Surgery 165 (2019) 1035–1045

Contents lists available at ScienceDirect

Surgery

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

Trauma/Critical Care

Comparing clinical judgment with the *MySurgeryRisk* algorithm for preoperative risk assessment: A pilot usability study

Meghan Brennan, MD, MS[a,b], Sahil Puri, MS[c], Tezcan Ozrazgat-Baslanti, PhD[a,d], Zheng Feng, MS[a,e], Matthew Ruppert, BS[a,d], Haleh Hashemighouchani, MD[a,d], Petar Momcilovic, PhD[a,f], Xiaolin Li, PhD[a,e], Daisy Zhe Wang, PhD[a,c], Azra Bihorac, MD, MS[a,d,*]

[a] *Precision and Intelligent Systems in Medicine (PRISMA^P), Division of Nephrology, Hypertension and Transplantation, University of Florida, Gainesville*
[b] *Department of Anesthesiology, University of Florida College of Medicine, Gainesville*
[c] *Department of Computer and Information Science and Engineering, University of Florida Herbert Wertheim College of Engineering, Gainesville*
[d] *Department of Medicine, University of Florida College of Medicine, Gainesville*
[e] *Department of Electrical and Computer Engineering, University of Florida Herbert Wertheim College of Engineering, Gainesville*
[f] *Department of Industrial and Systems Engineering, University of Florida Herbert Wertheim College of Engineering, Gainesville*

ARTICLE INFO

Article history:
Accepted 2 January 2019
Available online 18 February 2019

ABSTRACT

Background: Major postoperative complications are associated with increased cost and mortality. The complexity of electronic health records overwhelms physicians' abilities to use the information for optimal and timely preoperative risk assessment. We hypothesized that data-driven, predictive-risk algorithms implemented in an intelligent decision-support platform simplify and augment physicians' risk assessments.

Methods: This prospective, nonrandomized pilot study of 20 physicians at a quaternary academic medical center compared the usability and accuracy of preoperative risk assessment between physicians and *MySurgeryRisk,* a validated, machine-learning algorithm, using a simulated workflow for the real-time, intelligent decision-support platform. We used area under the receiver operating characteristic curve to compare the accuracy of physicians' risk assessment for six postoperative complications before and after interaction with the algorithm for 150 clinical cases.

Results: The area under the receiver operating characteristic curve of the *MySurgeryRisk* algorithm ranged between 0.73 and 0.85 and was significantly better than physicians' initial risk assessments (area under the receiver operating characteristic curve between 0.47 and 0.69) for all postoperative complications except cardiovascular. After interaction with the algorithm, the physicians significantly improved their risk assessment for acute kidney injury and for an intensive care unit admission greater than 48 hours, resulting in a net improvement of reclassification of 12% and 16%, respectively. Physicians rated the algorithm as easy to use and useful.

Conclusion: Implementation of a validated, *MySurgeryRisk* computational algorithm for real-time predictive analytics with data derived from the electronic health records to augment physicians' decision-making is feasible and accepted by physicians. Early involvement of physicians as key stakeholders in both design and implementation of this technology will be crucial for its future success.

A.B., T.O.B., P.M., X.L., and D.Z.W. were supported by RO1GM110240 from the NIGMS. A.B. and T.O.B. were supported by Sepsis and Critical Illness Research Center Award P50 GM-111152 from the NIGMS. T.O.B. has received a grant (97071) from Clinical and Translational Science Institute, University of Florida. This work was supported in part by the NIH/NCATS Clinical and Translational Sciences Award to the University of Florida UL1 TR000064. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the University of Florida. University of Florida and A.B., T.O.B., X.L., P.M., and Z.W. have a patent pending on real-time use of clinical data for surgical risk prediction using machine learning models in the *MySurgeryRisk* algorithm.

[*] Reprint requests: Reprints will not be available from the authors.
E-mail address: abihorac@ufl.edu (A. Bihorac).

https://doi.org/10.1016/j.surg.2019.01.002
0039-6060/© 2019 Elsevier Inc. All rights reserved.

Introduction

Postoperative complications increase the odds of 30-day mortality, lead to greater readmission rates, and require greater utilization of health care resources.[1–5] Prediction of postoperative complications for individual patients is increasingly complex because of the need for rapid decision making coupled with the constant influx of dynamic physiologic data in electronic health records (EHRs). Risk-communication tools and scores are continually being developed to convert the large amount of available EHR data into a usable format, but it is unclear whether these tools are able to change users' perceptions of risk.[6]

Two commonly used and validated risk scores for surgical patients, the National Surgical Quality Improvement score and the Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity, provide risk stratifications for selected postoperative complications.[7,8] Although these scoring systems have been proven reliable, they have not been automated or integrated uniformly into the EHR because they require elaborate data collection and calculations.[9] Other risk scores frequently integrated into EHR, such as the Modified Early Warning Score or Rothman Index, are designed to alert health care providers to all at-risk patients; however, these risk scores often have high false-positive rates and do not differentiate between risks of specific postoperative complications.[1,10] Interesting to note are a lack of studies comparing how physicians' clinical judgment compares with these risk models for predicting surgical complications.

Recently we validated the machine learning algorithm *MySurgeryRisk*, which predicts preoperative risk for major postoperative complications using EHR data. The algorithm is integrated into the clinical workflow through the intelligent, perioperative platform for real-time analytics of routine clinical data and prospective data collection for the model retraining.[11–13]

In this prospective pilot study we compared the usability and accuracy of preoperative risk assessment between physicians and the *MySurgeryRisk* algorithm, using a simulated workflow for the real-time, intelligent, decision-support platform. We tested the hypothesis that physicians will gain knowledge from interaction with the algorithm and improve the accuracy of their risk assessments.

Materials and Methods

The Institutional Review Board and Privacy Office of the University of Florida ([UF] Gainesville) approved this study (#2013-U-1338, #5-2009). Written informed consent was obtained from all participants.

