This is the accepted version (peer reviewed) of the article published in final form in the ***Journal of Trauma and Acute Care Surgery***, February 2023, DOI: https://doi.org/10.1097/ta.0000000000003904 PMID: 36726191

Prospective multicenter external validation of postoperative mortality prediction tools in patients undergoing emergency laparotomy

Authors:

Stamatios Kokkinakis, MD, MSc[1](ORCID: 0000-0003-2733-314X), Evangelos I. Kritsotakis, PhD, CStat[2](ORCID: 0000-0002-9526-3852), Konstantinos Paterakis, MD, MSc1, Garyfallia-Apostolia Karali, MD[1], Vironas Malikides, MD[1], Anna Kyprianou, MD[1], Melina Papalexandraki, MD[1], Charalampos S. Anastasiadis, MD, MSc[3], Odysseas Zoras, MD, PhD, FACS[3], Nikolas Drakos, MD[4], Ioannis Kehagias, MD, PhD[4], Dimitrios Kehagias, MD[4], Nikolaos Gouvas, MD, PhD[5], Georgios Kokkinos, MD[5], Ioanna Pozotou, MD[5], Panagiotis Papatheodorou, MD[5], Kyriakos Frantzeskou, MD5, Dimitrios Schizas, MD, PhD6, Athanasios Syllaios, MD, MSc6, Ifaistion M Palios, MD7, Konstantinos Nastos, MD, PhD8, Markos Perdikaris, MD8, Nikolaos V Michalopoulos, MD, MSc, PhD8, Ioannis Margaris, MD8, Evangelos Lolis, MD9, Georgia Dimopoulou, MD9, Dimitrios Panagiotou, MD10, Vasiliki Nikolaou, MD10, Georgios K. Glantzounis, MD, PhD11, George Pappas-Gogos, MD11, Kostas Tepelenis, MD11, Georgios Zacharioudakis, MD, PhD12, Savvas Tsaramanidis, MD12, Ioannis Patsarikas, MD12, Georgios Stylianidis, MD13, Georgios Giannos, MD13, Michail Karanikas, MD, MSc, PhD14, Konstantinia Kofina, MD14, Markos Markou, MD14, Emmanuel Chrysos, MD, PhD, FACS1, Konstantinos Lasithiotakis, MD, PhD1, FEBS, FRCS (ORCID: 0000-0002-6538-0951)

Affiliations

1. Department of General Surgery, University Hospital of Heraklion, University of Crete, School of Medicine, Greece.

2. Laboratory of Biostatistics, School of Medicine, University of Crete, Heraklion, Crete, Greece.

3. Department of Surgical Oncology, University Hospital of Heraklion, University of Crete, School of Medicine, Greece.

4. Department of Surgery, University General Hospital of Patras, School of Medicine, University of Patras, Patras, Greece.

5. Department of Surgery, General Hospital of Nicosia, School of Medicine, University of Cyprus, Nicosia, Cyprus.

6. First Department of Surgery, National and Kapodistrian University of Athens, Laikon General Hospital, Athens, Greece.

7. Second Propaedeutic Department of Surgery, National and Kapodistrian University of Athens, Laikon General Hospital, Athens, Greece.

8. Department of Surgery, University General Hospital Attikon, School of Medicine, University of Athens, Athens, Greece.

9. Department of Surgery, General Hospital of Volos, Volos, Greece.

10. Department of Surgery, General Hospital of Trikala, Trikala, Greece.

11. Department of Surgery, University Hospital of Ioannina, Greece.

12. Department of Surgery, Ippokrateion General Hospital of Thessaloniki, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece.

13. Second Department of Surgery, Evangelismos General Hospital, Athens, Greece.

14. Department of Surgery, University General Hospital of Alexandroupolis, School of Medicine, University of Thrace, Alexandroupolis, Greece.

Disclosure Information: Nothing to disclose.

Short title: Prognostic models for emergency laparotomy

Correspondence: K. Lasithiotakis, Department of General Surgery, University Hospital of Crete, Heraklion, 71110, Greece, Email: k.lasithiotakis@uoc.gr, Tel: +30 2810392676, Fax: +30 2810392380

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authorship

SK, EIK and KL contributed to the study conception and design, data analysis and interpretation and manuscript preparation. SK, KL, KP, GAK, VM, AK, MP, CSA, ND, DK, GK, IP, PP, KF, AS, IMP, MP, IM, GD, DP, VN, KT, ST, IP, GS, GG, KK and MM contributed to the acquisition of data. EIK contributed to analysis and interpretation of data (lead). SK, OZ, IK, NG, DS, KN, NVM, EL, GKG, GPG, GZ, MK, EC and KL contributed to the critical review/revision. KL is the principal investigator and takes primary responsibility for the manuscript. All authors read and approved the final version of the manuscript.

ABSTRACT

BACKGROUND: Accurate preoperative risk assessment in emergency laparotomy (EL) is valuable for informed decision-making and rational use of resources. Available risk prediction tools have not been validated adequately across diverse healthcare settings. Herein, we report a comparative external validation of 4 widely cited prognostic models.

METHODS: A multicenter cohort was prospectively composed of consecutive patients undergoing EL in 11 Greek hospitals from January 2020 to May 2021 using the National Emergency Laparotomy (NELA) audit inclusion criteria. 30-day mortality risk predictions were calculated using the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP), NELA, Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (P-POSSUM) and Predictive Optimal Trees in Emergency Surgery Risk (POTTER) tools. Surgeons' assessment of postoperative mortality using pre-defined cutoffs was recorded, and a surgeon-adjusted ACS-NSQIP prediction was calculated when the original model's prediction was relatively low. Predictive performances were compared using scaled Brier scores, discrimination and calibration measures and plots, and decision curve analysis. Heterogeneity across hospitals was assessed by random-effects meta-analysis.

RESULTS: 631 patients were included and 30-day mortality was 16.3%. The ACS-NSQIP and its surgeon-adjusted version had the highest scaled Brier scores. All models presented high discriminative ability, with concordance statistics ranging from 0.79 for P-POSSUM to 0.85 for NELA. However, except the surgeon-adjusted ACS-NSQIP (Hosmer-Lemeshow test p=0.742), all other models were poorly calibrated (p <0.001). Decision curve analysis revealed superior clinical utility of the ACS-NSQIP. Following recalibrations, predictive accuracy improved for all models but ACS-NSQIP retained the lead. Between-hospital heterogeneity was minimum for the ACS-NSQIP model and maximum for P-POSSUM.

CONCLUSION: The ACS-NSQIP tool was most accurate for mortality predictions after EL in a broad external validation cohort, demonstrating utility for facilitating preoperative risk management in the Greek healthcare system. Subjective surgeon assessments of patient prognosis may optimise ACS-NSQIP predictions.

Level of Evidence: Level II, Diagnostic test/criteria

Keywords: Laparotomy; prediction rule; mortality; risk; validation; clinical decision support.

BACKGROUND

Emergency laparotomy (EL) is a common procedure performed worldwide for a wide variety of abdominal pathologies. Despite documented advances in the modern era,[1] mortality following EL remains substantial worldwide, affecting up to 1 of every 5 patients in the first 30 postoperative days in high-quality health care systems.[2–5] Efforts to standardise perioperative care of EL patients through implementation of predetermined pathways have led to reduction in postoperative mortality.[6] Standardisation in contemporary practice requires calculation and consideration of the risks associated with EL before entering the operating room.[7] Preoperative risk stratification may result in rational utilisation of escalated levels of care postoperatively, higher level of consultant involvement in high-risk patients, improvement of communication between surgical disciplines and optimal shared decision-making.[8,9] For patients undergoing EL, factors such as age, comorbidity and waiting time from admission to operation have been associated with worse outcomes,[10] and have been subsequently combined in multivariable risk prediction models. Their use is embraced by modern guidelines but no specific recommendation on the best model has been made.[11]

There are few external validation studies directly comparing between risk prediction models in EL.[12,13] In a recent review of the applicability of risk stratification tools to emergency general surgery, the authors identified the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) tool as best fitting their definition of the ideal scoring tool.[14] Several other risk prediction tools have been variably proposed and widely cited, including the National Emergency Laparotomy Audit (NELA) tool, the Portsmouth-Physiologic and Operative Severity Score for the enUmeration of Mortality and morbidity (P-POSSUM) and the Predictive OpTimal Trees in Emergency Surgery Risk (POTTER). These tools have demonstrated excellent predictive performance in the populations in which they were developed (mainly in the UK and USA) but their broader transportability in diverse external settings has not been adequately validated. Previous reports have revealed significant differences in the management of EL in this population compared to the UK.[15,16]

The present study performed comparative external validation of 4 common risk prediction tools (ACS-NSQIP, NELA, P-POSSUM and POTTER), in a multicenter prospective cohort design, to identify the best tool for predicting 30-day mortality in Greek patients undergoing EL.

METHODS

Ethics and reporting

The study was approved by the Institutional Review Board and the Bioethics Committee of the participating institutions and is reported according to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement (**SDC Table S5**).[17] The study was registered in ClinicalTrials.gov (identifier NCT04615520).

Study design and participants

A multicenter cohort was prospectively assembled by enrolling consecutive patients undergoing EL in 10 hospitals in Greece and 1 hospital in Cyprus (1 secondary-care, 2 tertiary-care and 8 university-affiliated hospitals), from January 2020 to May 2021. All participating centers submitted prospectively collected anonymised data on patients undergoing EL. Each patient was followed up until the 30th postoperative day. Patient inclusion and exclusion criteria were similar to those used in the NELA audit (**SDC Table S1**). Briefly, all patients who had EL were eligible for this study except for appendectomy, cholecystectomy, negative diagnostic laparotomy or laparoscopy, biopsy, non-gastrointestinal surgery and elective gastrointestinal surgery. Only adults (>=18 years) were enrolled.

Outcomes and predictors

The primary endpoint was 30-day postoperative mortality. Demographic data, preoperative variables required for each prediction model (NELA, P-POSSUM, ACS-NSQIP and POTTER), type of operation and postoperative outcomes were prospectively recorded for each patient. The data were then entered into respective online calculators to make predictions for the risk of 30-day postoperative death for each patient from each model. Before surgery, attending surgeons answered the following question for each patient: "In your clinical judgement, what is the risk of death within 30 days?" with 4 ordered response options, namely <5%, 5-10%, 10-20% or >20%. The ACS-NSQIP online calculator allows clinicians to adjust for underestimation by increasing the estimated risks based on their subjective impression of the patient. We independently simulated this process by opting for adjustment on the ACS-NSQIP calculator whenever the 30-day mortality risk prediction was lower than the surgeon's preoperative assessment as shown **in SDC Table S2** (we refer to this as surgeon-adjusted ACS-NSQIP). We calculated the ACS-NSQIP predicted risk both with and without incorporating the subjective surgeon's assessment. For patients discharged prior to the 30-day mark, we scheduled office visits and follow-up calls from study personnel, in which relevant outcomes were documented.

Sample size

For external validation of prognostic models, a minimum of 100 outcome events is recommended to ensure adequate power to detect changes in predictive performance metrics in external datasets.[18,19] We therefore recruited patients for this study over 17 months until about 100 postoperative deaths occurred in our cohort.

Missing data

Data were readily available to calculate risk predictions for at least 98% of the patients using the NELA, P-POSSUM, and ACS-NSQIP models, but only for 486 (77%) patients based on POTTER. More than half of the missing data for POTTER originated from 2 hospitals, where POTTER predictions could not be calculated for 48% and 98% of their patients, respectively. This was because variables necessary for the POTTER tool, such as pre-operative albumin, were not routinely available preoperatively in those hospitals. As the prognostic models are aimed at being applied in clinic, we opted for complete case analysis for each model. Moreover, there was no evidence of association between missing POTTER prediction and mortality rate (15.3% in patients with missing POTTER vs 20% in those with non-missing POTTER, p=0.172), suggesting that complete case analysis may not bias our results.

Statistical methods

The analysis aimed to estimate and compare metrics of predictive performance and utility for decision-making of the selected prediction models of 30-day mortality, when applied to our independent cohort of patients undergoing EL. The subjective preoperative assessment of patient's prognosis by their surgeon was considered as a minimum benchmark that any useful prediction tool should outperform. In addition, we examined the possibility of improving predictive performance by recalibrating the models and assessed the heterogeneity of predictive accuracy between hospitals.

