Retrospective Cohort Study

INTERNATIONAL JOURNAL OF SURGERY

OPEN

Machine learning based peri-surgical risk calculator for abdominal related emergency general surgery: a multicenter retrospective study

Biao Chen, MD[a], Weiyong Sheng, MD[a,k], Zhixin Wu, MD[a,l], Bingqing Ma, MM[a], Nan Cao, PhD[h], Xushu Li, MD[n], Jia Yang, MD[g], Xiaowei Yuan, MD[i], Lizhao Yan, MD[b], Gaobo Zhu, MD[l], Yuanhong Zhou, MD[m], Zhonghua Huang, MD[n], Meiwei Zhu, MD[n], Xuehui Ding, MD[o], Hansong Du, MD[g], Yanqing Wan, MD[j], Xuan Gao, MD[c], Xing Cheng, MD[d], Peng Xu, MD[a], Teng Zhang, PhD[h], Kaixiong Tao, MD, PhD[e], Xiaoming Shuai, MD, PhD[e], Ping Cheng, MD, PhD[a], Yong Gao, MD, PhD[f], Jinxiang Zhang, MD, PhD[a,*]

Background: Currently, there is a lack of ideal risk prediction tools in the field of emergency general surgery (EGS). The American Association for the Surgery of Trauma recommends developing risk assessment tools specifically for EGS-related diseases. In this study, we sought to utilize machine learning (ML) algorithms to explore and develop a web-based calculator for predicting five perioperative risk events of eight common operations in EGS.

Method: This study focused on patients with EGS and utilized electronic medical record systems to obtain data retrospectively from five centers in China. Five ML algorithms, including Random Forest (RF), Support Vector Machine, Naive Bayes, XGBoost, and Logistic Regression, were employed to construct predictive models for postoperative mortality, pneumonia, surgical site infection, thrombosis, and mechanical ventilation >48 h. The optimal models for each outcome event were determined based on metrics, including the value of the Area Under the Curve, F1 score, and sensitivity. A comparative analysis was conducted between the optimal models and Emergency Surgery Score (ESS), Acute Physiology and Chronic Health Evaluation II (APACHE II) score, and American Society of Anesthesiologists (ASA) classification. A web-based calculator was developed to determine corresponding risk probabilities.

Result: Based on 10 993 patients with EGS, we determined the optimal RF model. The RF model also exhibited strong predictive performance compared with the ESS, APACHE II score, and ASA classification. Using this optimal model, the authors developed an online calculator with a questionnaire-guided interactive interface, catering to both the preoperative and postoperative application scenarios.

Conclusions: The authors successfully developed an ML-based calculator for predicting the risk of postoperative adverse events in patients with EGS. This calculator accurately predicted the occurrence risk of five outcome events, providing quantified risk probabilities for clinical diagnosis and treatment.

Keywords: abdomen, emergency general surgery, machine learning, risk prediction

Introduction

In the past decade, acute care surgery (ACS) has rapidly developed as an emerging specialty[1], comprising subspecialties of emergency general surgery (EGS), trauma surgery, and surgical critical care[2]. The burden of EGS has increased in recent years. Epidemiological data from 2001 to 2010 showed a 28% increase in hospitalizations for EGS in the United States, with over 27 million annual hospitalizations, of which ~28.8% (approximately 7.97 million cases) required emergency surgical

[a]Department of Emergency Surgery, [b]Department of Hand Surgery, [c]Department of Radiology, [d]Health Management Center, [e]Department of Gastrointestinal Surgery, [f]Computer Management Center, Union Hospital, [g]Department of Gastrointestinal Surgery, Central Hospital of Wuhan, Tongji Medical College, [h]School of Computer Science and Technology, [i]Department of Medical, [j]Department of General Surgery, Union Dongxihu Hospital, Huazhong University of Science and Technology, Wuhan, [k]Department of Cardiac Surgery, Yijishan Hospital, Wannan Medical College, Wuhu, [l]Department of Emergency Surgery, [m]Department of Science and Education, Central People's Hospital of Yichang, Three Gorges University, Yichang, [n]Department of General Surgery and [o]Department of Obstetrics and Gynecology, Central Hospital of Hefeng County, Hefeng, People's Republic of China

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

*Corresponding author. Address: Department of Emergency Surgery, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, People Republic of China. Tel.: +862 785 726 301. E-mail: zhangjinxiang@hust.edu.cn (J. Zhang).

International Journal of Surgery (2024) 110:3527–3535

Received 21 December 2023; Accepted 22 February 2024

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.lww.com/international-journal-of-surgery.

Published online 15 March 2024

http://dx.doi.org/10.1097/JS9.0000000000001276

intervention[3]. According to U.S. Census Bureau projections, the incidence of EGS and the public health insurance burden will increase at an annual rate of 45%, reaching $41.2 billion by 2060[4]. Additionally, compared with nonemergency surgeries, patients undergoing emergency surgery face higher risks of mortality and postoperative complications[5,6]. For the same type of procedure, the risk of death in emergency surgery patients can be up to eight times higher than that in elective surgery patients. Even after controlling for preoperative variables and surgical type, emergency surgery has proven to be an independent risk factor for postoperative mortality and complications[6]. Therefore, in this unique patient population, the development of appropriate perioperative risk prediction tools is crucial. These tools would not only guide clinical decision-making but also aid in resource allocation across different medical units and institutions[4].

The field of EGS encompasses a wide spectrum of complex diseases, with patients often presenting diverse physiological disturbances. Additionally, patients with EGS frequently require rapid decision-making owing to the urgency of their condition, yet information may be insufficient. Therefore, the construction of a risk assessment system for EGS faces unique challenges[1]. In 2014, the American Association for the Surgery of Trauma (AAST) recognized the urgency and unique challenges in developing risk assessment tools for the EGS population and conducted a review of existing surgical risk assessment tools[7]. They reviewed existing surgical-related risk assessment tools, including ICU risk assessment systems, the anesthesia-related American Society of Anesthesiologists (ASA) classification system, and other surgical-related risk stratification systems. The ICU risk assessment systems, such as the Acute Physiology and Chronic Health Evaluation II (APACHE II), Simplified Acute Physiologic Score (SAPS), Multiple Organ Dysfunction Score (MODS), Sequential Organ Failure Assessment (SOFA), and Mortality Prediction Model (MPM)[8–10], have been well validated in ICU patient populations but have not been validated specifically in the perioperative period, especially in patients undergoing emergency surgery. Therefore, their use in EGS populations outside of the ICU is not recommended[7]. The anesthesia-related ASA classification system categorizes patients into five grades based on the presence of mild-to-severe life-threatening systemic diseases. Currently, no research has explicitly demonstrated its applicability in the EGS population, and the grading system itself is subjective. Clinical studies conducted by nonanesthesia specialists or those unfamiliar with the classification system may have significant differences in determining the ASA classification[1]. Other surgical-related risk stratification systems, such as the Emergency Surgery Score (ESS), ACS-NSQIP Universal Surgical Risk Calculator, and Predictive Optimal Trees in Emergency Surgery Risk (POTTER) Calculator, have been validated in some centers with emergency surgical patients[11–13]. However, their predictive efficacy in specific EGS populations requires further validation. Based on a review of these various risk assessment tools, AAST developed and published the AAST EGS grading system for 16 common EGS diseases in 2016. Future research should explore the construction of relevant risk assessment tools by combining patient age, acute physiology, and comorbidity status[1,14,15].

