Machine learning risk prediction model of 90-day mortality after gastrectomy for cancer

Manuel Pera, MD, PhD[1*], Joan Gibert, PhD[2*], Marta Gimeno[1], Elisenda Garsot, MD, PhD[3], Emma Eizaguirre, MD, PhD[4], Mónica Miró, MD[5], Sandra Castro, MD[6], Coro Miranda, MD[7], Lorena Reka, MD[8], Saioa Leturio, MD[9], Marta González-Duaigües, MD[10], Clara Codony, MD[11], Yanina Gobbini, MD[12], Alexis Luna, MD PhD[13], Sonia Fernández, MD[14], Aingeru Sarriugarte, MD[15], Carles Olona, MD PhD[16], Joaquín Rodríguez-Santiago, MD, PhD[17], Javier Osorio, MD, PhD[5], Luis Grande, MD, PhD[1] on behalf of the Spanish EURECCA Esophagogastric Cancer Group

* These authors share first authorship criteria.

1. Section of Gastrointestinal Surgery, Hospital del Mar, Department of Surgery, Universitat Autònoma de Barcelona, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain.
2. Department of Pathology, Hospital Universitario del Mar, Cancer Research Program, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain.
3. Department of Surgery, Hospital Universitari Germans Trias i Pujol, Universitat Autònoma de Barcelona, Badalona, Barcelona, Spain.
4. Department of Surgery, Hospital Universitario de Donostia, Donostia, Spain.
5. Department of Surgery, Hospital Universitari de Bellvitge, L´Hospitalet de Llobregat, Barcelona, Spain.
6. Department of Surgery, Hospital Universitari Vall d´Hebron, Universitat Autónoma de Barcelona, Barcelona, Spain.
7. Department of Surgery, Hospital Universitario de Navarra, Pamplona, Spain.
8. Department of Surgery, Hospital Universitario de Araba, Vitoria, Spain.

9. Department of Surgery, Hospital Universitario de Basurto, Bilbao, Spain.
10. Department of Surgery, Hospital Universitari Arnau de Vilanova, Lleida, Spain.
11. Department of Surgery, Hospital Universitari Josep Trueta, Girona, Spain.
12. Department of Surgery, Hospital de Sant Joan Despí Moisès Broggi, Sant Joan Despí, Barcelona, Spain.
13. Department of Surgery, Hospital Universitari Parc Taulí de Sabadell, Sabadell, Barcelona, Spain.
14. Department of Surgery, Hospital de la Santa Creu I Sant Pau, Universitat Autònoma de Barcelona, Barcelona, Spain.
15. Department of Surgery, OSI EE-Cruces, UPV/EHU, IIS Biocruces, Bizkaia, Spain.
16. Department of Surgery, Hospital Universitari de Tarragona, Joan XXIII, Tarragona, Spain.
17. Department of Surgery, Hospital Universitari Mútua Terrassa, Terrassa, Barcelona, Spain.

Category of article: Original Research

Correspondence: Manuel Pera, Section of Gastrointestinal Surgery, Hospital Universitario del Mar, Barcelona, Spain. Email: pera@parcdesalutmar.cat

Reprints: Manuel Pera, Section of Gastrointestinal Surgery, Hospital Universitario del Mar, Barcelona, Spain. Email: pera@parcdesalutmar.cat

Funding: No funding was received for this work.

Running head: Mortality risk in gastric cancer surgery.

MINI-ABSTRACT

The 90-day mortality risk model derived using a large multi-institutional population-based data set and machine learning approaches provided excellent performance in quantifying the risk of mortality after gastrectomy with curative intent and will be useful for patients and surgeons in making informed decisions.

STRUCTURED ABSTRACT

Objective: To develop and validate a risk prediction model of 90-day mortality (90DM) using machine learning in a large multicenter cohort of patients undergoing gastric cancer resection with curative intent.

Summary Background Data: The 90DM rate after gastrectomy for cancer is a quality of care indicator in surgical oncology. There is a lack of well validated instruments for personalized prognosis of gastric cancer.

Methods: Consecutive patients with gastric adenocarcinoma who underwent potentially curative gastrectomy between 2014 and 2021 registered in the Spanish EURECCA Esophagogastric Cancer Registry database were included. The 90DM for all causes was the study outcome. Preoperative clinical characteristics were tested in four 90DM predictive models: Cross Validated Elastic regularized logistic regression method (cv-Enet), boosting linear regression (glmboost), random forest (RF), and an ensemble model. Performance was evaluated using the area under the curve (AUC) by 10-fold cross-validation.

Results: 3,182 and 260 patients from 39 institutions in six regions were included in the development and validation cohorts, respectively. The 90DM rate was 5.6% and 6.2%, respectively. The RF model showed the best discrimination capacity with a validated AUC of 0.844 (95% confidence interval [CI] 0.841-0.848) as compared with cv-Enet (0.796, 95% CI 0.784-0.808) glmboost (0.797, 95% CI 0.785-0.809), and ensemble model (0.847, 95% CI 0.836-0.858) in the development cohort. Similar discriminative capacity was observed in the validation cohort.

Conclusion: A robust clinical model for predicting the risk of 90DM after surgery of gastric cancer was developed. Its use may aid patients and surgeons in making informed decisions.

Keywords: Gastric cancer; Gastrectomy; Mortality; Prediction; Machine learning.

INTRODUCTION

Radical gastrectomy remains the primary potentially curative treatment for patients with resectable gastric cancer (1). However, despite improvements in surgical technique and perioperative care, this surgical procedure is still associated with high morbidity (20-45%) (2-5) and mortality (2%-7%) rates (2, 4, 5). Postoperative mortality rates have been consistently reported to be higher when assessed at 90 days as compared to 30 days (2, 3, 6). Two recent studies from the Western world based on the National Cancer Database in the United States (7) and the GASTRODATA registry (3) in Europe showed that gastrectomy-related mortality increased from 3.5% and 3.6% at 30 days to 6.7% and 4.5% at 90 days, respectively. In fact, 90-day mortality has emerged as a new quality of care measure that better reflects short-term outcome of patients undergoing complex digestive surgery (8), so that mortality prediction analysis should be moved toward this new benchmark.

Risk prediction is a critical factor to determine eligibility for surgery during multidisciplinary tumor board discussions and can help inform patients and surgeons in shared decision-making processes guiding treatment. Several risk prediction models for survival or adverse outcomes after curative gastrectomy have been developed (9,10), but 90-day mortality has been considered in only two of them (6, 11). On the other hand, almost all prediction models are based on logistic or Cox regression analysis (10), validation is generally limited, and the overall performance has been considered as fair (10). Machine learning approaches, a branch of artificial intelligence, have evolved and grown popularity, because they have greater flexibility to capture complex non-linear relationships especially when a very large number and loose data are used (12). New technologies, however, should follow the methodological standards of data treatment, avoiding common pitfalls such as insufficient sample size, risk of bias, unclear

descriptions of treatment and patients' characteristics, handling of missing data or lack of use of calibration models (13,14).

The Spanish EURECCA Esophagogastric Cancer Registry (SEEGCR) is a large and audited multi-institutional population-based registry launched in 2013 that provides real-world data on outcomes and quality measures of esophagogastric cancer surgery (15). The present study used information from this database and machine learning methods to develop a clinically useful risk prediction model of 90-day mortality after gastric cancer resection with curative intent, followed by internal-external validation of the model.

METHODS

This study conformed to the TRIPOID10 (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) reporting guidelines (16) (appendix 1).

