Table 2 presents the results of this test condition.

When participants were first presented with the list of cases that included just preoperative patient data, their predictive accuracy was equal to chance (50.00%). When these participants were then presented with the list of cases that included the risk metric,

Table 1
Logistic regression analysis statistics for the predictive model developed and used to drive the DSS.

Variable	Predictive Model				
	Coefficient	p	OR	95% OR CI	
				Lower limit	Upper limit
Constant	−5.164	0.004			
Smoking	0.183	0.670	1.200	0.518	2.782
BMI (kg/m^2)	0.111	0.001	1.118	1.044	1.197
Diabetes	0.905	0.210	2.471	0.601	10.153
Age	0.011	0.594	1.012	0.970	1.055
Sex	−0.431	0.386	0.650	0.245	1.720
Hypertension	0.010	0.983	1.010	0.407	2.503
Anemia	0.240	0.815	1.272	0.170	9.522
Model Chi-square (df, p)				16.242 (7, p = 0.023)	
% Correct Predictions					75.00
Nagelkerke R Square					0.165

Note: OR = odds ratio; CI = confidence interval.

```
              Observed Groups and Predicted Probabilities

       16 +                                                                                                   +
          I                                                                                                   I
          I                                                                                                   I
F         I             1                                                                                     I
R      12 +             0                                                                                     +
E         I             0                                                                                     I
Q         I             0  1                                                                                  I
U         I             0  1                                                                                  I
E       8 +             0  0                                                                                  +
N         I            10 10                                                                                  I
C         I            00 001                                                                                 I
Y         I          1 00 001                1          1                                                     I
        4 +        0 00001000   10   0       0          1                                                     +
          I        0100000000 1010 0 0 0     0   0      1                                                     I
          I        0000000000 0000 010001 0  0   01     1 1     1  1   10                                     I
          I       0000000000000000000000 0  101101  0010001  0   01 101 00 1                                  I
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
    Prob: 0        .1        .2        .3        .4        .5        .6        .7        .8        .9         1
   Group: 000000000000000000000000000000000000000000000000001111111111111111111111111111111111111111111111111111

          Predicted Probability is of Membership for 1.00
          The Cut Value is .50
          Symbols: 0 - Postoperative complication/s not experienced
                   1 - Postoperative complication/s experienced
          Each Symbol Represents 1 Case.
```

Fig. 3. Classification plot showing how the model classified cases that did and did not go on to experience postoperative complications.

Table 2
Predictive accuracy statistics for participants who were first presented with case data alone and then were presented with case data accompanied by the risk metric.

	Without risk metric – X(this list presented first)		With risk metric – Y (this list presented second)		p value*
	n	Proportion	n	Proportion	
Correct	200	0.5000	243	0.6075	0.0022
Incorrect	200	0.5000	157	0.3925	0.0022
TrueNegative	143	0.3575	193	0.4825	0.0003
TruePositive	57	0.1425	44	0.1100	0.1667
FalseNegative	47	0.1175	60	0.1500	0.1772
FalsePositive	153	0.3825	103	0.2575	0.0002
Hit rate		0.1425		0.1100	0.1667
False alarm rate		0.3825		0.2575	0.0002
Sensitivity (A′)		−0.0182		0.1151	0.0001
Bias (B″)		0.3181		0.3227	0.8892

* *Notes:* p values test differences in proportion statistics. The standard statistical significance threshold is used here (0.05).

their predictive accuracy improved significantly to 60.75% ($\chi^2 = 9.341$, $p < 0.01$).

The proportion of true negatives for these participants increased significantly between stimulus conditions. When presented with the list without the risk metric, the proportion of true negatives was 0.3575. When participants were subsequently presented with the list that included the risk metric, the proportion of true negatives increased significantly to 0.4825 ($\chi^2 = 12.812$, $p < 0.05$). This indicated an improved ability to accurately identify the patients who did not go on to experience postoperative complications. The proportion of true positives and false negatives did not change significantly in this condition. When participants were first presented with the list of cases without the risk metric, the proportion of false positives was 0.3825. This reduced significantly when participants were presented with the cases accompanied by the risk metric (0.2575; $\chi^2 = 13.343$, $p < 0.05$). A false positive occurs when a participant predicts that a patient will experience a postoperative complication and the patient does not. The risk metric significantly reduced this type of prediction error.

