A deep learning-based model to estimate pulmonary function from chest x-rays: multi-institutional model development and validation study in Japan

A deep learning-based model to estimate pulmonary function from chest x-rays: multi-institutional model development and validation study in Japan

Daiju Ueda, Toshimasa Matsumoto, Akira Yamamoto, Shannon L Walston, Yasuhito Mitsuyama, Hirotaka Takita, Kazuhisa Asai, Tetsuya Watanabe, Koji Abo, Tatsuo Kimura, Shinya Fukumoto, Toshio Watanabe, Tohru Takeshita, Yukio Miki

Summary

Background Chest x-ray is a basic, cost-effective, and widely available imaging method that is used for static assessments of organic diseases and anatomical abnormalities, but its ability to estimate dynamic measurements such as pulmonary function is unknown. We aimed to estimate two major pulmonary functions from chest x-rays.

Methods In this retrospective model development and validation study, we trained, validated, and externally tested a deep learning-based artificial intelligence (AI) model to estimate forced vital capacity (FVC) and forced expiratory volume in 1 s (FEV_1) from chest x-rays. We included consecutively collected results of spirometry and any associated chest x-rays that had been obtained between July 1, 2003, and Dec 31, 2021, from five institutions in Japan (labelled institutions A–E). Eligible x-rays had been acquired within 14 days of spirometry and were labelled with the FVC and FEV_1. X-rays from three institutions (A–C) were used for training, validation, and internal testing, with the testing dataset being independent of the training and validation datasets, and then x-rays from the two other institutions (D and E) were used for independent external testing. Performance for estimating FVC and FEV_1 was evaluated by calculating the Pearson's correlation coefficient (*r*), intraclass correlation coefficient (ICC), mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE) compared with the results of spirometry.

Findings We included 141 734 x-ray and spirometry pairs from 81 902 patients from the five institutions. The training, validation, and internal test datasets included 134 307 x-rays from 75 768 patients (37 718 [50%] female, 38 050 [50%] male; mean age 56 years [SD 18]), and the external test datasets included 2137 x-rays from 1861 patients (742 [40%] female, 1119 [60%] male; mean age 65 years [SD 17]) from institution D and 5290 x-rays from 4273 patients (1972 [46%] female, 2301 [54%] male; mean age 63 years [SD 17]) from institution E. External testing for FVC yielded *r* values of 0·91 (99% CI 0·90–0·92) for institution D and 0·90 (0·89–0·91) for institution E, ICC of 0·91 (99% CI 0·90–0·92) and 0·89 (0·88–0·90), MSE of 0·17 L^2 (99% CI 0·15–0·19) and 0·17 L^2 (0·16–0·19), RMSE of 0·41 L (99% CI 0·39–0·43) and 0·41 L (0·39–0·43), and MAE of 0·31 L (99% CI 0·29–0·32) and 0·31 L (0·30–0·32). External testing for FEV_1 yielded *r* values of 0·91 (99% CI 0·90–0·92) for institution D and 0·91 (0·90–0·91) for institution E, ICC of 0·90 (99% CI 0·89–0·91) and 0·90 (0·90–0·91), MSE of 0·13 L^2 (99% CI 0·12–0·15) and 0·11 L^2 (0·10–0·12), RMSE of 0·37 L (99% CI 0·35–0·38) and 0·33 L (0·32–0·35), and MAE of 0·28 L (99% CI 0·27–0·29) and 0·25 L (0·25–0·26).

Interpretation This deep learning model allowed estimation of FVC and FEV_1 from chest x-rays, showing high agreement with spirometry. The model offers an alternative to spirometry for assessing pulmonary function, which is especially useful for patients who are unable to undergo spirometry, and might enhance the customisation of CT imaging protocols based on insights gained from chest x-rays, improving the diagnosis and management of lung diseases. Future studies should investigate the performance of this AI model in combination with clinical information to enable more appropriate and targeted use.

Funding None.

Lancet Digit Health 2024; 6: e580–88

Published Online July 8, 2024 https://doi.org/10.1016/S2589-7500(24)00113-4

Department of Diagnostic and Interventional Radiology (D Ueda MD, T Matsumoto PhD, A Yamamoto MD, S L Walston MS, Y Mitsuyama MD, H Takita MD, Prof Y Miki MD), **Department of Artificial Intelligence** (D Ueda), **Department of Respiratory Medicine** (K Asai MD, Te Watanabe MD), **and Department of Premier Preventive Medicine** (T Kimura MD, S Fukumoto MD, Prof To Watanabe MD), **Graduate School of Medicine, Osaka Metropolitan University, Osaka, Japan; Central Clinical Laboratory, Osaka Metropolitan University Hospital, Osaka, Japan** (K Abo MS); **Department of Radiology, Osaka Habikino Medical Center, Osaka, Japan** (T Takeshita MD)

Correspondence to: Dr Daiju Ueda, Department of Artificial Intelligence, Graduate School of Medicine, Osaka Metropolitan University, Osaka 545-8585, Japan **ai.labo.ocu@gmail.com**

Introduction

Pulmonary function testing is an essential examination for diagnosis of obstructive or restrictive respiratory impairment. Forced vital capacity (FVC) and forced expiratory volume in 1 s (FEV_1) are representative values for pulmonary function and are measured with spirometry. Since spirometry was first implemented in clinical practice in 1846,[1] its importance has been proven in many areas.[2] Two of the most valuable areas for the clinical application of spirometry are chronic obstructive pulmonary disease (COPD) and asthma. Current international COPD guidelines, the Global Initiative for Chronic Obstructive Lung Disease guidelines, require FVC for the diagnosis of COPD in individuals with

Research in context

Evidence before this study

We evaluated the current state of knowledge regarding chest x-ray-based artificial intelligence (AI) estimation for pulmonary functions by searching PubMed, MEDLINE, and the Web of Science for publications from database inception up to March 1, 2024, using the keywords "artificial intelligence", "deep learning", "convolutional neural network", "pulmonary function", "spirometry", "spirogram", "chest x-ray", "chest radiography", and "chest radiograph." We did not identify any research to estimate measurements of spirometry from chest x-rays.

Added value of this study

To our knowledge, this is the first study to develop an AI model to predict pulmonary function from x-rays. The deep learning model developed in this study was able to accurately identify the spirometric values from chest x-rays with correlation coefficients of 0·91 and 0·90 for forced vital capacity and 0·91 and 0·91 for forced expiratory volume in 1 s in the two external test datasets, respectively. Our AI model, capable of estimating dynamic examination values from a static image, has revealed a new potential for chest x-ray in respiratory care.

