Surgical Risk Is Not Linear: Derivation and Validation of a Novel, User-friendly, and Machine-learning-based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) Calculator

Dimitris Bertsimas, PhD, Jack Dunn, PhD,* George C. Velmahos, MD, PhD,† and Haytham M. A. Kaafarani, MD, MPH, FACS†*

Introduction: Most risk assessment tools assume that the impact of risk factors is linear and cumulative. Using novel machine-learning techniques, we sought to design an interactive, nonlinear risk calculator for Emergency Surgery (ES).

Methods: All ES patients in the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) 2007 to 2013 database were included (derivation cohort). Optimal Classification Trees (OCT) were leveraged to train machine-learning algorithms to predict postoperative mortality, morbidity, and 18 specific complications (eg, sepsis, surgical site infection). Unlike classic heuristics (eg, logistic regression), OCT is adaptive and reboots itself with each variable, thus accounting for nonlinear interactions among variables. An application [Predictive OpTimal Trees in Emergency Surgery Risk (POTTER)] was then designed as the algorithms' interactive and user-friendly interface. POTTER performance was measured (c-statistic) using the 2014 ACS-NSQIP database (validation cohort) and compared with the American Society of Anesthesiologists (ASA), Emergency Surgery Score (ESS), and ACS-NSQIP calculators' performance.

Results: Based on 382,960 ES patients, comprehensive decision-making algorithms were derived, and POTTER was created where the provider's answer to a question interactively dictates the subsequent question. For any specific patient, the number of questions needed to predict mortality ranged from 4 to 11. The mortality c-statistic was 0.9162, higher than ASA (0.8743), ESS (0.8910), and ACS (0.8975). The morbidity c-statistics was similarly the highest (0.8414).

Conclusion: POTTER is a highly accurate and user-friendly ES risk calculator with the potential to continuously improve accuracy with ongoing machine-learning. POTTER might prove useful as a tool for bedside preoperative counseling of ES patients and families.

Keywords: artificial intelligence, complication, Emergency General Surgery, Emergency Surgery, machine-learning, morbidity, mortality, Optimal Classification Trees, POTTER, risk calculator, risk prediction

(*Ann Surg* 2018;268:574–583)

The burden of emergency surgical disease has continuously increased over the last 2 decades. Between 2001 and 2010, the United States alone reported >27 million Emergency Surgery (ES) admissions accounting for 7.1% of all hospitalizations.[1] The correlation between ES and adverse outcome has been studied extensively: when compared with similar elective surgery, ES carries a much higher risk of postoperative morbidity and mortality.[2–4] The ability to reliably predict postoperative risk is critical for surgical decision-making, counseling of patients and families, resource allocation, and quality benchmarking. The existent risk stratification models range from the simple and subjective, like the American Society of Anesthesiologists (ASA) classification,[5] to the comprehensive, like the Elixhauser[6] and Charlson[7] Comorbidity Indices. The American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) has also produced its own Surgical Risk Calculator (ACS-SRC).[8] Given that most of these models have been created with the elective surgical patient in mind, many studies have questioned their performance in ES.[9–10] Because of that concern, the Emergency Surgery Score (ESS) was recently suggested as a better predictive model of mortality and morbidity after ES.[11–13] All these aforementioned risk calculators (including ESS), although useful, assume that the variables in their models interact in a linear and additive fashion. The mathematical and medical realities, however, suggest that the interaction of comorbidities and markers of disease acuity are far from linear, and that some variables gain or lose significance due to the absence or presence of other variables.[14] Take, for example, 3 variables which have been repeatedly found to be independent predictors of postoperative mortality: age >70 years, cirrhosis, and use of steroids. In existing, linear, and predictive models, each of these variables is treated as "present" or "absent," and often assigned the same weight irrespective of the presence or absence of the other 2 risk factors. However, it is theoretically possible that, for patients >70 years, cirrhosis plays a role but the use of steroids does not; whereas in patients <70 years, cirrhosis does not play a role but use of steroids does. Therefore, in a nonlinear risk model, the age of the patient would determine whether cirrhosis or steroid use would be included in the prediction of outcomes. The inclusion of 1 of these 2 would then determine the next variable to be included, and this variable could be different for each of the 2 choices. For example, if cirrhosis was chosen, then temperature >100.4 could be the next variable added; if steroid use was chosen, then heart failure could be added. Therefore, in a linear model the surgical risk of these 2 ES patients would be established based on the presence or absence of the same set of variables. In a nonlinear model, the risk could be determined by 2 very different sets of variables. The latter arguably better represents the complexity, interactivity, and nonlinearity of real life.

In this paper, we sought to combine big data from a well-validated, national, surgical database with artificial intelligence (AI) to design and test a novel, interactive, and nonlinear risk calculator for ES. These machine-learning methods, namely, Optimal Classification Trees (OCT) and Optimal Imputation, promise a higher degree of accuracy, interpretability, and automatic integration into electronic health records (EHRs). If translated to user-friendly applications, they may be of real-time assistance to surgeons by the bedside.

From the *Massachusetts Institute of Technology, Cambridge, MA; and †Massachusetts General Hospital & Harvard Medical School, Boston, MA.

The authors report no conflicts of interest.

Reprints: Haytham M. A. Kaafarani, MD, MPH, FACS, Harvard Medical School, Division of Trauma, Emergency Surgery and Surgical Critical Care, 165 Cambridge Street, Suite 810, Boston, MA 02114.
E-mail: hkaafarani@mgh.harvard.edu.

ISSN: 0003-4932/18/26804-0574
DOI: 10.1097/SLA.0000000000002956

METHODS

Patient Population: Derivation and Validation Cohorts

We used the entire ACS-NSQIP 2007 to 2013 dataset for model derivation and algorithm training. The 2014 ACS-NSQIP dataset was used for model testing and validation. The dataset includes >150 preoperative, intraoperative, and postoperative variables.[15–16]

Data Variables

The preoperative variables were used to design our models, whereas the postoperative variables were used as the dependent variables or outcomes to predict. We restricted the dataset to those patients who underwent ES, indicated by the ‘‘Emergency‘‘ variable. We excluded variables that were not consistently collected between 2007 and 2014. We removed the ICD-9 and CPT codes because they are often unknown preoperatively and their complexity compromises use by the surgeon at the bedside. Also, their initial inclusion did not improve the model performance in preliminary testing. We left numerical variables, such as laboratory results, in their raw numeric form. When possible, we converted categorical variables and scales into numeric or ordinal ones (such as functional status and wound classification) to enhance model building. The ASA classification as a preoperative variable requires physician evaluation and arguably subjective judgment, and thus we opted to design 2 decision models, one without (OCT1) and one with (OCT2) the ASA variable. Our primary outcome, mortality, was formed from the ‘‘Days from Operation to Death’’ variable, with null values interpreted as survival and all others as death. Similarly, we formed 18 separate OCT algorithms and models to predict the 30-day postoperative ACS-NSQIP complications (eg, surgical site infection, postoperative pulmonary embolism, postoperative acute renal failure).

