Predicting in-hospital death in pneumonic COPD exacerbation *via* BAP-65, CURB-65 and machine learning

Akihiro Shiroshita [1], Yuya Kimura [2], Hiroshi Shiba [3], Chigusa Shirakawa [4], Kenya Sato [5], Shinya Matsushita [5], Keisuke Tomii [4] and Yuki Kataoka [6,7,8]

[1]Dept of Respiratory Medicine, Ichinomiyanishi Hospital, Ichinomiya, Japan. [2]Dept of Respiratory Medicine, Clinical Research Center, National Hospital Organization Tokyo National Hospital, Tokyo, Japan. [3]Post Graduate Education Center, Kameda Medical Center, Kamogawa, Japan. [4]Department of Respiratory Medicine, Kobe City Medical Center General Hospital, Kobe, Japan. [5]Dept of Thoracic Medicine, Saiseikai Yokohamashi Tobu Hospital, Yokohama, Japan. [6]Dept of Internal Medicine, Kyoto Min-Iren Asukai Hospital, Kyoto, Japan. [7]Section of Clinical Epidemiology, Dept of Community Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan. [8]Dept of Healthcare Epidemiology, Graduate School of Medicine/Public Health, Kyoto University, Kyota, Japan.

Corresponding author: Akihiro Shiroshita (akihirokun8@gmail.com)

Shareable abstract (@ERSpublications)
BAP-65, CURB-65 and the XGBoost model show low predictive performance for in-hospital death in pneumonic COPD exacerbation. Further large-scale studies with more variables are warranted to develop an ideal prognostic model. https://bit.ly/3m0ISLA

Cite this article as: Shiroshita A, Kimura Y, Shiba H, *et al*. Predicting in-hospital death in pneumonic COPD exacerbation *via* BAP-65, CURB-65 and machine learning. *ERJ Open Res* 2022; 8: 00452-2021 [DOI: 10.1183/23120541.00452-2021].

Received: 7 July 2021
Accepted: 4 Dec 2021

Abstract

Introduction There is no established clinical prediction model for in-hospital death among patients with pneumonic COPD exacerbation. We aimed to externally validate BAP-65 and CURB-65 and to develop a new model based on the eXtreme Gradient Boosting (XGBoost) algorithm.

Methods This multicentre cohort study included patients aged ⩾40 years with pneumonic COPD exacerbation. The input data were age, sex, activities of daily living, mental status, systolic and diastolic blood pressure, respiratory rate, heart rate, peripheral blood eosinophil count and blood urea nitrogen. The primary outcome was in-hospital death. BAP-65 and CURB-65 underwent external validation using the area under the receiver operating characteristic curve (AUROC) in the whole dataset. We used XGBoost to develop a new prediction model. We compared the AUROCs of XGBoost with that of BAP-65 and CURB-65 in the test dataset using bootstrap sampling.

Results We included 1190 patients with pneumonic COPD exacerbation. The in-hospital mortality was 7% (88 out of 1190). In the external validation of BAP-65 and CURB-65, the AUROCs (95% confidence interval) of BAP-65 and CURB-65 were 0.69 (0.66–0.72) and 0.69 (0.66–0.72), respectively. XGBoost showed an AUROC of 0.71 (0.62–0.81) in the test dataset. There was no significant difference in the AUROCs of XGBoost *versus* BAP-65 (absolute difference 0.054; 95% CI −0.057–0.16) or *versus* CURB-65 (absolute difference 0.0021; 95% CI −0.091–0.088).

Conclusion BAP-65, CURB-65 and XGBoost showed low predictive performance for in-hospital death in pneumonic COPD exacerbation. Further large-scale studies including more variables are warranted.

Introduction

COPD is a common respiratory disease that is characterised by airflow limitation due to chronic inflammation of the airways and lungs [1]. Patients with COPD often experience acute worsening of baseline symptoms, and with coexisting consolidation (pneumonic COPD exacerbation) on chest imaging, mortality is increased compared to non-pneumonic COPD exacerbation [2]. A previous study suggested that pneumonic COPD exacerbation might have a different inflammation profile from non-pneumonic COPD exacerbation [3, 4].