Implications of all the available evidence

Our AI model can estimate spirometry measurements from chest x-rays with excellent agreement with real values, indicating that static features on x-rays correlate well with results of this dynamic examination. Our model adds new value to chest x-rays and the rapidity and availability of x-rays provides further potential as a complement to spirometry. Future studies should investigate the performance of this AI model in combination with clinical information, and across more diverse populations, to enable more appropriate and targeted use.

relevant symptoms and risk factors.[3] They also recommend the use of FVC and FEV_1 for estimating the prognosis of COPD.[4,5] Spirometry is also important in the diagnosis and management of asthma.[6] Current international asthma guidelines, the Global Initiative for Asthma guidelines, strongly support continuous monitoring with regular spirometry.[6] FEV_1 is considered a strong independent predictor of the risk of asthma exacerbations.[7] Furthermore, respiratory impairment as detected with spirometry is an important risk factor not only for these respiratory diseases, but also for all-cause mortality and cardiovascular disease.[8–10] Although spirometry is a useful test, room remains for a complementary method that is easier to use in older individuals and young children who have difficulty following instructions.[11] Moreover, during the COVID-19 pandemic, the use of spirometry was restricted to protect staff and individual patients from infection.[12]

Chest x-ray is used worldwide,[13] and evidence has suggested that some findings from chest x-ray and chest CT correlate with FVC and FEV_1. For example, a lower position of the diaphragm in the chest cavity and an enlarged posterior sternal gap on chest x-ray in patients with COPD has been associated with a decrease in FEV_1.[14] Additionally, the extent of emphysema and bronchial wall thickening both correlate directly with decreases in FEV_1.[15] In patients with asthma evaluated with CT scans, bronchial wall thickening and air trapping have correlated with the FEV_1/FVC ratio.[16] We hypothesised that pulmonary function could be estimated from chest x-ray with these or other unknown findings.

Deep learning is a field of artificial intelligence (AI) that can automatically extract features from training data, while conventional machine learning needs the features to be manually defined.[17] Therefore, deep learning is more advantageous for tasks with complex or unknown features. A recent single-centre AI study identified an association between chest CT scan findings and respiratory function;[18] however, to our knowledge, the association between pulmonary function and findings from chest x-rays, which are more widely available, has yet to be investigated. We aimed to develop and validate a deep learning AI model to estimate FVC and FEV_1 from chest x-rays collected from multiple institutions in Japan.

Methods

Study design

In this retrospective model development and validation study, we trained, validated, and externally tested a deep learning model that estimated FVC and FEV_1 from chest x-rays. Chest x-rays were retrospectively collected from patients who had undergone spirometry at one of five institutions. After developing the AI model, we visualised the regions of the chest x-rays that were important for the AI to predict FVC and FEV_1 values.

This study complies with the Declaration of Helsinki. The ethics board of Osaka Metropolitan University reviewed and approved the protocol for the present study (approval ID: 2021-013). The need for informed consent was waived because the x-rays had been acquired during daily clinical practice. This manuscript was written in accordance with the TRIPOD guideline.[19]

Patients, examination, and x-ray acquisition

We collected consecutive spirometry data obtained between July 1, 2003, and Dec 31, 2021, at five institutions in Japan: Osaka Metropolitan University Hospital, Osaka (institution A), between May 1, 2007, and June 30, 2019; Habikino Medical Center, Habikino (institution B), between July 1, 2003, and Aug 31, 2020; MedCity21, Osaka (institution C), between Dec 1, 2014, and Dec 31, 2020; Higashisumiyoshi Morimoto Hospital,

Osaka (institution D), between Feb 1, 2018, and Dec 31, 2021; and Kashiwara Municipal Hospital, Kashiwara (institution E), between April 1, 2010, and Dec 31, 2021. If a patient underwent spirometry more than once during the data collection period, all examinations were included. Patients usually underwent both tests at the same institution. Chest x-rays with a posteroanterior view in the standing position taken within 14 days of the spirometry assessment were collected for this analysis. This window is smaller than those stipulated in guidelines recommending spirometry follow-up for pulmonary disease.[3,6] We chose this window to ensure temporal alignment of data, considering that substantial spirometric changes are less likely to occur during shorter intervals. Institution C is specialised for outpatient health check-ups, such that spirometry and x-ray were always done on the same day. If two or more x-rays were available within the collection period, the x-ray taken closest to the day of the spirometry examination was selected. Patients with no chest x-rays within the collection window were excluded. Hence, one chest x-ray was collected per spirometry assessment.

Ground truth labelling

Both FVC and FEV_1 were extracted from spirometry reports. The collected chest x-rays were labelled with these ground truth values. All spirometry was performed according to the American Thoracic Society/European Respiratory Society Task Force recommendations.[20] Information regarding COPD, asthma, interstitial lung disease, inactive tuberculosis, non-tuberculous mycobacteriosis, and lung cancer diagnoses were extracted from the electronic patient medical records using ICD-10 definitions.[21]

Data partitioning

Labelled chest x-rays from three institutions (institutions A, B, and C) were divided into training, validation, and internal test datasets on a patient basis in an 8:1:1 ratio. Additionally, we prepared external test datasets collected from the other two institutions (institutions D and E). We confirmed that there was no overlap of patients among the respective datasets. The training dataset was used to train the AI model and the validation dataset was used to internally tune the AI model. The internal test dataset was an independent dataset that was not used for training and tuning but was collected from the same institutions as the training and validation datasets, while the external test datasets were collected from different facilities in the region and exclusively used for performance evaluation.

Model development

We developed an AI model to estimate FVC and FEV_1 values from chest x-rays using ConvNeXt[22] as a feature extractor, followed by two classifiers to estimate the two values. Each classifier comprised a fully connected layer connected to a loss function. The two loss values from the two classifiers were summed and smeared to obtain the total loss value (ie, the two loss values from the two classifiers were aggregated to obtain the total loss value, which was then used to update the model's weights during training). As loss functions, root mean square error (RMSE), mean square error (MSE), and mean absolute error (MAE) were applied and compared. The highest performance was explored by testing with image resolutions of 128, 256, 512, and 1024 pixels. During this training, the AI model determined which features could predict FVC and FEV_1 from x-rays. The model in which the total loss value in the validation dataset was the smallest within 300 epochs was chosen as the best performing model. Every development process was performed using the PyTorch framework (version 2.0.1; The Linux Foundation, San Francisco, CA, USA). Detailed processes for development of the AI model, the machine environment, and an outline of the model are shown in the appendix (pp 2, 4).

For more on **PyTorch** see https://pytorch.org/

See **Online** for appendix

Model test

The prediction performance of the best performing model was assessed on the internal and external test datasets. Performance at estimating FVC and FEV_1 was evaluated by calculating the difference from the real spirometry results.

