Predicting outcomes after hospitalisation for COPD exacerbation using machine learning

Chih-Ying Wu[1], Chien-Ning Hsu[2,3], Charlotte Wang[4], Jung-Yien Chien[5], Chi-Chuan Wang[1,6,7] and Fang-Ju Lin[1,6,7]

[1]Graduate Institute of Clinical Pharmacy, College of Medicine, National Taiwan University, Taipei, Taiwan. [2]School of Pharmacy, College of Pharmacy, Kaohsiung Medical University, Kaohsiung, Taiwan. [3]Department of Pharmacy, Kaohsiung Chang Gung Memorial Hospital, Kaohsiung, Taiwan. [4]Institute of Health Data Analytics and Statistics, College of Public Health, National Taiwan University, Taipei, Taiwan. [5]Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan. [6]School of Pharmacy, College of Medicine, National Taiwan University, Taipei, Taiwan. [7]Department of Pharmacy, National Taiwan University Hospital, Taipei, Taiwan.

Corresponding author: Fang-Ju Lin (fjilin@ntu.edu.tw)

Shareable abstract (@ERSpublications)
Machine learning models significantly enhance the prediction of early readmission and mortality following COPD exacerbations. The identification of critical prognostic factors could facilitate more effective post-discharge care. https://bit.ly/3YqKQIs

Cite this article as: Wu C-Y, Hsu C-N, Wang C, *et al*. Predicting outcomes after hospitalisation for COPD exacerbation using machine learning. *ERJ Open Res* 2025; 11: 00651-2024 [DOI: 10.1183/23120541.00651-2024].

Received: 28 June 2024
Accepted: 6 Oct 2024

Abstract

Background Early readmission and death are critical adverse outcomes following hospitalisation due to exacerbation of chronic obstructive pulmonary disease (ECOPD). This study aimed to develop and validate machine learning models to enhance the prediction of these outcomes after ECOPD hospitalisation.

Methods Utilising a nationwide database, data from the index ECOPD hospitalisation and the preceding year were collected. Prediction models for 30-day readmission and death were developed using logistic lasso regression, random forest, extreme gradient boosting (XGBoost) and neural network, with the LACE index serving as a reference. Model performance was assessed with receiver operating characteristic (ROC) curves and calibration plots from the validation dataset. Key predictors were identified using SHapley Additive exPlanations.

Results The study included 101 011 hospitalisations in the development dataset and 17 565 in the validation dataset. The rates of 30-day readmission and death were 29.1% and 4.3%, respectively. XGBoost outperformed other models, achieving an area under the ROC curve of 0.721 (95% CI 0.713–0.729) for readmission and 0.809 (95% CI 0.794–0.824) for death, both exceeding the corresponding values for the LACE index (0.651 and 0.641). All machine learning models demonstrated good calibration. The number of hospitalisations in the previous year and the lowest haemoglobin level during the index hospitalisation were the top predictors of readmission and death, respectively.

Conclusions Applying machine learning techniques to large-scale data effectively improves the prediction of early readmission and death following ECOPD hospitalisation. Identifying critical prognostic factors could enhance targeted post-discharge care for this high-risk patient group.

Introduction

COPD significantly contributes to global morbidity and mortality [1]. Exacerbations of COPD (ECOPD), characterised by a sudden worsening of symptoms, frequently result in hospitalisations and are associated with increased systemic inflammation, escalating the risk of subsequent cardiovascular and respiratory complications. ∼18–20% of ECOPD hospitalisations lead to readmissions within 30 days [2, 3], and the post-discharge 30-day mortality rate ranges from 7% to 9% [4]. These events often precipitate accelerated declines in lung function, decreased quality of life and increased mortality.

Although early readmissions and deaths after ECOPD hospitalisation are potentially preventable, identifying the risk factors remains a challenge [5–12]. Traditional studies predominantly utilise logistic regression to explore a variety of patient, provider and system-related factors, yet these studies often show only moderate predictive performance and lack comprehensive validation [9, 10, 13, 14]. Machine learning approaches have recently gained prominence for their superior ability to analyse complex datasets and relationships [15], demonstrating enhanced predictive accuracy over traditional methods in various medical conditions [16–18]. However, only some studies have applied these advanced techniques to ECOPD [11, 12].

Recognising the gap, this study was designed to develop and validate machine learning models using comprehensive data from Taiwan's healthcare system, aiming to enhance the prediction of early readmission and death after ECOPD hospitalisations. By leveraging advanced analytics, this study seeks to improve post-discharge planning and ultimately patient outcomes.

Material and methods

An expanded version of the study methods is provided in the supplementary material.

Data source

This study utilised data from Taiwan's National Health Insurance Research Database (NHIRD) and individual patients' linked laboratory data between January 2014 and December 2019, which includes extensive healthcare data on over 99% of the population. The individuals' linked laboratory tests were uploaded by healthcare institutions beginning in January 2015, with a coverage rate of >80% for all test results [19]. Additionally, environmental data on weather and air quality were included.

Study cohort

We identified all hospitalisations of patients aged 40 years and older with a primary discharge diagnosis of COPD or with a primary diagnosis of respiratory failure and a secondary diagnosis of ECOPD between January 2015 and November 2019 (supplementary figure S1). The algorithm for identifying hospitalisation for ECOPD was based on readmission measures defined by the Centers for Medicare and Medicaid Services (CMS) [20], accommodating the transition from ICD-9-CM to ICD-10-CM coding systems in Taiwan in January 2016. Exclusions were applied to hospitalisations involving in-hospital mortality, terminal illness at discharge, missing discharge dates, or unreliable readmission data. Additionally, only hospitalisation records with available laboratory results were included. Transfers between hospitals were treated as a single hospitalisation.