Optimal Imputation

A significant number of values in the ACS-NSQIP dataset are missing. We imputed missing values using a recently developed and novel machine-learning method called Optimal Impute,[17] which formulates the imputation task as a family of optimization problems. Imputing the missing values in this way before building predictive models has been shown in multiple real-world datasets to lead to significant improvements in prediction accuracy compared with classical missing values imputation methods.

Machine-learning OCT

To create our AI-based decision-trees, we used a recently developed innovative machine-learning method called OCT.[18] Through OCT, we produced a set of predictive models for 30-day postoperative mortality, morbidity, and each one of the 18 individual postoperative complications of the ACS-NSQIP. We trained a separate decision-tree for each of the above postoperative outcomes. The OCT method is adaptive and reboots itself with each variable, accounting for nonlinear interactions among variables. Beyond its promise for higher accuracy, OCT also increases interpretability due to its tree structure which allows predictions through a few decision splits on a small number of high-importance variables, a characteristic not shared by other machine-learning methods such as neural networks or gradient boosted decision-trees, which are more opaque ‘‘black box" methods. Classical decision-tree methods typically cannot achieve the same level of accuracy as machine-learning methods. However, the early AI machine-learning trees often suffered from limited interpretability. Our novel OCT methodology is a recent advance in AI and machine-learning that trains a single-decision-tree, permitting high-accuracy predictions without sacrificing interpretability.[18] This high level of accuracy is achieved by leveraging modern optimization techniques to train decision-trees from the perspective of global optimality rather than using greedy heuristics like the classical methods.

To better understand OCT, an example of a decision-tree that estimates the risk of any complication (including mortality) after ES is displayed in Figure 1. The actual decision-tree is far more comprehensive but has been limited in this example to a maximum depth of 4 nodes for display purposes. The root node of the tree shows that there are approximately 313,000 patients in the dataset, and the overall risk of mortality or other complication is around 25%. The next decision-tree split refers to transfusion in the 72 hours before surgery. If none occurred, the algorithm leads to the left branch of the tree. There are 302,000 patients following this path, with an updated risk of 23%. If a transfusion occurred, the algorithm leads to the right branch of the tree, which analyzes 10,000 patients with transfusion and estimates the updated risk to 76%. The tree proceeds to further split both sides of the initial branch, and after each new split, the risk is recalculated. Importantly, the preoperative variables used by the tree are not the same at each level; the questions asked change based on the responses at the prior node. In this way, decision-trees can capture nonlinear interactions between variables rather than mandate that the variables interact in a linear and additive fashion, as classical logistic regression does.

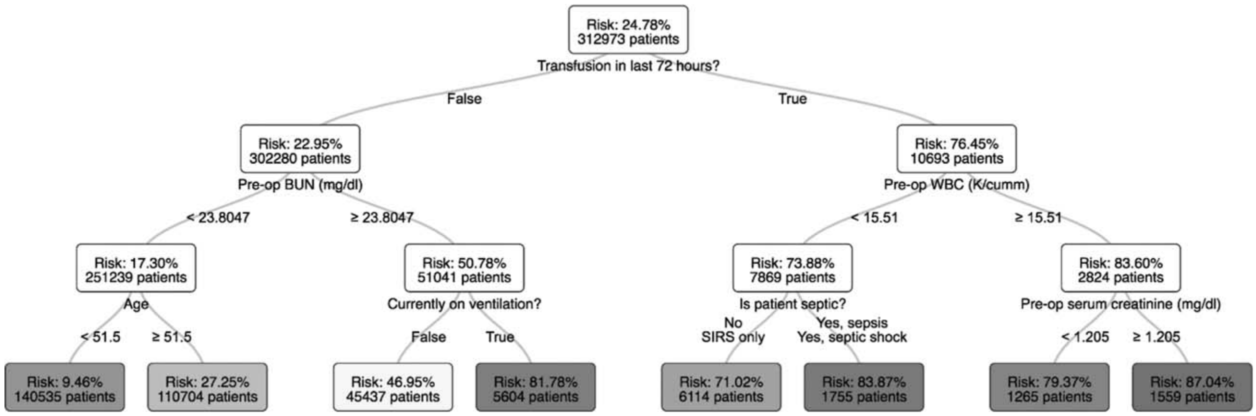


FIGURE 1. An illustrative example of a segment of a decision-tree to predict any complication (including mortality).

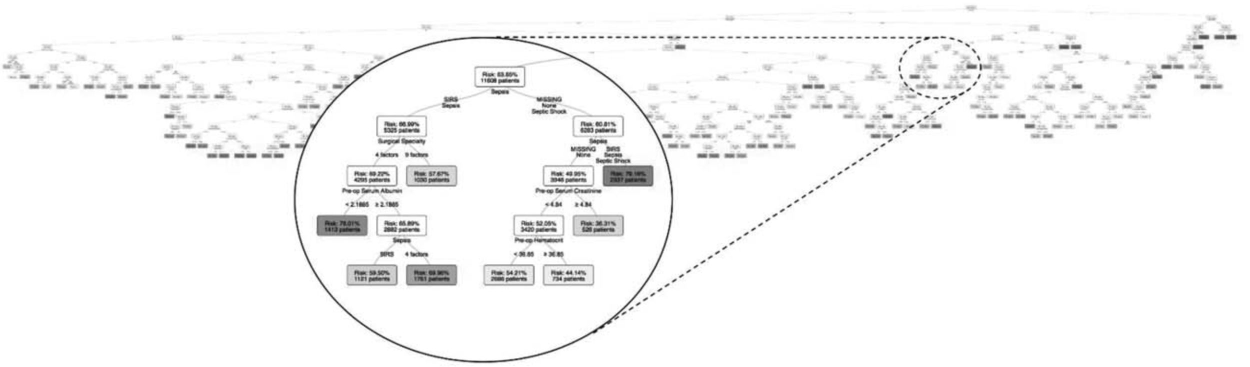


FIGURE 2. The comprehensive decision-tree to predict postoperative 30-day mortality with a random zoom in of one of its trees.

Measurement of Model Performance

The OCT algorithm performance and its ability to predict 30-day postoperative mortality, morbidity, and each of 18 postoperative complications was measured using the c-statistic, also known as the area under the curve (AUC). The c-statistic measures the ability of a model to discriminate between the outcomes of interest and has been used as a measure of model success in multiple prior risk-scoring development efforts.[8,19–21] The performance of our OCT models in predicting mortality was measured in the derivation, validation, and the entire ACS-NSQIP dataset against the performance of the ASA, ACS-SRC, and ESS. For the ACS-SRC comparison, we used the predictions of morbidity and mortality included in the ACS-NSQIP data under the MORBPROB and MORTPROB fields, respectively.

The User-friendly Interface

Using our trained decision-trees for each outcome, we built a dynamic online and phone application as the user-friendly interface of the algorithms for direct use by providers. The application presents the results in the form of an interactive questionnaire. The surgeon user, dealing with a specific ES patient in mind, is initially prompted to select the outcome for which the risk needs to be estimated. The user is then asked a series of simple questions about the presence or absence of certain preoperative variables. Like the machine-learning OCT algorithms, the questions are adaptive; the subject of each new question depends on the specific answer to the prior question. When all the questions are answered, the end-user receives the final risk estimate of the selected outcome for the particular patient. The application design allows easy integration into any EHR environment, so that it can potentially pull most available variables directly from the EHR database in an automated fashion. Once integrated into the EHR, the user would only be required to answer questions that cannot be pulled in automatically. If there is full EHR automation, the risk would be calculated at once.

