Original Paper

Deep Learning Estimation of Forced Expiratory Volume in One Second/Forced Vital Capacity and Obstructive Lung Disease Classification From Chest Radiographs With Subgroup Performance Analysis in a North American Cohort: Retrospective Cohort Study

Eptehal Nashnoush[1], BEng, MMA; Helen D'Couto[2], MD; Benjamin Fine[1,3], SM, MD; Leo Anthony Celi[4,5], MPH, MD; Mohamed Abdalla[1,6], PhD

[1]Trillium Health Centre, Mississauga, ON, Canada

[2]MedStar Health, Washington, DC, MD, United States

[3]Department of Medical Imaging, Department of Medical Imaging, University of Toronto, Toronto, ON, Canada

[4]Massachusetts Institute of Technology, Cambridge, MA, United States

[5]Beth Israel Deaconess Medical Center, Boston, MA, United States

[6]Department of Medicine, Faculty of Medicine & Dentistry, University of Alberta, Edmonton, AB, Canada

Corresponding Author:
Eptehal Nashnoush, BEng, MMA
Trillium Health Centre
2200 Eglinton Avenue West
Mississauga, ON, L5M2N1
Canada
Phone: 1 9058132200
Email: e.nashnoush@gmail.com

Abstract

Background: Spirometry is the standard physiological test defining airflow obstruction, the key criterion for diagnosing chronic obstructive pulmonary disease. It is underused in high-income settings and often unavailable in low- and middle-income countries, causing underdetection. Deep learning analysis of chest radiographs, which are widely available where spirometry is not, may complement spirometric screening, but its use in North American cohorts and across demographic strata has not been examined.

Objective: This study aimed to train a deep learning model to estimate the forced expiratory volume in 1 second (FEV_1)/forced vital capacity (FVC) ratio from chest radiographs and classify airflow obstruction (FEV_1/FVC <0.70), evaluate it on a held-out test set, and audit subgroup performance across age, sex, and surname-inferred ethnicity.

Methods: We conducted a retrospective cohort study of 3537 adults who underwent prebronchodilator spirometry and chest radiography within 30 days at a large hospital network in Ontario, Canada, between October 2020 and May 2023. A ConvNeXt-Base architecture pretrained on ImageNet was trained to predict FEV_1/FVC, with predictions classified using a 0.70 cutoff for binary airflow limitation. At the patient level, the cohort was divided into training (n=2263), validation (n=566), and held-out test (n=708 patients; 3273 examinations) sets. Performance was assessed using regression (mean absolute error [MAE], root mean squared error [RMSE], and Pearson r), classification (sensitivity, specificity, positive and negative predictive value [PPV and NPV], and likelihood ratios [LR+ and LR−]), calibration, and decision curve metrics, with 95% CIs from patient-level cluster bootstrap (1000 resamples). Subgroup analyses used Holm correction and two 1-sided tests.

Results: In the held-out test cohort, MAE was 0.08 (95% CI 0.07-0.09) and RMSE was 0.10 (95% CI 0.10-0.11). For binary obstruction, sensitivity was 0.70 (95% CI 0.65-0.74), specificity 0.72 (95% CI 0.67-0.76), PPV 0.71 (95% CI 0.65-0.76), NPV 0.71 (95% CI 0.66-0.76), LR+ 2.46 (95% CI 2.11-2.88), and LR− 0.42 (95% CI 0.36-0.49). Patient-level estimates were similar (sensitivity 0.69, 95% CI 0.66-0.72; specificity 0.74, 95% CI 0.71-0.78). Calibration was excellent for regression (slope=0.97; intercept=0.015) and mildly miscalibrated for the binary task (slope=1.41; intercept=0.04; Brier=0.195). Decision curve analysis showed net benefit at threshold probabilities of approximately 0.27 to 0.86. Sensitivity was meaningfully reduced in Asian patients

(0.43, 95% CI 0.29-0.56) compared with White patients (0.75, 95% CI 0.70-0.79; absolute difference −0.32; Holm P<.001), with accompanying differences in specificity, PPV, and LR−, and was lower in younger age groups, peaking at 65-74 years.

Conclusions: A deep learning model trained on routine chest radiographs estimated FEV_1 /FVC and identified airflow limitation in a North American cohort, with moderate discrimination, well-calibrated regression predictions, and positive net benefit. Performance was not uniform across demographic strata, with reduced sensitivity in Asian patients and younger age groups. Multisite external validation and subgroup-specific verification are important next steps.