Risk prediction tools can facilitate the accurate identification of high-risk patients, thereby guiding surgical decision-making, informed consent, or referral to units or institutions with higher levels of medical resource allocation. Given the broad spectrum of EGS diseases, lack of predictability in diagnosis and treatment, and absence of specialized teams trained in EGS, medical adverse events are more likely to occur[4]. An ideal EGS risk assessment system would be able to quickly and accurately predict the risk of death and postoperative complications in the early stages of patient management and guide the allocation of medical resources to improve quality of care. However, an ideal EGS risk prediction tool has not yet been developed[4]. In this study, we developed five algorithm models using machine learning (ML) algorithms that can predict the risk of five adverse events for eight EGS-related surgeries. Based on these models, we designed a web-based calculator that provides quantifiable risk probabilities for the diagnosis and treatment of such patients in a clinical setting with the aim of optimizing the quality of care.

HIGHLIGHTS

- Risk prediction tools is crucial to facilitate the accurate identification of high-risk patients, thereby guiding surgical decision-making, informed consent, or referral to units or institutions with higher levels of medical resource allocation.
- No accurate tool were developed to assess the risk of postoperative outcome events in emergency general surgery (EGS).
- Employed five machine learning algorithms to develop models for predicting five outcome events of EGS, and selected the optimal model - Random Forest.
- Developed five machine-learning-based predictive models for perioperative risk in EGS.
- Developed a web-based risk prediction calculator for patients undergoing EGS, providing quantified risk probabilities for five outcome events, including postoperative mortality, pneumonia, SSI, thrombosis, and mechanical ventilation > 48 h.

Methods

Data sources and patients

This study was approved by the Ethics Committee of Union Hospital, Tongji Medical College, Huazhong University of Science and Technology (Approval No. 2020 0516-01), and registered in Chinese Clinical Trial Registry (ChiCTR 2000039772). All participating sub-centers' ethics committees have approved this study.

According to the anatomical grading system for 16 common EGS diseases developed by the AAST, this study focused on eight diseases in the spectrum of abdominal emergency surgical patients: acute appendicitis, acute cholecystitis, hernia (including intra-abdominal and extra-abdominal hernias), acute intestinal obstruction, peptic ulcer perforation (gastric or duodenal ulcer perforation), mesenteric arterial thrombosis, acute colonic diverticulitis, and ulcerative colitis. This study included five centers, including Wuhan Union Hospital, Central Hospital of Wuhan, Central People's Hospital of Yichang, Union Dongxihu Hospital, and Central Hospital of Hefeng County. Using current diagnostic codes, all patients (> 18 years old) who underwent EGS between 2012 and 2022 were retrieved from the Electronic Medical Records System (EMRS). Patients who underwent EGS within 24 h were included in this study[16]. To minimize the complexity of the risk prediction model, specific calculators for individual surgeries were not included in this study. Furthermore, patients with known pregnancies, infectious diseases under fixed-point treatment, and those lacking surgical and discharge records were excluded from this study. This retrospective study has been reported in line with the strengthening the reporting of cohort, cross-sectional, and case–control studies in surgery (STROCSS) criteria[17] (Supplemental Digital Content 1, http://links.lww.com/JS9/C90).

Prediction variable selection

Seventy percent (7695 cases) of the cases were randomly selected and included in the derivation cohort, whereas the remaining 30% (3298 cases) were used for model testing. To improve the accuracy of the model, variables were selected from five dimensions, including age, anatomy, physiology, comorbidities, and surgery-related factors. Considering the diagnostic and treatment practices at each center and the variable definitions from the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) database, 33 research variables were defined. These variables included age, sex, BMI, history of alcohol consumption, smoking history, hypertension, diabetes, chronic obstructive pulmonary disease, renal disease history, previous abdominal surgery, neurological disorders, tumor, cardiac history, hormone use, preoperative sepsis, white blood cell (WBC), red blood cell (RBC), hemoglobin (Hb), platelet (PLT), hematocrit (Hct), alanine aminotransferase (ALT), aspartate aminotransferase (AST), total bilirubin (TBIL), albumin (ALB), creatinine (Crea), blood urea nitrogen (BUN), sodium ion (Na^+), potassium ion (K^+), total carbon dioxide (TCO_2), prothrombin time (PT), international normalized ratio (INR), AAST EGS Grade, and ASA classification. Regarding the definition of postoperative outcome events, Scarborough *et al.*[18] analyzed complication events in 7.91 million EGS surgical patients and found that the most common complications were postoperative bleeding (6.2%), surgical site infection (SSI) (3.4%), postoperative pneumonia (2.7%), postoperative urinary tract infection (1.5%), postoperative thromboembolism (1.1%), and postoperative myocardial infarction (0.5%). Based on previous reports and the research data, we ultimately defined five outcome variables: postoperative mortality, pneumonia, SSI, thrombotic events, and mechanical ventilation > 48 h[12,13,19,20]. The occurrences of death, postoperative SSI, and mechanical ventilation > 48 h refer to events that happened during the in-hospital postoperative period and were explicitly documented in medical records. Pneumonia and thrombotic events were diagnosed as such through imaging examinations after excluding their presence before surgery. Supplemental Table 1 (Supplemental Digital Content 2, http://links.lww.com/JS9/C91) shows the distribution of the demographic and clinical characteristics. We present the five hospitals using quartiles categorized into low, medium, high, and very high volume hospitals. Regarding the selection of variable dimensions for the model, the concept of anatomical and physiological construction has already been applied in trauma risk assessment models, such as the Trauma and Injury Severity Score (TRISS)[21]. In this study, we integrated the severity grading of disease anatomy with demographic, physiological, comorbidity, and surgery-related parameters based on existing literature.