Ethical Oversight

Institutional Review Board approval for the study was obtained.

RESULTS

Out of a total of 382,960 ES patients in the derivation cohort, comprehensive decision-making algorithms were derived, and a user-friendly application, the Predictive OpTimal Trees in Emergency Surgery Risk (POTTER), was created.

OCT and 30-day Postoperative Mortality

The OCT decision algorithms for 30-day mortality are depicted in Figure 2. Table 1 compares the performance of our OCT models on the derivation, validation, and entire datasets, compared with that of ASA, ESS, and the ACS-SRC. The OCT1 model does not use the ASA classification, whereas the OCT2 does. The ACS-SRC was not included in the derivation/validation cohorts' comparison; only the final predictions of the model were available and there was no ability to calibrate the model to the derivation set and evaluate out-of-sample on the validation set. In summary, the mortality c-statistic for the OCT2 algorithms was the highest at 0.9199, outperforming the ASA (0.8743), ESS (0.8910), and ACS-SRC (0.8979). In the validation dataset, the accuracy of the OCT1 and OCT2 models in detecting a risk of mortality $>50\%$ was 0.9533 and 0.9535, respectively.

OCT and 30-day Postoperative Morbidity

The OCT decision algorithms for 30-day postoperative morbidity (ie, the occurrence of any complication) are depicted in Figure 3. Table 2 compares the performance of our OCT models in predicting 30-day morbidity on the derivation and validation datasets, compared to that of ASA, ESS and ACS-SRC. In summary, the c-statistic of the OCT2 algorithms was the highest at 0.8414, outperforming the ASA (0.7842), ESS (0.7768), and the ACS-SRC (0.8063).

OCT and the 30-day Individual Postoperative Complications

Except for superficial surgical site infection, the OCT1 and OCT2 algorithms predicted the occurrence of individual 30-day postoperative complications with a moderate to extremely high accuracy (c-statistic range from 0.7358 to 0.9338) (Table 3). Their best performances were in predicting postoperative septic shock (c-statistic 0.9338), postoperative ventilator dependence for

TABLE 1. The Performance of Optimal Classification Trees (OCT) in the Predicting 30-day Postoperative Mortality, as Compared With Other Known Risk-Prediction Models

Model	Derivation Cohort	Validation Cohort	Entire Cohort
OCT1	0.9125	0.9064	0.8979
OCT2	0.9233	0.9199	0.9162
ASA	0.8743	0.8740	0.8743
ESS	0.8922	0.8838	0.8910
ACS-SRC	N/A	N/A	0.8975

ASA indicates American Society of Anesthesia; ASC-SRC, American College of Surgeons Surgical Risk Calculator; ESS, Emergency Surgery Score; OCT1, Optimal Classification Trees, excluding ASA; OCT2, Optimal Classification Trees, including ASA.

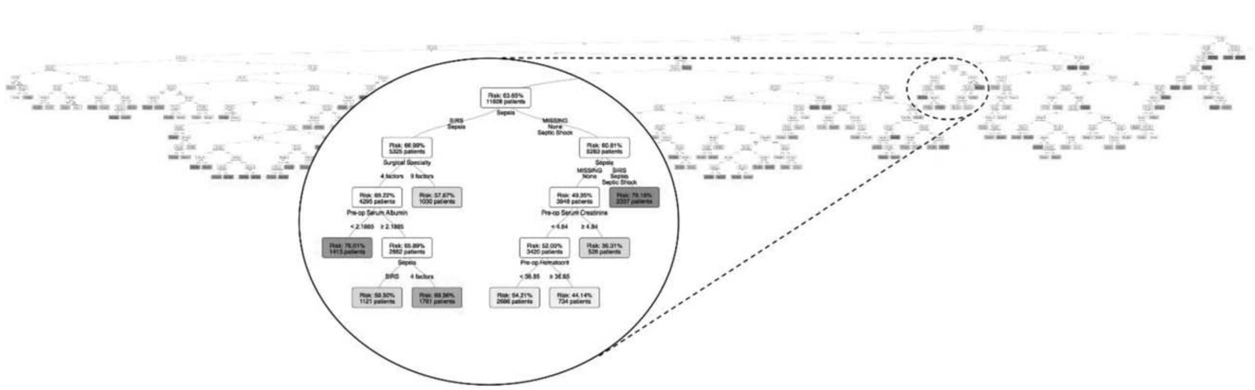


FIGURE 3. The comprehensive decision-tree to predict postoperative 30-day morbidity.

longer than 48 hours (0.9254), and postoperative renal failure (0.9126).

The Interface: POTTER

Using OCT1 and OCT2, a user-friendly, interactive, and comprehensive online and phone application was designed as the algorithms' end-user interface. With the POTTER calculator application, now available for free download in both android and iphone online stores, the provider's answer to a question interactively dictates the subsequent question. For any specific patient, the provider can predict the risk of 30-day postoperative mortality, overall postoperative morbidity or each of 18 individual postoperative complications, such as renal failure, respiratory failure, myocardial infarction, or deep vein thrombosis. For any specific patient, the number of questions needed to predict mortality ranged from 4 to 11, and a specific complication from 3 to 20. These numbers corresponded to the same number of ''clicks'' and typically consumed <1 minute. Figure 3 illustrates the simplicity and power of POTTER, where 2 screenshots depict how the change of answer to one question takes the decision-tree in a totally different direction with different questions and different variables required to predict the final risk of postoperative mortality. Figure 4 shows the almost completely different questions, variables, and decision-trees needed to predict different outcomes, in this case postoperative renal failure versus postoperative myocardial infarction Figure 5.

DISCUSSION

Combining the power of big data and the innovative logic of AI, we have designed POTTER, a novel calculator for the ES patient. As the interface of the OCT algorithms, POTTER offers the advantages of being (1) evidence-based, (2) accurate, (3) nonlinear/machine-learning-based, (4) user-friendly/interactive, (5) amenable to integration into existing EHR, and (6) potentially actionable.