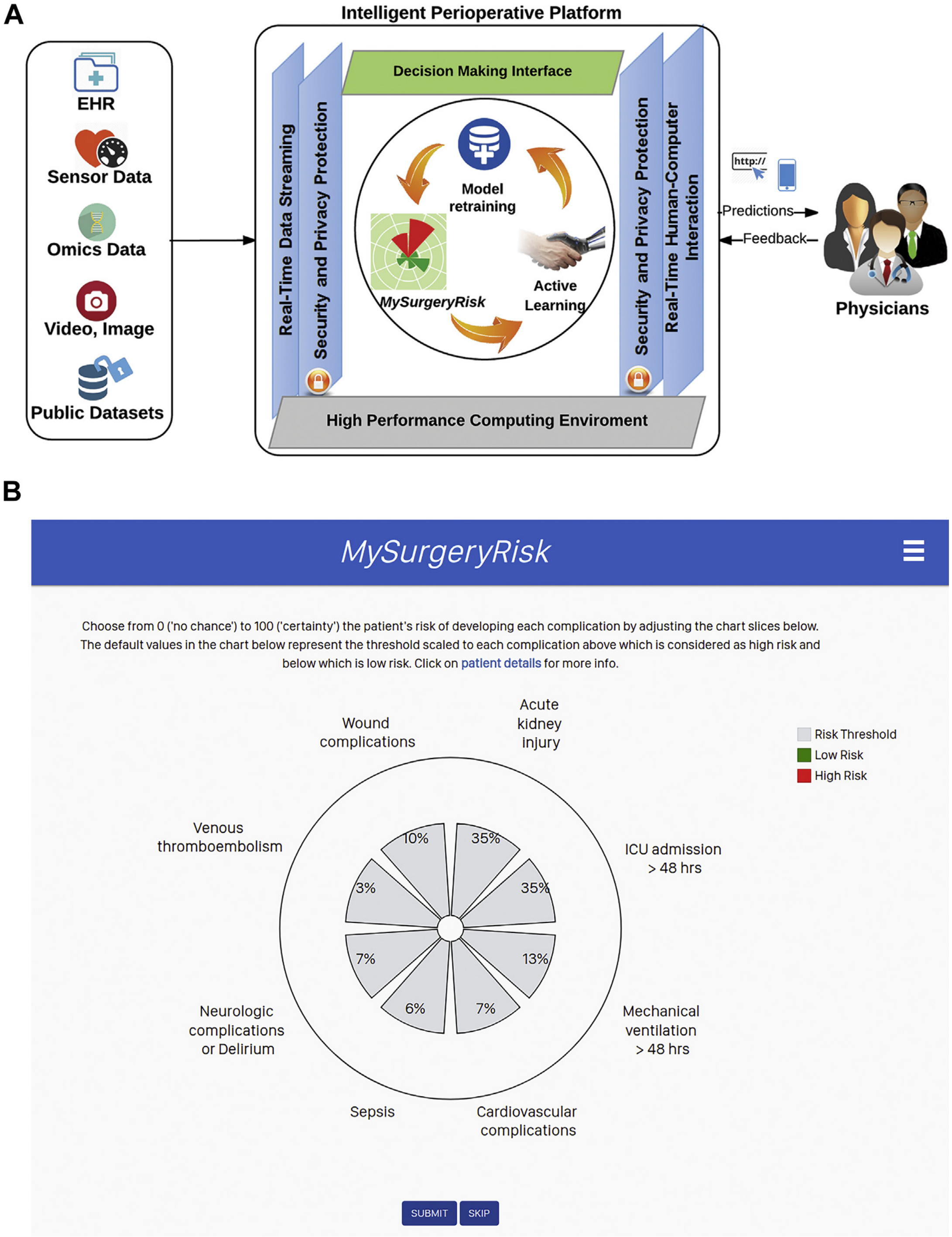
Study design

This prospective, nonrandomized, interventional pilot study of 20 surgical intensivists (attending physicians or trainees in anesthesiology and surgical fellowships) at a single academic quaternary care institution was designed to assess the usability and accuracy of the *MySurgeryRisk* algorithm for preoperative risk assessment, using a simulated workflow for the real-time, intelligent, decision-support platform.

The *MySurgeryRisk* algorithm is a validated, machine-learning algorithm that predicts preoperative risk for major complications, using existing clinical data in the EHR data with a high sensitivity and a high specificity.[11–13] Development and validation of the algorithm is described in detail in Bihorac et al.[11] We designed an intelligent platform such that the *MySurgeryRisk* algorithm can be implemented in real time to provide an augmented preoperative risk assessment for inpatient surgical cases at University of Florida. Before operation, the platform autonomously integrates and transforms existing EHR data to run the *MySurgeryRisk* algorithm in real time and calculates risk probabilities for major complications. The output of the algorithm is presented to the surgeons scheduled to perform the operations, using an interactive interface that resides on the Web portal within the platform and allows user feedback (Figs. 1, *A–D*).[11] This pilot study was performed before the launch of this real-time platform to evaluate its usability and performance. We simulated the real-time workflow of this platform for 150 patient cases to allow us to study the participants' interaction with the results of the algorithm in the same manner in which they would with fully a functional real-time platform.[11] We selected new cases from a large, retrospective, longitudinal database of adult patients 18 years of age or older admitted to University of Florida Health (UF Health) for greater than 24 hours after any type of inpatient surgical procedure 2000 to 2010.[11,12,14] The selected cases were not used for the development of the algorithm reported elsewhere.[11] For each case, we had a complete HER from which we used available preoperative data as an input for both the algorithm and the physicians' risk assessment, and the clinical data related to the hospitalization after the operation were used to determine whether complications occurred (a detailed description of the algorithm input data and assessment of complications is provided in Bihorac et al[11]). Physicians and algorithm were blinded to the observed outcomes of the cases.

Each physician evaluated 8 to 10 individual cases and provided a risk assessment for each complication both before and after seeing the scores of the *MySurgeryRisk* algorithm (Fig. 2). All evaluations were performed on a personal laptop during a single, "think aloud" individual session with a research coordinator who assisted with the use of the interactive interface of the platform. The physicians had access to all available preoperative data that were also used as an input for *MySurgeryRisk* algorithm. For each case, we summarized the preoperative clinical data as a brief clinical vignette similar to a progress note available in the patient's chart. On review of a case, the physicians were asked to assess the absolute risk for each of the 6 complications ranging 0% to 100%, using a sliding pie chart for data input. For each complication, we used a set threshold[11] to determine whether the assigned absolute risk classified the patient into the low-risk or high-risk group as reflected in the change of color from green to red on the pie chart (Fig. 1, *B*). After the initial risk assessment, physicians were presented with absolute-risk scores and risk groups calculated by *MySurgeryRisk* algorithm (Fig. 1, *C*). Each score was accompanied with the display of the top 3 features that were the most important contributors to the calculated risk for the individual patient. Finally, the physicians were asked to repeat their risk assessment for the same patient to assess whether the interaction with the algorithm would change their perception of the risk. They used a similar interactive pie chart to reenter the absolute risk cores for each complication (Fig. 1, *D*). At the end of the session. we surveyed the physicians regarding the usability of the algorithm and Web interface.

At study enrollment, we evaluated each physician's decision-making style and numeracy skills with a validated cognitive reflection test (CRT) and a numeracy assessment test.[15–17] The CRT consists of three questions and was validated against other measures of cognitive reflection.[17] A lesser score on the CRT indicated a more impulsive decision-making preference and a strong reliance on intuition, and a greater score indicated a reflective thinker with a more cautious decision-making preference and less reliance on intuition.[17] The numeracy assessment test measured the physician's ability to understand and use numbers, and a greater score indicated an increased ability to use numeric data.

