SCHOOL *of* CLINICAL SCIENCES

UNIVERSITY/BHF CENTRE FOR CARDIOVASCULAR SCIENCE

The University of Edinburgh
SU.305 Chancellor's Building
Royal Infirmary of Edinburgh
49 Little France Crescent
Edinburgh EH16 4SB
Email: Dimitrios.Doudesis@ed.ac.uk

Validation of Lenus Stratify's predictive models for chronic obstructive pulmonary disease risk: an independent evaluation

August 13th, 2024

Dr Dimitrios Doudesis, BSc, MSc, PhD

Senior Data Scientist & Lecturer - Health Data Science

BHF Centre for Cardiovascular Science, Royal Infirmary of Edinburgh

49 Little France Crescent, Edinburgh EH16 4SU

Website: dimitriosdoudesis.com

An online version of the report can be found here

Table of Contents

Table of Figures and Tables

Acknowledgements

This work uses data provided by patients and collected by the NHS as part of their care and support. This project has been facilitated by the DataLoch service (reference: DL_2022_024). DataLoch enables access to de-identified extracts of health care data from the South-East Scotland region to approved applicants: dataloch.org.

Approvals/Ethics summary

The project received approval through DataLoch, including: Caldicott Guardian approval, ACCORD sponsorship (AC23005), favourable ethical opinion under DataLoch's delegated authority (Reference: 22/NS/0093), and a public value assessment by the DataLoch Public Reference Group.

Analysis statement

All analyses were performed using de-identified data within a secure data environment approved by DataLoch.

Data availability statement

Data may be accessed through DataLoch (dataloch.org) following successful application and approvals.

Introduction

Lenus Health has developed a suite of risk-prediction models for chronic obstructive pulmonary disease (COPD) in collaboration with clinicians from the respiratory innovation team in NHS Greater Glasgow and Clyde (NHS GG&C) using routinely collected data from electronic health records. These models were originally trained on deidentified historic data for patients with a COPD diagnosis in NHS GG&C and were operationalised to be used in live clinical practice as part of the DYNAMIC-AI trial (NCT05914220). Further details of the model development approach and the evaluation of model performance by Lenus Health in NHS GG&C can be found here. Having proven their utility in NHS GG&C, variants of these models were developed in NHS Lothian using local deidentified data provided by DataLoch (a partnership between the University of Edinburgh and NHS Lothian). Lenus Health reports that this suite of models are performant in both NHS Lothian and NHS GG&C but to date there has not been an external independent evaluation into the performance of these models on unseen data.

To address this, the University of Edinburgh was contracted to evaluate the performance of three of the models (the 12-month mortality model, the 3-month all cause readmission model, and the 3-month respiratory related readmission model) on data that was not available to Lenus Health. The data science team from Lenus Health generated calibrated probability estimates for each of the models for individuals in the deidentified dataset to evaluate the performance of the models on unseen data. The data science team from Lenus Health trained the 12-month mortality model on data up to the end of 2022 and the 3-month readmission models on data up to the last discharge date participants had from the 18th to the 31st March 2023 (where the data available to Lenus Health was censored). Admissions data censored on the 30th June 2023 (3-

months following the latest possible discharge date) and mortality data for all of 2023 were provided to the independent researcher (Dr Dimitrios Doudesis) from the University of Edinburgh along with the models' output probabilities and source of truth data. The data science team from Lenus Health, didn't have access to the source of truth data of the unseen data, only to data necessary to compute the probabilities for each model.

This report summarises the findings from this external validation. The receiver operating characteristic (ROC), precision-recall (PR) and calibration curves for each of the three models will be presented. Additional analysis looking at the impact and consequences of decision threshold selection will also be shown for each of these models. This includes decision curve analysis and analysis of the numbers of correct and incorrect predictions at illustrative thresholds.