Evidence-based: POTTER is evidence-based because its OCT algorithms do not rely on any modeling assumptions, but are completely derived from patient level data including actual patient outcomes. In this case, the data used for both the derivation and the validation are from the national ACS-NSQIP database, arguably the largest, most reliable, and best validated database in surgery.[22–24] Several studies have suggested the superior accuracy of the ACS-NSQIP database to administrative databases such as the Nationwide Inpatient Sample or insurance claims databases.[25–29]

Accurate: Despite the significantly fewer number of questions needed to estimate the postoperative risk of a specific patient, its accuracy in predicting 30-day postoperative outcome, as measured

TABLE 2. The Performance of Optimal Classification Trees (OCT) in the Predicting 30-day Postoperative Morbidity, as Compared With Other Known Risk-Prediction Models

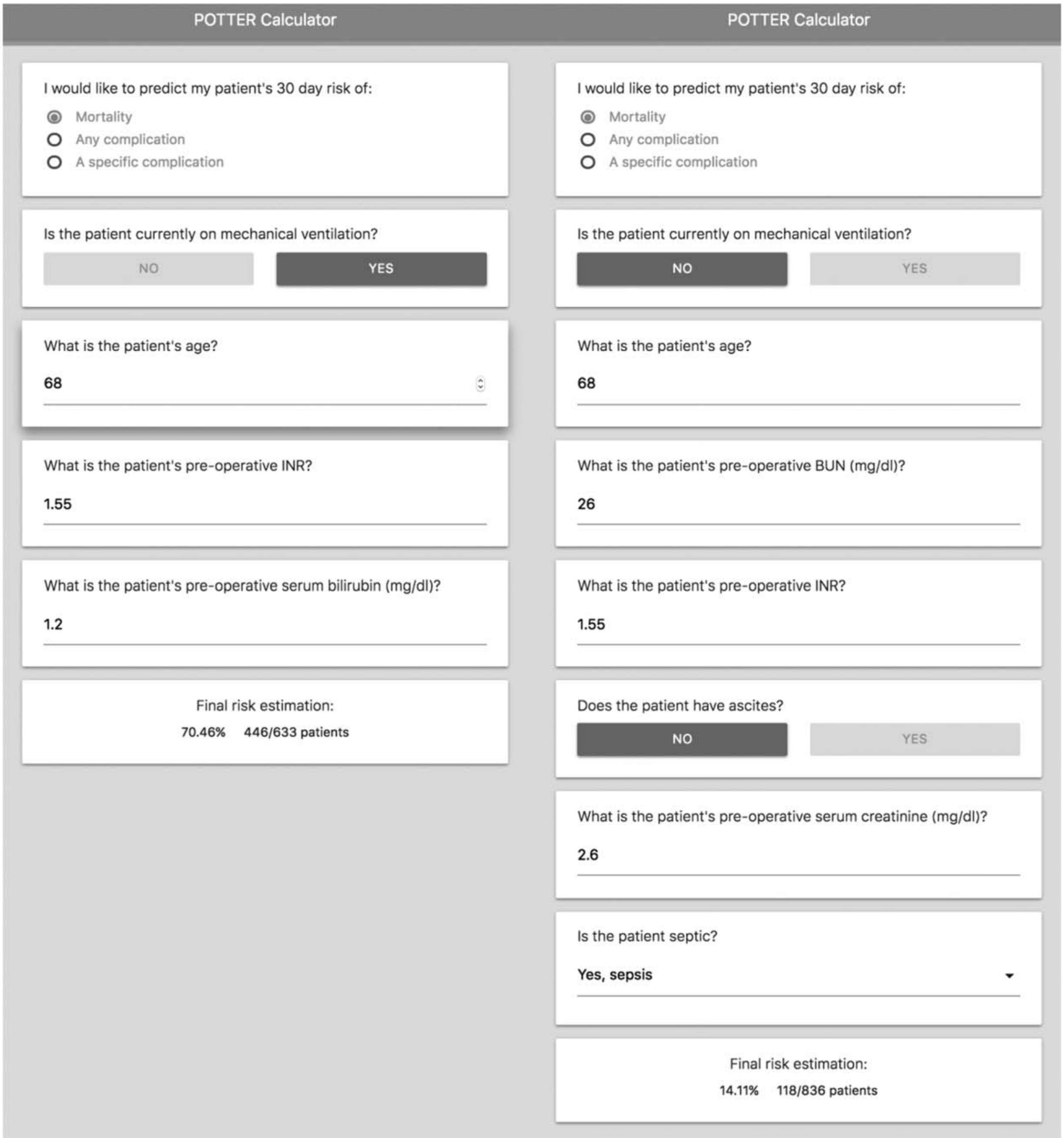
Model	Derivation Cohort (2007–2013)	Validation Cohort (2014)	Entire Cohort
OCT1	0.8366	0.8397	0.8187
OCT2	0.8471	0.8511	0.8414
ASA	0.7884	0.7673	0.7842
ESS	0.7906	0.7715	0.7768
ACS-SRC	N/A	N/A	0.8063

ASA indicates American Society of Anesthesia; ASC-SRC, American College of Surgeons Surgical Risk Calculator; ESS, Emergency Surgery Score; OCT1, Optimal Classification Trees, excluding ASA; OCT2, Optimal Classification Trees, including ASA.

TABLE 3. The Performance of Optimal Classification Trees (OCT) in the Predicting Individual 30-day Postoperative Complications

	Derivation Cohort		Validation Cohort	
Complication	OCT1	OCT2	OCT1	OCT2
Superficial SSI	0.6804	0.6859	0.6762	0.6808
Deep incisional SSI	0.7358	0.7446	0.7405	0.7540
Pulmonary embolism	0.7470	0.7595	0.7196	0.7333
Organ space SSI	0.7723	0.7789	0.7828	0.7860
Sepsis	0.7744	0.7860	0.8444	0.8448
Wound disruption	0.7749	0.7891	0.7689	0.7790
Urinary tract infection	0.7766	0.7778	0.7378	0.7396
DVT/thrombophlebitis	0.7995	0.8129	0.7787	0.7886
Progressive renal insufficiency	0.8315	0.8353	0.8210	0.8188
Myocardial infarction	0.8343	0.8467	0.8151	0.8240
Pneumonia	0.8365	0.8432	0.8364	0.8470
Unplanned intubation	0.8381	0.8462	0.8402	0.8493
Stroke/CVA	0.8536	0.8590	0.8300	0.8343
Cardiac arrest requiring CPR	0.8661	0.8838	0.8722	0.8882
Septic shock	0.8808	0.8888	0.9204	0.9338
Bleeding requiring transfusions	0.8969	0.8984	0.8974	0.9028
Acute renal failure	0.9002	0.9107	0.9025	0.9126
On ventilator >48 h	0.9094	0.9210	0.9107	0.9254

CPR indicates cardiopulmonary resuscitation; CVA, cerebrovascular accident; DVT, deep vein thrombosis; SSI, surgical site infection.



I would like to predict my patient's 30 day risk of:

● Mortality
○ Any complication
○ A specific complication

Is the patient currently on mechanical ventilation?

NO [✓ YES]

What is the patient's age?

68

What is the patient's pre-operative INR?

1.55

What is the patient's pre-operative serum bilirubin (mg/dl)?

1.2

Final risk estimation:

70.46% 446/633 patients

I would like to predict my patient's 30 day risk of:

● Mortality
○ Any complication
○ A specific complication

Is the patient currently on mechanical ventilation?

[✓ NO] YES

What is the patient's age?

68

What is the patient's pre-operative BUN (mg/dl)?

26

What is the patient's pre-operative INR?

1.55

Does the patient have ascites?

[✓ NO] YES

What is the patient's pre-operative serum creatinine (mg/dl)?

2.6

Is the patient septic?