The Brier score (mean squared error between observed and predicted probabilities) was used as a comparative measure of overall model performance, which we scaled by its maximum value under a non-informative null model to let it range up to a theoretical maximum of 100%. The scaled Brier score represents the amount of prediction error in a null model that is explained by the model under validation. Higher values of the scaled Brier score indicate better prediction accuracy (negative values indicate a potentially harmful model).[20] Bootstrap resampling (500 replications) was applied to compute 95% confidence intervals (CI) for the scaled Brier scores. The discriminatory ability of each model (to rank patients according to risk) was quantified by calculating the concordance c-statistic as the area under the receiver operating characteristic curve (AUC) with exact binomial 95% CI. The DeLong test for correlated data was employed to compare each model's AUC with the minimum benchmark.[21]

We emphasised on the calibration of the models (agreement between the predicted and observed numbers of outcome events), because adequate calibration ensures an accurate absolute risk is communicated to patients and physicians.[22,23] The Hosmer–Lemeshow goodness-of-fit test was used to broadly assess calibration in deciles of predicted risks (a small p-value indicates poor calibration). The ratio of expected to observed outcome events (E:O ratio) was calculated, which ideally should be 1 (E:O < 1 indicates underestimation and E:O > 1 indicates overestimation of the total number of deaths). In addition, a calibration regression line was made (plotting predicted against observed risks). The intercept and slope of the calibration regression line and their Wald-type 95% Cis were estimated using logistic regression.[22,23] The calibration intercept, also known as calibration-in-the-large (CITL), compares the average predicted risk with the overall event rate and, ideally, should be 0 (CITL < 0 indicates the predictions are systematically too high, whereas CITL > 0 indicates the predictions are systematically too low). The calibration slope evaluates the spread of the predicted risks and has a target value of 1. When CITL is close to 0, a slope close to 1 indicates that good calibration is also maintained across the range of individuals.[23,24] Additionally, lowess-smoothed calibration curves were constructed to allow for visual inspection of calibration.[25]

Decision-curve analysis (DCA) was employed to provide insight into the range of decision thresholds to label a patient as 'high risk for postoperative death' that would have highest net benefit (NB) for decision-making when using each risk prediction model for this purpose.[26] NB is defined as the difference between the proportion of true positives (labelled as high risk pre-operatively and then going on to die within 30-days of EL) and the proportion of false positives (labelled as high risk but not going on to die within 30-days) weighted by the odds of the selected threshold for the high-risk label. At any given threshold, the model with higher NB is the preferred model.[26] We examined risk thresholds between 5% and 50%.

As there was different case-mix between our cohort and the cohorts on which the risk prediction tools were originally developed, we examined whether adjusting (recalibrating) each model's intercept and slope would result in better calibrated predictions.[22,23] Finally, we examined the variability in performance across hospitals (heterogeneity) using random effects meta-analysis of hospital-specific scaled Brier scores. Heterogeneity was quantified using 95% prediction intervals, which indicate the dispersion in performance metrics that can be expected when applying the model in a new centre. [27]

All statistical analyses were performed using STATA v.17 (StataCorp, College Station, TX, USA).

Development vs Validation

The clinical setting and eligibility criteria in this study were similar to those of the NELA model, to allow for a reasonable comparison of outcomes.[28] NELA was initiated in 2014 in the UK as a specific tool for emergency laparotomies, focusing on 30-day postoperative mortality as main outcome.[28] The ACS-NSQIP was developed in a broader setting that included both elective and emergency procedures from a variety of surgical subspecialties in the USA between 2009 and 2012.[29] Only 59% of the development cohort for ACS-NSQIP were general surgery patients. Outcomes of interest in ACS-NSQIP were 30-day mortality, common postoperative complications, and procedure specific complications such as anastomotic leak.[29] The POTTER model was developed on a subset of the data from the ACS-NSQIP database, conditioning on patients who underwent emergency surgery between 2007 and 2013, to predict outcomes similar those targeted by the ACS-NSQIP model.[30] P-POSSUM was a modification to correct for over prediction of mortality from the initial POSSUM equation that was developed based on patients who underwent both elective and emergency surgery between 1993 and 1995, excluding day-surgery and pediatric cases.[31]

RESULTS

Patient characteristics

Over the 17-month recruitment period, 633 patients underwent EL in the 11 participating hospitals and were enrolled in the study. 2 patients were lost to follow-up and were excluded from analysis. The remaining 631 were included in the final analysis. Case ascertainment rate was high and it has been described in detail in the report of the Hellenic Emergency Laparotomy Study (HELAS).[15] The patients had a mean age of 66 years (range 19-99 years), 54% were male and 43% were classified as ASA status III/IV. The most common indication for EL was gastrointestinal obstruction (39%), followed by perforation (36%) and ischemia (15%). Demographic and clinical characteristics are detailed in **Table 1**.

Observed and predicted mortality rates

There were 103 deaths within 30-days of EL, an overall 30-day mortality rate of 16.3% (95%CI 13.5% to 19.4%). The surgeons provided subjective preoperative risk assessment for all but 21 (3.3%) patients. Compared to the actual mortality rate, the average predicted risks were lower for the POTTER (8.9%), NELA (10.5%) and ACS-NSQIP (12.2%) models, much higher than observed mortality for P-POSSUM (19.9%) and about similar to observed mortality (within CI limits) for the surgeon’s subjective assessment (14.9%) and the surgeon-adjusted ACS-NSQIP model (18.1%). All models assigned a significantly higher mean predicted risk to the group of patients who eventually died within 30-days of EL than those who survived (**Table 2**).

Case mix

SDC Table S3 compares the distribution of context-important clinical characteristics and outcomes between the present study cohort and the cohorts of patients on which the development of the NELA and ACS-NSQIP models were based. The current cohort appears to represent a different case-mix of patients with higher mortality compared to the original model development cohorts of NELA and ACS-NSQIP.

Predictive performance

Predictive performance metrics are shown in **Table 3**. The overall Brier scaled score was highest for ACS-NSQIP (22.4%, 95% CI 14.5% to 30.3%) and surgeon-adjusted ACS-NSQIP (20.6%, 95% CI 13.4% to 27.2%), and lowest for surgeon's assessment (10.6%, 95% CI 1.3% to 18.7%) and P-POSSUM (1.5%, 95% CI 0.0% to 13.1%). Discrimination was excellent for all models (**SDC Figure S1**), with AUC point estimates ranging from 0.79 (surgeon and P-POSSUM) to 0.85 (NELA). DeLong's tests showed that NELA and ACS-NSQIP had significantly higher AUCs than the minimum benchmark of the surgeon's preoperative assessment, whereas no statistically significant difference from the minimum benchmark was observed for P-POSSUM (p=0.868) and POTTER (p=0.081). The Hosmer-Lemeshow test indicated poor agreement of observed and predicted risks in decile groups for all models (all p<0.001), except the surgeon-adjusted ACS-NSQIP model (p=0.742; **SDC Figure S2**). As seen in **Table 3**, the CITL statistic indicated that POTTER, NELA and ACS-NSQIP produced (in this order of magnitude) predictions that were systematically too low (CILT CI limits above zero), whereas P-POSSUM systematically overestimated mortality (CILT CI limits below zero). In contrast, no significant deviation from the ideal CILT was seen for the surgeon-adjusted ACS-NSQIP model, which was the only model with an acceptable calibration slope (slope CI limits spanning 1). **Figure 1** shows smoothed calibration plots for each model confirming visually the superior calibration of the ACS-NSQIP model, especially its surgeon-adjusted version.

Decision curve analysis

Figure 2 shows that all models had positive NB for decision thresholds up to about 40% mortality risk, but best overall utility on wider ranges of thresholds was maintained for the ACS-NSQIP and surgeon-adjusted ACS-NSQIP models.

Model recalibration

After intercept and slope adjustments, scaled Brier scores and calibration metrics improved substantially for all models, with ACS-NSQIP retaining lead performance (**SDC Table S4**). As seen in **SDC Figure S3**, the flexible calibration curves of all updated models were much closer to the diagonal reference line of perfect calibration compared to the original unadjusted models. There was evident underestimation of mortality risks in very high-risk patients from the recalibrated NELA and P-POSSUM models. DCA on recalibrated models is shown on **SDC Figure S4** and confirmed that ACS-NSQIP had best overall clinical utility to select high-risk patients.

Heterogeneity

Random-effects meta-analysis revealed substantial and statistically significant heterogeneity of hospital-specific Brier scores for NELA, P-POSSUM and POTTER, whereas heterogeneity was low and non-significant for ACS-NSQIP and the surgeon-adjusted ACS-NSQIP (**SDC figures S5-**

S10). As seen from the 95% prediction intervals in **Figure 3**, future new studies would maintain their scaled Brier score within acceptable limits only for ACS-NSQIP and its surgeon-adjusted version.

DISCUSSION

We presented the results of a comprehensive external validation of 4 commonly cited prognostic models when applied on a large multicentre cohort of Greek patients who underwent EL over 17 months at 11 hospitals. Discordant case-mix between this cohort and cohorts where the models were originally developed implies that this study assesses broader transportability of the models in a different setting rather than mere reproducibility on patients similar to those in the original model development cohorts. The results of this assessment favor the use of the ACS-NSQIP model, including its surgeon-adjusted version, when assessing the prognosis of EL patients. ACS-NSQIP was seen to outperform all other models with respect to several metrics of predictive performance, demonstrated clinical utility on a wider range of risk thresholds for decision-making, and exhibited minimum heterogeneity across hospitals.

This is the first study to perform comparative external validation of different risk prediction tools for EL patients in a prospective design in Greece. In contrast to our findings, previously performed comparative external validations for emergency abdominal surgery in other countries have mostly favoured the NELA risk prediction tool.[12,13,32,33] In a cohort of 758 EL patients in New Zealand, NELA presented superior discrimination and calibration, compared to ACS-NSQIP, P-POSSUM and APACHE II.[13] Lai et al compared NELA to P-POSSUM in an Asian population and concluded that NELA predicts 30-day mortality more accurately.[32] A multicentre Australian study concluded that the NELA was highly sensitive and comparable with the P-POSSUM and

ACS-NSQIP models in EL patients.[33] A recent analysis of 650 EL patients of the NELA database favoured the discriminative power of the NELA compared to that of the P-POSSUM.[12] Aforementioned studies were all retrospective in nature, and 30-day mortality was reported in no more than 60 patients (range 47-60), implying external validation statistics may have been relatively imprecise. Only 2 of those studies involved comparisons with the ACS-NSQIP model.[13,33] A recent meta-analysis of the accuracy of ACS-NSQIP in emergency abdominal surgery also pointed out that the existing literature consists of exclusively retrospective studies that are mostly underpowered.[34]

The predictive ability of ACS-NSQIP in emergency surgical patients has been seen to be inferior to that in elective cases, with reported underestimation of the mortality risk of patients undergoing emergency surgery.[35] Use of subjective surgeon assessment and subsequent utilisation of surgeon-adjusted risk scores has been reported in geriatric patients undergoing lumbar surgery[36]. For patients judged to have "somewhat higher than estimated" risks according to the Surgeon Adjustment Score (SAS), the ACS-NSQIP prediction of postoperative mortality was accurate, while in patients with "significantly higher than estimated" risks, the model accurately predicted the risks of surgical site infection and reoperation.[36] The importance of combining subjective assessment with an objective risk prediction tool was emphasized in a recent study validating prediction models for surgical patients, which reported that the combination of the best predictive model with the surgeon's subjective assessment was superior than any predictive model alone.[37] Similarly we found that calibrating the ACS-NSQIP prediction of 30-day mortality on the basis of preoperative subjective assessment improved predictive performance in our cohort of patients undergoing EL.

External validations should use standardised performance metrics and adhere to guidelines for reporting model performance to allow for comprehensive and informative comparisons of prognostic models in different patient populations. Reporting of multivariable prediction models has been shown to be insufficient before the implementation of the TRIPOD statement.[38] A recent systematic review emphasized the lack of prospective cohorts in external validation studies and revealed that reporting of key performance measures was largely incomplete with median completeness of the TRIPOD checklist at only 61%.[39] Methodological issues, such as poor assessment of calibration, have also been pointed out in a systematic review of risk assessment tools for EL[40]. Therefore, we strictly followed the TRIPOD guidelines and utilised multiple metrics of predictive performance to thoroughly compare the prediction models in this study.

Good discrimination and calibration metrics do not necessarily warrant that use of a model will aid decision-making.[26] We therefore performed DCA to identify the model that demonstrates superiority in selecting a “high-risk” patient with highest net benefit upon use in clinical practice.[26] The results of the DCA are to be interpreted with caution. Superiority of the ACS-NSQIP implies that it should be used in everyday practice as part of a shared decision-making in our setting, without choosing a specific threshold.[41] In a cohort of patients undergoing hepatopancreaticobiliary surgery, DCA performed on ACS-NSQIP and POTTER favored their use for guiding interventions on important outcomes, such as readmission and venous thromboembolism.[42] In that cohort, ACS-NSQIP had superior discrimination, but POTTER demonstrated net benefit for a wider range of risk thresholds for venous thromboembolism risk, and the authors pointed out the importance of not relying solely on metrics, such as the concordance statistic (or AUC).[42] Assessing the performance of prediction models solely on the basis of concordance statistics may lead to mislead conclusions, because a model’s discriminatory performance is bound to be lower in a homogenous sample with restricted case-mix and does not provide information about the calibration of predicted risks with observed events.[43,44] The results of DCA in this study indicate that using the ACS-NSQIP model and its surgeon-adjusted version may guide the surgeon to modify interventions and the net benefit is maintained over a wider range of thresholds for defining a "high-risk" patient compared to the other models assessed here.

Our study has a number of limitations. First, it is important to acknowledge that this study represents a self-selected group of hospitals, mostly university or tertiary-care hospitals, and our patient cohort might not be a true population-based or nationally representative sample of EL patients. Second, while missing data were minimal for the NELA, P-POSSUM, and ACS-NSQIP models, risk calculations from POTTER were not possible for most patients in 2 participating centres, where data necessary for applying this tool are not part of the routine preoperative workup of EL patients. Multiple imputation has been suggested as the preferred approach for handling missing predictor data,[45] but we opted for complete case analysis so that our results come from a pragmatic cohort of patients for whom risks can be readily estimated from preoperative data. Fourth, our analysis showed that recalibrating the models resulted in improved predictive accuracy for all models in our cohort. However, full model revision (re-estimation) from our dataset was not possible, as the selected risk prediction tools are proprietary and have their equations undisclosed. Finally, not all available risk prediction tools could be possibly validated in a single study, and we chose to examine 4 well-known and commonly cited models for which comparative validation studies are lacking for EL patients.