Source of data

For this multi-institutional population-based cohort study, data were retrieved from the SEEGCR linked to the EURECCA Upper gastrointestinal (GI) network. The registry collects prospective clinical data from all patients with primary esophageal, gastro-esophageal junction (GEJ), and gastric cancer undergoing resection with curative intent in 39 public hospitals from six regions in Spain covering nearly a population of 14 million people. Ninety-six variables are collected from each patient by the principal investigator at each institution. The online data cover information about the whole care process and the short- and long-term follow-up. Detailed definitions of comorbidities, postoperative complications (17, 18), histopathology findings, and quality outcome measures are used since January 2014 and have been previously described (15). The SEEGCR database was recently audited (period 2014-2017) with completeness estimated at 97% and data accuracy at 95% (15).

Ethics

The SEEGCR project was approved by the Parc de Salut Mar Clinical Research Ethics Committee (code 2013/5047/I) in 2013 for processing health care information, and subsequently by the local ethics committee of each participating hospital. All patients received a specific patient information sheet and signed the informed consent. The present study was a retrospective analysis of anonymized data recorded in the SEEGCR

database and the institutional review board waived the requirement to obtain informed consent.

Eligibility and primary outcome

All patients with primary gastric or GEJ cancer who underwent partial or total gastrectomy with curative intent from January 2014 to October 2021 were eligible. The primary outcome was 90-day mortality defined as all-cause mortality within 90 days after surgery. Data were collected from January 2014 to December 2020 for the construction and internal validation of the model and from January to September 2021 for internal-external validation.

Predictor characteristics

All preoperative variables available in the registry have been used in the prediction model and any variable selection method was used during the training-validation process (Table 1). Age, body mass index (BMI), hemoglobin and albumin serum levels, and hospital volume activity (number of gastrectomies per center per year) were considered as continuous variables. The remaining variables were categorized as dichotomous variables by using one-hot encoding (19). Missing data were imputed by including a separate category of predictor variables that had missing values (20). Descriptive statistics were presented as mean and standard deviation or number and percentages for continuous and categorical variables, respectively. Differences between groups surviving or dying within 90 days were assessed using the Fisher exact test for categorical variables or the Kolmogorov-Smirnov test for continuous variables. Statistical significance was set at $p<0.05$.

Model building and validation

Cross Validated Elastic net regularized logistic regression (cv-Enet), random forest (RF), and glmboost were used to predict 90-day mortality. cv-Enet find the optimal coefficients for lasso and ridge penalties by using internal cross validation (21). RF is a combination of decision trees built on a subsampling of the dataset (22). Glmboost is a generalized linear model fitted against a boosting algorithm (23). As other boosting machines, it can be seen as a combination of models to generate a better one while adding some type of regularization to minimize overfitting (24).

To address the class imbalance problem, which could lead to a severely imbalanced degree of performance, synthetic minority technique (SMOTE) (25) was applied for model training. Hyperparameter tuning was performed using nested resampling, which consists in using a k-fold cross-validation procedure for model hyperparameter optimization nested inside the k-fold cross-validation procedure for model selection (26). For 90DM prediction, we used 1000-evaluation random search 10-fold cross-validation with 5 repeats for each model, respectively. The best models of the three strategies were then combined to generate an ensemble model by generating a linear blend of predicted probabilities using logistic regression.

Discrimination of the models was assessed using the area under the receiver operator characteristic (ROC) curve (AUC). Internal validation was performed using 10-fold cross-validation. Calibration was assessed visually and formally with the Hosmer-Lemeshow test. Finally, isotonic regression was used to scale probabilities of the models as previously described (27).

The mean contribution of each variable was calculated as a percentage contribution to the AUC relative to the full model using the "model parts" function from the DALEX package (28).

Data analysis was performed using R software version 3.5.3 (R Foundation for Statistical Computing, Vienna, Austria). Models were trained using mlr3 package (29).

The calibrated final model is available freely at **https://gastrohmar.shinyapps.io/rf_eurecca_model/**

RESULTS

Study development cohort

A total of 3,197 patients were included in the SEEGCR database over a 7-year period. After excluding 15 patients (palliative surgery in 12 and duplicates in 3), the study cohort included 3,182 patients. Table 1 shows comparative data on comorbidity, performance status, nutrition, clinical tumor staging, and surgical approach between patients who died within 90-days postoperatively and those who survived. The overall rate of missing data for variables was 0.6% (677 items in 101.824 cells). The most frequently missing characteristics were preoperative albumin (n=441 [13.9%]) and BMI (n=88 [2.7%]).

In relation to mortality, 110 patients (3.4%) died within 30 days and 179 (5.6%) within 90 days. A total of 133 patients were classified as Clavien-Dindo V (4.1%). Patients in the 90-day mortality group were significantly more likely to be male and older, had worse ECOG performance status and ASA grade, and lower levels of hemoglobin and albumin. Other more common findings were myocardial infarction, congestive heart failure, chronic pulmonary disease, peripheral vascular disease, cerebrovascular disease, dementia, diabetes mellitus (end-organ damage), malignant lymphoma, moderate-to-severe renal disease, neoadjuvant therapy, and being operated on a non-elective basis and using open surgery.

Model performance: discrimination

From the single models, RF showed the best discrimination capacity with a validated AUC of 0.844 (95% confidence interval [CI] 0.841-0.848) as compared with cv-Enet (AUC of 0.796, 95% CI 0.784-0.808) and glmboost (AUC of 0.797, 95% CI 0.785-

0.809). The ensemble model did not show a significantly better AUC (0.847, 95% CI 0.836-0.858) as compared with the RF model alone (Figure 1).

Model performance: inter-external validation

Comparative data between the development and the internal-external validation (n=260) cohorts are shown in Supplementary Table 1. The AUC for RF, cv-Enet, glmboost, and ensemble model was 0.829 (95% CI 0.743-0.916), 0.784 (95% CI 0.677-0.891), 0.789 (95% CI 0.680-0.897), and 0.829 (95% CI 0.741-0.916), respectively. Similar discriminative capacity was observed in both the development and the validation cohorts.

Model performance: calibration

Calibration of the models was assessed visually by plotting the fitted values against the actual average values. The RF and the ensemble models presented the most unfavorable distributions compared to the other models (Figure 2, Supplementary Figure 1), although all of them showed a poor calibration on the training set when tested using the Hosmer Lemeshow test ($p<2.2e-16$ for RF, $p=2.5e-8$ for cv-Enet, $p=2.9e-11$ for glmboost and $p<2.2e-16$ for ensemble). Probabilities generated for all the models were scaled using isotonic regression. This correction improves the distribution of the values for all the models used (Figure 2, Supplementary Figure 1).

Variable importance

Feature importance analysis was assessed as a percentage contribution for each variable to each model. For RF, the most important factors were age (1.45% AUC), hospital volume activity (1.11% AUC), preoperative albumin level (1.10% AUC), preoperative hemoglobin level (0.09% AUC), and total/subtotal gastrectomy (0.85% AUC) (Figure 3). Although age was also the most important factor for the other models (5.51% for cv-

Enet, 6.25% for glmboost and 1.15% for ensemble), RF and ensemble models showed a broader contribution of factors (Supplementary Figure 2).

DISCUSSION

Using a large population-based, multicenter, and audited clinical database and machine learning techniques, a prediction model for 90-day mortality in patients undergoing gastric cancer resection with curative intent showed an excellent performance (AUC 0.844). The model uses routine and easily available preoperative clinical data and provides relevant information to inform the patient and to obtain the informed consent.

