Experimental Research

INTERNATIONAL JOURNAL OF SURGERY

OPEN

90-day mortality prediction in elective visceral surgery using machine learning: a retrospective multicenter development, validation, and comparison study

Christoph Riepe[a,*], Robin van de Water, MSc[b], Axel Winter, MD[a], Bjarne Pfitzner, MSc[b], Lara Faraj[c], Robert Ahlborn, MSc[d], Maximilian Schulze, BSc[a], Daniela Zuluaga, MD[a], Christian Schineis, MD[e], Katharina Beyer, MD[e], Johann Pratschke, MD[a], Bert Arnrich, MSc[b], Igor M. Sauer, MD[a], Max M. Maurer, MD[a,f]

Background: Machine Learning (ML) is increasingly being adopted in biomedical research, however, its potential for outcome prediction in visceral surgery remains uncertain. This study compares the potential of ML methods for preoperative 90-day mortality (90DM) prediction of an aggregated multi-organ approach to conventional scoring systems and individual organ models.
Methods: This retrospective cohort study enrolled patients undergoing major elective visceral surgery between 2014 and 2022 across two tertiary centers. Multiple ML models for preoperative 90DM prediction were trained, externally validated and benchmarked against the American Society of Anesthesiologists (ASA) score and revised Charlson Comorbidity Index (rCCI). Areas under the receiver operating characteristic (AUROC) and precision recall curves (AUPRC) including standard deviations were calculated. Additionally, individual models for esophageal, gastric, intestinal, liver, and pancreatic surgery were developed and compared to an aggregated approach.
Results: 7711 cases encompassing 78 features were included. Overall 90DM was 4% (n = 309). An XBoost classifier demonstrated the best performance and high robustness following external validation (AUROC: 0.86 [0.01]; AUPRC: 0.2 [0.04]). All models outperformed the ASA score (AUROC: 0.72; AUPRC: 0.08) and rCCI (AUROC: 0.81; AUPRC: 0.11). rCCI, patient age and C-reactive protein emerged as most decisive model weights. Models for gastric (AUROC: 0.88 [0.13]; AUPRC: 0.24 [0.26]) and intestinal surgery (AUROC: 0.87 [0.05]; AUPRC: 0.17 [0.09]) revealed the highest organ-specific performances, while pancreatic surgery yielded the lowest results (AUROC: 0.66 [0.08]; AUPRC: 0.22 [0.12]). A combined multi-organ approach (AUROC: 0.84 [0.04]; AUPRC: 0.21 [0.06]) demonstrated superiority over the weighted average across all organ-specific models (AUROC: 0.82 [0.07]; AUPRC: 0.2 [0.13]).
Conclusion: ML offers robust preoperative risk stratification for 90DM in elective visceral surgery. Leveraging training across multi-organ cohorts may improve accuracy and robustness compared to organ-specific models. Prospective studies are needed to confirm the potential of ML in surgical outcome prediction.

Keywords: 90-day mortality, artificial intelligence, explainable AI, machine learning, preoperative risk stratification, visceral surgery

Introduction

Considering an estimated lifetime risk of 44% in the Western world, almost one in two people will experience abdominal

HIGHLIGHTS

- Machine Learning (ML) effectively predicts 90DM in preoperative patients undergoing elective visceral surgery, demonstrating high robustness and generalizability following external validation (AUROC: 0.86; AUPRC: 0.2).
- The revised Charlson Comorbidity Index (rCCI), patient age and C-reactive protein are the most influential model weights.
- ML models can outperform conventional risk scores, including the American Society of Anesthesiologists (ASA) score (AUROC: 0.72; AUPRC: 0.08) and the rCCI (AUROC: 0.81; AUPRC: 0.11).
- ML prediction performance benefits from an aggregated multi-organ approach compared to organ-specific models. This may enable limited sample sizes to be overcome in the future.

[a]Charité- Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Surgery, Berlin, Germany, [b]Hasso Plattner Institute (HPI), Universität Potsdam, Digital Health Cluster, Potsdam, Germany, [c]Einstein Center for Neurosciences (ECN), Charité-Universitätsmedizin Berlin, Berlin, Germany, [d]Charité- Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Medical Informatics, Berlin, Germany, [e]Charité- Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of General and Abdominal Surgery, Berlin, Germany and [f]Berlin Institute of Health (BIH), Charité- Universitätsmedizin Berlin, Berlin, Germany

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

*Corresponding Author. Address: Charité- Universitätsmedizin Berlin, Department of Surgery, Augustenburger Platz 1, 13353 Berlin, Germany.
E-mail: christoph.riepe@charite.de (C. Riepe)

International Journal of Surgery (2025) 111:3742–3751
Received 22 November 2024; Accepted 20 March 2025
Published online 1 April 2025
http://dx.doi.org/10.1097/JS9.0000000000002372

surgery during their lifespan[1]. Although perioperative care has continuously improved over the past decades,[2-4] major procedures are still associated with perioperative morbidity as high as 37.8% and mortality rates reaching up to 11.7%.[5-7]

Improving risk stratification prior to surgery stands out as a pivotal strategy for reducing morbidity and mortality, as high-risk patients may opt for alternative strategies or intensified postoperative monitoring settings[8]. Various approaches are available, ranging from rather experience-based methods such as the American Society of Anesthesiologists (ASA) score to more objective tools like the Preoperative Score to Predict Postoperative Mortality (POSPOM) or the American College of Surgeons (ACS) risk calculator.[9-11]

However, a more holistic characterization of cumulative surgical risk needs to account for higher-dimensional relationships and modulating interactions rather than mere linear dependencies of individual risk parameters[12]. Machine Learning (ML) as a subfield of Artificial Intelligence (AI) provides a promising approach for surgical risk stratification since it enables multidimensional pattern recognition[13,14].

Yet, the development and application of robust ML models in surgery is particularly challenging: First, mortality prediction is usually associated with highly imbalanced data, requiring special considerations regarding model assessment metrics[15]. Additionally, ML results can be distorted by overfitting, making external validation essential. However, most studies to date followed a single-center design, leaving the potential of ML in visceral surgery uncertain.[12,16-26] Furthermore, 90-day mortality (90DM) is increasingly recognized as a more favorable benchmark for perioperative outcome assessment[27,28], whereas previous research has predominantly investigated in-house (IHM) or 30-day mortality (30DM).[12,18,20-22,25,29] Finally, model development requires a considerable amount of data for training. Most pilot studies encounter this challenge, as they typically focus on individual procedures, thereby naturally limiting the available amount of training samples[17,21,23,24,26,30]. The aggregation of multiple different procedures or even organ systems could therefore provide a novel approach to considerably amplify model training, although potentially at the expense of dataset specificity.