CURB-65 (confusion, blood urea nitrogen >19 mg·dL^{-1}, respiratory rate ⩾30 breaths·min^{-1}, systolic blood pressure <90 mmHg or diastolic blood pressure ⩽60 mmHg and age ⩾65 years) is a simple prediction

model in patients with community-acquired pneumonia and has been validated internally and externally [5, 6]. On the other hand, BAP-65 (blood urea nitrogen $\geqslant$25 mg·dL^{-1}, altered mental status, heart rate $\geqslant$109 beats·min^{-1} and age $\geqslant$65 years) is an easily computable prediction model in patients with COPD exacerbation that has shown good performance in internal validation and external validation cohorts [7, 8]. However, we could not evaluate how many patients with pneumonic COPD exacerbation were included in those studies. Another study showed that CURB-65 had poor predictive ability for death in pneumonic COPD exacerbation [9].

To date, there is no established clinical prediction model specifically for the population with pneumonic COPD exacerbation. It is also unclear whether BAP-65 and CURB-65 can be applied to patients with pneumonic COPD exacerbation [10]. Our study had two purposes: 1) the external validation of BAP-65 and CURB-65 for predicting in-hospital death among patients with pneumonic COPD exacerbation; and 2) the development of a high-performance clinical prediction model using a modern machine learning algorithm that is gaining ground in the medical field [11].

Methods

Study design

Our study was a multicentre retrospective cohort study conducted across five acute care hospitals in Japan. To maximise patient capture, patient data were collected during different periods in each hospital between April 1, 2008, and July 31, 2020.

Pneumonic COPD exacerbation is diagnosed when the criteria for both pneumonia and COPD exacerbation are met [4, 12, 13]. To select patients with pneumonic COPD exacerbation, we used the validated algorithm based on the 10th revision of the International Classification of Diseases and Related Health Problems (supplementary eFigure S1) [4]. First, patients aged $\geqslant$40 years who had both pneumonia and COPD exacerbation were selected. Patients with other differential diagnoses mimicking pneumonic COPD exacerbation were excluded, including heart failure, pneumothorax, asthma exacerbation and obstructive pneumonia.

This study was approved by the institutional review board of each hospital (approval number 200811). This article was reported according to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement (supplementary eTable S1) [14].

Input and output data

The following input data on the day of admission were collected from the data warehouse or electronic medical records in each hospital: age, sex, the activities of daily living status (full support or not), mental status (altered mental status or not), vital signs (systolic and diastolic blood pressure, respiratory rate and heart rate), laboratory results (peripheral blood eosinophil count and blood urea nitrogen) and presence of tracheal intubation. Activities of daily living were defined as full support when the Barthel index was zero, and altered mental status was defined as a Japan Coma Scale score $\geqslant$1. These two variables are used for administrative purposes in the Japanese original case-mix classification system or Diagnosis Procedure Combination [15, 16]. The Ministry of Health, Labour and Welfare regularly evaluates the trend, quality and cost of the healthcare system using Diagnosis Procedure Combination data. We extracted data from the database containing Diagnosis Procedure Combination data submitting the anonymised patient data to the Ministry of Health, Labour and Welfare. Our variable selection was based on existing clinical prediction models of pneumonia or COPD exacerbation [5, 7, 17]. We did not collect data on other comorbidities from the Diagnosis Procedure Combination database because these variable codes were not fully externally validated. The primary outcome was in-hospital death, which was derived from the electronic medical records in each hospital.

Statistical analysis

The study process is illustrated in figure 1. Patient characteristics were summarised as means for continuous variables and as percentages for categorical variables. All statistical analyses were performed using R software version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria). The scripts are available in the GitHub repository (https://github.com/AkihiroShiroshita/Prediction-model-for-Pnuemonic-COPD-exacerbation.git).