Postoperative mortality at 90 days has become a new quality measure in oncological surgery. Several studies have shown that 30-day mortality significantly underestimates the rate of mortality after complex oncological surgery, including esophagogastric cancer surgery (2, 3). For this reason, the aim of this work was to develop a prediction model of 90-day mortality in a large series of patients undergoing gastric cancer resection. Our series incorporated 3,182 patients, 110 (3.4%) of which died within 30 days and 179 (5.6%) within 90 days. These 30- and 90-day mortality rates are similar to data recently in other Western observational studies using national data sets (2, 3, 7).

Of 47 clinical prediction models in esophageal and gastric cancer surgery analyzed in a recent systematic review (10), only one based on a large cohort of patients who underwent distal gastrectomy in Japan was developed to predict “operative mortality” (11). Operative mortality was defined as death during the index hospitalization, regardless of the length of the hospital stay (< 90 days), as well as death after hospital discharge and 30 days or less from the surgery date. Besides this systematic review, another study from Japan using the same national registry dataset described a prediction model of "operative mortality" in a large cohort of patients undergoing total gastrectomy (6). Although the results obtained in both studies can be considered good from the predictive point of view, with AUC of 0.798 and 0.824, respectively, both models were developed using regression-based methods and in one of them the discriminatory

accuracy of the prediction model was qualified as a fair level of evidence in a recent systematic review (10). In this setting we explored different machine learning approaches (random forest, cv-Enet or glmboost) in order to improve the discrimination capability. The best result was obtained with the RF technique, with an AUC of 0.844 that compares favorably with the two previous prediction models (6, 11). An ensemble model using other machine learning approaches did not improve significantly the discriminative capacity obtained with the RF method.

In our risk prediction model, the six more important variables contributing to 90-day mortality were age, hospital volume activity, preoperative albumin and hemoglobin levels, type of gastrectomy, and history of chronic obstructive pulmonary disease. In all models, age was consistently identified as the most important factor associated with 90-day mortality. Previous observational studies have shown that chronological age should be considered as an individual risk factor, with more pronounced incremental risk after ages of 75 and 80 years (30, 31). However, the magnitude of increased risk might be further modulated by associated comorbidities. Hospital volume activity arises as the second contributing variable to the model. Previous population-based studies have shown the influence of hospital volume on mortality after gastrectomy for cancer (4). Other contributing factors included modifiable factors, such as hemoglobin and albumin levels which have also been identified in previous risk models of 30- and 90-day mortality after complex gastrointestinal procedures, including gastric cancer resection (6, 11).

The major strength of this study lies in the use of a large prospective, multicenter, and audited population-based database in which 39 hospitals from six Spanish regions participated. This guarantees the quality of the data, but also represents the real-life practice in our country. In this scenario, the number of missing data does not reach 1%

in a database with more than 100,000 cells. Moreover, the registry includes a large number of centers with substantial differences in infrastructure and technology that represents the real-life setting. Another strength of the study is having followed the recommendations of the TRIPOD guides regarding reporting.

Some limitations should be mentioned. Although machine learning techniques allow the construction of predictive models at the same time as internal validation, one limitation of the study is the lack of external validation. However, an additional internal-external validation was performed showing the same levels of reliability, with an AUC of 0.829 (95% CI 0.743-0.916). A class-imbalance data with a low number of observed events (179 of 3,182 patients) may be another limitation of the study (32), but the SMOTE oversampling technique was used to balance medical data (25, 33).

In conclusion, a robust model for prediction of 90-day mortality after gastrectomy for cancer with curative intent has been obtained using machine learning techniques. This tool can provide relevant information to inform the patient and to assist in the decision-making process in clinical practice.

Author contributions

***Writing group*:** Manuel Pera, Luis Grande, Joan Gibert, and Marta Gimeno conceived the idea, designed the study, collected and analyzed the data, drafted the manuscript and revised it critically.

All authors were involved in acquisition of data and gave final approval of the version to be published.

Acknowledgements: The authors are grateful to collaborators of the Spanish EURECCA Esophagogastric Cancer Group who also collected patient data and contributed to the care of the study patients: Dulce Momblan (Hospital Clinic, Barcelona); Aurora Aldeano (Hospital General de Granollers, Granollers, Barcelona); Judit Hermoso (Hospital Universitari de Vic, Vic, Barcelona); Juan J. Sánchez-Cano (Hospital Universitari de Sant Joan, Reus, Tarragona), Laura Pulido (Hospital de Mataró, Consorci Sanitari del Maresme, Mataró, Barcelona); Rafael Gil-Albarellos (Hospital de San Pedro, Logroño); Mercè Güell (Altaia Xarxa Assistencial i Universitaria de Manresa, Manresa); Jaume Tur (Hospital Universitari de Igualada, Igualada, Barcelona). The authors would like to thank Marta Pulido, MD, for editorial assistance.

Declaration of interests: The authors declare there are no conflicts of interest.

Data availability: All code and scripts to reproduce the experiments of this paper are available at: https://github.com/Tato14/rf_gastro

REFERENCES

1. De Manzoni G, Marrelli D, Baiocchi GL, et al. The Italian Research Group for Gastric Cancer (GIRG) guidelines for gastric cancer staging and treatment: 2015. *Gastric Cancer*. 2017;20:20-30.
2. Challine A, Voron T, Dousset B, et al. Postoperative outcomes after laparoscopic or open gastrectomy: A national cohort study of 10,343 patients. *Eur J Surg Oncol*. 2021;47:1985-1995.
3. Baiocchi GL, Giacopuzzi S, Reim D, et al. Incidence and grading of complications after gastrectomy for cancer using the GASTRODATA Registry. *Ann Surg*. 2020;272:807-813.
4. Voeten DM, Busweiler LAD, van der Werf LR, et al. Outcomes of esophagogastric cancer surgery during eight years of surgical auditing by the Dutch Upper Gastrointestinal Cancer Audit (DUCA). *Ann Surg*. 2021;274:866-873.
5. Papenfuss WA, Kukar M, Oxenberg J, et al. Morbidity and mortality associated with gastrectomy for gastric cancer. *Ann Surg Oncol*. 2014;21:3008-3014.
6. Watanabe N, Miyata H, Gotoh M, et al. Total gastrectomy risk model: data from 20,011 Japanese patients in a nationwide internet-based database. *Ann Surg*. 2014;260:1034-1039.
7. Sallehi O, Vega EA, Kutlu OC, et al. Western population-based study of oncologic surgical quality and outcomes of laparoscopic versus open gastrectomy for gastric adenocarcinoma. *Surg Endosc*. 2021;35:4786-4793.
8. D´Journo XB, Boulate D, Fourdrain A, et al. Risk prediction model of 90-day mortality after esophagectomy for cancer. *JAMA Surg*. 2021;156:836-845.
9. Fischer C, Lingsma H, Hardwick R, et al. Risk adjustment models for short-term outcomes after surgical resection for oesophagogastric cancer. *Br J Surg*.

2016;103:105-106.

10. Van den Boorn HG, Engelhardt EG, van Kleef J, et al. Prediction models for patients with esophageal or gastric cancer: A systematic review and meta-analysis. *PLoS One*. 2018;13:e0192310. doi: 10.1371/journal.pone.0192310.
11. Kurita N, Miyata H, Gotoh M, et al. Risk model for distal gastrectomy when treating gastric cancer on the basis of data from 33,917 Japanese patients collected using a nationwide web-based data entry system. *Ann Surg*. 2015;262:295-230.
12. Collins GS, Dhiman P, Andaur Navarro C, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11:e048008. doi: 10.1136/bmjopen-2020-048008.
13. Dhiman P, Ma J, Andaur-Navarro C, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J Clin Epidemiol.* 2021;138:60-72.
14. Andaur-Navarro CL, Damen JA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: a systematic review. *BMJ*. 2021; 375:n281. doi: 10.1136/bmj.n2281.
15. Dal Cero M, Rodríguez-Santiago J, Miro M, et al. Evaluation of data quality in the Spanish EURECCA Esophagogastric Cancer Registry. *Eur J Surg Oncol*. 2021;47:3081-3087.
16. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med.* 2015;162:55-63.
17. Baiocchi GL, Giacopuzzi S, Marrelli D, et al. International consensus on a complications list after gastrectomy for cancer. *Gastric Cancer*. 2019;22:172-189.