Given these persisting constraints, this study aims to evaluate the potential of ML for preoperative 90DM prediction in major elective visceral surgery across multiple organ systems. Various ML models are trained and results are validated externally as well as compared to conventional risk scores. Additionally, a comparison between a combined multi-organ approach against organ-specific models for esophageal, gastric, intestinal, liver, and pancreatic surgery is conducted. The results may provide novel insights regarding the aforementioned limitations of previous research and considerably improve perioperative care of visceral surgical patients.

Methods

This study was approved by the Institutional Ethics Committee on 26 July 2023 and conducted in accordance with the Declaration of Helsinki. An additional data protection consultation was performed by the supervising institution (DSO_888) and the study adheres to the TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods[31,32].

Setting

This retrospective cohort study enrolled patients of legal age undergoing major elective visceral surgery at two independent, high-volume, tertiary, university affiliated surgical centers between January 2014 and December 2022. Center 1 (C1) served as the internal training cohort while Center 2 (C2) was used to perform independent validation. Major visceral surgery encompassed procedures of the esophagus, stomach, (large and small) intestine, liver, and pancreas defined based on the German adaptation of the International Classification of Procedures in Medicine (OPS; Supplementary Table 1, available at: http://links.lww.com/JS9/E42)[33,34]. Surgery involving the esophagus, stomach, liver and pancreas was performed by surgeons who subspecialized in these respective organ systems, whereas intestinal surgery was handled by general surgeons. 90DM was selected as the primary and 30DM as the secondary study endpoint. Exclusion criteria were specified as missing information regarding an endpoint or data completeness below 75% across all features (Supplementary Figure 1, available at: http://links.lww.com/JS9/E42).

Data

A total of 94 preoperatively available features were selected for analysis, encompassing binary, numerical, and categorical data referring to patient characteristics, pre-existing conditions, intervention characteristics, and laboratory values. Pre-existing conditions and the revised Charlson Comorbidity Index (rCCI)[35,36] were obtained retrospectively as outlined by Quan *et al*[37]. Data preprocessing included the elimination of features with completeness below 50% (Supplementary Table 2, available at: http://links.lww.com/JS9/E42) and the definition of upper and lower boundaries for numerical features. Categorical parameters were ordinally or one-hot encoded, based on the presence of an intrinsic order. Prior to ML analysis, missing values were imputed using the k-Nearest Neighbors (KNN) imputer, and numerical parameters were normalized. Finally, binary features were added to indicate missing values. Data processing and analysis were performed using Python (v3.10, Python Software Foundation)[38], including various libraries (Supplementary Table 3, available at: http://links.lww.com/JS9/E42).

Patient data was compared between the two included centers. Continuous variables are presented as mean values with standard deviation (SD) and were analyzed using Student's t-test[39] or Mann-Whitney U-test[40], depending on the distribution determined using the Anderson-Darling test[41]. Binary and categorical features are shown as class frequencies and were juxtaposed using the Chi-Square test[42]. Statistical significance was defined at $P < 0.05$. The discriminatory performance of all models was assessed using the area under the receiver operating characteristic (AUROC), the area under the precision-recall curve (AUPRC), the Matthews correlation coefficient (MCC), and the F_1-Score. Where different folds or seeds led to several results, these are given as an average with standard deviation. A schematic illustration of the methodology is presented in Fig. 1.

Prediction models

Conventional scoring systems

The ASA score and rCCI were used as conventional indicators of perioperative risk. To calculate the prediction performance, all score samples of C1, including imputed values, were used to determine their respective mortality rates by class. These were

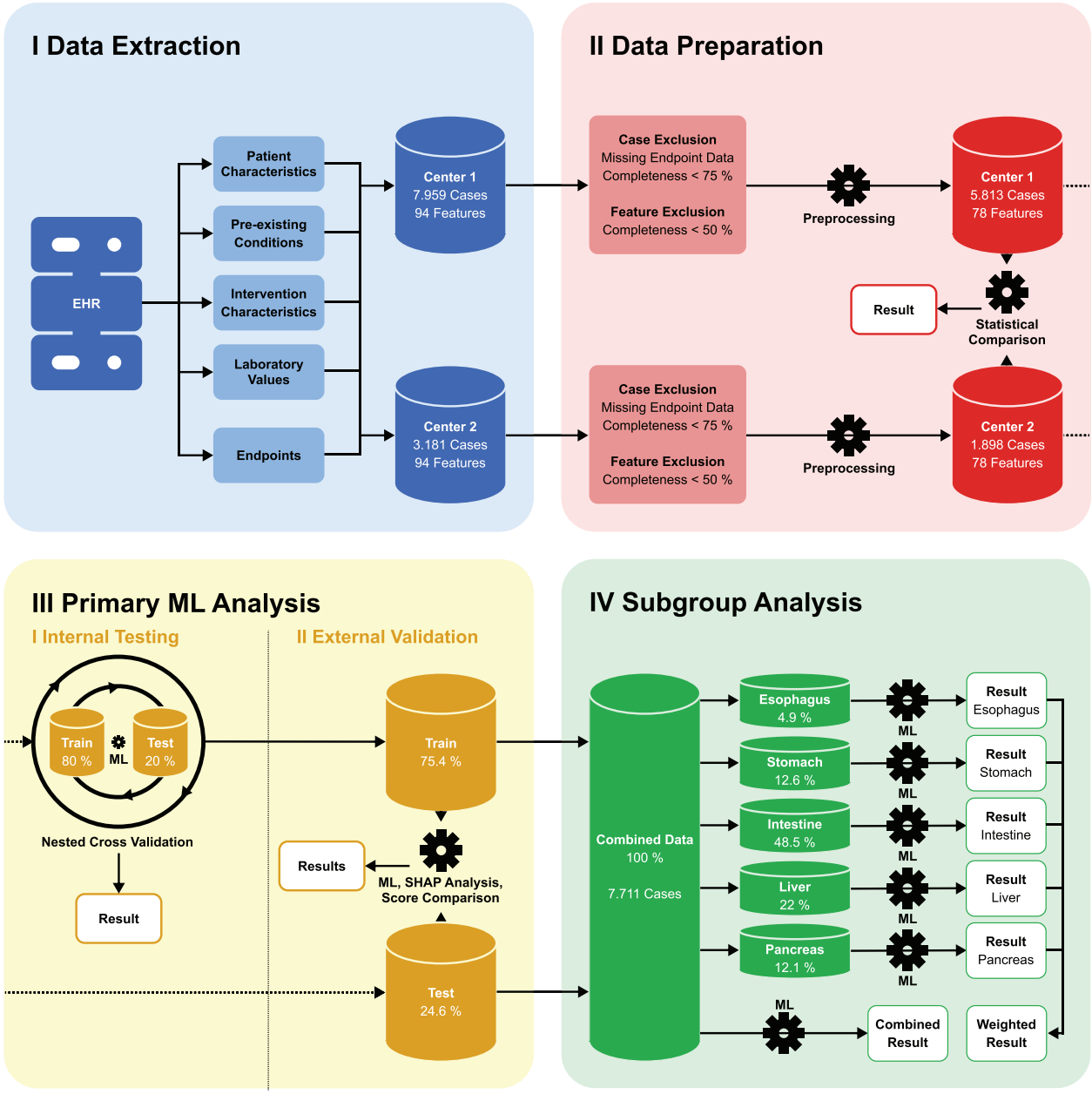

Figure 1. Schematic illustration of the study concept divided into four different phases: data extraction (I), data preparation (II), primary ML analysis (III), and subgroup analysis (IV). EHR = Electronic Health Record, ML = Machine Learning, SHAP = SHapley Additive exPlanations.

subsequently applied to the outcome data of C2 as an external validation to calculate the described evaluation metrics and assess the predictive performance in an independent cohort.