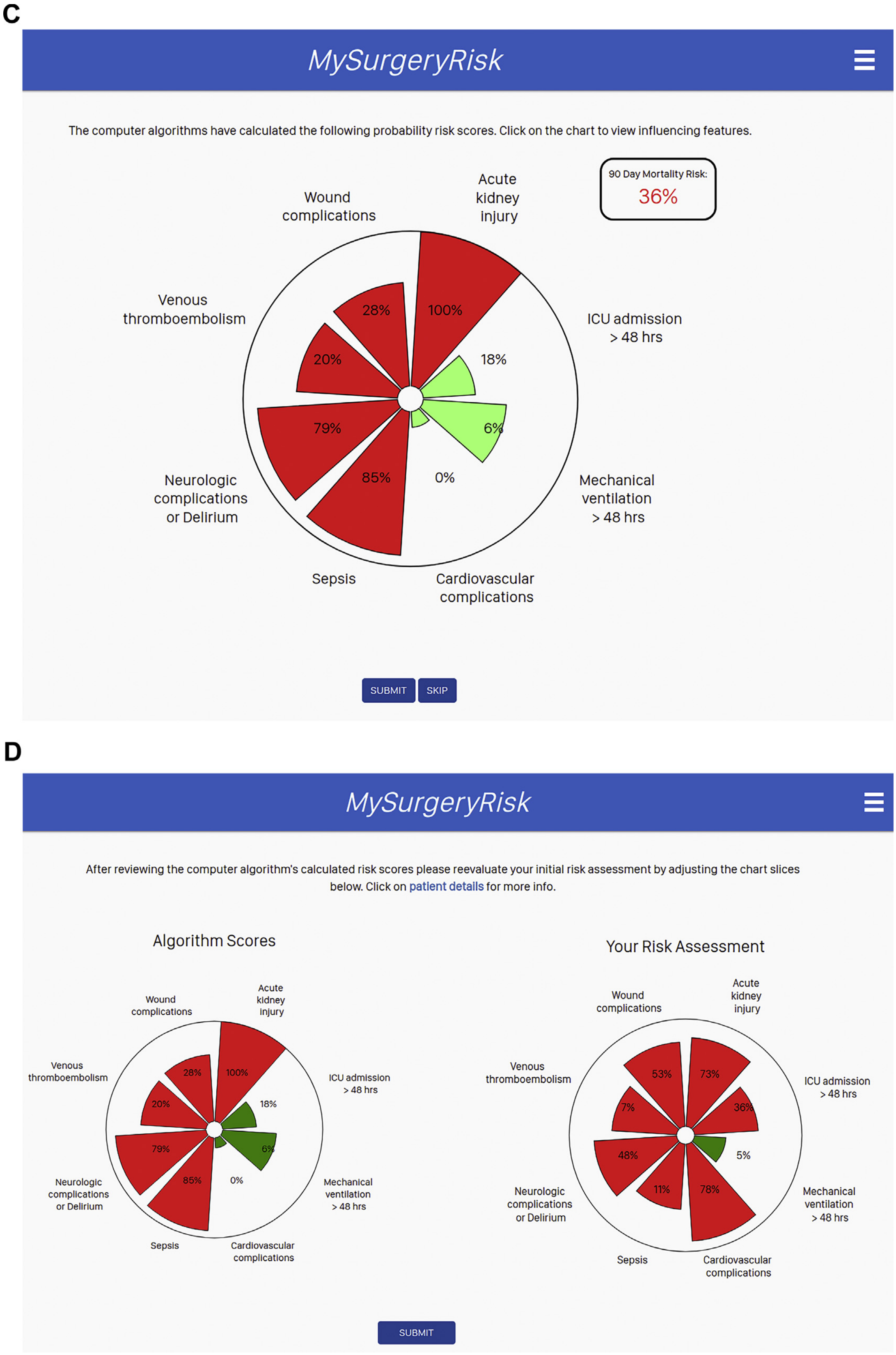


Choose from 0 ('no chance') to 100 ('certainty') the patient's risk of developing each complication by adjusting the chart slices below. The default values in the chart below represent the threshold scaled to each complication above which is considered as high risk and below which is low risk. Click on patient details for more info.

Fig. 1. (*A*) Design of Intelligent Perioperative Platform that hosts *MySurgeryRisk* algorithm, reprinted with permission from *Annals of Surgery*, Bihorac et al.[11] (*B*) The interactive interface for physicians to input their initial assessment of absolute risk for each complication. (*C*) The interactive interface displaying absolute risk scores and risk groups calculated by *MySurgeryRisk* algorithm to physicians. Each score was accompanied with the display of the top three features that were the most important contributors to calculated risk for the individual patient. (*D*) The interactive interface for physicians to input their repeated assessment of the absolute risk after reviewing *MySurgeryRisk* scores for the same case.

Statistical analysis

To increase the number of cases for complications with a low prevalence in the original cohort, we selected patient cases with observed 30-day mortality and matched them with patient cases without an observed 30-day mortality in a 5:1 ratio. This approach allowed us to increase the number of cases with observed 30-day mortality in testing the cohort to 16% compared with 3% in original cohort. As expected, this strategy also resulted in an increase of the prevalence for other complications (Table I). Each of the 150 cases was treated as an independent observation. The study had an 80% power to detect at least a 10% difference between the algorithm and the physician risk assessments, assuming a standard deviation of 10%.[11,12,14]



The computer algorithms have calculated the following probability risk scores. Click on the chart to view influencing features.

After reviewing the computer algorithm's calculated risk scores please reevaluate your initial risk assessment by adjusting the chart slices below. Click on patient details for more info.

