Assessing the utility of deep neural networks in predicting postoperative surgical complications: a retrospective study

Alexander Bonde, Kartik M Varadarajan, Nicholas Bonde, Anders Troelsen, Orhun K Muratoglu, Henrik Malchau, Anthony D Yang, Hasan Alam, Martin Sillesen

Summary

Background Early detection of postoperative complications, including organ failure, is pivotal in the initiation of targeted treatment strategies aimed at attenuating organ damage. In an era of increasing health-care costs and limited financial resources, identifying surgical patients at a high risk of postoperative complications and providing personalised precision medicine-based treatment strategies provides an obvious pathway for reducing patient morbidity and mortality. We aimed to leverage deep learning to create, through training on structured electronic health-care data, a multilabel deep neural network to predict surgical postoperative complications that would outperform available models in surgical risk prediction.

Methods In this retrospective study, we used data on 58 input features, including demographics, laboratory values, and 30-day postoperative complications, from the American College of Surgeons (ACS) National Surgical Quality Improvement Program database, which collects data from 722 hospitals from around 15 countries. We queried the entire adult (≥18 years) database for patients who had surgery between Jan 1, 2012, and Dec 31, 2018. We then identified all patients who were treated at a large midwestern US academic medical centre, excluded them from the base dataset, and reserved this independent group for final model testing. We then randomly created a training set and a validation set from the remaining cases. We developed three deep neural network models with increasing numbers of input variables and so increasing levels of complexity. Output variables comprised mortality and 18 different postoperative complications. Overall morbidity was defined as any of 16 postoperative complications. Model performance was evaluated on the test set using the area under the receiver operating characteristic curve (AUC) and compared with previous metrics from the ACS-Surgical Risk Calculator (ACS-SRC). We evaluated resistance to changes in the underlying patient population on a subset of the test set, comprising only patients who had emergency surgery. Results were also compared with the Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) calculator.

Findings 5 881 881 surgical patients, with 2941 unique Current Procedural Terminology codes, were included in this study, with 4 694 488 in the training set, 1 173 622 in the validation set, and 13 771 in the test set. The mean AUCs for the validation set were 0·864 (SD 0·053) for model 1, 0·871 (0·055) for model 2, and 0·882 (0·053) for model 3. The mean AUCs for the test set were 0·859 (SD 0·063) for model 1, 0·863 (0·064) for model 2, and 0·874 (0·061) for model 3. The mean AUCs of each model outperformed previously published performance metrics from the ACS-SRC, with a direct correlation between increasing model complexity and performance. Additionally, when tested on a subgroup of patients who had emergency surgery, our models outperformed previously published POTTER metrics.

Interpretation We have developed unified prediction models, based on deep neural networks, for predicting surgical postoperative complications. The models were generally superior to previously published surgical risk prediction tools and appeared robust to changes in the underlying patient population. Deep learning could offer superior approaches to surgical risk prediction in clinical practice.

Funding The Novo Nordisk Foundation.

Copyright © 2021 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY-NC-ND 4.0 license.

Lancet Digit Health 2021; 3: e471–85

Published Online June 29, 2021 https://doi.org/10.1016/S2589-7500(21)00084-4

Department of Surgical Gastroenterology and Transplantation, Rigshospitalet (A Bonde MD, M Sillesen PhD), **Center for Surgical Translational and Artificial Intelligence Research, Rigshospitalet** (A Bonde, M Sillesen), **and Department of Orthopedics** (N Bonde MD, Prof A Troelsen PhD), **Hvidovre Hospital, Copenhagen University Hospital, Copenhagen, Denmark; Harris Orthopedics Laboratory, Massachusetts General Hospital, Boston, MA, USA** (K M Varadarajan PhD, O K Muratoglu PhD, Prof H Malchau PhD); **Department of Orthopaedics, Sahlgrenska University Hospital, Gothenburg, Sweden** (Prof H Malchau); **Surgical Outcomes and Quality Improvement Center, Department of Surgery, Northwestern University Feinberg School of Medicine, Chicago, IL, USA** (A D Yang MD); **Department of Surgery, Northwestern Memorial Hospital, Chicago, IL, USA** (Prof H Alam MD)

Correspondence to: Dr Kartik M Varadarajan, Harris Orthopedics Laboratory, Massachusetts General Hospital, Boston, MA 02114, USA **kmangudivaradarajan@mgh.harvard.edu**

Introduction

Worldwide, approximately 234 million major surgical procedures are done each year,[1] addressing around 11% of the global burden of disease.[2] An estimated 4% of patients die as a direct result of surgery and 15% have a protracted recovery due to complications.[1] In addition to the obvious impact on individual patients, surgical postoperative complications carry a considerable socioeconomic burden, increasing treatment costs by 119–172% compared with uncomplicated recoveries.[3] In an era of increasing health-care costs and limited financial resources, identifying patients at risk of postoperative complications and providing personalised precision medicine-based treatment strategies provides an obvious pathway for reducing patient morbidity and mortality, and health care-related costs in the surgical setting.

Research in context

Evidence before this study

Deep learning is a prominent type of artificial intelligence that has shown notable success in several medical applications.

We searched Google Scholar for articles published in English between database inception and Oct 8, 2020, using the search terms ("deep learning" OR "machine learning" OR "artificial intelligence" AND "postoperative complications"). Our search did not identify any articles that explored multilabel deep neural networks for surgical risk prediction that were trained on multi-institutional data and validated on an independent cohort.

We identified several studies on the development and validation of deep neural networks for surgical risk prediction based on single medical centres, single procedures, or single outcomes. Traditional machine learning techniques, such as logistic regression, K-nearest neighbours, and tree-based methods, were more common in the prediction of postoperative complications than were deep neural networks. We identified no standard public surgery datasets that could be directly compared with our algorithms' performance.

Added value of this study

We developed multilabel deep learning models to predict postoperative complications, which were trained on the largest multi-institutional dataset of surgical cases and validated on an independent cohort to show evidence of potential in clinical practice. We did not exclude rare procedures so as to not inflate our performance metrics. The mean areas under the receiver operating characteristic curve of each of our models outperformed previously published performance metrics, with a direct correlation between increasing model complexity and performance. Our models also retained predictive power despite substantial changes in the underlying patient population.

Implications of all the available evidence

Our deep learning models were superior to previously published surgical risk prediction tools, despite the increasingly rigorous standards for model validation. Our algorithms might be used by clinicians to help guide future preoperative, intraoperative, and postoperative risk management, serving as an important step towards personalised medicine in surgery. A clinical trial is required to identify whether the use of deep learning models can help to reduce the incidence of surgical postoperative complications.

Efficient, precision medicine-based approaches in surgery are, however, fraught with difficulty. Although risk stratification tools have been developed using large groups of surgical patients,[4–6] these tools are often not robust when the analysis is applied to different cohorts.[7] Even large-scale efforts, such as the American College of Surgeons Surgical Risk Calculator (ACS-SRC), developed using data from more than 4·3 million surgical patients,[8] often fail to retain predictive power when used on other cohorts, such as emergency versus elective procedures.[9] Considerable controversy thus still surrounds these predictive models,[10] limiting the transition to clinical practice.

For the **hospitals and countries currently participating in the ACS NSQIP** see https://www.facs.org/search/nsqip-participants?allresults=

Deep learning, a prominent type of artificial intelligence, has shown notable success on unstructured health-care data, with techniques such as computer vision and natural language processing.[11–14] In deep learning, layers of artificial neurons are combined into deep multilayer and non-linear neural networks. These networks can then be trained to model very complex relationships between inputs and outputs. Although deep learning has the potential to transform health care as we know it, this technique remains underexplored on structured health-care data such as that from the ACS National Surgical Quality Improvement Program (ACS NSQIP) database.[15,16] We hypothesise that deep learning can be leveraged to create a state-of-the-art, personalised risk prediction tool for surgical adverse events through training on structured surgical data. Specifically, we aimed to develop a multilabel deep neural network that would outperform the currently available models in surgical risk prediction.

See Online for appendix

Methods

Data source

For this retrospective study, institute review board approval was granted under an expedited review process. Our study used data manually curated by trained and certified surgical clinical reviewers across 722 hospitals across approximately 15 countries participating in the ACS NSQIP. This multi-institutional database contains structured electronic health records data on patients undergoing major surgical procedures within multiple surgical fields, including general, cardiothoracic, orthopaedic, gynaecological, urology, neurosurgery, otolaryngology, and vascular surgery. The ACS NSQIP encompasses more than 150 variables, including demographics, laboratory values, comorbidities, and 30-day postoperative complications. Postoperative complications were defined as per the ACS NSQIP definitions presented in the appendix (p 9). We queried the entire adult (≥18 years) ACS NSQIP database for patients who underwent surgery between Jan 1, 2012, and Dec 31, 2018. The query resulted in a total of 5 881 881 surgical cases that were all included in this study (appendix p 3). We then identified all patients who were treated at a large midwestern US academic medical centre (13 771 [0·23%] of 5 881 881 [the entire adult ACS NSQIP database]). This patient cohort was excluded from the base dataset and reserved for final model testing. We did so by matching selected column values from an independent dataset (a local copy of the data submitted to the ACS NSQIP by the midwestern hospital) to the column values in the overall ACS NSQIP dataset, which allowed us to identify all but