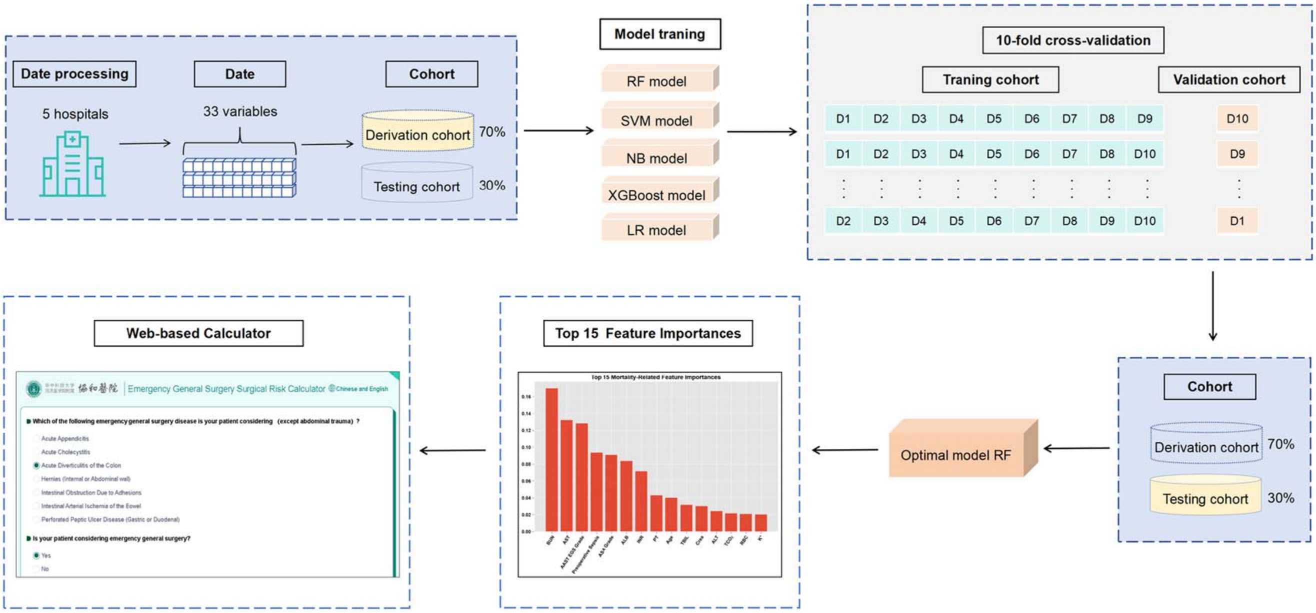
Data preprocessing

Because the data originated from various real-world hospital databases, they inevitably contained some missing values (1.54%, 6782/439720). To address this, we followed common data analysis practices described in the literature by filling missing values with the median of the corresponding feature[22,23]. Additionally, we performed min–max normalization on the entire dataset by mapping all features to the range of [0, 1] using the maximum and minimum values of each feature. This normalization eliminates scale differences among different features and ensures comparability and analysis within the same numerical range, thereby enhancing the performance of the algorithm model.

In ML modeling, a sample class ratio lower than 9:1 is generally considered imbalanced. In our dataset, the proportion of each outcome event was significantly lower than this ratio, indicating an extremely unbalanced class distribution. To address this issue, we employed various techniques such as up-sampling with SMOTE, Borderline SMOTE, random under-sampling, TomekLink under-sampling, and a combination of SMOTE and down-sampling methods. Ultimately, we determined that random under-sampling provided the highest recognition accuracy for handling imbalanced data; therefore, we used this method to obtain a balanced dataset for modeling. This approach involves randomly removing samples from the majority class to achieve class balance.

Development of model

Figure 1 presents the model construction process, while Supplemental File 1 (Supplemental Digital Content 3, http://links.lww.com/JS9/C92) presents the code involved in model processing. Building a prognostic model for disease outcomes requires the consideration of individual-specific time-point risks, which is a classical ML classification problem. Commonly used modeling algorithms include Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGBoost)[24,25]. In this study, the data were divided into the derivation and testing cohorts. The derivation cohort was used for model development, whereas the testing cohort was used to evaluate the generalization performance of the model. Five ML algorithms were employed to develop models for predicting five outcome events: postoperative mortality, pneumonia, SSI, thrombotic events, and mechanical ventilation for > 48 h. The models were evaluated using 10-fold cross-validation to improve performance and enhance generalization ability. The feature importance was studied and ranked using the RF model to explore the importance of various features in real-world scenarios. The effectiveness of the modeling was tested based on the top five, 10, 15, and 20 important features. The results showed that using the top 15 important features yielded a modeling performance comparable to that obtained using all features. Therefore, graphical representations of the top 15 ranked features by importance were presented for each outcome event. Using a prebuilt optimal algorithm model, we developed a web interface based on questionnaire-guided responses (https://www.bychjh.com/severe-calculator-pc/#/). This interface allows clinicians to perform real-time online calculations of risk probabilities for different prognostic outcomes in patient populations with relevant diseases.



Emergency General Surgery Surgical Risk Calculator 🌐Chinese and English

Which of the following emergency general surgery disease is your patient considering (except abdominal trauma) ?

○ Acute Appendicitis
○ Acute Cholecystitis
● Acute Diverticulitis of the Colon
○ Hernias (Internal or Abdominal wall)
○ Intestinal Obstruction Due to Adhesions
○ Intestinal Arterial Ischemia of the Bowel
○ Perforated Peptic Ulcer Disease (Gastric or Duodenal)

Is your patient considering emergency general surgery?

● Yes
○ No

Figure 1. Model construction process.

Statistical analysis

Data analysis was performed using R Studio 3.4.3 and Python 3.8, utilizing various packages, including XGBoost, GLM, and MI. Graphs were generated using R or GraphPad Prism, version 9.2.0. Continuous variables were described as mean ± SD (X ± SD) or median interquartile range (IQR), and between-group comparisons were conducted using the Student's *t*-test or Mann–Whitney *U* test. Categorical variables are presented as frequencies (%), and between-group comparisons were assessed using Pearson's χ^2 test or Fisher's exact test. Statistical significance was defined as $P < 0.05$. We evaluated the model performance and selected the optimal model using metrics including the Area Under the Curve (AUC), F1 score, sensitivity, specificity, accuracy, and precision. Given the focus of this study on the model's ability to identify endpoint events in medical data, where the proportion of endpoint events is small, we placed greater emphasis on the model's performance in terms of the AUC value, F1 score, and sensitivity. The corresponding risk probabilities were presented quantitatively using an interactive web-based calculator.