Participants presented with just patient data first were not able to discriminate between cases that were likely to experience postoperative complications and those that were not ($A' = -0.0182$). When these participants were then presented with the list of cases that included the risk metric, their ability to discriminate between complicated and uncomplicated cases improved significantly ($A' = 0.1151$; $\chi^2 = 30.150$, $p < 0.01$). Response bias (B'') did not differ between conditions.

5.3. The Y-X condition: Cases with the risk metric (Y), then blinded cases (X)

The four participants in this condition were first presented with the list of 100 cases that included patient data and the DSS risk metric. These participants were subsequently presented with the blinded list of 100 cases, which did not include the DSS risk metric. Table 3 presents the results from this test condition.

When participants were first presented with the list of cases that included patient data along with the risk metric, their predictive accuracy was significantly higher than chance (50.00% compared to 60.50%; $\chi^2 = 8.907$, $p < 0.01$). When these participants were then presented with the list of cases that did not include the risk metric, their predictive accuracy remained significantly higher than chance (50.00% compared to 63.50%; $\chi^2 = 14.832$, $p < 0.05$), but did not improve significantly (60.50% vs. 63.50%).

The proportion of true negatives, true positives, false negatives and false positives did not change significantly when comparing participant performance on each of the stimulus lists in this Y-X condition.

Participants who were first presented with the list of patient data that included the risk metric demonstrated a low level of sensitivity ($A' = 0.1265$). This improved significantly, though, when they were subsequently presented with the list of case data that did not include the risk metric ($A' = 0.2279$; $\chi^2 = 14.087$, $p < 0.05$). Participants were more able to discriminate between groups in the blinded condition. This result was counterintuitive and suggested a carry-over effect

Table 3
Predictive accuracy statistics for participants who were firstly presented with case data accompanied by the risk metric and then were presented with case data alone.

	With risk metric – Y(this list presented first)		Without risk metric – X(this list presented second)		p value*
	n	Proportion	n	Proportion	
Correct	242	0.6050	254	0.6350	0.3824
Incorrect	158	0.3950	146	0.3650	0.3824
TrueNegative	201	0.5025	210	0.5250	0.5246
TruePositive	41	0.1025	44	0.1100	0.7309
FalseNegative	63	0.1575	60	0.1500	0.7689
FalsePositive	95	0.2375	86	0.2150	0.4472
Hit rate		0.1025		0.1100	0.7309
False alarm rate		0.2375		0.2150	0.4472
Sensitivity (A′)		0.1265		0.2279	0.0002
Bias (B″)		0.2658		0.3263	0.0611

* *Notes:* p values test differences in proportion statistics. The standard statistical significance threshold is used here (0.05).

associated with the risk metric to the blinded condition. Response bias (B″) did not differ between conditions.

5.4. Analysis between X-Y and Y-X presentation conditions

In the X-Y condition, when participants were presented with the blinded stimuli first (patient data only), followed by the stimuli including the risk metric, their predictive performance on the blinded list was equal to chance (0.5000). This suggested that these participants were truly blind and that they experienced no risk-metric-related problem solving advantage.

In the Y-X condition, however, when participants were presented with the blinded stimuli second, after having been exposed to the stimuli including the risk metric, their predictive performance was significantly higher than those completing the same list in the X-Y condition (0.5000 vs. 0.6350; $\chi^2 = 14.832$, $p < 0.01$). This suggested that these participants were not truly blind when presented with the blinded stimulus list that did not include the risk metric. It appears that these participants were able to carry over a problem solving advantage to the blinded stimuli, after having been exposed to the risk metric in the previous stimulus list.