	Training set (n=4 694 488)	Validation set (n=1 173 622)	Test set (n=13 771)
Input variables for model 1			
Current Procedural Terminology codes	2924/2941 (99·4%)	2739/2941 (93·1%)	712/2941 (24·2%)
Age (continuous), years	58 (45–69)	58 (45–69)	59 (47–68)
Height (continuous), inches	66 (63–69)	66 (63–69)	65 (63–69)
Weight (continuous), lb	180 (151–215)	180 (151–215)	177 (147–209)
Sex			
Female	2 672 405/4 694 485 (56·9%)	667 945/1 173 621 (56·9%)	8063/13 771 (58·6%)
Male	2 022 080/4 694 485 (43·1%)	505 676/1 173 621 (43·1%)	5708/13 771 (41·4%)
Functional status			
Independent	4 537 730/4 663 238 (97·3%)	1 134 340/1 165 757 (97·3%)	13 587/13 698 (99·2%)
Partially dependent	103 742/4 663 238 (2·2%)	25 976/1 165 757 (2·2%)	102/13 698 (0·7%)
Totally dependent	21 766/4 663 238 (0·5%)	5441/1 165 757 (0·5%)	9/13 698 (0·1%)
Emergency case			
No	4 270 662/4 694 465 (91·0%)	1 067 775/1 173 618 (91·0%)	13 017/13 771 (94·5%)
Yes	423 803/4 694 465 (9·0%)	105 843/1 173 618 (9·0%)	754/13 771 (5·5%)
ASA physical status class*			
1 (healthy)	411 210/4 682 151 (8·8%)	102 278/1 170 568 (8·7%)	1167/13 718 (8·5%)
2 (mild systemic disease)	2 101 261/4 682 151 (44·9%)	524 764/1 170 568 (44·8%)	7436/13 718 (54·2%)
3 (severe systemic disease)	1 884 113/4 682 151 (40·2%)	471 849/1 170 568 (40·3%)	4734/13 718 (34·5%)
4 (life-threatening severe systemic disease)	277 125/4 682 151 (5·9%)	69 602/1 170 568 (5·9%)	366/13 718 (2·7%)
5 (moribund person)	8442/4 682 151 (0·2%)	2075/1 170 568 (0·2%)	15/13 718 (0·1%)
Steroid use for chronic condition			
No	4 523 819/4 694 486 (96·4%)	1 131 095/1 173 621 (96·4%)	12 820/13 771 (93·1%)
Yes	170 667/4 694 486 (3·6%)	42 526/1 173 621 (3·6%)	951/13 771 (6·9%)
Ascites within 30 preoperative days			
No	4 678 510/4 694 486 (99·7%)	1 169 586/1 173 622 (99·7%)	13 738/13 771 (99·8%)
Yes	15 976/4 694 486 (0·3%)	4036/1 173 622 (0·3%)	33/13 771 (0·2%)
Systemic sepsis within 48 preoperative h			
None	4 439 171/4 694 479 (94·6%)	1 110 186/1 173 621 (94·6%)	13 199/13 771 (95·8%)
Systemic inflammatory response syndrome	140 034/4 694 479 (3·0%)	34 623/1 173 621 (3·0%)	360/13 771 (2·6%)
Sepsis	96 609/4 694 479 (2·1%)	24 198/1 173 621 (2·1%)	161/13 771 (1·2%)
Septic shock	18 665/4 694 479 (0·4%)	4614/1 173 621 (0·4%)	51/13 771 (0·4%)
Ventilator-dependent			
No	4 679 251/4 694 486 (99·7%)	1 169 884/1 173 621 (99·7%)	13 733/13 771 (99·7%)
Yes	15 235/4 694 486 (0·3%)	3737/1 173 621 (0·3%)	38/13 771 (0·3%)
Disseminated cancer			
No	4 587 844/4 694 486 (97·7%)	1 147 050/1 173 622 (97·7%)	13 337/13 771 (96·8%)
Yes	106 642/4 694 486 (2·3%)	26 572/1 173 622 (2·3%)	434/13 771 (3·2%)
Diabetes			
No	3 968 722/4 694 483 (84·5%)	991 726/1 173 622 (84·5%)	11 843/13 771 (86·0%)
Yes (not on insulin)	455 741/4 694 483 (9·7%)	113 795/1 173 622 (9·7%)	1141/13 771 (8·3%)
Yes (on insulin)	270 020/4 694 483 (5·8%)	68 101/1 173 622 (5·8%)	787/13 771 (5·7%)
Hypertension requiring medication			
No	2 588 707/4 694 486 (55·1%)	646 331/1 173 622 (55·1%)	7873/13 771 (57·2%)
Yes	2 105 779/4 694 486 (44·9%)	527 291/1 173 622 (44·9%)	5898/13 771 (42·8%)
Congestive heart failure in 30 preoperative days			
No	4 654 152/4 694 486 (99·1%)	1 163 661/1 173 622 (99·2%)	13 725/13 771 (99·7%)
Yes	40 334/4 694 486 (0·9%)	9961/1 173 622 (0·8%)	46/13 771 (0·3%)

(Table 1 continues on next page)

	Training set (n=4 694 488)	Validation set (n=1 173 622)	Test set (n=13 771)
(Continued from previous page)			
Dyspnoea			
No	4 432 091/4 694 481 (94·4%)	1 107 779/1 173 622 (94·4%)	13 619/13 771 (98·9%)
Yes (on moderate exertion)	241 555/4 694 481 (5·1%)	60 650/1 173 622 (5·2%)	148/13 771 (1·1%)
Yes (at rest)	20 835/4 694 481 (0·4%)	5193/1 173 622 (0·4%)	4/13 771 (<0·1%)
Current smoker (within 1 year of surgery)			
No	3 871 599/4 694 486 (82·5%)	967 551/1 173 621 (82·4%)	12 054/13 771 (87·5%)
Yes	822 887/4 694 486 (17·5%)	206 070/1 173 621 (17·6%)	1717/13 771 (12·5%)
History of COPD			
No	4 486 535/4 694 486 (95·6%)	1 121 734/1 173 622 (95·6%)	13 428/13 771 (97·5%)
Yes	207 951/4 694 486 (4·4%)	51 888/1 173 622 (4·4%)	343/13 771 (2·5%)
Dialysis			
No	4 632 265/4 694 486 (98·7%)	1 157 869/1 173 622 (98·7%)	13 588/13 771 (98·7%)
Yes	62 221/4 694 486 (1·3%)	15 753/1 173 622 (1·3%)	183/13 771 (1·3%)
Acute renal failure			
No	4 678 096/4 694 486 (99·7%)	1 169 622/1 173 621 (99·7%)	13 747/13 771 (99·8%)
Yes	16 390/4 694 486 (0·3%)	3999/1 173 621 (0·3%)	24/13 771 (0·2%)
Operation year			
2012	433 836/4 694 488 (9·2%)	108 195/1 173 622 (9·2%)	1854/13 771 (13·5%)
2013	519 551/4 694 488 (11·1%)	130 008/1 173 622 (11·1%)	2381/13 771 (17·3%)
2014	598 512/4 694 488 (12·7%)	150 175/1 173 622 (12·8%)	2250/13 771 (16·3%)
2015	707 057/4 694 488 (15·1%)	176 586/1 173 622 (15·0%)	1859/13 771 (13·5%)
2016	798 349/4 694 488 (17·0%)	200 145/1 173 622 (17·1%)	1899/13 771 (13·8%)
2017	821 553/4 694 488 (17·5%)	205 330/1 173 622 (17·5%)	1830/13 771 (13·3%)
2018	815 630/4 694 488 (17·4%)	203 183/1 173 622 (17·3%)	1698/13 771 (12·3%)
Additional input variables for model 2			
Preoperative serum sodium concentration (continuous), mEq/L	139 (137–141)	139 (137–141)	138 (136–140)
Preoperative blood urea nitrogen concentration (continuous), mg/dL	15 (11–19)	15 (11–19)	15 (12–20)
Preoperative serum creatinine concentration (continuous), mg/dL	0·85 (0·70–1·02)	0·85 (0·70–1·02)	0·87 (0·72–1·05)
Preoperative serum albumin concentration (continuous), g/dL	4·0 (3·6–4·3)	4·0 (3·6–4·3)	4·1 (3·7–4·4)
Preoperative total bilirubin concentration (continuous), mg/dL	0·5 (0·4–0·7)	0·5 (0·4–0·7)	0·5 (0·4–0·7)
Preoperative serum GOT concentration (continuous), U/L	21 (17–28)	21 (17–28)	20 (16–26)
Preoperative alkaline phosphatase concentration (continuous), U/L	77 (62–97)	77 (62–97)	71 (56–91)
Preoperative white blood cell count (continuous), 10^3 per µL	7·3 (5·9–9·4)	7·3 (5·9–9·4)	6·8 (5·6–8·6)
Preoperative haematocrit (continuous), %	40·0% (36·8–43·0)	40·0% (36·7–43·0)	39·6% (36·1–42·6)
Preoperative platelet count (continuous), 10^3 per µL	240 (198–290)	240 (198–290)	251 (206–305)
Preoperative PTT (continuous), s	29·1 (26·9–32·1)	29·1 (26·9–32·1)	30·2 (28·0–32·6)
Preoperative INR (continuous)	1·0 (1·0–1·1)	1·0 (1·0–1·1)	1·0 (1·0–1·1)
Preoperative prothrombin time (continuous), s	12·3 (10·9–13·6)	12·3 (10·9–13·6)	..