(JMIR AI 2026;5:e87770) doi: 10.2196/87770

KEYWORDS

deep learning; chest radiography; pulmonary function tests; obstructive lung disease; forced expiratory volume in one second/forced vital capacity ratio; FEV1/FVC ratio

Introduction

Burden of Obstructive Lung Disease

According to the World Health Organization's Global Burden of Disease estimates, chronic obstructive pulmonary disease (COPD) was the third leading cause of death worldwide in 2020 [1,2]. Although age-standardized COPD incidence and mortality rates have generally fallen over the past decade, absolute case and death counts have risen as populations age and survive earlier-life respiratory exposures [3]. The Burden of Obstructive Lung Disease study estimated the prevalence of chronic airflow obstruction at 11.2% in men and 8.6% in women, with smoking having the highest mean attributable risk globally [4]. Other risk factors include occupational exposures and outdoor and indoor air pollution [1,2]. The macroeconomic burden of COPD has been estimated at approximately US $4.33 trillion (in 2017 international dollars) from 2020 to 2050 [5].

Estimates of obstructive lung disease burden in resource-limited settings and low- and middle-income countries (LMICs) are limited because of variable and poorly defined risk factors, low awareness of obstructive lung disease, and limited reliability of diagnostic tools [6,7]. In high-income countries, tobacco smoke is the leading cause of COPD. In LMICs, there is growing recognition of the role of indoor air pollution from cooking fuels and lighting sources [2,6], which likely explains the higher incidence of COPD in women in LMICs despite lower smoking rates and the earlier onset of disease resulting from childhood exposure to inhaled pollutants [6].

Diagnostic Gaps in Spirometry

Airflow obstruction is defined by spirometry: forced vital capacity (FVC) and forced expiratory volume in 1 second (FEV_1) are the 2 measurements used. The American Thoracic Society, European Respiratory Society, and the Global Initiative for Chronic Obstructive Lung Disease (GOLD) recommend identifying obstruction using an FEV_1 /FVC ratio below the lower limit of normal or below the GOLD fixed threshold of 0.70 [8]. A clinical diagnosis of COPD requires this spirometric criterion in combination with appropriate clinical context, symptoms, exposure history, and exclusion of alternative diagnoses [8].

The availability and standardization of spirometry vary substantially. In LMICs, spirometry frequently requires calibrated equipment and specialist expertise that are not available, and substitute approaches such as symptom-based questionnaires and peak expiratory flow have historically had limited sensitivity and specificity and lack universally accepted diagnostic thresholds [7]. More recently, machine learning models trained on questionnaire data have shown encouraging discrimination for the long-term risk of airflow obstruction [9], suggesting that data-driven approaches to underserved populations are feasible. Spirometry is also limited by the patient's ability to perform the test [6,10].

In practice, spirometry remains underused even in high-income settings. A population-based study in Ontario found that only approximately 42% of adults newly diagnosed with asthma underwent pulmonary function testing within the period from 1 year before to 2.5 years after diagnosis [11]. One Ontario regional hospital performs approximately 600 spirometry tests annually [12], whereas more than 20 million diagnostic X-ray procedures are conducted across Canada each year [13]. These figures highlight the scarcity of spirometry relative to the ubiquity of chest radiography and motivate alternative approaches to assess airflow limitation.

Deep Learning, Chest Imaging, and Prior Work

Quantitative measurements on cross-sectional chest imaging, including emphysema, air trapping, and airway wall thickening on high-resolution computed tomography (CT), have been shown to correlate with FEV_1 and other pulmonary function parameters [14,15]. Deep learning has been applied to thoracic imaging tasks, including detection of pulmonary findings on chest radiographs such as nodule classification [16]. Deep learning has also been used for spirometry quality assurance [17] and structural phenotyping of COPD from spirometric curves [18,19]. More recently, convolutional neural network models have successfully estimated FEV_1 and FVC from low-dose chest CT scans [20-22]. However, CT availability is low in resource-limited and rural settings, and chest radiography is far more prevalent [23,24]. Prior work has shown that deep learning models can estimate spirometric indexes directly from chest radiographs in Japanese cohorts [25]. Whether the underlying radiographic signal is present in other populations and how performance varies across demographic strata remain unclear.

In this study, we train a deep learning model based on a pretrained ConvNeXt architecture to estimate FEV_1 /FVC from chest radiographs and classify airflow obstruction; evaluate the resulting model on a held-out test set from a North American

cohort; and conduct an exploratory subgroup performance audit across age, sex, and surname-inferred ethnicity. We treat this work as an early exploratory study.