To show the region of interest for each classifier as it discriminated each image, saliency maps for images in the external test dataset from institution E with the top 10% and bottom 10% of FVC and FEV_1 prediction were obtained. The images from each 10% set were added together and divided by the total number of included images to create an averaged image per variable. To make each saliency map, SHapley Additive exPlanations (SHAP) was applied.[23] SHAP is a method based on cooperative game theory that is used to increase the transparency and interpretability of machine learning models. In the case of this study, the player is the pixel in the image and the game outcome is the prediction of the model. A detailed explanation of the saliency map generation model is in the appendix (p 2). Independent radiologists assessed clinically valuable findings relevant to FVC and FEV_1 with the saliency maps generated from the external dataset from institution D. A detailed explanation of this process is in the appendix (p 2).

Statistical analysis

To evaluate the regression performance of the AI model, the Pearson correlation coefficient (r), intraclass correlation coefficient (ICC), RMSE, MSE, and MAE between model-predicted values and spirometry values were calculated. The r, ICC, RMSE, MSE, and MAE were further assessed on the basis of the patient's sex, age, and disease aetiology. The AI classification performance for FVC of less than 80% predicted, FEV_1 of less than 80% predicted, and FEV_1/FVC ratio of less than 70% for

	Training dataset			Validation dataset			Internal test dataset			External test dataset	
	Institution A	Institution B	Institution C	Institution A	Institution B	Institution C	Institution A	Institution B	Institution C	Institution D	Institution E
Total number of x-ray and spirometry examination pairs	16580	54712	37074	1821	6671	4688	1789	6302	4670	2137	5290
Patients	11692	28044	21273	1317	3405	2659	1340	3379	2659	1861	4273
Sex											
Male	7105 (61%)	13648 (49%)	9949 (47%)	787 (60%)	1639 (48%)	1239 (47%)	817 (61%)	1636 (48%)	1230 (46%)	1119 (60%)	2301 (54%)
Female	4587 (39%)	14396 (51%)	11324 (53%)	530 (40%)	1766 (52%)	1420 (53%)	523 (39%)	1743 (52%)	1429 (54%)	742 (40%)	1972 (46%)
Age, years	64 (14)	54 (23)	51 (11)	63 (14)	54 (23)	52 (11)	64 (13)	54 (23)	51 (11)	65 (17)	63 (17)
Time between spirometry and x-ray, days	5 (7)	2 (6)	0 (0)	5 (7)	2 (6)	0 (0)	6 (7)	2 (6)	0 (0)	1 (5)	1 (5)
Forced vital capacity, L	3·0 (0·93)	2·6 (0·89)	3·5 (0·84)	3·0 (0·93)	2·6 (0·89)	3·5 (0·84)	3·0 (0·9)	2·6 (0·89)	3·5 (0·83)	2·9 (1·0)	2·8 (0·92)
Forced expiratory volume in 1 s, L	2·2 (0·77)	2·0 (0·80)	2·8 (0·69)	2·2 (0·77)	2·0 (0·80)	2·8 (0·69)	2·1 (0·76)	2·0 (0·82)	2·8 (0·69)	2·2 (0·87)	2·2 (0·79)
Diagnosis*											
Chronic obstructive pulmonary disease	2620 (16%)	6525 (12%)	254 (1%)	308 (17%)	716 (11%)	31 (1%)	259 (14%)	695 (11%)	53 (1%)	249 (12%)	92 (2%)
Asthma	2636 (16%)	25244 (46%)	308 (1%)	292 (16%)	3021 (45%)	21 (<1%)	249 (14%)	2844 (45%)	33 (1%)	267 (12%)	118 (2%)
Interstitial lung disease	791 (5%)	1773 (3%)	40 (<1%)	80 (4%)	207 (3%)	1 (<1%)	94 (5%)	193 (3%)	6 (<1%)	121 (6%)	14 (<1%)
Inactive tuberculosis	250 (2%)	1475 (3%)	30 (<1%)	32 (2%)	178 (3%)	4 (<1%)	36 (2%)	193 (3%)	6 (<1%)	12 (1%)	2 (<1%)
Non-tuberculous mycobacteria	156 (1%)	1290 (2%)	46 (<1%)	16 (1%)	154 (2%)	3 (<1%)	23 (1%)	170 (3%)	13 (<1%)	4 (<1%)	9 (<1%)
Lung cancer	2812 (17%)	8453 (15%)	120 (<1%)	300 (16%)	933 (14%)	29 (1%)	358 (20%)	1071 (17%)	14 (<1%)	264 (12%)	23 (<1%)

Data are n, n (%), or mean (SD). *Diagnostic data were only collected for the listed conditions; percentages were calculated using the total number of x-ray and spirometry pairs.

Table 1: **Dataset demographic and clinical data**

the external datasets were analysed by receiver operating curve analysis comparing spirometry and model-predicted values;[3,24] these values were chosen on the basis of available guidelines.[3] We quantified the frequency at which the model could accurately predict FVC and FEV_1 values within ranges, compared with spirometry results. To reveal any clustering effects from using multiple spirometry and x-rays from the same patient, we re-calculated the results after patient duplication removal.

All analyses were done in SciPy using Python (version 3.8.1). p values are not reported because we aimed to evaluate the AI model's performance in estimating pulmonary function values rather than comparing groups or testing specific hypotheses. Instead of using p values, we calculated 99% CIs of the performance metrics, and we estimated these using bootstrapping (repeatedly sampling from the original dataset with replacement to create multiple simulated datasets of the same size).

Role of the funding source

There was no funding source for this study.

Results

141734 x-ray and spirometry matched pairs from 81902 patients were included in our analysis. The training, validation, and internal test datasets included 134307 x-rays from 75768 patients (37718 [50%] were female and 38050 [50%] were male; mean age 56 years [SD 18]). The training dataset included 108366 x-rays from 61009 patients (30307 [50%] were female and 30702 [50%] were male; mean age 54 years [SD 17; range 6–99]) and the validation dataset included 13180 x-rays from 7381 patients (3716 [50%] were female and 3665 [50%] were male; mean age 54 years [SD 17; range 7–96]) from institutions A, B, and C. The internal test dataset included 12761 x-rays from 7378 patients (3695 [50%] were female and 3683 [50%] were male; mean age 54 years [SD 17; range 7–94]) from the same three institutions. The external test datasets included 2137 x-rays from 1861 patients (742 [40%] were female and 1119 [60%] were male; mean age 65 years [SD 17; range 7–98]) from institution D and 5290 x-rays from 4273 patients (1972 [46%] were female and 2301 [54%] were male; mean age 63 years [SD 17; range 4–99]) from institution E. Data on race and ethnicity were not available. Demographic and clinical data for the datasets are shown in table 1 and a flowchart of the dataset criteria is in the appendix (p 5).