The findings of this study are promising for the use of the ACS-NSQIP model in the Greek healthcare system, demonstrating its broad transportability in a system different from the USA where the model was developed. More comparative validations of different risk prediction tools should be performed at national levels to determine which model might best fit each healthcare setting. Our results imply that combining the subjective assessment of the attending surgeon with a proper tool yields the most accurate prediction, therefore future validations could focus on such combinations. Complete model re-estimation with adjustment of key variables and re-validation in a new sample of patients may generate adjusted versions of existing models, which best fit specific patient populations.

CONCLUSIONS

The ACS-NSQIP tool was most accurate for mortality predictions after EL in a broad external validation cohort, outperforming the surgeon's preoperative risk assessment and the NELA, P-POSSUM and POTTER tools in several comparative prediction metrics and demonstrating utility for facilitating preoperative risk management in the Greek healthcare system. The surgeon's subjective risk assessment may help optimise ACS-NSQIP predictions.

Supplementary digital content (SDC): Supplementary material is available at ***supplement_JTACS_R1*** file.

- Table S5: Checklist for transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)
- Table S1. Inclusion and Exclusion criteria
- Table S2. Adjustment of the ACS-NSQIP calculator according to the surgeon's preoperative assessment

- Table S3. Comparison of the distribution of important variables Hospitals with development data of the NELA and ACS-NSQIP
- Figure S1. Receiver Operating Characteristic (ROC) Curve showing the discriminating performance of the models when predicting 30-day post-operative death in emergency laparotomies
- Figure S2. Calibration of prognostic models when predicting 30-day postoperative death
- Table S4. Predictive performance measures of prognostic models for 30-day postoperative death after updating of calibration intercept and slope
- Figure S3. Calibration of prognostic models when predicting 30-day postoperative death after updating of intercept and slope (recalibration)
- Figure S4. Decision curves showing the net benefit in clinical decision-making of using each prognostic model of 30-day postoperative mortality after updating of intercept and slope (recalibration)
- Figure S5. Forest plot with hospital-specific Brier scores of the Surgeon's assessment prognostic model and overall pooled Brier score based on random-effects meta-analysis
- Figure S6. Forest plot with hospital-specific Brier scores of the NELA prognostic model and overall pooled Brier score based on random-effects meta-analysis
- Figure S7. Forest plot with hospital-specific Brier scores of the P-POSSUM prognostic model and overall pooled Brier score based on random-effects meta-analysis
- Figure S8. Forest plot with hospital-specific Brier scores of the POTTER prognostic model and overall pooled Brier score based on random-effects meta-analysis
- Figure S9. Forest plot with hospital-specific Brier scores of the ACS-NSQIP prognostic model and overall pooled Brier score based on random-effects meta-analysis

- Figure S10. Forest plot with hospital-specific Brier scores of the ACS-NSQIP adjusted prognostic model and overall pooled Brier score based on random-effects meta-analysis

References

1. McLean RC, Brown LR, Baldock TE, O'Loughlin P, McCallum IJ. Evaluating outcomes following emergency laparotomy in the North of England and the impact of the National Emergency Laparotomy Audit – A retrospective cohort study. *Int J Surg*. 2020;77:154-162.
2. Fagan G, Barazanchi A, Coulter G, Leeman M, Hill AG, Eglinton TW. New Zealand and Australia emergency laparotomy mortality rates compare favourably to international outcomes: a systematic review. *ANZ J Surg*. 2021;91(12):2583-2591.
3. Jansson Timan T, Hagberg G, Sernert N, Karlsson O, Prytz M. Mortality following emergency laparotomy: a Swedish cohort study. *BMC Surg*. 2021;21(1):322.
4. Tolstrup M-B, Watt SK, Gögenur I. Morbidity and mortality rates after emergency abdominal surgery: an analysis of 4346 patients scheduled for emergency laparotomy or laparoscopy. *Langenbeck's Arch Surg*. 2017;402(4):615-623.
5. Tan BHL, Mytton J, Al-Khyatt W, Aquina CT, Evison F, Fleming F, et al. A Comparison of Mortality Following Emergency Laparotomy Between Populations From New York State and England. *Ann Surg*. 2017;266(2):280-286.
6. Huddart S, Peden CJ, Swart M, McCormick B, Dickinson M, Mohammed MA, et al. Use of a pathway quality improvement care bundle to reduce mortality after emergency laparotomy. *Br J Surg*. 2014;102(1):57-66.
7. Sivarajah V, Walsh U, Malietzis G, Kontovounisios C, Pandey V, Pellino G. The

importance of discussing mortality risk prior to emergency laparotomy. *Updates Surg*. 2020;72(3):859-865.

8. Mak M, Hakeem A, Chitre V. Pre-NELA vs NELA – has anything changed, or is it just an audit exercise? *Ann R Coll Surg Engl*. 2016;98(8):554-559.

9. Harris EP, MacDonald DB, Boland L, Boet S, Lalu MM, McIsaac DI. Personalized perioperative medicine: a scoping review of personalized assessment and communication of risk before surgery. *Can J Anesth Can d'anesthésie*. 2019;66(9):1026-1037.

10. Smith MTD, Bruce JL, Clarke DL. Using Machine Learning to Establish Predictors of Mortality in Patients Undergoing Laparotomy for Emergency General Surgical Conditions. *World J Surg*. 2022;46(2):339-346.

11. Peden CJ, Aggarwal G, Aitken RJ, Anderson ID, Bang Foss N, Cooper Z, et al. Guidelines for Perioperative Care for Emergency Laparotomy Enhanced Recovery After Surgery (ERAS) Society Recommendations: Part 1-Preoperative: Diagnosis, Rapid Assessment and Optimization. *World J Surg*. 2021;45(5):1272-1290.

12. Thahir A, Pinto-Lopes R, Madenlidou S, Daby L, Halahakoon C. Mortality risk scoring in emergency general surgery: Are we using the best tool? *J Perioper Pract*. 2021;31(4):153-158.

13. Barazanchi A, Bhat S, Palmer-Neels K, Macfater WS, Xia W, Zeng I, et al. Evaluating and improving current risk prediction tools in emergency laparotomy. *J Trauma Acute Care Surg*. 2020;89(2):382-387.

14. Havens JM, Columbus AB, Seshadri AJ, Brown CVR, Tominaga GT, Mowery NT, et al. Risk stratification tools in emergency general surgery. *Trauma Surg acute care open*. 2018;3(1):e000160.

15. Lasithiotakis K, Kritsotakis EI, Kokkinakis S, Petra G, Paterakis K, Karali GA, et al. The Hellenic Emergency Laparotomy Study (HELAS): A Prospective Multicentre Study on the Outcomes of Emergency Laparotomy in Greece. *World J Surg.* Epub 2022 Sep 15

16. Zacharis G, Seretis C. Letter to the Editor: The Hellenic Emergency Laparotomy Study (HELAS): A Prospective Multicentre Study on the Outcomes of Emergency Laparotomy in Greece. *World J Surg.* Epub 2022 Oct 25

17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med*. 2015;162(1):55-63.

18. Pavlou M, Qu C, Omar RZ, Seaman SR, Steyerberg EW, White IR, et al. Estimation of required sample size for external validation of risk models for binary outcomes. *Stat Methods Med Res*. 2021;30(10):2187-2206.

19. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med*. 2016;35(2):214-226.

20. Kattan MW, Gerds TA. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagnostic Progn Res*. 2018;2(1):7.

21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.

22. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.

23. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2015;68(3):279-289.

24. Stevens RJ, Poppe KK. Validation of clinical prediction models: what does the "calibration slope" really measure? *J Clin Epidemiol*. 2020;118:93-99.

25. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med*. 2014;33(3):517-535.

26. Vickers AJ, Holland F. Decision curve analysis to evaluate the clinical benefit of prediction models. *Spine J*. 2021;21(10):1643-1648.

27. Snell KIE, Ensor J, Debray TPA, Moons KGM, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res*. 2018;27(11):3505-3522.

28. Eugene N, Oliver CM, Bassett MG, Poulton TE, Kuryba A, Johston C, et al. Development and internal validation of a novel risk adjustment model for adult patients undergoing emergency laparotomy surgery: the National Emergency Laparotomy Audit risk model. *Br J Anaesth*. 2018;121(4):739-748.

29. Bilimoria KY, Liu Y, Paruch JL, Zhou L, Kmiecik TE, Ko CY, 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(5):833-42.e1-3.

30. 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(4):574-583.

31. Prytherch DR, Whiteley MS, Higgins B, Weaver PC, Prout WG, Powell SJ. POSSUM and Portsmouth POSSUM for predicting mortality. *Br J Surg*. 2003;85(9):1217-1220.

32. Lai CPT, Goo TT, Ong MW, Prakash PS, Lim WW, Drakeford PA. A Comparison of the P-POSSUM and NELA Risk Score for Patients Undergoing Emergency Laparotomy in Singapore. *World J Surg*. 2021;45(8):2439-2446.

33. Hunter Emergency Laparotomy Collaborator Group, Hunter Emergency Laparotomy Collaborator Group. High-Risk Emergency Laparotomy in Australia: Comparing NELA, P-POSSUM, and ACS-NSQIP Calculators. *J Surg Res*. 2020;246:300-304.

34. Parkin CJ, Moritz P, Kirkland O, Glover A. What is the Accuracy of the ACS-NSQIP Surgical Risk Calculator in Emergency Abdominal Surgery? A Meta-Analysis. *J Surg Res*. 2021;268:300-307.

35. Hyder JA, Reznor G, Wakeam E, Nguyen LL, Lipsitz SR, Havens JM. Risk Prediction Accuracy Differs for Emergency Versus Elective Cases in the ACS-NSQIP. *Ann Surg*. 2016;264(6):959-965.

36. Wang X, Hu Y, Zhao B, Su Y. Predictive validity of the ACS-NSQIP surgical risk calculator in geriatric patients undergoing lumbar surgery. *Medicine (Baltimore)*. 2017;96(43):e8416.

37. Wong DJN, Harris S, Sahni A, Bedford JR, Cortes L, Shawyer R, et al. Developing and validating subjective and objective risk-assessment measures for predicting mortality after major surgery: An international prospective cohort study. Menon D, ed. *PLoS Med*. 2020;17(10):e1003253.

38. Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, et al. Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. *BMC Med*. 2018;16(1):120.

39. Groot OQ, Bindels BJJ, Ogink PT, Kapoor ND, Twining PK, Collins AK, et al. Availability and reporting quality of external validations of machine-learning prediction models with orthopedic surgical outcomes: a systematic review. *Acta Orthop*. 2021;92(4):385-393.

40. Oliver CM, Walker E, Giannaris S, Grocott MPW, Moonesinghe SR. Risk assessment tools validated for patients undergoing emergency laparotomy: a systematic review. *Br J Anaesth*. 2015;115(6):849-860.

41. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res.* 2019 Oct 4;3:18.

42. Dadashzadeh ER, Bou-Samra P, Huckaby L V., Nebbia G, Handzel RM, Varley PR, et al. Leveraging Decision Curve Analysis to Improve Clinical Application of Surgical Risk Calculators. *J Surg Res*. 2021;261:58-66.

43. Merkow RP, Hall BL, Cohen ME, Dimick JB, Wang E, Chow WB, et al. Relevance of the C-Statistic When Evaluating Risk-Adjustment Models in Surgery. *J Am Coll Surg*. 2012;214(5):822-830.

44. Cohen ME, Liu Y, Ko CY, Hall BL. An Examination of American College of Surgeons NSQIP Surgical Risk Calculator Accuracy. *J Am Coll Surg.* 2017;224(5):787-795.e1.

45. Hoogland J, van Barreveld M, Debray TPA, Reitsma JB, Verstraelen TE, Dijkgraaf MGW, et al. Handling missing predictor values when validating and applying a prediction model to new patients. *Stat Med*. 2020;39(25):3591-3607.