Fig. 1. Continued

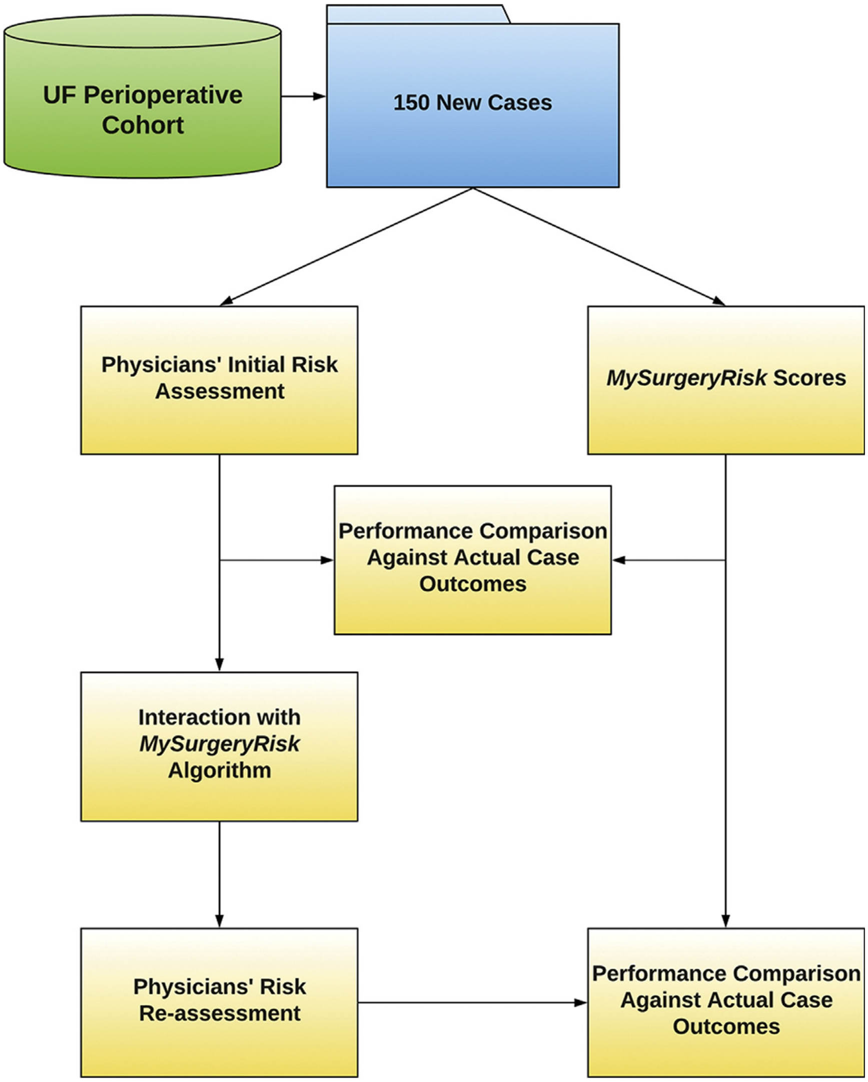

UF Perioperative Cohort
150 New Cases
Physicians' Initial Risk Assessment
MySurgeryRisk Scores
Performance Comparison Against Actual Case Outcomes
Interaction with *MySurgeryRisk* Algorithm
Physicians' Risk Re-assessment
Performance Comparison Against Actual Case Outcomes

Fig. 2. Flowchart of the study design.

We calculated the area under the receiver operating characteristic curve (AUC) to test the accuracy of both the *MySurgeryRisk* algorithm and the physicians' risk assessments for predicting the occurrence of each of the six complications separately. For each case, we compared both the initial and repeated (after reviewing the *MySurgeryRisk* scores) physician risk assessments to the *MySurgeryRisk* scores against the true occurrence of each complication. The change in accuracy between the initial and the repeated physician's risk assessments were compared using the DeLong test.[18] The net improvement in reclassification was calculated to measure the improvement in the physicians' risk reassessments after interaction with the algorithm.[19] We calculated the misclassification rate as a proportion of the cases misclassified in a wrong risk group, based on the observed outcome for each complication. A case was considered as misclassified if the physician's assessment of absolute risk classified the patient into a low-risk group for positive cases where the complication was observed or into a high-risk group for negative cases where the complication was not observed. We determined thresholds of absolute risk, separating low-risk and high-risk groups based on the prevalence of each complication in the original cohort, with values similar to the thresholds reported elsewhere[11] (0.32 for intensive care unit [ICU] admission greater than 48 hours, 0.26 for acute kidney injury, 0.13 for mechanical ventilation greater than 48 hours, 0.07 for cardiovascular complications, 0.05 for severe sepsis, and 0.034 for 30-day mortality). The proportion of misclassified cases between physicians stratified by specialty or training status were compared using the Fisher exact test. The correlation between years of practice and the average misclassification rate for physicians was calculated using the Spearman correlation. A *t* test was used to test the hypothesis that the mean absolute difference in risk-assessment score, after interaction with the *MySurgeryRisk* algorithm, was different than 0. All statistical analyses were performed using SAS software (v 9.4, SAS Institute, Cary, NC) and R software (v 3.4.0, https://www.r-project.org/).

Results

Comparison between physicians' initial risk assessment and the MySurgeryRisk algorithm

A total of 20 physicians provided risk-assessment scores for 6 postoperative complications for a total of 150 patient cases. Of the physicians, 14 were attending physicians, the remainder were

Table I
Summary of reference cohort and testing cohort

	Overall reference cohort ($N = 10{,}000$)	Testing cohort ($n = 150$)
Demographic features		
Age (years), median (25th–75th)	56 (43, 68)	62 (51–70)
Female sex, *n* (%)	4,898 (49)	73 (49)
Race, *n* (%)		
White	8,032 (80)	125 (83)
African American	1,235 (12)	19 (13)
Hispanic	338 (3)	4 (3)
Other	210 (2)	1 (1)
Primary insurance group, *n* (%)		
Medicare	3,845 (38)	71 (47)
Medicaid	1,294 (13)	17 (11)
Private	4,088 (41)	57 (38)
Uninsured	773 (8)	5 (3)
Socioeconomic features		
Neighborhood characteristics		
Rural area, *n* (%)	3,156 (32)	50 (33)
Total population, median (25th–75th)	17,176 (10,002, 27,782)	16,192 (9,812–25,971)
Median income (US$), median (25th–75th)	33,221 (28,400, 40,385)	33,293 (28,589–41,410)
Total proportion of African Americans (%), mean (SD)	15 (15)	16 (16)
Total proportion of Hispanic (%), mean (SD)	6 (6)	5 (6)
Population proportion below poverty (%), mean (SD)	15 (8)	14 (8)
Distance from residency to hospital (km), median (25th–75th)	54.1 (26.7, 117.7)	47.8 (21.9–99.4)
Comorbidity features		
Charlson comorbidity index (CCI), median (25th–75th)	1 (0, 2)	
Cancer, *n* (%)	2,025 (20)	27 (18)
Diabetes, *n* (%)	1,620 (16)	25 (17)
Chronic pulmonary disease, *n* (%)	1,573 (16)	30 (20)
Peripheral vascular disease, *n* (%)	1,131 (11)	26 (17)
Cerebrovascular disease, *n* (%)	823 (8)	28 (19)
Congestive heart failure, *n* (%)	781 (8)	20 (13)
Myocardial infarction, *n* (%)	650 (7)	17 (11)
Liver disease, *n* (%)	503 (5)	10 (7)
Operative features		
Admission		
Weekend admission, *n* (%)	142 (14)	26 (17)
Admission source, *n* (%)		
Emergency room	2,637 (27)	45 (30)
Outpatient setting	5,900 (60)	76 (51)
Transfer	1,334 (14)	29 (19)
Emergent surgery status, *n* (%)	4,545 (45)	80 (53)
Surgery type, *n* (%)		
Cardiothoracic surgery	1,284 (13)	22 (15)
Noncardiac general surgery	3,905 (39)	54 (36)
Neurologic surgery	1,713 (17)	30 (20)
Specialty surgery	2,959 (30)	41 (27)
Other surgery	139 (1)	3 (2)
Preoperative and admission day laboratory results		
Estimated reference glomerular filtration rate (mL/min/1.73 m^2), median (25th–75th)	91.9 (71.6, 107.8)	85.8 (56.2–102.6)
Automated urinalysis, urine protein (mg/dL), *n* (%)		
Missing	8,330 (83)	124 (83)
<30	1,109 (11)	15 (10)
≥30	561 (6)	11 (7)
Complete blood count tests, *n* (%)		
0	2,954 (30)	34 (23)
1	5,526 (55)	88 (59)
≥2	1,520 (15)	28 (18)
Hemoglobin, g/dL median (25th–75th)	11.7 (10.2, 13.3)	11.4 (9.7–13.0)
Hematocrit, median (25th–75th)	34.4 (30.1, 38.6)	33.4 (29.1–37.6)
Admission day medications		
Admission day medication groups (top 10 categories), *n* (%)		
β-blockers	2,333 (23)	42 (28)
Diuretics	1,124 (11)	19 (13)
Statins	1,122 (11)	23 (15)
Aspirin	671 (7)	12 (8)
Angiotensin-converting–enzyme inhibitors	979 (10)	13 (9)
Vasopressors and inotropes	551 (6)	15 (10)
Bicarbonate	406 (4)	15 (10)
Antiemetics	5,717 (57)	84 (56)
Aminoglycosides	687 (7)	5 (3)
Corticosteroids	1,218 (12)	23 (15)