The best performing model was obtained with an RMSE loss function of 0·39, and an image size of 1024 pixels at 182 epochs. In the FVC determination using external test datasets, *r* values for institutions D and E were 0·91 (99% CI 0·90–0·92) and 0·90 (0·89–0·91), respectively (table 2; figure 1A). ICC values were 0·91 (99% CI 0·90–0·92) and 0·89 (0·88–0·90), MSE values were 0·17 L^2 (99% CI 0·15–0·19) and 0·17 L^2 (0·16–0·19), RMSE values were 0·41 L (99% CI 0·39–0·43) and 0·41 L (0·39–0·43), and MAE values were 0·31 L (99% CI

	Internal test dataset			External test dataset	
	Institution A	Institution B	Institution C	Institution D	Institution E
Overall					
FVC					
Pearson correlation coefficient	0·92 (0·91–0·94)	0·91 (0·90–0·92)	0·94 (0·93–0·94)	0·91 (0·90–0·92)	0·90 (0·89–0·91)
Intraclass correlation coefficient	0·92 (0·91–0·93)	0·90 (0·90–0·91)	0·94 (0·93–0·94)	0·91 (0·90–0·92)	0·89 (0·88–0·90)
Mean square error, L^2	0·12 (0·11–0·14)	0·14 (0·13–0·15)	0·08 (0·07–0·09)	0·17 (0·15–0·19)	0·17 (0·16–0·19)
Root mean square error, L	0·35 (0·33–0·38)	0·37 (0·36–0·39)	0·28 (0·27–0·29)	0·41 (0·39–0·43)	0·41 (0·39–0·43)
Mean absolute error, L	0·27 (0·25–0·28)	0·29 (0·28–0·29)	0·22 (0·21–0·22)	0·31 (0·29–0·32)	0·31 (0·30–0·32)
FEV_1					
Pearson correlation coefficient	0·91 (0·90–0·92)	0·92 (0·91–0·92)	0·92 (0·91–0·92)	0·91 (0·90–0·92)	0·91 (0·90–0·91)
Intraclass correlation coefficient	0·90 (0·89–0·91)	0·91 (0·90–0·91)	0·91 (0·91–0·92)	0·90 (0·89–0·91)	0·90 (0·90–0·91)
Mean square error, L^2	0·10 (0·09–0·11)	0·11 (0·10–0·12)	0·08 (0·07–0·08)	0·13 (0·12–0·15)	0·11 (0·10–0·12)
Root mean square error, L	0·32 (0·30–0·34)	0·33 (0·32–0·34)	0·28 (0·27–0·29)	0·37 (0·35–0·38)	0·33 (0·32–0·35)
Mean absolute error, L	0·25 (0·23–0·26)	0·25 (0·25–0·26)	0·22 (0·21–0·22)	0·28 (0·27–0·29)	0·25 (0·25–0·26)
Chronic obstructive pulmonary disease					
FVC					
Pearson correlation coefficient	0·92 (0·89–0·94)	0·84 (0·81–0·87)	0·77 (0·62–0·87)	0·81 (0·74–0·86)	0·78 (0·66–0·87)
Intraclass correlation coefficient	0·91 (0·89–0·92)	0·82 (0·80–0·85)	0·76 (0·65–0·84)	0·79 (0·73–0·84)	0·74 (0·64–0·83)
Mean square error, L^2	0·14 (0·11–0·18)	0·18 (0·16–0·21)	0·17 (0·11–0·25)	0·29 (0·21–0·37)	0·25 (0·16–0·37)
Root mean square error, L	0·38 (0·33–0·42)	0·43 (0·40–0·46)	0·42 (0·33–0·50)	0·53 (0·46–0·61)	0·50 (0·40–0·61)
Mean absolute error, L	0·29 (0·26–0·33)	0·34 (0·32–0·37)	0·33 (0·25–0·42)	0·40 (0·34–0·46)	0·38 (0·30–0·48)
FEV_1					
Pearson correlation coefficient	0·89 (0·85–0·92)	0·89 (0·85–0·91)	0·87 (0·77–0·92)	0·83 (0·76–0·89)	0·83 (0·72–0·90)
Intraclass correlation coefficient	0·88 (0·85–0·90)	0·87 (0·85–0·89)	0·80 (0·71–0·87)	0·79 (0·73–0·85)	0·82 (0·73–0·89)
Mean square error, L^2	0·11 (0·09–0·14)	0·12 (0·10–0·14)	0·18 (0·10–0·25)	0·16 (0·12–0·20)	0·12 (0·07–0·17)
Root mean square error, L	0·34 (0·30–0·38)	0·34 (0·31–0·38)	0·42 (0·32–0·50)	0·40 (0·35–0·45)	0·35 (0·27–0·41)
Mean absolute error, L	0·27 (0·24–0·31)	0·26 (0·24–0·28)	0·34 (0·25–0·42)	0·31 (0·27–0·36)	0·26 (0·21–0·33)
Asthma					
FVC					
Pearson correlation coefficient	0·93 (0·91–0·95)	0·90 (0·89–0·91)	0·88 (0·79–0·93)	0·89 (0·85–0·93)	0·87 (0·79–0·93)
Intraclass correlation coefficient	0·93 (0·91–0·95)	0·90 (0·89–0·91)	0·87 (0·81–0·92)	0·88 (0·84–0·91)	0·85 (0·79–0·91)
Mean square error, L^2	0·09 (0·07–0·12)	0·16 (0·14–0·18)	0·12 (0·06–0·18)	0·22 (0·16–0·29)	0·24 (0·15–0·40)
Root mean square error, L	0·31 (0·26–0·35)	0·40 (0·38–0·42)	0·34 (0·25–0·43)	0·47 (0·40–0·54)	0·49 (0·38–0·64)
Mean absolute error, L	0·24 (0·21–0·27)	0·30 (0·29–0·31)	0·29 (0·20–0·37)	0·35 (0·30–0·40)	0·38 (0·30–0·47)
FEV_1					
Pearson correlation coefficient	0·92 (0·90–0·94)	0·92 (0·91–0·93)	0·81 (0·61–0·92)	0·90 (0·86–0·93)	0·87 (0·77–0·93)
Intraclass correlation coefficient	0·92 (0·90–0·94)	0·91 (0·90–0·92)	0·75 (0·56–0·86)	0·88 (0·85–0·90)	0·86 (0·80–0·91)
Mean square error, L^2	0·09 (0·07–0·11)	0·12 (0·11–0·13)	0·14 (0·07–0·24)	0·16 (0·13–0·21)	0·16 (0·09–0·26)
Root mean square error, L	0·30 (0·26–0·33)	0·34 (0·33–0·37)	0·38 (0·26–0·49)	0·40 (0·36–0·45)	0·40 (0·30–0·51)
Mean absolute error, L	0·23 (0·20–0·26)	0·26 (0·25–0·27)	0·28 (0·17–0·39)	0·31 (0·27–0·35)	0·30 (0·23–0·36)

Data are n, with 99% CIs in parentheses. FVC=forced vital capacity. FEV_1=forced expiratory volume in 1 s.