Study outcome

The outcome measures of this study were all-cause readmission and death within 30 days following an index discharge from hospitalisation for ECOPD. For readmission, we followed the definition used in CMS readmission measures, excluding readmissions planned for procedures such as transplant surgery, maintenance chemotherapy, rehabilitation and other scheduled treatments. The causes of readmission were grouped according to the Clinical Classifications Software (CCS) categories [6]. For death events, patients were classified as deceased if they were documented as such in the beneficiary registry within the 30-day post-discharge period.

Predictors and observation windows

Baseline characteristics were assessed using data from 1 year before or during the index hospitalisation. Candidate predictors were selected based on expert opinions and a scoping review of previous studies [14]. These included demographic characteristics, clinical data (discharge season, care setting, in-hospital laboratory results, history of ECOPD, medication usage, procedures, comorbidities, frailty) and environmental conditions (weather, air quality). Comorbidity and frailty were quantified using the Charlson Comorbidity Index (CCI) and the multimorbidity frailty index (mFI), respectively [21, 22]. For in-hospital laboratory results, analyses focused on the worst test results. These variables and their specific details are listed in supplementary table S1.

Statistical analysis

Data preparation and initial analysis

We assessed baseline characteristics of the included patients with ECOPD hospitalisations, comparing those with and without specific outcomes and between development and validation datasets. t-test, Wilcoxon rank sums test or Kolmogorov–Smirnov test was used for continuous variables, while Chi-square test or Fisher's exact test was used for categorical variables. All analyses were conducted using a two-sided approach with a significance level of 0.05.

Handling multicollinearity and missing data

Multicollinearity of predictor variables was evaluated using a correlation matrix heatmap, retaining the more clinically significant variables from pairs with a Pearson's correlation coefficient above 0.7. Variables with ⩽30% missing data were imputed using missForest [23]. For variables with >30% missing data, such as serum albumin, blood gas and C-reactive protein levels – potentially absent due to less severe conditions or oversight – we categorised the available test values and treated the missing values as the reference category.

Model development and validation

Several statistical and machine learning methods, including logistic lasso regression, random forest, artificial neural network and extreme gradient boosting (XGBoost), were utilised to predict the risk of adverse outcomes. The models incorporated all available variables, with built-in algorithms used to select relevant variables and assess their importance. Additionally, we utilised logistic regression with items from the LACE index – including length of stay, emergent admission, CCI and number of emergency department visits in the previous 6 months – as a reference model [24]. The LACE index was chosen as the reference owing to its widespread use and availability of predictors in our data sources.

The study cohort was split into the development dataset, consisting of hospital discharges between January 2015 and December 2018, and the validation dataset, comprising discharges between January 2019 and November 2019. Models were trained in the development dataset to estimate the probability of outcome occurrence, and performance was evaluated in the validation dataset. All the models were trained using a five-fold cross-validation. In addition, we applied undersampling to alleviate the imbalance in the data due to the rarity of mortality outcomes.

Model performance and prediction validity

Model discrimination was evaluated by plotting receiver operating characteristic (ROC) curves and measuring the area under the curve (AUROC). A higher AUROC indicates better model performance in distinguishing high-risk and low-risk individuals. The sensitivity and specificity at the threshold with the maximum Youden index were calculated. Furthermore, to address potential misinterpretation of the AUROC in the presence of imbalanced outcomes, we generated a precision-recall (PR) curve and evaluated its area under the curve (AUPRC). The calibration of the models was evaluated using a calibration plot, which divides the predicted risks into deciles and calculates the mean observed risk within each decile.

Examination of variable importance

The SHapley Additive exPlanations (SHAP) method was used to determine variable importance within the top-performing machine learning model [25]. SHAP offers a unified solution for accurately estimating the impact of each variable by assessing its contribution across all possible combinations of other variables. The SHAP summary plot displays the most critical variables along with their respective SHAP values, and the partial SHAP dependence plot visualises how the attributed importance of a variable changes with varying values.

Sensitivity analyses

To verify the clinical relevance of the features in patients with COPD, an extensive analysis was conducted on the subset of patients enrolled in the nationwide COPD pay-for-performance (P4P) programme, incorporating additional data linked from the programme registry. The additional covariates included body mass index (BMI), smoking status, Global Initiative for Obstructive Lung Disease (GOLD) severity grade, modified Medical Research Council dyspnoea score, COPD assessment test score, and forced expiratory volume in 1 s after bronchodilator, with data points closest to the discharge date being used. Further sensitivity analyses involved exploring data-driven features to examine the granularity of the predictors included. We grouped diagnoses and medications that shared the first three digits of their codes (specifically using Anatomical Therapeutic Chemical (ATC) codes for medications), and items with a prevalence >2% were added to the candidate predictors. Moreover, the baseline covariate collection window was reduced from 1 year to 6 months. All sensitivity analyses were performed using the best-performing machine learning model from the main analysis.

Statistical analysis was conducted using SAS Enterprise Guide (version 7.1) and R (version 4.0.3).

Results

Cohort characteristics

A total of 118 576 ECOPD hospitalisations between January 2015 and November 2019 were included after applying the inclusion and exclusion criteria (supplementary figure S2). Throughout this period, the 30-day all-cause readmission rate ranged from 27.6% to 29.1%, and the 30-day all-cause mortality rate ranged from 4.2% to 4.3%. The daily number of readmission events decreased over time after discharge, as shown in supplementary figure S3. Of all the readmissions, 36.3% were due to re-exacerbation, while pneumonia, respiratory failure, septicaemia and urinary tract infection were other common causes (supplementary table S2), defined according to the CCS categories. A comparison of the characteristics between individuals with and without outcome occurrence is presented in supplementary table S3.

