공학석사 학위논문

Development and validation of a deep learning model for prediction of the 30-day mortality of patients with community-acquired pneumonia from chest X-ray

지역사회 획득 폐렴 환자의 30일 사망률 예측을 위한 흉부 방사선 영상 기반 딥 러닝 모델 개발

2023년 2월

서울대학교 대학원
공과대학 협동과정 바이오엔지니어링전공
김 찬 기

Development and validation of a deep learning model for prediction of the 30-day mortality of patients with community-acquired pneumonia from chest X-ray

지도 교수 박 창 민, 최 진 욱

이 논문을 공학석사 학위논문으로 제출함

2022년 12월

서울대학교 대학원

공과대학 협동과정 바이오엔지니어링전공

김 찬 기

김찬기의 공학석사 학위논문을 인준함

2022년 12월

위 원 장	이 재 성	(인)
부위원장	박 창 민	(인)
부위원장	최 진 욱	(인)
위 원	이 정 찬	(인)

Master's Thesis

Development and validation of a deep learning model for prediction of the 30-day mortality of patients with community-acquired pneumonia from chest X-ray

지역사회 획득 폐렴 환자의 30일 사망률 예측을 위한 흉부 방사선 영상 기반 딥 러닝 모델 개발

December 2022

Interdisciplinary program in Bioengineering
The Graduate School
Seoul National University

Changi Kim

Development and validation of a deep learning model for prediction of the 30-day mortality of patients with community-acquired pneumonia from chest X-ray

December 2022

Interdisciplinary program in Bioengineering

The Graduate School

Seoul National University

Changi Kim

Confirming the master's thesis written by

Changi Kim

Chair	Jae Sung Lee	(Seal)
Vice Chair	Chang Min Park	(Seal)
Vice Chair	Jinwook Choi	(Seal)
Examiner	Jung Chan Lee	(Seal)

Abstract

In order to improve the prognosis of community-acquired pneumonia and reduce the burden of costs due to pneumonia, accurate risk prediction to take appropriate action according to the severity of each patient is important. Although indicators have been developed to predict the prognosis of related disease groups (e.g. CURB-65), there are limitations in that there is difficulty in actual use due to unsatisfactory performance or many factors included in the indicators. In this retrospective study, a DL model was developed to predict the risk of death within 30 days of the diagnosis of CAP from the initial CR, using data from patients diagnosed with CAP in a single institution between 2013 and 2019. The DL model was evaluated in consecutive patients who visited the emergency department of the same institution due to CAP between January and December 2020 (test cohort A), and two different institutions (test cohorts B and C). The discrimination of the DL model was evaluated using area under receiver operating characteristic curves (AUCs). The added value of DL model prediction to the CURB-65 score, an established risk prediction tool, was evaluated using continuous net reclassification improvement (NRI) and integrated discrimination improvement (IDI). In test cohorts A (947 patients; mean age, 71 years ± 14; 597 men), B (467 patients; mean age, 73 years ± 15; 296 men), and C (381 patients; mean age, 71 years ± 14; 243 men), the 30-day mortality rates were 18%, 8%, and 11%, respectively. The DL model exhibited AUCs of 0.77, 0.80, and 0.80 in test cohorts A, B, and C, respectively. Adding DL model prediction to the CURB-65 score improved discrimination in all external test cohorts

(continuous NRI, 0.30-0.74; IDI, 0.08-0.12). In conclusion, a deep learning-based model could predict 30-day mortality in patients with community-acquired pneumonia from chest radiographs. Adding deep learning model prediction to the CURB-65 score led to improved discrimination. Evaluation of CXRs of patients with CAP using the DL model for mortality prediction may help improve risk stratification and clinical decision-making for hospitalization or intensive care.

Keyword : Deep learning, Convolutional neural network, Community-acquired pneumonia, Chest X-ray, Survival prediction
Student Number : 2021-21892

Table of Contents

Chapter 1. Introduction

1.1. Background

Pneumonia is a potentially fatal infectious disease and a major cause of death. In 2020, 47,601 people died of pneumonia (14.4 per 100,000 population) in the United States [1]. Among infectious diseases, it was the second most common cause of death after coronavirus disease in 2020 and was the number one cause of death in 2019 [1]. In addition, it is a major burden on health resource utilization. In 2019, approximately 1.8 million people visited the emergency department due to pneumonia in the United States [2]. Community-acquired pneumonia (CAP), caused by infection outside the healthcare system, is the most common type of pneumonia [3].

Prediction of adverse outcomes in patients with CAP is essential for appropriate treatment[4-8]. Identification of high-risk patients for hospitalization and intensive treatment, including intravenous administration of antibiotics or respiratory support, may help improve patient prognosis. Furthermore, early discharge to home and conservative treatment for low-risk patients may help reduce unnecessary utilization of medical resources. In this regard, there are available tools for predicting adverse outcomes in patients with CAP based on clinical risk factors (e.g., CURB-65 score [confusion, blood urea nitrogen level, respiratory rate, blood pressure, age 65 years or older] [7] and pneumonia severity index [4]).

Chest radiography (CR) is an essential tool for the diagnosis of CAP[5,8,9]. Since most patients with CAP undergo CR at the time of diagnosis, it can be used for risk stratification. However, it has been difficult to incorporate the findings of CR in a risk prediction tool because the interpretation of CR is prone to inter-reader variability [10, 11], and it is difficult to obtain objective and quantitative biomarkers from CR for risk prediction [12, 13]. Recently, deep learning (DL) technology has been widely applied to

the evaluation of medical images, including CRs. In addition to the detection of abnormal findings or diagnosis of specific diseases, a DL algorithm can also be applied to the prediction of future events, such as adverse patient outcomes [14–16].

Therefore, this study aimed to develop a DL model to predict the 30-day mortality of patients with CAP using their initial CRs, validate the performance of the DL model in patients from different institutions, compare the performance of the model with that of an established prediction tool (CURB-65), and investigate the added value of the model to the existing prediction tool.

1.2. Purpose of Research

Therefore, the objectives of the present study were a) to develop a deep learning model to predict the 30-day mortality of patients with community-acquired pneumonia using their initial chest radiographs, b) to validate the performance of the deep learning model in patients from different institutions, c) to compare the performance of the deep learning model with that of an established prediction tool (CURB-65), and d) to investigate the added value of deep learning model to the existing prediction tool.

Chapter 2. Materials and Methods

2.1. Patient Selection

This retrospective study was approved by the institutional review boards of all participating institutions (Seoul National University Hospital [2101-175-1192], Boramae Medical Center [30-2021-127], Chung-Ang University Hospital [2203-021-19412]). The requirement for informed consent from patients was waived by the institutional review boards.