Table I (*continued*)

	Overall reference cohort (N = 10,000)	Testing cohort (n = 150)
Outcomes		
Postoperative complications, n (%)		
Acute kidney injury during hospitalization	3,869 (39)	82 (55)
Intensive care unit admission >48 hours	3,161 (32)	74 (49)
Mechanical ventilation >48 hours	1,313 (13)	55 (37)
Wound complications	1,085 (11)	17 (11)
Neurologic complications or delirium	773 (8)	26 (17)
Cardiovascular complications	724 (7)	43 (29)
Sepsis	536 (5)	39 (26)
Venous thromboembolism	287 (3)	11 (7)
Mortality, n (%)		
30-day mortality	335 (3)	24 (16)
3-month mortality	668 (7)	41 (27)
6-month mortality	919 (9)	46 (31)
12-month mortality	1,261 (13)	48 (32)
24-month mortality	1,687 (17)	55 (37)

Table II
Physician characteristics

Physician characteristics	N = 20
Female sex, n (%)	5 (25)
Age (years), n (%)	
≤30	2 (10)
31–40	10 (50)
41–50	4 (20)
>50	4 (20)
Attending physicians, n (%)	14 (70)
Specialty, n (%)	
Anesthesiology/emergency medicine	16 (80)
Surgery	4 (20)
Years since graduation, mean (SD)	13 (10)
High numeracy score (≥9),[*] n (%)	18 (90)
Cognitive reflection test score,[†] n (%)	
Low score (0): Nonreflective thinker with unquestioning reliance on intuition	3 (15)
Intermediate score (1–2): Balance between reflective thinking and intuition	14 (70)
High score (3): Reflective thinker whose initial intuition is tempered by analysis	3 (15)

[*] The numeracy assessment is a measure of ability to understand and use numeric data.

[†] The cognitive reflection test is a validated measure of decision-making preference—differentiating impulsive decision-making preferences from reflective decision-making preferences.

residents or fellows, with an average of 13 years of experience. A total of 90% had high numeracy skills on the numeracy assessment. The majority, 70%, scored in the intermediate range for the decision-making style, reflecting a balance between intuitive and reflective decision making. Only 15% of physicians scored in the impulsive decision-making range, with a strong reliance on intuition, and the other 15% were in the reflective thinker range, with a more cautious decision-making preference and less reliance on intuition (Table II).

As expected, the prevalence of postoperative complications among the 150 cases was greater than in the reference population as a result of selection process and ranged between 16% for 30-day mortality and 49% for ICU admission >48 hours (Table I). The *MySurgeryRisk* algorithm (AUCs ranged from 0.64 to 0.85) was more accurate in predicting the risk for complications compared with the initial physicians' risk assessments (AUCs ranged between 0.47 and 0.69) with greater AUCs for predicted absolute risks for all complications ($P < .002$ each) except cardiovascular (Table III). Compared with the *MySurgeryRisk* algorithm, the physicians were more likely to underestimate the risk of an ICU stay and acute kidney injury (AKI) for positive cases, for which complications occurred (their assessment of absolute risk was less than the *MySurgeryRisk* score). In contrast, the physicians overestimated the risk of mortality, cardiovascular complications, and severe sepsis for negative cases, for which complications did not occur (their assessment of absolute risk was greater than of *MySurgeryRisk*).

Among physicians, the rate of misclassifying patients in a wrong risk category based on the observed outcome (low-risk group for patients with observed outcome and high-risk group for patients without outcome) ranged from 28% for severe sepsis to 64% for the 30-day mortality (Table IV). We observed no significant difference in the misclassification rate between attending physicians and trainees. The years of practice correlated with the misclassification rate for predicting risk of AKI, with more experienced physicians having a lesser rate ($r = -0.63$, $P = .01$). The proportion of cases where the physicians' assessment of absolute risk was more accurate than the algorithm (for cases where the complication was observed, the physicians estimated a greater absolute risk and for cases where the complication was not observed, the physicians estimated lesser absolute risk) ranged widely from 20% when predicting mortality to 46% when predicting risk for AKI (29% for severe sepsis, 33% for mechanical ventilation greater than 48 hours, 35% for cardiovascular complications, and 39% for ICU admission greater than 48 hours). We observed no difference based on physician specialty or training status.