18. Dindo D, Demartines N, Clavien PA. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Ann Surg*. 2004;240:205–213.
19. Hancock JT, Khoshgoftaar TM. Survey on categorical data for neural networks. *J Big Data*. 2020;7:28. 10.1186/s40537-020-00305-w-
20. Ding Y, Simonoff JS. An investigation of missing data methods for classification trees applied to binary response data. *J Mach Learn Res*. 2010;11:131-170.
21. Friedman J, Hastie T, Tibshirani. Regularization paths for generalized linear models vi coordinate descent. *J Stat Softw*. 2010;33:1-2.
22. Breiman L. Random Forest. *Mach Learning*. 2001;45:5-32.
23. Bühlmann P, Yu B. Boosting with the L_2 loss: Regression and classification. *J Am Statist Assoc*. 2003;98:324-339.
24. Schapire RE. The strength of weak learnability. *Mach Learning*. 1990;5:197-227.
25. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. *J Art Intell Res*. 2002:6:321-357.
26. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*. 2010;11:2079-2107.
27. Chen W, Sahiner B, Samuelson F, et al. Calibration of medical diagnostic classifier scores to the probability of disease. *Stat Methods Med Res*. 2018;27:1394-1409.
28. Biecek P. DALEX: Explainers for complex predictive models in R. *J Mach Learn Res*. 2018;19:1-5.
29. Lang M, Binder M, Richter J, et al. mlr3: A modern object-oriented machine learning framework in R. *J Open Source Software*. 2019;4:1903.
30. Shannon AB, Straker RJ, Fraker DL, et al. Ninety-day mortality after total gastrectomy for gastric cancer. *Surgery*. 2021;170:603-609.
31. Hamilton TD, Mahar A, Haas B, et al. The impact of advanced age on short-term

outcomes following gastric cancer resection: an ACS-NSQIP analysis. *Gastric Cancer*. 2018;21:710-719.

32. Iswaran H, O´Brien R. Commentary: The problem of class imbalance in biomedical data. *J Thorac Cardiovasc Surg*. 2021;161:1940-1941.
33. Khushi M, Shaukat K, Alam TM, et al. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*. doi 10.1109/ACCESS.2021.3102399.

FIGURE LEGEND

Figure 1. Model discrimination. Receiver operating characteristic (ROC) curves for A. Random Forest (area under the curve [AUC] 0.844, 95% CI 0.841-0.848), B. cv-Enet (AUC 0.796, 95% CI 0.784-0.808), C. glmboost (AUC 0.797, 95% CI 0.785-0.809) and D. ensemble (AUC 0.847, 95% CI 0.836-0.858). The shaded area represents the 95% confidence interval.

Figure 2. RF model before and after isotonic regression. A. Unscaled calibration (intercept - 0.866, slope 3.183) and B. scaled calibration (intercept 0.013, slope 1.001). The shaded areas represent two standard errors (SE).

Figure 3. Variable importance of RF model. Mean AUC contribution of all factors on the RF model after 1e4 permutations.

Supplementary Figure 1. **Models before and after isotonic regression.** Unscaled calibration (left) and scaled calibration (right) for cv-Enet, glmboost and ensemble. The shaded areas represent two standard errors.

Supplementary Figure 2. **Variable importance of models.** Mean AUC contribution of all factors on the cv-Enet, glmboost and ensemble models after 1e4 permutations.

Figure 1

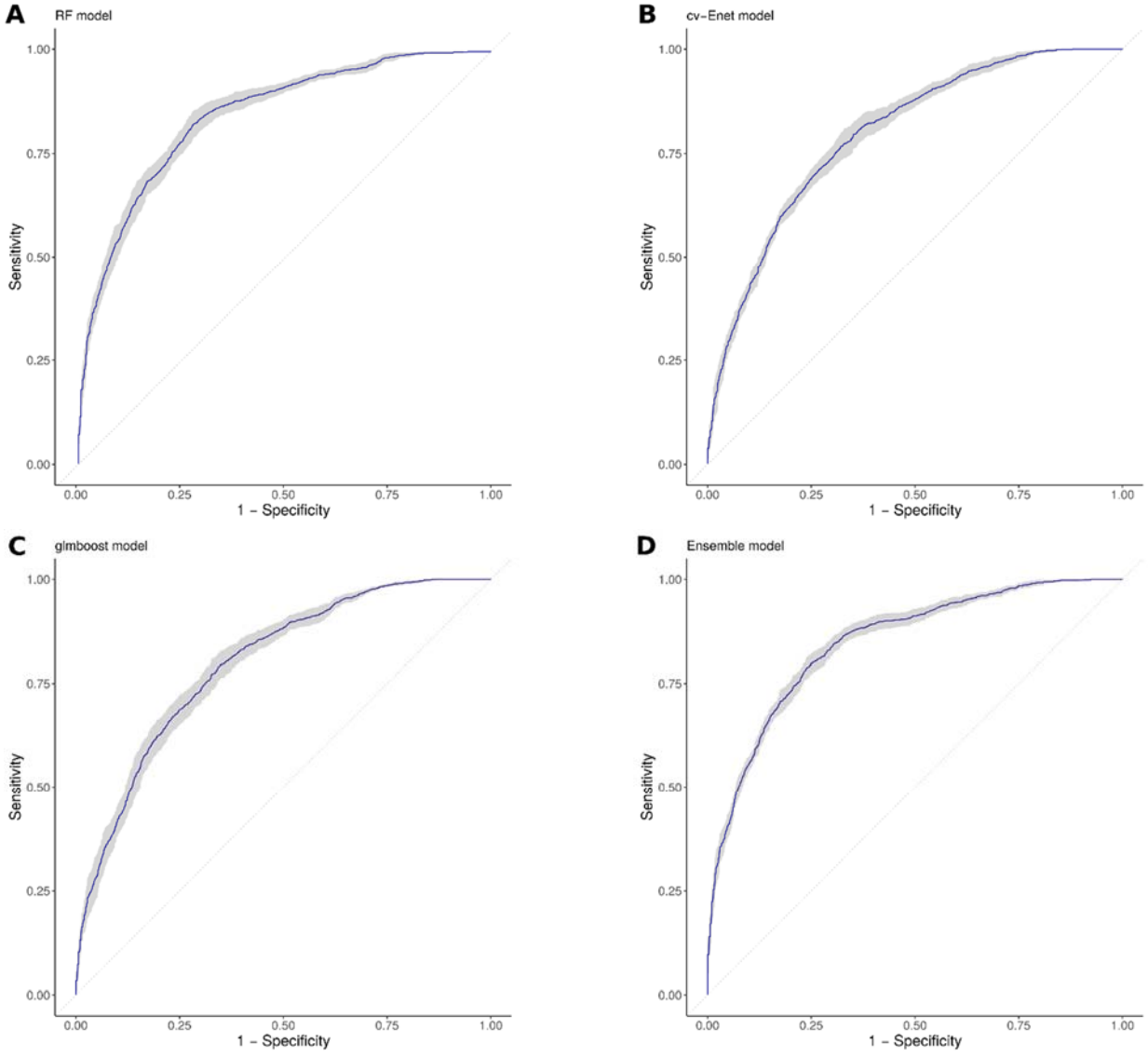

A
RF model
B
cv−Enet model
C
gimboost model
D
Ensemble model
Sensitivity
1 − Specificity
0.00
0.25
0.50
0.75
1.00