TABLES

Table 1. Demographics and clinical characteristics in 631 patients undergoing emergency laparotomy in 11 Greek Hospitals

Characteristic	Total (n=631)	Survived (n=528)	Died (n=103)	p-value
Age (years)	66.2 ± 16.7	64.3 ± 16.8	75.8 ± 12.4	<0.001
Male sex	340 (53.9)	287 (54.4)	53 (51.5)	0.59
Body mass index (kg/m2)	26.5 ± 5.3	26.5 ± 5.1	26.6 ± 6.3	0.97
ASA class				<0.001
I	150 (23.8)	145 (27.5)	5 (4.9)	
II	200 (31.7)	188 (35.6)	12 (11.7)	
III	171 (27.1)	141 (26.7)	30 (29.1)	
IV	103 (16.3)	52 (9.8)	51 (49.5)	
V	6 (1.0)	1 (0.2)	5 (4.9)	
Missing	1 (0.2)	1 (0.2)	0 (0.0)	
Preoperative functional status				<0.001
Independent	443 (70.2)	402 (76.1)	41 (39.8)	
Partially dependent	154 (24.4)	105 (19.9)	49 (47.6)	
Fully dependent	32 (5.1)	20 (3.8)	12 (11.7)	
Missing	2 (0.3)	1 (0.2)	1 (1.0)	
Anticipated severity of malignancy				0.19
None	461 (73.1)	394 (74.6)	67 (65.0)	
Primary	78 (12.4)	63 (11.9)	15 (14.6)	
Nodal metastasis	19 (3.0)	14 (2.7)	5 (4.9)	
Distant metastasis	72 (11.4)	56 (10.6)	16 (15.5)	
Missing	1 (0.2)	1 (0.2)	0 (0.0)	
Diabetes mellitus	103 (16.3)	72 (13.6)	31 (30.4)	<0.001
Cardiac comorbidity	264 (42.0)	204 (38.7)	60 (58.8)	<0.001
Chronic steroid use	54 (8.7)	38 (7.3)	16 (15.7)	0.006
Ascites	81 (12.9)	54 (10.2)	27 (26.2)	<0.001
Borderline cardiomegaly chest x-ray	38 (6.0)	33 (6.3)	5 (4.9)	0.59
Respiratory History				0.005
No dyspnoea	569 (90.2)	485 (91.9)	84 (81.6)	
Dyspnoea on exertion or limiting	36 (5.7)	24 (4.5)	12 (11.7)	
Dyspnoea at rest or long-term oxygen	18 (2.9)	13 (2.5)	5 (4.9)	
Missing	8 (1.3)	6 (1.1)	2 (1.9)	
Smoking	185 (29.3)	165 (31.3)	20 (19.4)	0.016
Haemodialysis or CVVH	10 (1.6)	6 (1.1)	4 (3.9)	0.042
Preoperative acute renal failure	77 (12.2)	50 (9.5)	27 (26.5)	<0.001
Sepsis within 48h prior to surgery				<0.001
None	308 (48.8)	284 (53.8)	24 (23.3)	
Two SIRS criteria	218 (34.5)	186 (35.2)	32 (31.1)	

Severe sepsis	80 (12.7)	48 (9.1)	32 (31.1)	
Septic shock	25 (4.0)	10 (1.9)	15 (14.6)	
Preoperative diagnosis				0.032
Perforation	225 (35.7)	187 (35.4)	38 (36.9)	
Obstruction	247 (39.1)	218 (41.3)	29 (28.2)	
Ischemia	94 (14.9)	74 (14.0)	20 (19.4)	
Other	65 (10.3)	49 (9.3)	16 (15.5)	
Operation type				0.080
Adhesiolysis	75 (11.9)	72 (13.6)	3 (2.9)	
Small bowel resection	130 (20.6)	106 (20.1)	24 (23.3)	
Colectomy right	58 (9.2)	48 (9.1)	10 (9.7)	
Hartmann's procedure	73 (11.6)	59 (11.2)	14 (13.6)	
Strangulated hernia with bowel resection	38 (6.0)	35 (6.6)	3 (2.9)	
Peptic ulcer repair	75 (11.9)	63 (11.9)	12 (11.7)	
Colectomy other	50 (7.9)	38 (7.2)	12 (11.7)	
Stoma formation	41 (6.5)	33 (6.3)	8 (7.8)	
Other	91 (14.4)	74 (14.0)	17 (16.5)	

Data are presented as mean ± SD for continuous measures, and n (%) for categorical measures.

ASA, American Society of Anesthesiologists; SIRS, systemic inflammatory response syndrome; CVVH, continuous veno-venous hemofiltration.

Table 2. Distributions of mortality risk predictions in patients undergoing emergency laparotomy in 11 Greek Hospitals

30-day mortality predictions	**Total (n=631)**	**Survived (n=528)**	**Died (n=103)**	**p-value**
Surgeon's preoperative assessment, n (%)				<0.001
<5%	172 (27.3)	166 (31.4)	6 (5.8)	
5 - 10%	220 (34.9)	205 (38.8)	15 (14.6)	
11 - 20%	137 (21.7)	94 (17.8)	43 (41.7)	
>20%	81 (12.8)	45 (8.5)	36 (35.0)	
Unable to assess	21 (3.3)	18 (3.4)	3 (2.9)	
Mean predicted risk ± SD, %				
Surgeon *	14.9 ± 18.3	12.0 ± 15.6	29.5 ± 23.3	<0.001
NELA	10.5 ± 15.2	7.6 ± 12.2	25.8 ± 19.3	<0.001
P-POSSUM	19.9 ± 25.1	16.0 ± 22.3	40.3 ± 28.4	<0.001
POTTER	8.9 ± 11.4	6.7 ± 9.3	21.0 ± 14.4	<0.001
ACS-NSQIP	12.2 ± 17.6	8.4 ± 13.6	31.6 ± 22.3	<0.001
ACS-NSQIP surgeon-adjusted	18.1 ± 16.8	14.8 ± 14.0	35.5 ± 19.0	<0.001

* Point estimates of risk prediction provided by clinicians were taken as the midpoint of the predicted risk intervals (i.e., 2.5% for the interval <5%, 7.5% for the interval 5%-10%, and so on).

Table 3. Predictive performance measures of prognostic models for 30-day postoperative death

Prognostic model	**N**	**Overall fit**	**Discrimination**		**Calibration**				**Clinical utility**		
		Brier scaled % (95% CI)	AUC (95% CI)	DeLong p-value	E:O ratio	CITL (95% CI)	Slope (95% CI)	HL-GOF p-value	NB, 5%	NB, 10%	NB, 20%
Surgeon	610	10.6 (1.3, 18.7)	0.79 (0.75, 0.82)	Ref.	0.91	0.16 (-0.09, 0.41)	0.74 (0.57, 0.91)	<0.001	0.12	0.10	0.04
NELA	623	17.2 (10.2, 23.9)	0.85 (0.82, 0.88)	0.005	0.65	0.68 (0.43, 0.93)	0.86 (0.68, 1.05)	<0.001	0.13	0.11	0.07
P-POSSUM	622	1.5 (-13.1, 13.1)	0.79 (0.75, 0.82)	0.868	1.23	-0.41 (-0.68, -0.13)	0.48 (0.37, 0.60)	<0.001	0.13	0.10	0.06
POTTER	486	15.4 (9.3, 21.8)	0.84 (0.81, 0.87)	0.081	0.55	0.75 (0.47, 1.03)	0.98 (0.73, 1.23)	<0.001	0.12	0.09	0.05
ACS-NSQIP	618	22.4 (14.5, 30.3)	0.84 (0.81, 0.87)	0.030	0.75	0.49 (0.23, 0.75)	0.75 (0.59, 0.91)	<0.001	0.12	0.10	0.07
ACS-NSQIP adjusted	618	20.6 (13.4, 27.2)	0.83 (0.79, 0.86)	0.057	1.12	-0.16 (-0.39, 0.08)	1.12 (0.85, 1.40)	0.742	0.13	0.10	0.07

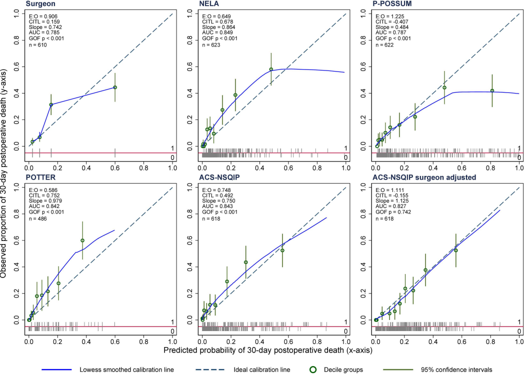
N, number of patients in the analysis; CI, confidence interval; AUC, area under the curve; E:O, ratio of expected and observed events; CITL, calibration-in-the-large; Slope, calibration slope; HL-GOF, Hosmer-Lemeshow goodness-of-fit test; NB, net benefit (calculated at decision thresholds 5%, 10% and 20%).

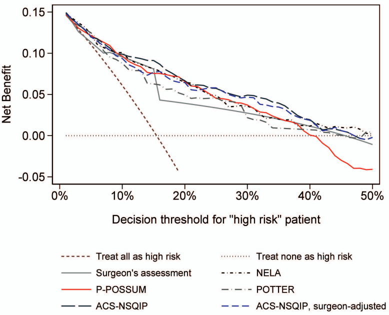
Figure legends

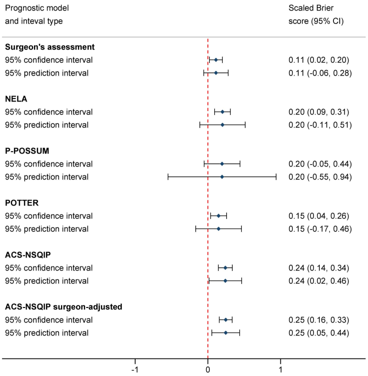
Figure 1. Calibration of prognostic models when predicting 30-day postoperative death. The blue line is a smoothed locally weighted regression (lowess) line that shows the agreement between predicted probabilities and observed proportions of 30-day mortality. The dashed diagonal line indicates perfect calibration. The circled points represent mean risks in decile groups of predicted probabilities, with vertical lines representing 95% confidence intervals. The spike plot on the x-axis summarises the density of patients in the range of predicted risks of 30-day death. E:O, ratio of expected and observed deaths; CITL, calibration-in-the-large; Slope, calibration slope; AUC, area under the curve; GOF goodness-of-fit.

Figure 2. Decision curves showing the net benefit in clinical decision-making of using each prognostic model of 30-day postoperative mortality.

Figure 3. Summary forest plot with overall scaled Brier score for each prognostic model based on the results of random-effects meta-analysis of hospital-specific data. The confidence interval quantifies the precision in estimating the average Brier score in this study whereas the prediction interval quantifies the dispersion of the Brier score value in future single-center studies by accounting for between-hospitals heterogeneity.


Surgeon
E:O = 0.906
CITL = 0.159
Slope = 0.742
AUC = 0.785
GOF p < 0.001
n = 610
NELA
E:O = 0.649
CITL = 0.678
Slope = 0.864
AUC = 0.849
GOF p < 0.001
n = 623
P-POSSUM
E:O = 1.225
CITL = -0.407
Slope = 0.484
AUC = 0.787
GOF p < 0.001
n = 622
POTTER
E:O = 0.586
CITL = 0.752
Slope = 0.979
AUC = 0.842
GOF p < 0.001
n = 486
ACS-NSQIP
E:O = 0.748
CITL = 0.492
Slope = 0.750
AUC = 0.843
GOF p < 0.001
n = 618
ACS-NSQIP surgeon adjusted
E:O = 1.111
CITL = -0.155
Slope = 1.125
AUC = 0.827
GOF p = 0.742
n = 618
Observed proportion of 30-day postoperative death (y-axis)
Predicted probability of 30-day postoperative death (x-axis)
0.0
0.2
0.4
0.6
0.8
1.0
1
0
Lowess smoothed calibration line
Ideal calibration line
Decile groups
95% confidence intervals


Net Benefit
0.15
0.10
0.05
0.00
-0.05
0%
10%
20%
30%
40%
50%
Decision threshold for "high risk" patient
Treat all as high risk
Treat none as high risk
Surgeon's assessment
NELA
P-POSSUM
POTTER
ACS-NSQIP
ACS-NSQIP, surgeon-adjusted


Prognostic model
and inteval type
Scaled Brier
score (95% CI)
Surgeon's assessment
95% confidence interval
0.11 (0.02, 0.20)
95% prediction interval
0.11 (-0.06, 0.28)
NELA
95% confidence interval
0.20 (0.09, 0.31)
95% prediction interval
0.20 (-0.11, 0.51)
P-POSSUM
95% confidence interval
0.20 (-0.05, 0.44)
95% prediction interval
0.20 (-0.55, 0.94)
POTTER
95% confidence interval
0.15 (0.04, 0.26)
95% prediction interval
0.15 (-0.17, 0.46)
ACS-NSQIP
95% confidence interval
0.24 (0.14, 0.34)
95% prediction interval
0.24 (0.02, 0.46)
ACS-NSQIP surgeon-adjusted
95% confidence interval
0.25 (0.16, 0.33)
95% prediction interval
0.25 (0.05, 0.44)
-1
0
1

SUPPLEMENTARY MATERIAL

Prospective multicentre external validation of postoperative mortality prediction tools in patients undergoing emergency laparotomy

Table S1: Inclusion and Exclusion criteria

Inclusion criteria:

- Age >18yrs
- Emergency laparotomy (operation simultaneously with resuscitation usually within one hour) or
- urgent (operation as soon as possible after resuscitation, within 24hrs)
- Operation in the gastrointestinal tract:
  - Open or laparoscopic, or laparoscopically assisted procedures.
  - Procedures involving the stomach, small or large bowel, or rectum for conditions such as
  - perforation, ischaemia, abdominal abscess, bleeding or obstruction
  - Wash out/evacuation of intraperitoneal abscess or haematoma
  - Bowel resection/repair due to incarcerated/incisional hernias
  - Bowel resection or repair due to incarcerated umbilical, inguinal or femoral hernias
  - Open or laparoscopic adhesiolysis
  - Laparotomy/laparoscopy with inoperable pathology
  - Return to theatre for repair of a substantial dehiscence of major abdominal wound (i.e.“burst abdomen”)
  - Return to theatre after any operation (including vascular, gynaecology, urology, cardiac) meeting the criteria above
  - In the case of multiple procedures in the abdominopelvic cavity the patient is included if the main procedure is a general surgical one (i.e. if bowel resection happens during an open aneurysm repair it should not be included)
  - Any intra-abdominal procedure not identifiable within exclusion criteria should be included.

