Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

External validation and clinical utility of NELA and P-POSSUM risk models in an ageing Singapore emergency laparotomy cohort

Fang Ju Beatrice Koh

beatrice.koh.fangju@singhealth.com.sg

Singapore General Hospital

Charmian Chong

National Heart Centre Singapore

Gek Hsiang Lim

Singapore General Hospital

Sachin Mathur

Singapore General Hospital

Research Article

Keywords: Emergency laparotomy, risk stratification, NELA, ageing

Posted Date: September 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10555815/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Purpose

Emergency laparotomy remains one of the highest-risk general surgical procedures, particularly in an ageing population, with increasing comorbidity and physiological vulnerability. Although the National Emergency Laparotomy Audit (NELA) and Portsmouth Physiological and Operative Severity Score (P-POSSUM) models are widely adopted for perioperative risk assessment, their transportability and clinical utility in Asian healthcare systems remains largely uncertain. We aim to externally validate the NELA and P-POSSUM models in a Singapore emergency laparotomy cohort by evaluating discrimination, calibration and clinical utility.

Methods

Adult patients who had undergone emergency laparotomy in our centre from October 2020 to May 2022 were retrospectively identified. Primary outcome was all-cause death within 30 days following an emergency laparotomy. Receiver operating characteristic analysis was used to measure NELA and P-POSSUM model's discriminatory ability. Calibration was assessed by the Hosmer–Lemeshow goodness-of-fit test and calibration plot. Clinical utility was assessed for both models using decision curve analysis across clinically relevant mortality thresholds.

Results

A total of 393 patients were included (median age 69 years; 57.5% male), with an observed 30-day mortality of 8.1%. NELA demonstrated superior discriminatory performance compared with P-POSSUM (87.07, 95% CI: 81.42, 92.73) and is a better calibrated model. Decision curve analysis demonstrated greater clinical net benefit for NELA across clinically relevant threshold probabilities, whereas the net benefit of P-POSSUM declined progressively at higher decision thresholds.

Conclusion

NELA demonstrated superior discrimination and greater clinical utility than P-POSSUM in this ageing Singapore emergency laparotomy population. Although both models remained useful for perioperative risk stratification, differences in calibration and decision curve performance suggest that model transportability cannot be assumed across healthcare systems with differing patient populations and perioperative care pathways. NELA appears better suited to support risk stratification and perioperative decision-making in our setting.

Background

Emergency laparotomy continues to carry substantial postoperative morbidity and mortality, with reported 30-day mortality rates ranging from 6% to 20% across international cohorts despite improvements in perioperative care [1–3]. In Singapore, rapid population ageing means that an increasing proportion of patients requiring emergency abdominal surgery are older adults with multiple comorbidities and reduced physiological reserve. These patients often present with advanced intra-abdominal pathology with limited opportunity for preoperative optimisation, making accurate estimation of perioperative risk difficult.

Several risk prediction models have been developed to support perioperative risk assessment in emergency laparotomy. Among these, P-POSSUM has been one of the most widely adopted tools internationally, although concerns regarding calibration and systemic overestimation of mortality have been reported [3, 4]. In contrast, the NELA risk model, developed from over 38,000 cases in the United Kingdom, has shown superior calibration and discrimination in multiple external validation studies including in Australian and European populations [5, 6].

Despite increasing international validation, prediction models derived in one healthcare system cannot be assumed to perform identically in another. Although a previous multicentre Singapore study demonstrated good discrimination for both NELA and P-POSSUM, model performance is unlikely to remain uniform across institutions. Differences in patient demographics, referral patterns, disease severity, operative case mix, perioperative pathways may influence calibration, even within the same healthcare system [7]. Consequently, external validation in individual institutions remains important before these models are adopted for clinical decision-making.

We therefore performed an external validation of the NELA and P-POSSUM models in an emergency laparotomy cohort managed within a tertiary Acute Care Surgery service in Singapore. In addition to evaluating discrimination and calibration, we assessed clinical utility using decision curve analysis to determine whether either model provide greater net benefit across clinically relevant mortality thresholds used to guide perioperative care.

Methods

Study Design

This is a retrospective cohort study conducted at the largest tertiary hospital in Singapore, the Singapore General Hospital. All patients aged over 18 who underwent emergency laparotomy between October 2020 and May 2022 were retrospectively identified from the electronic hospital database. Inclusion and exclusion criteria mirrored that of the NELA study (Table 1).