Figure 2

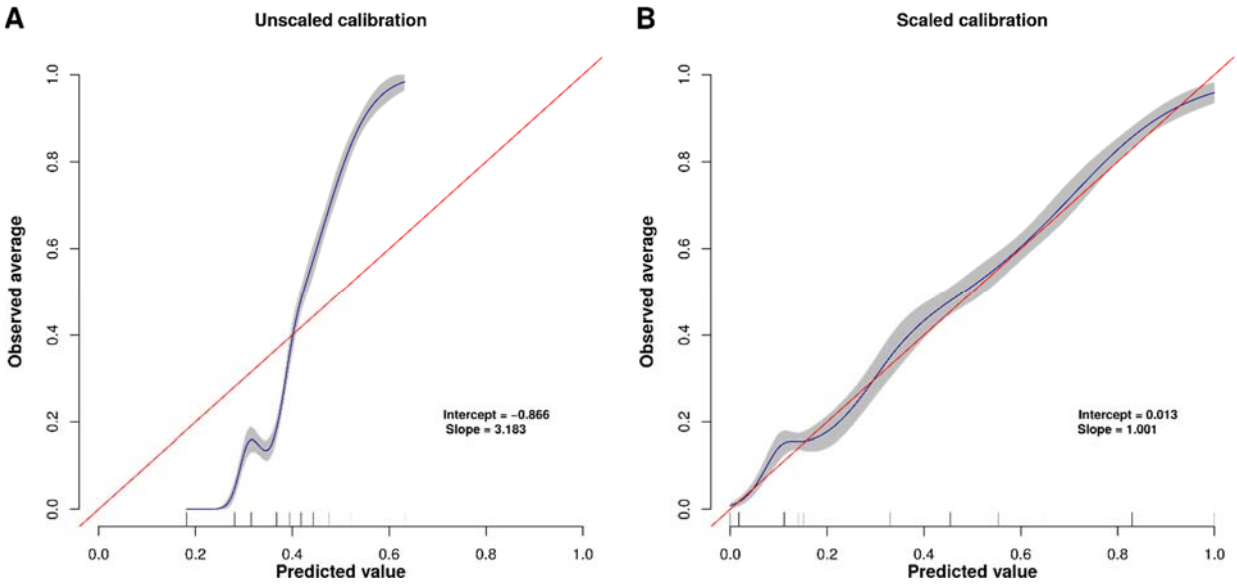
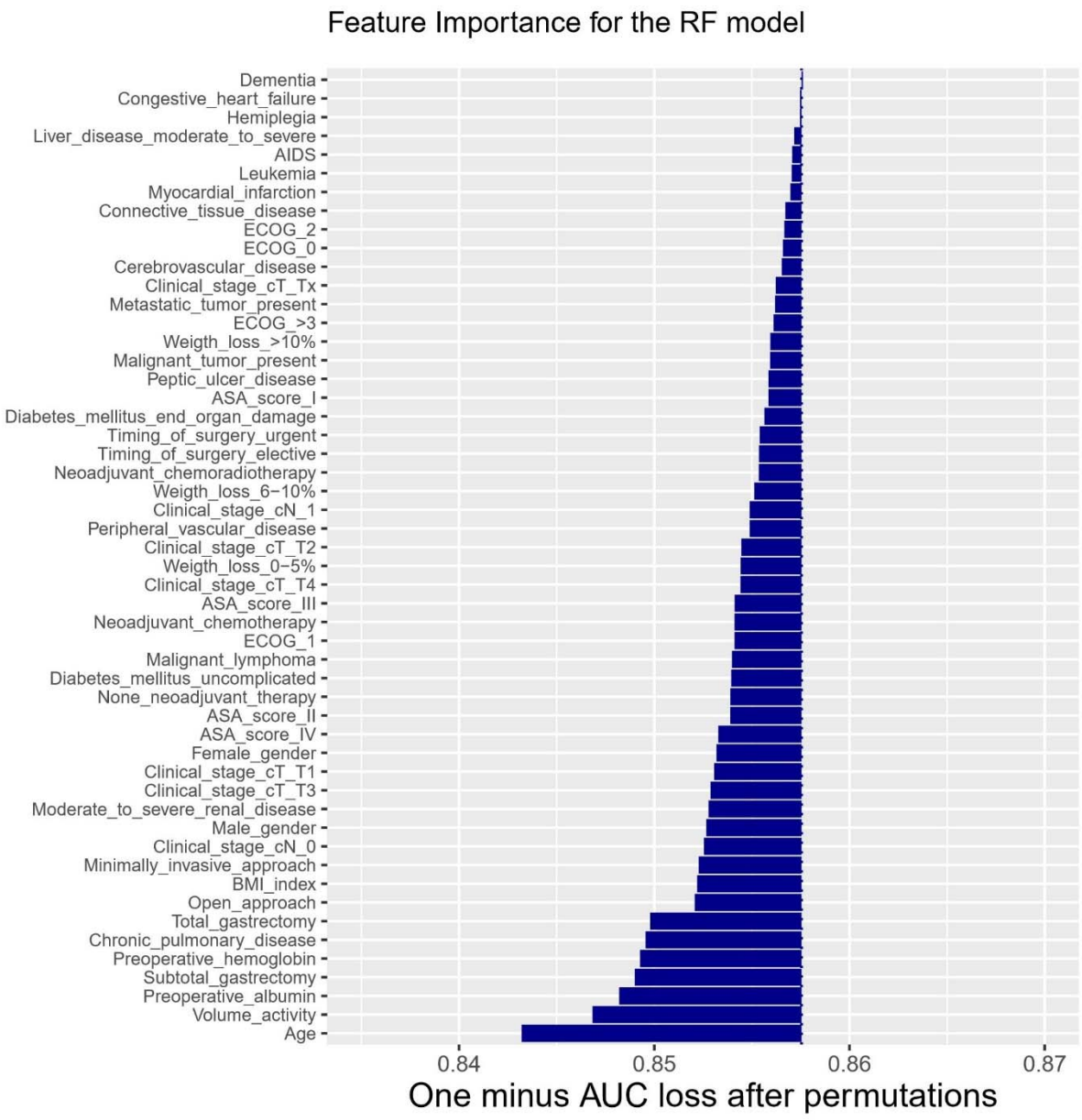
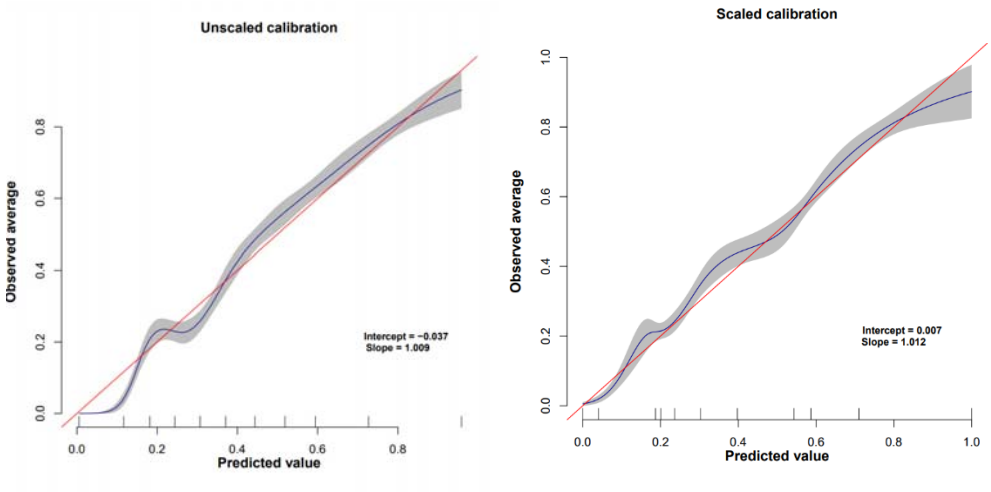

A
Unscaled calibration
Observed average
Predicted value
Intercept = −0.866
Slope = 3.183
B
Scaled calibration
Observed average
Predicted value
Intercept = 0.013
Slope = 1.001

Figure 3.