Results

Populations and outcomes

For the 12-month mortality model evaluation there were 22,941 patients in total (those alive in the dataset available to the Lenus Health team at the end of 2022). A total of 1618 individuals died over 2023. Individuals who died within the 3-month prediction window for the readmission models were removed from the evaluation to match the context that these models were trained on and to reflect the intended use of these models alongside a mortality prediction model. For the 3-month all cause readmission model there were 456 individuals included in total based on the number of individuals with an all-cause hospital admission occurring within the last two weeks of March 2023 and excluding 77 individuals who died within the prediction window. A total of 199 out of 456 individuals readmitted within the 3-month prediction window. For the 3-month respiratory related readmission model there were 163 individuals included based on the number of patients with a respiratory related hospital admission occurring within the last two weeks of March 2023 and excluding 50 individuals who died within the prediction window. A total of 48 out of 163 individuals had a respiratory related admission within the 3-month prediction window.

Model calibration

The calibration of each of the 3 models is shown below for the validation data. The Brier score for the 12-month mortality model (*Figure 1-A*) was 0.055, while for the 3-month all cause readmission model (*Figure 1-B*) and 3-month respiratory related readmission model (*Figure 1-C*) were 0.184 and 0.17, respectively.

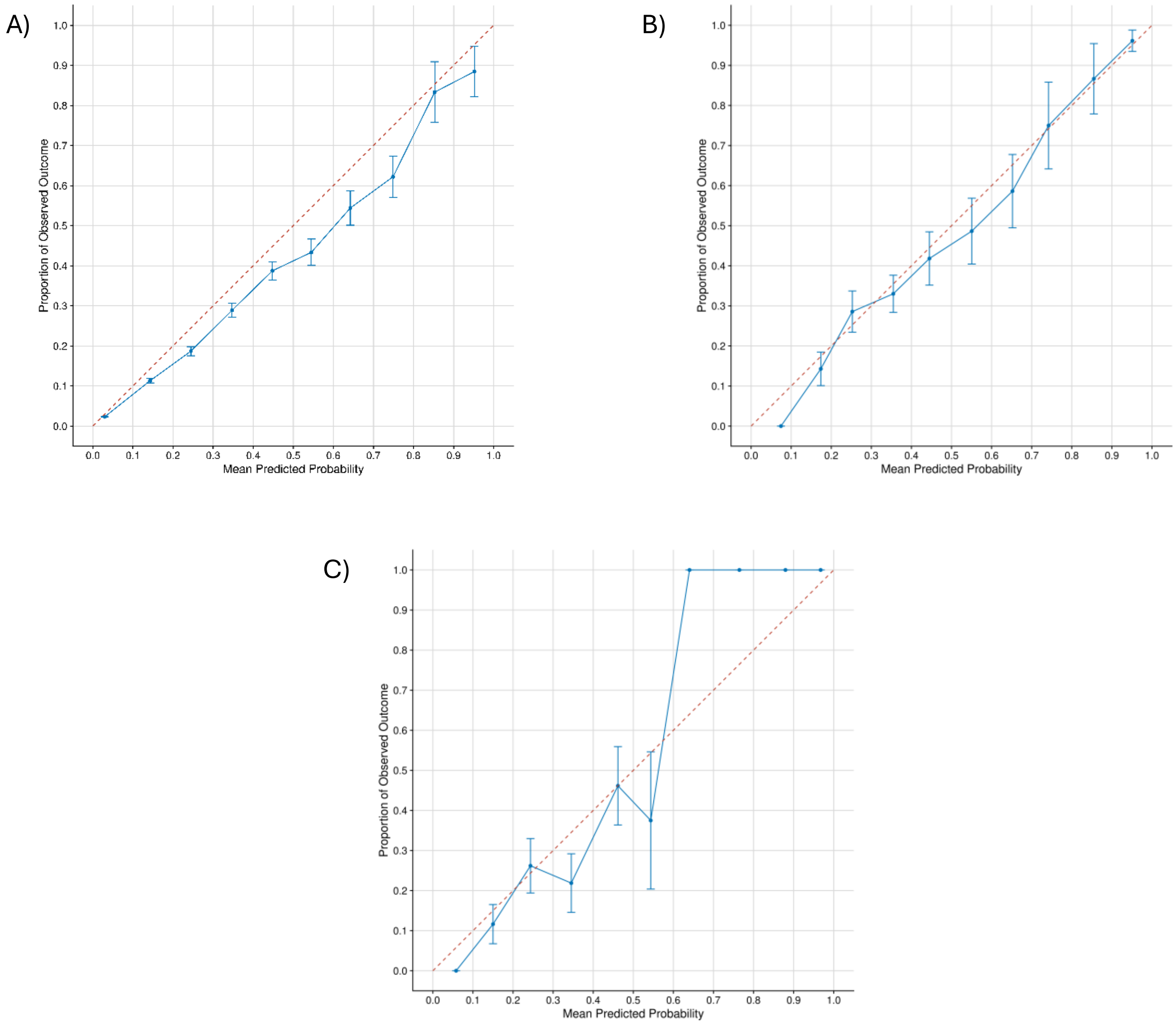


Figure 1. Calibration curves for the 12-month mortality model (panel A), 3-month all cause readmission model (panel B) and 3-month respiratory related readmission model (panel C).