External validation of BAP-65 and CURB-65

We conducted the external validation of BAP-65 and CURB-65 with respect to the entire dataset to evaluate their performance in a large sample size. We calculated the sensitivity and specificity using each total score as the cut-off point. To assess the calibration ability, we summarised the mortality according to

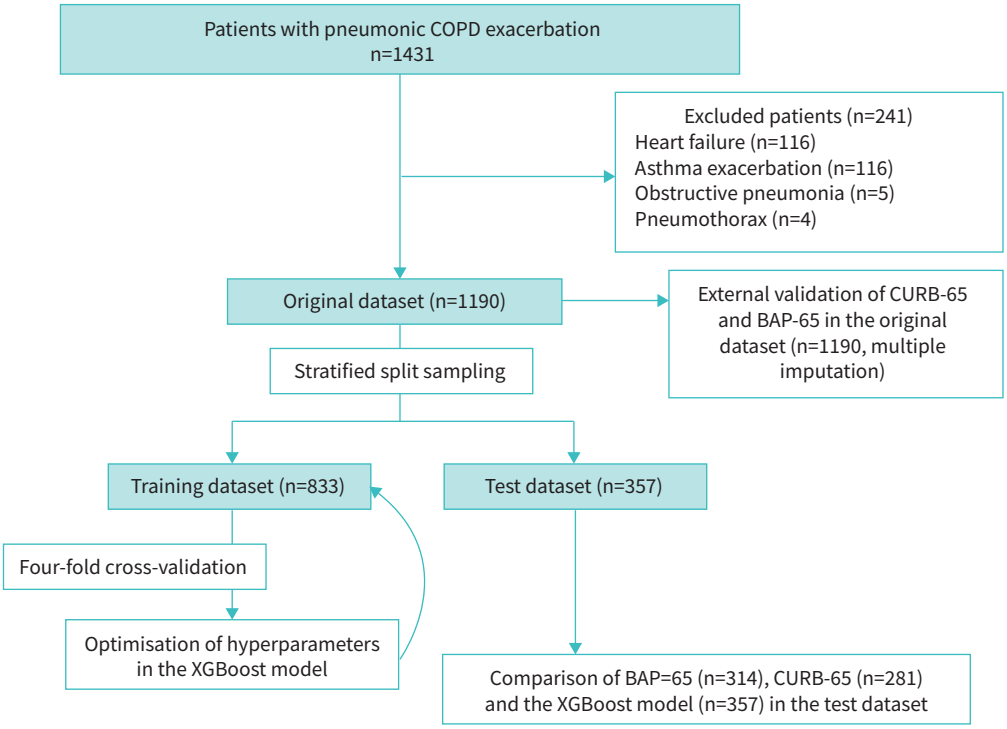

FIGURE 1 Patient selection flow and framework of the study process. XGBoost: eXtreme Gradient Boosting.

each total risk score. To assess the discriminatory ability of the two models, we calculated the area under the receiver operating characteristic curve (AUROC). We used multiple imputation to cope with missing data [18]. We created a total of 100 datasets using multiple imputation with chained equations and calculated the AUROC within each dataset. Thereafter, we combined the estimates of AUROC using Rubin's combining rule [19, 20].

Model development via machine learning

We used the eXtreme Gradient Boosting (XGBoost) algorithm to develop a clinical prediction model for in-hospital death among patients with pneumonic COPD exacerbation. The XGBoost algorithm is a powerful ensemble method of machine learning that combines a set of weak learners of the decision tree [21]. Its parallel computation enables the efficient and accurate development of a prediction model. Because it extracts variable importance, imputation of missing data, scaling or normalisation is not required. What is required in the algorithm is the proper tuning of the hyperparameters, which are parameters that control the behaviour of the model. In our study, the original data were first partitioned into training and test datasets. We used the stratified sampling method with a 7:3 ratio for data splitting, which allowed the two datasets to have similar in-hospital mortality. Second, we developed prediction models using a training dataset. We performed a grid search with four-fold cross-validation to obtain the optimal hyperparameters for maximising the mean AUROC (supplementary eFigure 2) [21]. In the grid search, the hyperparameter candidates for max_depth (maximum tree depth) was {2, 4, 6, 8, 10}, and min_child_weight (minimum degree of impurity needed in a node) was {1, 2, 3, 4, 5}. After fixing max_depth and min_child_weight, we searched the maximum number of trees based on the cross-validation. We set the remaining hyperparameters as default. Third, for external validation, we validated the trained model using the test dataset. We used the AUROC as an index to validate the model. Finally, the importance of the variables based on the impurity metric was plotted. Impurity is the degree of misclassification. It displays the degree to which each input dataset influences the output in our XGBoost model.