Change in the physicians' risk-assessment after the interaction with the MySurgeryRisk algorithm

To assess whether physicians changed their perception of absolute risk after reviewing the *MySurgeryRisk* scores, we compared their initial and repeated risk assessments. In greater than 75% of clinical cases, physicians responded to interaction with the algorithm by changing their risk-assessment score, and the majority of the new scores were closer to the *MySurgeryRisk* score. The average change in the physicians' absolute risk–perception ranged between 8% and 10% (Table V). Compared with their initial risk assessment, the accuracy of the physician's repeated risk assessments improved after the interaction with the algorithm, with an increase in AUC between 2% and 5% for all complications except 30-day mortality. The improvement in AUC for predicting cardiovascular complications before and after their interaction was the only complication that was statistically significant, increasing by 5% (Table VI). The

Table III
Comparison between physicians' initial risk assessments and *MySurgeryRisk* algorithm prediction

Postoperative complications	Prevalence of positive cases with complications among 150 cases, *n* (%)	Physicians' first risk assessment AUC (95% CI)	*MySurgeryRisk* algorithm risk score AUC (95% CI)	*P* value for difference in AUC
Intensive care unit admission longer than 48 hours	74 (49)	0.69 (0.61–0.77)	0.84 (0.78, 0.90)	< .001
Acute kidney injury	57 (38)	0.65 (0.56–0.74)	0.79 (0.72, 0.87)	.002
Mechanical ventilation greater than 48 hours	55 (37)	0.66 (0.57–0.75)	0.85 (0.79, 0.91)	< .001
Cardiovascular complications	43 (29)	0.54 (0.44–0.65)	0.64 (0.55, 0.73)	.09
Severe sepsis	39 (26)	0.54 (0.44–0.64)	0.78 (0.69, 0.87)	< .001
30-day mortality	24 (16)	0.47 (0.36–0.57)	0.73 (0.64, 0.83)	< .001

CI, confidence interval.
Note: *MySurgeryRisk* algorithm and physicians evaluated absolute risk for complications ranging from 0 (no risk) to 100 (complete certainty of risk).

Table IV
Misclassification rate of physicians' initial assessments of absolute risk for postoperative complications

	Intensive care unit admission greater than 48 hours	Acute kidney injury	Mechanical ventilation greater than 48 hours	Cardiovascular complications	Severe sepsis	30-day mortality
Misclassification rate						
Physicians' initial risk assessment						
Overall	37 (4)	41 (0.04)	43 (4)	57 (4)	63 (4)	65 (4)
By physician's specialty						
Surgery	38 (8)	59 (8)*	44 (8)	69 (7)	67 (8)	77 (7)
Anesthesiology/emergency medicine	37 (5)	35 (5)	42 (5)	52 (5)	61 (5)	60 (5)
By training status						
Attending physicians	36 (5)	38 (5)	40 (5)	52 (5)	61 (5)	61 (5)
Trainees	41 (8)	51 (8)	49 (8)	68 (7)	68 (7)	73 (7)
Correlation between years of practice and misclassification rate, r (*P* value)	−0.13 (.63)	−0.63 (.01)	−0.27 (.30)	−0.14 (.60)	0.03 (.92)	−0.43 (.09)

Note: Data represent the proportion of misclassified cases as a percent and its standard error in parenthesis. The correlation between years of practice and the average misclassification rate for physicians was calculated using the Spearman correlation. A case was considered as misclassified if the *MySurgeryRisk* algorithm risk score or the physician's assessment of absolute risk classified the patient into a low-risk group for positive cases, where the complication was observed or classified into the high-risk group for negative cases, where the complication was not observed. Thresholds separating low-risk and high-risk groups were 0.32 for intensive care unit admission greater than 48 hours, 0.26 for acute kidney injury, 0.13 for mechanical ventilation greater than 48 hours, 0.07 for cardiovascular complications, 0.05 for severe sepsis, and 0.034 for 30-day mortality.
* *P* < .05 using the Fisher exact test. No significant difference (*P* > .05) was observed in proportion of misclassified cases between attending physicians and trainees.

calculated net improvement in reclassification (net percentages of correctly reclassified cases after the interaction with the algorithm) showed a statistically significant improvement for AKI and ICU admission greater than 48 hours, with 12% and 16% cases correctly reclassified, respectively.

Although the study size was too small for formal a comparison, decision-making attitudes as classified by the CRT appear to play a role in physician interaction with the algorithm. Reflective decision makers changed their scores more frequently than intuitive decision makers. This change was most noticeable in cases in which complications did not occur. Half of the physicians completed a written posttest survey, with both Likert scale and free-response questions administered to assess the usability of the *MySurgeryRisk* algorithm in a simulated workflow for the real-time, intelligent decision-support platform. Half of the respondents found the algorithm helpful with the decision-making process, and 25% were neutral (Table VII), whereas the majority listed tablet and Web site–based applications during clinics and ICU rounds as the best way to access the algorithm. Two physicians reported they would use the *MySurgeryRisk* algorithm for counseling patients preoperatively.

Discussion

In this pilot usability study, the validated *MySurgeryRisk* algorithm implemented in a simulated workflow for a real-time intelligent platform predicted postoperative complications with equal or greater accuracy than our sample of physicians using readily available clinical data from the EHR. Of note, physicians were more likely than the algorithm to both underestimate the risk of postoperative complications for cases where complications actually occurred and overestimate risk for cases where complications did not occur. Although lacking in statistical significance for all complications, the interaction with the *MySurgeryRisk* algorithm resulted in a change in the physicians' risk-perceptions and improvement in the AUC and net scores for reclassification for the tested postoperative complications. Establishing users' attention, facilitating information processing, and updating risk perceptions remains a challenge for all types of risk-assessment tools.[6] It appears the algorithm was able to address these challenges, because in a majority of the cases, physicians changed their risk assessments in response to *MySurgeryRisk*. We attribute this success to the trust instilled by the transparent nature of the *MySurgeryRisk* interface, which highlights important clinical variables contributing to the calculated risks of postoperative complications. The algorithm is deployed currently in a real-time, intelligent platform integrated in the clinical workflow for autonomous surgical risk prediction as a part of a single-center, prospective clinical trial at the UF.[13]

The physicians' abilities to predict postoperative outcomes and the comparison of physicians' risk scores to that of the automated predictive risk scores and systems have not been studied extensively and the existing studies have produced mixed findings.[12,13,15,16,20,21] Among several studies that compared differential diagnosis generators, symptom checkers, and automated electrocardiograms with physicians, the algorithms showed improved accuracy in "less acute" and more "common" scenarios, but in general, physicians had better diagnostic accuracy.[20,21] Studies specific to colorectal and hepatobiliary surgery showed that the

Table V
Change in physicians' perceptions of absolute risk after interaction with *MySurgeryRisk* algorithm

Postoperative complications	Physician changed absolute risk assessment after interaction with *MySurgeryRisk* algorithm, *n* (%)	Physician changed absolute risk assessment closer to *MySurgeryRisk* algorithm, *n* (%)	Difference in physicians' assessment of absolute risk after interaction with *MySurgeryRisk* algorithm, mean (SD)
Intensive care unit admission greater than 48 hours	122 (81)	91 (75)	10 (13)[*]
Acute kidney injury	123 (82)	88 (72)	8 (11)[*]
Mechanical ventilation greater than 48 hours	120 (80)	85 (71)	10 (13)[*]
Cardiovascular complications	124 (83)	94 (76)	8 (10)[*]
Severe sepsis	120 (80)	91 (76)	8 (11)[*]
30-day mortality	114 (76)	83 (73)	7 (10)[*]

Note: The *MySurgeryRisk* algorithm and physicians evaluated absolute risk for complications ranging from 0 (no risk) to 100 (complete certainty of risk). Data represent the number and percentages of cases satisfying the criteria.