This proposition is also supported by considering the A′ metrics between the X-Y versus the Y-X conditions. The A′ statistic in the X-Y condition for the list that included the risk metric (A′ = 0.1151) was essentially equivalent to the A′ statistic in the Y-X condition that included the risk metric (A′ = 0.1265). It did not differ significantly. However, the A′ statistic in the Y-X condition for the blinded list (no risk metric presented; A′ = 0.2279) was significantly higher than the A′ statistic in the X-Y condition for the same blinded list (A′ = −0.0182; $\chi^2 = 81.401$, $p < 0.01$). This, again, suggests that a problem solving advantage was carried over in the Y-X condition when participants were required to complete the list of cases that included the risk metric first, and then were required to complete the list of cases that did not include the risk metric.

When interviewed after the experimental tasks, some participants stated that they were able to spot trends in the data when they were provided with the risk metric along with patient data (e.g., cases with diabetes had a high risk metric). They were able to develop problem solving strategies that they could then employ when considering the subsequent blinded list.

6. Discussion

Predictive modeling has previously been applied to high-risk surgical procedures [18], although these efforts appear to rarely be translated into usable DSS to effectively support clinical decision making. Systems that have been created in other fields are often complex and involve an intermediate scoring system. Furthermore, they often yield output that is not readily interpretable [13,28,29].

This study was designed to develop a predictive model and an efficiently usable DSS that could accurately predict the likelihood that complex spine surgery patients would experience complications. The study was also designed to determine whether or not this predictive model-driven DSS improved the decision making quality and problem solving performance of senior physicians involved in the delivery of complex spine surgical care. The experiment was focused on evaluating the effect of providing a cognitive aid (a quantified risk metric) on the problem solving process (risk prediction). Results supported the proposed hypotheses and the core proposition of cognitive fit theory.

When physicians were provided the probabilistic risk metric, in addition to preoperative patient information (BMI, sex, age, and diabetes, anemia, hypertension and smoking status), they were more able to accurately predict whether or not patients went on to experience postoperative complications than when surgeons were presented with preoperative patient information alone. The model-driven DSS alone performed better than expert physicians alone, who only had access to the preoperative patient data and did not have access to the risk metric, in correctly identifying the surgical cases that went on to experience postoperative complications. The DSS also performed better than expert physicians even when the physicians had access to the risk metric. The ability of complex spine surgeons to discriminate between cases that went on to experience complications and those that did not, improved when they were exposed to the DSS risk metric. Error avoidance was also improved when surgeons had access to the risk metric.

Cognitive fit theory proposes that when people are presented with a stimulus that aligns with the problem solving domain and task, their problem solving performance improves [30,31]. The probabilistic risk metric is a quantitative synthesis of patient risk factors. This study suggested that it afforded a powerful problem solving advantage for physicians. Results supported the core proposition of cognitive fit theory. When participants were presented with the list of cases without the risk metric, followed by the list of cases with the risk metric, their predictive accuracy improved significantly. Additionally, when they were presented with the list of cases that included the risk metric and then the list of cases without it, their performance on both lists was significantly better than chance. These results suggested that the risk metric helped with the problem solving process, even at a later time when it was not present in a subsequent stimulus list. It appears that the risk metric allowed physicians to spot trends in the data and develop problem solving heuristics that could then be employed at a later time.

The risks of complex spine surgery can be broken down into intraoperative, short-term post-operative, and long-term risks.

Intraoperative complications include severe blood loss, surgeon error, coagulopathy, blindness, neurologic injury, hypotensive sequelae and death [32,33]. Short-term complications (within 30 days of surgery) include infection, thromboembolism, reoperation, poor wound healing, hardware-related problems, neurologic problems, and complications arising from comorbid conditions. Long-term complications (more than 90 days after surgery) include infection, pseudoarthrosis, proximal and distal junctional failure, and hardware failure [15,29,34–39]. This study focused on complications presenting within 30 days after surgery because these complications have a direct impact on patient morbidity, mortality, and length of stay [8,40].