Results

Study population

Baseline characteristics of the derivation and testing cohorts are presented in Table 1. A total of 10 993 EGS patients from five medical institutions in China were included in this study. Among them, there were 153 (1.39%) cases of postoperative in-hospital death, 632 (5.75%) cases of postoperative pneumonia, 531 (4.83%) cases of postoperative SSI, 616 (5.60%) cases of postoperative thrombosis, and 674 (6.13%) cases with mechanical ventilation > 48 h. The dataset was randomly divided into a derivation cohort of 7695 cases and testing cohort of 3298 cases in a 7:3 ratio.

Postoperative in-hospital mortality

Table 2 compares the performances of the five models in the testing cohort. Figure 2 shows the ROC curves of the five models for predicting in-hospital mortality. The RF model had the highest AUC value (0.8961) and F1 score in the testing set, indicating its superior discriminative ability for postoperative mortality events. In addition, the RF model exhibited the highest sensitivity and precision, suggesting strong identification capabilities for mortality events. Therefore, the RF model demonstrated the best predictive ability of postoperative in-hospital mortality. Regarding the prediction of in-hospital mortality, the top five relevant features according to the variable importance distribution of the RF model were BUN level, preoperative sepsis status, AAST EGS grading, ALB, and age (Fig. 3).

Postoperative complications

We employed five algorithms to model four postoperative outcomes: postoperative pneumonia, SSI, thrombosis, and mechanical ventilation > 48 h. Table 3 presents the performance of these five models in the testing cohort. The RF model demonstrated higher AUC values (Fig. 2) and F1 scores than the other four models, indicating its superior classification ability. Although the RF model exhibited lower sensitivity than the XGBoost model for postoperative pneumonia and mechanical ventilation > 48 h, it displayed a better AUC value, F1 score, accuracy, and precision, suggesting a stronger capability to classify and identify these events. Therefore, the RF model performed better in predicting postoperative pneumonia and mechanical ventilation > 48 h. For the outcome of postoperative SSI, the RF model showed the same sensitivity as the XGBoost model, but outperformed the other four models in terms of AUC value, F1 score, accuracy, and precision, indicating its superior classification and recognition capability. Regarding postoperative thrombosis, the RF model demonstrated stronger classification and outcome recognition

Table 1

Baseline characteristic in the derivation and validation cohort.

	All (n = 10 993)	Derivation cohort (n = 7695)	Testing cohort (n = 3298)	P
Age (years), median (IQR)	53.00 (30.00)	53.00 (31.00)	53 (29.00)	0.2882
Sex (n, %)				
Male	5864 (53.34)	4097 (53.24)	1767 (53.58)	0.7466
Female	5129 (46.66)	3598 (46.76)	1531 (46.42)	
BMI, median (IQR)	21.26 (4.00)	21.25 (3.68)	21.28 (3.89)	0.3315
Drink (n, %)	971 (8.83)	649 (8.43)	322 (9.76)	0.0244
Smoke (n, %)	1158 (10.53)	795 (10.33)	363(11.01)	0.2906
Abdominal surgery history (n, %)	2222 (20.21)	1560 (20.27)	662(20.07)	0.8108
Hypertension (n, %)	1299 (11.82)	899 (11.68)	400 (12.13)	0.5071
Diabetes (n, %)	712 (6.48)	497 (6.46)	215 (6.52)	0.9062
COPD (n, %)	136 (1.24)	91 (1.18)	45 (1.36)	0.4292
Chronic kidney disease (n, %)	168 (1.53)	123 (1.60)	45 (1.36)	0.3594
Nervous system disease (n, %)	196 (1.78)	136 (1.77)	60 (1.82)	0.8505
Cancer (n, %)	437 (3.98)	304 (3.95)	133 (4.03)	0.8399
Cardiovascular disease (n, %)	533 (4.85)	377 (4.90)	156 (4.73)	0.7052
Steroid use (n, %)	101 (0.92)	74 (0.96)	27 (0.82)	0.4715
Surgical sepsis				
None (n, %)	10167 (92.49)	7127 (92.62)	3040 (92.18)	0.2146
Sepsis only (n, %)	559 (5.09)	394 (5.12)	165 (5.00)	
Septic shock (n, %)	267 (2.43)	174 (2.26)	93 (2.82)	
Laboratory test, median (IQR)				
WBC (10^9 /l)	9.20 (6.81)	9.20 (6.79)	9.20 (6.88)	0.8410
RBC (10^9 /l)	4.29 (0.84)	4.29 (0.84)	4.29 (0.86)	0.9879
PLT (10^9 /l)	198.00 (85.50)	198.00 (86.00)	197.00 (84.00)	0.2929
Hct (%)	38.70 (7.30)	38.80 (7.30)	39.60 (7.40)	0.5308
Hb (g/l)	129.00 (27.00)	129.00 (27.00)	129.00 (27.00)	0.5146
ALT (U/l)	20.00 (17.00)	20.00 (17.00)	20.00 (17.00)	0.7527
AST (U/l)	21.00 (11.00)	21.00 (10.00)	21.00 (11.00)	0.3773
TBil (μmol/l)	16.37 (12.60)	16.31 (12.60)	16.40 (12.70)	0.8059
ALB (g/l)	39.70 (7.00)	39.62 (7.00)	39.76 (7.10)	0.4501
Crea (μmol/l)	69.30 (26.50)	69.30 (26.90)	69.30 (25.80)	0.4927
BUN (mmol/l)	5.25 (2.90)	5.27 (2.90)	5.20 (2.92)	0.8997
Na^+ (mmol/l)	139.80 (4.00)	139.80 (4.00)	139.80 (3.90)	0.6138
K^+ (mmol/l)	3.91 (0.54)	3.91 (0.54)	3.92 (0.52)	0.9869
TCO_2 (mmol/l)	23.30 (3.50)	23.30 (3.50)	23.30 (3.50)	0.3941
PT (s)	13.60 (1.80)	13.60 (1.80)	13.60 (1.90)	0.1379
INR	1.08 (0.18)	1.08 (0.18)	1.08 (0.18)	0.3484
AAST EGS Grade (n, %)				
1	6073 (55.24)	4274 (55.54)	1799 (54.55)	0.7571
2	1151 (10.47)	811 (10.54)	340 (10.31)	
3	1501 (13.65)	1032 (13.41)	469 (14.22)	
4	1093 (9.94)	763 (9.29)	330 (10.01)	
5	1175 (10.69)	815 (10.59)	360 (10.92)	
ASA classification (n, %)				
1	658 (5.99)	441 (5.73)	217 (6.58)	0.1999
2	7358 (66.93)	5193 (67.49)	2165 (65.65)	
3	2796 (25.43)	1939 (25.20)	857 (25.99)	
4	167 (1.52)	111 (1.44)	56 (1.70)	
5	14 (0.13)	11 (0.14)	3 (0.09)	
ESS, median (IQR)	2.00 (2.00)	2.00 (2.00)	2.00 (2.00)	0.6812
APACHE II, median (IQR)	4.00 (4.00)	4.00 (4.00)	4.00 (4.00)	0.9143
In-hospital events				
Postoperative Mortality (n, %)	153 (1.39)	107 (1.39)	46 (1.39)	0.9860
Postoperative pneumonia (n, %)	632 (5.75)	432 (5.61)	200 (6.06)	0.3527
Postoperative SSI (n, %)	531 (4.83)	390 (5.07)	141 (4.28)	0.0756
Postoperative thrombosis (n, %)	616 (5.60)	427 (5.55)	189 (5.73)	0.7043
Postoperative mechanical ventilation > 48 h (n, %)	674 (6.13)	470 (6.11)	204 (6.19)	0.8763