Primary ML analysis

Internal testing utilized a repeated nested cross-validation setup to split the data of C1 into a training and test subset at a ratio of 80:20[43]. This approach nests the hyperparameter optimization cross-validation (k = 5) inside a second cross-validation (k = 5) loop, thereby reducing the risk of overfitting[43]. The optimal hyperparameters, including those related to cost-sensitive learning for handling class imbalance, were subsequently selected from a predefined range based on the average performance across 30 iterations. The best-performing model on this inner training data was then evaluated on the left-out test fold. The process was repeated 5 times for each fold and the entire experiment was repeated for 5 random seeds to prevent bias from initial fold-splitting. The following classifiers were trained: eXtreme Gradient Boosting (XGBoost), Logistic Regression (LR), and Balanced Random Forest (BRF). Finally, it was tested whether the Synthetic Minority Over-sampling Technique (SMOTE)[44] could further increase performance.

External validation used the complete C1 cohort for model training across 5 random seeds to ensure the maximum amount

of available training data. Subsequently, the model was assessed across all C2 samples. The code is publicly available at https://github.com/rvandewater/CASS-PROPEL.

Additionally, feature importance analysis was performed during external validation using SHapley Additive exPlanations (SHAP) values[45]. At its core, this approach investigates the alteration in model performance by deleting each feature individually, thus allowing the calculation of the outcome contribution of features separately within the final model[45,46]. Feature weights were averaged across 10 random seeds and normalized for final reporting.

Subgroup analysis

To evaluate the potential of a combined large-scale model against individual organ-specific approaches, the data from both centers was merged. Using the XGBoost classifier, individual models for esophageal, gastric, intestinal, liver, and pancreatic surgery as well as a joint multi-organ model were trained and validated analogously to preceding internal testing. The organ-specific performance results were then weighted according to the proportion of their respective organ system to obtain a separately trained weighted average across all individual models. This ensures the same organ distribution as in the combined multi-organ approach, against which the weighted result was then compared. The weighted average of the standard deviation was derived mathematically from the calculated weighted average of the variance.

Results

A total of 11 140 cases met the inclusion criteria. Following the application of the exclusion criteria, 7711 (n_{C1} = 5813; n_{C2} = 1898) cases remained eligible and were enrolled in the analysis. The overall completeness across the 78 included features was 93.2% (C1: 93.3%; C2: 93%).

Baseline characteristics

Statistical key figures between the two centers are given in Table 1 (Supplementary Tables 4–8, available at: http://links.lww.com/JS9/E42). Most patients in both cohorts were male (C1: 54.2% [n = 3153]; C2: 56.9% [n = 1079]) without significant distribution differences (P = 0.27). Patients in the internal training cohort were significantly older (C1: 58.8 [SD = 14.8]; C2: 55.2 [SD = 17.8]; P < 0.01). Consistently in all organ systems, most cases were oncologic patients (esophagus: 96.1% [n = 366]; stomach: 49% [n = 474]; intestine: 54.6% [n = 2041]; liver: 79% [n = 1338]; pancreas: 85.5% [n = 795]), yet the overall proportion was significantly higher in the internal (71% [n = 4128]) than the external cohort (46.7% [n = 886]; P < 0.01). The predominant comorbidities in the internal training cohort were arterial hypertension (45.6% [n = 2651]), followed by metastasis of a solid tumor (36.7% [n = 2135]) and diabetes without chronic complications (21.4% [n = 1241]). The rCCI differed significantly with a median of 5 and an interquartile range (IQR) of 6 in the internal training cohort, compared to a median of 3 (IQR = 6) in the external validation cohort (P < 0.01). ASA 2 was the predominant category in both cohorts (C1: 42%; C2: 55.1%).

The distribution of organ systems diverged significantly between the two cohorts (P < 0.01). The internal training cohort showcased higher proportions for gastric (C1: 14.8% [n = 862]; C2: 5.6% [n = 106]), pancreatic (C1: 14.1% [n = 822]; C2: 5.7% [n = 108]) and particularly liver surgery (C1: 26.8% [n = 1560]; C2: 7%

Table 1

Statistical analysis of selected features and statistical comparison between the two included centers (numerical features are given as an average with a standard deviation)

#	Feature	Completeness (%)	Combined (C1 + C2)	Internal (C1)	External (C2)	*P*
1	Gender	94.71				0.27
	Male		54.88% (n = 4232)	54.24% (n = 3153)	56.85% (n = 1079)	
	Female		39.83% (n = 3071)	38.74% (n = 2252)	43.15% (n = 819)	
2	Age *(y)*	100	57.93 (SD = 15.69)	58.82 (SD = 14.83)	55.21 (SD = 17.79)	<0.01
3	Height *(m)*	85.25	172.91 (SD = 10.09)	172.91 (SD = 10.17)	172.89 (SD = 9.86)	0.39
4	Weight *(kg)*	85.44	79.85 (SD = 25.09)	81.97 (SD = 26.9)	73.52 (SD = 17.2)	<0.01
5	BMI *(kg / m^2)*	85.25	26.62 (SD = 7.71)	27.32 (SD = 8.3)	24.52 (SD = 5.04)	<0.01
6	ASA score	88.33				<0.01
	1		4.92% (n = 379)	4.27% (n = 248)	6.90% (n = 131)	
	2		45.18% (n = 3484)	41.96% (n = 2439)	55.06% (n = 1045)	
	3		36.77% (n = 2835)	40.63% (n = 2362)	24.92% (n = 473)	
	4		1.45% (n = 112)	1.60% (n = 93)	1.00% (n = 19)	
	5		0.01% (n = 1)	0.02% (n = 1)	0.00% (n = 0)	
23	Malignancy	100	65.02% (n = 5014)	71.01% (n = 4128)	46.68% (n = 886)	<0.01
46	Primary system	100				<0.01
	Esophagus		4.94% (n = 381)	4.94% (n = 287)	4.95% (n = 94)	
	Stomach		12.55% (n = 968)	14.83% (n = 862)	5.58% (n = 106)	
	Intestine		48.49% (n = 3739)	39.26% (n = 2282)	76.77% (n = 1457)	
	Liver		21.96% (n = 1693)	26.84% (n = 1560)	7.01% (n = 133)	
	Pancreas		12.06% (n = 930)	14.14% (n = 822)	5.69% (n = 108)	
69	CRP *(mg / L)*	63.05	19.19 (SD = 40.67)	17.15 (SD = 38.3)	27.73 (SD = 48.46)	<0.01
70	Hemoglobin *(g / dL)*	99.47	12.53 (SD = 2.09)	12.60 (SD = 2.05)	12.29 (SD = 2.2)	<0.01
76	Erythrocytes	99.31	4.39 (SD = 0.66)	4.40 (SD = 0.65)	4.35 (SD = 0.69)	0.03
79	90-day mortality	100	4.01% (n = 309)	4.04% (n = 235)	3.90% (n = 74)	0.83
80	30-day mortality	100	2.02% (n = 156)	1.96% (n = 114)	2.21% (n = 42)	0.56