(Table 1 continues on next page)

three patients, who, for some reason, were not present in the base set, and exclude them from model training and validation. After separating the 13 771 patients for the test set, we randomly split the remaining 5 868 110 patients into a training set (n=4 694 488; 79·81% of 5 881 881) and a validation set (n=1 173 622; 19·95% of 5 881 881). The training dataset was used for preliminary testing, model development, and training. The validation set was used

	Training set (n=4 694 488)	Validation set (n=1 173 622)	Test set (n=13 771)
(Continued from previous page)			
Additional input variables for model 3			
Race			
White	3 360 754/4 005 071 (83·9%)	839 295/1 001 259 (83·8%)	9 910/12 705 (78·0%)
Black or African American	466 233/4 005 071 (11·6%)	117 453/1 001 259 (11·7%)	2 327/12 705 (18·3%)
Asian	133 069/4 005 071 (3·3%)	33 307/1 001 259 (3·3%)	434/12 705 (3·4%)
American Indian or Alaska Native	26 772/4 005 071 (0·7%)	6 673/1 001 259 (0·7%)	29/12 705 (0·2%)
Native Hawaiian or Pacific Islander	18 243/4 005 071 (0·5%)	4 531/1 001 259 (0·5%)	5/12 705 (<0·1%)
Hispanic ethnicity			
No	3 660 496/4 019 863 (91·1%)	915 006/1 004 740 (91·1%)	11 885/12 908 (92·1%)
Yes	359 367/4 019 863 (8·9%)	89 734/1 004 740 (8·9%)	1023/12 908 (7·9%)
Principal anaesthesia technique			
Epidural	6610/4 693 592 (0·1%)	1616/1 173 395 (0·1%)	277/13 771 (2·0%)
General	4 186 692/4 693 592 (89·2%)	1 046 759/1 173 395 (89·2%)	10 490/13 771 (76·2%)
Local	11 107/4 693 592 (0·2%)	2733/1 173 395 (0·2%)	28/13 771 (0·2%)
Monitored anaesthesia care (intravenous sedation)	218 927/4 693 592 (4·7%)	54 644/1 173 395 (4·7%)	1344/13 771 (9·8%)
None	796/4 693 592 (<0·1%)	161/1 173 395 (<0·1%)	1/13 771 (<0·1%)
Other	3556/4 693 592 (0·1%)	825/1 173 395 (0·1%)	7/13 771 (0·1%)
Regional	30 721/4 693 592 (0·7%)	7595/1 173 395 (0·6%)	131/13 771 (1·0%)
Spinal	235 183/4 693 592 (5·0%)	59 062/1 173 395 (5·0%)	1493/13 771 (10·8%)
Surgical specialty			
Cardiac surgery	21 673/4 694 458 (0·5%)	5513/1 173 610 (0·5%)	14/13 771 (0·1%)
General surgery	2 150 022/4 694 458 (45·8%)	537 513/1 173 610 (45·8%)	4622/13 771 (33·6%)
Gynaecology	377 543/4 694 458 (8·0%)	94 444/1 173 610 (8·0%)	1238/13 771 (9·0%)
Interventional radiologist	803/4 694 458 (<0·1%)	179/1 173 610 (<0·1%)	0/13 771
Neurosurgery	240 819/4 694 458 (5·1%)	60 437/1 173 610 (5·1%)	691/13 771 (5·0%)
Orthopaedics	1 034 175/4 694 458 (22·0%)	258 692/1 173 610 (22·0%)	2946/13 771 (21·4%)
Other	22/4 694 458 (<0·1%)	7/1 173 610 (<0·1%)	0/13 771
Otolaryngology	128 016/4 694 458 (2·7%)	31 678/1 173 610 (2·7%)	131/13 771 (1·0%)
Plastics	135 671/4 694 458 (2·9%)	33 995/1 173 610 (2·9%)	1467/13 771 (10·7%)
Thoracic	56 684/4 694 458 (1·2%)	14 243/1 173 610 (1·2%)	132/13 771 (1·0%)
Urology	264 613/4 694 458 (5·6%)	65 363/1 173 610 (5·6%)	1041/13 771 (7·6%)
Vascular	284 417/4 694 458 (6·1%)	71 546/1 173 610 (6·1%)	1489/13 771 (10·8%)
Wound classification according to NSQIP definitions			
1 (clean)	2 638 426/4 694 485 (56·2%)	659 842/1 173 621 (56·2%)	7924/13 771 (57·5%)
2 (clean/contaminated)	1 532 329/4 694 485 (32·6%)	382 563/1 173 621 (32·6%)	4524/13 771 (32·9%)
3 (contaminated)	289 157/4 694 485 (6·2%)	72 528/1 173 621 (6·2%)	869/13 771 (6·3%)
4 (dirty/infected)	234 573/4 694 485 (5·0%)	58 688/1 173 621 (5·0%)	454/13 771 (3·3%)
Open wound			
No	4 558 095/4 694 486 (97·1%)	1 139 561/1 173 622 (97·1%)	13 311/13 771 (96·7%)
Yes	136 391/4 694 486 (2·9%)	34 061/1 173 622 (2·9%)	460/13 771 (3·3%)
Elective surgery			
No	948 202/4 688 803 (20·2%)	237 265/1 172 229 (20·2%)	2034/13 766 (14·8%)
Yes	3 740 601/4 688 803 (79·8%)	934 964/1 172 229 (79·8%)	11 732/13 766 (85·2%)
Bleeding disorder			
No	4 499 562/4 694 486 (95·8%)	1 124 713/1 173 622 (95·8%)	12 870/13 771 (93·5%)
Yes	194 924/4 694 486 (4·2%)	48 909/1 173 622 (4·2%)	901/13 771 (6·5%)
Preoperative weight loss			
No	4 636 833/4 694 486 (98·8%)	1 159 238/1 173 621 (98·8%)	13 500/13 771 (98·0%)
Yes	57 653/4 694 486 (1·2%)	14 383/1 173 621 (1·2%)	271/13 771 (2·0%)

(Table 1 continues on next page)

	Training set (n=4 694 488)	Validation set (n=1 173 622)	Test set (n=13 771)
(Continued from previous page)			
Preoperative blood transfusion			
No	4 651 859/4 694 486 (99·1%)	1 162 934/1 173 621 (99·1%)	13 668/13 771 (99·3%)
Yes	42 627/4 694 486 (0·9%)	10 687/1 173 621 (0·9%)	103/13 771 (0·7%)
Inpatient or outpatient status			
Inpatient	2 750 105/4 694 487 (58·6%)	688 551/1 173 622 (58·7%)	10 013/13 771 (72·7%)
Outpatient	1 944 382/4 694 487 (41·4%)	485 071/1 173 622 (41·3%)	3758/13 771 (27·3%)
Transfer status to the hospital where surgery was done			
From acute care hospital inpatient	69 561/4 688 943 (1·5%)	17 407/1 172 237 (1·5%)	332/13 771 (2·4%)
Not transferred (admitted from home)	4 484 143/4 688 943 (95·6%)	1 121 245/1 172 237 (95·7%)	13 265/13 771 (96·3%)
Nursing home or chronic care facility or intermediate care unit	43 894/4 688 943 (0·9%)	10 894/1 172 237 (0·9%)	69/13 771 (0·5%)
Outside emergency department	77 958/4 688 943 (1·7%)	19 348/1 172 237 (1·7%)	76/13 771 (0·6%)
Transfer from somewhere else	13 387/4 688 943 (0·3%)	3343/1 172 237 (0·3%)	29/13 771 (0·2%)
Superficial incisional surgical site infection PATOS			
No	4 689 919/4 694 467 (99·9%)	1 172 455/1 173 619 (99·9%)	13 765/13 771 (100·0%)
Yes	4548/4 694 467 (0·1%)	1164/1 173 619 (0·1%)	6/13 771 (<0·1%)
Deep incisional surgical site infection PATOS			
No	4 688 534/4 694 467 (99·9%)	1 172 138/1 173 619 (99·9%)	13 765/13 771 (100·0%)
Yes	5933/4 694 467 (0·1%)	1481/1 173 619 (0·1%)	6/13 771 (<0·1%)
Organ-space surgical site infection PATOS			
No	4 674 569/4 694 467 (99·6%)	1 168 694/1 173 619 (99·6%)	13 699/13 771 (99·5%)
Yes	19 898/4 694 467 (0·4%)	4925/1 173 619 (0·4%)	72/13 771 (0·5%)
Pneumonia PATOS			
No	4 683 885/4 694 467 (99·8%)	1 170 962/1 173 619 (99·8%)	13 756/13 771 (99·9%)
Yes	10 582/4 694 467 (0·2%)	2657/1 173 619 (0·2%)	15/13 771 (0·1%)
On ventilator for >48 h at the time of surgery			
No	4 685 196/4 694 467 (99·8%)	1 171 330/1 173 619 (99·8%)	13 748/13 771 (99·8%)
Yes	9271/4 694 467 (0·2%)	2289/1 173 619 (0·2%)	23/13 771 (0·2%)
Urinary tract infection PATOS			
No	4 684 571/4 694 467 (99·8%)	1 171 193/1 173 619 (99·8%)	13 755/13 771 (99·9%)
Yes	9896/4 694 467 (0·2%)	2426/1 173 619 (0·2%)	16/13 771 (0·1%)
Sepsis PATOS			
No	4 654 136/4 694 467 (99·1%)	1 163 360/1 173 619 (99·1%)	13 631/13 771 (99·0%)
Yes	40 331/4 694 467 (0·9%)	10 259/1 173 619 (0·9%)	140/13 771 (1·0%)
Septic shock PATOS			
No	4 676 334/4 694 467 (99·6%)	1 169 111/1 173 619 (99·6%)	13 725/13 771 (99·7%)
Yes	18 133/4 694 467 (0·4%)	4508/1 173 619 (0·4%)	46/13 771 (0·3%)
Total operation time, min	85 (50–140)	85 (50–140)	146 (95–229)
Time from hospital admission to operation, days	0 (0–0)	0 (0–0)	0 (0–0)
Work relative value units,† units	15·37 (10·05–20·82)	15·37 (10·05–20·82)	20·72 (13·99–26·49)

(Table 1 continues on next page)

for continuous validation of the model and to prevent overfitting. The test set was reserved for model testing after successful development and training, and thus served as the basis for evaluating the performance of the final model.