Model comparison

We compared the model performances of the three prediction models using the test dataset to allow comparison on a one-to-one basis. To evaluate the discriminatory performance, we described the receiver operating characteristic (ROC) curves of the three prediction models. Thereafter, we estimated the

TABLE 1 Patient characteristics

Characteristics	Survivors	Non-survivors	Total	p-values
Subjects n	1102	88	1190	
Age years	77±8	80±7	77±8	0.006
Male	974 (88)	85 (97)	1059 (89)	0.029
Full support in activities of daily living	188 (17)	23 (26)	211 (18)	0.045
Altered mental status	132 (12)	39 (44)	171 (14)	<0.001
Missing data	8 (1)	0 (0)	8 (1)	
Systolic blood pressure mmHg	133±26	125±25	132±26	0.010
Missing data	149 (14)	5 (6)	154 (13)	
Diastolic blood pressure mmHg	75±17	72±16	75±17	0.130
Missing data	155 (14)	5 (6)	160 (13)	
Respiratory rate breaths·min^{-1}	25±6	27±7	25±7	<0.001
Missing data	219 (20)	8 (9)	227 (19)	
Heart rate beats·min^{-1}	102±19	106±21	102±19	0.061
Missing data	142 (13)	3 (3)	145 (12)	
Blood urea nitrogen mg·dL^{-1}	20±11	30±21	21±12	<0.001
Missing data	13 (1)	1 (1)	14 (1)	
Blood eosinophil count·μL^{-1}	99±169	53±123	96±167	0.068
Missing data	409 (37)	41 (47)	450 (38)	

Data are presented as mean±SD or n (%) unless otherwise stated.

differences in AUROCs using bootstrap sampling (BAP-65 *versus* XGBoost, and CURB-65 *versus* XGBoost) [22]. The XGBoost model can take into account missing data while BAP-65 and CURB-65 cannot. The XGBoost model used the whole test dataset while BAP-65 and CURB-65 only used the patient data without missing values.

Results

The patient selection flowchart is shown in figure 1. We initially selected 1431 patients. After excluding 241 patients with other diagnoses, 1190 patients with pneumonic COPD exacerbation were included in our analysis. Patient characteristics are summarised in table 1. The in-hospital mortality rate was 88 out of 1190 (7%). The number of intratracheal intubations was 16 out of 1190 (1%) and median length of hospital stay was 12 (interquartile range: 8–18) days.

External validation of BAP-65 and CURB-65

Table 2 presents a summary of the number of patients with each total score. The calibration performances of both prediction models were low. The AUROC of BAP-65 was 0.69 (95% CI 0.66–0.72) and that of CURB-65 was 0.69 (95% CI 0.66–0.72). The discriminatory performance of both prediction models was also low.

TABLE 2 Risk scores and in-hospital mortality of BAP-65 and CURB-65

Risk scores	Patients n	In-hospital mortality n (%)
BAP-65 Class		
1	20	0 (0)
2	455	20 (4)
3	404	27 (7)
4	120	32 (27)
5	22	5 (23)
CURB-65		
0	21	0 (0)
1	306	16 (5)
2	351	14 (4)
3	191	31 (16)
4	57	13 (23)
5	8	2 (25)