[n = 133]). Surgery on the intestine, including colorectal surgery, made up the majority of operations in both cohorts. However, they were particularly prevalent in the external validation cohort (C1: 39.3% [n = 2282]; C2: 76.8% [n = 1457]).

Overall, 30DM reached 2% (n = 156; C1: 2%; C2: 2.2%) and almost doubled to 4% regarding 90DM (n = 309; C1: 4%; C2: 3.9%). No significant differences were found between the two cohorts (P_{90DM} = 0.83; P_{30DM} = 0.56).

Prediction results

Conventional scoring systems

The ASA score achieved an AUROC of 0.72 and AUPRC of 0.08 for 90DM as well as 0.75 and 0.05 for 30DM, respectively. In comparison, the rCCI demonstrated higher discriminatory performance in terms of both AUROC (90DM: 0.81; 30DM: 0.78) and AUPRC (90DM: 0.11; 30DM: 0.05). Due to the low mortality prevalence of less than 50% across all categories, the default MCC and F_1-Score defined at a decision boundary of 0.5 equaled 0 in all cases. To bypass this, the maximum F_1-Score was employed, resulting in 0.15 for the ASA score and 0.2 for the rCCI, both with a threshold of 0.06 for 90DM. Regarding 30DM, 0.1 with a threshold of 0.03 was achieved for the ASA score, while the rCCI reached 0.11 for a 0.02 threshold.

Primary ML analysis

During internal testing, all classifiers yielded comparable results for 90DM with an AUROC of 0.85 across all models and AUPRC results ranging from 0.23 (0.04; XGBoost, BRF) to 0.24 (0.05; LR). The MCC varied between 0.28 (0.03; BRF) and 0.31 (0.05; LR), whereas the highest F_1-Score was 0.34 (0.05, LR). Prediction models for 30DM showed higher AUROC values (0.87 [0.03]; LR), though lower AUPRC results (0.17 [0.06]; LR). The SMOTE technique did not further increase performance. Overall, LR demonstrated the best model performance during internal testing (Supplementary Table 9, available at: http://links.lww.com/JS9/E42).

Results of external validation demonstrated high reliability and robustness with only minor performance decreases compared to preceding internal testing. For 90DM, the XGBoost classifier performed best across all metrics with an AUROC of 0.86 (0.01), AUPRC of 0.2 (0.04), MCC of 0.28 (0.01) and F_1-Score of 0.3 (0.01). For 30DM, the discriminatory power was lower, reaching a maximum AUROC of 0.83 (0.01; BRF) and AUPRC of 0.11 (0.01; XGBoost). Detailed results are given in Table 2 and corresponding ROC and PR curves are shown in Fig. 2.

Feature importance analysis identified the rCCI as the highest impact parameter for 90DM prediction (SHAP = 0.17). Thereafter, patient age (SHAP = 0.14), C-reactive protein (CRP; SHAP = 0.14), hemoglobin (SHAP = 0.11) and erythrocytes (SHAP = 0.09) showed high feature weights. Pancreatic surgery was identified as the most relevant organ-related feature (SHAP = 0.06). A complete overview of the 20 most important features according to SHAP values is presented in Fig. 3.

Subgroup analysis

Organ-specific ML performance varied substantially across the investigated subdomains. Gastric surgery achieved the best result for 90DM (AUROC: 0.88 [0.13]; AUPRC: 0.24 [0.26]), although the particularly high SD must be considered. Contrary, 90DM prediction for pancreatic surgery yielded the lowest model performance (AUROC: 0.66 [0.08]; AUPRC: 0.22 [0.12]). Considering 30DM, the organ-specific model for surgery of the intestine performed best, as indicated by the highest AUROC of 0.85 (0.07) and good AUPRC of 0.18 (0.11), especially considering the low mortality rate. Conversely, surgery on the liver (AUROC: 0.72 [0.13]; AUPRC: 0.09 [0.11]) and pancreas (AUROC: 0.62 [0.13]; AUPRC: 0.17 [0.12]) demonstrated a rather weak prediction performance. Finally, the weighted average across the five organ-specific models yielded an AUROC of 0.82 (0.07) and AUPRC of 0.2 (0.13) for 90DM as well as 0.79 (0.12) and 0.17 (0.17) for 30DM, respectively.

In contrast to organ-specific training, an aggregated multi-organ approach using all samples achieved an AUROC of 0.84 (0.04) and an AUPRC of 0.21 (0.06) for 90DM as well as 0.83 (0.04) and 0.14 (0.08) for 30DM, respectively. Hence, all performance metrics for 90DM and the AUROC for 30DM of the aggregated approach are superior compared to the weighted organ-specific average. Moreover, the combined approach demonstrated a considerably lower SD compared to organ-specific training across all metrics. Detailed results of the subgroup analysis are presented in Table 3.

Table 2:

Model performance for both endpoints following external validation, shown as mean and standard deviation across all seeds, and comparison with the conventional scoring systems ASA and rCCI

Endpoint	Classifier/score	AUROC	AUPRC	MCC	F_1-score
90DM	XGBoost	0.86 (0.01)	0.20 (0.04)	0.28 (0.01)	0.30 (0.01)
	Logistic regression	0.84 (<0.01)	0.19 (<0.01)	0.27 (<0.01)	0.29 (<0.01)
	Balanced random forest	0.84 (0.01)	0.18 (0.01)	0.26 (0.01)	0.28 (0.01)
	ASA score	0.72	0.08	0.00	0.00
	rCCI	0.81	0.11	0.00	0.00
	Chance level	0.50	0.04	0.00	0.04
30DM	XGBoost	0.81 (0.01)	0.11 (0.01)	0.18 (0.02)	0.18 (0.02)
	Logistic regression	0.78 (0.02)	0.10 (<0.01)	0.16 (0.01)	0.17 (<0.01)
	Balanced random forest	0.83 (0.01)	0.09 (0.01)	0.17 (0.01)	0.17 (0.01)
	ASA score	0.75	0.05	0.00	0.00
	rCCI	0.78	0.05	0.00	0.00
	Chance level	0.50	0.02	0.00	0.02