Table 2: **Model performance results**

0·29–0·32) and 0·31 L (0·30–0·32), respectively. In the FEV_1 determination using external test datasets, *r* values for institutions D and E were 0·91 (99% CI 0·90–0·92) and 0·91 (0·90–0·91), respectively. ICC values were 0·90 (99% CI 0·89–0·91) and 0·90 (0·90–0·91), MSE values were 0·13 L^2 (99% CI 0·12–0·15) and 0·11 L^2 (0·10–0·12), RMSE values were 0·37 L (99% CI 0·35–0·38) and 0·33 L (0·32–0·35), and MAE values were 0·28 L (99% CI 0·27–0·29) and 0·25 L (0·25–0·26), respectively. Patients with COPD had *r* values of 0·81 (99% CI 0·74–0·86) and 0·78 (0·66–0·87) for FVC and of 0·83 (0·76–0·89) and 0·83 (0·72–0·90) for FEV_1, for institutions D and E, respectively. Patients with asthma had *r* values of 0·89 (0·85–0·93) and 0·87 (0·79–0·93) for FVC and 0·90 (99% CI 0·86–0·93) and 0·87 (0·77–0·93) for FEV_1, for institutions D and E, respectively. Model regression metrics by sex, age, and the presence of other diseases are in the appendix (pp 8–10).

The area under the receiver operating characteristic curve for classifying FVC as less than 80% predicted was

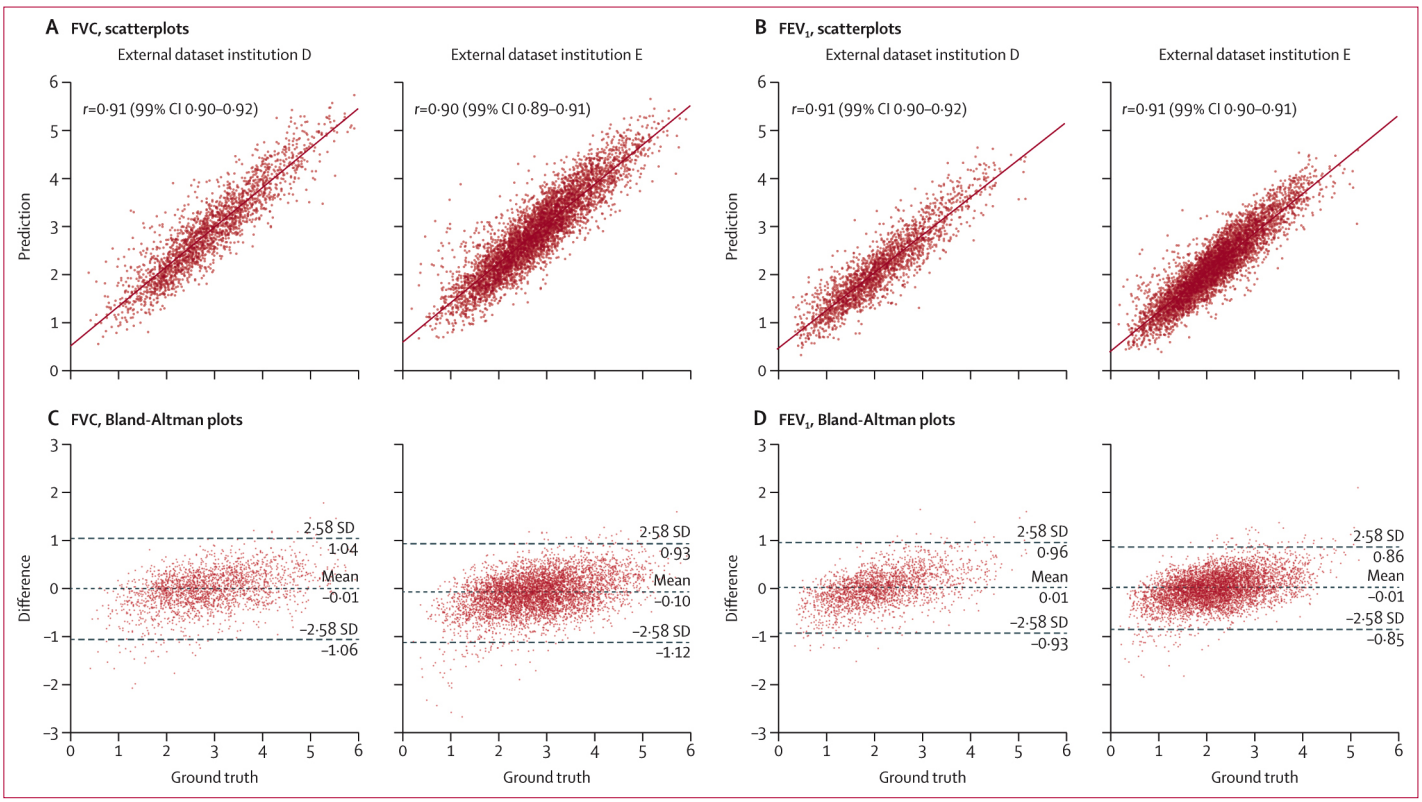

Figure 1: **Predicted *vs* ground truth FVC (A,C) and FEV_1 (B,D) in external test datasets**
Panels A and B show scatterplots of the ground truth and AI predicted values, and panels C and D show Bland-Altman plots. Each scatterplot also includes a regression line. Similar plots for internal test datasets are in the appendix (p 6). AI=artificial intelligence. FVC=forced vital capacity. FEV_1= forced expiratory volume in 1 s.

0·88 (99% CI 0·86–0·90) for institution D and 0·85 (0·83–0·86) for institution E, for FEV_1 of less than 80% predicted was 0·87 (99% CI 0·85–0·89) for institution D and 0·87 (0·85–0·88) for institution E, and for FEV_1/FVC ratio of less than 70% was 0·83 (99% CI 0·80–0·86) for institution D and 0·87 (0·85–0·89) for institution E (figure 2). Accuracies for model-predicted values compared with spirometry results with error ranges and unique patient metrics after duplicate removal are in the appendix (pp 11–12).