Table 1
Inclusion and exclusion criteria of the NELA study

NELA inclusion criteria:
• Procedures involving the stomach, small or large bowel, or rectum for conditions such as perforation, ischaemia, abdominal abscess, bleeding or obstruction. • Washout/evacuation of intra-peritoneal abscess. • Bowel resection/repair due to incarcerated umbilical, inguinal and femoral hernias (but not hernia repair without bowel resection/repair). • Return to theatre for repair of substantial dehiscence of major abdominal wound (i.e. 'burst abdomen').
NELA exclusion criteria:
• Procedures for appendicitis, cholecystitis, vascular surgery, trauma, obstetric emergencies. • Emergency creation of colostomy or ileostomy as either trephine or laparoscopy.

Data Collection

Preoperative and intraoperative parameters needed to determine the NELA and P-POSSUM scores were extracted and analysed. The NELA [8] and P-POSSUM [9] scores were calculated as per the risk calculator. The primary outcome was all-cause death within 30 days following an emergency laparotomy. For patients who underwent repeat emergency laparotomy procedures, the parameters from the first index procedure were used while that of subsequent procedures were excluded.

Statistical Analysis

Descriptive statistics were used to summarize the data. Continuous variables were reported as mean (standard deviation [SD]) or median (interquartile range [IQR]), as appropriate according to their distribution, while categorical variables were reported as frequencies and percentages.

The predictive performance of the NELA and P-POSSUM models for mortality was evaluated in terms of discrimination, calibration, and clinical utility. Discriminatory ability was assessed separately for each model using receiver operating characteristic (ROC) curve analysis, with the area under the ROC curve (AUC) and corresponding 95% confidence intervals reported.

Calibration was assessed using multiple complementary approaches. Calibration plots comparing predicted and observed mortality risks across the range of predicted probabilities were constructed for each model. Goodness-of-fit was further evaluated using the Hosmer–Lemeshow test. The ratio of observed-to-expected (O:E) deaths was calculated and reported across categories of predicted mortality risk for each model. Overall predictive accuracy was assessed using the Brier score, with lower values indicating better agreement between predicted probabilities and observed outcomes.

The potential clinical usefulness of each prediction model was evaluated using decision curve analysis (DCA), which estimates the net benefit of applying a model across a range of clinically relevant threshold probabilities for mortality risk.

Statistical analyses were performed using Stata 17 (StataCorp. 2021. Stata Statistical Software: Release 17. College Station, TX, StataCorp LLC).

Ethics

This study was approved by the SingHealth Centralised Institutional Review Board (CIRB). The authors declare no conflict of interest. This research did not receive any grant from any funding agency.

Results

Baseline patient characteristics

A total of 393 patients underwent emergency laparotomy between October 2020 and May 2022. The median age of patients was 69 years (IQR: 59, 76), and 57.5% were male. The ASA classification is shown in Table 2.

Table 2 Patient characteristics

Patient characteristics	n = 393
Age; median (IQR)	69 (59, 76)
Gender; n (%)	
Male	226 (57.5)
Female	167 (42.5)
Race; n (%)	
Chinese	326 (82.95)
Malay	32 (8.14)
Indian	18 (4.58)
Others	17 (4.33)
BMI; median (IQR)	22.1 (19.5, 25.2)
Missing	2
ASA score; n (%)	
I	14 (3.56)
II	123 (31.30)
III	150 (38.17)
IV	105 (26.72)
V	1 (0.25)
Urgency of operation; n (%)	
Immediate (<2 hours)	9 (2.29)
Urgent (2-6 hours)	244 (62.09)
Urgent (6-18 hours)	16 (4.07)
Expedited (>18 hours)	124 (31.55)
Mode of surgery; n (%)	
Emergency (within 24 hours), resuscitation	383 (97.46)
Emergency (within 2 hours)	10 (2.54)
POSSUM mortality risk calculation; median (IQR)	9.4 (4.4, 20.9)
NELA pre-op mortality risk calculation; median (IQR)	5.16 (1.73, 11.58)

Observed death within 30 days; n (%)	32 (8.14)

Comparison between models

The observed 30-day mortality was 8.14% (n=32). The observed/expected (O/E) ratio was 0.93 for NELA and 0.48 for P-POSSUM; where the expected 30-day mortality was calculated by aggregating the predicted mortality risk of all patients based on each model. This indicated that both models over-predicted mortality, with NELA over-predicting less than P-POSSUM. The median mortality risk was estimated as 9.40% based on the P-POSSUM model and 5.16% based on the NELA model. AUC for the P-POSSUM model was 80.02 (95% CI: 71.97, 88.07) (Fig. 1A) and the AUC for NELA was 87.07 (95% CI: 81.42, 92.73) (Fig. 1B), indicating that NELA is a better predictor for mortality within 30 days as compared to POSSUM. The Hosmer-Lemeshow goodness of fit statistic indicated that both the POSSUM and NELA model fit well to the observed data.