Yes, sepsis

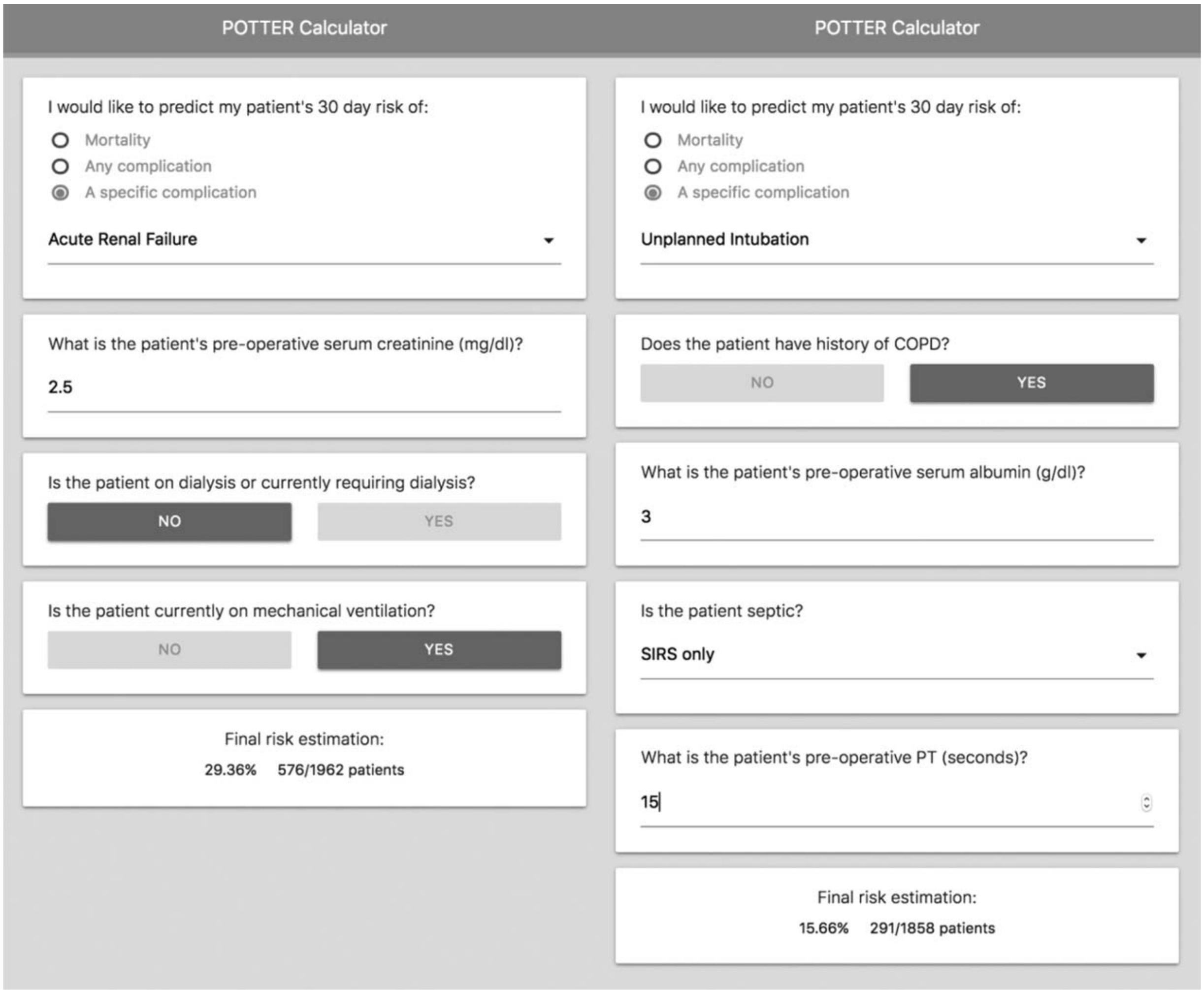
Final risk estimation:

14.11% 118/836 patients

FIGURE 4. An example illustrating how POTTER is interactive, and the answer to a question dictates the next question. In this specific example, whether the provider answers yes to no to the question regarding mechanical ventilation takes the algorithm and questions in a different direction.

using the AUC and c-statistics, was significantly higher than the currently existing methods such as the ASA, ESS, and the ACS-SRC.

Nonlinear and machine-learning-based: Machine-learning is an application of AI where machines are enabled to recognize patterns and learn from their own experiences without being explicitly programmed to do so.[30–32] It is particularly useful to detect subtle intervariable complex relationships that are typically imperceivable to the human eye or mind. The current literature suggests that machine-learning algorithms in general and Classification and Regression Trees (CARTs) in particular can significantly improve the accuracy of classical risk prediction models based on multivariable analyses.[33–35] This is due to the fact that surgical risk is simply not linear, and the impact of a certain variable is dependent on the absence or presence of another variable upstream along the decision-tree. However, CART takes a top-down approach to determining the partitions: starting from the root node, a split is determined by solving an optimization problem before proceeding to repeat the efforts at the level of the 2 resulting new nodes. Such a top-down



POTTER Calculator

I would like to predict my patient's 30 day risk of:
○ Mortality
○ Any complication
● A specific complication

Acute Renal Failure

What is the patient's pre-operative serum creatinine (mg/dl)?
2.5

Is the patient on dialysis or currently requiring dialysis?
[✓ NO] YES

Is the patient currently on mechanical ventilation?
NO [✓ YES]

Final risk estimation:
29.36% 576/1962 patients

POTTER Calculator

I would like to predict my patient's 30 day risk of:
○ Mortality
○ Any complication
● A specific complication

Unplanned Intubation

Does the patient have history of COPD?
NO [✓ YES]

What is the patient's pre-operative serum albumin (g/dl)?
3

Is the patient septic?
SIRS only

What is the patient's pre-operative PT (seconds)?
15

Final risk estimation:
15.66% 291/1858 patients

FIGURE 5. An example illustrating how POTTER interactively uses completely different algorithms, and thus different questions to predict different postoperative complications. In this specific example, we see different questions needed to predict the risk of developing postoperative acute renal failure versus requiring an unplanned intubation.

approach has been criticized because each tree split is determined in isolation without reconsidering the possible impact of future splits in the tree, and typically in practice leads to decision-trees having worse performance than alternative methods. In contrast, the OCT methodology used here and recently developed and validated by our team, constructs the entire decision-tree in a single step, yielding the single best decision-tree for the training data.[18] OCT has been suggested to outperform the accuracy of CART or Random Forest techniques by up to 7%.[18]

Interactive and user-friendly: Because of the complexity of the OCT decision-trees, we have created an interactive interface that starts by asking the providers what outcome they would like to predict on a specific patient. Through a series of short, specific questions where one chooses a yes/no answer or an answer from a drop-down menu, or simply enters an actual laboratory value, the provider quickly receives a specific percentage of risk, sometimes with as little as 3 questions. As a result of the machine-learning methodology used, the risk model is interactive in real time. A provider's answer to the first question will dictate what the next one will be, the answer to the second question will dictate the third, and so on. Each interaction with the application corresponds to a unique decision-tree node and is based on the specific patient characteristics and selected outcome (mortality, morbidity, or a specific complication).