Exclusion criteria:

- Patients under 18
- Elective operation
- Diagnostic laparoscopy or laparotomy where no other procedure is performed (NB, if no procedure is performed due to inoperable pathology, then include)
- Appendicectomy with or without drainage of localized abscess
- Cholecystectomy with or without drainage of localized abscess
- Hernia repair without bowel resection
- Minor abdominal wound revision
- Vascular surgery
- Gynaecological surgery – c-section – ruptured ectopic pregnancy
- Surgery relating to organ transplantation

Table S2: Adjustment of the ACS-NSQIP calculator according to the surgeon's preoperative assessment

ACS-NSQIP predicted probability of 30-day mortality	Surgeon's preoperative assessment of mortality risk	Chosen option of adjustment (for underestimation) on the ACS-NSQIP online calculation
<0.05	<5%	1 – No adjustment necessary
<0.05	5 - 10%	2 – Risk somewhat higher then estimate
<0.05	11 - 20%	3 – Risk significantly higher than estimate
<0.05	>20%	3 – Risk significantly higher than estimate
0.05 – 0.10	<5%	1 – No adjustment necessary
0.05 – 0.10	5 - 10%	1 – No adjustment necessary
0.05 – 0.10	11 - 20%	2 – Risk somewhat higher then estimate
0.05 – 0.10	>20%	3 – Risk significantly higher than estimate
0.10 – 0.20	<5%	1 – No adjustment necessary
0.10 – 0.20	5 - 10%	1 – No adjustment necessary
0.10 – 0.20	11 - 20%	1 – No adjustment necessary
0.10 – 0.20	>20%	2 – Risk somewhat higher then estimate
>0.20	Any	1 – No adjustment necessary

Table S3. Comparison of the distribution of important variables of 631 patients undergoing emergency laparotomy in 11 Greek Hospitals with development data of the NELA and ACS-NSQIP risk prediction models.

Characteristic	Greek cohort, n(%)	NELA, n(%)	ACS-NSQIP, n(%)
Age group			
≤60	201 (32)	13121 (34)	-
>60	430 (68)	25709 (66)	-
Male Sex	341 (54)	18 740 (48)	604016 (43)
ASA class			
1-2	351 (56)	17190 (44)	777115 (55)
3	171 (27)	13706 (35)	541404 (38)
4-5	110 (17)	7934 (21)	95487 (7)
Steroid use (1)	55 (9)	-	43296 (3)
Ascites (2)	81 (13)	-	8697 (1)
Preoperative functional status			
Independent	444 (70)	-	1344929 (95)
Partially dependent	154 (25)	-	52500 (4)
Totally dependent	33 (5)	-	16577 (1)
Sepsis (3)			
SIRS	218 (34)	-	55090 (4)
Sepsis	81 (13)	-	33725 (2)
Septic shock	25 (4)	-	8546 (1)
Ventilator dependent	16 (3)	-	10119 (1)
Smoking (4)	186 (29)	-	272322 (19)
Dyspnoea			
Moderate exertion	36 (6)	6210 (16)	110720 (8)
At rest	19 (3)	4632 (12)	15571 (1)
Haemodialysis or CVVH	10 (2)	-	22829 (2)
Acute renal failure	77 (12)	-	7103 (1)
Anticipated severity of malignancy			
None	462 (73)	29774 (77)	-
Primary	78 (12)	4496 (12)	-
Nodal metastasis	19 (3)	1655 (4)	-
Distant metastasis	73 (12)	2905 (7)	28173 (2) (5)
Diabetes mellitus	103 (16)	-	215180 (15)
Urgency of operation			
Expedited (>18 hours)	71 (11)	6405 (17)	-
Urgent (6-18 hours)	179 (28)	11735 (30)	-
Urgent (2-6 hours)	206 (33)	15051 (39)	-
Immediate (<2 hours)	177 (28)	5639 (14)	-
Operative severity			
Major	388 (62)	24453 (63)	-
Major+	243 (38)	14377 (37)	-
Peritoneal soiling			
None	171 (27)	14537 (37)	-

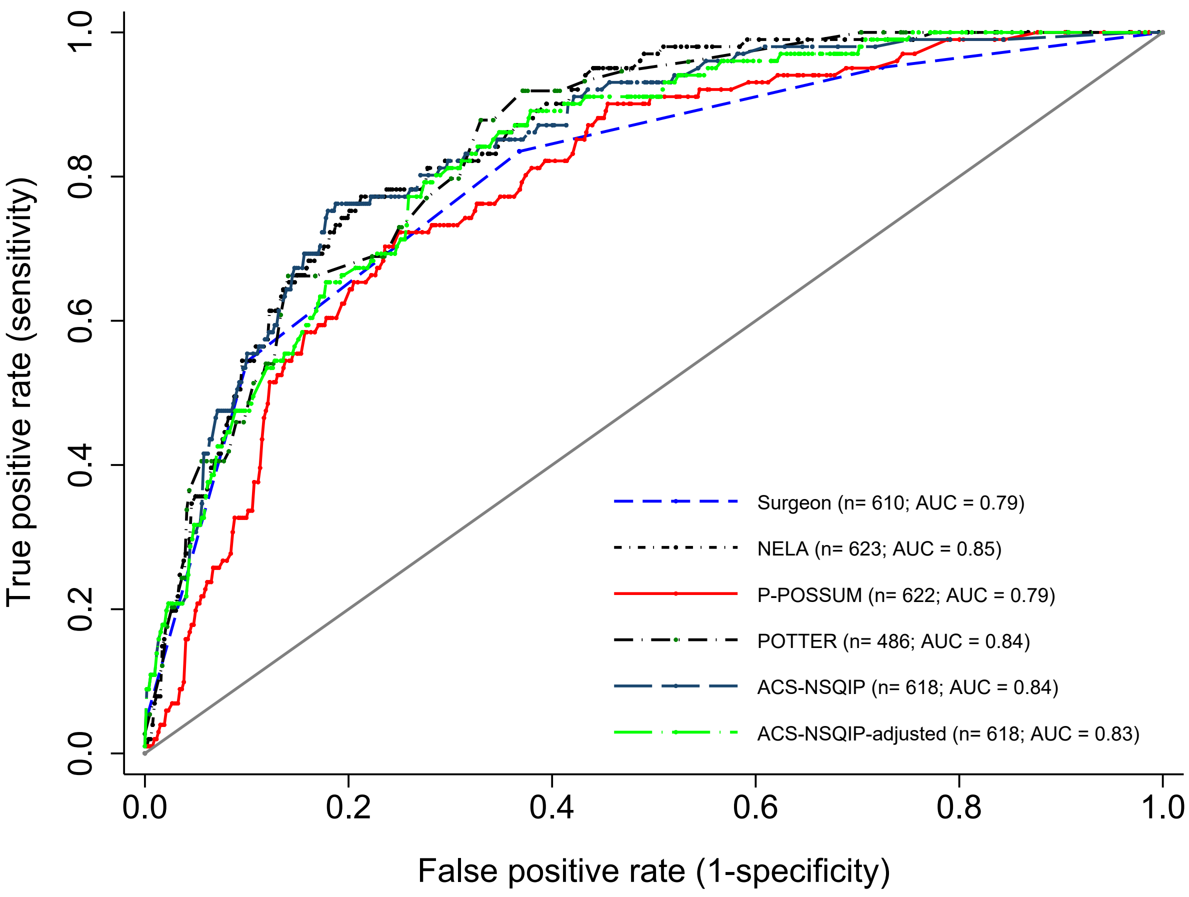
Serous fluid	234 (37)	9992 (26)	-
Localised pus	54 (9)	4183 (11)	-
Free bowel content, pus, or blood	172 (27)	10118 (26)	-
Intraoperative blood loss			
<100ml	357 (56)	18380 (47)	-
101-500ml	237 (38)	17463 (45)	-
501-999 ml	25 (4)	2001 (5)	-
>1000ml	12 (2)	986 (3)	-
30-day mortality	103 (16)	4458 (12)	18,909 (1)

NA, not applicable; ASA, American Society of Anesthesiologists; CVVH, continuous veno-venous hemofiltration; BMI, body mass index. Dashes (-) imply that data for these variables were not available in the original model development publications.

Notes.

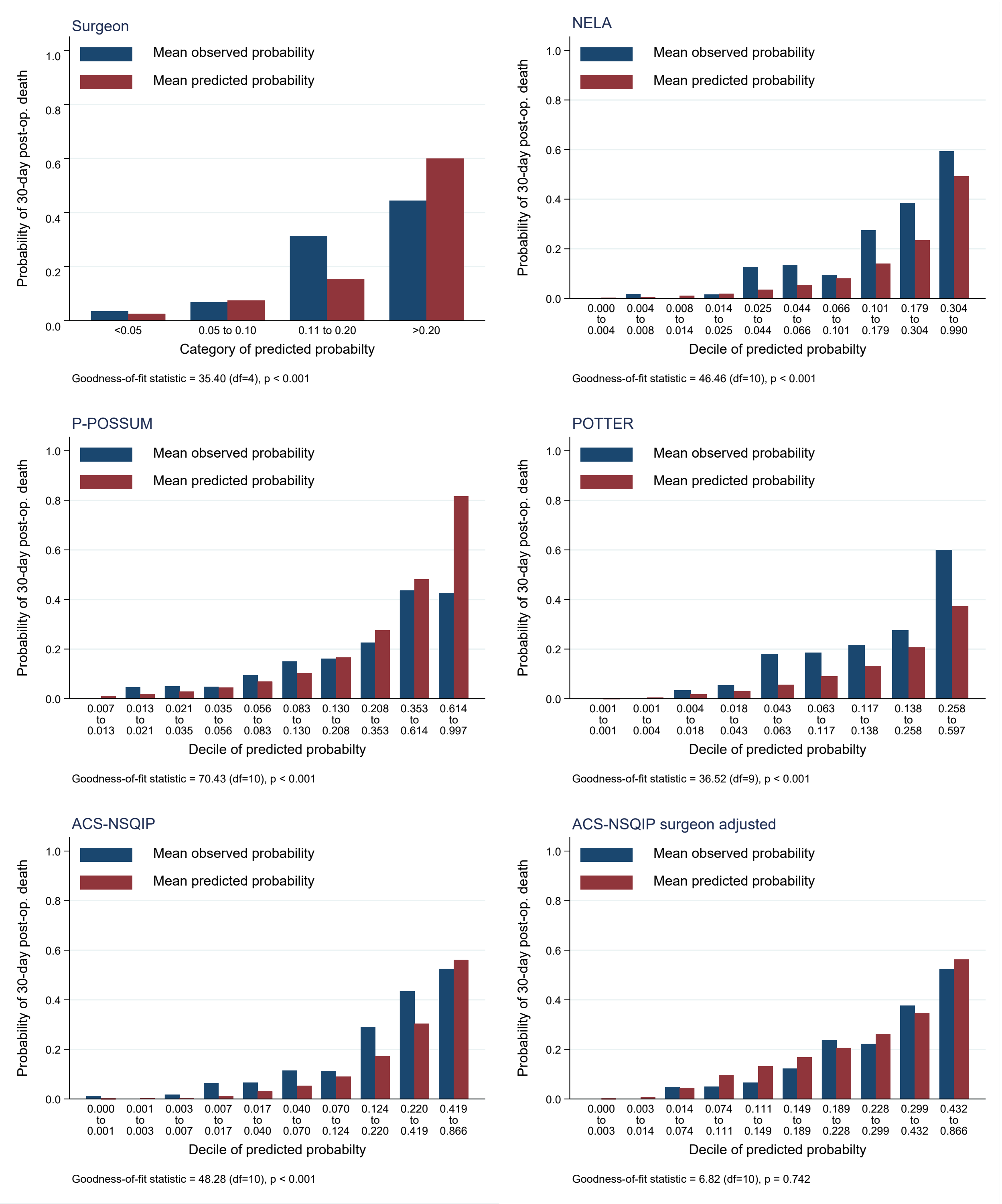
(1) Steroid use for chronic condition
(2) Ascites within 30 days prior to surgery
(3) Systemic sepsis within 48h from surgery
(4) Smoking within 12 months from surgery
(5) Disseminated cancer

Figure S1. Receiver Operating Characteristic (ROC) Curve showing the discriminating performance of the models when predicting 30-day post-operative death in emergency laparotomies



<u>Note</u>: For NELA, P-POSSUM, POTTER, ACS-NSQIP and ACS-NSQIP-adjusted, the ROC curves were plotted by calculating the sensitivity and specificity for all values ranging from 0 to 1, to construct a curve, and for the surgeon's prediction, they were calculated for each of the four categories, and the points were combined to form the curve.

Figure S2. Calibration of prognostic models when predicting 30-day postoperative death.



Surgeon

Goodness-of-fit statistic = 35.40 (df=4), p < 0.001

NELA

Goodness-of-fit statistic = 46.46 (df=10), p < 0.001

P-POSSUM

Goodness-of-fit statistic = 70.43 (df=10), p < 0.001

POTTER

Goodness-of-fit statistic = 36.52 (df=9), p < 0.001

ACS-NSQIP

Goodness-of-fit statistic = 48.28 (df=10), p < 0.001

ACS-NSQIP surgeon adjusted

Goodness-of-fit statistic = 6.82 (df=10), p = 0.742

Table S4. Predictive performance measures of prognostic models for 30-day postoperative death after updating of calibration intercept and slope

Prognostic model	**N**	**Overall fit**	**Discrimination**		**Calibration**				**Clinical utility**		
		Brier scaled % (95% CI)	AUC (95% CI)	DeLong p-value	E:O ratio	CITL (95% CI)	Slope (95% CI)	HL-GOF p-value	NB, 5%	NB, 10%	NB, 20%
Surgeon	610	10.6 (1.3, 18.7)	0.79 (0.75, 0.82)	Ref.	0.91	0.16 (-0.09, 0.41)	0.74 (0.57, 0.91)	<0.001	0.12	0.10	0.04
NELA RCM	623	22.2 (13.4, 29.5)	0.85 (0.82, 0.88)	0.005	1.00	-0.00 (-0.24, 0.24)	1.00 (0.79, 1.21)	0.560	0.13	0.11	0.08
P-POSSUM RCM	622	10.9 (4.5, 16.8)	0.79 (0.75, 0.82)	0.868	1.00	-0.00 (-0.23, 0.23)	1.00 (0.75, 1.25)	0.199	0.12	0.10	0.06
POTTER RCM	486	20.9 (10.9, 29.1)	0.84 (0.81, 0.87)	0.081	0.93	-0.00 (-0.28, 0.28)	1.00 (0.74, 1.26)	0.853	0.12	0.10	0.06
ACS-NSQIP RCM	618	24.3 (16.4, 31.6)	0.84 (0.81, 0.87)	0.030	1.01	0.00 (-0.24, 0.24)	1.00 (0.79, 1.21)	0.806	0.13	0.10	0.08

Note. RCM, re-calibrated model; N, number of patients in the analysis; CI, confidence interval; AUC, area under the curve; E:O, ratio of expected and observed events; CITL, calibration-in-the-large; Slope, calibration slope; HL-GOF, Hosmer-Lemeshow goodness-of-fit test; NB, net benefit (calculated at decision thresholds 5%, 10% and 20%).