Input variables

We created a total of three deep learning models with increasing levels of complexity (ie, increasing numbers of input variables). Table 1 summarises the input and output variables for each of the three models. For comparative purposes, the first model was built on the same 21 input variables as the ACS-SRC.[4] To allow automatic differentiation between variables that were changed over the study period, we added the year of operation as an input feature. Model 2 was trained on the same input variables as model 1, with the addition of 13 different preoperative laboratory values. Model 3 was trained on the same input variables as model 2, with the addition of the remaining 23 preoperative

	Training set (n=4 694 488)	Validation set (n=1 173 622)	Test set (n=13 771)
(Continued from previous page)			
Output features for all models			
Mortality			
No	4 648 123/4 694 488 (99·0%)	1 161 987/1 173 622 (99·0%)	13 671/13 771 (99·3%)
Yes	46 365/4 694 488 (1·0%)	11 635/1 173 622 (1·0%)	100/13 771 (0·7%)
Morbidity			
No	4 382 377/4 694 488 (93·4%)	1 094 854/1 173 622 (93·3%)	12 434/13 771 (90·3%)
Yes	312 111/4 694 488 (6·6%)	78 768/1 173 622 (6·7%)	1337/13 771 (9·7%)
Superficial incisional surgical site infection			
No	4 629 545/4 694 488 (98·6%)	1 157 064/1 173 622 (98·6%)	13 534 /13 771 (98·3%)
Yes	64 943/4 694 488 (1·4%)	16 558/1 173 622 (1·4%)	237/13 771 (1·7%)
Deep incisional surgical site infection			
No	4 677 242/4 694 488 (99·6%)	1 169 169/1 173 622 (99·6%)	13 724/13 771 (99·7%)
Yes	17 246/4 694 488 (0·4%)	4453/1 173 622 (0·4%)	47/13 771 (0·3%)
Organ-space surgical site infection			
No	4 653 081/4 694 488 (99·1%)	1 163 180/1 173 622 (99·1%)	13 517/13 771 (98·2%)
Yes	41 407/4 694 488 (0·9%)	10 442/1 173 622 (0·9%)	254/13 771 (1·8%)
Wound disruption			
No	4 676 753/4 694 488 (99·6%)	1 169 168/1 173 622 (99·6%)	13 691/13 771 (99·4%)
Yes	17 735/4 694 488 (0·4%)	4454/1 173 622 (0·4%)	80/13 771 (0·6%)
Pneumonia			
No	4 649 943/4 694 488 (99·1%)	1 162 517/1 173 622 (99·1%)	13 591/13 771 (98·7%)
Yes	44 545/4 694 488 (0·9%)	11 105/1 173 622 (0·9%)	180/13 771 (1·3%)
Unplanned intubation			
No	4 659 213/4 694 488 (99·2%)	1 164 807/1 173 622 (99·2%)	13 609/13 771 (98·8%)
Yes	35 275/4 694 488 (0·8%)	8815/1 173 622 (0·8%)	162/13 771 (1·2%)
Pulmonary embolism			
No	4 679 109/4 694 488 (99·7%)	1 169 805/1 173 622 (99·7%)	13 679/13 771 (99·3%)
Yes	15 379/4 694 488 (0·3%)	3817/1 173 622 (0·3%)	92/13 771 (0·7%)
On ventilator for >48 h within 30 postoperative days			
No	4 659 223/4 694 488 (99·2%)	1 164 758/1 173 622 (99·2%)	13 622/13 771 (98·9%)
Yes	35 265/4 694 488 (0·8%)	8864/1 173 622 (0·8%)	149/13 771 (1·1%)
Progressive renal insufficiency			
No	4 682 890/4 694 488 (99·8%)	1 170 715/1 173 622 (99·8%)	13 731/13 771 (99·7%)
Yes	11 598/4 694 488 (0·2%)	2907/1 173 622 (0·2%)	40/13 771 (0·3%)
Acute renal failure			
No	4 681 964/4 694 488 (99·7%)	1 170 427/1 173 622 (99·7%)	13 725/13 771 (99·7%)
Yes	12 524/4 694 488 (0·3%)	3195/1 173 622 (0·3%)	46/13 771 (0·3%)
Urinary tract infection			
No	4 643 983/4 694 488 (98·9%)	1 160 700/1 173 622 (98·9%)	13 556/13 771 (98·4%)
Yes	50 505/4 694 488 (1·1%)	12 922/1 173 622 (1·1%)	215/13 771 (1·6%)
Stroke			
No	4 685 413/4 694 488 (99·8%)	1 171 383/1 173 622 (99·8%)	13 729/13 771 (99·7%)
Yes	9075/4 694 488 (0·2%)	2239/1 173 622 (0·2%)	42/13 771 (0·3%)
Cardiac arrest requiring cardiopulmonary resuscitation			
No	4 680 473/4 694 488 (99·7%)	1 170 023/1 173 622 (99·7%)	13 712/13 771 (99·6%)
Yes	14 015/4 694 488 (0·3%)	3599/1 173 622 (0·3%)	59/13 771 (0·4%)
Myocardial infarction			
No	4 677 500/4 694 488 (99·6%)	1 169 367/1 173 622 (99·6%)	13 694/13 771 (99·4%)
Yes	16 988/4 694 488 (0·4%)	4255/1 173 622 (0·4%)	77/13 771 (0·6%)

(Table 1 continues on next page)

	Training set (n=4 694 488)	Validation set (n=1 173 622)	Test set (n=13 771)
(Continued from previous page)			
Thrombophlebitis (including deep vein thrombosis)			
No	4 668 257/4 694 488 (99·4%)	1 167 084/1 173 622 (99·4%)	13 527/13 771 (98·2%)
Yes	26 231/4 694 488 (0·6%)	6538/1 173 622 (0·6%)	244/13 771 (1·8%)
Sepsis			
No	4 658 607/4 694 488 (99·2%)	1 164 606/1 173 622 (99·2%)	13 658/13 771 (99·2%)
Yes	35 881/4 694 488 (0·8%)	9016/1 173 622 (0·8%)	113/13 771 (0·8%)
Septic shock			
No	4 675 151/4 694 488 (99·6%)	1 168 818/1 173 622 (99·6%)	13 707/13 771 (99·5%)
Yes	19 337/4 694 488 (0·4%)	4804/1 173 622 (0·4%)	64/13 771 (0·5%)
Bleeding requiring transfusions			
No	4 448 873/4 694 488 (94·8%)	1 111 477/1 173 622 (94·7%)	12 386/13 771 (89·9%)
Yes	245 615/4 694 488 (5·2%)	62 145/1 173 622 (5·3%)	1385/13 771 (10·1%)

Data are n/N (%) or median (IQR). ACS NSQIP=American College of Surgeons National Surgical Quality Improvement Program. ASA=American Society of Anesthesiologists. COPD=chronic obstructive pulmonary disease. GOT=glutamic oxaloacetic aminotransferase. INR=international normalised ratio. PATOS=present at time of surgery. PTT=partial thromboplastin time. *Patients with an ASA class of 6 are those who have been declared brain dead and whose organs are being removed. These patients are not accrued in the ACS NSQIP. †A measure of resource requirements and thus, indirectly, a measure of the severity of surgical procedures.

Table 1: **Input and output variables for all three models and baseline characteristics of patients in the training, validation, and test sets**

For the **2018 ACS NSQIP user guide** see https://www.facs.org/-/media/files/quality-programs/nsqip/nsqip_puf_userguide_2018.ashx

ACS NSQIP variables from the 2018 ACS NSQIP user guide, except for information regarding additional procedures other than the primary procedure and the number of days between when the preoperative laboratory values were sampled and the surgery. These variables were excluded because of their high cardinality and risk of overfitting. Model 3 had a total of 58 input features, which turned into 76 input features when preprocessing the continuous variables (creating a new binary input feature for those containing missing values). To minimise data leakage, we chose to exclude variables that were collected after the completion of each surgical procedure. Input variables were divided into categorical and continuous data. Categorical variables were embedded according to the number of categories, as implemented in the fast.ai library. In practice, this approach meant attributing a random vector of a certain length to each of the categorical values.[17] The parameters in each vector were randomly initialised and updated with training. Training the embedding matrices along with the linear layers allowed us to capture complex and multidimensional relationships between categories. Missing and unknown values were kept as distinct categories within each variable. For continuous data, we replaced missing values with the median of the group while simultaneously creating a new binary column indicating whether a variable was missing or not. Subsequently, continuous variables were normalised by subtraction of the mean and division by the SD.

For more on the **fast.ai library** see https://www.fast.ai

Multilabel output variables

Neural networks intrinsically support multilabel tasks in their design, with each label being another node in the output layer. However, the traditional SoftMax activation evaluates each class against the rest. This approach only works if classes are mutually exclusive, such as if a patient dies or survives. These two classes are mutually exclusive, as a patient cannot both die and survive. Therefore, we used the sigmoid activation function, which evaluates each class separately, allowing for multilabel risk prediction (eg, the risk of mortality and the risk of sepsis).

For comparative purposes, we chose the same output variables as two previously published models.[4,6] These output variables comprised mortality and 18 different postoperative complications (table 1). In the original paper describing the ACS-SRC, the overall morbidity was defined as any of 16 postoperative complications (ie, superficial surgical site infection, deep surgical site infection, organ-space surgical site infection, wound disruption, pneumonia, unplanned intubation, pulmonary embolism, on a ventilator for more than 48 h, progressive renal insufficiency, acute renal failure, urinary tract infection, stroke, cardiac arrest requiring cardiopulmonary resuscitation, myocardial infarction, thrombophlebitis [including deep vein thrombosis], and systemic sepsis). Cardiac events were defined as either cardiac arrest or myocardial infarction, and renal failure as either progressive renal insufficiency or acute renal failure.[4] The same definitions were used for this study.