The McNemar's test was used to compare whether there was any difference between NELA and POSSUM predicted mortality risk. The predicted mortality risk was classified into ≥5% (high risk) and <5% (not high risk), following the recommendations of the NELA project team and the Royal College of Surgeons. It was observed that there was a significant difference in the mortality risk prediction between NELA and POSSUM in patients who survived beyond 30 days (p<0.001) (Table 3). Notably, 103 (28.6%) patients were classified as non-high risk by NELA but classified as high risk by POSSUM among patients who survived beyond 30 days, suggesting that POSSUM tended to provide a higher mortality risk than NELA. Table 4 indicated that there was no significant difference in the risk prediction between NELA and POSSUM in patients who did not survive beyond 30 days (p=0.16).

Table 3 Patients who survived beyond 30 days

	POSSUM	
NELA	**Not high risk**	**High risk**
Not high risk	**89**	**103**
High risk	**20**	**148**

McNemar's chi-square test p<0.001

Table 4 Patients who did not survive beyond 30 days

	POSSUM	
NELA	**Not high risk**	**High risk**
Not high risk	**0**	**2**
High risk	**0**	**30**

McNemar's chi-square test p=0.16

Calibration plots showed that NELA was generally well calibrated across the range of predicted risks, with only slight underestimation of mortality at higher predicted probabilities. Conversely, P-POSSUM demonstrated a tendency to overpredict mortality, particularly among patients in the moderate- and high-risk strata, indicating poorer calibration compared with NELA.

The NELA model also demonstrated superior overall predictive accuracy, with a lower Brier score than P-POSSUM (0.0606 vs. 0.0783). This indicates that NELA generated more accurate mortality predictions, with less prediction error and better agreement between predicted and observed outcomes.

Decision curve analysis demonstrated that the NELA model provided greater clinical utility than P-POSSUM across most clinically relevant threshold probabilities. NELA consistently achieved a higher net benefit than both the "treat-all" and "treat-none" strategies over a broad range of thresholds, whereas the net benefit of P-POSSUM declined more rapidly and became negligible or negative at higher threshold probabilities. These findings suggest that NELA may offer superior clinical value for risk stratification and decision-making in this patient population.

Discussion

In this study, we externally validated the predictive performance of the NELA and P-POSSUM scoring systems in a cohort of patients undergoing emergency laparotomy. Our results demonstrate that both models exhibited good calibration via the Hosmer-Lemeshow goodness-of-fit test, indicating an acceptable fit between predicted and observed mortality. However, discrimination, as measured by the area under the receiver operating characteristic curve (AUC), was higher for the NELA score [AUC 87.1 (95% CI: 81.4, 92.7)] compared to the P-POSSUM score [AUC 80.0 (95% CI: 72.0, 88.1)]. While discrimination is an important measure of predictive performance, it does not necessarily indicate whether use of a model leads to better clinical decision-making. Decision curve analysis therefore provides complementary information by quantifying net clinical benefit achieved across different risk thresholds. NELA consistently demonstrated greater net benefit than P-POSSUM across clinically relevant threshold probabilities of approximately 5%–20%, which encompass the range commonly used to guide escalation of perioperative care, consultant involvement, and postoperative critical care planning. These findings suggest that beyond superior statistical discrimination, NELA is more likely to support appropriate risk-based clinical decisions while reducing unnecessary escalation of care.

Several characteristics of our healthcare system may also contribute to differences in model performance compared with the original UK derivation cohort. Our institution operates a consultant-led Acute Care Surgery service, allowing early senior surgical review and expedited decision-making for emergency laparotomy patients. In addition, rapid access to intensive care beds, standardised perioperative pathways involving anaesthetists, and a well-resourced healthcare system may improve postoperative outcomes beyond those expected from patient-related risk factors alone. These organisational differences are not explicitly captured within either of the models studied and may

partially explain the observed tendency towards mortality overprediction in our cohort despite acceptable discrimination.