For development of a deep learning-based prediction model (DL-model), we retrospectively included patients with following inclusion criteria: a) patients diagnosed with CAP in a single tertiary-referral institution (Seoul National University Hospital; SNUH) between March 2013 and December 2019; and b) patients who underwent CR for the diagnosis of CAP (Development cohort, hereafter). For validation of the DL-model, we separately included patients with following inclusion criteria: a) patients diagnosed with CAP after vising emergency department of one tertiary-referral institution (SNUH) and two secondary-referral institutions (Boramae Medical Center and Chung-Ang University Hospital; BMC and CAUH) between January and March 2020; and b) patients who underwent CR for the diagnosis of CAP (external test cohorts A, B, and C for patients from SNUH, BMC, and CAUH, respectively). Patients without available information regarding 30-day mortality since the diagnosis of CAP were excluded from the study. Patients with multiple episodes of CAP, data for the first episode were included in the study (Figure 1).

Patients in the development cohort were randomly assigned to training, validation, and internal test datasets at a ratio of 3:1:1, for the training of the DL-model, the optimization of hyperparameters of the model, and the in-house testing of performance of the model (Figure 1).

CRs of included patients obtained at the timing of the diagnosis of CAP were retrospectively collected. CRs were obtained using

various scanners, including both fixed and portable scanners. CRs from fixed scanners were obtained in a erect position with posteroanterior projections, while CRs from portable scanners were obtained in a supine position with anteroposterior projections.

Regarding the outcome of patients, we investigated all-cause mortality within 30 days from the diagnosis of CAP. Mortality information were confirmed by electronic medical records or death registry data from the Ministry of the Interior and Safety, Republic of Korea.

As a benchmark in the evaluation of the performance of DL-model for the prediction of 30-day mortality, we used the CURB-65 score [7]. The CURB-65 scores were calculated for the patients in the internal test dataset of the development cohort and external test cohorts. Each variable of CURB-65 score (presence of new onset confusion, blood urea nitrogen level, respiratory rate, blood pressure, and age of patient) at the timing of diagnosis were retrospectively obtained from the electronic medical records.

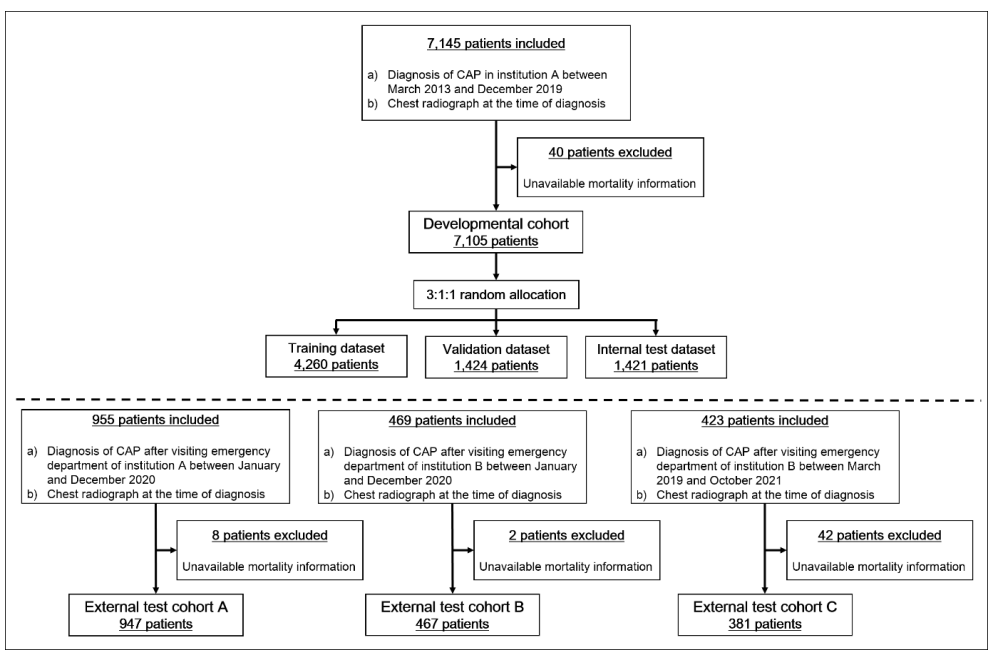

Figure 1. Overall data description

2.2. Design of Survival Prediction Model

A convolutional neural network (CNN) for the 30-day mortality prediction since CAP diagnosis from CR images was developed using CRs from patients in the developmental cohort. Before input into the CNN, the images were resized to 256×256, while maintaining the ratio of the original image with zero padding. Random brightness, random contrast, random gamma, motion blur, median blur, Gaussian noise, image flipping, and image rotation were used for data augmentation. All preprocessing algorithms were conducted in Python (version 3.6; Python Software Foundation, Del) by using the Albumentations (https://albumentations.ai/)

We adopted a previously reported CNN architecture for the survival prediction (Nnet-survival) [17]. The model adopts a negative log-likelihood loss function and incorporates non-proportional hazards. Pre-trained weights for the CNN were adopted from the CNN that can classify CRs with five different classes (normal, lung cancer, pneumonia, tuberculosis, and pneumothorax) [18]. This model was designed using DenseNet-121 backbone [19, 20]. The outputs of the CNN included conditional probabilities of survival in different time intervals. However, since the primary aim of our study was to predict 30-day mortality, the final output of the interest was the probability for the 30-day mortality (Figure 2).

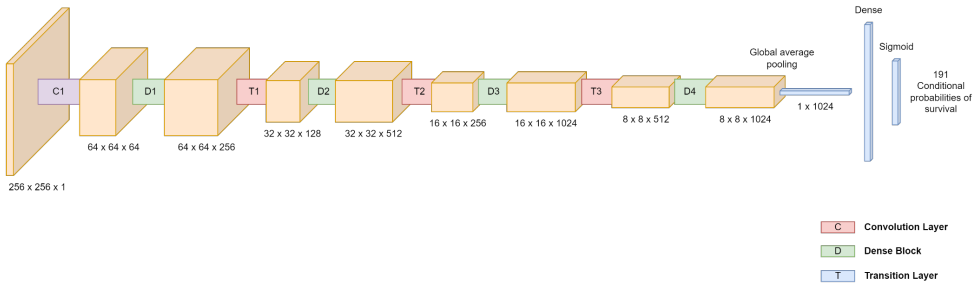

Figure 2. Architecture of the convolutional neural network for the 30-day mortality prediction since CAP diagnosis from CR images

To improve the calibration (agreement between the predicted probability and observed probability) of the DL-model, we conducted logistic recalibration of the model output in the internal test dataset [21]. Therefore, the final output from the model for the validation was recalibrated predicted probability for the 30-day mortality.