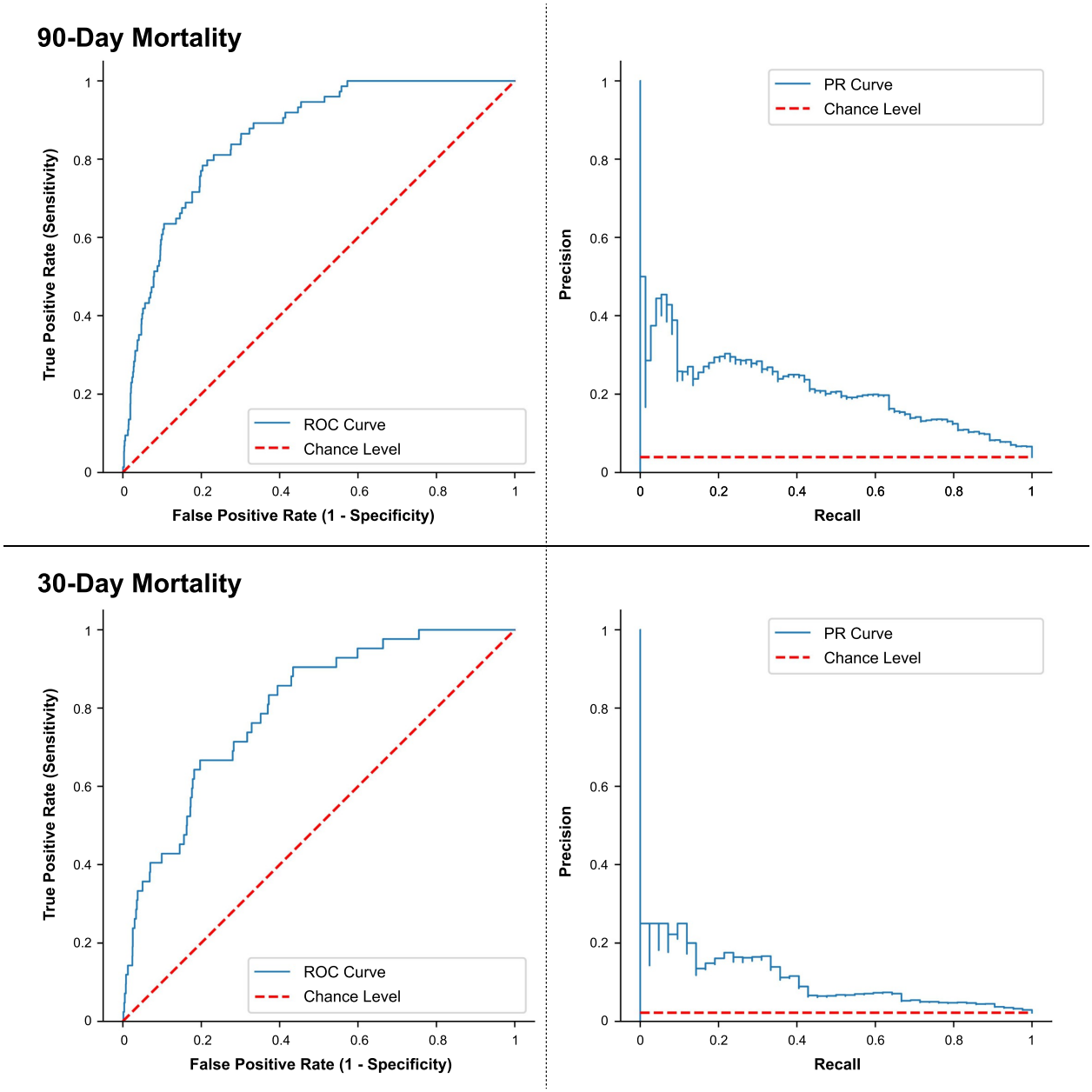

Figure 2. Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for 90-day and 30-day mortality during external validation using the eXtreme Gradient Boosting classifier, for a single fold, in comparison with the chance level defined as random guessing.

Discussion

This study aimed to evaluate, validate, and compare the potential of ML in preoperative risk stratification for major elective visceral surgery across all as well as for five individual organ systems. To date, 90DM was rarely considered for prediction tasks,[12,18,20-22,25,29] although it is increasingly recognized as a more accurate benchmark for perioperative risk assessment compared to IHM or 30DM.[27,28], Moreover, the lack of external validation still presents a major shortcoming of many surgical ML studies.[12,16-26]

ML analysis for 90DM achieved AUROC results of 0.85 for internal testing and 0.86 during external validation, indicating high robustness and generalizability of the trained models. Although it is still the most frequently reported metric for ML model performance, AUROC assessment is critically limited in discrimination tasks of highly imbalanced data[15]. Consequently, the AUPRC provides a more elaborate evaluation since it ignores the large proportion of true negative predictions[15]. Notably, the AUPRC reached 0.2 during external validation, demonstrating a discriminatory performance five times higher than baseline chance level of 0.04 (mortality rate). The XGBoost classifier achieved the best results, supporting recent findings that gradient-boosted decision trees are particularly suitable for this use case[47]. Model performance for 30DM was considerably lower across all metrics, despite its shorter