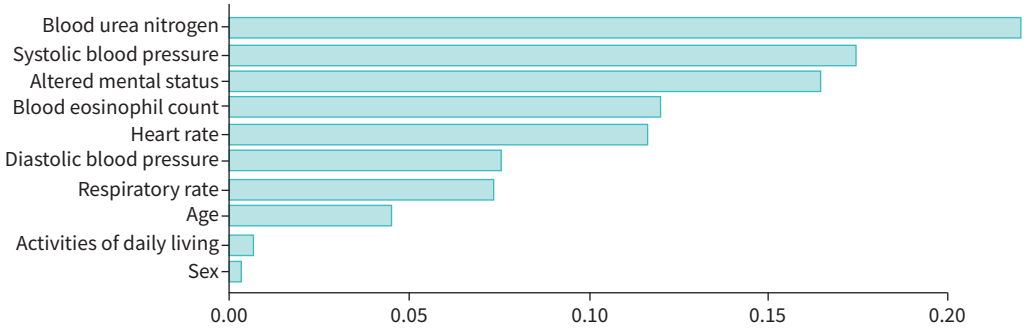

FIGURE 2 Important variables based on the impurity metric. Blood urea nitrogen was the most important feature. Activities of daily living and sex were of little importance.

Model development via machine learning

Based on the results of the grid search, we set up the hyperparameters as follows: max_depth (maximum tree depth) = 4, min_child_weight (minimum degree of impurity needed in a node) = 2, eta (learning rate) = 0.1, subsample (the proportion of cases to be randomly sampled for each tree) = 0.8, colsample_bytree (the proportion of predictor variables sampled for each tree) = 0.8, gamma (minimal loss to expand on a leaf node) = 0, lambda (L2 regularisation term on weights) = 1, alpha (L1 regularisation term on weights) = 0, scale_pos_weight (balance of positive and negative weights) = 1 and maximum number of trees = 37. Cross-validation of the developed model revealed a mean AUROC of 0.76, and external validation in the test dataset revealed an AUROC of 0.72 (95% CI 0.62–0.82). Feature importance is illustrated in figure 2, which shows that blood urea nitrogen was the most important factor for predicting in-hospital death. Systolic blood pressure and altered mental status also had important roles in the XGBoost model. On the contrary, activities of daily living and sex showed little importance.

Model comparison

We performed model comparisons using the test data. The XGBoost model used the whole test dataset (n=357), while BAP-65 and CURB-65 used the data of 314 and 281 patients, respectively, because of missing values. Figure 3 shows the ROC curves of the BAP-65, CURB-65 and XGBoost models. There was no significant difference in AUROCs between the XGBoost model and BAP-65 (absolute difference 0.054; 95% CI −0.057–0.16) or between the XGBoost model and CURB-65 (absolute difference 0.0021; 95% CI −0.091–0.088).

Discussion

Our study revealed that contrary to the study results for either pneumonia or COPD exacerbation, all three models (BAP-65, CURB-65 and XGBoost model) had low discriminatory ability for predicting in-hospital