Neural network architecture

Model architecture was established during preliminary testing. A graphic overview of the final neural network architecture is presented in the appendix (p 4). The only architectural difference between model 1 and the other models was the number of embedding matrices. The three models were developed with 1 134 080, 1 186 327, and 1 264 846 trainable variables and 21, 35, and 57 embedding layers, respectively. The remaining

	Model 1		Model 2		Model 3	
	Validation set	Test set	Validation set	Test set	Validation set	Test set
Mortality	0·946 (0·943–0·949)	0·942 (0·910–0·974)	0·952 (0·949–0·955)	0·946 (0·915–0·977)	0·955 (0·952–0·957)	0·951 (0·921–0·980)
Superficial incisional surgical site infection	0·786 (0·782–0·790)	0·816 (0·782–0·849)	0·789 (0·785–0·793)	0·816 (0·783–0·849)	0·803 (0·799–0·807)	0·825 (0·793–0·858)
Deep incisional surgical site infection	0·815 (0·807–0·823)	0·782 (0·704–0·861)	0·818 (0·810–0·826)	0·779 (0·701–0·858)	0·834 (0·827–0·842)	0·797 (0·720–0·873)
Organ-space surgical site infection	0·852 (0·847–0·856)	0·851 (0·821–0·881)	0·854 (0·849–0·858)	0·852 (0·822–0·882)	0·869 (0·864–0·873)	0·871 (0·842–0·899)
Wound disruption	0·837 (0·830–0·845)	0·832 (0·776–0·887)	0·839 (0·832–0·847)	0·830 (0·775–0·886)	0·852 (0·845–0·859)	0·828 (0·772–0·883)
Pneumonia	0·869 (0·865–0·873)	0·862 (0·828–0·896)	0·875 (0·871–0·879)	0·861 (0·826–0·895)	0·885 (0·881–0·890)	0·872 (0·839–0·905)
Unplanned intubation	0·896 (0·891–0·900)	0·887 (0·853–0·920)	0·903 (0·899–0·908)	0·890 (0·857–0·923)	0·914 (0·910–0·918)	0·902 (0·870–0·933)
Pulmonary embolism	0·789 (0·781–0·798)	0·745 (0·686–0·803)	0·798 (0·789–0·806)	0·752 (0·694–0·810)	0·813 (0·805–0·821)	0·763 (0·705–0·820)
On ventilator for >48 h within 30 postoperative days	0·936 (0·933–0·940)	0·918 (0·888–0·949)	0·942 (0·938–0·945)	0·920 (0·890–0·950)	0·952 (0·949–0·955)	0·935 (0·907–0·962)
Progressive renal insufficiency	0·872 (0·864–0·880)	0·920 (0·862–0·978)	0·887 (0·879–0·894)	0·935 (0·882–0·988)	0·895 (0·888–0·903)	0·923 (0·866–0·980)
Acute renal failure	0·927 (0·921–0·934)	0·934 (0·884–0·984)	0·943 (0·938–0·949)	0·937 (0·888–0·986)	0·949 (0·944–0·955)	0·942 (0·895–0·989)
Urinary tract infection	0·775 (0·770–0·780)	0·760 (0·722–0·797)	0·778 (0·773–0·782)	0·763 (0·726–0·801)	0·789 (0·784–0·794)	0·771 (0·734–0·808)
Stroke	0·883 (0·874–0·893)	0·922 (0·866–0·978)	0·886 (0·877–0·895)	0·923 (0·867–0·979)	0·895 (0·886–0·903)	0·928 (0·874–0·982)
Cardiac arrest requiring cardiopulmonary resuscitation	0·905 (0·899–0·912)	0·913 (0·864–0·963)	0·913 (0·906–0·919)	0·911 (0·860–0·961)	0·919 (0·913–0·926)	0·922 (0·875–0·969)
Myocardial infarction	0·880 (0·873–0·887)	0·880 (0·83–0·929)	0·885 (0·878–0·891)	0·883 (0·833–0·932)	0·893 (0·886–0·899)	0·898 (0·852–0·945)
Thrombophlebitis (including deep vein thrombosis)	0·806 (0·799–0·812)	0·774 (0·740–0·809)	0·814 (0·808–0·821)	0·779 (0·744–0·814)	0·827 (0·821–0·833)	0·795 (0·761–0·829)
Sepsis	0·832 (0·827–0·837)	0·811 (0·763–0·860)	0·838 (0·832–0·843)	0·815 (0·767–0·863)	0·854 (0·849–0·859)	0·847 (0·802–0·892)
Septic shock	0·907 (0·902–0·913)	0·908 (0·859–0·957)	0·914 (0·908–0·919)	0·906 (0·856–0·955)	0·925 (0·920–0·930)	0·918 (0·871–0·964)
Bleeding requiring transfusions	0·901 (0·899–0·903)	0·862 (0·849–0·874)	0·927 (0·926–0·929)	0·895 (0·884–0·906)	0·943 (0·941–0·944)	0·924 (0·914–0·933)
Mean performance	0·864 (0·053)	0·859 (0·063)	0·871 (0·055)	0·863 (0·064)	0·882 (0·053)	0·874 (0·061)

Data are area under the receiver operating characteristic curve (95% CI) or mean (SD).

Table 2: **Performance metrics for the three models on the validation and test sets**

parts of the neural networks were similar across the three models. This remaining architecture included three batch norm layers, two rectified linear unit layers, and three linear layers (appendix p 4). In each of the three models, the first linear layer had 1000 activations and the second linear layer had 500 activations. The final layer of each model was a linear layer with 19 outputs: one for mortality and one for each of the 18 postoperative complications. Hyperparameters included dropouts of 0·04 after the embedding layers, 0·001 after the first linear layer, and 0·01 after the second linear layer. To allow for multiple output variables, we one-hot encoded all output variables and transformed them into multicategory tensors. We then used the Adam optimiser and a flattened binary cross entropy loss function for multilabel risk prediction. All models were trained for five epochs, with a batch size of 1024, a learning rate of $3e^{-3}$, and a weight decay of 0·2.

Deep learning model prediction

The overall model performance was evaluated on both the validation and the test set. We calculated the area under the receiver operating characteristic curve (AUC), with 95% CIs, and the Brier score, for each of the 19 output variables (mortality and the 18 postoperative complications). These values were compared with the published

performance metrics from the ACS-SRC.[4] In addition to evaluating the model on the validation and test sets, we evaluated resistance to changes in the underlying patient population on a subset of the test set, comprising only the patients who had emergency surgery. Results were compared with the Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) calculator, which has been designed specifically for emergency surgery cohorts.[6]

	ACS-SRC	Model 1	Model 2	Model 3
Mortality	0·944	0·942 (0·910–0·974)	0·946 (0·915–0·977)	0·951 (0·921–0·980)
Morbidity	0·816	0·863 (0·852–0·873)	0·865 (0·855–0·875)	0·876 (0·866–0·886)
Pneumonia	0·870	0·862 (0·828–0·896)	0·861 (0·826–0·895)	0·872 (0·839–0·905)
Cardiac	0·895	0·897 (0·862–0·932)	0·898 (0·862–0·933)	0·910 (0·877–0·943)
Surgical site infection	0·817	0·851 (0·830–0·871)	0·851 (0·830–0·871)	0·864 (0·844–0·883)
Urinary tract infection	0·806	0·760 (0·722–0·797)	0·763 (0·726–0·801)	0·771 (0·734–0·808)
Thrombophlebitis (eg, deep vein thrombosis)	0·819	0·774 (0·740–0·809)	0·779 (0·744–0·814)	0·795 (0·761–0·829)
Renal failure	0·903	0·927 (0·889–0·965)	0·934 (0·898–0·970)	0·934 (0·898–0·970)
Mean performance	0·859 (0·052)	0·860 (0·066)	0·862 (0·066)	0·872 (0·063)

Data are area under the receiver operating characteristic curve (95% CI) or mean (SD). We only compare the categories included in the orginal ACS-SRC paper.[4] ACS-SRC=American College of Surgeons Surgical Risk Calculator.

Table 3: **Performance metrics for all surgeries for the three models on the test set compared with those from the ACS-SRC**

Shapley additive explanations (SHAP) for feature importance

One of the disadvantages of a deep neural network is the limited interpretability of the complex underlying multilayer and non-linear structure.[18] To get insights into the workings of our models, we calculated SHAP for all patients in the test set using previously published methods.[19] Based on game theory, this approach allowed us to estimate the mean impact on model output magnitude for each of the input features.[20] The calculated SHAP values were then used to create feature importance plots.

Personalised risk prediction

To illustrate the potential for personalised risk prediction, we present four examples of risk predictions based on the input features for model 3. For illustrative purposes, we randomly chose a patient that did not develop a deep surgical site infection (patient A), a patient that did (patient B), a patient that did not develop thrombophlebitis (patient C), and a patient that did (patient D).