AAST, American Association for the Surgery of Trauma; ALB, albumin; ALT, alanine aminotransferase; APACHE II, Acute Physiology and Chronic Health Evaluation II; ASA, American Society of Anesthesiologists; AST, aspartate aminotransferase; BUN, blood urea nitrogen; COPD, chronic obstructive pulmonary disease; Crea, creatinine; EGS, emergency general surgery; ESS, Emergency Surgery Score; Hb, hemoglobin; Hct, hematocrit; INR, international normalized ratio; IQR, interquartile range; K^+, potassium ion; Na^+, sodium ion; PLT , platelet; PT, prothrombin time; RBC, red blood cell; SSI, surgical site infection; TBIL, total bilirubin; TCO_2, total carbon dioxide; WBC, white blood cell.

capabilities. Figure 3 provides a list of the top 15 most relevant features associated with the variable importance distribution of the RF model for predicting postoperative pneumonia, SSI, thrombosis, and mechanical ventilation > 48 h events.

Comparison of RF model with ESS, APACHE II score, and ASA classification

Table 4 presents a performance comparison of the RF model with the ESS, APACHE II score, and ASA classification. In predicting postoperative thrombotic events, the RF model demonstrated a significantly higher AUC value, F1 score, and sensitivity than ESS, APACHE II score, and ASA classification (Table 4), which are commonly utilized scoring tools in literature reports and clinical practice. This indicates the better discriminative ability and outcome prediction capability of the RF model. In addition, the RF model accurately identified cases in which the outcome event did not occur. Therefore, compared with other models, the RF model shows superior performance in identifying outcome events and achieving higher accuracy in their recognition.

Questionnaire-guided interactive interface online calculator

The RF model was ultimately selected as the algorithm tool for the web-based calculator on our website (http://athenaai.jvlei.

Table 2

Performance evaluation of five algorithm models in predicting postoperative in-hospital mortality.

	Testing cohort				
Metrics	**XGBoost**	**RF**	**LR**	**SVM**	**NB**
AUC	0.8504	0.8961	0.7893	0.7705	0.8024
F1	0.1214	0.1284	0.0955	0.1235	0.0904
Se	0.8500	0.9500	0.7500	0.6500	0.8000
Sp	0.8508	0.8422	0.8287	0.8910	0.8047
Pr	0.0654	0.0688	0.0510	0.0682	0.0479
Acc	0.8508	0.8435	0.8277	0.8881	0.8047

Acc, Accuracy; AUC, Area Under the Curve; F1, F1 score; LR, Logistic Regression; NB, Naive Bayes; Pr, Precision; RF, Random Forest; Se, Sensitivity; Sp, Specificity; SVM, Support Vector Machine; XGBoost, eXtreme Gradient Boosting.

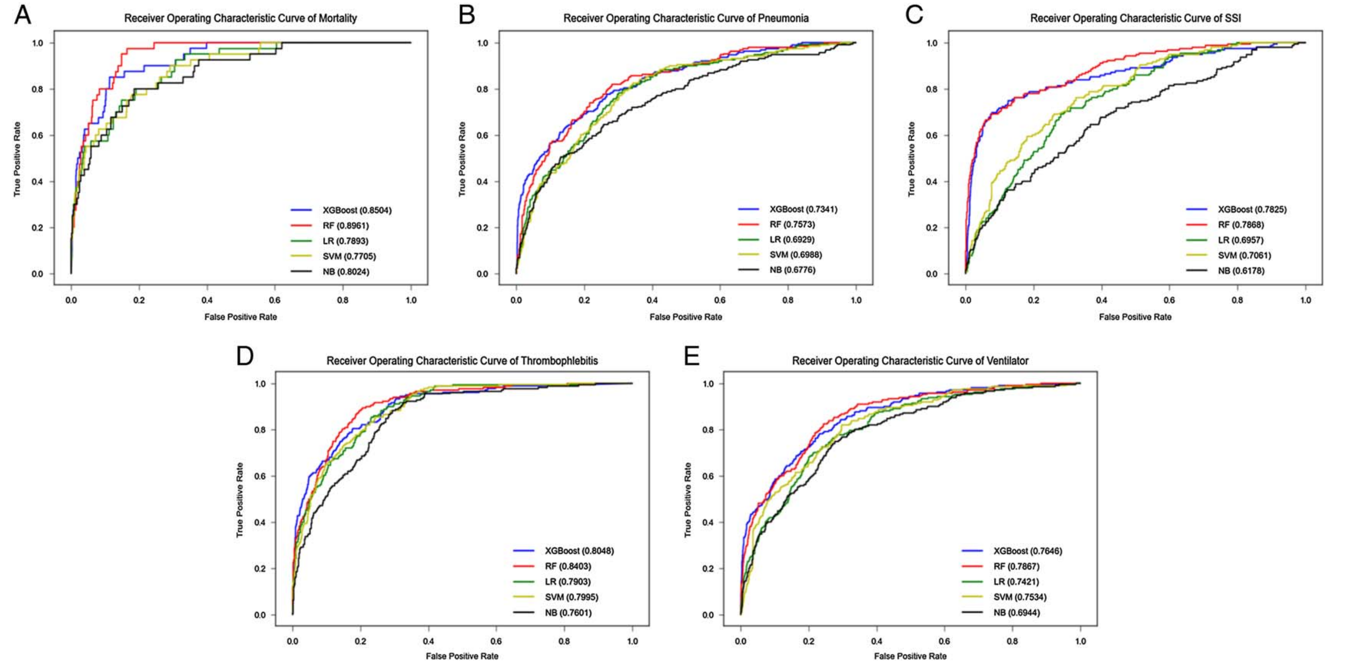


Figure 2. ROC curves of five algorithm models predicting five outcome events in the test cohort. A, Postoperative Mortality. B, Postoperative Pneumonia. C, Postoperative SSI. D, Postoperative Thrombosis. E, Postoperative Mechanical Ventilation > 48 h.