To be utilized in the clinical decision making, pre-defined cut-off value might be required. Considering that CURB-65 score of 2 or greater are considered for the criteria of hospitalization, we defined the binary cut-off value of the DL-model score to classify the same number of patients positively with the CURB-65 score of 2 or greater, in the internal test dataset.

For visualization of DL-model output, we used gradient-weighted class activation mapping (Figures 3 and 4) [22]. All codes used for the development of DL-model are available in the GitHub (github.com/Fr2zyRoom/CAP_DeepSP).

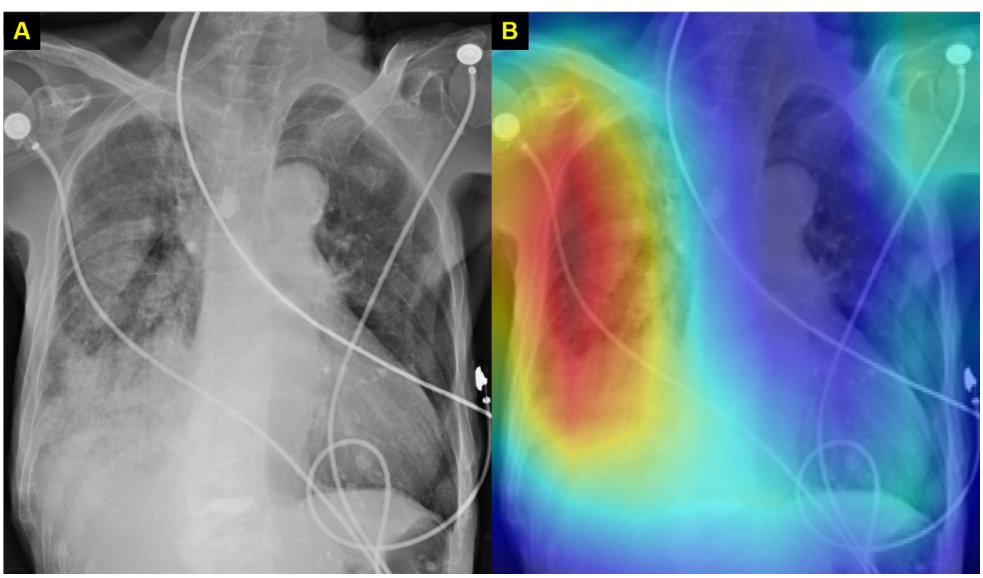

Figure 3. Gradient-weighted class activation map of DL-model output. Chest radiograph of a 78-year-old woman with CAP shows diffuse consolidation involving the right lung (A). The risk of 30-day mortality predicted by the DL model was 42%, and the gradient-weighted class activation map (B) shows that the prediction of the model was influenced by the area of pneumonia in the chest radiograph. The CURB-65 score of the patient was 2. The patient died 11 days after the diagnosis of pneumonia.

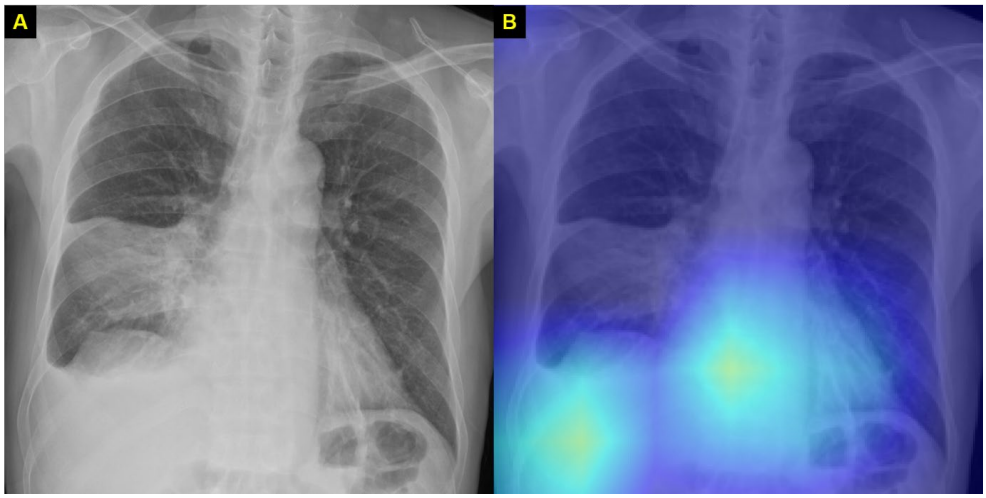

Figure 4. Gradient-weighted class activation map of DL-model output. Chest radiograph of a 69-year-old man with CAP shows an area of consolidation involving the right lower lung field with right pleural effusion (A). The risk of 30-day mortality predicted by the DL model was 9%, and the gradient-weighted class activation map (B) shows that the prediction of the model was not influenced by the area of pneumonia in the chest radiograph. The CURB-65 score of the patient was 4. The patient survived 30 days after the diagnosis of pneumonia.

2.3. Combination of CURB-65 Score and DL Model

To investigate the added value of DL-model result to the CURB-65 score, we built a logistic regression model to predict 30-day mortality using CURB-65 score and DL-model output (Combined model, hereafter), in the internal test dataset. The cut-off value for binary classification of the output of combined model was defined to classify the same number of patients positively with the CURB-65 score of 2 or greater, in the internal test dataset.

2.4. Training Environment

Hardware specification

CPU: Intel Xeon Gold 5220 2.20GHz
GPU: Tesla V100-SXM2 32GB
RAM: 16GB x 16

Software specification

Deep learning libraries:
Pytorch – 1.8.0 with cuda 10.1 and cudnn 7.6
Python libraries (version – 3.6)
Numpy – 1.19.5 for
Pandas – 1.1.5 for
OpenCV – 4.1.2
Albumentations – 1.0.3
Pycox – 0.2.3
Torchtuples – 0.2.2

2.5. Validation of Prediction Models and Statistical Analysis

To validate the DL and combined models, we applied the DL and combined models developed in the developmental cohort to the CRs from the patients in three external test cohorts, as well as the internal test dataset. To evaluate discriminative performances of prediction models to predict 30-day mortality, area under the receiver operating characteristic curves (AUCs) were used. To evaluate discriminative performance for the binary classifications, specificity, positive predictive value (PPV), and negative predictive value (NPV) were evaluated at the same sensitivity with a CURB-65 score of ≥2 (criterion for hospitalization of patients with

pneumonia) [5]. The improvement in the discriminative performance of the combined model compared to the CURB-65 score was evaluated using continuous net reclassification improvement (NRI) [23] and integrated discrimination improvement (IDI) [24]. The method suggested by DeLong et al. was used to compare the AUCs between the prediction models [25]. Sensitivities and specificities were compared using the McNemar test, while PPVs and NPVs were compared using the method suggested by Leisenring et al. [26].