We developed a predictive model to assess this risk of complications based on a collection of routinely collected preoperative variables. These variables are easily, affordably, routinely and reliably measured [41]. The model provided good predictive differentiation between high and low risk patients. Our findings align with previous research [16,18]. BMI is a predictor of various complications for patients undergoing spine surgery [42,43] and diabetes is a predictor for the development of postoperative infection after spine surgery [44]. The remaining predictive variables were age, gender, smoking history, preoperative anemia, and hypertension. These variables have been linked to poorer outcomes after surgical intervention [12,17,45,46]. Negative outcomes may not arise directly from any one factor but may instead be the result of interactions amongst a collection of risk factors [13]. The interaction amongst these factors and their differential weighting can be captured in the multivariate predictive modeling process. Despite publications describing an increased risk of complications in patients with a prior history of depression and anxiety [16,46], the addition of these variables into the model decreased the accuracy of predicting postoperative complications in this study. This may have been due to insufficient data. Further studies investigating the predictive power of psychological variables are warranted.

In order to make the predictive statistical model usable for clinicians, a DSS was created. Key design principles guiding the development of our DSS were usability, efficiency, and clarity. The use of well-designed DSS can improve the quality of decision making, facilitate rapid insight, and aid accurate interpretation and planning [47,48]. This DSS generates a real-time, empirically-based, probabilistic estimate of a patient's risk of post-surgical complications with high accuracy.

The complexity of surgical decision making, particularly with regard to the assessment of risk, may lead to the use of cognitive heuristics [49], wherein a limited number of familiar or otherwise salient variables are considered more strongly, based on experience and preference, at the expense of others. Focusing on a small collection of risk factors may yield an inaccurate overall surgical risk assessment and result in suboptimal medical decision making [13]. Human reasoning and decision-making processes in the healthcare setting are often based on the use of heuristics and are compromised by cognitive and affective biases and errors. Consistency of judgment can be low [14,50,51], as biases influence assessments of surgical risk and the nature of the recommendations made to patients [52–54]. Mitigating these biases and errors is an important goal [14]. Factors like fatigue, sleep deprivation and cognitive overload are important determinants that predispose decision makers to the inadvertent tendency towards bias and the increased likelihood of error [14].

Evidence-based medicine involves the application of decision theory to mitigate cognitive limitations and reduce systematic biases and errors [55]. The application of the DSS tool developed here significantly improved the ability of physicians to accurately predict whether or not patients would be likely to experience postoperative complications, suggesting that the DSS was able to positively influence the quality of clinical judgment. By providing a clear prediction of risk, it may allow the surgeon and preoperative surgical review team to allocate more cognitive resources to other necessary considerations that may be more difficult to quantify, including social environment factors, and the specific needs of the patient and their family. This tool may also provide objective evidence of risk to help guide discussion in multidisciplinary preoperative clearance-for-surgery conferences [32,56]. Use of this tool adds negligible cost to the care of a complex spine patient, has the potential to improve outcomes, and is likely to increase the overall *value* of complex spine care, which may have reimbursement and competitive ramifications in the changing healthcare market [16,19,57]. Finally, this type of tool can facilitate the efficient and clear communication of risks to patients, thereby enhancing the informed consent process.