External validation studies have demonstrated considerable variability in the performance of both NELA and P-POSSUM across different healthcare settings. In the Egyptian cohort reported by Alabbasay et al. [10], both models underestimated the risk of 30-day mortality despite demonstrating acceptable discrimination (preoperative NELA AUC 0.774; P-POSSUM AUC 0.763). Conversely, our study and that of Lai et al. [7] found that both models tended to overestimate mortality, with P-POSSUM exhibiting greater overprediction than NELA. These contrasting findings suggest that the direction and magnitude of model miscalibration are highly population-dependent and may reflect differences in patient characteristics, disease severity, perioperative care pathways, and healthcare systems. Similarly, Barazanchi et al. [2] compared NELA, P-POSSUM, ACS-NSQIP and APACHE-II, demonstrating that NELA achieved the highest discrimination (AUC 0.83), which further improved to 0.88 when combined with preoperative weight-loss status, highlighting the potential value of incorporating additional patient-level factors. Consistent with these observations, the recent systematic review by Hewitt et al. [1] concluded that no single risk prediction model consistently outperformed the others across all populations and recommended local validation before widespread clinical implementation.

Most existing risk stratification models assume linear interactions between patient and procedure risk factors, however, that may not always be true. The Predictive Optimal Trees in Emergency Surgery Risk algorithm (POTTER) was developed using the American College of Surgeons (ACS) NSQIP database which allows non-linear interactions between variables and had been validated in multiple studies [11, 12]. A key limitation in applying the POTTER algorithm retrospectively lies in its use of adaptive questioning, wherein the selection of subsequent input variables depends on prior responses. This design enhances clinical usability but poses challenges in retrospective analyses, as many of the context-specific variables required for risk calculation may not be routinely or uniformly documented in existing records. Consequently, missing data limits the completeness and accuracy of POTTER-based predictions in retrospective cohorts.

Limitations of our study include its single-institution design, which may limit external generalizability and low number of deaths in our patient cohort could also affect the power of the study. Furthermore, although Hosmer-Lemeshow testing showed good model fit, this test can be sensitive to sample size and grouping methods, potentially masking minor miscalibrations. An additional limitation is that the study period coincided with the COVID-19 pandemic. Although emergency laparotomy services continued throughout this period, perioperative practice may have been influenced by pandemic-related factors including changes in operating theatre workflows, critical care resource allocation and infection control measures. These factors may have affected postoperative outcomes and therefore the observed performance of both risk prediction models. We also did not include a comparison with newer tools such as POTTER as a significant number of factors required for the calculation were not collected as part of our audit which would result in incompleteness of data.

Conclusion

In conclusion, both NELA and P-POSSUM demonstrated acceptable calibration in this contemporary Singapore emergency laparotomy cohort; however, NELA provided superior discrimination, less mortality overprediction, and greater clinical utility as demonstrated by decision curve analysis. These findings support the use of NELA as the preferred perioperative risk stratification tool in our institution. Nevertheless, differences in healthcare systems, perioperative pathways, and patient populations highlight the importance of local external validation before widespread adoption of risk prediction models. Future multicentre studies involving contemporary post-pandemic Asian cohorts and evaluation of emerging machine learning-based prediction models may further improve individualized perioperative risk assessment.

Abbreviations

P-POSSUM Portsmouth Physiological and Operative Severity Score for the enumeration of Mortality and Morbidity
NELA National Emergency Laparotomy Audit

AUC Area under the curve

POTTER Predictive Optimal Trees in Emergency Surgery Risk algorithm

Declarations

Ethics approval and consent to participate

All methods were carried out in accordance with relevant regulations. This study was performed in line with the principles of the Declaration of Helsinki. Due to the retrospective nature of the study and anonymized data use, institutional review board waiver (CIRB reference number: 2023/2141) was obtained.

Consent to publish

Not applicable.

Competing interest

The authors declare no competing interests.

Funding

No funding was received for conducting this study.