To evaluate the calibration of the prediction models, we used calibration plots and Spiegelhalter's Z-test [27]. Finally, we conducted decision curve analyses to evaluate the benefit of using prediction models with different weightings between the benefit of hospitalization of high-risk patients and the cost of hospitalization of low-risk patients [28].

All statistical analyses were conducted using R (version 4.2.0, R Project for Statistical Computing, Vienna, Austria). Statistical significance was set at $p<0.05$.

Chapter 3. Results

3.1. Development Prediction Model and In-house Performance

A total of 7,105 patients (mean age, 73 years ± 15 [standard deviation]; 4417 men) who were diagnosed with CAP were included in the developmental cohort. The 30-day mortality rate in the developmental cohort was 11% (807/7105). Among the patients in the developmental cohort, 1,421 (mean age, 68 years ± 15; 882 men; 30-day mortality rate, 11% [162/1421]) were randomly assigned to the internal test dataset (Figure 1, Table 1).

Before logistic recalibration, the DL model result exhibited an AUC of 0.83 (95% confidence interval [CI], 0.80–0.87), showing a significant underestimation of mortality risk (P<.001, Spiegelhalter's Z-test) in the internal test dataset (Figure 5). Logistic recalibration of the DL model in the internal test dataset led to improved calibration (calibration slope, 1.46; calibration intercept, 1.80; P=.822, Spiegelhalter's Z-test) (Figure 5).

	Internal Test Dataset (n=1,421)*	External Test Cohort A (n=947)	External Test Cohort B (n=467)	External Test Cohort C (n=381)
Age (years)	68±15	71±14	73±15	71±14
Male patients	882 (62%)	597 (63%)	296 (63%)	243 (64%)
Chest radiographs from fixed scanner	973 (69%)	259 (27%)	144 (31%)	129 (34%)
30-day mortality	162 (11%)	167 (18%)	39 (8%)	41 (11%)
CURB-65 scores				
Score 0	291 (21%)	110 (12%)	70 (15%)	71 (19%)
Score 1	507 (36%)	271 (29%)	186 (40%)	99 (26%)
Score 2	412 (29%)	280 (30%)	145 (31%)	128 (34%)
Score 3	150 (11%)	209 (22%)	48 (10%)	61 (16%)
Score 4	42 (3%)	63 (7%)	17 (4%)	20 (5%)
Score 5	15 (1%)	14 (2%)	1 (1%)	2 (1%)

Table 1. Demographic and clinical characteristics of patients

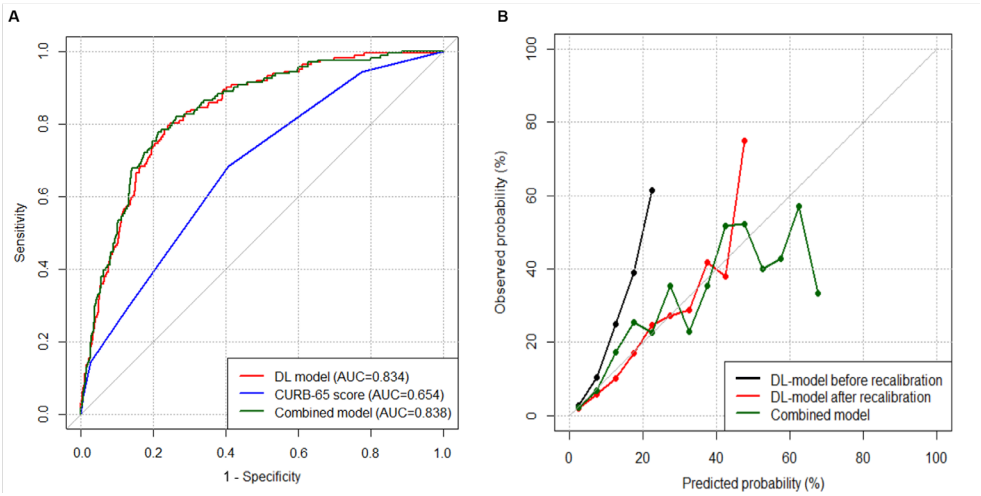

Figure 5. Receiver operating characteristic curves of DL-model, CURB-65, and combined model. Receiver operating characteristic curves obtained in the internal test dataset (A) show the DL model exhibited better discrimination (AUC, 0.83) for the prediction of 30-day mortality compared to the CURB-65 score (AUC, 0.65). The discrimination of combined model of DL-model prediction and CURB-65 score (AUC, 0.84) was better than that of the CURB-65 score and similar with that of the DL-model. Calibration plots obtained in the internal validation dataset (B) show the initial prediction of the DL-model tended to underestimate the risk of 30-day mortality, while the calibration was improved after the logistic recalibration. The combined model of DL-model prediction and CURB-65 score exhibited acceptable calibration.

After excluding four patients without CURB-65 score information, the CURB-65 score exhibited an AUC of 0.679 (95% CI, 0.64–0.72), which was significantly lower than that of the DL model (0.83, P<.001) (Table 2, Figure 5). At the same sensitivity level with a CURB-65 score ≥2 (sensitivity, 68%), the DL model exhibited higher specificity (84% vs. 59%; P<.001), PPV (35% vs. 18%; P<.001), and NPV (95% vs. 94%; P=.033) than the CURB-65 score (Table 2).

3.2. Performance of Prediction Models in Rxternal Test Cohorts

A total of 947 (male-to-female ratio, 597:350; mean age ± standard deviation, 71 ± 14 years; 30-day mortality rate, 17.6% [167/947]), 467 (male-to-female ratio, 296:171; mean age ± standard deviation, 73 ± 15 years; 30-day mortality rate, 8.4% [39/467]), and 381 (male-to-female ratio, 243:138; mean age ± standard deviation, 71 ± 14 years; 30-day mortality rate, 10.8% [41/381]) patients were included in external test cohorts A, B, and C, respectively. Table 1 show demographic and clinical characteristics of patients in external test cohorts.