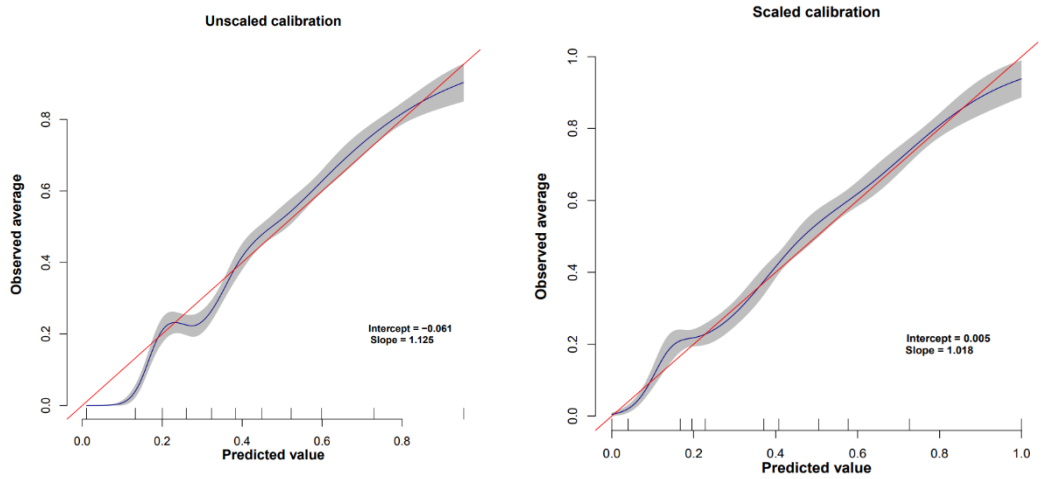

Feature Importance for the RF model
Dementia
Congestive_heart_failure
Hemiplegia
Liver_disease_moderate_to_severe
AIDS
Leukemia
Myocardial_infarction
Connective_tissue_disease
ECOG_2
ECOG_0
Cerebrovascular_disease
Clinical_stage_cT_Tx
Metastatic_tumor_present
ECOG_>3
Weigth_loss_>10%
Malignant_tumor_present
Peptic_ulcer_disease
ASA_score_I
Diabetes_mellitus_end_organ_damage
Timing_of_surgery_urgent
Timing_of_surgery_elective
Neoadjuvant_chemoradiotherapy
Weigth_loss_6−10%
Clinical_stage_cN_1
Peripheral_vascular_disease
Clinical_stage_cT_T2
Weigth_loss_0−5%
Clinical_stage_cT_T4
ASA_score_III
Neoadjuvant_chemotherapy
ECOG_1
Malignant_lymphoma
Diabetes_mellitus_uncomplicated
None_neoadjuvant_therapy
ASA_score_II
ASA_score_IV
Female_gender
Clinical_stage_cT_T1
Clinical_stage_cT_T3
Moderate_to_severe_renal_disease
Male_gender
Clinical_stage_cN_0
Minimally_invasive_approach
BMI_index
Open_approach
Total_gastrectomy
Chronic_pulmonary_disease
Preoperative_hemoglobin
Subtotal_gastrectomy
Preoperative_albumin
Volume_activity
Age
0.84
0.85
0.86
0.87
One minus AUC loss after permutations