Amenable to integration with HER: We have designed POTTER to be easily amenable to integration into an EHR environment, so that many of the answers can be pulled automatically from the EHR. In an advanced EHR, one can envision the immediate and automated generation of multiple risk estimates for mortality, morbidity, and the specific complications. Prior studies have shown promise in the ability of EHR-integrated, machine-learning algorithms to aid bedside decision-making.

Actionable: POTTER can equip surgeons with personalized and highly accurate risk estimates that will allow them to counsel ES patients and families before surgery. Such information might give the objective data needed to forgo surgery in the patient with little risk for meaningful survival instead of the surgeon having to rely on gestalt or preference.[36–37] Even if surgery will be pursued, using specific risk estimates of mortality and morbidity helps set the right expectations for recovery and what its journey entails. For example, the surgeon might choose to estimate the risk of respiratory failure in the COPD patient while stressing the risk of myocardial infarction in the patient with severe heart disease. In addition, only a small subset of the variables are used in the trees, reducing the need of the physician to know or plug all of a patient's information before receiving a risk estimate. As importantly, it is plausible that POTTER is able to identify "break points" in the early perioperative patient care where a specific clinical care intervention can favorably alter the eventual outcome of a specific patient. Our team has also recently developed another mathematical machine-learning methodology, the *OPT*.[38] OPT is a promising tool that will learn from existing data to recommend/prescribe the best personalized care intervention for each patient that can effectively reduce the risk of postoperative complications or mortality. The challenge remains, of course, in identifying the specific actionable variables that are not mere indicators of the severity of illness or the acuteness of disease, but essential "break point" factors that directly impact patient outcome, if modified in a timely fashion.

Central to the limitations of our study lies the fact that the power of machine-learning prediction depends on the accuracy and comprehensiveness of the data it uses, in this case the ACS-NSQIP database.[14,39] As such, systematic biases resulting from the ACS-NSQIP data collection methodology and its changes over the the multiple years of data might exist. A second issue of our study is the exclusion of ICD and CPT codes from the model. Although it is difficult to accept that the risk of postoperative mortality in ES is not dependent on the diagnosis or the type of surgery performed, our analyses showed otherwise. The inclusion of these codes did not enhance the accuracy of the model, possibly because the type of surgery needed is theoretically reflected in the preoperative derangements that are included in the algorithms. For example, an ES who needs a simple incarcerated inguinal hernia repair might not be showing the same chemical, hematological, and coagulopathic derangements as the patient with perforated viscus or bleeding spleen requiring ES. A third limitation refers to causality between the variables and the outcomes, which is still not proven despite the high degree of connectivity between the 2. Therefore, interpretability and actionability on the relevant variables are controversial. For example, if the mortality decision-tree of a specific patient included a low sodium level, correcting it might not necessarily improve mortality. The decision-tree might simply change in a different direction, and ultimately estimate the same mortality risk.

CONCLUSION

We have developed POTTER, a highly accurate ES risk calculator that outperforms, in accuracy and user-friendliness, all the current existing risk prediction tools. POTTER might prove useful as an evidence-based, adaptive, and interactive tool for bedside preoperative counseling of ES patients and families. Further studies are needed to explore the ability of POTTER to identify actionable "break points" in preoperative patient care that can effectively favorably alter their postoperative outcome.

REFERENCES

1. Gale SC, Shafi S, Dombrovskiy VY, et al. The public health burden of emergency general surgery in the United States: A 10-year analysis of the Nationwide Inpatient Sample—2001 to 2010. *J Trauma Acute Care Surg*. 2014;77:202–208.
2. Ingraham AM, Cohen ME, Bilimoria KY, et al. Comparison of 30-day outcomes after emergency general surgery procedures: potential for targeted improvement. *Surgery*. 2010;148:217–238.
3. Havens JM, Peetz AB, Do WS, et al. The excess morbidity and mortality of emergency general surgery. *J Trauma Acute Care Surg*. 2015;78: 306–311.
4. Havens JM, Do WS, Kaafarani H, et al. Explaining the excess morbidity of emergency general surgery: packed red blood cell and fresh frozen plasma transfusion practices are associated with major complications in nonmassively transfused patients. *Am J Surg*. 2016;211:656–663.
5. Wolters U, Wolf T, Stützer H, et al. ASA classification and perioperative variables as predictors of postoperative outcome. *Br J Anaesth*. 1996;77: 217–222.
6. Elixhauser A, Steiner C, Harris DR, et al. Comorbidity measures for use with administrative data. *Med Care*. 1998;36:8–27.
7. Charlson M, Szatrowski TP, Peterson J, et al. Validation of a combined comorbidity index. *J Clin Epidemiol*. 1994;47:1245–1251.
8. Bilimoria KY, Liu Y, Paruch JL, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *J Am Coll Surg*. 2013;217:833–842.
9. Bohnen JD, Ramly EP, Sangji NF, et al. Perioperative risk factors impact outcomes in emergency versus nonemergency surgery differently: time to separate our national risk-adjustment models? *J Trauma Acute Care Surg*. 2016;81:122–130.
10. Hyder JA, Reznor G, Wakeam E, et al. Risk prediction accuracy differs for emergency versus elective cases in the ACS-NSQIP. *Ann Surg*. 2016;264: 959–965.
11. Sangji NF, Bohnen JD, Ramly EP, et al. Derivation and validation of a novel Emergency Surgery Acuity Score (ESAS). *J Trauma Acute Care Surg*. 2016;81:213–220.
12. Nandan AR, Bohnen JD, Sangji NF, et al. The Emergency Surgery Score (ESS) accurately predicts the occurrence of postoperative complications in emergency surgery patients. *J Trauma Acute Care Surg*. 2017;83:84–89.
13. Peponis T, Bohnen JD, Sangji NF, et al. Does the emergency surgery score accurately predict outcomes in emergent laparotomies? *Surgery*. 2017;162: 445–452.
14. Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N Engl J Med*. 2017;376:2507–2509.
15. ACS NSQIP. Data collection, analysis, and reporting. Available at: https://www.facs.org/quality-programs/acs-nsqip/program-specifics/data. Accessed June 22, 2017.
16. ACS NSQIP. User guide for the 2014 participant use data. American College of Surgeons, 2015. Available at: https://www.facs.org/quality-programs/acs-nsqip/program-specifics/participant-use. Accessed at: March 27, 2018.
17. Bertsimas D, Pawlowski C, Zhuo YD. From predictive methods to missing data imputation: an optimization approach. *Journal of Machine Learning Research*. 2018;18:1–39.
18. Bertsimas D, Dunn J. Optimal classification trees. *Machine Learning*. 2017;106:1039–1082.
19. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29–36.
20. Sangji NF, Bohnen JD, Ramly EP, et al. Derivation and validation of a novel Physiological Emergency Surgery Acuity Score (PESAS). *World J Surg*. 2017;41:1782–1789.
21. Ladha KS, Zhao K, Quraishi SA, et al. The Deyo-Charlson and Elixhauser-van Walraven Comorbidity Indices as predictors of mortality in critically ill patients. *BMJ Open*. 2015;5:e008990.
22. Khuri SF, Daley J, Henderson W, et al. The Department of Veterans Affairs' NSQIP: the first national, validated, outcome-based, risk-adjusted, and peer-controlled program for the measurement and enhancement of the quality of surgical care. National VA Surgical Quality Improvement Program. *Ann Surg*. 1998;228:491–507.
23. Fink AS, Campbell DA Jr, Mentzer RM Jr, et al. The National Surgical Quality Improvement Program in non-veterans administration hospitals: initial demonstration of feasibility. *Ann Surg*. 2002;236:344–353.
24. Hall BL, Hamilton BH, Richards K, et al. Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: an evaluation of all participating hospitals. *Ann Surg*. 2009;250: 363–376.
25. Bohl DD, Basques BA, Golinvaux NS, et al. Nationwide Inpatient Sample and National Surgical Quality Improvement Program give different results in hip fracture studies. *Clin Orthop Relat Res*. 2014;472:1672–1680.
26. Lawson EH, Zingmond DS, Hall BL, et al. Comparison between clinical registry and medicare claims data on the classification of hospital quality of surgical care. *Ann Surg*. 2015;261:290–296.
27. Kaafarani HM, Rosen AK. Using administrative data to identify surgical adverse events: an introduction to the Patient Safety Indicators. *Am J Surg*. 2009;198(5 Suppl.):S63–S68.
28. Bedard NA, Pugely AJ, McHugh MA, et al. Big data and total hip arthroplasty: how do large databases compare? *J Arthroplasty*. 2018;33:41–45. e3.
29. Somani S, Di Capua J, Kim JS, et al. Comparing national inpatient sample and national surgical quality improvement program: an independent risk factor analysis for risk stratification in anterior cervical discectomy and fusion. *Spine (Phila Pa 1976)*. 2017;42:565–572.