Author Contribution

Koh Fang Ju Beatrice: study conception and design, article writing, literature review.Charmian Chong: data collection, article writing, literature review. Lim Gek Hsiang: statistics, data evaluation. Sachin Mathur: study conception and design. All authors read and approved the final manuscript

Data Availability

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

References

1. Hewitt JN, Milton TJ, Jeanes J, Murshed I, Nann S, Wells S, et al. Emergency laparotomy preoperative risk assessment tool performance: A systematic review. Surg Pract Sci. 2024;31(19):100264. https://doi.org/10.1016/j.sipas.2024.100264.
2. 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):382–7. https://doi.org/10.1097/TA.0000000000002745.
3. Darbyshire AR, Kostakis I, Pucher PH, Prytherch D, Mercer SJ. P-POSSUM and the NELA score overpredict mortality for laparoscopic emergency bowel surgery: an analysis of the NELA database. World J Surg. 2022;46(3):552–60. https://doi.org/10.1007/s00268-021-06404-w.
4. Barghash M, Iskandar A, Fawzy SI, Effiom D, Huck C, Hajibandeh S, et al. Predictive performance of NELA Versus P-POSSUM mortality scores: are we underestimating the risk of mortality following emergency laparotomy? Cureus. 2022;14(12):e32859. https://doi.org/10.7759/cureus.32859.
5. Hunter Emergency Laparotomy Collaborator Group. High-risk emergency laparotomy in Australia: comparing NELA, P-POSSUM, and ACS-NSQIP calculators. J Surg Res. 2020;246:300–4. https://doi.org/10.1016/j.jss.2019.09.024.
6. Kokkinakis S, Kritsotakis EI, Paterakis K, Karali GA, Malikides V, Kyprianou A, et al. Prospective multicenter external validation of postoperative mortality prediction tools in patients undergoing emergency laparotomy. J Trauma Acute Care Surg. 2023;94(6):847–56. https://doi.org/10.1097/TA.0000000000003904.
7. 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):2439–46. https://doi.org/10.1007/s00268-021-06120-5.
8. Calculation of NELA score. Available on this link https://nela-risk-tool.apps.crownaudit.org
9. Calculation of P-POSSUM score. Available on this link https://www.mdapp.co/p-possum-score-calculator-428/

10. Alabbasy MM, Elsisy AAE, Mahmoud A, Alhanafy SS. Comparison between P-POSSUM and NELA risk score for patients undergoing emergency laparotomy in Egyptian patients. BMC Surg. 2023;23(1):286. https://doi.org/10.1186/s12893-023-02189-y.
11. Panossian VS, Argandykov D, Arnold SC, Gebran A, Paranjape CN, Hwabejire JO, et al. Validation of Artificial Intelligence-based POTTER calculator in emergency general surgery patients undergoing laparotomy: prospective, bi-Institutional study. J Am Coll Surg. 2025;240(3):254–62. https://doi.org/10.1097/XCS.0000000000001234.
12. Ribeiro Junior MAF, Smaniotto R, Gebran A, Zamudio JP, Mohseni S, Rodrigues JMDS, et al. The use of POTTER (Predictive Optimal Trees in Emergency Surgery Risk) calculator to predict mortality and complications in patients submitted to emergency surgery. Rev Col Bras Cir. 2023;50:e20233624. https://doi.org/10.1590/0100-6991e-20233624-en.