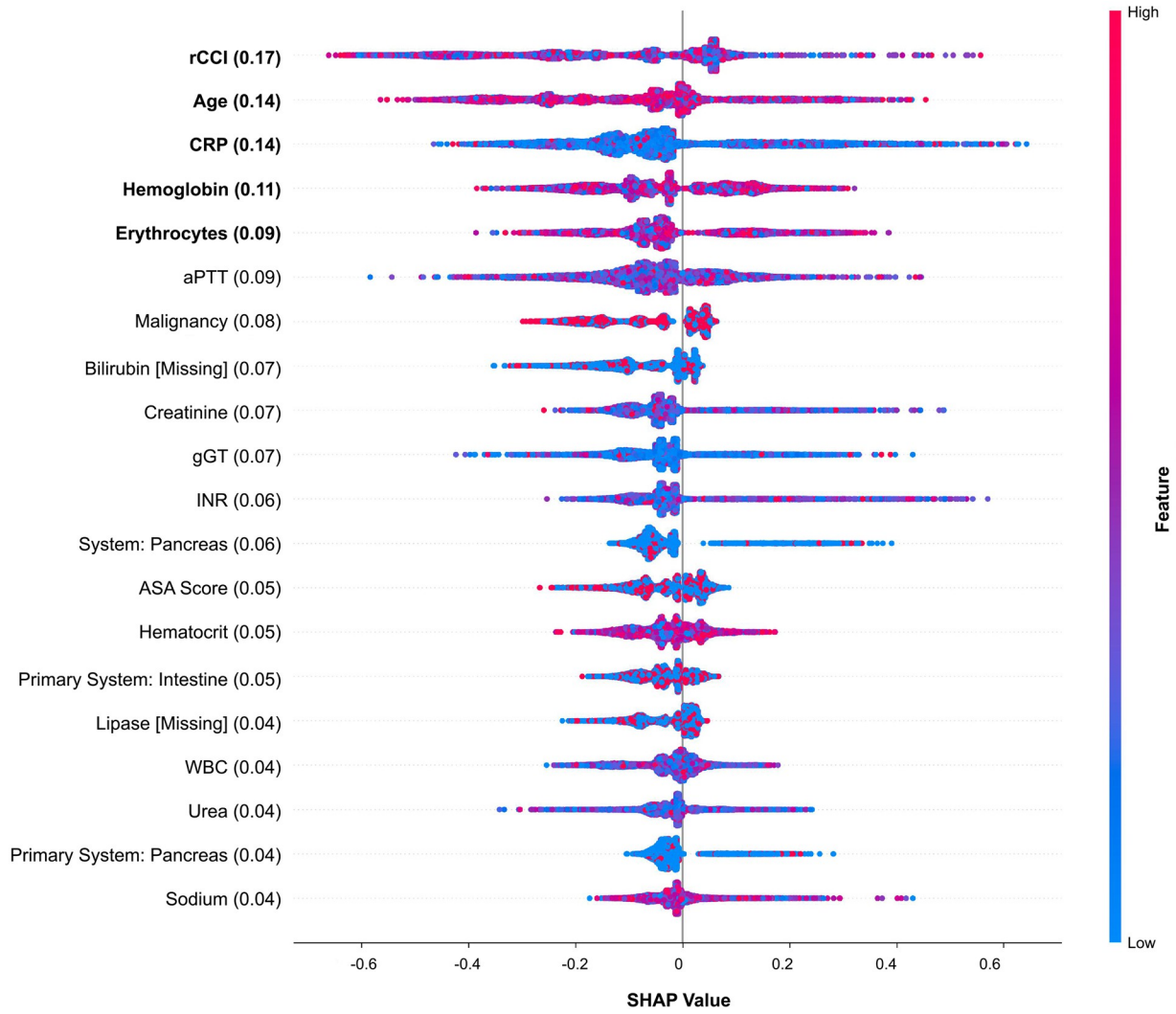

Figure 3. Results of SHapley Additive exPlanations (SHAP) analysis with the 20 most important features for 90-day mortality prediction during external validation (eXtreme Gradient Boosting classifier), shown as a Beeswarm plot with corresponding mean SHAP values. rCCI = revised Charlson Comorbidity Index, CRP = C-Reactive Protein, aPTT = activated Partial Thromboplastin Time, gGT = gamma-Glutamyl Transpeptidase, INR = International Normalized Ratio, ASA = American Society of Anesthesiologists, WBC = White Blood Count.

prediction time frame. This finding might be attributable to the naturally lower proportion of positive samples compared to 90DM, which has negative effects on model training and thus may outweigh the shorter prediction horizon. Remarkably, all ML models surpassed the ASA score and rCCI as conventional risk stratification methods for both endpoints. This observation is in accordance with previous studies suggesting that ML can indeed outperform conventional risk scores like the ASA score,[12,18-22,29] POSPOM[18,21], ACS risk calculator[12,19,22], and rCCI[16,18,21] for mortality prediction.

Large-scale ML models for outcome prediction in the domain of surgery are still scarce. An analysis including almost 500 000 patients after non-cardiac surgery yielded a higher AUROC of 0.94 for 30DM.[29] However, this study also enrolled minor surgeries, resulting in a very low mortality rate of below one percent[29], which is unfavorably in terms of AUROC assessment. Contrary, the reported AUPRC of 0.16 was lower than in this study's model[29], though the lower mortality rate must be considered for comparison. Furthermore, two studies investigating IHM with cohorts exceeding 50 000 cases also achieved AUROC results over 0.9[18,25]. However, they did not differentiate between major and minor procedures, thus yielding a considerable class imbalance. While Graeßner *et al* reported a model AUPRC of 0.11[25], this metric is not given by Hill *et al*[18], making the evaluation of results based on this unbalanced data challenging. Moreover, both studies did not consider external validation[18,25].

Interestingly, organ-specific training showed considerable performance differences across individual organ systems, yet different cohort sizes and mortality rates must be noted. The best model was achieved for gastric surgery, outperforming the results of two large coherent multicenter ML studies for 90DM after oncologic gastrectomy in terms of AUROC[30,48] and AUPRC[48]. Likewise, the individual model for 30DM prediction in patients undergoing surgery of the intestine demonstrated a high AUROC of 0.85, surpassing the performance of a large

Table 3

Results of the subgroup analysis for both endpoints analogous to internal testing using the eXtreme Gradient Boosting classifier, presented as mean and standard deviation across all folds and seeds

Endpoint	Organ system	Weight	Positive	AUROC	AUPRC	MCC	F_1-score
90DM	Esophagus	0.05 (n = 381)	0.05	0.83 (0.04)	0.20 (0.05)	0.24 (0.07)	0.27 (0.06)
	Stomach	0.13 (n = 968)	0.02	0.88 (0.13)	0.24 (0.26)	0.29 (0.30)	0.30 (0.28)
	Intestine	0.48 (n = 3739)	0.03	0.87 (0.05)	0.17 (0.09)	0.24 (0.14)	0.25 (0.12)
	Liver	0.22 (n = 1693)	0.05	0.78 (0.07)	0.24 (0.10)	0.31 (0.10)	0.30 (0.10)
	Pancreas	0.12 (n = 930)	0.09	0.66 (0.08)	0.22 (0.12)	0.25 (0.14)	0.32 (0.11)
	Weighted result	1 (n = 7711)	0.04	0.82 (0.07)	0.20 (0.13)	0.26 (0.16)	0.28 (0.14)
	Combined result	1 (n = 7711)	0.04	0.84 (0.04)	0.21 (0.06)	0.28 (0.06)	0.29 (0.05)
30DM	Esophagus	0.05 (n = 381)	0.02	0.84 (0.04)	0.16 (0.1)	0.24 (0.14)	0.22 (0.13)
	Stomach	0.13 (n = 968)	0.01	0.84 (0.21)	0.30 (0.38)	0.18 (0.33)	0.19 (0.32)
	Intestine	0.48 (n = 3739)	0.01	0.85 (0.07)	0.18 (0.11)	0.26 (0.17)	0.25 (0.14)
	Liver	0.22 (n = 1693)	0.02	0.72 (0.13)	0.09 (0.11)	0.09 (0.13)	0.10 (0.10)
	Pancreas	0.12 (n = 930)	0.05	0.62 (0.13)	0.17 (0.12)	0.23 (0.22)	0.26 (0.17)
	Weighted result	1 (n = 7711)	0.02	0.79 (0.12)	0.17 (0.17)	0.21 (0.20)	0.21 (0.17)
	Combined result	1 (n = 7711)	0.02	0.83 (0.04)	0.14 (0.08)	0.20 (0.11)	0.19 (0.09)

nationwide ML study in colorectal cancer surgery[21]. Yet, a more elaborate comparison is again challenging as no AUPRC is given[21]. In contrast, organ-specific models for patients undergoing liver and pancreatic surgery showed rather weak discriminatory power for both endpoints. A study applying a multilayer perceptron (MLP) network for surgical outcome prediction after hepatocellular carcinoma resections reported a higher AUROC of 0.84, however, comparison is difficult as the study is only limited to IHM[49].