30. Carlos RC, Kahn CE, Halabi S. Data science: big data, machine learning, and artificial intelligence. *J Am Coll Radiol*. 2018;15(3 Pt. B):497–498.
31. Syeda-Mahmood T. Role of big data and machine learning in diagnostic decision support in radiology. *J Am Coll Radiol*. 2018;15(3 Pt. B):569–576.
32. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2:230–243.
33. Soguero-Ruiz C, Fei WM, Jenssen R, et al. Data-driven temporal prediction of surgical site infection. *AMIA Annu Symp Proc*. 2015;2015:1164–1173.
34. Wang PS, Walker A, Tsuang M, et al. Strategies for improving comorbidity measures based on Medicare and Medicaid claims data. *J Clin Epidemiol*. 2000;53:571–578.
35. Breiman L, Friedman J, Olshen R, et al. *Classification and Regression Trees*. Wadsworth and Brooks; 1984.
36. Kaafarani HM. Surgeon preference and variation of surgical care. *Am J Surg*. 2011;201:709–711.
37. Kaafarani HM, Hawn MT, Itani KM. Individual surgical decision-making and comparative effectiveness research. *Surgery*. 2012;152:787–789.
38. Bertsimas D, Dunn J, Mundru N. Optimal prescriptive trees. INFORMS Journal on Optimization. In Press
39. Chen JH1, Goldstein MK, Asch SM, et al. Dynamically evolving clinical practices and implications for predicting medical decisions. *Pac Symp Biocomput*. 2016;21:195–206.

DISCUSSANTS

Dr Paul C. Kuo (Tampa, FL):

I would like to thank the Association for the privilege of commenting on this paper, and also my thanks to the authors for sending their manuscript well ahead of time.

In this paper, the authors apply a big data technique to create a predictive model for outcomes after emergency surgery using the ACS NSQIP dataset from 2007 to 2013. They examine a variety of outcomes, including mortality. This dataset includes 382,000 patients and >150 perioperative variables. They use a classification tree methodology (OCT) and an imputation technique (Optimal Impute) for missing variables.

They compare this model with established models such as ASA and ESS. The AUC of their OCT models is better than that of the comparison models. They also present a user application, POTTER, that implements OCT and may be used at the bedside.

I love your title. I agree with the authors that not only is surgical risk not linear, but also life is not linear. As a result of newer techniques, larger datasets, and more powerful computing capacity, we no longer need to limit ourselves to linear approaches to analysis.

In the same way that protein chemistry started with just a linear consideration of amino acid sequences and advanced to considerations of tertiary and quaternary structures, data science and analysis have similarly evolved.

As a big data nerd wannabe, I will confine my comments to methodology. Big data predictive modeling is essentially hypothesis generating. The equivalent of prospective clinical trials is required to determine applicability. In addition, these models, although potentially powerful, suffer from lack of consideration of local environmental factors, such as hospital resources, that might either further enhance performance or, alternatively, indicate overfitting of your model.

I ask, have the authors begun to implement OCT in their own institution to determine performance prospectively? And, certainly, beyond AUC, although that is the common statistic used to compare these kinds of models, it would be nice to see accuracy, sensitivity, and specificity for each of the models.

The methodology used for imputing missing variables is novel, but, unfortunately, the reference is under review. As a simple test of proof of principle, I ask if the authors created a data subset in which known values were deleted, applied their imputation technique, and determined accuracy of Optimal Impute. How does it compare with alternative approaches for missing variables, which are replete throughout the big data literature?

As we understand and implement machine-learning techniques, the approach overall is empiric. It is a tool. Performance trumps elegance. So I ask if the authors have compared OCT with other techniques such as traditional regression, random forest, gradient boosting, neural networks, and so on. Perhaps it is the imputation methodology rather than a classification approach.

This has to do with the paper. I was a bit confused by the description of POTTER. Has it been used yet? The paper refers to POTTER performance. If POTTER has been used in the context of this modeling approach, a clarification in this paper would have been nice. Otherwise, I think OCT performance has been measured and not POTTER.

I am going to throw in a question I did not send to you. That is, if I enter my patient's data into POTTER, do you retain that data?

It would be of interest to see the relevant independent variables and weights that comprise the trees for mortality and the various complications.

Lastly, as a simple rhetorical academic flourish, when a biologist publishes a paper, there is an agreement that requires that the substrate and/or biologics become available to other researchers. For the purposes of a purely academic discussion, do you think publishing your paper would require that you make your code available to your readership?

In closing the paper, I think the paper emphasizes the need to include contemporary methodologies as we address ongoing clinical problems. I congratulate the authors on your work. Very good. Thank you.

Dr Haytham M. Kaafarani:

Thank you, Dr Kuo, for your kind remarks. These are extremely insightful questions. We agree that our data generate as many hypotheses as it answers questions. The advantage of machine-learning techniques resides in the fact that they can continue to improve performance as we add more data, and as the variables become more comprehensive. Although we have recently started at MGH using the POTTER calculator in our daily AM sign out rounds when discussing ES patients, and in the ED when consulting on such patients, we have not yet started a prospective study evaluating its performance. Our team is currently discussing with the hospital leadership adopting it in our EHR and automatically evaluating its performance. A multi-institutional prospective study is also on our agenda.