The development dataset included 101 011 hospitalisations (21.5% female, median age 76 years), and the validation dataset included 17 565 hospitalisations (20.7% female, median age 75 years). In the development cohort, 29 380 (29.1%) early readmissions and 4298 (4.3%) early deaths occurred. In the validation cohort, 4840 (27.6%) early readmissions and 734 (4.2%) early deaths occurred. The comprehensive demographic and clinical profiles for both datasets are summarised in table 1.

Model performance: early readmission

After assessing multicollinearity through a heatmap (supplementary figure S4), 235 factors were retained as candidate predictors. XGBoost demonstrated the highest performance in both the development and validation datasets. Specifically, in the validation dataset, XGBoost achieved an AUROC of 0.721 (95% confidence interval (CI) 0.713–0.729), and the reference model yielded the lowest AUROC (0.651; 95% CI 0.642–0.660). Figure 1a displays the ROC curves for all the machine learning and reference models, while table 2 shows their discriminatory capabilities. Supplementary figure S5 (left) illustrates the PR curves and AUPRC, revealing that all the machine learning models outperformed the reference model.

Calibration plots in supplementary figure S6–1 show that XGBoost, the best-performing model, provided well-calibrated predictions, unlike the reference model, which displayed a narrower range of predicted risks

TABLE 1 Selected baseline characteristics and predictor values for individuals in the development and validation datasets

Variables	Development dataset	Validation dataset	p-value
Individuals, n	101 011	17 565	
Age years, median (Q1–Q3)	76 (66–84)	75 (65–82)	**<0.001**
Female, n (%)	21 691 (21.5)	3634 (20.7)	0.019
Charlson comorbidity index, median (Q1–Q3)	3 (1–4)	3 (1–4)	**0.005**
Care setting, n (%)			
Medical centre	19 592 (19.4)	3548 (20.2)	**<0.001**
Regional hospital	53 596 (53.1)	8722 (49.7)	
District hospital	27 823 (27.5)	5295 (30.2)	
Top 20 important factors on early readmission, excluding age and sex, which are presented above			
Multimorbidity frailty index, median (Q1–Q3)	0.19 (0.09–0.25)	0.16 (0.09–0.25)	**<0.001**
Fit, n (%)	15 943 (15.8)	3108 (17.7)	**<0.001**
Mild, n (%)	21 499 (21.3)	3983 (22.7)	
Moderate, n (%)	22 668 (22.4)	3980 (22.7)	
Severe, n (%)	40 901 (40.5)	6494 (37.0)	
History of pneumonia, n (%)	47 587 (47.1)	8197 (46.7)	0.276
Cumulative dose of previous inhaled short-acting bronchodilator uses (DDD), median (Q1–Q3)	43.4 (0.3–180.8)	45.0 (0.3–184.8)	0.202
Presence of severe ECOPD within 30 days before admission, n (%)	18 360 (18.2)	3163 (18.0)	0.592
Highest in-hospital blood eosinophil level %, median (Q1–Q3)	1.5 (0.4–3.5)	1.6 (0.4–3.6)	**<0.001**
Lowest in-hospital haemoglobin level g·dL^{-1}, median (Q1–Q3)	12.3 (10.6–13.9)	12.6 (10.8–14.1)	**<0.001**
Highest in-hospital serum creatinine level mg·dL^{-1}, median (Q1–Q3)	1.0 (0.8–1.3)	1.0 (0.8–1.3)	**<0.001**
Highest in-hospital blood urea nitrogen level mg·dL^{-1}, median (Q1–Q3)	21.0 (15.0–30.0)	21.3 (15.7–30.3)	**<0.001**
Daily dose of in-hospital systemic corticosteroids (prednisone equivalent in mg), median (Q1–Q3)	53.8 (24.4–83.8)	56.9 (25.0–86.9)	**<0.001**
Duration of in-hospital systemic corticosteroids days, median (Q1–Q3)	6 (3–10)	7 (3–10)	**<0.001**
Daily dose of in-hospital systemic antibiotics (DDD), median (Q1–Q3)	1.2 (0.7–1.8)	1.2 (0.8–1.8)	**<0.001**
Duration of in-hospital oxygen therapy days, median (Q1–Q3)	6 (3–10)	6 (2–10)	**<0.001**
Duration of in-hospital nasogastric tube feeding days, median (Q1–Q3)	0 (0–0)	0 (0–0)	**0.002**
In-hospital diuretics use, n (%)	36 918 (36.6)	5898 (33.6)	**<0.001**
In-hospital cough suppressants use, n (%)	47 601 (47.1)	8340 (47.5)	0.383
Length of stay days, median (Q1–Q3)	9 (6–13)	9 (6–13)	**0.006**
Number of previous hospitalisations, median (Q1–Q3)	1 (0–3)	1 (0–3)	0.278
Number of previous emergency department visits, median (Q1–Q3)	1 (0–4)	1 (0–4)	0.705

Bold type for p-values denotes statistical significance. DDD: defined daily dose; ECOPD: exacerbation of chronic obstructive pulmonary disease.