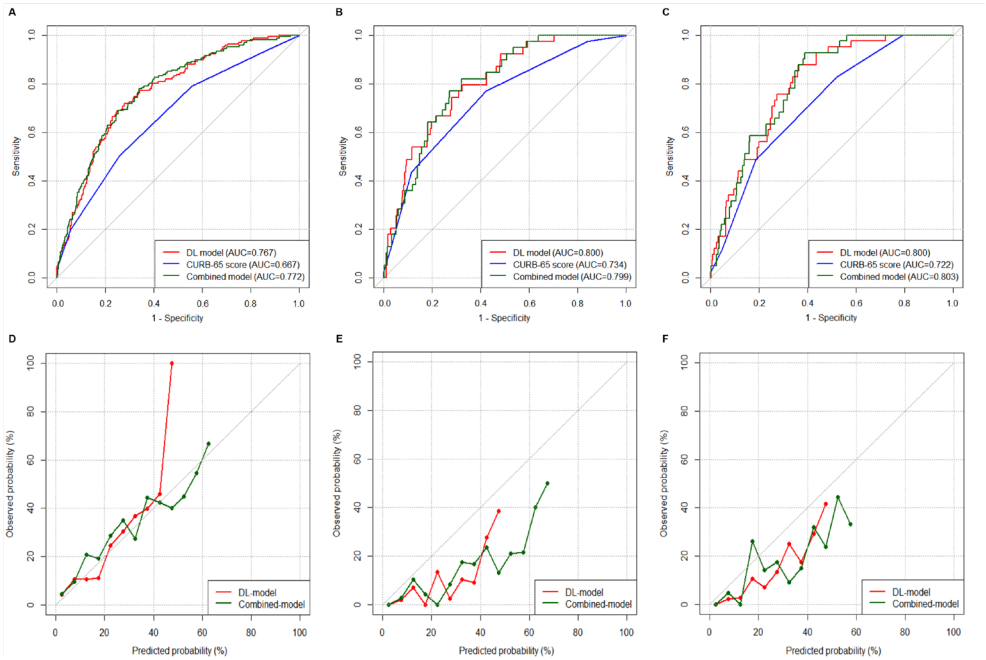

Figure 6. Receiver operating characteristic curves and calibration plots of DL-model, CURB-65, and combined model. Receiver operating characteristic curves obtained in the external test cohorts A (A), B (B), and C (C) show the DL model exhibited consistent discrimination for the prediction of 30-day mortality (AUC, 0.77–0.80). The discriminations of the DL model were better than those of CURB-65 score (AUC, 0.67–0.73). The combined model (AUC, 0.77–0.80) exhibited better discrimination compared to the CURB-65 score and similar discrimination compared to the DL-model. Calibration plots obtained in the external test cohort A (D) show acceptable calibration of the DL-model and combined model. Meanwhile in the external test cohorts B (E) and C (F), both the DL-model and the combined model overestimated the risk.

The DL model exhibited AUCs of 0.77 (95% CI: 0.73–0.81), 0.80 (95% CI: 0.74–0.86), and 0.80 (95% CI: 0.74–0.86) in external test cohorts A, B, and C, respectively (Table 2, Figure 6). In terms of calibration, the DL model exhibited fair calibration in external test cohort A (P=.159, Spiegelhalter's Z-test), while it significantly overestimated the risk of 30-day mortality in external test cohorts B and C (P<.001, Spiegelhalter's Z-test) (Figure 6).

The CURB-65 score exhibited AUCs of 0.67 (95% CI: 0.62–0.71), 0.734 (95% CI: 0.65–0.81), and 0.72 (95% CI: 0.65–0.79) in external test cohorts A, B, and C, respectively (Figure 6). The DL model exhibited higher AUCs than the CURB-65 score in the external test cohorts, while evidence of difference was found only in external test cohort A (P<.001). At the same sensitivity levels with a CURB-65 score ≥2, the DL model exhibited higher specificity and PPV than the CURB-65 score in all external test cohorts (Table 2).

	Internal Test Dataset	External Test Cohort A	External Test Cohort B	External Test Cohort C
DL model				
AUC	0.83 (0.80–0.87)	0.77 (0.73–0.81)	0.80 (0.74–0.86)	0.80 (0.74–0.86)
P–value from Spiegelhalter's Z–test	.822	.159	<.001	<.001
Sensitivity	68% (110/161) (61%, 76%)	79% (132/167) (73%, 85%)	77% (30/39) (64%, 90%)	83% (34/41) (71%, 94%)
Specificity	84% (1049/1256) (81%, 86%)	61% (476/780) (58%, 64%)	69% (295/428) (65%, 73%)	66% (225/340) (61%, 71%)
PPV	35% (110/317) (29%, 40%)	30% (132/436) (26%, 35%)	18% (30/163) (12%, 24%)	23% (34/149) (16%, 30%)
NPV	95% (1049/1100) (94%, 97%)	93% (476/511) (91%, 95%)	97% (295/304) (95%, 99%)	97% (225/232) (95%, 99%)
CURB–65 score				
AUC	0.68 (0.64–0.72)	0.67 (0.62–0.71)	0.73 (0.65–0.81)	0.72 (0.65–0.79)
P–value	<.001	<.001	.194	.081
Sensitivity	68%	79%	77% (30/39)	83% (34/41)

	(110/161)	(132/167)	(64%, 90%)	(71%, 94%)
	(61%, 76%)	(73%, 85%)		
P-value	>.999	>.999	>.999	>.999
Specificity	59%	44%	58%	48%
	(747/1256)	(346/780)	(247/428)	(163/340)
	(57%, 62%)	(41%, 48%)	(53%, 62%)	(43%, 53%)
P-value	<.001	<.001	<.001	<.001
PPV	18%	23%	14%	16%
	(110/619)	(132/566)	(30/211)	(34/211)
	(15%, 21%)	(20%, 27%)	(10%, 19%)	(11%, 21%)
P-value	<.001	<.001	.035	.002
NPV	94%	91%	96%	96%
	(747/798)	(346/381)	(247/256)	(163/170)
	(92%, 95%)	(88%, 94%)	(94%, 99%)	(93%, 99%)
P-value	.033	.118	.650	.527

Table 2. Performance of DL model and CURB-65 score

	Coefficient	Odds Ratio	*P*-Value
DL model prediction	0.08 (0.01)	1.08 (1.07-1.10)	<.001
CURB-65 score (reference: score 0)			
Score 1	0.56 (0.39)	1.76 (0.81-3.80)	.151
Score 2	0.82 (0.39)	2.27 (1.06-4.86)	.035
Score 3	0.44 (0.44)	1.55 (0.66-3.65)	.319
Score 4	1.10 (0.51)	3.00 (1.09-8.21)	.033
Score 5	2.07 (0.68)	1.08 (1.07-1.10)	.002
Intercept	−4.07 (0.36)	0.02	<.001

Table 3. Multivariate logistic regression for 30-day mortality with CURB-65 score and DL-model result

	Internal Test Dataset	External Test Cohort A	External Test Cohort B	External Test Cohort C
AUC	0.84 (0.81–0.87)	0.77 (0.73–0.81)	0.80 (0.74–0.86)	0.80 (0.75–0.86)
P–value (vs. CURB–65 score)	<.001	<.001	.164	.081
P–value (vs. DL –model)	.484	.462	.959	.702
P–value from Spiegelhalter's Z–test	.642	.003	<.001	<.001
Sensitivity	68% (110/161) (61%, 76%)	79% (132/167) (73%, 85%)	77% (30/39) (64%, 90%)	83% (34/41) (71%, 94%)
P–value (vs. CURB–65 score)	>.999	>.999	>.999	>.999
P–value (vs. DL model)	>.999	>.999	>.999	>.999
Specificity	84% (1058/1256) (82%, 86%)	64% (500/780) (61%, 67%)	73% (312/428) (69%, 77%)	65% (222/340) (60%, 70%)
P–value (vs. CURB–65 score)	<.001	<.001	<.001	<.001
P–value (vs. DL model)	.095	<.001	<.001	.317