Regarding sensitivity and specificity: This is not only statistically possible, but also relatively easy to do. The only caveat is that we would need to choose a threshold (eg, if the probability of mortality >5% or >10%). We would be happy to provide that in the manuscript.

Regarding the Optimal Impute methodology: We have indeed tested it across 95 real-world datasets by removing some known values and comparing their imputation technique against other methods. We observed an overall improvement of 10% to 15% in imputation accuracy for Optimal Impute compared to the best of the other methods.

Regarding the OCT methodology itself, this is another great question with a lot of insight into AI in general and machine-learning methodologies in specific. The key advantage of OCT in this case is the interpretability of the method. We sought to develop a risk calculator that was easy for physicians to both use and understand. The decision-tree structure means that few variables (typically 5 to 10) are needed to make a prediction for a patient, whereas for these alternative approaches such as random forests or gradient boosting, we would need to enter values for each of the >150 variables into a black-box methodology that does not make the logic clear in attaining the risk percentage. With OCT, you can follow the logic of the AI as it goes because as it takes you from one question to the other, there is a numerator and a denominator that tells you why this risk is evolving in that specific direction.

The other question you had was about the terminology. POTTER is just simply the interface that we use so people do not have to look at the algorithms and guess the risks of POTTER, and the optimal classification algorithms are one and the same. There is no difference between the 2. I apologize if that was not clear in the manuscript, and we will correct that.

Regarding the weights for each variable, each decision-tree has different variables with different weights because of the OCT methodology. As such, the nonlinear nature of the trees make it hard to come up with meaningful averages because the importance of each variable is so dependent on the previous answers. For example, the weight of diabetes is really dependent on the specific patient and might be different from one patient to the other.

The question regarding publishing the codes: In principle, we totally agree. It would aid in reproducing results or applying approaches to new data sources. The realities are slightly more complicated, introducing concerns such as the quality of the code being published and how usable it is required to be, whether the code is required to be supported or kept up to date by the author as new versions of software come out.

You asked me whether we have the ability to retain the data in our application. As of now, we have not done that. Again, the application was just approved by iPhone and Android 2 weeks ago for Android and last week for iPhone. That is something to look into, so that we can continue to improve the algorithms as multiple people around the country use it.

Dr Henry Pitt (Philadelphia, PA):

I would like to congratulate the authors on applying machine-learning to NSQIP data. I have 2 questions.

A year ago we published a paper in JACS evaluating the ACS NSQIP risk calculator in elective and emergent colorectal patients at Temple. The risk calculator was more accurate in the elective than in the emergent patients, and one of the areas of inaccuracy was the ability of the risk calculator to predict whether a patient would be discharged to a skilled nursing facility (SNF). I did not see that outcome in your data. Thus, my question is, can POTTER predict the ability to be discharged to a SNF? Which actually is very important for some of these end-of-life patients.

Also, at Temple University Hospital we do 100% mortality review. The majority of the patients who have an emergent operation and do not survive have a cardiac or a vascular operation. As you know, the procedures captured in NSQIP are not preferred by cardiac surgeons and vascular surgeons because of the existence of STS and SVS databases. Thus, my second question is whether your POTTER NSQIP data can really be applied to emergent cardiac and vascular surgery patients. Thank you.

Dr Haytham M. Kaafarani:

Thank you for your questions. Addressing the first one: Functional outcome is probably the most important outcome besides mortality. A patient may survive like you could survive, but with a predicted risk of complications as high as 95% or 97%. That suggests that, for those who survive, they might survive with a very poor functional outcome and be discharged to a nursing facility. We have not done the predictive models for discharge venues or functional outcomes, but that is a really good suggestion, and we probably should. Thank you.

To answer your second question, yes, the OCT algorithms can be applied to other databases. They need a lot of patients to distill some of the noise in the data. But if you get us the STS database or other the cardiothoracic, vascular, or transplant databases, we can run the OCT algorithms similarly to what we did. The bigger the data and the more accurate the data and the more comprehensive the data, the better the algorithms will be.

Dr Adil Haider (Boston, MA):

Dr Kaafarani, fantastic presentation and congratulations on getting this through the Apple iPhone app store. I know how difficult it is to get a medical app through the app store, so congratulations.

My question is regarding the area under the curve or the c-statistic that you have used to determine the discriminatory ability of the POTTER score. I noticed that when you look at any complication, your discriminatory ability is actually quite good. But when you look at the individual complications, it's not as good as even the previous scores that you have determined. Can you talk about the ability of this machine-learning algorithm to discriminate between one complication versus multiple complications?

Also, does the POTTER score have anything to do with a certain very famous book, the POTTER piece of that?

Dr Haytham M. Kaafarani:

Let me answer the first question first. You are absolutely correct. The machine-learning methodology really depends on the data that we plug in. If there is a certain baseline variable we are not measuring that affects outcome, then that undermines the c-statistics.

What we found in POTTER, which is amazing, is that for life threatening complications, the performance was really above everything that we know. The c-statistic is impressively high.

However, the worst c-statistic was for superficial surgical site infection. I think it is because the NSQIP itself does not collect enough data to accurately predict surgical site infection. In general, if you look at most of the literature in predicting surgical site infection, you can see the c-statistics tend to be much lower because the data we collect tend to be the data that predicts how patients do overall—whether they live or do not live, and whether they get major complications or do not.

Your second question: I have a 9-year-old who is obsessed and reading Harry Potter over and over and over again. When I was playing with the perioperative optimal classification trees, the letters were matching to include P, O, T, E, and R, and that is the mnemonic that naturally came to my head.

Dr Ari Leppaniemi (Helsinki, Finland):

The concept of failure to rescue has been used, as you know, to compare the outcomes of patients' after complications. Now, the problem with that is sometimes how do you define the denominator? How do you include patients in the analysis?

My question is, how do you use this risk assessment as a tool to determine, for example, that everybody whose risk of dying is >50% would be included in the analysis? And then you actually look at what happened, and, therefore, have a case-mix adjusted calculator of your outcomes.

Second, briefly, I checked the application. Some of the units that are used in the United States, is it available for units to be used in Europe? Like millimoles and stuff like that.

Dr Haytham M. Kaafarani:

Let me start with the second question. That is a good suggestion by our friends from Finland: We should be able to automatically do the conversion to every unit that every country wants to use. It should be a pretty easy I.T. problem to fix.

For your first question, which is the failure to rescue question, can we identify a high-risk population and then we can look at those specific risks? Yes, in the algorithms, we can probably draw certain threshold lines, say, mortality >50%, to identify the final tree nodes that lead to such mortality. Then, we can go back and see those patients, who they are, and we can do the analysis only on this subset of patients. I think it is technically doable.

But I actually think the more fascinating suggestion you had is that regarding looking at the subset of patients who had a complication and analyzing how they die. We are experimenting with another AI tool called optimal prescriptive trees, which can help us identify tree nodes at which interventions can alter the outcome following complications to prevent the patient's clinical situation from spiraling to death.