Analysis of validity

Multiple measures have been taken to ensure the validity of this study. Data leakage has been minimised by

	POTTER	Model 1	Model 2	Model 3
Mortality	0·9199	0·8943 (0·8307–0·9579)	0·9049 (0·8442–0·9657)	0·9115 (0·8525–0·9704)
Morbidity	0·8511	0·8608 (0·8342–0·8873)	0·8715 (0·8458–0·8972)	0·8780 (0·8528–0·9032)
Superficial incisional surgical site infection	0·6808	0·7460 (0·6312–0·8608)	0·7497 (0·6353–0·8640)	0·7562 (0·6426–0·8698)
Deep incisional surgical site infection	0·7540	0·8654 (0·6791–1·0000)	0·8463 (0·6505–1·0000)	0·8222 (0·6163–1·0000)
Organ-space surgical site infection	0·7860	0·7815 (0·6507–0·9123)	0·8059 (0·6797–0·9321)	0·8625 (0·7507–0·9743)
Wound disruption	0·7790	0·7545 (0·5941–0·9149)	0·7720 (0·6147–0·9293)	0·8184 (0·6716–0·9652)
Pneumonia	0·8470	0·8406 (0·7543–0·9269)	0·8358 (0·7486–0·9231)	0·8167 (0·7261–0·9074)
Unplanned intubation	0·8493	0·8640 (0·7928–0·9351)	0·8755 (0·8068–0·9443)	0·8776 (0·8093–0·9459)
Pulmonary embolism	0·7333	0·8619 (0·6559–1·0000)	0·9023 (0·7228–1·0000)	0·8494 (0·6366–1·0000)
On ventilator for >48 h within 30 postoperative days	0·9254	0·8915 (0·8306–0·9524)	0·9023 (0·8440–0·9606)	0·9176 (0·8635–0·9717)
Progressive renal insufficiency	0·8188	0·7983 (0·6320–0·9645)	0·8208 (0·6608–0·9809)	0·8099 (0·6468–0·9731)
Acute renal failure	0·9126	0·9219 (0·8358–1·0000)	0·9442 (0·8702–1·0000)	0·9530 (0·8847–1·0000)
Urinary tract infection	0·7396	0·7293 (0·5653–0·8934)	0·7631 (0·6042–0·9221)	0·7575 (0·5976–0·9174)
Stroke	0·8343	0·9078 (0·7772–1·0000)	0·9357 (0·8242–1·0000)	0·9208 (0·7985–1·0000)
Cardiac arrest requiring cardiopulmonary resuscitation	0·8882	0·9131 (0·8144–1·0000)	0·9047 (0·8021–1·0000)	0·9147 (0·8167–1·0000)
Myocardial infarction	0·8240	0·8303 (0·7161–0·9445)	0·8609 (0·7546–0·9671)	0·8788 (0·7780–0·9795)
Thrombophlebitis (including deep vein thrombosis)	0·7886	0·8277 (0·7294–0·9260)	0·8234 (0·7242–0·9225)	0·8245 (0·7256–0·9235)
Sepsis	0·8448	0·7954 (0·6285–0·9624)	0·7792 (0·6085–0·9499)	0·8419 (0·6887–0·9952)
Septic shock	0·9338	0·9060 (0·8101–1·0000)	0·9166 (0·8255–1·0000)	0·9417 (0·8640–1·0000)
Bleeding requiring transfusions	0·9028	0·8908 (0·8578–0·9238)	0·9283 (0·9010–0·9557)	0·9396 (0·9143–0·9648)
Mean performance	0·8307 (0·0714)	0·8441 (0·0593)	0·8572 (0·0611)	0·8646 (0·0587)

Data are area under the receiver operating characteristic curve (95% CI) or mean (SD). POTTER=Predictive OpTimal Trees in Emergency Surgery Risk.

Table 4: **Performance metrics for emergency surgeries for the three models on the test set compared with those from POTTER models**

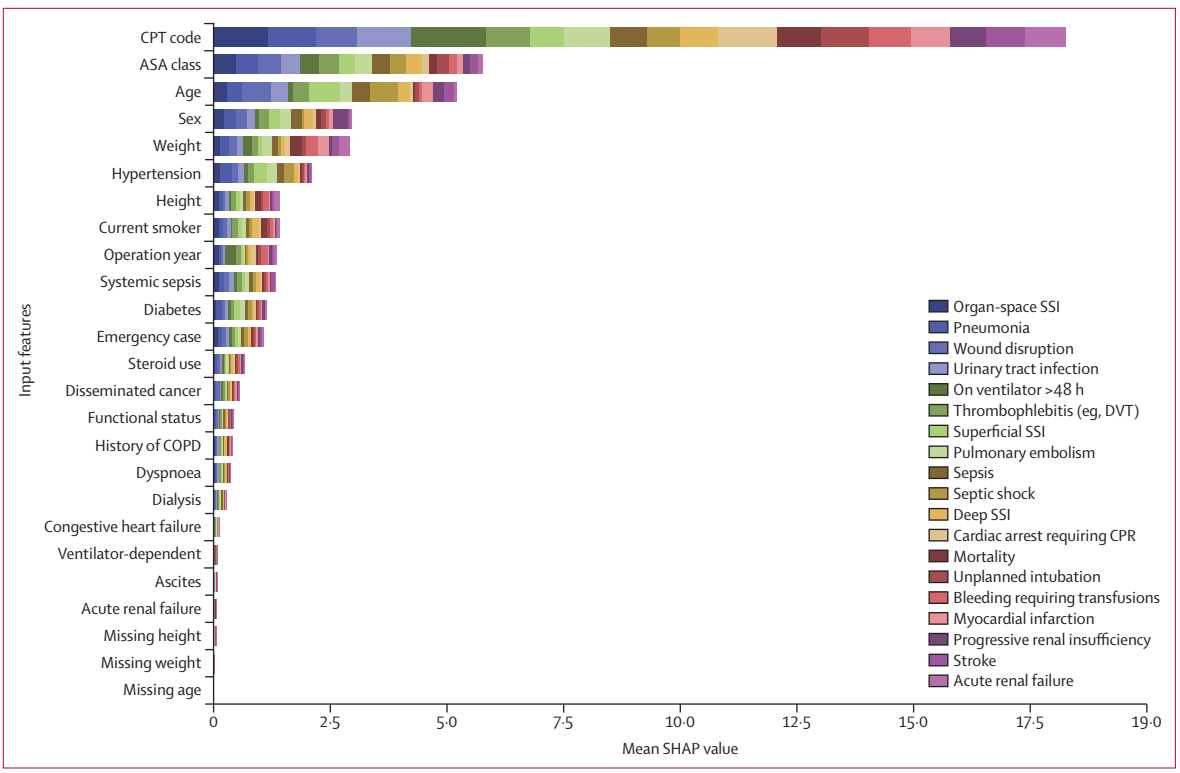

Figure 1: **SHAP feature importance plot for model 1 in the test set**
Plots for model 2 and model 3 can be found in the appendix (pp 1–2). The values on the x-axis indicate the mean impact on model output magnitude based on SHAP. The input features on the y-axis are ordered by descending importance and the horizontal bars are colour-coded to match the multilabel output variables. The larger the coloured area, the more impact the feature had on the corresponding outcome. A current smoker was defined as those who had smoked cigarettes at any point within the 12 months before admission for surgery. ASA=American Society of Anesthesiologists. COPD=chronic obstructive pulmonary disease. CPR=cardiopulmonary resuscitation. CPT=Current Procedural Terminology. DVT=deep vein thrombosis. SHAP=Shapley additive explanations. SSI=surgical site infection.

exclusively choosing input variables that can be captured at the end of the primary surgical procedure (eg, the total operation time). Additionally, the ACS NSQIP data definitions are designed to minimise the risk of coded postoperative complications being present preoperatively. Moreover, we double checked the feature importance plots for any signs of data leakage. As an additional validation process, we chose not to count events that were coded as present at the time of surgery as postoperative complications.

Implementation

Models were implemented using Python (version 3.7.7), PyTorch (verson 1.6.0),[21] and fast.ai (version 2.0.11).[22] Performance metrics were calculated using scitkit-learn (version 0.23.1).[23] Personalised risk predictions were estimated with SHAP (version 0.35.0).[19] Implementations of feature importance and personalised risk predictions were based on code from Zachary Mueller.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

We included 5 881 881 adults who had surgery during the 6-year study period (2012–18), spanning 2941 unique Current Procedural Terminology (CPT) codes. Summary statistics for all the input and output features in the training, validation, and test sets are presented in table 1. Overall, the training and validation sets were very similar, indicating that random selection worked. As expected, some baseline characteristics from the test set deviated from the other datasets (table 1). Across all datasets, we found a mortality of 58 100 (0·99%) of 5 881 881 and a morbidity of 392 216 (6·7%).

The performance of each model on the validation and the test sets is presented in table 2. Additionally, we have added the Brier score of each model in the appendix (p 5). The receiver operating characteristic curve plots with associated AUCs for the test set are presented in the appendix (pp 6–8). We observed an increase in AUC with an increase in the number of input variables. The mean performance was only modestly lower for the test set compared with the validation set. The mean AUCs for the validation set were 0·864 (SD 0·053) for model 1, 0·871 (0·055) for model 2, and 0·882 (0·053) for model 3. The mean AUCs for the test set were 0·859 (SD 0·063)