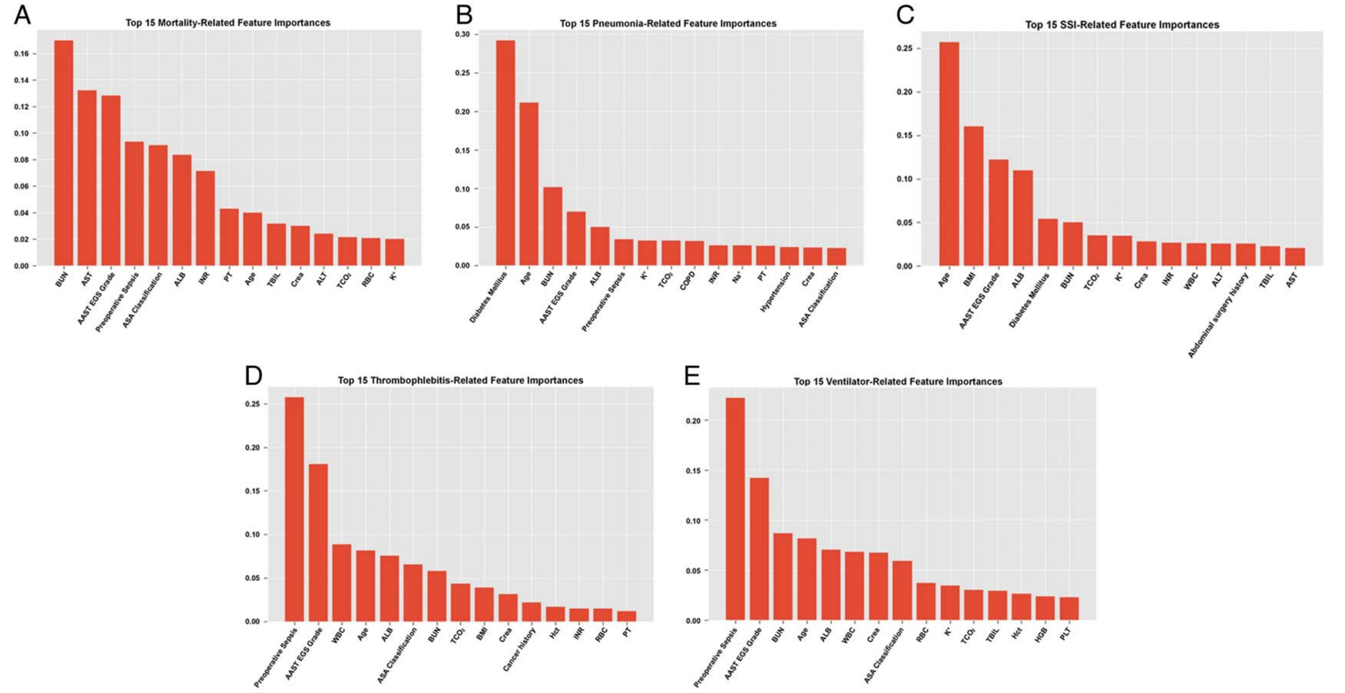


Figure 3. The relative importance ranking of predictive variables for five outcome events in the optimal model RF. A, For postoperative mortality-related events, the top 5 most relevant features in the distribution of variable importance are BUN, AST, preoperative sepsis, AAST EGS Grade and ALB. Following them are ALB, INR, PT, age, TBIL, Crea, ALT, TCO_2, RBC and K^+. B, For postoperative pneumonia events, the top 5 most relevant features in the distribution of variable importance are diabetes mellitus, age, BUN, AAST EGS Grade and ALB. Following them are preoperative sepsis, K^+, TCO_2, chronic obstructive pulmonary disease (ASA), INR, Na^+, PT, hypertension, Crea and ASA Grade. C, For postoperative SSI outcome events, the top 5 most relevant features in the distribution of variable importance are age, BMI, AAST EGS Grade, ALB, and diabetes mellitus. Following them are BUN, TCO_2, WBC, K^+, Crea, INR, WBC, ALT, abdominal surgery history, TBIL and AST. D, For postoperative thrombosis events, the top 5 most relevant features in the distribution of variable importance are preoperative sepsis, AAST EGS Grade, WBC, age, and ALB. Following them are ASA Grade, BUN, TCO_2, BMI, Crea, cancer history, Hct, INR, RBC, and PT. E, For postoperative events with Ventilation > 48 h, the top 5 most relevant features in the distribution of variable importance are preoperative sepsis, AAST EGS Grade, BUN, age, and ALB. Following them are WBC, Crea, ASA Grade, RBC, K^+, TCO2, TBIL, Hct, and PLT.

Table 3

Evaluation of predictive performance of five algorithm models for postoperative complications.

		Testing cohort				
Complication	Metric	XGBoost	RF	LR	SVM	NB
Pneumonia	AUC	0.7341	0.7573	0.6929	0.6988	0.6776
	F1	0.2175	0.2710	0.2374	0.2497	0.2363
	Se	0.8085	0.7394	0.5904	0.5851	0.5372
	Sp	0.6597	0.7752	0.7954	0.8125	0.8179
	Pr	0.1256	0.1659	0.1486	0.1587	0.1514
	Acc	0.6682	0.7731	0.7837	0.7995	0.8019
SSI	AUC	0.7825	0.7868	0.6957	0.7061	0.6178
	F1	0.2470	0.2539	0.1726	0.1884	0.1424
	Se	0.7935	0.7935	0.7161	0.6903	0.5226
	Sp	0.7715	0.7801	0.6754	0.7218	0.7129
	Pr	0.1463	0.1511	0.0981	0.1091	0.0824
	Acc	0.7725	0.7807	0.6773	0.7204	0.7040
Thrombosis	AUC	0.8048	0.8403	0.7903	0.7995	0.7601
	F1	0.2676	0.3180	0.2951	0.2816	0.2557
	Se	0.8764	0.8933	0.7809	0.8315	0.7584
	Sp	0.7332	0.7874	0.7996	0.7676	0.7618
	Pr	0.1579	0.1934	0.1819	0.1695	0.1538
	Acc	0.7410	0.7931	0.7986	0.7710	0.7616
Ventilation > 48 h	AUC	0.7646	0.7867	0.7421	0.7534	0.6944
	F1	0.2512	0.2864	0.2642	0.2556	0.2551
	Se	0.8762	0.8524	0.7524	0.8190	0.6000
	Sp	0.6531	0.7211	0.7318	0.6877	0.7888
	Pr	0.1466	0.1721	0.1602	0.1514	0.1620
	Acc	0.6673	0.7295	0.7331	0.6961	0.7768