[*] $P < .001$ when testing the hypothesis that the mean absolute difference in risk assessment score after interaction with *MySurgeryRisk* algorithm is different than 0 using a *t* test.

Table VI
Comparison between physicians' initial and repeated absolute risk assessments after interaction with *MySurgeryRisk* algorithm

Postoperative complications	Physicians' initial risk assessment AUC (95% CI)	Physicians' risk reassessment AUC (95% CI)	*P* value for difference in AUC	Net reclassification improvement, % (95% CI)
Intensive care unit admission greater than 48 hours	0.69 (0.61–0.77)	0.71 (0.62–0.79)	.452	16.0 (3.0–29.6)[*]
Acute kidney injury	0.65 (0.56–0.74)	0.69 (0.60–0.77)	.064	12.4 (1.0–23.8)[*]
Mechanical ventilation greater than 48 hours	0.66 (0.57–0.75)	0.70 (0.61–0.80)	.074	0.8 (−10.9 to 9.3)
Cardiovascular complications	0.54 (0.44–0.65)	0.59 (0.49–0.69)	.039[†]	5.1 (−2.9 to 13.1)
Severe sepsis	0.54 (0.44–0.64)	0.59 (0.50–0.69)	.063	7.8 (−5.9 to 21.6)
30-day mortality	0.47 (0.36–0.57)	0.49 (0.39–0.60)	.276	−1.0 (−11.6 to 9.6)

CI, confidence interval.
Note: The *MySurgeryRisk* algorithm and physicians evaluated absolute risk for complications ranging from 0 (no risk) to 100 (complete certainty of risk).

[*] *P* value < .05 for testing null hypothesis of net improvement in reclassification being equal to 0 using χ^2 distribution with 1 degree of freedom.

[†] *P* value < .05. The change in AUC for repeated physicians' risk assessment after interaction with *MySurgeryRisk* algorithm was tested using the DeLong test.

Table VII
Participant feedback survey

	Physicians' feedback			
Usability questions	Agree	Neutral	Disagree	
Easy to use	8	1	1	
Helps with decision making	5	2	3	
Would use a version of the algorithm	5	3	2	
Algorithm was helpful and innovative	5	2	3	
	Personal computer	Tablet/phone	Web site	EHR
Best device for this application	1	6	6	1
	Clinic/ICU	Preoperatively	Postoperatively	Home/office
Best location to use the algorithm[*]	7	2	1	1

[*] Based on a summary of free response text.

surgeon's gut feeling was more accurate than the Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity score to predict postoperative mortality.[22,23] Ivanov et al[24] showed that physicians tended to overestimate the risk of postoperative mortality and prolonged ICU stay in patients undergoing coronary artery bypass surgery when compared with a statistical model.[24] Detsky et al[25] showed that the accuracy for ICU physicians' prediction of inhospital mortality, return to home at 6 months, and 6-month cognitive function varied considerably and was only slightly better than random guessing. The computational algorithm as a greater-capacity and lesser-cost information processing service is a logical next step to augment physicians' decision making for rapid identification of patients at risk in the perioperative period.[26] Our algorithm outperformed the physicians' initial risk assessment of postoperative complications (except for cardiovascular complications) in the majority of cases. When examining the rates of misclassification of events and nonevents, physicians tended to overestimate the risks of sentinel events, such as 30-day mortality and cardiovascular events likely attributable to the associated emotional, personal, and professional consequences of failing to recognize these risks.

Physicians' decision-making style can influence their perception of risk and the use of information from decision tools such as algorithms. In spite of our small sample size, we observed that physicians who scored at the extreme of the cognitive reflection test reacted differently in response to interaction with the algorithm. The reasons for this are unknown, but studies of decision-making preferences by Frederick et al[17] suggest that this observation could be attributable to time spent on the risk assessment and the willingness to reassess the decision-making process as assessed by the CRT that measures people's ability to resist their first instinct. A high score indicates a reflective thinker whose initial intuition is tempered by analysis and who takes more time to reflect on the risk probabilities and the information provided by the algorithm. Although our participants had high numeracy scores, it has been demonstrated that even those individuals are likely to make numeric errors on relatively simple questions.[15,16]

Our study has several limitations. Data used for the *MySurgeryRisk* algorithm, although more than sufficient in size to have well-fitting and precise models, were collected from a single center. The results may not be generalizable where patient characteristics differ dramatically. Second, the number of physician participants ($n = 20$) was small and homogeneous, making it difficult to provide a robust statistically significant estimate of physician decision-making preferences based on the CRT and numeric assessment. We did not compare the accuracy of physician risk assessments between subgroups, such as the sex or experience level of the physician, because of the small sample size and the inability to make inference based on those comparisons. Third, it is plausible that physician risk assessment was positively influenced by the *MySurgeryRisk* patient case presentation in summary format, because it is often difficult to find relevant predictive information in the large amount of data in the EHR. Whether an even more optimal way to present numeric clinical data would improve risk assessment should be clarified in future studies. The *MySurgeryRisk* algorithm performance is independent of the enrichment of the testing cohort with positive cases; however, physician assessment may be improved because of the increased number of rare complications and specifically for sentinel events. Although physicians also may or may not have estimated the risks more accurately if they had exposure to an increased amount of patient data, we specifically designed our study to reflect everyday practice and a routine preoperative risk-assessment environment.