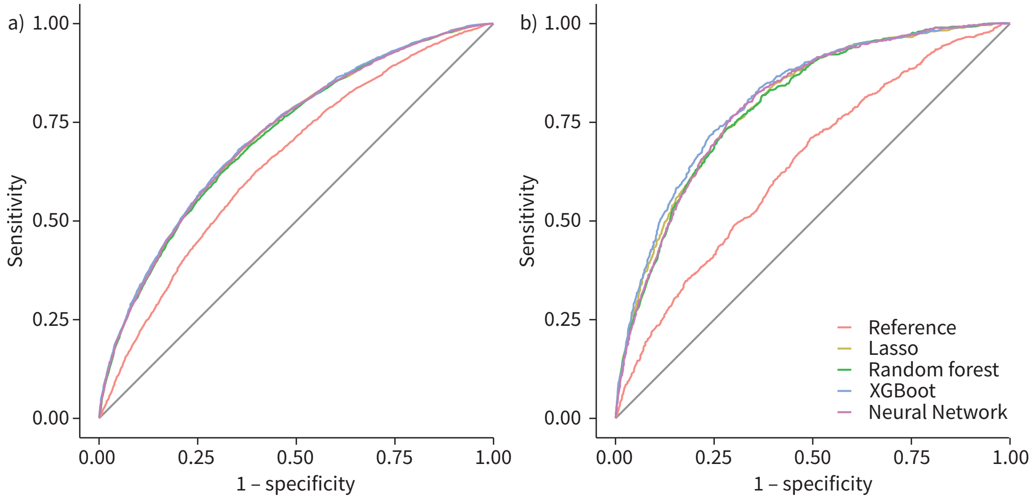

FIGURE 1 Receiver operating characteristic curves of models in the validation dataset: a) early readmission prediction; b) early death prediction.

in the validation dataset. The final hyperparameter settings for the machine learning models used to predict early readmissions are provided in supplementary table S4–1.

Model performance: early death

Like early readmission predictions, all the machine learning models outperformed the reference model in predicting early death. In the validation dataset, XGBoost demonstrated superior performance, achieving an AUROC of 0.809 (95% CI 0.794–0.824), significantly higher than the reference model, which had the lowest AUROC of 0.641 (95% CI 0.621–0.661). The ROC curves, discriminatory measures and PR curves are shown in figure 1b, table 3 and supplementary figure S5 (right), respectively.

All the machine learning models were well calibrated and had a broad range of predicted risks (supplementary figure S6–2). The final hyperparameter settings for these models, which are tailored to predicting early death, are provided in supplementary table S4–2.

Top important predictors

Figure 2 displays the 20 most important variables based on the SHAP values in the XGBoost model. The impact of each top predictor on the risk of adverse outcomes is shown in supplementary figure S7. A SHAP value exceeding zero signifies an increased risk of outcome occurrence. The number of previous hospitalisations was identified as the most significant predictor of early readmission, significantly outperforming the other factors. For early death, the lowest in-hospital haemoglobin level was the most important predictor, followed by greater age and the highest in-hospital blood urea nitrogen (BUN) and blood eosinophil percentage (EOS) levels. Among the top 20 predictors for early readmission and death, in-hospital laboratory results accounted for four and six of the factors, respectively.

Sensitivity analyses

The prediction performance decreased when additional clinical factors were included for the COPD P4P programme cohort (predicting early readmission: AUROC 0.708, 95% CI 0.675–0.742; early death: AUROC 0.716, 95% CI 0.691–0.831). Among the additional factors, BMI was the most important predictor of readmission. On the other hand, when data-driven features were added, the prediction performance marginally improved (predicting early readmission: AUROC 0.723, 95% CI 0.714–0.731; early death: AUROC 0.816, 95% CI 0.802–0.831), and previous use of perfusion solution (ATC code: B05) was the most important predictor of readmission among these features. When the analysis was repeated with a baseline covariate collection window of 6 months, the results were unchanged (predicting early readmission: AUROC 0.722, 95% CI 0.714–0.730; early death: AUROC 0.813, 95% CI 0.798–0.828). The ROC curves, calibration plots and variable importance of these sensitivity analyses are available in supplementary figures S8–S11.

TABLE 2 Discriminatory performance of machine learning models for early readmissions

Algorithm	Development dataset (2015–2018)	Validation dataset (2019)					
	AUROC# (95% CI)	AUROC (95% CI)	Balanced threshold¶ (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)
LACE logistic regression (as reference model)	0.647 (0.643–0.651)	0.651 (0.642–0.660)	0.274 (0.267–0.283)	0.626 (0.575–0.658)	0.603 (0.569–0.657)	0.375 (0.366–0.389)	0.809 (0.800–0.816)
Logistic lasso regression (LR)	0.708 (0.704–0.711)	0.717 (0.708–0.725)	0.261 (0.249–0.310)	0.681 (0.570–0.716)	0.638 (0.604–0.747)	0.417 (0.405–0.466)	0.840 (0.819–0.849)
Extreme gradient boosting (XGBoost)	**0.712 (0.709–0.715)**	**0.721 (0.713–0.729)**	**0.278 (0.271–0.312)**	**0.680 (0.606–0.698)**	**0.647 (0.632–0.719)**	**0.423 (0.416–0.455)**	**0.842 (0.826–0.848)**
Random forest (RF)	0.706 (0.703–0.710)	0.714 (0.706–0.723)	0.310 (0.279–0.314)	0.620 (0.605–0.700)	0.693 (0.612–0.705)	0.434 (0.405–0.445)	0.827 (0.822–0.844)
Artificial neural network (ANN)	0.707 (0.704–0.711)	0.717 (0.709–0.726)	0.293 (0.274–0.328)	0.680 (0.621–0.719)	0.641 (0.602–0.699)	0.419 (0.406–0.445)	0.841 (0.828–0.851)
Sensitivity analyses							
COPD-pay-for-performance participants (XGBoost, n=5932)	0.701 (0.685–0.718)	0.708 (0.675–0.742)	0.232 (0.223–0.320)	0.755 (0.529–0.796)	0.571 (0.548–0.808)	0.393 (0.378–0.497)	0.864 (0.817–0.885)
Adding data-driven features (XGBoost)	0.713 (0.710–0.717)	0.723 (0.714–0.731)	0.298 (0.262–0.313)	0.637 (0.599–0.717)	0.689 (0.611–0.725)	0.438 (0.411–0.458)	0.833 (0.825–0.851)
Changing baseline window to 6 months (XGBoost)	0.712 (0.708–0.715)	0.722 (0.714–0.730)	0.292 (0.250–0.329)	0.637 (0.568–0.734)	0.685 (0.588–0.756)	0.435 (0.402–0.435)	0.832 (0.822–0.855)