Figures

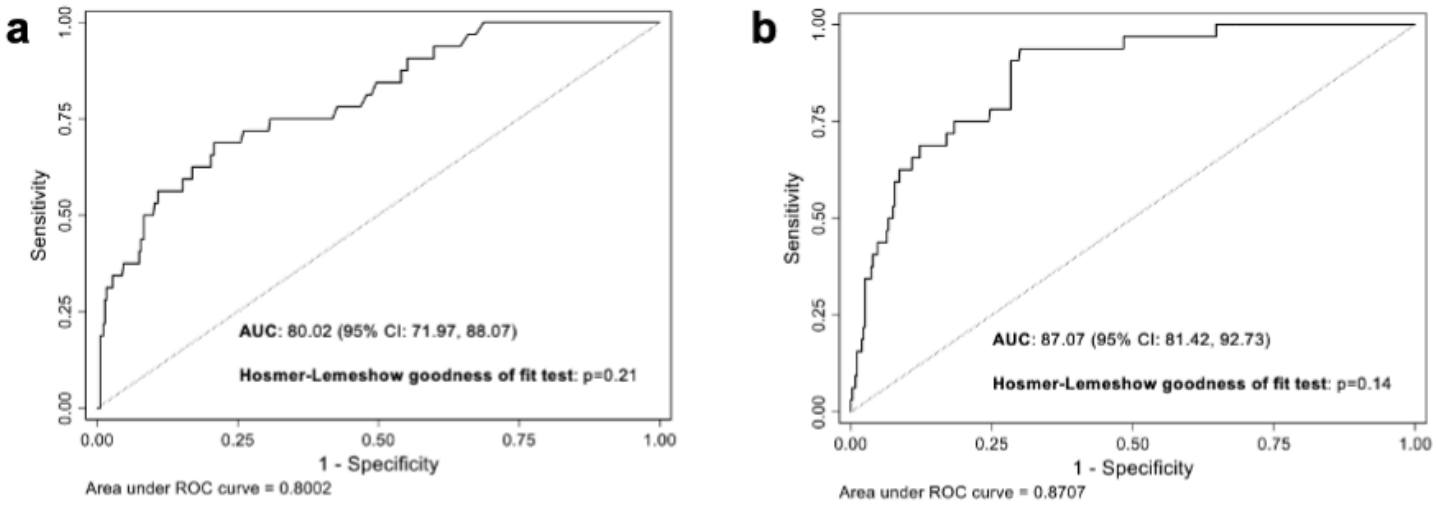


Figure 1

ROC analysis of P-POSSUM model **(a)** and NELA model **(b)** against observed death within 30 days.

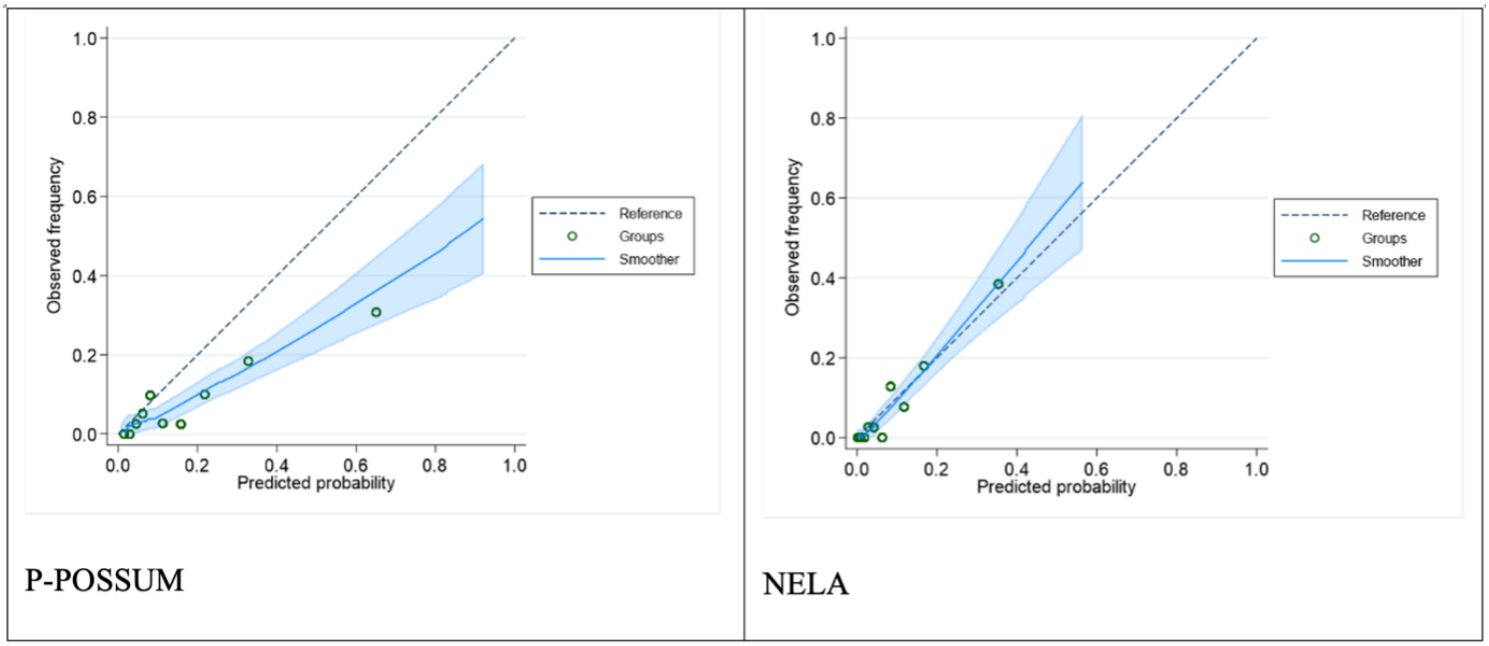


Figure 2

Calibration plots for P-POSSUM and NELA model predicting 30-day mortality following emergency laparotomy. The x-axis represents the predicted probability of 30-day mortality, while the y-axis represents the observed mortality. The 45-degree reference line represents perfect agreement between predicted and observed risk. Deviation from this line reflects model miscalibration.