Model discrimination

The area under the curve of the Receiver Operating Characteristic (AUC-ROC) and the area under the curve of the Precision-Recall (AUC-PR) curves are shown for each model for the validation data. The AUC-ROC for the 12-month mortality model (Figure 2-A) was 0.842 (95% CI: 0.832 to 0.852). The AUC-ROC values for the 3-month all cause readmission model (Figure 3-A) and the 3-month respiratory-related readmission model (Figure 4-A) were 0.770 (95% CI: 0.726 to 0.814) and 0.722 (95% CI: 0.631 to 0.813), respectively. Regarding the AUC-PR, the 12-month mortality model (Figure 2-B) achieved a score of 0.365, while the 3-month all cause readmission model (Figure 3-B) and the 3-month respiratory related readmission model (Figure 4-B) scored 0.766 and 0.615, respectively.

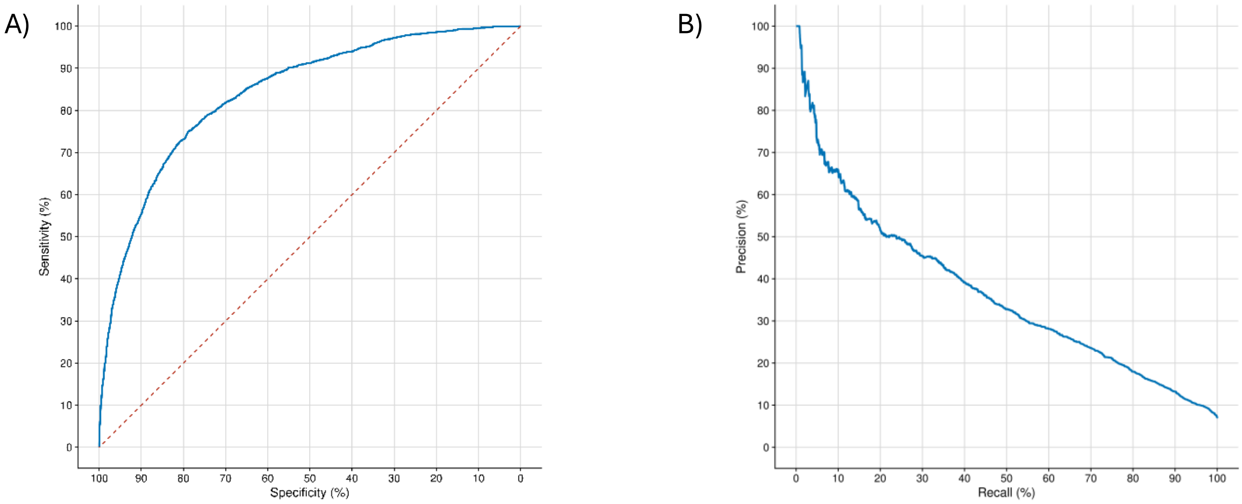


Figure 2. Receiver Operating Characteristic and Precision-Recall Curves (panel A) for the 12-month mortality model. The AUC-ROC curve is 0.842 (95% CI: 0.832 to 0.852). The AUC-PR curve (panel B) is 0.365.