Statistics for XGBoost are presented in bold to highlight its superior performance compared to other models. AUROC: area under the receiver operating characteristic curve; PPV: positive predictive value; NPV: negative predictive value. #: performance metrics derived from cross-validation of the development dataset are reported; ¶: balanced thresholds were determined by the maximum Youden index.

TABLE 3 Discriminatory performance of machine learning models for early death

Algorithm	**Development dataset (2015–2018)[#]**	**Validation dataset (2019)**					
	AUROC (95% CI)	**AUROC (95% CI)**	**Balanced threshold[¶] (95% CI)**	**Sensitivity (95% CI)**	**Specificity (95% CI)**	**PPV (95% CI)**	**NPV (95% CI)**
LACE logistic regression (as reference model)	0.642 (0.633–0.650)	0.641 (0.621–0.661)	0.038 (0.038–0.044)	0.708 (0.522–0.744)	0.506 (0.488–0.699)	0.059 (0.057–0.068)	0.976 (0.971–0.978)
Logistic lasso regression (LR)	0.796 (0.790–0.803)	0.798 (0.783–0.813)	0.040 (0.026–0.043)	0.721 (0.691–0.860)	0.732 (0.596–0.758)	0.105 (0.084–0.113)	0.984 (0.982–0.990)
Extreme gradient boosting (XGBoost)	**0.810 (0.804–0.816)**	**0.809 (0.794–0.824)**	**0.039 (0.025–0.042)**	**0.726 (0.699–0.857)**	**0.752 (0.625–0.770)**	**0.113 (0.089–0.121)**	**0.984 (0.985–0.990)**
Random forest (RF)	0.790 (0.784–0.797)	0.794 (0.778–0.809)	0.048 (0.036–0.051)	0.726 (0.700–0.841)	0.724 (0.614–0.746)	0.103 (0.086–0.111)	0.984 (0.983–0.989)
Artificial neural network (ANN)	0.787 (0.780–0.794)	0.798 (0.783–0.813)	0.047 (0.032–0.047))	0.766 (0.737–0.843)	0.703 (0.633–0.724)	0.101 (0.089–0.108)	0.986 (0.984–0.989)
Sensitivity analyses							
COPD-pay-for-performance participants (XGBoost, n=5932)	0.720 (0.669–0.770)	0.716 (0.601–0.831)	0.052 (0.037–0.108)	0.727 (0.455–1.000)	0.649 (0.408–0.900)	0.038 (0.028–0.095)	0.992 (0.952–1.000)
Adding data-driven features (XGBoost)	0.815 (0.809–0.821)	0.816 (0.802–0.831)	0.035 (0.026–0.038)	0.768 (0.729–0.845)	0.729 (0.647–0.756)	0.110 (0.094–0.119)	0.986 (0.984–0.990)
Changing baseline window to 6 months (XGBoost)	0.808 (0.802–0.814)	0.813 (0.798–0.828)	0.043 (0.029–0.046)	0.730 (0.698–0.837)	0.755 (0.643–0.775)	0.115 (0.092–0.117)	0.985 (0.983–0.989)

Statistics for XGBoost are presented in bold to highlight its superior performance compared to other models. AUROC: area under the receiver operating characteristic curve; PPV: positive predictive value; NPV: negative predictive value. [#]: performance metrics derived from cross-validation of the development dataset are reported; [¶]: balanced thresholds were determined by the maximum Youden index.