com/severe-calculator-pc/#/), considering its superior predictive performance compared with the XGBoost, SVM, NB, and LR models. The calculator utilizes a questionnaire-guided approach, allowing clinicians to calculate real-time risk probabilities for five selected common adverse outcomes. This tool is suitable for preoperative and intraoperative scenarios and enables predictions under different circumstances. Patients and their families can make informed decisions regarding surgery based on risk probabilities, and prepare for possible postoperative complications. Clinicians can tailor the postoperative treatment plans based on the indicated risks for each outcome. During the intraoperative procedure, risk prediction probabilities can be updated based on the AAST EGS anatomical grading system, facilitating immediate communication between medical professionals and informed decisions regarding the postoperative ICU transition. The web interface provides options in both Chinese and English, making it accessible to a broader population (Supplemental Fig. 1, Supplemental Digital Content 4, http://links.lww.com/JS9/C93). As future multicenter studies will collect more extensive data, the model developed in this research can be further optimized and updated. It can also be integrated into the EMRS, enabling automated risk probability prediction immediately after patient examination.

Table 4

Comparison between RF Model and ESAS, APACHE II, ASA.

		Testing cohort			
Complication	Metric	RF	ESS	APACHE II	ASA
Death	AUC	0.8961	0.7510	0.7147	0.6747
	F1	0.1284	0.0679	0.0692	0.0541
	Se	0.9500	0.6087	0.5000	0.5652
Pneumonia	AUC	0.7341	0.6497	0.6337	0.5609
	F1	0.2175	0.1758	0.1625	0.1320
	Se	0.9500	0.4000	0.2450	0.3550
SSI	AUC	0.7864	0.6110	0.5991	0.5964
	F1	0.2497	0.0868	0.0654	0.1162
	Se	0.8258	0.1773	0.1064	0.4255
Thrombosis	AUC	0.8048	0.7074	0.6955	0.7131
	F1	0.2676	0.2115	0.2226	0.2140
	Se	0.8933	0.5238	0.4286	0.6349
Ventilation > 48 h	AUC	0.7646	0.6620	0.6471	0.6710
	F1	0.2512	0.2025	0.1743	0.2147
	Se	0.8524	0.4706	0.3431	0.5735

Discussion

We developed an emergency surgical risk-prediction model for a common spectrum of diseases in EGS. The model was developed using a large dataset that covered different geographical locations and considered variations in healthcare resource allocation. Five ML algorithms were used to construct predictive models for selected common outcome events during EGS. The predictive performances of the models were compared, and our results demonstrated that the RF model had a better predictive ability for all endpoint events. The RF model belongs to the supervised learning category, and is based on decision tree models, which are tree-like predictive models. Each branch of the tree corresponds to a feature split, and each leaf node represents a set of samples that satisfy all constraints along the path. The constraints along the path can be regarded as rules, providing good interpretability for the decision tree results. The decision tree model selects the field with the maximum information gain in the data samples as the node of the tree, and establishes different branch nodes based on different field values. This process was repeated for each branch to form a decision tree[26].

Based on the RF model, we designed a web-based interactive interface that allows the inclusion of links on mobile devices. This interface provides real-time, convenient access to clinical emergency physicians.

The ESS has been considered suitable for risk prediction in emergency surgical patients, whereas the APACHE II score and ASA classification are commonly used in clinical practice. Compared to these models, our model still demonstrated an excellent discriminative ability. We also used the POTTER calculator, which is based on ML algorithms for emergency surgical populations in developed countries. However, the POTTER calculator has limitations that render it unsuitable for clinical practice in China. These limitations include discrepancies in unit conversion for laboratory data, lack of support for input parameters specific to the Chinese population, and the inability to accommodate certain admission pathways and referral parameters, such as transfers from community hospitals or emergency departments[13]. Additionally, the inclusion of hospitals in the ACS-NSQIP database requires conditional screening by the American College of Surgeons[27], making it difficult for hospitals in resource-limited areas to contribute data to the database. This restricts the widespread applicability of POTTER calculators.

The limitations of this study lie primarily in the fact that the predictive ability of ML depends on the accuracy and comprehensiveness of the data used. Although we employed data filtering and extraction based on EMRS to avoid biases introduced by manual data entry, variations in the scope of data selection and the accuracy of real-world data from different centers can still introduce biases. The sampling method using ML in this study specifically targeted retrospective big data for the intervention. Additionally, we introduced disease anatomical severity grading as an indicator, which helped to define the spectrum of diseases studied. Multiple studies have shown good consistency in the grading criteria before, during, and after surgery[28–30], greatly enhancing the applicability of this indicator. However, in practical applications, familiarity with this grading system is required by clinicians, and it relies on subjective judgment, introducing a certain degree of operator bias. To reflect patient-related nutritional indicators such as nutritional scoring scales and weakness indices, it was not possible to calculate these indicators during retrospective data collection and relied on data from future prospective studies.

Finally, our study is not only based on clinical questions, but also explores the existing gaps in the development of the EGS discipline. The ultimate significance of constructing a clinical prediction model lies in observing whether the use of this model in clinical practice changes physicians' and patients' diagnostic and treatment behaviors, improves patient outcomes, and is cost effective. This is what we refer to as the impact study of a clinical prediction model.

Conclusions

In this study, we combined multicenter clinical big data from different geographical locations with artificial intelligence algorithms to develop a web-based risk prediction calculator for patients undergoing emergency surgery for abdominal diseases in general surgery. This tool provides quantified risk probability for events, including postoperative mortality, pneumonia, SSI, thrombosis, and mechanical ventilation > 48 h. Furthermore, it can be integrated into healthcare systems to enhance its practicality.

Ethical approval

This study was approved by the Ethics Committee of Union Hospital, Tongji Medical College, Huazhong University of Science and Technology (Approval No. 2020 0516-01).

Consent

This is a retrospective study, conducted based on existing medical records or data for analysis and investigation. Therefore, we have waived the process of obtaining informed consent. However, we strictly adhere to ethical principles and requirements to protect the rights and interests of the participants. The data has been anonymized to ensure that it does not directly involve the privacy and personal information of the participants.