Averaged saliency maps for institution E are shown in figure 3. We grouped patients into a high group and a low group, with the high group for each of FVC and FEV_1 consisting of the averaged images of the 10% of patients with the highest values and the low group consisting of the averaged images of the 10% of patients with the lowest values. The AI model primarily focused on lung regions in the x-rays. The averaged saliency maps for both FVC and FEV_1 show that the model gives lower weight to the peripheral lung fields and higher weight to features in the central lung fields. From saliency maps generated using the external dataset from institution D, radiologists observed that hyperinflation of the lung fields and bronchial wall thickening were features of the chest x-rays that were associated with a decrease in FEV_1. Furthermore, they determined that volume loss in the lung fields and reticular shadows at the lung field periphery were findings linked to a decrease in FVC. Detailed reader results and representative saliency images are shown in the appendix (pp 3, 7, 13).

Discussion

We developed and validated a deep learning-based AI model for estimating FVC and FEV_1 from chest x-rays collected from multiple institutions. To our knowledge, this is the first model to estimate FVC and FEV_1 from conventional chest x-rays. This AI model predicts pulmonary function without requiring active patient participation and provides good generalisability across cohorts of patients in Japan.

Compared with previous studies of models that have attempted to estimate pulmonary function from chest imaging, our AI model has several advantages. One recent single-centre AI study estimated FVC and FEV_1 from chest CT images.[18] The correlation coefficients of this study were 0·94 for FVC and 0·91 for FEV_1, which are very similar to those generated from our model. This previous study and our current study indicate that chest imaging, such as chest x-ray and chest CT, are strongly correlated with a dynamic outcome, pulmonary function, despite the fact that imaging is static. Our study differs from the previous study in that we used chest x-ray, which

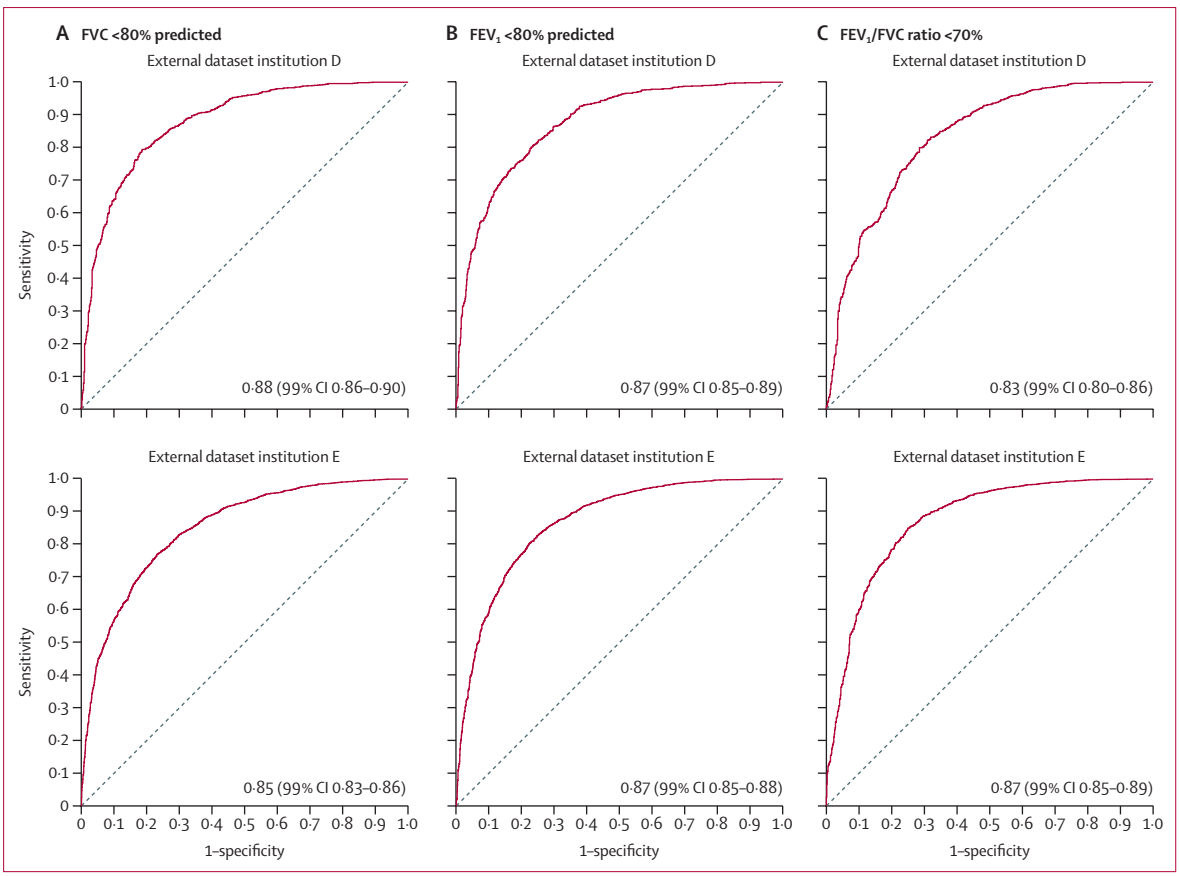

Figure 2: **Receiver operating characteristic curves for classification of FVC <80% predicted (A), FEV_1 <80% predicted (B) and FEV_1/FVC ratio <70% (C) by the AI-based model for external datasets**

AI=artificial intelligence. FVC=forced vital capacity. FEV_1=forced expiratory volume in 1 s.

is more widely available than chest CT, and our research was conducted in a multicentre setting, in institutions across one region of Japan. Other previous studies have reported the use of dynamic digital radiography to determine pulmonary function.[25–28] Dynamic digital radiography is an examination in which chest x-rays are taken during both the expiratory and inspiratory phases. These previous reports have suggested that dynamic chest x-rays correlate with FVC and FEV_1, but the correlation coefficient in each case was 0·9 or less, and these single-centre studies included fewer than 300 participants.[25–28] Although direct comparison with these studies is difficult because the test datasets differed substantially in terms of cohort demographic and clinical characteristics as well as methods used, our model had equal or better performance. Another advantage of our model is that it is applicable to general posterior-anterior view chest x-rays and can be used without changing the usual practice routine. Furthermore, our model requires less exposure to radiation than dynamic digital radiography, which requires sequential x-rays.

One clinical implication of this research is the potential use of our model as a complementary tool to spirometry. First, this model could be an alternative method for patients who can undergo chest x-ray but not spirometry, such as young children, older people who are more likely to be contraindicated, and those with physical or cognitive disabilities for whom spirometry is difficult.[11] Spirometry is the gold standard for evaluating pulmonary function,[3,6] but it requires patient cooperation and the ability to follow specific instructions. Chest x-ray is less time-consuming and more reproducible than spirometry. These advantages enable the use of x-ray for patients for whom estimation of respiratory function with spirometry has been difficult. Of course, as with spirometry, it is important to narrow down the target patient cohort for which the AI model will be used. This is because the impact of overestimation and underestimation cannot be ignored when performed on a broad population with different disease prevalence ratios.[29] Future studies should investigate the performance of this AI model in combination with clinical information to enable more appropriate and targeted use.