The predictive DSS developed here was designed for use in adult spinal deformity patients and was derived from a sample of patients at a single institution, limiting the ability to generalize this predictive algorithm to other institutions. It is important to note and consider, though, that inter-institutional generalizability was not the goal of this study. Each institution has its own way of delivering complex spine care and each institution has its own surgery-related risk profile [58–60]. Some institutions implement systematic care processes that have been shown to reduce risk and improve patient safety. Other institutions do not. Examples of these risk reduction processes include multidisciplinary patient case review conferences, a dual surgeon approach in the operating room and intraoperative coagulopathy monitoring [32]. Surgical outcomes and systemic risk profiles are likely to differ between institutions due to various factors including perioperative organizational processes, surgeon skill and the degree and quality of postoperative support. Because risk profiles vary by institution, the only forward-looking ways to accurately quantify surgical risk are to either (1) minimize inter-institutional risk variability by ensuring consistent care processes across institutions and then build general predictive models, or (2) generate institution-specific predictive risk models to account for each institution's own local risk profile. Large-scale, generalized predictive models based on large datasets from many institutions may be a useful low-fidelity risk assessment tool for institutions unable to create their own local risk models. However, we cannot be confident in the accuracy of these large general models at the local level, unless appropriate validation studies are conducted. While variability in care delivery exists, institution-specific risk modeling is a useful way to accurately quantify risk and confidently provide patients with the most accurate quantified risk assessment information. This analysis underscores the need for each healthcare system to perform similar analyses to maximize the quality of their own risk stratification processes.

A limitation of this study was our classification of the smoking variable, which was split into the categories of "smoker" and "never smoked." Patients in the smoker category were people who had smoked at any point in their life and may have stopped smoking well before their operation. The predictive strength of this variable may be increased by increasing smoking status categorization granularity, as research suggests that health status can improve after smoking cessation [61,62]. This study did not include neurologic complications. The frequency of these outcomes was very low and difficult to characterize. Future studies with larger samples would do well to include this variable.

As the complexity of medical decision making increases [63], this type of evidence-based data-driven DSS tool facilitates accurate risk stratification in complex spine surgery in a way that is clinically useful. DSS can improve the quality, value and safety of complex spine surgery care. The DSS tool's usability, simplicity and accuracy allow it to rapidly become an element of standard practice and to sharpen the accuracy of clinical decision making in favour of patient safety. We advocate for the development of similar predictive DSS at other institutions and for their application as an integral component of a broader systematic approach to patient evaluation.

Disclosures

All authors have reviewed and approved this manuscript and have no relevant financial or other conflicts of interest with regard to this research and its publication.