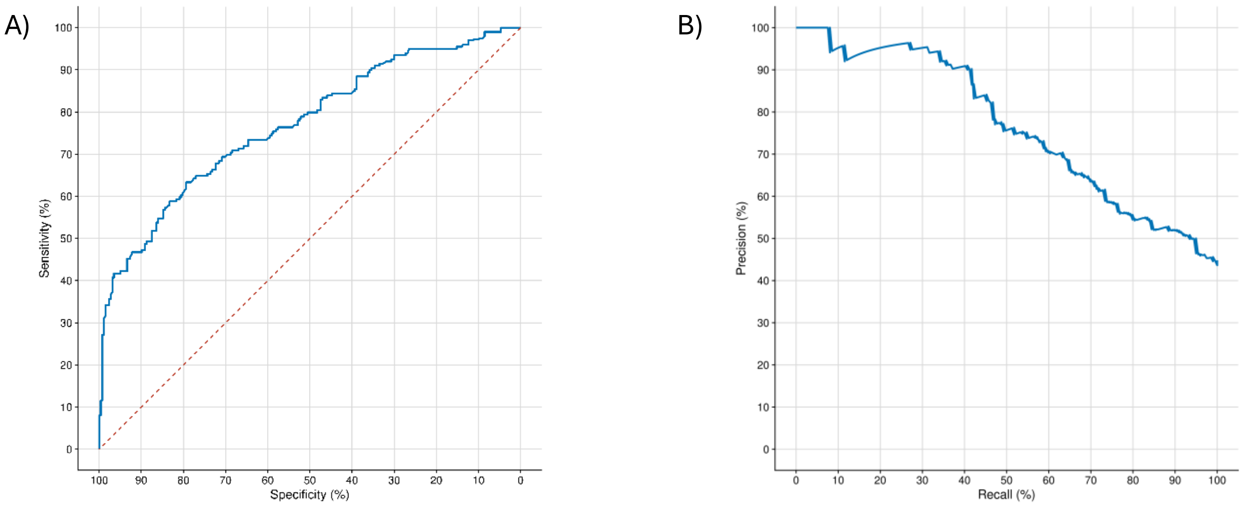


Figure 3. Receiver Operating Characteristic and Precision-Recall Curves (panel A) for the 3-month all cause readmission model. The AUC-ROC curve is 0.770 (95% CI: 0.726 to 0.814). The AUC-PR curve (panel B) is 0.766.

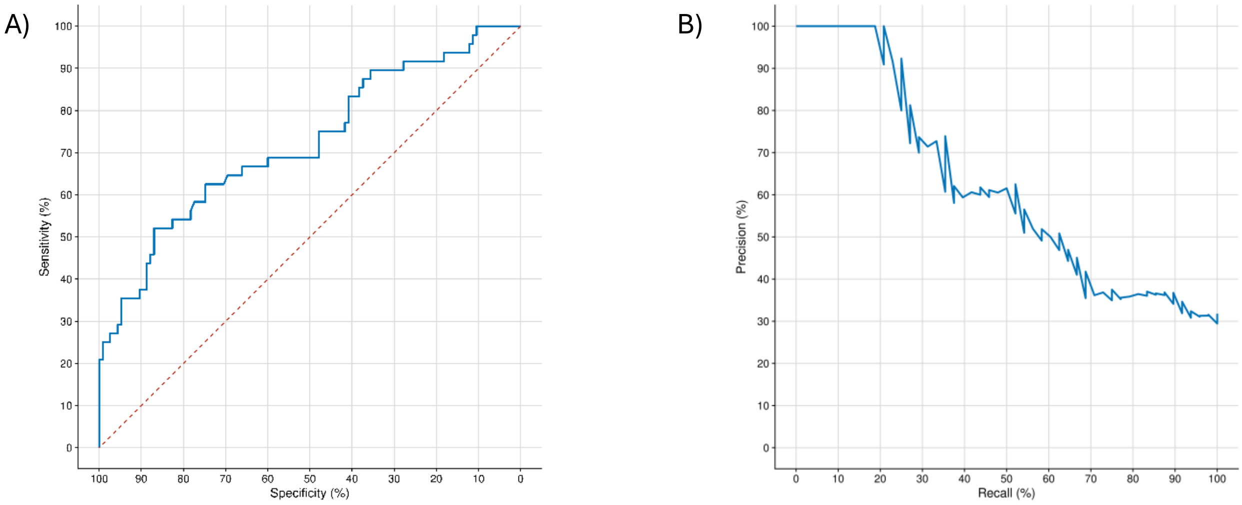
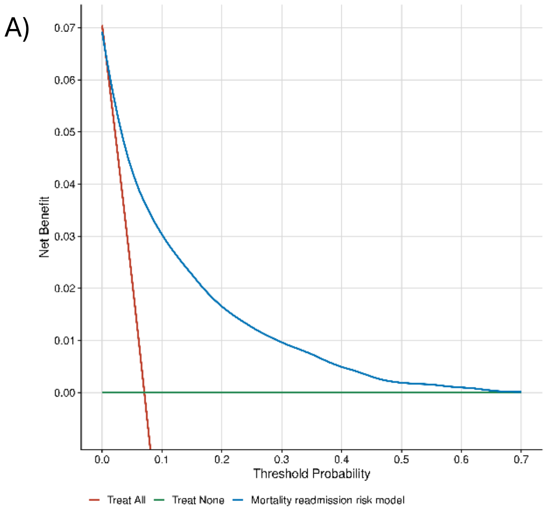
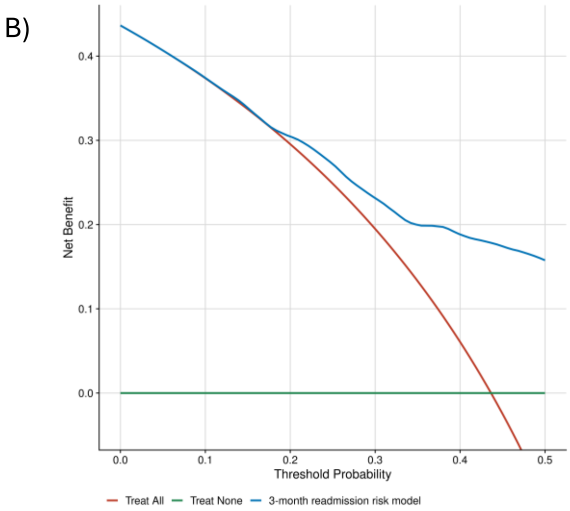