Figure S3. Calibration of prognostic models when predicting 30-day postoperative death after updating of intercept and slope (recalibration). The blue line is a smoothed locally weighted regression (lowess) line that shows the agreement between predicted probabilities and observed proportions of 30-day mortality. The dashed diagonal line indicates perfect calibration. The circled points represent mean risks in decile groups of predicted probabilities, with vertical lines representing 95% confidence intervals. The spike plot on the x-axis summarises the density of patients in the range of predicted risks of 30-day death. RCM, re-calibrated model; E:O, ratio of expected and observed deaths; CITL, calibration-in-the-large; Slope, calibration slope; AUC, area under the curve; GOF goodness-of-fit.

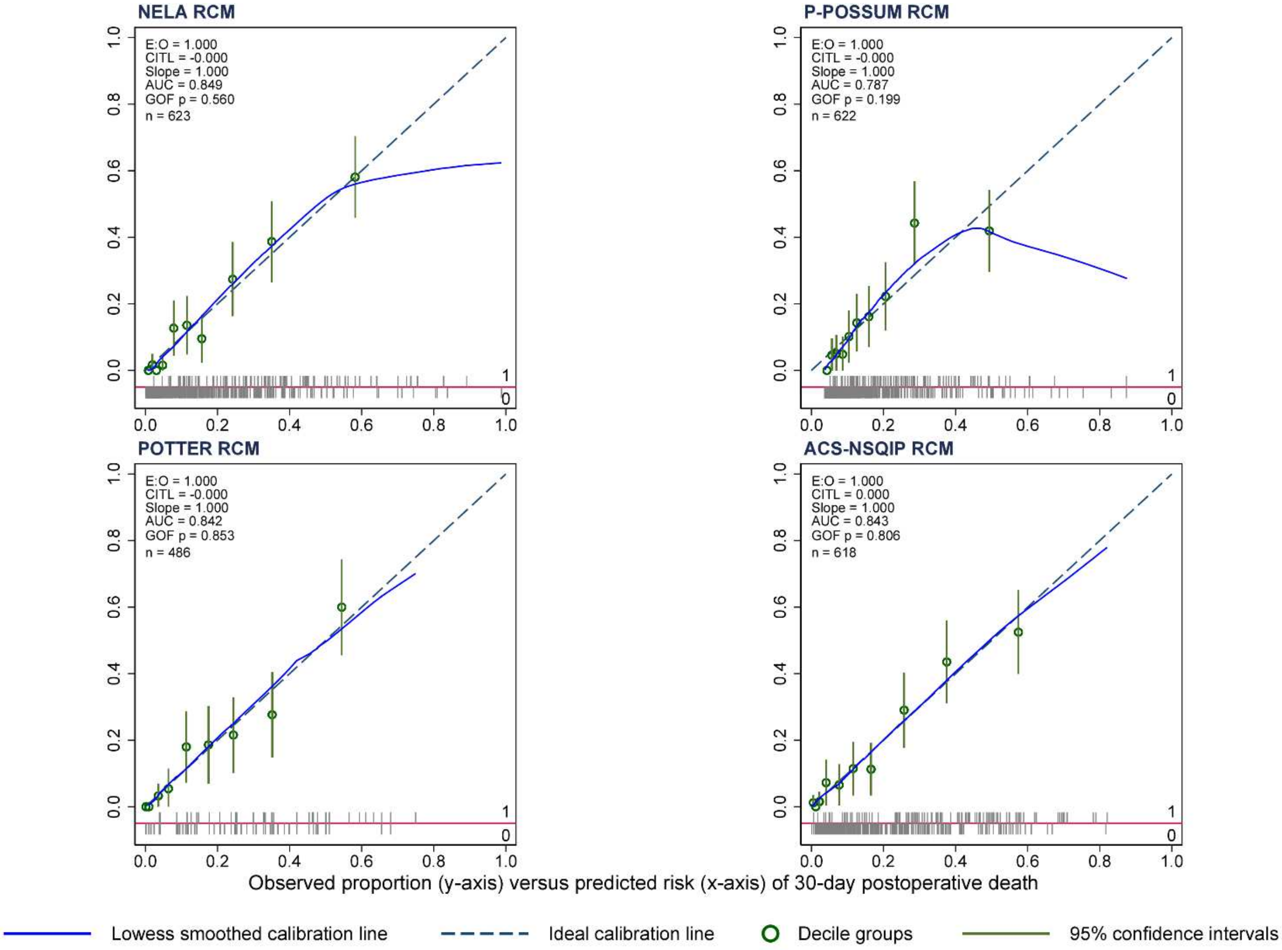


Figure S4. Decision curves showing the net benefit in clinical decision-making of using each prognostic model of 30-day postoperative mortality after updating of intercept and slope (recalibration).

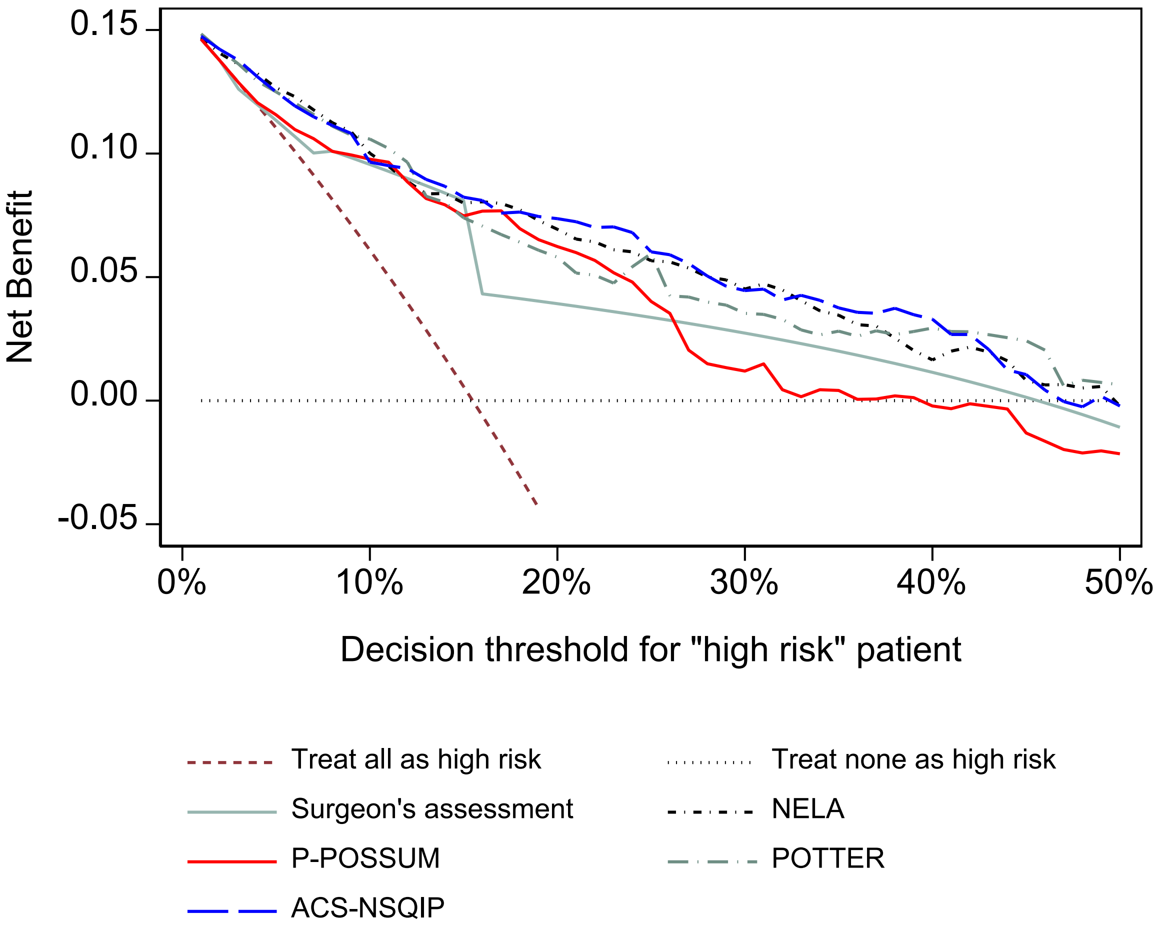
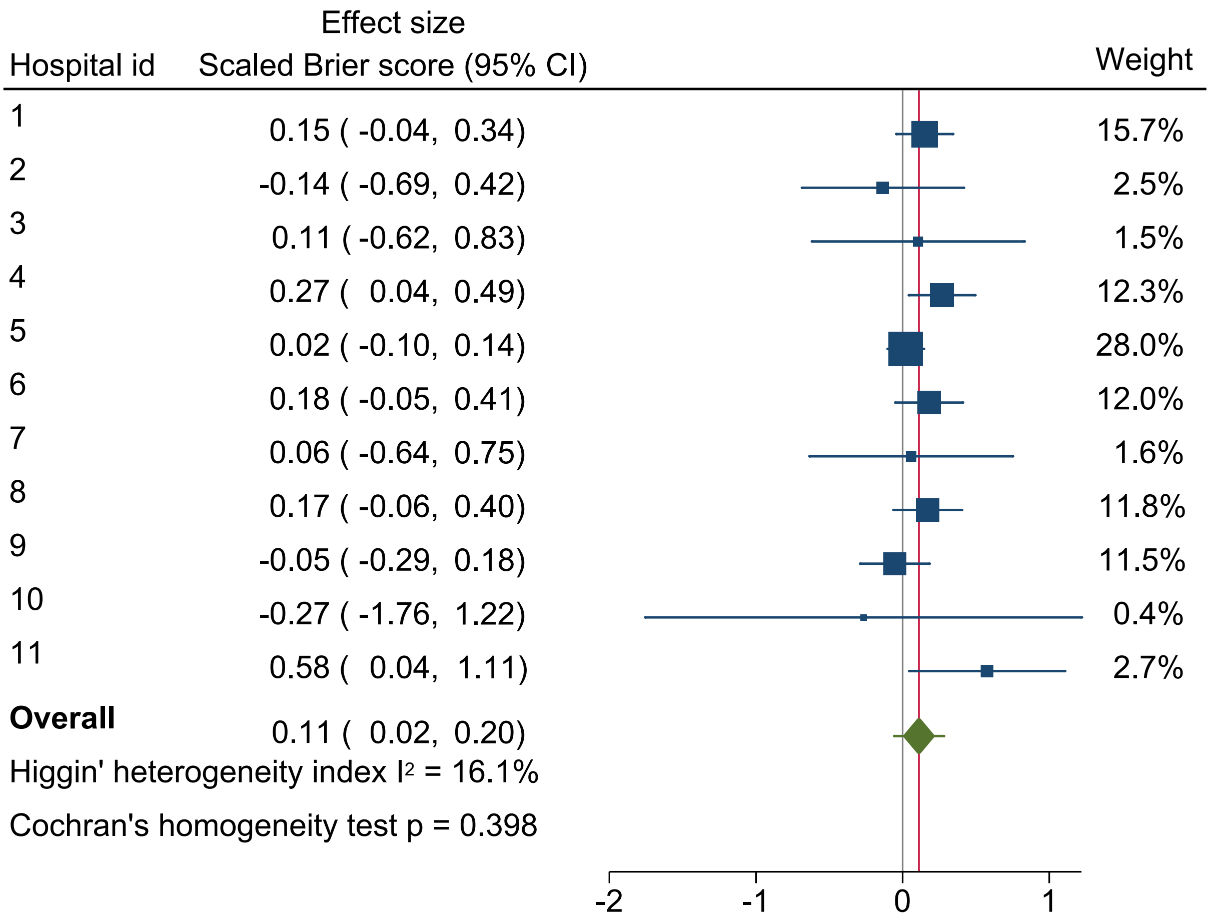


Figure S5. Forest plot with hospital-specific Brier scores of the Surgeon's assessment prognostic model and overall pooled Brier score based on random-effects meta-analysis.