PPV	36% (110/308) (30%, 41%)	32% (132/412) (28%, 37%)	21% (30/146) (14%, 27%)	22% (34/152) (16%, 29%)
P-value (vs. CURB-65 score)	<.001	<.001	.004	.003
P-value (vs. DL model)	.230	.013	.079	.689
NPV	95% (1058/1109) (94%, 97%)	93% (500/535) (91%, 96%)	97% (312/321) (95%, 99%)	97% (222/229) (95%, 99%)
P-value (vs. CURB-65 score)	.027	.068	.553	.544
P-value (vs. DL model)	.878	.591	.802	.963
Continuous NRI (to CURB-65 score)	0.93 (0.78, 1.09)	0.74 (0.58, 0.90)	0.30 (0.10, 0.51)	0.35 (0.10, 0.60)
IDI (to CURB-65 score)	0.13 (0.11, 0.16)	0.08 (0.06, 0.11)	0.11 (0.04, 0.18)	0.12 (0.07, 0.17)

Table 4. Performance of combined model

3.3. Added Value to the CURB-65 Score

The coefficients and odds ratios for the combined model built in the internal test dataset are listed in Table 3. Prediction by the DL model was a significant predictor of 30-day mortality (odds ratio, 1.08 for 1% increase in predicted risk [95% CI, 1.07 to 1.10]; P<.001) after adjustment for the CURB-65 score. In the internal test dataset, the combined model exhibited an AUC of 0.84 (95% CI,

0.81–0.87), which was significantly higher than that of the CURB-65 score (0.68; P<.001) and similar to that of the DL model (0.83; P=.484) (Table 4, Figure 5). At the same sensitivity level with a CURB-65 score ≥2 (sensitivity, 68%), the combined model exhibited higher specificity (84% vs. 59%; P<.001), PPV (36% vs. 18%; P<.001), and NPV (95% vs. 94%; P=.033) than the CURB-65 score (Table 4). The continuous NRI and IDI for the combined model compared with the CURB-65 score were 0.93 (95% CI, 0.78–1.09) and 0.13 (95% CI, 0.11–0.16), respectively. The combined model exhibited acceptable calibration (P=.642, Spiegelhalter's Z-test) (Table 3, Figure 5).

The combined model exhibited AUCs of 0.77 (95% CI: 0.73–0.81), 0.80 (95% CI: 0.74–0.86), and 0.80 (95% CI: 0.75–0.86) in external test cohorts A, B, and C, respectively (Figure 6). In comparison to the CURB-65 score, the combined model exhibited higher AUCs, while evidence of difference was found only in external test cohort A (P<.001). Meanwhile, DL-model and combined models exhibited similar AUCs in all external test cohorts. At the same sensitivity levels with a CURB-65 score ≥2, the combined model exhibited higher specificity and PPV than the CURB-65 score in all external test cohorts (Table 4). The combined model exhibited a significant improvement in discrimination compared to the CURB-65 score in terms of continuous NRI and IDI (Table 4).

In terms of calibration, the combined model exhibited fair calibration in external cohort A (P=.003, Spiegelhalter's Z-test) and overestimated the risk of 30-day mortality in external cohorts B and C (P<.001, Spiegelhalter's Z-test) (Figure 6).

3.4. Decision Curve Analyses

Figure 7 shows decision curves of CURB-65 score, DL-model, and combined model in internal test dataset and external test

cohorts. The DL-model and combined model exhibited higher net benefit than the CURB-65 score in internal test dataset and external test cohort A when the benefit of hospitalization of high-risk patients is greater than the cost of hospitalization of low-risk patients. In external test cohorts B and C, similar patterns of decision curves were observed, while the magnitude of improved net benefit for DL-model and combined model was only modest.

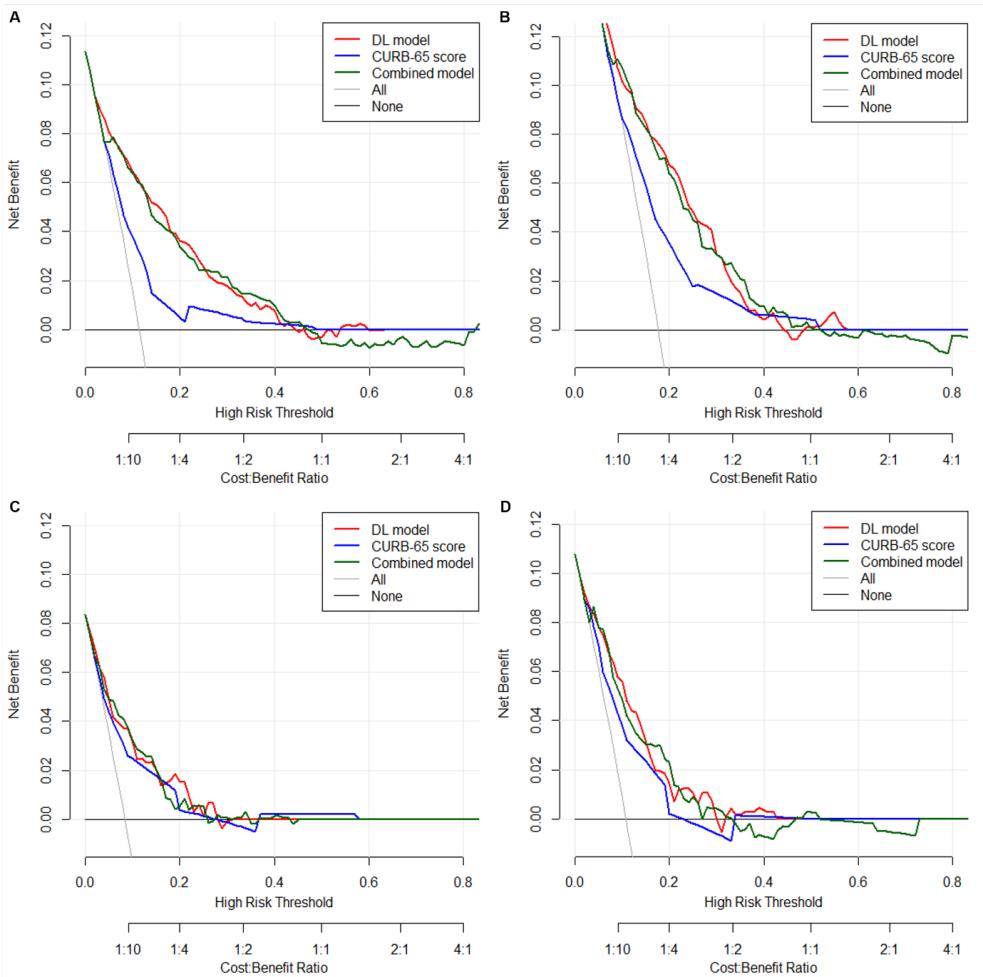

Figure 7. Decision curves of DL-model, CURB-65, and combined model. Decision curves obtained in the internal test dataset (A) and external test cohort A (B) show higher net benefit of the DL model and the combined model compared to the CURB-65 score, when the benefit of hospitalization of high-risk patient is greater than the cost of hospitalization of low risk patients. In external test cohorts B (C) and C (D), similar pattern of decision curves were observed, while the magnitude of improved net benefit for DL model and combined model was only modest.