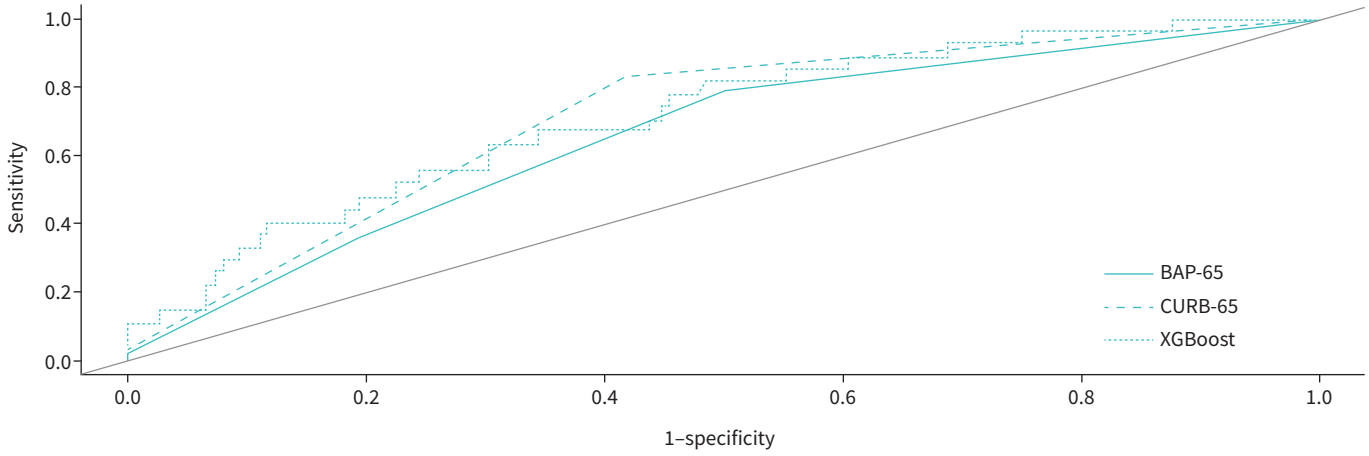

FIGURE 3 The receiver operating characteristic curves of BAP-65, CURB-65 and the eXtreme Gradient Boosting (XGBoost) model in the test dataset. The XGBoost model showed the best discriminatory performance.

death among patients with pneumonic COPD exacerbation. Further large-scale studies are needed to develop a specific clinical prediction model for pneumonic COPD exacerbation.

The two simple scoring systems, *i.e.* BAP-65 and CURB-65, showed low predictive performance in our dataset of patients with pneumonic COPD exacerbation. Although our study did not contrast their predictive abilities in either pneumonia or COPD exacerbation with pneumonic COPD exacerbation, CURB-65 was externally validated for the Japanese population and BAP-65 was validated for the Chinese population [23, 24]. Contrary to the results in either pneumonia or COPD exacerbation, CURB-65 or BAP-65 was not externally validated in our patient cohort. Our results were consistent with those of a previous retrospective cohort study that revealed CURB-65 had poor performance for predicting death in pneumonic COPD exacerbation, while it had high performance for non-pneumonic COPD exacerbation [9]. Our target population included patients with a specific category of pneumonic COPD exacerbation. The disease spectrum of pneumonic COPD exacerbation, which fulfils the diagnostic criteria for both pneumonia and COPD exacerbation, may differ from that of COPD exacerbation and pneumonia. A specific clinical prediction model for pneumonic COPD exacerbation is warranted.

A strength of our study was the use of a powerful machine learning technique that can overcome the drawbacks of the development processes of BAP-65 and CURB-65. However, the XGBoost model also had a low predictive performance for in-hospital deaths in pneumonic COPD exacerbation. Contrary to the recursive partition that was used in the development of BAP-65, the XGBoost model avoids model instability [5, 25]. In addition, unlike logistic regression, which was used in the development of CURB-65, the XGBoost model is not based on the assumption of linearity and does not require the categorisation of continuous variables [5]. It can also find the optimal interaction terms between variables. Despite its great ability, our XGBoost model did not show high performance.

However, our study had several weaknesses. First, our sample may have been too small to develop internally and externally validated prediction models. The number of events required for model development is at least 10 events per variable [26]. Our input data included 10 variables, and at least 100 events were required; however, there were only about 60 events in our training dataset. Although we used the XGBoost model, which may require a smaller sample size, we could not overcome the problem in our dataset [27]. Second, other missing variables should have been included in the model. For example, in a previous study, the DECAF score, a simple and validated scoring system for predicting outcomes in COPD exacerbation, tended to have a higher predictive performance than CURB-65 in pneumonic COPD exacerbation [9]. However, we could not collect the values for the Extended Medical Research Council Dyspnoea Scale, arterial blood gas analysis results or atrial fibrillation because they were not routinely collected in our clinical site, and they could be additional candidates for future prediction models. According to a systematic review of prediction models for COPD exacerbation, patients' baseline characteristics such as body mass index, forced expiratory volume in 1 s, and previous COPD exacerbation were used in studies with a low risk of bias [28]. These could also be additional candidates for future models.