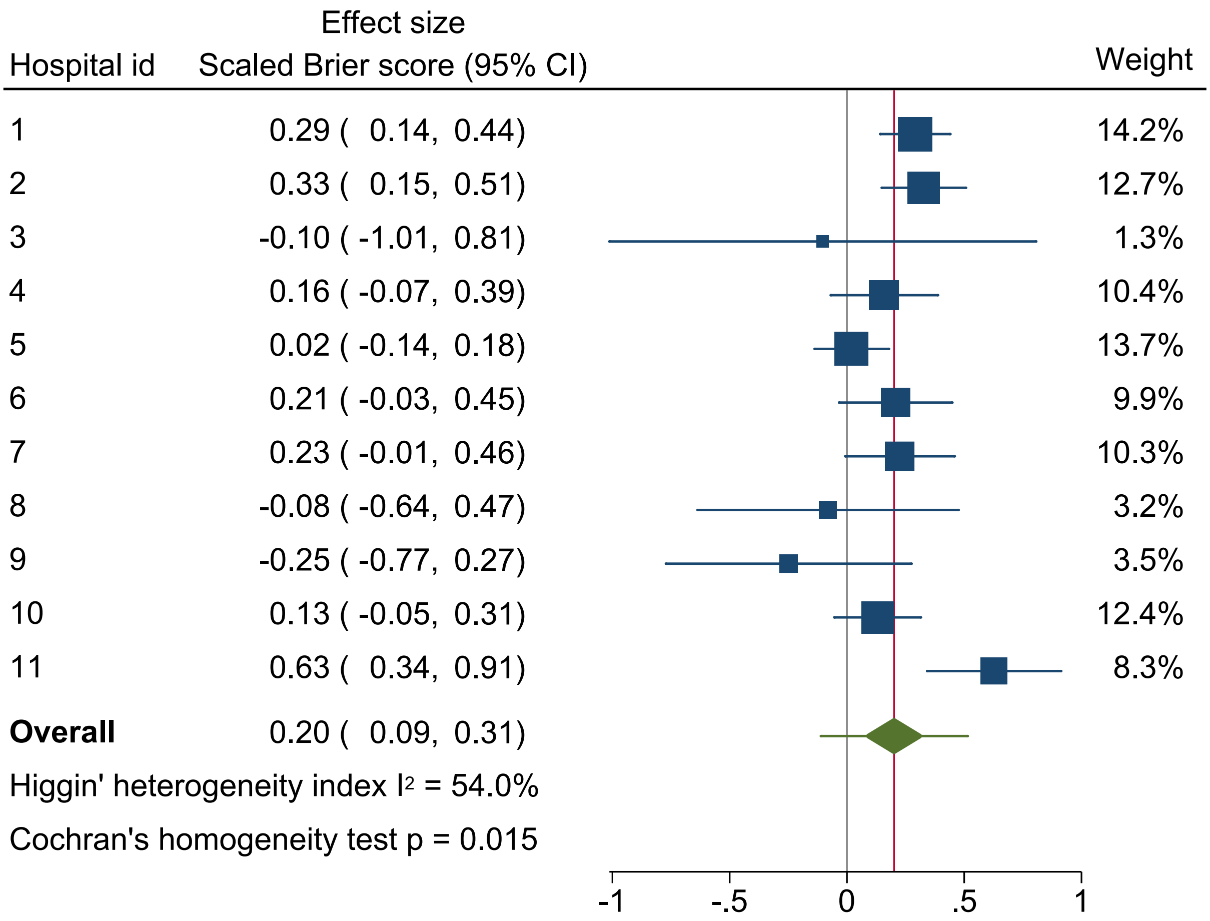
Hospital id	Effect size Scaled Brier score (95% CI)	Weight
1	0.15 (-0.04, 0.34)	15.7%
2	-0.14 (-0.69, 0.42)	2.5%
3	0.11 (-0.62, 0.83)	1.5%
4	0.27 (0.04, 0.49)	12.3%
5	0.02 (-0.10, 0.14)	28.0%
6	0.18 (-0.05, 0.41)	12.0%
7	0.06 (-0.64, 0.75)	1.6%
8	0.17 (-0.06, 0.40)	11.8%
9	-0.05 (-0.29, 0.18)	11.5%
10	-0.27 (-1.76, 1.22)	0.4%
11	0.58 (0.04, 1.11)	2.7%
Overall	0.11 (0.02, 0.20)	

Higgin' heterogeneity index I^2 = 16.1%

Cochran's homogeneity test p = 0.398

Note. The overall meta-analyzed scaled Brier score is represented by the diamond centered on its estimated value with the diamond width corresponding to the length of the confidence interval. The green whiskers extending from the overall diamond represent the prediction interval, which provides a plausible range for the scaled Brier score in a future, new study.

Figure S6. Forest plot with hospital-specific Brier scores of the NELA prognostic model and overall pooled Brier score based on random-effects meta-analysis.



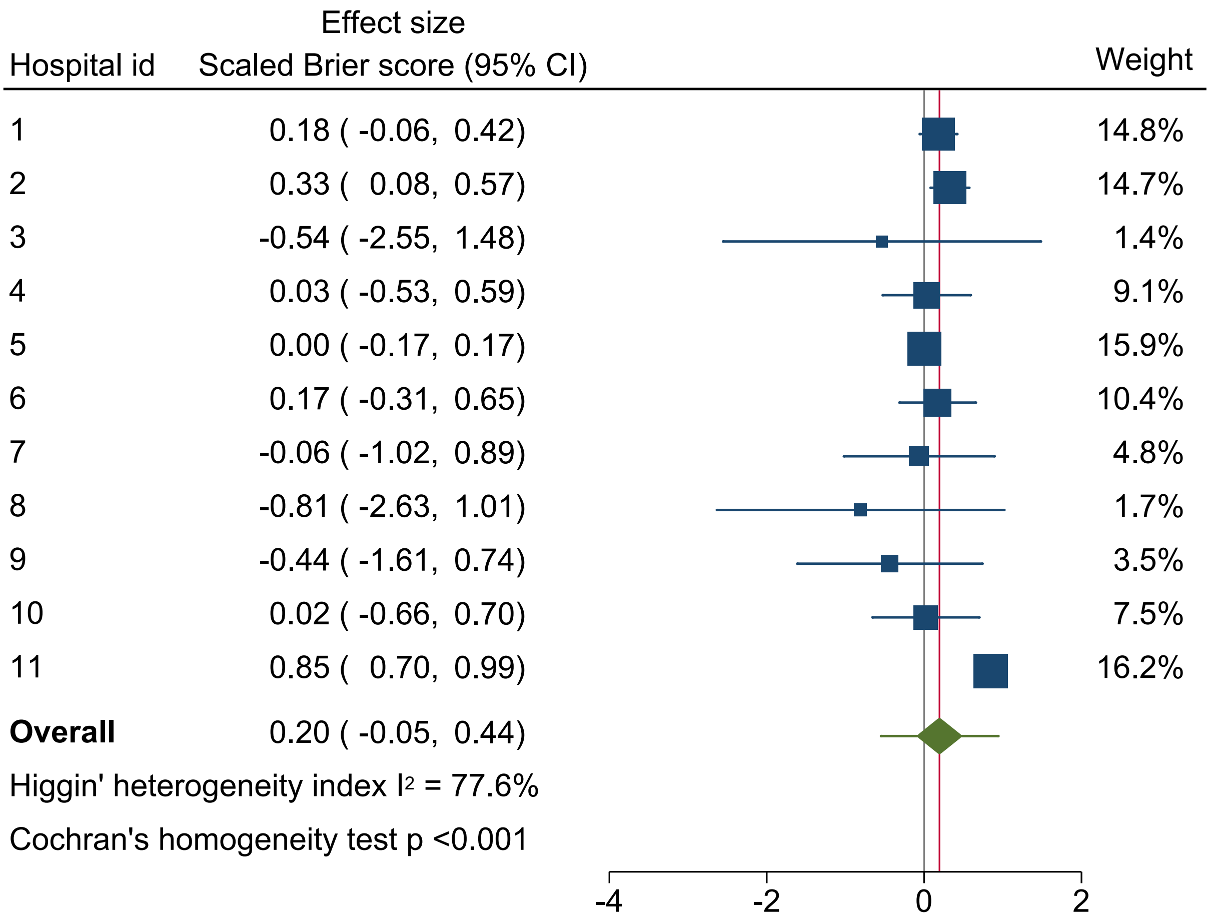
Hospital id	Effect size Scaled Brier score (95% CI)	Weight
1	0.29 (0.14, 0.44)	14.2%
2	0.33 (0.15, 0.51)	12.7%
3	-0.10 (-1.01, 0.81)	1.3%
4	0.16 (-0.07, 0.39)	10.4%
5	0.02 (-0.14, 0.18)	13.7%
6	0.21 (-0.03, 0.45)	9.9%
7	0.23 (-0.01, 0.46)	10.3%
8	-0.08 (-0.64, 0.47)	3.2%
9	-0.25 (-0.77, 0.27)	3.5%
10	0.13 (-0.05, 0.31)	12.4%
11	0.63 (0.34, 0.91)	8.3%
Overall	0.20 (0.09, 0.31)	

Higgin' heterogeneity index I^2 = 54.0%

Cochran's homogeneity test p = 0.015

Note. The overall meta-analyzed scaled Brier score is represented by the diamond centered on its estimated value with the diamond width corresponding to the length of the confidence interval. The green whiskers extending from the overall diamond represent the prediction interval, which provides a plausible range for the scaled Brier score in a future, new study.

Figure S7. Forest plot with hospital-specific Brier scores of the P-POSSUM prognostic model and overall pooled Brier score based on random-effects meta-analysis.



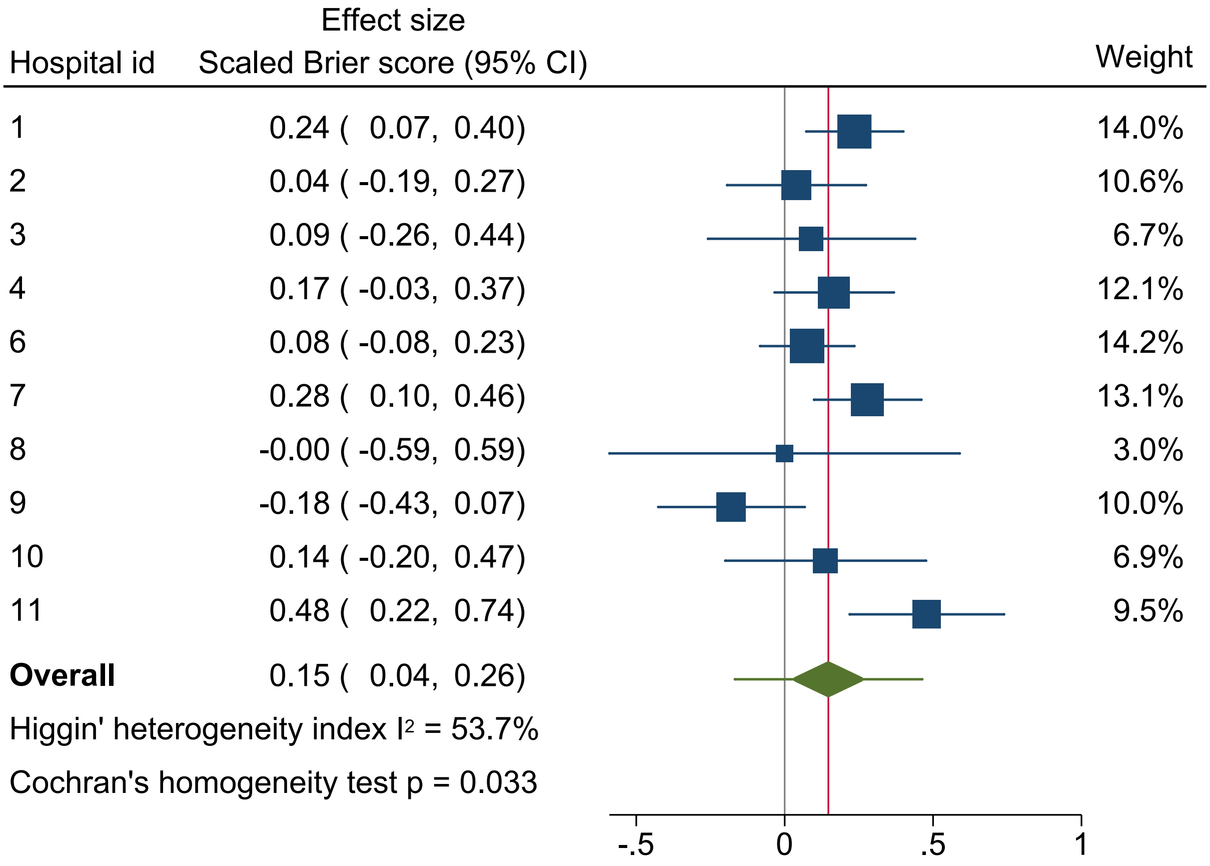
Hospital id	Effect size Scaled Brier score (95% CI)	Weight
1	0.18 (-0.06, 0.42)	14.8%
2	0.33 (0.08, 0.57)	14.7%
3	-0.54 (-2.55, 1.48)	1.4%
4	0.03 (-0.53, 0.59)	9.1%
5	0.00 (-0.17, 0.17)	15.9%
6	0.17 (-0.31, 0.65)	10.4%
7	-0.06 (-1.02, 0.89)	4.8%
8	-0.81 (-2.63, 1.01)	1.7%
9	-0.44 (-1.61, 0.74)	3.5%
10	0.02 (-0.66, 0.70)	7.5%
11	0.85 (0.70, 0.99)	16.2%
Overall	0.20 (-0.05, 0.44)	

Higgin' heterogeneity index I^2 = 77.6%

Cochran's homogeneity test p <0.001

Note. The overall meta-analyzed scaled Brier score is represented by the diamond centered on its estimated value with the diamond width corresponding to the length of the confidence interval. The green whiskers extending from the overall diamond represent the prediction interval, which provides a plausible range for the scaled Brier score in a future, new study.