Figure 4. Receiver Operating Characteristic and Precision-Recall Curves (panel A) for the 3-month respiratory related readmission model. The AUC-ROC curve is 0.722 (95% CI: 0.631 to 0.813). The AUC-PR curve (panel B) is 0.615.

Decision curve analysis

Decision curve analysis for each of the models is shown below. Each figure presents the decision curves, illustrating the net benefit across a range of threshold probabilities. The analysis compares the clinical usefulness of the models by showing the net benefit of making decisions based on the model predictions, relative to treating all patients or treating none. Each curve represents the respective model's performance in balancing the trade-offs between true positives and false positives at various thresholds, providing insight into their practical application in clinical decision-making.





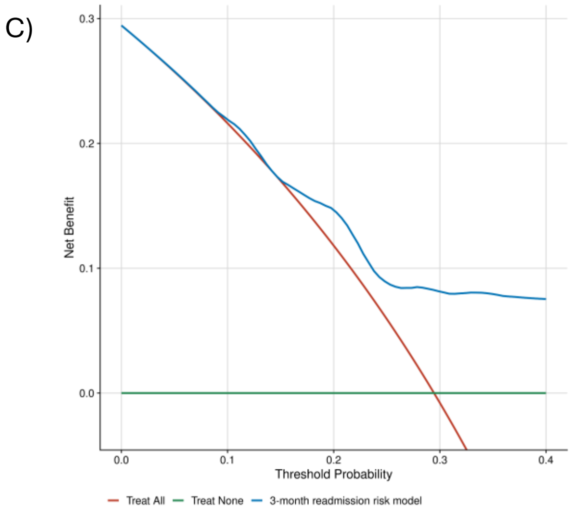


Figure 5. Decision Curve Analysis for the 12-month mortality model, 3-month all cause readmission model, and 3-month respiratory related readmission model.

Confusion matrices

The tables below presents the confusion matrix for each model at different thresholds.

Table 1. Confusion matrix for the 12-month mortality model at various thresholds. The table presents the true negative (TN), false negative (FN), false positive (FP), and true positive (TP) counts at four different decision thresholds: 0.15, 0.25, 0.30, and 0.40.

	True Negative	*False Negative*	*False Positive*	*True Positive*
	TN	*FN*	*FP*	*TP*
Threshold of 0.15	18,333	569	2,990	1,049
Threshold of 0.25	19,946	870	1,377	748
Threshold of 0.3	20,341	977	982	641
Threshold of 0.4	20,806	1,166	517	452

Table 2. Confusion matrix for the 3-month all cause readmission model at various thresholds. The table presents the true negative (TN), false negative (FN), false positive (FP), and true positive (TP) counts at four different decision thresholds: 0.25, 0.4, 0.45, and 0.6.