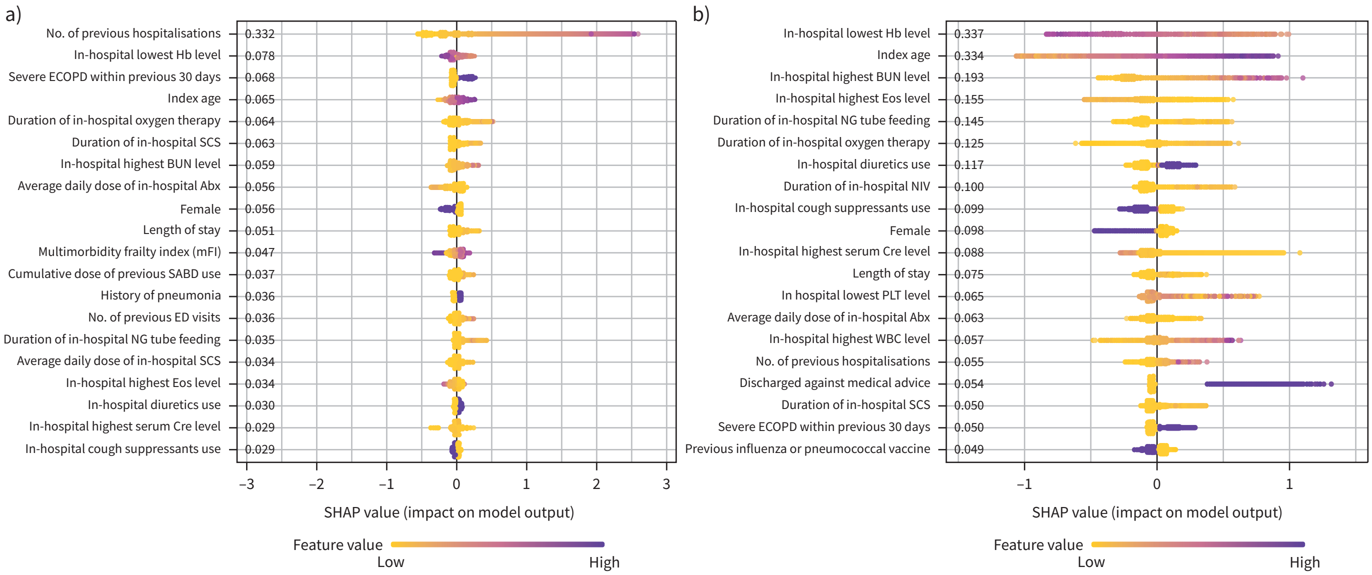

FIGURE 2 SHAP summary plots of the top 20 important factors in the XGBoost models: a) early readmission prediction; b) early death prediction. Abx: antibiotics; BUN: blood urea nitrogen; Cre: creatinine; ECOPD: exacerbation of chronic obstructive pulmonary disease; ED: emergency department; Eos: blood eosinophil; Hb: haemoglobin; NG tube: nasogastric tube; NIV: noninvasive ventilator; PLT: platelet; SABD: short-acting bronchodilators; SCS: systemic corticosteroids; SHAP: SHapley Additive exPlanations; WBC: white blood cell.

Discussion

This study utilised a nationwide administrative database to characterise the patterns of early readmission and death following 118 576 hospitalisations for ECOPD. Various machine learning approaches were employed to assess risk based on routinely collected data, and these methods helped us address complex interactions and nonlinearity among predictors that traditional statistical methods typically fail to capture [15]. A comprehensive set of predictors was included, particularly medication use, laboratory results and environmental data, which may have yet to be considered in previous studies. A temporal validation was performed using data from the same source but collected at different times. Our findings showed that XGBoost was the best model for predicting both early readmission and death, outperforming the LACE index with its greater discriminatory ability. Notably, the number of hospitalisations in the previous year and the lowest haemoglobin level during hospitalisation emerged as the most critical predictors for early readmission and death, respectively.

In our analysis, the XGBoost model achieved an AUROC of 0.721 for predicting early readmission and 0.809 for early death, both superior to the reference model. These results highlight the limitation of the LACE index, which, with only four scoring items, may not fully capture the complexities of patient conditions or include COPD-specific information. Moreover, our machine learning models demonstrated potentially greater predictive accuracy than previous works. Established clinical tools like CODEX and PEARL, designed to predict readmission or death following ECOPD hospitalisation, reported AUROCs of 0.65 and 0.70 for 30-day outcomes, respectively [9, 10]. However, these tools depend on pulmonary function tests or clinical symptom scores, which are not routine measurements. Previous studies using administrative data with logistic regression reported AUROCs ranging from 0.636 to 0.717 for 30-day readmission [5, 7], while other studies employing machine learning approaches yielded AUROCs between 0.61 and 0.72 [11, 12, 26–28].

The machine learning models developed for predicting early readmission in our study demonstrated suboptimal predictive ability compared to those for predicting early death. Similar findings were demonstrated by Frizzell *et al.* [29] in a study of 56 477 Medicare patients, which revealed no improvement in discriminatory ability (AUROC 0.62) for 30-day readmission after hospitalisation for heart failure despite using machine learning approaches. Unlike predictions related to early deaths, early readmissions are more likely influenced by a variety of factors, including socioeconomic and psychosocial elements. For instance, positive social support from family members has been linked to reduced readmissions [30]. COPD-specific factors, such as inhaler technique and pulmonary rehabilitation, can also impact readmissions [31]. Further research into these multifaceted and sometimes intangible factors will enhance our understanding of readmission dynamics.

This study pinpointed several predictors for early readmission and death among COPD patients. Frequent previous hospitalisations emerged as the most critical factor for early readmission, aligning with previous studies indicating that a history of COPD-related and non-COPD-related hospitalisations increases the risk of 30-day readmission by 53–56% and 60–64%, respectively [32, 33]. Notably, this study is the first to include nationwide laboratory tests in assessing variable importance and identified the lowest haemoglobin level during hospitalisation as a predictor for both early readmission and death. COPD patients with anaemia may have lower oxygen-carrying capacity and be more susceptible to dyspnoea, leading to exacerbations and a greater risk of hospitalisation and mortality [34, 35]. Hence, monitoring haemoglobin levels could be a cost-effective strategy for early intervention. Furthermore, higher in-hospital BUN levels were consistently among the top three predictors of short-term mortality, suggesting an association with worse outcomes in pulmonary disease, although the underlying mechanism remains unclear [36]. Interestingly, high EOS levels were associated with decreased early death risk, possibly due to a better response to corticosteroid treatment for eosinophilic exacerbations [37], while exacerbations with low EOS levels may indicate severe bacterial infection, leading to poorer prognosis [38].