The feature importance plot in our study highlighted the importance of blood urea nitrogen, systolic blood pressure and altered mental status. These variables should be included in a new clinical prediction model for pneumonic COPD exacerbation. On the contrary, activities of daily living and sex were of little importance in the model. In our study, the activities of daily living were categorised as full support or not, and this might have led to the loss of notable information. ∼90% of the included patients were men, which may have led to the unimportance of sex as a variable. Our study revealed some candidates for the included variables in developing a new model.

In our study, we could not conclude which clinical model was superior. In our test dataset, the number of patients was ∼300, and the in-hospital mortality was 25, which was much smaller than the necessary sample size for precise external validation [29, 30]. The results of the external validation of BAP-65 and CURB-65 in the whole dataset showed that the predictive ability for in-hospital death appeared to be similar for CURB-65 and BAP-65. Physicians who use either BAP-65 or CURB-65 will not have to change their practice based on our study.

Our study had several limitations. First, as pointed out above, our sample size was small. Because we could not include additional patients after the patient enrolment period or incorporate another patient cohort, we could not address the problem. Second, only Japanese patients were included, suggesting a lack of generalisability. Third, the primary outcome in our study was in-hospital mortality, and long-term outcomes could not be evaluated. Fourth, we could not set aside an additional dataset for external

validation before splitting the dataset. Because we performed external validation on the split dataset, the AUROC in the test dataset may have been overestimated. Fifth, altered mental status was evaluated based on the Japan Coma Scale. Although this scale has been widely used in Japan because of its simplicity, its accuracy has not been validated in patients with COPD. Sixth, we could not collect data on the patients' code status (do-not-intubate or not). In our study, the rate of tracheal intubation was lower than mortality. This may have decreased the generalisability of our study results to intensive care units. To overcome these limitations, large-scale studies from different regions are needed.

Conclusion

BAP-65, CURB-65 and the XGBoost model showed poor performance in predicting in-hospital death among patients with pneumonic COPD exacerbation. Further large-scale studies with more variables are needed to develop a new prognostic model for pneumonic COPD exacerbation.

Provenance: Submitted article, peer reviewed.

Author contributions: A. Shiroshita, K. Yuya, H. Shiba, C. Shirakawa, K. Sato, S. Matsushita, K. Tomii and L. Yuki contributed to the conception and design of the work. A. Shiroshita, H. Shiba, C. Shirakawa, K. Sato, S. Matsushita and K. Tomii contributed to data acquisition. A. Shiroshita, K. Yuya and K. Yuki contributed to the data analysis and interpretation. A. Shiroshita, K. Yuya and K. Yuki drafted the manuscript. All authors revised the manuscript critically and approved the final version of the manuscript. A. Shiroshita, K. Yuya, H. Shiba, C. Shirakawa, K. Sato, S. Matsushita, K. Tomii and K. Yuki agreed to be accountable for all aspects of any part of the work.

Data availability statement: The datasets generated and/or analysed during the current study are not publicly available due to the privacy issues but are available from the corresponding author on reasonable request.

Conflict of interest: None declared.