	True Negative	*False Negative*	*False Positive*	*True Positive*
	TN	*FN*	*FP*	*TP*
Threshold of 0.25	89	18	168	181
Threshold of 0.4	186	66	71	133
Threshold of 0.45	207	80	50	119
Threshold of 0.6	237	107	20	92

Table 3. Confusion matrix for the 3-month respiratory related readmission model at various thresholds. The table presents the true negative (TN), false negative (FN), false positive (FP), and true positive (TP) counts at four different decision thresholds: 0.15, 0.2, 0.25, and 0.40.

	True Negative	*False Negative*	*False Positive*	*True Positive*
	TN	*FN*	*FP*	*TP*
Threshold of 0.15	20	3	95	45
Threshold of 0.2	40	5	75	43
Threshold of 0.25	56	15	59	33
Threshold of 0.4	96	23	19	25

The tables below summarises diagnostic metrics for each model at different thresholds.

Table 4. Predictive performance metrics for the 12-month mortality model at various thresholds. The table presents the negative predictive value (NPV), recall/sensitivity, specificity, precision/positive predictive value (PPV), F1 score for positive and negative predictions at four different decision thresholds: 0.15, 0.25, 0.30, and 0.40.

	NPV	*Recall/ Sensitivity*	*Specificity*	*Precision/ PPV*	*F1 Score positive*	*F1 Score negative*
Threshold of 0.15	0.97	0.65	0.86	0.26	0.37	0.91
Threshold of 0.25	0.96	0.46	0.94	0.35	0.4	0.95
Threshold of 0.3	0.95	0.4	0.95	0.39	0.4	0.95
Threshold of 0.4	0.95	0.28	0.98	0.47	0.35	0.96

Table 5. Predictive performance metrics for the 3-month all cause readmission model at various thresholds. The table presents the negative predictive value (NPV), recall/sensitivity, specificity, precision/positive predictive value (PPV), F1 score for positive and negative predictions at four different decision thresholds: 0.25, 0.4, 0.45, and 0.6.

	NPV	*Recall/ Sensitivity*	*Specificity*	*Precision/ PPV*	*F1 Score positive*	*F1 Score negative*
Threshold of 0.25	0.83	0.91	0.35	0.52	0.66	0.49
Threshold of 0.4	0.74	0.66	0.73	0.65	0.66	0.73
Threshold of 0.45	0.72	0.6	0.81	0.7	0.65	0.7
Threshold of 0.6	0.69	0.46	0.92	0.82	0.59	0.79

Table 6. Predictive performance metrics for the 3-month respiratory related readmission model at various thresholds. The table presents the negative predictive value (NPV), recall/sensitivity, specificity, precision/positive predictive value (PPV), F1 score for positive and negative predictions at four different decision thresholds: 0.15, 0.20, 0.25, and 0.40**.**

	NPV	*Recall/ Sensitivity*	*Specificity*	*Precision/ PPV*	*F1 Score positive*	*F1 Score negative*
Threshold of 0.15	0.87	0.94	0.17	0.32	0.48	0.29
Threshold of 0.2	0.89	0.9	0.35	0.36	0.52	0.5
Threshold of 0.25	0.79	0.69	0.49	0.36	0.47	0.6
Threshold of 0.4	0.81	0.52	0.83	0.57	0.54	0.82

Interpretation of the results

Calibration curves and Brier scores

12-month mortality model

The calibration plot for the 12-month mortality model demonstrates that the predicted probabilities are reasonably well-aligned with the observed outcomes, particularly in the low to mid-range probabilities, where the model is expected to be used. There are slight deviations at the mid to high-range probabilities, where the model overestimates risk. This suggests that while the model is generally well-calibrated, some refinement could improve its performance in predicting mid to high-risk cases if the scope of the model changes in the future. The Brier score indicates an overall moderate accuracy in probabilistic predictions, reflecting a balance between calibration and sharpness.