Sources of funding

This study was supported by the Clinical Research Fund of Union Hospital, Tongji Medical College, Huazhong University of Science and Technology (2021xhlcyj13). The funding body was not involved in the design of the study, the collection, analysis and interpretation of the data, or the writing of the paper.

Author contribution

Z.-J.X.: conceived and designed this study, promoted communication, and collaboration among multiple centers, and provided guidance throughout the analysis and writing process; C.-B. and S.-W.Y.: participated in the project design and collection of data from multiple centers, contributed to the data analysis and interpretation, and completed the writing of the manuscript; W.-Z.X.: participated in data collection and critically revised the manuscript. As experts in computer science, C.-N., Y.-L.Z., G.-X., and Z.-T.: provided guidance and assistance in building models; M.-B.Q., C.-X., X.-P., T.-K.X., S.-X.M., C.-P., and G.-Y.: assisted with data collection and critically revised the manuscript; L.-X.S., Y.-J., Y.-X.W., Z.-G.B., Z.-Y.H., H.-Z.H., Z.-M.W., D.-X.H., D.-H.S., and W.-Y.Q.: were important personnel involved in the sub-center data collection. C.-B., S.-W.Y., and W.-Z.X.: contributed equally to this work.

Conflicts of interest disclosure

The authors declare that they have no financial conflicts of interest with regard to the content of this report.

Research registration unique identifying number (UIN)

This was registered in Chinese Clinical Trial Registry (ChiCTR2000039772).

Guarantor

Jinxiang Zhang is the guarantor of this work.

Data availability statement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Provenance and peer review

This paper was not published by invitation.

Presentation

None.

Acknowledgements

Assistance with the study: none.

References

[1] Shafi S, Aboutanos M, Brown CV, *et al*. Measuring anatomic severity of disease in emergency general surgery. J Trauma Acute Care Surg 2014;76: 884–7.

[2] Davis KA, Rozycki GS. Acute care surgery in evolution. Crit Care Med 2010;38(9 suppl):S405–10.

[3] 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;77:202–8.

[4] Havens JM, Neiman PU, Campbell BL, *et al*. The future of emergency general surgery. Ann Surg 2019;270:221–2.

[5] Ingraham AM, Cohen ME, Bilimoria KY, *et al*. Comparison of 30-day outcomes after emergency general surgery procedures: potential for targeted improvement. Surgery 2010;148:217–38.

[6] Havens JM, Peetz AB, Do WS, *et al*. The excess morbidity and mortality of emergency general surgery. J Trauma Acute Care Surg 2015;78: 306–11.

[7] Havens JM, Columbus AB, Seshadri AJ, *et al*. Risk stratification tools in emergency general surgery. Trauma Surg Acute Care Open 2018;3:e000160.

[8] Vincent JL, Moreno R, Takala J, *et al*. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med 1996;22:707–10.

[9] Knaus WA, Draper EA, Wagner DP, *et al*. APACHE II: a severity of disease classification system. Crit Care Med 1985;13:818–29.

[10] Vincent JL, Moreno R. Clinical review: scoring systems in the critically ill. Crit Care 2010;14:207.

[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;81:213–20.

[12] 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.e831-833.

[13] Bertsimas D, Dunn J, Velmahos GC, *et al*. 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.

[14] Tominaga GT, Staudenmayer KL, Shafi S, *et al*. The American Association for the Surgery of Trauma grading scale for 16 emergency general surgery conditions: Disease-specific criteria characterizing anatomic severity grading. J Trauma Acute Care Surg 2016;81:593–602.

[15] Crandall ML, Agarwal S, Muskat P, *et al*. Application of a uniform anatomic grading system to measure disease severity in eight emergency general surgical illnesses. J Trauma Acute Care Surg 2014;77:705–8.

[16] Levtzion-Korach O, Murphy KG, Madden S, *et al*. For urgent and emergent cases, which one goes to the OR first? OR Manager 2010;26:11–3.

[17] Mathew G, Agha R, Albrecht J, *et al*. STROCSS 2021: strengthening the reporting of cohort, cross-sectional and case-control studies in surgery. Int J Surg 2021;96:106165.

[18] Scarborough JE, Schumacher J, Pappas TN, *et al*. Which complications matter most? prioritizing quality improvement in emergency general surgery. J Am Coll Surg 2016;222:515–24.

[19] Coccolini F, Catena F, Pisano M, *et al*. Open versus laparoscopic cholecystectomy in acute cholecystitis. Systematic review and meta-analysis. Int J Surg 2015;18:196–204.

[20] Lambrichts DPV, Vennix S, Musters GD, *et al*. Hartmann's procedure versus sigmoidectomy with primary anastomosis for perforated diverticulitis with purulent or faecal peritonitis (LADIES): a multicentre, parallel-group, randomised, open-label, superiority trial. Lancet Gastroenterol Hepatol 2019;4:599–610.

[21] Boyd CR, Tolson MA, Copes WS. Evaluating trauma care: the TRISS method. Trauma Score and the Injury Severity Score. J Trauma 1987;27:370–8.

[22] Ponraj DN, Jenifer ME, Poongodi P, *et al*. A survey on the preprocessing techniques of mammogram for the detection of breast cancer. J Emerg Trends Comput Information Sci 2011;2:656–64.

[23] Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems; 2012. (1).

[24] Moore LJ, McKinley BA, Turner KL, *et al*. The epidemiology of sepsis in general surgery patients. J Trauma 2011;70:672–80.

[25] Hansun S, Argha A, Liaw ST, *et al*. Machine and deep learning for tuberculosis detection on chest X-rays: systematic literature review. J Med Internet Res 2023;25:e43154.

[26] Chen J, Wu L, Liu K, *et al*. EDST: a decision stump based ensemble algorithm for synergistic drug combination prediction. BMC Bioinformatics 2023;24:325.

[27] NSQIP A. User guide for the 2019 participant use data. American College of Surgeons; 2020. Accessed 27 January 2021. https://www.facs.org/quality-programs/acs-nsqip/program-specifics/participant-use

[28] Hernandez MC, Thorn MJ, Kong VY, *et al*. Validation of the AAST EGS grading system for perforated peptic ulcer disease. Surgery 2018;164:738–45.

[29] Hernandez MC, Birindelli A, Bruce JL, *et al*. Application of the AAST EGS grade for adhesive small bowel obstruction to a multi-national patient population. World J Surg 2018;42:3581–8.

[30] Savage SA, Klekar CS, Priest EL, *et al*. Validating a new grading scale for emergency general surgery diseases. J Surg Res 2015;196:264–9.