Supplementary Figure 1

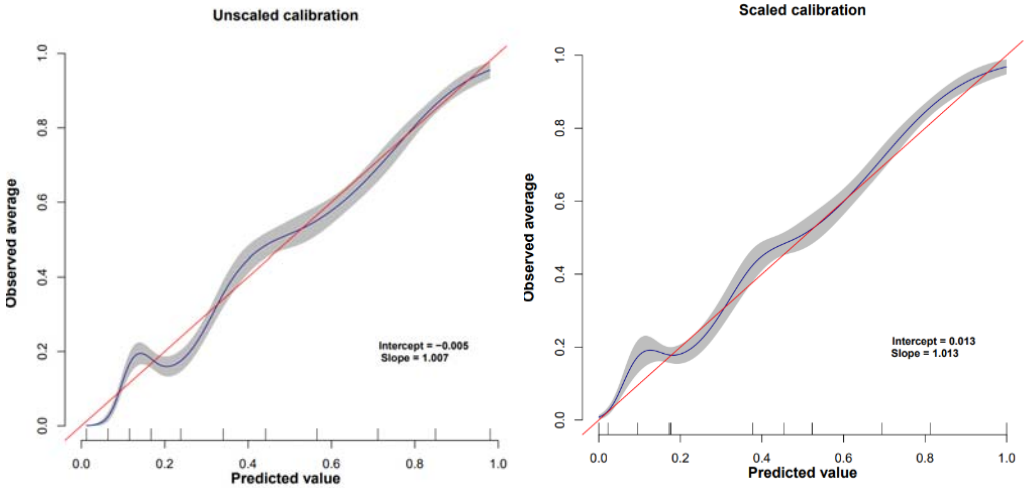
cv-Enet



glmboost



Ensemble



Supplementary Figure 2

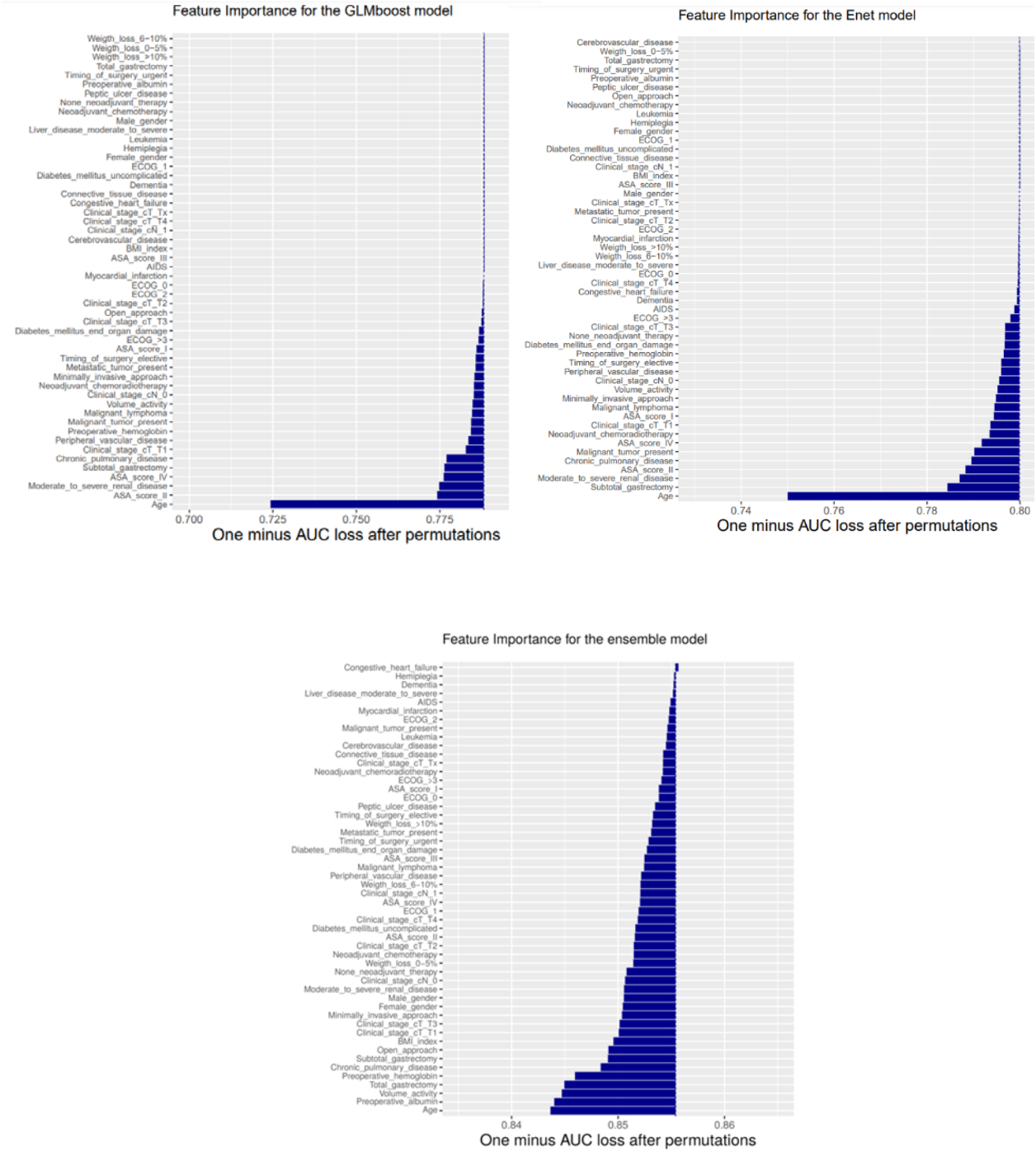


Table 1. Association between different potential risk factors of mortality in the development cohort

	Deceased ≤ 90 days (n=179)	Alive > 90 days (n=3003)	p-value
Gender, n (%)			0.006
Male	129 (72.1)	1849 (61.6)	
Female	50 (27.9)	1154 (38.4)	
Age, years, n (%)	76 (9.6)	70 (12.0)	<0.001
Body mass index*, kg/m^2, mean (SD)	26 (4.8)	26.0 (4.6)	0.170
Missing	8 (4.5)	80 (2.7)	
ECOG performance status, n (%)			<0.001
0	35 (19.6)	1150 (38.3)	
1	98 (54.7)	1563 (52.0)	
2	33 (18.4)	236 (7.9)	
≥3	12 (6.7)	39 (1.3)	
Missing	1 (0.6)	15 (0.5)	
ASA index, n (%)			<0.001
I	0 (0)	110 (3.7)	
II	41 (22.9)	1394 (46.4)	
III	107 (59.8)	1403 (46.7)	
IV	31 (17.3)	96 (3.2)	
Weight loss*, %, n (%)			<0.001
0-5%	109 (60.9)	2055 (68.4)	
6-10	39 (21.8)	564 (18.8)	
>10%	29 (16.2)	361 (12.0)	
Missing	2 (1.1)	23 (0.8)	
Preoperative hemoglobin, gr/dL, mean (SD)	11.0 (1.9)	12.1 (1.9)	<0.001
Missing	2 (1.1)	22 (0.7)	
Preoperative albumin, mg/dL, mean (SD)	35 (7.3)	38 (6)	<0.001
Missing	22 (12.3)	419 (14.0)	
Myocardial infarction, n (n)			<0.001
Yes	29 (16.2)	224 (7.5)	
No	150 (83.8)	2779 (92.5)	
Congestive heart failure, n (%)			0.040
Yes	17 (9.5)	166 (5.5)	
No	162 (90.5)	2837 (94.5)	
Chronic pulmonary disease, n (%)			<0.001
Yes	46 (25.7)	404 (13.5)	
No	133 (74.3)	2599 (86.5)	
Connective tissue disease, n (%)			1.000
Yes	3 (1.7)	44 (1.5)	
No	176 (97.3)	2959 (98.5)	
Peripheral vascular disease, n (%)			<0.001
Yes	31 (11.7)	195 (6.5)	
No	148 (82.7)	2808 (83.5)	
Cerebrovascular disease, n (%)			0.003
Yes	21 (11.7)	179 (6.0)	
No	158 (88.3)	2824 (94.0)	
Dementia, n (%)			0.045
Yes	5 (2.8)	28 (0.9)	
No	174 (97.2)	2975 (99.1)	
Peptic ulcer disease, n (%)			0.570
Yes	11 (6.1)	147 (4.9)	
No	168 (93.9)	2856 (95.1)	
Diabetes mellitus (uncomplicated), n (%)			0.880
Yes	28 (16.2)	491 (16.4)	
No	151 (83.8)	2512 (83.6)	
Diabetes mellitus (end-organ damage), n (%)			<0.001
Yes	21 (11.7)	116 (3.9)	
No	158 (88.3)	2887 (96.1)	

Leukemia, n (%)			1.000
Yes	1 (0.6)	15 (0.5)	
No	178 (99.4)	2988 (99.5)	
Malignant lymphoma, n (%)			0.007
Yes	6 (3.4)	28 (0.9)	
No	173 (96.6)	2975 (99.1)	
Liver disease.moderate-to-severe, n (%)			0.059
Yes	9 (5.0)	73 (2.4)	
No	170 (95.0)	2930 (97.6)	
Hemiplegia, n (%)			0.940
Yes	1 (0.6)	7 (0.2)	
No	178 (99.4)	2996 (99.8)	
Metastatic tumor, n (%)			0.720
Yes	5 (2.8)	31 (1.0)	
No	174 (97.2)	2972 (99.0)	
Renal disease, moderate-to-severe, n (%)			<0.001
Yes	31 (17.3)	131 (4.4)	
No	148 (82.7)	2872 (95.6)	
AIDS, n (%)			0.770
Yes	1 (0.5)	5 (0.2)	
No	178 (99.4)	2998 (99.8)	
Timing of surgery, n (%)			0.005
Elective	160 (89.4)	2842 (94.6)	
Emergency	19 (10.6)	161 (5.4)	
Tumor location, n (%)			0.450
Antrum-pylorus	79 (44.1)	1451 (48.3)	
Corpus-fundus	76 (42.5)	1200 (40.0)	
Linitis	2 (1.1)	31 (1.0)	
Stump	8 (4.5)	73 (2.4)	
Gastro-esophageal junction	14 (7.8)	245 (8.2)	
Missing	0 (0)	3 (0.1)	
Tumor cT stage[&], n (%)			<0.001
T1	14 (7.8)	514 (17.1)	
T2	33 (18.5)	759 (25.3)	
T3	74 (41.3)	1008 (33.6)	
T4	39 (21.8)	530 (17.6)	
Tx	14 (7.8)	159 (5.3)	
Missing	5 (2.8)	33 (1.1)	
Tumor cN stage[&], n (%)			0.160
Negative	88 (49.1)	1683 (56.0)	
Positive	85 (47.5)	1292 (43.0)	
Missing	6 (3.4)	28 (1.0)	
Neoadjuvant therapy, n (%)			<0.001
None	147 (82.1)	2085 (69.5)	
Chemoradiotherapy	0 (0)	54 (1.8)	
Chemotherapy	31 (17.3)	857 (28.5)	
Missing	1 (0.6)	7 (0.2)	
Surgical approach, n (%)			<0.001
Open	121 (67.6)	1585 (52.8)	
Laparoscopic	58 (32.4)	1418 (47.2)	
Type of gastrectomy, n (%)			0.170
Total	86 (48.0)	1278 (42.6)	
Partial	93 (52.0)	1725 (57.4)	
Volume activity, mean/year/center, mean (SD)	23 (10)	24 (10)	0.250

*At the time of diagnosis; [&]according to the 7th edition of the AJCC; Abbreviations: ECOG, Eastern Cooperative Oncology Group; ASA, American Society of Anesthesiologists; AIDS: Acquired Immune Deficiency Syndrome.