Importantly, however, the combined multi-organ training approach conducted in this study surpassed the weighted average of organ-specific training across all metrics for 90DM and the AUROC for 30DM. Although the benefit is minor, it suggests that it may be advantageous to aggregate several heterogeneous groups into one large cohort for the sake of training sample size. More importantly, the combined approach also showed a substantially lower SD compared to the weighted average of organ-specific training. This demonstrates the increased reliability and robustness following this strategy. To optimize combined approaches more effectively toward individual organ systems in the future, the concept of transfer learning (TL) might be a particularly promising approach. Here, a source model can be derived from training based on a large though less specific cohort with subsequent organ tailored fine tuning for each subdomain. Additionally, collaborative techniques such as federated learning (FL) could be employed to establish even larger cohorts in a privacy-sensitive environment in the future.

To date, decision support tools are considered by less than a quarter of surgeons to aid their decision-making process[8,50]. Low confidence in accuracy as well as a lack of transparency regarding internal result calculation have been identified as major barriers to a wider acceptance and clinical implementation[8]. Therefore, it is crucial to provide medical professionals with more detailed model insights to increase confidence.

In this study, the rCCI was identified as the most decisive model factor, which aligns with previous clinical studies using conventional statistics[51,52]. Additionally, advanced age, being the second most important feature, is a well-recognized risk factor for postoperative mortality[53,54]. Notably, hemoglobin and erythrocytes both ranked among the top five model weights. This is particularly interesting, as preoperative anemia has been associated with increased perioperative morbidity. Implementing optimized perioperative blood management has been shown to improve patient outcomes[55]. Finally, pancreatic surgery was identified as a feature of major importance, which is consistent with the well-documented high morbidity and mortality rates associated with these procedures both in this study and previous research[1]. Comparing the identified predictors with established clinical evidence reveals a high level of concordance between the respective risk factors. This may increase confidence in these models and promote integration into clinical practice.

This study encountered several limitations. The retrospective study design may have had quantitative and qualitative effects on the data. However, automated data extraction methods were employed to minimize human error and ensure a dataset of high quality and completeness. Additionally, only patients providing follow-up data for at least 90 days were considered for analysis. As patients with uncomplicated courses are more likely not to attend follow-up appointments after discharge, the true 90DM rate may be lower than apparent in our cohorts. Moreover, corresponding ASA values were not available for all cases, resulting in missing values to be imputed, therefore no longer reflecting actual medical assessments performed by a physician. Finally, this study does not differentiate between different socioeconomic subgroups, as advocated by the TRIPOD+AI guideline. However, given the low-threshold access to healthcare in Germany due to generalized public health insurance, we assume that the local population structure is well represented in the data.

Conclusion

ML presents a promising preoperative risk stratification approach for the prediction of 90DM in patients undergoing major elective visceral surgery, surpassing conventional risk stratification approaches. Leveraging training across multiple organ cohorts may increase ML performance and especially model robustness, holding potential for advancing both further research and clinical applications. However, future research and particularly prospective clinical application studies are needed to fully assess the potential of such models.

Ethical approval

This study was approved by the Ethics Committee of Charité – University Medicine Berlin on 26 July 2023 (EA4/238/21) and conducted in accordance with the Declaration of Helsinki.

Consent

Patients have given their informed consent to the use of their medical data for retrospective research purposes as part of their treatment contract.

Sources of funding

This research received no external funding. M.M.M. was funded by the BIH Charité Digital Clinician Scientist Program. A.W., R. v.d.W., and B.P. received funding from the Federal Joint Committee (G-BA), 01VSF20015. L.F. was funded by the Einstein Center for Neurosciences.

Author's contribution

C.R., A.W., R.v.d.W., M.M.: conceptualization; C.R., A.W., R. A., M.M.: data curation; C.R., R.v.d.W.: formal analysis; C.R., R.v.d.W., A.W., M.S., M.M.: methodology; A.W., B.A., I.S., M. M.: project administration; C.R., R.v.d.W., B.P.: software; C.S., K.B., J.P. B.A., I.S., M.M.: supervision; R.v.d.W.: validation; C. R., R.v.d.W.: visualization; C.R.: writing – original draft; R.v.d. W., B.P., M.S., L.F., D.Z., I.S., M.M.: writing – review & editing. All authors have read and approved the current version of the manuscript.

Conflicts of interest disclosure

None.

Research registration unique identifying number (UIN)

This study is registered with the German Clinical Trials Register (Deutsches Register Klinischer Studien, DRKS) under the Unique Identification Number (UIN) DRKS00033658 and is publicly available at https://drks.de/search/de/trial/DRKS00033658.

Guarantor

Christoph Riepe, Max Maurer, and Igor Sauer.

Provenance and peer review

Not commissioned, externally peer-reviewed.

Data availability statement

The data set used in this study is not publicly accessible due to institutional data protection regulations but can be made available upon reasonable request subject to the consent of the supervising institution.

Code availability

The code is publicly available at https://github.com/rvandewater/CASS-PROPEL.