Support statement: Funding for English language editing was obtained from Ichinomiyanishi Hospital. The funder played no role in the study design, study execution, data analyses, data interpretation, or decision to submit the report. 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. 2020. Available from: http://goldcopd.org/ Date last accessed: 5 March 2021.
2 Saleh A, López-Campos JL, Hartl S, *et al.* The effect of incidental consolidation on management and outcomes in COPD exacerbations: data from the European COPD Audit. *PLoS ONE* 2015; 10: e0134004.
3 Huerta A, Crisafulli E, Menéndez R, *et al.* Pneumonic and nonpneumonic exacerbations of COPD: inflammatory response and clinical characteristics. *Chest* 2013; 144: 1134–1142.
4 Shiroshita A, Shiba H, Tanaka Y, *et al.* Effectiveness of steroid therapy on pneumonic chronic obstructive pulmonary disease exacerbation: a multicenter, retrospective cohort study. *Int J Chron Obstruct Pulmon Dis* 2020; 15: 2539–2547.
5 Lim WS, van der Eerden MM, Laing R, *et al.* Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax* 2003; 58: 377–382.
6 Ilg A, Moskowitz A, Konanki V, *et al.* Performance of the CURB-65 Score in predicting critical care interventions in patients admitted with community-acquired pneumonia. *Ann Emerg Med* 2019; 74: 60–68.
7 Tabak YP, Sun X, Johannes RS, *et al.* Mortality and need for mechanical ventilation in acute exacerbations of chronic obstructive pulmonary disease: development and validation of a simple risk score. *Arch Intern Med* 2009; 169: 1595–1602.
8 Shorr AF, Sun X, Johannes RS, *et al.* Validation of a novel risk score for severity of illness in acute exacerbations of COPD. *Chest* 2011; 140: 1177–1183.
9 Echevarria C, Steer J, Heslop-Marshall K, *et al.* Validation of the DECAF score to predict hospital mortality in acute exacerbations of COPD. *Thorax* 2016; 71: 133–140.
10 Trethewey SP, Hurst JR, Turner AM. Pneumonia in exacerbations of COPD: what is the clinical significance? *ERJ Open Res* 2020; 6: 00282-2019.
11 Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019; 380: 1347–1358.
12 Anthonisen NR, Manfreda J, Warren CP, *et al.* Antibiotic therapy in exacerbations of chronic obstructive pulmonary disease. *Ann Intern Med* 1987; 106: 196–204.
13 Shindo Y, Ito R, Kobayashi D, *et al.* Risk factors for drug-resistant pathogens in community-acquired and healthcare-associated pneumonia. *Am J Respir Crit Care Med* 2013; 188: 985–995.

14 Collins GS, Reitsma JB, Altman DG, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015; 162: 55–63.

15 Shigematsu K, Nakano H, Watanabe Y. The eye response test alone is sufficient to predict stroke outcome—reintroduction of Japan Coma Scale: a cohort study. *BMJ Open* 2013; 3: e002736.

16 Yasunaga H, Matsui H, Horiguchi H, *et al.* Clinical epidemiology and health services research using the diagnosis procedure combination database in Japan. *Asian Pac J Dis Manage* 2013; 7: 19–24.

17 Steer J, Gibson J, Bourke SC. The DECAF Score: predicting hospital mortality in exacerbations of chronic obstructive pulmonary disease. *Thorax* 2012; 67: 970–976.

18 White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011; 30: 377–399.

19 Toutenburg H, Rubin DB Multiple imputation for nonresponse in surveys. *Stat Pap* 1990; 31: 180.

20 Snell KI, Ensor J, Debray TP, *et al.* 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: 3505–3522.

21 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *arXiv* 2016; preprint [https://doi.org/10.1145/2939672.2939785].

22 Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med* 2000; 19: 1141–1164

23 Usui K, Tanaka Y, Noda H, *et al.* Comparison of three prediction rules for prognosis in community acquired pneumonia: Pneumonia Severity Index (PSI), CURB-65, and A-DROP. *Nihon Kokyuki Gakkai Zasshi* 2009; 47: 781–785.

24 Huang W, Cui M, Jiang Y, *et al.* A prospective validation of NEWS, CREWS and BAP-65 among patients with AECOPD. *Chin J Nursing* 2017; 12: 381–384.

25 Li R-H, Belford GG. Instability of decision tree classification algorithms. *In*: KDD '02: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, Association for Computing Machinery, 2002; pp. 570–575.

26 Riley RD, Snell KI, Ensor J, *et al.* Minimum sample size for developing a multivariable prediction model: PART II – binary and time-to-event outcomes. *Stat Med* 2019; 38: 1276–1296.

27 Floares AG, Ferisgan M, Onita D, *et al.* The smallest sample size for the desired diagnosis accuracy. *Int J Oncol Cancer Therapy* 2017; 2: 13–19.

28 Bellou V, Belbasis L, Konstantinidis AK, *et al.* Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. *BMJ* 2019; 367: 15358.

29 Vergouwe Y, Steyerberg EW, Eijkemans MJC, *et al.* Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005; 58: 475–483.

30 Steyerberg EW. Validation in prediction research: the waste by data splitting. *J Clin Epidemiol* 2018; 103: 131–133.