Figure S8. Forest plot with hospital-specific Brier scores of the POTTER prognostic model and overall pooled Brier score based on random-effects meta-analysis.



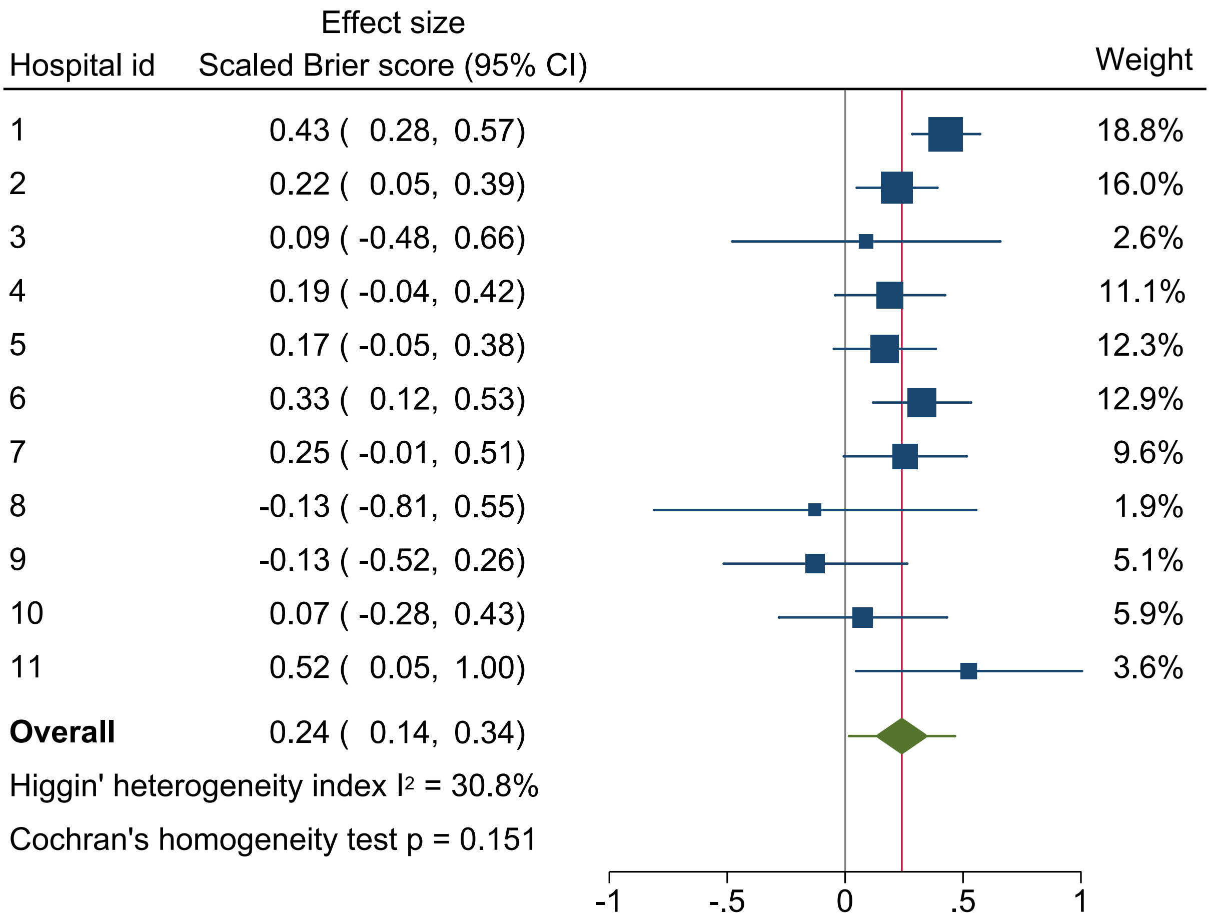
Hospital id	Effect size Scaled Brier score (95% CI)	Weight
1	0.24 (0.07, 0.40)	14.0%
2	0.04 (-0.19, 0.27)	10.6%
3	0.09 (-0.26, 0.44)	6.7%
4	0.17 (-0.03, 0.37)	12.1%
6	0.08 (-0.08, 0.23)	14.2%
7	0.28 (0.10, 0.46)	13.1%
8	-0.00 (-0.59, 0.59)	3.0%
9	-0.18 (-0.43, 0.07)	10.0%
10	0.14 (-0.20, 0.47)	6.9%
11	0.48 (0.22, 0.74)	9.5%
Overall	0.15 (0.04, 0.26)	

Higgin' heterogeneity index I^2 = 53.7%

Cochran's homogeneity test p = 0.033

Note. The overall meta-analyzed scaled Brier score is represented by the diamond centered on its estimated value with the diamond width corresponding to the length of the confidence interval. The green whiskers extending from the overall diamond represent the prediction interval, which provides a plausible range for the scaled Brier score in a future, new study.

Figure S9. Forest plot with hospital-specific Brier scores of the ACS-NSQIP prognostic model and overall pooled Brier score based on random-effects meta-analysis.



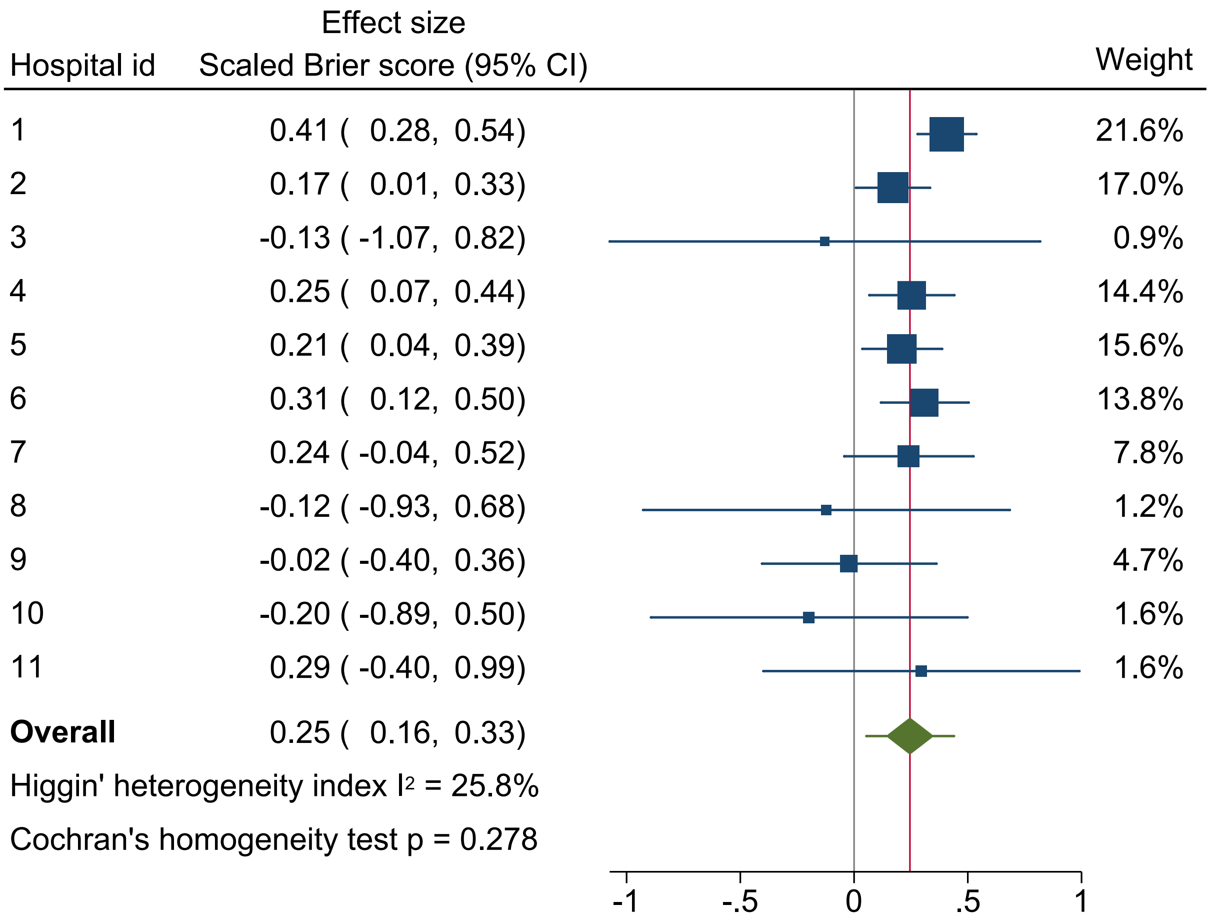
Hospital id	Effect size Scaled Brier score (95% CI)	Weight
1	0.43 (0.28, 0.57)	18.8%
2	0.22 (0.05, 0.39)	16.0%
3	0.09 (-0.48, 0.66)	2.6%
4	0.19 (-0.04, 0.42)	11.1%
5	0.17 (-0.05, 0.38)	12.3%
6	0.33 (0.12, 0.53)	12.9%
7	0.25 (-0.01, 0.51)	9.6%
8	-0.13 (-0.81, 0.55)	1.9%
9	-0.13 (-0.52, 0.26)	5.1%
10	0.07 (-0.28, 0.43)	5.9%
11	0.52 (0.05, 1.00)	3.6%
Overall	0.24 (0.14, 0.34)	

Higgin' heterogeneity index I^2 = 30.8%

Cochran's homogeneity test p = 0.151

Note. The overall meta-analyzed scaled Brier score is represented by the diamond centered on its estimated value with the diamond width corresponding to the length of the confidence interval. The green whiskers extending from the overall diamond represent the prediction interval, which provides a plausible range for the scaled Brier score in a future, new study.

Figure S10. Forest plot with hospital-specific Brier scores of the ACS-NSQIP adjusted prognostic model and overall pooled Brier score based on random-effects meta-analysis.



Note. The overall meta-analyzed scaled Brier score is represented by the diamond centered on its estimated value with the diamond width corresponding to the length of the confidence interval. The green whiskers extending from the overall diamond represent the prediction interval, which provides a plausible range for the scaled Brier score in a future, new study.

Table S5: Checklist for transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)

Section/Topic	Item	Checklist Item	Page
Title and abstract			
Title	1	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	Title Page file, 1
Abstract	2	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	Abstract file, 1-2
Introduction			
Background and objectives	3a	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	Manuscript file, 1-2
	3b	Specify the objectives, including whether the study describes the development or validation of the model or both.	2
Methods			
Source of data	4a	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.	2, 6-7
	4b	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	2
Participants	5a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	2
	5b	Describe eligibility criteria for participants.	3, and Supplementary Table S1
	5c	Give details of treatments received, if relevant.	2-3
Outcome	6a	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	3
	6b	Report any actions to blind assessment of the outcome to be predicted.	2-3
Predictors	7a	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	2-3, 7
	7b	Report any actions to blind assessment of predictors for the outcome and other predictors.	Not Applicable
Sample size	8	Explain how the study size was arrived at.	3-4
Missing data	9	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	4
Statistical analysis methods	10c	For validation, describe how the predictions were calculated.	3-4
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	4-6
	10e	Describe any model updating (e.g., recalibration) arising from the validation, if done.	6
Risk groups	11	Provide details on how risk groups were created, if done.	Not done
Development vs. validation	12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	6-7 and Supplementary Table S3
Results			

Participants	13a	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.	7
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	7, Tables 1 and 2
	13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	8, Supplementary Table S3
Model performance	16	Report performance measures (with CIs) for the prediction model.	8-10, Table 3, Figures 1-3, Supplementary Figure S1 – S10
Model-updating	17	If done, report the results from any model updating (i.e., model specification, model performance).	9-10, Supplementary Figure S3, Table S3 and Figure S4
Discussion			
Limitations	18	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	13-14
Interpretation	19a	For validation, discuss the results with reference to performance in the development data, and any other validation data.	10-13
	19b	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	10-13
Implications	20	Discuss the potential clinical use of the model and implications for future research.	14
Other information			
Supplementary information	21	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	14
Funding	22	Give the source of funding and the role of the funders for the present study.	15

Prospective multicenter external validation of postoperative mortality prediction tools in patients undergoing emergency laparotomy

Emergency Laparotomy patients in 11 centers (n=631)

30-day follow up

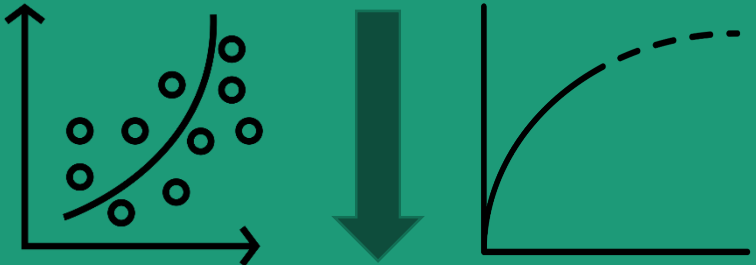
Prospective recording of risk factors and outcomes

30-day mortality → 16.3%

Data entry in 4 risk prediction tools (ACS-NSQIP, NELA, P-POSSUM and POTTER)

Comparison of multiple performance metrics

- **Scaled Brier score**
- **Discrimination**
- **Calibration**
- **Decision Curve Analysis**
- **Heterogeneity across hospitals**



Superiority of the *ACS-NSQIP* and the *surgeon-adjusted ACS-NSQIP* for prediction of 30-day mortality

@JTraumAcuteSurg

The Journal of
Trauma and
Acute Care Surgery®