Another clinical implication of the model is that the ability to understand pulmonary function from chest x-rays might improve diagnostic scrutiny by allowing appropriate customisation of examination sequences,

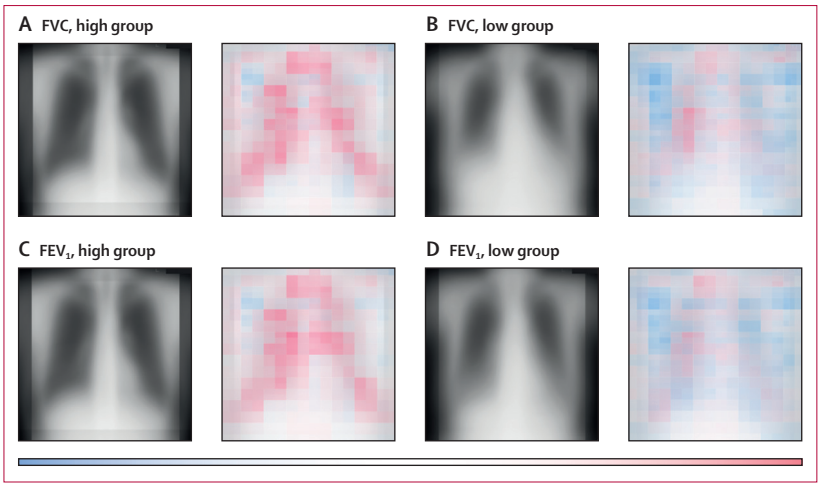

Figure 3: **Averaged saliency maps for external test dataset from institution E**
The top row shows averaged chest x-rays and saliency maps of the FVC estimation, with an averaged chest x-ray and an averaged saliency map of the top 10% of the high FVC images (A) and an averaged chest x-ray and an averaged saliency map of the bottom 10% of the low FVC images (B). The bottom row shows averaged chest x-rays and saliency maps of the FEV_1 estimation, with an averaged chest x-ray and an averaged saliency map of the top 10% of the high FEV_1 images (C) and an averaged chest x-ray and an averaged saliency map of the bottom 10% of the low FEV_1 images (D). The colour bar at the bottom shows the colour ranges, with blue areas showing those that contributed to lowering the values, and the red areas showing those that contributed to raising the metric values. FVC=forced vital capacity. FEV_1=forced expiratory volume in 1 s.

such as subsequent CT imaging.[30] Abnormal FVC and FEV_1 results can suggest airway diseases or interstitial lung diseases, and a customised CT protocol is essential for imaging evaluation in such patients. For large airway disease, volumetric CT scans using axial, coronal, and sagittal views help to determine the structure and range of airway issues (eg, masses, stenosis, wall thickness, and bronchiectasis). For interstitial lung disease, a high-resolution chest CT remains the primary diagnostic tool.[31] The presence of air trapping can be noted by examining images taken during peak inhalation and at the completion of exhalation. By providing an estimate of respiratory function without spirometry, our AI model can aid in properly customising CT scans and could aid in understanding the diagnosis, extent, and severity of lung disease.

When comparing the averaged saliency maps of both FVC and FEV_1 across high and low value groups, we observed a consistent trend: nearly the entire lung field, with the exception of the hilum, transitioned from red (contributing to increasing the metric) to blue (contributing to lowering the metric) regions as severity increased. These findings indicate that our AI model detects changes primarily in the peripheral regions of the lung fields to ascertain reductions in FVC and FEV_1 values. Delving deeper into these observations, over-inflation of the lung fields and the thickening of bronchial walls were seen to correlate with a decrease in FEV_1 by radiologists. Such patterns are commonly observed in conditions such as asthma and COPD, making their association with FEV_1 reasonable.[14–16] Similarly, reductions in lung field volume and the presence of reticular shadows surrounding the lung fields have been linked with a decrease in FVC.[32] These patterns are often observed in patients with interstitial lung disease, aligning well with established medical understanding.[32] Notably, the AI model, although solely trained on chest x-rays paired with their respective FEV_1 and FVC values, demonstrated a capability to discern lung field alterations commonly associated with COPD, asthma, and interstitial lung diseases, emphasising the intricate association between them. Although we postulate that the AI model might be able to recognise additional alterations in chest x-rays, the congruence of identified patterns with changes in the saliency map reinforces the credibility of our model and underscores its potential for interpretability in clinical settings.

We used wide eligibility criteria to improve the AI model's adaptability for a diverse range of patients and diseases. Although this process made the model more versatile, it also posed challenges in maintaining consistent performance due to class imbalance.[33] Our results highlight the importance of refining the model to cater for variability in patient populations.[34] One possible approach is to create AI models tailored for specific diseases. This could lead to more accurate lung function estimates by incorporating disease-specific factors. Additionally, including the disease name as part of the input data might help the AI model to provide better and more precise estimations for different diseases.

Our study has some limitations. First, because this was a retrospective study, a prospective design is needed to evaluate the model more rigorously. Although our multicentre approach strengthens the model's generalisability, the selection of institutions for internal and external testing might introduce bias. Second, the AI model was developed and validated using x-rays collected in Japan, and the dataset most likely comprised Asian patients (data not available). Therefore, assessment of the applicability of the model to different ethnic and racial groups is necessary. Future studies should consider alternative partitioning strategies to mitigate this generalisability and further validate the model's performance across diverse settings.[34] Third, we acknowledge that the current AI model's predictions exhibit variability compared with traditional spirometry, particularly when assessing subgroups of patients with lung diseases (eg, COPD and asthma) rather than the overall patient population. This limitation underscores the necessity for ongoing research to improve the accuracy and reliability of AI-based pulmonary function estimations. Fourth, use of this AI model requires chest x-ray, which involves an exposure of about 0·05 mSv of x-ray radiation. Although the annual natural radiation exposure is 2–3 mSv—about 50 times that of a chest x-ray—repeated examinations can cause cumulative damage. In any clinical setting, unnecessary chest x-rays should be avoided. Finally, some data for this study were collected during the COVID-19 pandemic and so the number of patients presenting for x-ray and spirometry and disease

prevalence ratios of the population during this period were probably affected.

Pulmonary function tests provide quantitative information and are used to elucidate the pathogenesis of respiratory symptoms, assess disease severity, and track the course of disease. Therefore, the use cases for this testing are wide. In this study, we created an AI model that can estimate FVC and FEV_1 with good performance from chest x-rays. The use of this model might provide additional value to chest x-ray through the estimation of pulmonary functions.