A majority of respondents to the posttest survey found the system easy to use, helpful for decision making, and appropriate for the clinical environment. The reasons for physician nonresponse to the written posttest survey are unclear and may be related to physician opinion of the algorithm or simply reflect the lack of time needed to complete the questionnaire. We have integrated the posttest survey into physician use of the algorithm in our current, prospective follow-study to streamline it further and obtain a greater response rate. We continue to further refine the algorithm, allowing participants to input their own assessments into the computational algorithm to facilitate a two-way knowledge transfer and allow models to “learn” from participants. We anticipate expanding our range of complications to allow for a greater personalization specific to individual patients and to include the algorithm risk-assessment scores into the EHR. Ongoing is our larger, prospective clinical evaluation of the algorithm in multiple real-time environments to assess algorithm and participant performance, ease of use in clinical decision-making, and the potential for further decreases in postoperative complications.[13]

The prediction of major postoperative complications is complex and multifactorial. In this pilot study, we have demonstrated that the implementation of a validated, computational algorithm for real-time predictive analytics with EHR data to augment physicians decision-making is feasible and accepted by physicians. Although our study suggests that the low-cost, high-capacity, information-processing power of computational algorithms within an EHR may augment the accuracy of physicians’ risk assessment, larger studies will be needed to confirm this assumption. The implementation of an autonomous platform for real-time analytics and communication with physicians in a perioperative clinical workflow would greatly simplify and augment the perioperative risk assessment and stratification of patients. With the advance of data science and digitalization of medical records, this type of advanced analytics[27,28] is coming of age for perioperative medicine, and early involvement of physicians as key stakeholders in both design and implementation of this technology will be crucial for its success.

Acknowledgments

We would like to acknowledge our research colleagues Dmytro Korenkevych, PhD and Paul Thottakkara who have assisted us in this project with development of algorithm risk scores, and Daniel W Freeman for graphical art. We want to thank Gigi Lipori and University of Florida Integrated Data Repository team for assistance with data retrieval. Meghan Brennan, Sahil Puri, and Tezcan Ozrazgat-Baslanti contributed equally to the manuscript. A.B., T.O.B., P.M., X.L., and D.Z.W. were supported by RO1GM110240 from the NIGMS. A.B. and T.O.B. were supported by Sepsis and Critical Illness Research Center Award P50 GM-111152 from the NIGMS. T.O.B. has received a grant (97071) from Clinical and Translational Science Institute, University of Florida. This work was supported in part by the NIH/NCATS Clinical and Translational Sciences Award to the University of Florida UL1 TR000064. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Univerity of Florida.

References

1. Tepas JJ, Rimar JM, Hsiao AL, Nussbaum MS. Automated analysis of electronic medical record data reflects the pathophysiology of operative complications. *Surgery*. 2013;154:918–924.
2. Hobson C, Ozrazgat-Baslanti T, Kuxhausen A, et al. Cost and mortality associated with postoperative acute kidney injury. *Ann Surg*. 2015;261:1207–1214.
3. Lagu T, Rothberg MB, Shieh MS, Pekow PS, Steingrub JS, Lindenauer PK. Hospitalizations, costs, and outcomes of severe sepsis in the United States 2003 to 2007. *Crit Care Med*. 2012;40:754–761.
4. Silber JH, Rosenbaum PR, Trudeau ME, et al. Changes in prognosis after the first postoperative complication. *Med Care*. 2005;43:122–131.
5. Hobson C, Singhania G, Bihorac A. Acute kidney injury in the surgical patient. *Crit Care Clin*. 2015;31:705–723.
6. Harle C, Padman R, Downs J. The impact of web-based diabetes risk calculators on information processing and risk perceptions. *AMIA Annu Symp Proc*. 2008: 283–287.
7. Copeland GP, Jones D, Walters M. POSSUM: A scoring system for surgical audit. *Br J Surg*. 1991;78:355–360.
8. Gawande AA, Kwaan MR, Regenbogen SE, Lipsiz SA, Zinner MJ. An Apgar score for surgery. *J Am Coll Surg*. 2007;204:201–208.
9. American College of Surgeons. ACS NSQIP Surgical Risk Calculator [Internet]. Chicago, IL: American College of Surgeons National Surgical Quality Improvement Program; 2007 [cited: 2018 Jan 12]. Available from: https://riskcalculator.facs.org/RiskCalculator/.
10. Hollis RH, Graham LA, Lazenby JP, et al. A role for the early warning score in early identification of critical postoperative complications. *Ann Surg*. 2016;263: 918–923.
11. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, et al. MySurgeryRisk: Development and validation of a machine-learning risk algorithm for major complications and death after surgery. *Ann Surg*. 2018. https://doi.org/10.1097/SLA.0000000000002706.
12. Thottakkara P, Ozrazgat-Baslanti T, Hupf BB, et al. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PLoS One*. 2016;11:e0155705.
13. Feng Z, Bhat R, Yuan X, Freeman D, et al. Intelligent perioperative system: Towards real-time big data analytics in surgery risk assessment. *DASC PICom DataCom CyberSciTech*. 2017;2017:1254–1259.
14. United States Census Bureau. American FactFinder [Internet]. Washington, D.C.: United States Census Bureau; 2010 [cited 2018 Jan 12]. Available from: https://factfinder.census.govhttps://factfinder.census.gov.
15. Lipkus IM, Samsa G, Rimer BK. General performance on a numeracy scale among highly educated samples. *Med Decis Making*. 2001;21:37–44.
16. Schwartz LM, Woloshin S, Black WC, Welch HG. The role of numeracy in understanding the benefit of screening mammography. *Ann Intern Med*. 1997;127:966–972.
17. Frederick S. Cognitive reflection and decision making. *J Econ Perspect*. 2005;19: 25–42.
18. Delong ER, Delong DM, Clarkepearson DI. Comparing the areas under 2 or more correlated receiver operating characteristic curves—A nonparametric approach. *Biometrics*. 1988;44:837–845.
19. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011;30:11–21.
20. Poon K, Okin PM, Kligfield P. Diagnostic performance of a computer-based ECG rhythm algorithm. *J Electrocardiol*. 2005;38:235–238.

21. Semigran HL, Levine DM, Nundy S, Mehrotra A. Comparison of physician and computer diagnostic accuracy. *JAMA Intern Med*. 2016;176:1860–1861.
22. Markus PM, Martell J, Leister I, Horstmann O, Brinker J, Becker H. Predicting postoperative morbidity by clinical assessment. *Brit J Surg*. 2005;92: 101–106.
23. Hartley MN, Sagar PM. The surgeon's 'gut feeling' as a predictor of postoperative outcome. *Ann R Coll Surg Engl*. 1994;76:277–278.
24. Ivanov J, Borger MA, David TE, Cohen G, Walton N, Naylor CD. Predictive accuracy study: Comparing a statistical model to clinicians' estimates of outcomes after coronary bypass surgery. *Ann Thorac Surg*. 2000;70:162–168.
25. Detsky ME, Harhay MO, Bayard DF, et al. Discriminative accuracy of physician and nurse predictions for survival and functional outcomes 6 months after an ICU admission. *JAMA*. 2017;317:2187–2195.
26. Liao L, Mark DB. Clinical prediction models: Are we building better mousetraps? *J Am Coll Cardiol*. 2003;42:851–853.
27. Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. *JAMA*. 2018;320:1107–1108.
28. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J Biomed Health Inform*. 2018;22:1589–1604.