Sensitivity analyses were conducted to evaluate the impact of employing different study approaches and including additional factors in the prediction. However, when additional clinical factors were added to the prediction model for participants in the COPD P4P programme, the prediction performance worsened, possibly due to the limited sample size. It is also possible that other predictors already captured some of the information contained in these additional factors. Moreover, altering the baseline collection window did not affect the model performance, indicating that a short covariate collection window may suffice to represent patients' health status. Data-driven approaches resulted in only a marginal improvement in prediction performance. A more extensive data-driven approach, such as natural language processing of electronic health records, might provide better information and improve performance, as MIN *et al.* [39] suggested.

Our study is subject to several limitations. First, important data such as smoking history, lung function tests and COPD symptom scores are not available in the NHIRD. However, when we attempted to incorporate more clinical information through sensitivity analysis in a subset of patients, this addition did not significantly impact the prediction outcomes. Secondly, while we applied the CMS-validated readmission algorithm used in previous studies, its external validity in Taiwan remains to be determined. Additionally, like many administrative databases, the NHIRD may contain recording inaccuracies. Furthermore, external validation has not been conducted; the predictive models developed have not yet been tested in other healthcare settings, potentially limiting the generalisability of the results. To establish the robustness of these models, further validation studies in diverse healthcare settings are essential.

Conclusions

This study establishes that machine learning approaches, coupled with analysing large-scale administrative and laboratory data, can substantially improve the prediction of early readmission and death following ECOPD hospitalisation. It also identifies critical prognostic factors that could shape targeted interventions to enhance post-discharge care. In the future, it is essential to implement these models within clinical settings to rigorously evaluate their real-world effectiveness in reducing readmission and mortality rates among this at-risk patient cohort.

Acknowledgement: We are grateful for the data provided by the Applied Health Research Data Integration Service from the National Health Insurance Administration in Taiwan. This study was presented in part at the 38th International Conference on Pharmacoepidemiology and Therapeutic Risk Management in Copenhagen, Denmark, in August, 2022.

Data availability: This study was conducted in part using data obtained from the national claims database, laboratory databases and the COPD pay-for-performance (P4P) programme database, all maintained by the National Health Insurance Administration in Taiwan. These data cannot be made publicly available due to legal restrictions.

Provenance: Submitted article, peer reviewed.

Ethics statement: This study was approved by the Research Ethics Committee of the National Taiwan University Hospital (number 202007082W). Informed consent was waived due to the retrospective nature of the study and the use of deidentified data. This research was conducted in compliance with the ethical standards of the Declaration of Helsinki and the institutional guidelines of the Research Ethics Committee.

Author contributions: C-Y. Wu and F-J. Lin were granted complete access to all the data within the study, ensuring the integrity and accuracy of the data analysis. They were both actively involved in the drafting and critical editing of the manuscript. Additionally, C-Y. Wu, C-N. Hsu, C. Wang, J-Y. Chien, C-C. Wang and F-J. Lin contributed significantly to the study's conceptualisation and design, the interpretation of the data, and the iterative refinement of the manuscript.

Conflict of interest: The authors have nothing to disclose.

Support statement: This work was partially supported by Taiwan's National Science and Technology Council (grant number 111–2628-B-002–026-MY3) and National Taiwan University (grant number: NTU-CDP-111L7776). The funders had no role in the study design, data collection, analysis, result interpretation, publication decision or manuscript preparation. Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global strategy for the diagnosis, management and prevention of COPD. 2023. Date last accessed: 8 December 2023. www.goldcopd.org

2 Myers LC, Faridi MK, Hasegawa K, *et al.* The hospital readmissions reduction program and readmissions for chronic obstructive pulmonary disease, 2006–2015. *Ann Am Thorac Soc* 2020; 17: 450–456.

3 Ruan H, Zhao H, Wang J, *et al.* All-cause readmission rate and risk factors of 30- and 90-day after discharge in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis. *Ther Adv Respir Dis* 2023; 17: 17534666231202742.

4 Chang CL, Sullivan GD, Karalus NC, *et al.* Predicting early mortality in acute exacerbation of chronic obstructive pulmonary disease using the CURB65 score. *Respirology* 2011; 16: 146–151.

5 Sharif R, Parekh TM, Pierson KS, *et al.* Predictors of early readmission among patients 40 to 64 years of age hospitalized for chronic obstructive pulmonary disease. *Ann Am Thorac Soc* 2014; 11: 685–694.

6 Shah T, Churpek MM, Coca Perraillon M, *et al.* Understanding why patients with COPD get readmitted: a large national study to delineate the Medicare population for the readmissions penalty expansion. *Chest* 2015; 147: 1219–1226.

7 Yu TC, Zhou H, Suh K, *et al.* Assessing the importance of predictors in unplanned hospital readmissions for chronic obstructive pulmonary disease. *Clinicoecon Outcomes Res* 2015; 7: 37–51.

8 Jo YS, Rhee CK, Kim KJ, *et al.* Risk factors for early readmission after acute exacerbation of chronic obstructive pulmonary disease. *Ther Adv Respir Dis* 2020; 14: 1753466620961688.

9 Almagro P, Soriano JB, Cabrera FJ, *et al.* Short- and medium-term prognosis in patients hospitalized for COPD exacerbation: the CODEX index. *Chest* 2014; 145: 972–980.

10 Echevarria C, Steer J, Heslop-Marshall K, *et al.* The PEARL score predicts 90-day readmission or death after hospitalisation for acute exacerbation of COPD. *Thorax* 2017; 72: 686–693.

11 Goto T, Jo T, Matsui H, *et al.* Machine learning-based prediction models for 30-day readmission after hospitalization for chronic obstructive pulmonary disease. *COPD* 2019; 16: 338–343.