References

[1] Nunoo-Mensah JW, Rosen M, Chan LS, Wasserberg N, Beart RW. Prevalence of intra-abdominal surgery: what is an individual's lifetime risk? South Med J 2009; 102:25–29.
[2] Gustafsson UO, Scott MJ, Hubner M, *et al*. Guidelines for perioperative care in elective colorectal surgery: Enhanced Recovery After Surgery (ERAS(®)) society recommendations: 2018. World J Surg 2019;43: 659–95.
[3] Joliat GR, Kobayashi K, Hasegawa K, *et al*. Guidelines for perioperative care for liver surgery: Enhanced Recovery After Surgery (ERAS) society recommendations 2022. World J Surg 2023;47:11–34.
[4] Melloul E, Lassen K, Roulin D, *et al*. Guidelines for perioperative care for pancreatoduodenectomy: Enhanced Recovery After Surgery (ERAS) recommendations 2019. World J Surg 2020;44:2056–84.
[5] Baum P, Diers J, Lichthardt S, *et al*. Mortality and complications following visceral surgery: a nationwide analysis based on the diagnostic categories used in German hospital invoicing data. Dtsch Arztebl Int 2019;116:739–46
[6] Anger F, Wellner U, Klinger C, *et al*. The effect of day of the week on morbidity and mortality from colorectal and pancreatic surgery. Dtsch Arztebl Int 2020;117:521–27
[7] Tolstrup MB, Watt SK, Gögenur I. Morbidity and mortality rates after emergency abdominal surgery: an analysis of 4346 patients scheduled for emergency laparotomy or laparoscopy. Langenbecks Arch Surg 2017; 402:615–23.
[8] Pradhan N, Dyas AR, Bronsert MR, *et al*. Attitudes about use of preoperative risk assessment tools: a survey of surgeons and surgical residents in an academic health system. Patient Saf Surg 2022;16:13
[9] RD D. New classification of physical status. Anesthesiology 1963; 24:111.
[10] Le Manach Y, Collins G, Rodseth R, *et al*. Preoperative Score to Predict Postoperative Mortality (POSPOM): derivation and validation. Anesthesiology 2016;124:570–79.
[11] 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.
[12] 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.
[13] Deo RC. Machine learning in medicine. Circulation 2015;132:1920–30.
[14] Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023;388:1201–08.
[15] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10:e0118432.
[16] Ehlers AP, Roy SB, Khor S, *et al*. Improved risk prediction following surgery using machine learning algorithms. EGEMS (Wash DC) 2017;5:3.
[17] Chen D, Afzal N, Sohn S, *et al*. Postoperative bleeding risk prediction for patients undergoing colorectal surgery. Surgery 2018;164:1209–16.
[18] Hill BL, Brown R, Gabel E, *et al*. An automated machine learning-based model predicts postoperative mortality using readily-extractable preoperative electronic health record data. Br J Anaesth 2019;123:877–86.
[19] Merath K, Hyer JM, Mehta R, *et al*. Use of machine learning for prediction of patient risk of postoperative complications after liver, pancreatic, and colorectal surgery. J Gastrointest Surg 2020;24: 1843–51.
[20] Chiew CJ, Liu N, Wong TH, Sim YE, Abdullah HR. Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. Ann Surg 2020; 272:1133–39.
[21] van den Bosch T, Warps AK, de Nerée Tot Babberich Mpm, *et al*. Predictors of 30-day mortality among Dutch patients undergoing colorectal cancer surgery, 2011-2016. JAMA Netw Open 2021;4:e217737.
[22] Gao J, Merchant AM. A machine learning approach in predicting mortality following emergency general surgery. Am Surg 2021; 87:1379–85.

[23] Chen KA, Joisa CU, Stitzenberg KB, *et al*. Development and validation of machine learning models to predict readmission after colorectal surgery. J Gastrointest Surg 2022;26:2342–50.
[24] Chen KA, Joisa CU, Stem JM, Guillem JG, Gomez SM, Kapadia MR. Prediction of ureteral injury during colorectal surgery using machine learning. Am Surg 2023; 89:5702–10.
[25] Graeßner M, Jungwirth B, Frank E, *et al*. Enabling personalized perioperative risk prediction by using a machine-learning model based on preoperative data. Sci Rep 2023;13:7128.
[26] Jung JO, Pisula JI, Bozek K, *et al*. Prediction of postoperative complications after oesophagectomy using machine-learning methods. Br J Surg 2023;110:1361–66.
[27] Resio BJ, Gonsalves L, Canavan M, *et al*. Where the other half dies: analysis of mortalities occurring more than 30 days after complex cancer surgery. Ann Surg Oncol 2021;28:1278–86.
[28] Joung RH, Merkow RP. Is it time to abandon 30-day mortality as a quality measure? Ann Surg Oncol 2021; 28:1263–64.
[29] Lee SW, Lee HC, Suh J, *et al*. Multi-center validation of machine learning model for preoperative prediction of postoperative mortality. NPJ Digit Med 2022;5:91.
[30] Pera M, Gibert J, Gimeno M, *et al*. Machine learning risk prediction model of 90-day mortality after gastrectomy for cancer. Ann Surg 2022;276:776–83.
[31] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55–63.
[32] Collins GS, Moons KGM, Dhiman P, *et al*. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. Bmj 2024: 385:e078378.
[33] World Health Organization. International Classification of Procedures in Medicine. Vol 1, 1978.
[34] Operationen-Und Prozedurenschlüssel (Bundesinstitut Für Arzneimittel Und Medizinprodukte) 2022
[35] Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40:373–83.
[36] Quan H, Li B, Couris CM, *et al*. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. Am J Epidemiol 2011;173:676–82.
[37] Quan H, Sundararajan V, Halfon P, *et al*. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005;43:1130–39.
[38] van Rossum G. Python reference manual. Cwi 1995;
[39] Student. The probable error of a mean. Biometrika 1908;6:1–25.
[40] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Statist 1947;18:50–60,11.
[41] Anderson TW, Darling DA. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. Ann Math Statist 1952;23: 193–212,20.
[42] Pearson KX,On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. London Edinburgh Dublin Philos Mag J Sci 1900;50:157–75.
[43] Cawley GCT, Nicola LC. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 2010;11:2079–107.
[44] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Int Res 2002;16: 321–57.
[45] Lundberg SM, Lee S-I A unified approach to interpreting model predictions. Presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA: 2017.
[46] Shapley LS. A Value for n-Person Games. Princeton University Press. 1953 Contributions to the Theory of Games II 307–17.
[47] McElfresh DaK S, Valverde J, Vishak Prasad C, *et al*. When do neural nets outperform boosted trees on tabular data? arXiv 2023;36: 76336-69.
[48] Dal Cero M, Gibert J, Grande L, *et al*. International external validation of risk prediction model of 90-day mortality after gastrectomy for cancer using machine learning. Cancers (Basel) 2024;16:2463.
[49] Shi HY, Lee KT, Lee HH, *et al*. Comparison of artificial neural network and logistic regression models for predicting in-hospital mortality after primary liver cancer surgery. PLoS One 2012;7:e35781.
[50] Bloomstone JA, Houseman BT, Sande EV, *et al*. Documentation of individualized preoperative risk assessment: a multi-center study. Perioper Med (Lond) 2020;9:28.
[51] Bhattacharjee HK, Kaviyarasan MP, Singh KJ, *et al*. Age adjusted Charlson comorbidity index (a-CCI) AS a tool to predict 30-day post-operative outcome in general surgery patients. ANZ J Surg 2023;93:132–38.
[52] Chang CM, Yin WY, Wei CK, *et al*. Adjusted age-adjusted Charlson comorbidity index score as a risk measure of perioperative mortality before cancer surgery. PLoS One 2016;11:e0148076.
[53] Pearse RM, Harrison DA, James P, *et al*. Identification and characterisation of the high-risk surgical population in the United Kingdom. Crit Care 2006;10:R81.
[54] Jhanji S, Thomas B, Ely A, Watson D, Hinds CJ, Pearse RM. Mortality and utilisation of critical care resources amongst high-risk surgical patients in a large NHS trust. Anaesthesia 2008; 63:695–700.
[55] Luo X, Li F, Hu H, *et al*. Anemia and perioperative mortality in non-cardiac surgery patients: a secondary analysis based on a single-center retrospective study. BMC Anesthesiology 2020;20:112.