Chapter 4. Discussion

4.1. Research Significance

Although chest radiography is crucial in the diagnosis of community-acquired pneumonia, its role in predicting the prognosis of these patients is limited. We developed a deep learning-based model for predicting 30-day mortality in patients with community-acquired pneumonia using their initial chest radiography. The prediction model exhibited robust discrimination performance in three external test cohorts (AUC, 0.77–0.80) and higher specificity (44–58% vs. 61–69%; all Ps<.001) at the same sensitivity as the CURB-65 score, an established risk prediction tool used in the daily practice. Finally, the combination of deep learning-based risk and the CURB-65 score led to improved discrimination compared to the CURB-65 score (continuous net reclassification improvement, 0.30–0.74; integrated discrimination improvement, 0.08–0.12).

Radiographic findings of CAP may provide prognostic information. For example, the presence of pleural effusion indicates worse prognosis [4, 29, 30]. However, the prognostic value of CR in CAP has rarely been investigated because it is difficult to obtain objective and quantitative prognostic biomarkers from CR. Recently, DL models exhibited the potential for predicting future outcomes. A study reported a DL-based prediction of mortality in patients with CAP using CRs [31]. Similar to our study, Quah et al. reported that the discriminative performance of the DL model, CURB-65 score, and the DL model combined with CURB-65 score for the prediction of 30-day mortality were AUCs of 0.79, 0.76, and 0.83, respectively [31]. Comparable discrimination of the DL model with the CURB-65 score and improved discrimination by combination with the CURB-65 score suggests the potential of the DL model as a decision support tool in CAP management. The high specificities at the same sensitivities compared to the CURB-65 score observed in our study suggest that the DL model may help reduce

unnecessary hospitalization or invasive treatment for low-risk patients.

Contrary to Quah et al.'s study, which developed and validated a DL model using single-institution data [31], we validated the DL model in three external test cohorts (one temporally separated cohort and two cohorts from other institutions) to evaluate the model's generalizability. Regarding discrimination, the model exhibited consistent performance in external test cohorts. The model exhibited higher AUCs (0.80) in the test cohorts from different institutions than in the temporally separated cohort (0.77). This difference in discrimination may be due to differences in the baseline characteristics of patients (e.g., tertiary referral institution vs. secondary referral institution) since the performance of the CURB-65 score exhibited a similar tendency. Regarding calibration, both the DL model and combined model overestimated the risk in external test cohorts B and C. This miscalibration might also be due to differences in patient characteristics between the developmental and external test cohorts. Recalibration of the risk predicted by the DL model before application to patients with different characteristics may improve model calibration [21, 32].

Studies have reported the feasibility of the DL model for predicting mortality in patients with coronavirus disease pneumonia using their CRs [33, 34]. Compared to models specifically targeting coronavirus disease, the strength of our model is that it can be applied to patients with CAP regardless of the causative pathogen, making it more valuable than models for coronavirus disease in the post-pandemic era.

The advantage of a DL model using CR as an input compared to models using clinical variables is that automated processing might be feasible and is not influenced by subjective evaluation by physicians [31]. However, an important shortcoming of the DL model is the difficulty in explaining the logical background of prediction. In our study, class activation maps suggested that predictions of the DL model tended to be influenced by the area of

pneumonia in cases of high predicted risk, whereas the model tended not to focus on the area of pneumonia in cases of low predicted risk (Figures 3 and 4).

4.2. Limitations

This study has several limitations. First, since our study was retrospective, we could not evaluate whether the prediction of the DL model can influence the management of patients with CAP. Second, since clinical variables and risk factors were collected retrospectively, we could not evaluate clinical risk factors other than the CURB-65 score. The pneumonia severity index, another established risk-scoring system for CAP, could not be obtained. The pneumonia severity index was not practically used in the management of patients in our study because more variables are included in the pneumonia severity index and the application is relatively complex. Finally, we evaluated only 30-day all-cause mortality (including death due to both CAP and other causes) as an outcome of this study, and other clinical outcomes such as the length of hospitalization were not evaluated.

Chapter 5. Conclusion

5.1. Conclusion

In conclusion, a deep learning-based model could predict the 30-day mortality in patients with community-acquired pneumonia from their initial chest radiographs with higher specificity at the same sensitivity compared to the CURB-65 score. Adding the deep learning model prediction to the CURB-65 score led to improved discrimination in predicting 30-day mortality. A prospective study is required to evaluate whether the deep learning model can contribute to the management of these patients.

Bibliography

[1] Cemters for Disease Control and Prevention, National Center for Health Statistics. National Vital Statistics System, Mortality 1999-2020 on CDC WONDER Online Database https://wonder.cdc.gov/controller/datarequest/D76;jsessionid=9ACC9395F6EBA67CC532E7CF3B69. Published 2022. Accessed 2022 October 17.

[2] Cairns C, Kang K. National Hospital Ambulatory Medical Care Survey: 2019 emergency department summary tables. https://www.cdc.gov/nchs/data/nhamcs/web_tables/2019-nhamcs-ed-web-tables-508.pdf. Published 2022. Accessed 2022 October 17.

[3] Jain, Seema, et al. "Community-acquired pneumonia requiring hospitalization among US adults." New England Journal of Medicine 373.5 (2015): 415-427.

[4] Fine, Michael J., et al. "A prediction rule to identify low-risk patients with community-acquired pneumonia." New England journal of medicine 336.4 (1997): 243-250.

[5] Lim, Wei Shen, et al. "BTS guidelines for the management of community acquired pneumonia in adults: update 2009." Thorax 64.Suppl 3 (2009): iii1-iii55.

[6] Lim, W. S., S. Lewis, and J. T. Macfarlane. "Severity prediction rules in community acquired pneumonia: a validation study." Thorax 55.3 (2000): 219-223.

[7] Lim, W. S., et al. "Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study." Thorax 58.5 (2003): 377-382.