For more on **code from Zachary Mueller** see https://github.com/muellerzr/

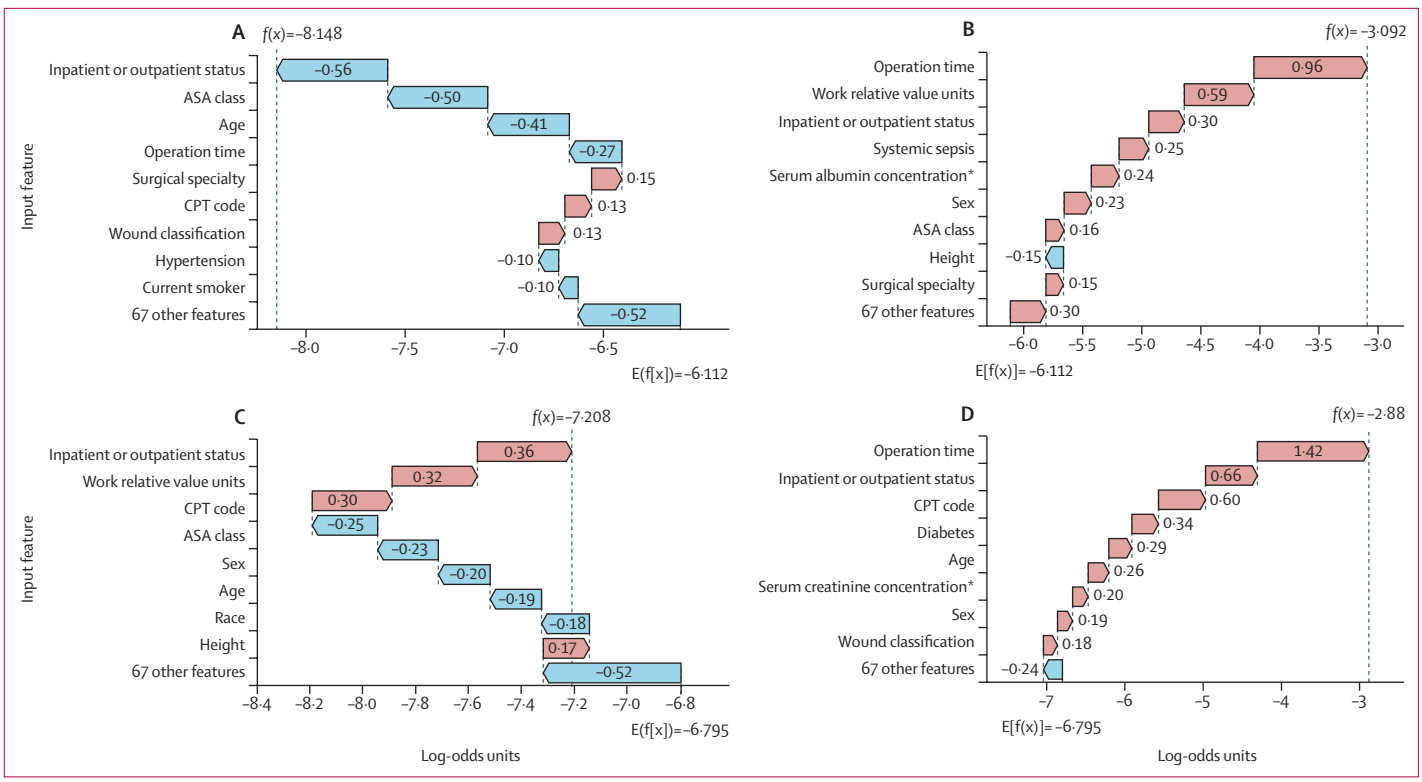

Figure 2: **Personalised risk prediction using model 3**

Waterfall plots (showing the cumulative effects of sequentially introduced positive and negative values) for a patient who did not develop a deep surgical site infection (A), a patient that did (B), a patient that did not develop thrombophlebitis (C), and a patient that did (D). The x-axes are log-odds units, so negative values imply probabilities of less than 0·5 that the person will develop the complication. The y-axes represent the input features ordered by descending importance. The background risk in the test set was 0·34% for deep surgical site infection and 1·77% for deep vein thrombosis. The x-axis represents the model output before it is converted to percentages. E[f(X)] is the model output for the background population, equalling −6·112 for deep surgical site infection and −6·795 for deep vein thrombosis, before both were passed through a sigmoid layer, converting them to the 0·34% and 1·77% background risks, respectively. f(x) is the personalised model output for a patient. If f(x) is smaller than E[f(x)], the patient has a lower risk of a complication relative to the background population. Conversely, if f(x) is larger than E[f(x)], the patient has an higher risk of a complication relative to the background population. Each arrow represents how a specific feature increases (red) or decreases (blue) the patients' risk for a specific complication. ASA=American Society of Anesthesiologists. CPT=Current Procedural Terminology. *Preoperative laboratory values.

for model 1, 0·863 (0·064) for model 2, and 0·874 (0·061) for model 3.

The mean AUCs of each of our models outperformed the mean AUC of previously published data from the ACS-SRC, again increasing with increasing model complexity (table 3). Compared with model performance on the test set, the ACS-SRC outperformed our models for two of nine complications (urinary tract infection and deep vein thrombosis; table 3). All models appeared to retain predictive power despite changes in the underlying patient population: model performance on the subset of patients in the test set who had emergency surgery is presented in table 4. Model 1 outperformed the POTTER calculator on ten of 20 outcomes, and had a higher mean AUC (0·8441 [SD 0·0593]) than did the POTTER calculator (0·8307 [0·0714]; table 4). Again, model performance increased with increasing levels of complexity, with model 2 (mean AUC 0·8572 [0·0611]) outperforming the POTTER calculator on 14 outcomes and model 3 (0·8646 [0·0587]) outperforming the POTTER calculator on 15 outcomes.

The features' importances were largely ranked consistently across the three models (figure 1; appendix pp 1–2). However, as expected, the mean impact on model output magnitude per feature decreased as more input features were added (figure 1; appendix pp 1–2). For model 3, the top five most important input features towards surgical risk prediction were the CPT code, total operation time, inpatient or outpatient status, age, and American Society of Anesthesiologists (ASA) classification. The most important preoperative laboratory value was the haematocrit (appendix pp 1–2), largely driven by its role in predicting the risk of being on a ventilator for more than 48 postoperative h.

To illustrate the potential for personalised risk prediction, we present four examples of risk predictions based on the input features for model 3 (figure 2). We randomly chose a patient that did not develop a deep surgical site infection (patient A), a patient that did (patient B), a patient that did not develop thrombophlebitis (patient C), and a patient that did (patient D). Patient A was a 26-year-old woman scheduled for an outpatient laparoscopic cholecystectomy in general surgery and a non-smoker, had an ASA class of 1, a wound classification of 2 (clean/contaminated), and a total operation time of

62 min, and did not have hypertension requiring medication. Patient B was a 58-year-old, 6-feet-tall man scheduled for inpatient pelvic exenteration for colorectal malignancy in general surgery. He had preoperative systemic sepsis, an ASA class of 3, a preoperative serum albumin concentration of 2·8 g/dL, a total operation time of 582 min, and a total work relative value units (a measure of resource requirements and thus, indirectly, a measure of the severity of surgical procedures) of 49·1 units. Patient C was a 5-feet-tall, Asian woman scheduled for inpatient laparoscopic colectomy. She had an ASA class of 2, a total operation time of 214 min, and a total work relative value unit of 31·92 units. Patient D was a 70-year-old man scheduled for an inpatient repair of an abdominal wall hernia. He had a wound classification of 3 (contaminated), diabetes requiring insulin, an ASA class of 3, a total operation time of 462 min, and a preoperative creatinine concentration of 1·83 mg/dL.

The model correctly identified both negative controls to have a lower risk of either deep surgical site infection (patient A risk prediction 0·0% risk) or deep vein thrombosis (patient C risk prediction 0·2% risk) than the background population (the test set). The two cases were also correctly classified as being at high risk of either a deep surgical site infection (patient B risk prediction 3·8% risk) or deep vein thrombosis (patient D risk prediction 2·6% risk). The most important risk factors for patient B were the total operating time, the work relative value units, inpatient or outpatient status, systemic sepsis within 48 preoperative h, and his serum albumin concentration of 2·8 g/L (figure 2B). The most important risk factors for patient D again included the total operation time and the inpatient or outpatient status, but also included the surgery itself, as denoted by the CPT code (abdominal wall hernia), and the fact that he had diabetes requiring insulin and was an older adult (70-years-old; figure 2D).

Discussion

In this study, deep neural networks were leveraged to train state-of-the-art models on structured electronic health-care data for surgical risk prediction. We created three different models that had an increasing number of input variables. The first model was built for comparative purposes and was trained on the same input variables as the ACS-SRC, with the addition of operation year. For the second model, we added 13 additional preoperative laboratory values. For the third model, we added all additional input variables that were considered relevant, resulting in a total of 76 input features. We showed how a single multilabel model reliably outperformed the current gold standard in surgical risk prediction, retaining predictive power even when used on emergency procedures. Additionally, we showed that increasing the number of input variables increased performance. Thus, these results further support other encouraging findings from machine learning studies on unstructured electronic health record data.[6,13,14,24] Compared with these previous works, our findings indicate that large-scale repositories of structured surgical data (eg, the ACS NSQIP) can be used to develop deep learning models, and that accurate risk prediction for multiple outcomes can be achieved without the need to train individual models.

Deep learning could provide several advantages if incorporated into clinical practice. First, the high level of performance has the potential to increase the accuracy of preoperative risk assessments, equipping physicians with the tools for optimised preoperative patient counselling and decision making. As such, using the combination of the risk prediction with the feature importance analyses would allow clinicians to assess both postoperative risks and potentially amendable driving factors. Second, personalised risk predictions have the potential to engage patients as stakeholders by empowering them with advanced knowledge of their personal surgical risk profiles. Using the feature importance analysis would allow patients to assess driving factors for their personal risk profile, and thus allow for the opportunity to address these factors, at least in the elective surgical setting. Finally, the ability of deep neural networks to incorporate high-dimensional input variables enables unbiased investigations into the driving factors behind postoperative complications, while factoring in the often non-linear associations between inputs and outcomes. As such, post-hoc analysis of a model incorporating all available datapoints enables analysis of features that the network highlights as important for postoperative complications. This analysis could be used for preoperative patient counselling and the ability to highlight important risks might be helpful in hypothesis generation for validation in future studies.

The fact that these deep learning models attain high accuracy in predicting the risk of postoperative complications suggests that postoperative complications are, to a certain degree, predictable events regulated by a small number of preoperative factors. This notion is further supported by the fact that these predictions were done by a singular network, rather than individual networks for each complication. As such, this could suggest that a common pathophysiological mechanism is shared between postoperative complications. Although our study did not explore this hypothesis, we speculate that shared factors, such as patient frailty and the gravity of the surgical stress response, could be key players.