References

[1] Aizawa T et al. Increasing incidence of degenerative spinal diseases in Japan during 25 years: the registration system of spinal surgery in Tohoku University Spine Society. Tohoku J Exp Med 2016;238(2):153–63.
[2] McCarthy I et al. Health economic analysis of adult deformity surgery. Neurosurg Clin North Am 2013;24(2):293–304.
[3] Porter ME. A strategy for health care reform—toward a value-based system. N Engl J Med 2009;361(2):109–12.
[4] Kepler CK et al. Cost-utility analysis in spine care: a systematic review. Spine J 2012;12(8):676–90.
[5] Lenke LG et al. Prospective, multicenter assessment of acute neurologic complications following complex adult spinal deformity surgery: the Scoli-Risk-1 Trial. Spine J 2013;13(9):S67.
[6] Cho SK et al. Major complications in revision adult deformity surgery: risk factors and clinical outcomes with 2-to 7-year follow-up. Spine 2012;37(6):489–500.
[7] Bertram W, Harding I. Complicatons of spinal deformity and spinal stenosis surgery in adults greater than 50 years old. J Bone Joint Surg (British Volume) 2012;94(Suppl. X):105.
[8] Schwab FJ et al. Risk factors for major peri-operative complications in adult spinal deformity surgery: a multi-center review of 953 consecutive patients. Eur Spine J 2012;21(12):2603–10.
[9] Daubs MD et al. Adult spinal deformity surgery: complications and outcomes in patients over age 60. Spine 2007;32(20):2238–44.
[10] Glassman SD et al. The impact of perioperative complications on clinical outcome in adult deformity surgery. Spine 2007;32(24):2764–70.
[11] Tormenti MJ et al. Perioperative surgical complications of transforaminal lumbar interbody fusion: a single-center experience: clinical article. J Neurosurg Spine 2012;16(1):44–50.
[12] Acosta Jr FL et al. Morbidity and mortality after spinal deformity surgery in patients 75 years and older: complications and predictive factors: clinical article. J Neurosurg Spine 2011;15(6):667–74.
[13] Cooney MT, Dudina AL, Graham IM. Value and limitations of existing scores for the assessment of cardiovascular risk: a review for clinicians. J Am Coll Cardiol 2009;54(14):1209–27.
[14] Croskerry P, Singhal G, Mamede S. Cognitive debiasing 1: origins of bias and theory of debiasing. BMJ Qual Saf 2013;22(Suppl. 2):ii58–64.
[15] Scheer J 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. Global Spine J 2016;6(S 01):GO304.
[16] Akins PT et al. Risk factors associated with 30-day readmissions after instrumented spine surgery in 14,939 patients. Spine 2015;40(13):1022–32.
[17] Chitale R et al. International classification of disease clinical modification 9 modeling of a patient comorbidity score predicts incidence of perioperative complications in a nationwide inpatient sample assessment of complications in spine surgery. J Spinal Disord Tech 2015;28(4):126–33.
[18] Boachie-Adjei O et al. Surgical risk stratification based on preoperative risk factors in severe pediatric spinal deformity surgery. Spine Deform 2014;2(5):340–9.
[19] Bilimoria KY et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. J Am Coll Surg 2013;217(5):833–42. e3.
[20] Bekelis K et al. A predictive model of complications after spine surgery: the National Surgical Quality Improvement Program (NSQIP) 2005–2010. Spine J 2014;14(7):1247–55.
[21] Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Münster (PROCAM) study. Circulation 2002;105(3):310–5.
[22] Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.
[23] Wang MQ, Eddy JM, Fitzhugh EC. Application of odds ratios and logistic models in epidemiology and health research. Health Values 1995;19(1):59–62.
[24] Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. John Wiley & Sons; 2013.
[25] Hosmer DW, Lemeshow S. Applied logistic regression. John Wiley & Sons; 2004.
[26] D'Agostino Sr RB et al. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. Jama 2001;286(2):180–7.
[27] Stanislaw H, Todorov N. Calculation of signal detection theory measures. Behav Res Methods Instrum Comput 1999;31(1):137–49.
[28] Jacobs JP et al. the society of thoracic surgeons congenital heart surgery database mortality risk model: part 2—clinical application. Ann Thorac Surg 2015;100(3):1063–70.
[29] O'Brien SM et al. The society of thoracic surgeons congenital heart surgery database mortality risk model: part 1—statistical methodology. Ann Thorac Surg 2015;100(3):1054–62.
[30] Shaft TM, Vessey I. The role of cognitive fit in the relationship between software comprehension and modification. Mis Q 2006:29–55.
[31] Vessey I. Cognitive fit: a theory-based analysis of the graphs versus tables literature. Decis Sci 1991;22(2):219–40.