Contributors

All authors contributed to the study conception and design. Material preparation, data collection, and data analysis were done by DU, KAb, TK, SF, ToW, and TT. All data were accessed and verified by DU, YMit, and HT. The model was developed by TM and DU. The first draft of the manuscript was written by DU, YMit, and HT, and revised by AY, KAs, and TeW. The manuscript was proofread by SLW. All processes were supervised by YMik. All authors commented on the draft. All authors read and approved the final manuscript. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

We declare no competing interests.

Data sharing

The study protocol and data are available from DU. The source code is available online (https://github.com/xp-spiro/Nervus). Chest x-rays are not available because participating hospitals have withheld them to protect patient privacy.

Acknowledgments

There was no funding for this study. We are grateful to MedCity21, Higashisumiyoshi Morimoto Hospital, Kashiwara Municipal Hospital, and Habikino Medical Center for participating in this research.

References

1 Hutchinson J. On the capacity of the lungs, and on the respiratory functions, with a view of establishing a precise and easy method of detecting disease by the spirometer. *Med Chir Trans* 1846; **29:** 137–252.
2 Kouri A, Dandurand RJ, Usmani OS, Chow C-W. Exploring the 175-year history of spirometry and the vital lessons it can teach us today. *Eur Respir Rev* 2021; **30:** 210081.
3 Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management and prevention of chronic obstructive pulmonary disease (2023 report). Global Initiative for Chronic Obstructive Lung Disease, 2022.
4 Doherty DE. A review of the role of FEV_1 in the COPD paradigm. *COPD* 2008; **5:** 310–18.
5 Burrows B. Airways obstructive diseases: pathogenetic mechanisms and natural histories of the disorders. *Med Clin North Am* 1990; **74:** 547–59.
6 Global Initiative for Asthma. Global strategy for asthma management and prevention: updated 2022. Global Initiative for Asthma, 2022.
7 Kitch BT, Paltiel AD, Kuntz KM, et al. A single measure of FEV1 is associated with risk of asthma attacks in long-term follow-up. *Chest* 2004; **126:** 1875–82.
8 Hole DJ, Watt GCM, Davey-Smith G, Hart CL, Gillis CR, Hawthorne VM. Impaired lung function and mortality risk in men and women: findings from the Renfrew and Paisley prospective population study. *BMJ* 1996; **313:** 711–15.
9 Bang KM, Gergen PJ, Kramer R, Cohen B. The effect of pulmonary impairment on all-cause mortality in a national cohort. *Chest* 1993; **103:** 536–40.
10 Schünemann HJ, Dorn J, Grant BJ, Winkelstein W Jr, Trevisan M. Pulmonary function is a long-term predictor of mortality in the general population: 29-year follow-up of the Buffalo Health Study. *Chest* 2000; **118:** 656–64.
11 Nishi SP, Wang Y, Kuo YF, Goodwin JS, Sharma G. Spirometry use among older adults with chronic obstructive pulmonary disease: 1999–2008. *Ann Am Thorac Soc* 2013; **10:** 565–73.
12 Crimi C, Impellizzeri P, Campisi R, Nolasco S, Spanevello A, Crimi N. Practical considerations for spirometry during the COVID-19 outbreak: literature review and insights. *Pulmonology* 2021; **27:** 438–47.
13 Mettler FA Jr, Bhargavan M, Faulkner K, et al. Radiologic and nuclear medicine studies in the United States and worldwide: frequency, radiation dose, and comparison with other radiation sources—1950–2007. *Radiology* 2009; **253:** 520–31.
14 Burki NK, Krumpelman JL. Correlation of pulmonary function with the chest roentgenogram in chronic airway obstruction. *Am Rev Respir Dis* 1980; **121:** 217–23.
15 Aziz ZA, Wells AU, Desai SR, et al. Functional impairment in emphysema: contribution of airway abnormalities and distribution of parenchymal disease. *AJR Am J Roentgenol* 2005; **185:** 1509–15.
16 Arakawa H, Fujimoto K, Fukushima Y, Kaji Y. Thin-section CT imaging that correlates with pulmonary function tests in obstructive airway disease. *Eur J Radiol* 2011; **80:** e157–63.
17 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521:** 436–44.
18 Park H, Yun J, Lee SM, et al. Deep learning-based approach to predict pulmonary function at chest CT. *Radiology* 2023; **307:** e221488.
19 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; **350:** g7594.
20 Miller MR, Hankinson J, Brusasco V, et al. Standardisation of spirometry. *Eur Respir J* 2005; **26:** 319–38.
21 WHO. The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research. Geneva: World Health Organization, 1993.
22 Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. *Proc CVPR IEEE* 2022; **2022:** 11976–86.
23 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017; **30:** 4765–74.
24 Kubota M, Kobayashi H, Quanjer PH, Omori H, Tatsumi K, Kanazawa M. Reference values for spirometry, including vital capacity, in Japanese adults calculated with the LMS method and compared with previous values. *Respir Investig* 2014; **52:** 242–50.
25 Ueyama M, Hashimoto S, Takeda A, et al. Prediction of forced vital capacity with dynamic chest radiography in interstitial lung disease. *Eur J Radiol* 2021; **142:** 109866.
26 Hino T, Hata A, Hida T, et al. Projected lung areas using dynamic x-ray (DXR). *Eur J Radiol Open* 2020; **7:** 100263.
27 Ohkura N, Tanaka R, Watanabe S, et al. Chest dynamic-ventilatory digital radiography in chronic obstructive or restrictive lung disease. *Int J Chron Obstruct Pulmon Dis* 2021; **16:** 1393–99.
28 Ohkura N, Kasahara K, Watanabe S, et al. Dynamic-ventilatory digital radiography in air flow limitation: a change in lung area reflects air trapping. *Respiration* 2020; **99:** 382–88.
29 Salmi LR, Coureau G, Bailhache M, Mathoulin-Pélissier S. To screen or not to screen: reconciling individual and population perspectives on screening. *Mayo Clin Proc* 2016; **91:** 1594–605.
30 Tseng HJ, Henry TS, Veeraraghavan S, Mittal PK, Little BP. Pulmonary function tests for the radiologist. *Radiographics* 2017; **37:** 1037–58.
31 Raghu G, Collard HR, Egan JJ, et al. An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med* 2011; **183:** 788–824.
32 Clukers J, Lanclus M, Mignot B, et al. Quantitative CT analysis using functional imaging is superior in describing disease progression in idiopathic pulmonary fibrosis compared to forced vital capacity. *Respir Res* 2018; **19:** 213.
33 Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *J Big Data* 2019; **6:** 27.
34 Ueda D, Kakinuma T, Fujita S, et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn J Radiol* 2023; **43:** 3–15.