12 Cavailles A, Melloni B, Motola S, *et al.* Identification of patient profiles with high risk of hospital re-admissions for acute COPD exacerbations (AECOPD) in France using a machine learning model. *Int J Chron Obstruct Pulmon Dis* 2020; 15: 949–962.

13 Njoku CM, Alqahtani JS, Wimmer BC, *et al.* Risk factors and associated outcomes of hospital readmission in COPD: a systematic review. *Respir Med* 2020; 173: 105988.

14 Alqahtani JS, Njoku CM, Bereznicki B, *et al.* Risk factors for all-cause hospital readmission following exacerbation of COPD: a systematic review and meta-analysis. *Eur Respir Rev* 2020; 29: 190166.

15 Doupe P, Faghmous J, Basu S. Machine learning for health services researchers. *Value Health* 2019; 22: 808–815.

16 Hsieh MH, Lin SY, Lin CL, *et al.* A fitting machine learning prediction model for short-term mortality following percutaneous catheterization intervention: a nationwide population-based study. *Ann Transl Med* 2019; 7: 732.

17 Ju C, Zhou J, Lee S, *et al.* Derivation of an electronic frailty index for predicting short-term mortality in heart failure: a machine learning approach. *ESC Heart Fail* 2021; 8: 2837–2845.

18 MacKay EJ, Stubna MD, Chivers C, *et al.* Application of machine learning approaches to administrative claims data to predict clinical outcomes in medical and surgical patient populations. *PLoS One* 2021; 16: e0252585.

19 Lee PC, Kao FY, Liang FW, *et al.* Existing data sources in clinical epidemiology: the Taiwan National Health Insurance Laboratory Databases. *Clin Epidemiol* 2021; 13: 175–181.

20 Yale New Haven Health Services Corporation/Center for Outcomes Research & Evaluation. Measures updates and specifications report hospital-level 30-day risk-standardized readmission measures, 2013. Date last

accessed: 14 March 2021. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/Downloads/Chronic-Obstructive-Pulmonary-Disease-COPD-Readmission-Updates.zip

21 Glasheen WP, Cordier T, Gumpina R, *et al.* Charlson Comorbidity Index: ICD-9 update and ICD-10 translation. *Am Health Drug Benefits* 2019; 12: 188–197.

22 Wen YC, Chen LK, Hsiao FY. Predicting mortality and hospitalization of older adults by the multimorbidity frailty index. *PLoS One* 2017; 12: e0187825.

23 Stekhoven DJ, Bühlmann P. MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012; 28: 112–118.

24 van Walraven C, Dhalla IA, Bell C, *et al.* Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *CMAJ* 2010; 182: 551–557.

25 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *In:* Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, Long Beach, CA, USA.

26 Amalakuhan B, Kiljanek L, Parvathaneni A, *et al.* A prediction model for COPD readmissions: catching up, catching our breath, and improving a national problem. *J Community Hosp Intern Med Perspect* 2012; 2: https://doi.org/10.3402/jchimp.v2i1.9915

27 Agarwal A, Baechle C, Behara R, *et al.* A natural language processing framework for assessing hospital readmissions for patients with COPD. *IEEE J Biomed Health Inform* 2018; 22: 588–596.

28 Verma VK, Lin WY. Machine learning-based 30-day hospital readmission predictions for COPD patients using physical activity data of daily living with accelerometer-based device. *Biosensors (Basel)* 2022; 12: 605.

29 Frizzell JD, Liang L, Schulte PJ, *et al.* Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA Cardiol* 2017; 2: 204–209.

30 Schultz BE, Corbett CF, Hughes RG, *et al.* Scoping review: social support impacts hospital readmission rates. *J Clin Nurs* 2022; 31: 2691–2705.

31 Puhan MA, Gimeno-Santos E, Cates CJ, *et al.* Pulmonary rehabilitation following exacerbations of chronic obstructive pulmonary disease. *Cochrane Database Syst Rev* 2016; 12: CD005305.

32 Nguyen HQ, Chu L, Amy Liu IL, *et al.* Associations between physical activity and 30-day readmission risk in chronic obstructive pulmonary disease. *Ann Am Thorac Soc* 2014; 11: 695–705.

33 Nguyen HQ, Rondinelli J, Harrington A, *et al.* Functional status at discharge and 30-day readmission risk in COPD. *Respir Med* 2015; 109: 238–246.

34 Yohannes AM, Ershler WB. Anemia in COPD: a systematic review of the prevalence, quality of life, and mortality. *Respir Care* 2011; 56: 644–652.

35 Ergan B, Ergün R. Impact of anemia on short-term survival in severe COPD exacerbations: a cohort study. *Int J Chron Obstruct Pulmon Dis* 2016; 11: 1775–1783.

36 Barakat MF, McDonald HI, Collier TJ, *et al.* Acute kidney injury in stable COPD and at exacerbation. *Int J Chron Obstruct Pulmon Dis* 2015; 10: 2067–2077.

37 Bafadhel M, McKenna S, Terry S, *et al.* Blood eosinophils to direct corticosteroid treatment of exacerbations of chronic obstructive pulmonary disease: a randomized placebo-controlled trial. *Am J Respir Crit Care Med* 2012; 186: 48–55.

38 Choi J, Oh JY, Lee YS, *et al.* The association between blood eosinophil percent and bacterial infection in acute exacerbation of chronic obstructive pulmonary disease. *Int J Chron Obstruct Pulmon Dis* 2019; 14: 953–959.

39 Min X, Yu B, Wang F. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on COPD. *Sci Rep* 2019; 9: 2362.