[8] Olson, Gregory, and Andrew M. Davis. "Diagnosis and treatment of adults with community-acquired pneumonia." Jama 323.9 (2020): 885-886.

[9] Mandell, Lionel A., et al. "Infectious Diseases Society of America/American Thoracic Society consensus guidelines on the management of community-acquired pneumonia in adults." Clinical

infectious diseases 44.Supplement_2 (2007): S27-S72.
[10] Albaum, Michael N., et al. "Interobserver reliability of the chest radiograph in community-acquired pneumonia." Chest 110.2 (1996): 343-350.
[11] Loeb, Mark B., et al. "Interobserver reliability of radiologists' interpretations of mobile chest radiographs for nursing home–acquired pneumonia." Journal of the American Medical Directors Association 7.7 (2006): 416-419.
[12] Kessler, Larry G., et al. "The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions." Statistical methods in medical research 24.1 (2015): 9-26.
[13] Sullivan, Daniel C., et al. "Metrology standards for quantitative imaging biomarkers." Radiology 277.3 (2015): 813.
[14] Lu, Michael T., et al. "Deep learning to assess long-term mortality from chest radiographs." JAMA network open 2.7 (2019): e197416-e197416.
[15] Nam, Ju Gang, et al. "Deep Learning Prediction of Survival in Patients with Chronic Obstructive Pulmonary Disease Using Chest Radiographs." Radiology (2022): 212071.
[16] Lu, Michael T., et al. "Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model." Annals of Internal Medicine 173.9 (2020): 704-713.
[17] Gensheimer, Michael F., and Balasubramanian Narasimhan. "A scalable discrete-time survival model for neural networks." PeerJ 7 (2019): e6257.
[18] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[19] Hwang, Eui Jin, et al. "Development and validation of a deep learning–based automated detection algorithm for major thoracic diseases on chest radiographs." JAMA network open 2.3 (2019): e191095-e191095.
[20] Saporta, Adriel, et al. "Benchmarking saliency methods for

chest X-ray interpretation." Nature Machine Intelligence 4.10 (2022): 867-878.
[21] Steyerberg, Ewout W., et al. "Validation and updating of predictive logistic regression models: a study on sample size and shrinkage." Statistics in medicine 23.16 (2004): 2567-2586.
[22] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.
[23] Pencina, Michael J., Ralph B. D'Agostino Sr, and Ewout W. Steyerberg. "Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers." Statistics in medicine 30.1 (2011): 11-21.
[24] Cook, Nancy R. "Comments on'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond'by MJ Pencina et al., Statistics in Medicine." Statistics in medicine 27.2 (2008): 191-195.
[25] DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson. "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach." Biometrics (1988): 837-845.
[26] Leisenring, Wendy, Todd Alono, and Margaret Sullivan Pepe. "Comparisons of predictive values of binary medical diagnostic tests for paired designs." Biometrics 56.2 (2000): 345-351.
[27] Spiegelhalter, David J. "Probabilistic prediction in patient management and clinical trials." Statistics in medicine 5.5 (1986): 421-433.
[28] Vickers, Andrew J., and Elena B. Elkin. "Decision curve analysis: a novel method for evaluating prediction models." Medical Decision Making 26.6 (2006): 565-574.
[29] Hasley, Peggy B., et al. "Do pulmonary radiographic findings at presentation predict mortality in patients with community-acquired pneumonia?." Archives of Internal Medicine 156.19 (1996): 2206-2212.
[30] Dean, Nathan C., et al. "Pleural effusions at first ED encounter

predict worse clinical outcomes in patients with pneumonia." Chest 149.6 (2016): 1509-1515.
[31] Quah, Jessica, et al. "Chest radiograph-based artificial intelligence predictive model for mortality in community-acquired pneumonia." BMJ open respiratory research 8.1 (2021): e001045.
[32] Hwang, Eui Jin, et al. "Automated identification of chest radiographs with referable abnormality with deep learning: need for recalibration." European Radiology 30.12 (2020): 6902-6912.
[33] Mushtaq, Junaid, et al. "Initial chest radiographs and artificial intelligence (AI) predict clinical outcomes in COVID-19 patients: analysis of 697 Italian patients." European radiology 31.3 (2021): 1770-1779.
[34] Au-Yong, Iain, et al. "Chest radiograph scoring alone or combined with other risk scores for predicting outcomes in COVID-19." Radiology 302.2 (2022): 460.

Abstract

지역사회획득폐렴 환자의 예후를 개선하고 폐렴으로 인한 비용 부담을 줄이기 위해서는 정확한 위험도 예측이 필요하다. 관련 질환군의 예후를 예측하기 위한 지표(CURB-65)가 개발되었으나 성능이 만족스럽지 못하거나 지표에 포함된 인자 획득의 어려움으로 실제 사용에 한계가 있다. 본 연구에서는 2013년부터 2019년 사이 단일 기관에서 지역사회획득폐렴으로 진단받은 환자 데이터를 활용해 진단 시 촬영한 흉부방사선영상에서 지역사회획득폐렴 진단 후 30일 이내 사망 위험을 예측하는 딥러닝 모델을 개발하고 검증했다. 제안하는 딥러닝 모델은 2020년 1월부터 12월 사이에 지역사회획득폐렴으로 같은 기관의 응급실을 방문한 환자(테스트 코호트 A)와 2개의 다른 기관(테스트 코호트 B, C)에서 평가되었다. 본 모델의 성능 평가를 위해 area under receiver operating characteristic curves (AUCs)를 사용했다. 기존의 위험 예측 지표인 CURB-65 점수에 대한 딥러닝 모델의 부가 가치는 순 재분류 지수(NRI)와 통합 판별 개선(IDI)을 이용하여 평가되었다. 검사 코호트 A, B, C에서는 30일간 사망률이 각각 18%, 8%, 11%였다. 딥러닝 모델은 테스트 코호트 A, B, C에 각각 0.77, 0.80, 0.80의 AUC를 나타냈다. 모든 외부 테스트 코호트(연속 NRI, 0.30-0.74, IDI, 0.08-0.12) 점수에 DL 모델 예측을 추가하였을 때 성능이 향상되었다. 본 연구를 통해 딥러닝 기반 모델이 흉부방사선영상을 통해 지역사회획득폐렴 환자의 30일 사망률을 예측할 수 있음을 확인하였다. 딥러닝 기반의 사망률 예측 모델을 사용하여 지역사회획득폐렴 환자의 흉부방사선영상을 평가하는 것은 입원 또는 집중 치료를 위한 위험 계층화와 임상 의사 결정을 개선하는 데 도움이 될 수 있다.

주요어: 딥러닝, 컨볼루션 신경망, 지역사회획득폐렴, 흉부방사선영상, 생존 예측

학 번: 2021-21892