Supplementary Table 1. Comparison data between the development and validation cohorts

	Development set (n=3,182)	Validation set (n=260)	*P* value
90DM, n (%)	179 (5.6)	16 (6.2)	0.830
Gender, n (%)			0.031
Male	1978 (62.1)	179 (68.8)	
Female	1204 (37.9)	80 (30.8)	
Missing	0 (0)	1 (0.64)	
Age, years, n (%)	70 (12.0)	71 (11.0)	0.370
Body mass index*, kg/m^2, mean (SD)	26 (4.6)	27 (4.3)	0.050
Missing	88 (2.8)	9 (3.4)	
ECOG performance status, n (%)			0.070
0	1185 (37.2)	100 (38.5)	
1	1661 (52.2)	119 (45.8)	
2	269 (8.5)	31 (11.9)	
≥3	51 (1.6)	7 (2.7)	
Missing	16 (0.5)	3 (1.1)	
ASA index, n (%)			0.110
I	110 (3.4)	8 (3.0)	
II	1435 (45.1)	97 (37.3)	
III	1510 (47.5)	141 (54.3)	
IV	127 (4.0)	12 (4.6)	
Missing	0 (0)	2 (0.8)	
Weight loss*, %, n (%)			0.150
0-5%	2164 (68.0)	186 (71.6)	
6-10	603 (19.0)	37 (14.2)	
>10%	390 (12.3)	36 (13.8)	
Missing	25 (0.7)	1 (0.4)	
Preoperative hemoglobin, gr/dL, mean (SD)	12.0 (1.9)	12.1 (2.0)	0.830
Missing	24 (0.7)	1 (0.4)	
Preoperative albumin, mg/dL, mean (SD)	38 (6.2)	38 (5.9)	0.150
Missing	441 (13.9)	26 (10.0)	
Myocardial infarction, n (n)			0.110
Yes	253 (8.0)	13 (5.0)	
No	2929 (92.0)	247 (95.0)	
Congestive heart failure, n (%)			0.700
Yes	183 (5.8)	17 (6.5)	
No	2999 (94.2)	243 (93.5)	
Chronic pulmonary disease, n (%)			0.910
Yes	450 (14.1)	38 (14.6)	
No	2732 (85.9)	222 (85.4)	
Connective tissue disease, n (%)			0.090
Yes	47 (1.5)	0 (0)	
No	3135 (98.5)	260 (100)	
Peripheral vascular disease, n (%)			0.650
Yes	226 (7.1)	16 (6.2)	
No	2956 (92.9)	244 (93.8)	
Cerebrovascular disease, n (%)			0.050
Yes	200 (6.3)	25 (9.6)	
No	2982 (93.7)	235 (90.4)	
Dementia, n (%)			0.490
Yes	33 (1.0)	1 (0.4)	
No	3149 (99.0)	259 (99.6)	
Peptic ulcer, n (%)			0.220
Yes	158 (5.0)	8 (3.1)	
No	3024 (95.0)	252 (96.9)	
Diabetes mellitus (uncomplicated), n (%)			0.330
Yes	519 (16.3)	49 (18.8)	
No	2663 (83.7)	211 (81.2)	

Diabetes mellitus (end-organ damage), n (%)			0.940
Yes	137 (4.3)	12 (4.6)	
No	3045 (95.7)	248 (95.4)	
Leukemia, n (%)			0.900
Yes	16 (8.8)	2 (0.8)	
No	3166 (91.2)	258 (99.2)	
Malignant lymphoma, n (%)			1.000
Yes	34 (1.1)	3 (1.2)	
No	3148 (98.9)	257 (98.8)	
Liver disease, moderate-to-severe, n (%)			0.950
Yes	82 (2.6)	6 (2.3)	
No	3100 (97.4)	254 (97.7)	
Hemiplegia, n (%)			0.370
Yes	8 (0.3)	2 (0.8)	
No	3174 (99.7)	258 (99.2)	
Metastatic tumor, n (%)			0.820
Yes	36 (1.1)	2 (0.8)	
No	3146 (98.9)	258 (99.2))	
Renal disease, moderate-to-severe, n (%)			0.260
Yes	162 (5.1)	18 (6.9)	
No	3020 (94.9)	242 (93.1)	
AIDS, n (%)			1.000
Yes	6 (0.2)	0 (0)	
No	3176 (99.8)	260 (100)	
Timing of surgery, n (%)			0.031
Elective	3002 (94.3)	254 (97.7)	
Emergency	180 (5.7)	6 (2.3)	
Tumor location, n (%)			0.033
Antrum-pylorus	1530 (48.2)	131 (50.4)	
Corpus-fundus	1276 (40.1)	91 (35.0)	
Linitis	33 (1.0)	8 (3.1)	
Stump	81 (2.5)	8 (3.14)	
Gastro-esophageal junction	259 (8.1)	22 (8.5)	
Missing	3 (0.1)	0 (0)	
Tumor cT stage[&], n (%)			0.100
T1	528 (16.6)	36 (13.8)	
T2	792 (24.9)	65 (25.0)	
T3	1082 (34.0)	80 (30.8)	
T4	569 (17.9)	57 (21.9)	
Tx	173 (5.4)	22 (8.5)	
Missing	38 (1.2)	0 (0)	
Tumor cN stage[&], n (%)			0..830
Negative	1771 (55.7)	148 (56.9)	
Positive	1377 (43.3)	111 (42.7)	
Missing	34 (1.0)	1 (0.4)	
Neoadjuvant therapy, n (%)			<0.001
None	2232 (70.1)	155 (59.6)	
Chemoradiotherapy	54 (1.7)	0 (0)	
Chemotherapy	888 (27.9)	103 (39.6)	
Missing	8 (0.3)	2 (0.8)	
Surgical approach, n (%)			<0.001
Open	1706 (67.6)	101 (38.8)	
Laparoscopic	1476 (31.3)	155 (59.6)	
Missing	0 (0)	4 (1.6)	
Type of gastrectomy, n (%)			<0.001
Total	1364 (42.9)	83 (31.9)	
Partial	1818 (57.1)	177 (68.2)	

*At the moment of the diagnosis; [&]according to the 7th edition of the AJCC; Abbreviations: 90DM, 90-day mortality; ECOG, Eastern Cooperative Oncology Group; ASA, American Society of Anesthesiologists; AIDS: Acquired Immune Deficiency Syndrome.

TRIPOD Checklist: Prediction Model Development and Validation

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	0
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	1-2
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	3-4
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	4
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	5
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	6
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5
	5b	D;V	Describe eligibility criteria for participants.	6
	5c	D;V	Give details of treatments received, if relevant.	5-6
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	6
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	6
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	NA
Sample size	8	D;V	Explain how the study size was arrived at.	9
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	6
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	7
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	7-8
	10c	V	For validation, describe how the predictions were calculated.	7
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	7
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	7-8
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	NA
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	6
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	9
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	9, 23-24
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	30-31
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	9
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	10
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	8
	15b	D	Explain how to the use the prediction model.	8
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	10
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	10
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	14
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	12-13
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	12-15
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	15
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	8,28-31
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	NA

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.