[32] Sethi RK et al. The Seattle Spine Team approach to adult deformity surgery: a systems-based approach to perioperative care and subsequent reduction in perioperative complication rates. Spine Deform 2014;2(2):95–103.
[33] Huang Y-H, Ou C-Y. Significant blood loss in lumbar fusion surgery for degenerative spine. World Neurosurg 2015;84(3):780–5.
[34] Yu X et al. Prediction of massive blood loss in scoliosis surgery from preoperative variables. Spine 2013;38(4):350–5.
[35] Guay J et al. Predicting blood loss in surgery for idiopathic scoliosis. Can J Anaesth 1994;41(9):775–81.
[36] Baldus CR et al. Can we safely reduce blood loss during lumbar pedicle subtraction osteotomy procedures using tranexamic acid or aprotinin?: A comparative study with controls. Spine 2010;35(2):235–9.
[37] Modi HN et al. Intraoperative blood loss during different stages of scoliosis surgery: a prospective study. Scoliosis Spinal Disord 2010;5(1):1.
[38] Elgafy H et al. Blood loss in major spine surgery: are there effective measures to decrease massive hemorrhage in major spine fusion surgery? Spine 2010;35(9S):S47–56.
[39] Emami A et al. Outcome and complications of long fusions to the sacrum in adult spine deformity: luque-galveston, combined iliac and sacral screws, and sacral fixation. Spine 2002;27(7):776–86.
[40] Lee MJ et al. Risk factors for medical complication after lumbar spine surgery: a multivariate analysis of 767 patients. Spine 2011;36(21):1801.
[41] Sehestedt T, Olsen MH. Traditional versus new models of risk prediction, in Early Vascular Aging (eva). Academic Press Incorporated; 2015.
[42] McClendon Jr J et al. The impact of body mass index on hospital stay and complications after spinal fusion. Neurosurgery 2014;74(1):42–50.
[43] Jiang J et al. Does obesity affect the surgical outcome and complication rates of spinal surgery? A meta-analysis. Clin Orthop Relat Res® 2014;472(3):968–75.
[44] Salvetti DJ et al. Preoperative prealbumin level as a risk factor for surgical site infection following elective spine surgery. Surg Neurol Int 2015;6(Suppl. 19):S500.
[45] Nerland US et al. The risk of getting worse: predictors of deterioration after decompressive surgery for lumbar spinal stenosis: a multicenter observational study. World Neurosurg 2015;84(4):1095–102.
[46] Chapin L, Ward K, Ryken T. Preoperative depression, smoking, and employment status are significant factors in patient satisfaction after lumbar spine surgery. J Spinal Disord Tech 2015.
[47] Few S. Information dashboard design. Sebastopol: O'Reilly Media Inc.; 2006.
[48] Kawamoto K et al. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. Bmj 2005;330(7494):765.
[49] Gigerenzer G, Gaissmaier W. Heuristic decision making. Annu Rev Psychol 2011;62:451–82.
[50] Dovidio JF, Fiske ST. Under the radar: how unexamined biases in decision-making processes in clinical interactions can contribute to health care disparities. Am J Public Health 2012;102(5):945–52.
[51] Hilbert M. Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. Psychol Bull 2012;138(2):211.
[52] Stone J, Moskowitz GB. Non-conscious bias in medical decision making: what can be done to reduce it? Med Educ 2011;45(8):768–76.
[53] Dawson NV, Arkes HR. Systematic errors in medical decision making. J Gen Intern Med 1987;2(3):183–7.
[54] Chapman GB, Elstein AS. Cognitive processes and biases in medical decision making. Decision making in health care: theory, psychology, and applications. Cambridge: Cambridge University Press; 2000.
[55] Elstein AS, Schwarz A. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. BMJ 2002;324(7339):729.
[56] Buchlak QD et al. Complication avoidance with pre-operative screening: insights from the Seattle spine team. Curr Rev Musculoskelet Med 2016:1–11.
[57] Kassin MT et al. Risk factors for 30-day hospital readmission among general surgery patients. J Am Coll Surg 2012;215(3):322–30.
[58] Krumholz HM et al. Comparing AMI mortality among hospitals in patients 65 years of age and older evaluating methods of risk adjustment. Circulation 1999;99(23):2986–92.
[59] Goodney PP et al. Hospital volume, length of stay, and readmission rates in high-risk surgery. Annals Surg 2003;238(2):161–7.

[60] Ghaferi AA, Birkmeyer JD, Dimick JB. Variation in hospital mortality associated with inpatient surgery. N Engl J Med 2009;361(14):1368–75.
[61] Fiorini T et al. Is there a positive effect of smoking cessation on periodontal health? A systematic review. J Periodontol 2014;85(1):83–91.
[62] Truntzer J et al. Smoking cessation and bone healing: optimal cessation timing. Eur J Orthop Surg Traumatol 2015;25(2):211–5.
[63] Hunink MM et al. Decision making in health and medicine: integrating evidence and values. Cambridge University Press; 2014.