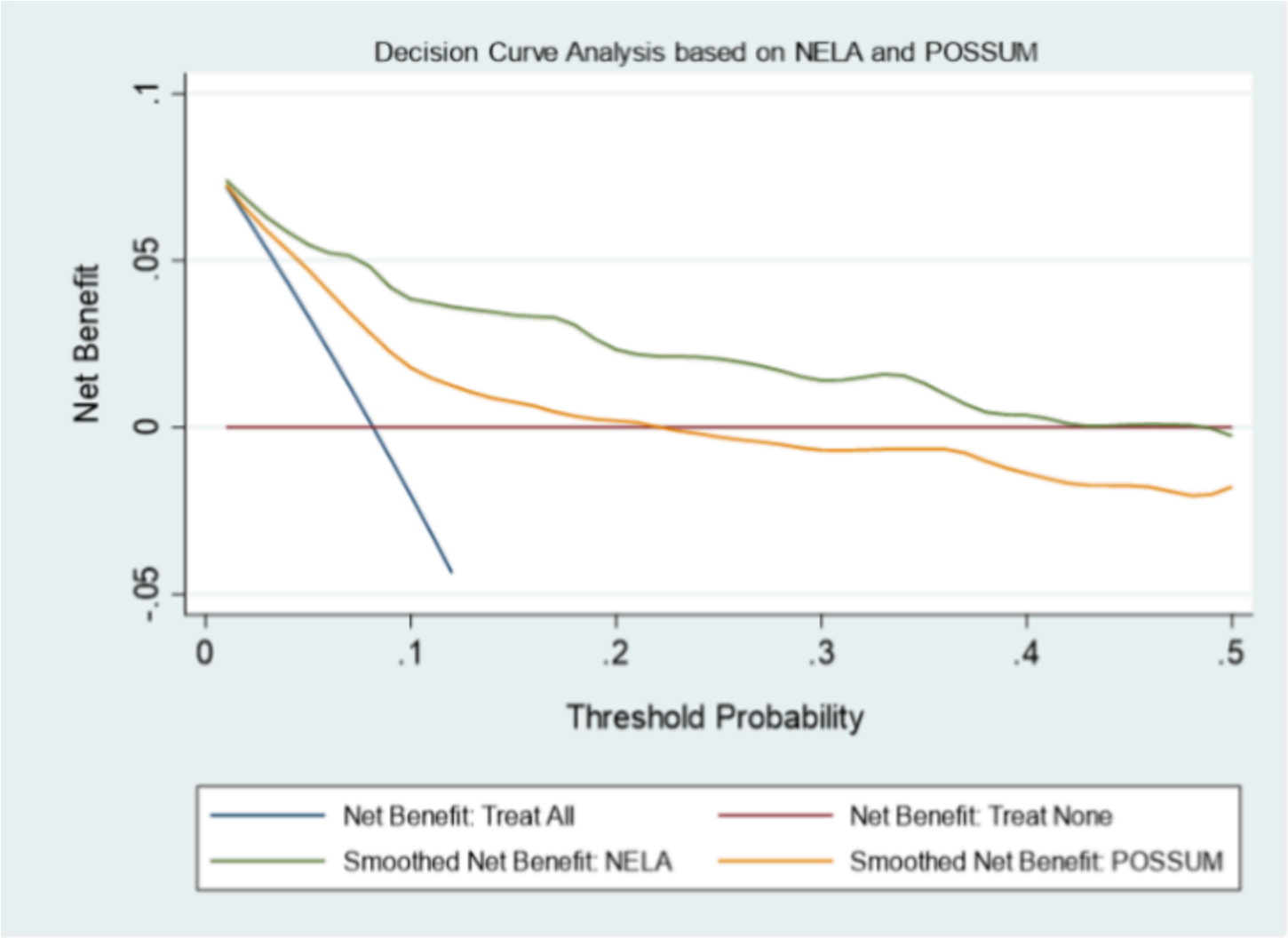


Figure 3

Decision curve analysis of P-POSSUM model and NELA model demonstrating net benefit across different threshold probabilities