3-month all cause readmission model

The calibration curve for the 3-month all cause readmission model indicates strong alignment between predicted probabilities and observed outcomes across most probability ranges. The model is particularly well-calibrated in the mid to high probability ranges (expected use), suggesting it is reliable for predicting patients at a higher risk of all cause readmission. The Brier score is the lowest among the three models, indicating the highest accuracy in probabilistic predictions. This suggests that not only performs well in terms of discrimination but also provides reliable probability estimates.

3-month respiratory related readmission model

The calibration plot for the 3-month respiratory related readmission model shows a more pronounced deviation from perfect calibration, particularly in the extreme probability ranges, probably because of the available sample size, where the model tends to underestimate the likelihood of respiratory-related readmissions. This suggests that the model may benefit from further calibration to improve its reliability in predicting these events accurately. The Brier score indicates a slightly higher error in the predictions compared to the other models, reflecting the observed calibration issues. However, despite these challenges, the model demonstrates good utility in the lower probability ranges, where its predictions align more closely with observed outcomes, making it potentially useful in identifying patients at lower risk for respiratory-related readmissions.

ROC and PR curves

12-Month Mortality Model

The ROC curve for the 12-month mortality model shows an AUC of 0.842, reflecting moderate discriminative ability. This indicates that the model can distinguish between patients who will and will not experience mortality within 12 months. The PR curve, with an AUC of 0.365, indicates that the model has limitations in maintaining high precision, especially in the context of imbalanced data where the number of positive cases (mortality) is much smaller than the number of negative cases. This suggests that while the model can identify true positives, it may also yield a higher number of false positives, particularly at lower threshold settings.

3-month all cause readmission model

The ROC curve for the 3-month all cause readmission models shows an AUC of 0.770, indicating moderate-to-strong discriminative ability. This suggests that the model is effective in distinguishing between patients at risk of all-cause readmission and those who are not. The PR curve with an AUC of 0.766, reflects the model's robust performance in maintaining high precision while accurately identifying true positives. This high AUC-PR value indicates that the model is particularly useful in contexts where it is critical to correctly identify high-risk patients without a significant increase in false positives.

3-month respiratory related readmission model

The ROC curve for the 3-month respiratory related readmission model, with an AUC of 0.722, indicates a similar level of discriminative ability as the 12-month mortality model. This suggests that the model can moderately differentiate between patients who are at risk of respiratory-related readmission and those who are not. The PR curve, with an AUC of 0.615, shows that the model performs reasonably well in identifying true positives while maintaining a balance with precision.

Decision curve analysis

12-month mortality model

The decision curve for the 12-month mortality model demonstrates that it provides a significant net benefit across a wide range of threshold probabilities compared to both treating all patients and treating no patients. This indicates that the model is effective in identifying patients at varying levels of risk for mortality within 12 months, offering substantial clinical value in making informed decisions about interventions.

3-month all cause readmission model

The decision curve for the 3-month all cause readmission model exhibits a net benefit, above the threshold 0.2. This suggests that the model could support clinical decision-making, especially in identifying patients who are at a general risk of readmission. The net benefit remains higher than both extremes (treating all or no patients), making this model potentially useful in clinical practice.

3-month respiratory related readmission model

The decision curve for the 3-month respiratory related readmission model shows a more modest net benefit. The curve indicates that the model provides a net benefit primarily at thresholds of 0.2 or greater, which is useful in scenarios where it is crucial to identify as many high-risk patients as possible, even if this comes with a higher number of false positives.

Impact of threshold changes

The selection of a decision threshold is crucial in balancing the trade-offs between false positives and false negatives. As seen in the provided confusion matrix values for various thresholds, altering the threshold significantly impacts these metrics. The choice of threshold should be guided by the specific clinical context and the relative importance of avoiding false negatives versus false positives. Lower thresholds are more appropriate for scenarios where missing a high-risk patient is unacceptable, while higher thresholds are better suited for prioritising those patients most likely to benefit from intervention, particularly in resource-constrained environments. The data clearly illustrate that there is no one-size-fits-all threshold; rather, the optimal choice depends on the clinical objectives and the resources available for patient care. The Lenus Health team could consider using two different thresholds to optimise the models' utility and maximise their performance. A lower threshold could be used to identify patients for whom no intervention is necessary, while a higher threshold could be employed to select patients who would benefit from intervention.