Nevertheless, as the complexity of the model increases, so does the need for multiple additional datapoints. Although previous models have reported good predictive performances while focusing on provider-level data inputs,[4–6] even when limited to eight variables,[5] complex deep learning models would be unsuitable in a setting in which provider-level input was required. Our first model does, however, use input data identical to that of the widely used online ACS-SRC; the subsequent models

use increasingly complex data and direct implementation into the electronic health record system would be required to fully capitalise on these models' potentials. A set-up achieving automated aggregation of surgical data has previously been proposed.[25]

Using a deep learning approach also offers additional advantages in the clinical setting. For instance, missing or incomplete data are a common occurrence in the clinical reality. Deep learning models are, however, robust against this issue, as indicated by a study that used real-world data from intensive care units and modelled missing data, and still produced acceptably performing models.[26] Furthermore, by contrast to other model architectures, deep learning offers the potential of dynamic retraining. With implementation into the electronic health record system, the concept of transfer learning, as used in fields such as clinical gene transcriptomics[27] and radiology[28] and electronic health record data,[29] allows for dynamic retraining of the network on local data. This retraining, in turn, would allow the network to learn general features on large-scale datasets, such as the ACS NSQIP database, with subsequent retraining on local data allowing the network to learn features important for geographical site-specific risk prediction. Neural networks tend to be underexplored on structured data due, in part, to their continuous nature. Classically, tree-based approaches have been better than neural networks at handling highly categorical data structures such as the ACS NSQIP database. One of the crucial elements in our approach to outperforming tree-based methods is the embedding of categorical variables. Replacing the traditional one-hot encoding, entity embeddings in our study helped to reveal the continuity of categorial variables, thereby retaining informative relations.[17] This approach allows us to map similar features (eg, CPT codes) next to each other in an embedding space. Thereby, we do not treat two procedures, like cholecystectomy and cholecystectomy with cholangiography, as two completely different categories. Instead, our embeddings learn that the two procedures are indeed very similar. Likewise, for example, they learn that a laparoscopic cholecystectomy and a laparoscopic appendicectomy share common features, and that a laparoscopic appendicectomy is closer to an open appendicectomy than an open cholecystectomy. These similarities are mapped in a multidimensional space for all 2941 CPT codes, and, in fact, for all categorical variables. Entity embeddings allowed us to capture complex and multidimensional relationships between categories, which we believe partially explains why our neural networks achieved a higher performance than traditional approaches.

In other studies, researchers have found that adding CPT codes to their models did not improve performance.[6] We believe that this finding reflects the fact that many modern machine learning applications are ill equipped at handling high cardinality features, such as CPT codes, and not that the CPT code itself is irrelevant for risk prediction.[17]

Future studies could further explore the addition of entity embeddings from deep neural networks to other machine learning applications, such as gradient-boosted trees and K-nearest neighbours. Doing so could improve the performance of traditional machine learning techniques.

Despite the promising results, our study has some inherent limitations. First, the ACS NSQIP database is comprised of abstracted clinical registry data; therefore, our models might not be accounting for important unmeasured factors, such as intraoperative data, comorbidity, pharmacological data, and lifestyle data (eg, a sedentary lifestyle), that are not captured in the ACS NSQIP database. Future studies could circumvent this problem by incorporating both additional clinical registry data and unstructured electronic health-care record data, which could potentially increase the granularity of inputs and predictions. Here, deep learning methods have the added advantage of being able to learn directly from unstructured clinical information, reducing the need for manual processing of input features and data. Second, missing data at the preoperative visit might present a challenge, especially in the hyperacute setting. On initial admission to a hospital, patients are likely to have sparse data compared with those presenting with longer treatment histories. Because most of the datapoints we used (eg, diagnoses codes and blood work) are part of a standard preoperative work-up, we believe that this limitation would mostly present in hyperacute settings. Third, although the performances of our models were generally superior to the performance of the ACS-SRC, the size of that difference is sometimes quite small, and the size of the difference in the AUC that represents a clinically significant difference in predictive capabilities is unknown. It is also important to note that the ACS-SRC outperformed our model for risk predictions for deep vein thrombosis and urinary tract infection. As such, combining the predictions from several machine learning model architectures (ensemble machine learning models) could be a more optimal approach. Finally, although all models appeared resistant to changes in the underlying patient population, whether these models will perform comparably well when used for hospitals outside the ACS NSQIP and the USA is still in question. This study was retrospective. The obvious next step would be to conduct a follow-up study assessing model performance on a prospective international cohort.

Our deep neural networks outperformed previously published surgical risk prediction tools and appeared resistant to changes in the underlying patient population. Physicians might benefit from the new models through more accurate preoperative risk assessment, patient counselling, and decision making, and patients might be empowered as stakeholders who can understand and directly alter their personal surgical risk profiles by

making changes to their lifestyle (eg, quitting smoking or exercising more). For future efforts, identifying patients at high risk of postoperative complications for a personalised medicine approach (eg, high-dose thromboprophylaxis for patients at an increased risk of deep vein thrombosis) and including complex biomarkers (eg, the so-called omics) in the prediction models could present a promising strategy to next-generation surgical precision medicine.

Contributors

AB, KMV, NB, AT, OKM, HM, ADY, HA, and MS participated in study conception and formulating the general outline. KMV, AB, NB, and MS developed and tested the model. AB, KMV, and MS accessed and verified the underlying data. AB and MS drafted the original manuscript, which was edited by critical review by KVM, NB, AT, OKM, ADY, HA, and HM. All authors read and approved the final manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

AT reports personal fees and grants from Zimmer Biomet and Pfizer Denmark, outside the submitted work. All other authors declare no competing interests.

Data sharing

The ACS NSQIP database can be requested through the ACS website at https://facs.org. Data recipients must comply with the terms and conditions set forth in the data use agreement, provide contact information, and complete a short online questionnaire. Once this information is received and processed by ACS NSQIP staff, a website address will be submitted electronically to the data recipient. The data recipient will then have 10 days (240 h) to visit the website and download the data file. We have published five Jupyter notebooks: one for data preprocessing, one for each of the three models, and one for calculating personalised risk predictions. With publication, the code and the notebooks will be freely available at https://github.com/alexbonde/NSQIP. For privacy and security reasons, we cannot publish the identification numbers of the patients in our test set.

Acknowledgments

The study was funded by a grant from the Novo Nordisk Foundation (grant #NNF19OC0055183) to MS. The ACS NSQIP and the hospitals participating in the ACS NSQIP are the source of the data used herein; they have not verified, and are not responsible for, the statistical validity of the data analysis or the conclusions derived by the authors.

References

1 Weiser TG, Regenbogen SE, Thompson KD, et al. An estimation of the global volume of surgery: a modelling strategy based on available data. *Lancet* 2008; **372:** 139–44.

2 Ozgediz D, Jamison D, Cherian M, McQueen K. The burden of surgical conditions and access to surgical care in low- and middle-income countries. *Bull World Health Organ* 2008; **86:** 646–47.

3 Healy MA, Mullard AJ, Campbell DA Jr, Dimick JB. Hospital and payer costs associated with surgical complications. *JAMA Surg* 2016; **151:** 823–30.

4 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:** 833–42.e1–3.

5 Meguid RA, Bronsert MR, Juarez-Colunga E, Hammermeister KE, Henderson WG. Surgical Risk Preoperative Assessment System (SURPAS): III. accurate preoperative prediction of 8 adverse outcomes using 8 predictor variables. *Ann Surg* 2016; **264:** 23–31.

6 Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA. 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. *Ann Surg* 2018; **268:** 574–83.

7 Cohn SL, Fernandez Ros N. Comparison of 4 cardiac risk calculators in predicting postoperative cardiac complications after noncardiac operations. *Am J Cardiol* 2018; **121:** 125–30.

8 American College of Surgeons. About the ACS risk calculator. https://riskcalculator.facs.org/RiskCalculator/about.html (accessed June 18, 2021).

9 Lubitz AL, Chan E, Zarif D, et al. American College of Surgeons NSQIP risk calculator accuracy for emergent and elective colorectal operations. *J Am Coll Surg* 2017; **225:** 601–11.

10 Cohen ME, Liu Y, Ko CY, Hall BL. An examination of American College of Surgeons NSQIP surgical risk calculator accuracy. *J Am Coll Surg* 2017; **224:** 787–95.e1.

11 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; **316:** 2402–10.

12 Charoentong P, Finotello F, Angelova M, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep* 2017; **18:** 248–62.

13 Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit Med* 2018; **1:** 18.

14 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:** 1589–604.

15 Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med* 2019; **25:** 24–29.

16 Hinton G. Deep learning—a technology with the potential to transform health care. *JAMA* 2018; **320:** 1101–02.

17 Guo C, Berkhahn F. Entity embeddings of categorical variables. *arXiv* 2016; published online April 22. https://arxiv.org/abs/1604.06737 (preprint).

18 Lei D, Chen X, Zhao J. Opening the black box of deep learning. *arXiv* 2018; published online May 22. https://arxiv.org/abs/1805.08355 (preprint).

19 Lundberg S, Lee S-I. A unified approach to interpreting model predictions. *arXiv* 2017; published online May 22. https://arxiv.org/abs/1705.07874 (preprint).

20 Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018; **2:** 749–60.

21 Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. 2017. https://openreview.net/pdf/25b8eee6c373d48b84e5e9c6e10e7cbbbce4ac73.pdf (accessed June 18, 2021).

22 Howard J, Gugger S, Howard J. fastai: a layered API for deep learning. *Information (Basel)* 2020; **11:** 108.

23 Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; **12:** 2825–30.

24 Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep* 2016; **6:** 26094.

25 Corey KM, Kashyap S, Lorenzi E, et al. Development and validation of machine learning models to identify high-risk surgical patients using automatically curated electronic health record data (Pythia): a retrospective, single-site study. *PLoS Med* 2018; **15:** e1002701.

26 Thorsen-Meyer H-C, Nielsen AB, Nielsen AP, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *Lancet Digit Health* 2020; **2:** e179–91.

27 López-García G, Jerez JM, Franco L, Veredas FJ. Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. *PLoS One* 2020; **15:** e0230536.

28 Kuba K, Imai Y, Rao S, et al. A crucial role of angiotensin converting enzyme 2 (ACE2) in SARS coronavirus-induced lung injury. *Nat Med* 2005; **11:** 875–79.

29 Si Y, Roberts K. Patient representation transfer learning from clinical notes based on hierarchical attention network. *AMIA Jt Summits Transl Sci Proc* 2020; **2020:** 597–606.