Methods

Study Design

We conducted a retrospective cohort study at a large community-based, academically affiliated hospital network in Mississauga, Ontario, Canada, serving a diverse metropolitan catchment area of more than 1 million residents. Eligible patients were adults aged ≥18 years who underwent both a prebronchodilator spirometry test and a chest radiograph within 30 days of each other between October 2020 and May 2023. When multiple pulmonary function tests were available within 30 days of a chest radiograph, only the pair with the shortest time interval was retained. Patients whose chest radiographs were obtained during invasive or noninvasive mechanical ventilation and those with a history of prior lung surgery were excluded. Spirometry reports, including FEV_1 and FVC, were generated by the hospital's pulmonary function testing system. A total of 3537 unique patients met the inclusion criteria. At the patient level, the cohort was divided into training (n=2263, 64%), validation (n=566, 16%), and held-out test (n=708, 20%) sets, with no overlap of patients across the 3 sets. The held-out test set comprised 3273 examinations from 708 unique patients.

Ethical Considerations

This retrospective study was approved by the Trillium Health Partners Research Ethics Board (1177). The requirement for individual informed consent was waived because of the retrospective design and the use of deidentified data. This study is reported in accordance with the TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Plus Artificial Intelligence) statement [26]; the completed checklist is provided in Multimedia Appendix 1.

Data Collection and Preprocessing

Chest radiographs were obtained in DICOM (Digital Imaging and Communications in Medicine) format from the hospital's imaging system and converted to JPEG for model input. The cohort comprised predominantly posteroanterior views, followed by lateral; left lateral; anteroposterior; and, rarely, right anterior oblique projections. Each radiograph was linked to its paired spirometry report by patient and examination identifiers. FEV_1/FVC was computed from the prebronchodilator FEV_1 and FVC values; examinations with missing or invalid values were excluded.

All images were preprocessed with the PyTorch transformation pipeline. Images were resized to 384×384 pixels to match the ConvNeXt input resolution and normalized using the ImageNet mean (0.485, 0.456, 0.406) and SD (0.229, 0.224, 0.225). Training images were augmented with random rotations up to 15°, random horizontal flips, random affine translations of up to 10% in each direction, random resized cropping with scale 0.9 to 1.1, and random color jitter (brightness and contrast up to 0.2). Validation and held-out test images underwent only deterministic preprocessing (resizing and normalization).

Model Architecture and Training

We used a ConvNeXt-Base architecture pretrained on ImageNet-22k and subsequently fine-tuned on ImageNet-1k by the original developers [27] at 384×384 input resolution. ConvNeXt-Base was selected as a strong convolutional architecture with state-of-the-art ImageNet performance [27] and well-validated transfer learning behavior on medical imaging tasks [28]. The model was accessed through the Hugging Face Transformers library [29]. The original 1000-class classification head was replaced with an identity layer, and a single fully connected layer with 1 output node was added on top of the resulting 1024-dimensional global feature embedding to predict the continuous FEV_1/FVC ratio.

The model was trained end-to-end with all parameters updateable, using the Adam optimizer (learning rate 1×10^{-4}), a batch size of 16, and mean squared error loss between the predicted and observed FEV_1/FVC values. A fixed random seed was set to support reproducibility. Training proceeded for 50 epochs on a single NVIDIA V100 graphics processing unit, and the model checkpoint with the lowest validation loss was retained for evaluation on the held-out test set. For the binary airflow obstruction task, predicted FEV_1/FVC values were classified using a 0.70 cutoff, the GOLD fixed criterion for airflow obstruction [8]. The GOLD criterion is defined on postbronchodilator spirometry, whereas this dataset contained prebronchodilator measurements; prebronchodilator spirometry detects somewhat more airflow obstruction but is concordant with postbronchodilator classification in most individuals [30]. The 0.70 threshold reflects this clinical convention applied to the model's continuous FEV_1/FVC output.

Statistical Analysis

Performance Metrics

Model performance on the held-out test set was characterized using both regression and binary classification metrics. Regression metrics were reported on the FEV_1/FVC scale: mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation coefficient. Binary classification metrics were derived from thresholding the predicted ratio at 0.70 and included sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F_1-score, positive likelihood ratio (LR+), and negative likelihood ratio (LR−). Likelihood ratios were reported in addition to predictive values because they are prevalence independent, whereas PPV and NPV depend on the test cohort's obstruction prevalence.

Patient-Level Analyses

All 95% CIs were estimated using a patient-level cluster bootstrap with 1000 resamples, with the patient (rather than individual examinations) as the resampling unit to account for within-patient correlation arising from multiple examinations per patient. CIs were reported as the 2.5th and 97.5th percentiles of the bootstrap distribution.

To complement examination-level estimates, a patient-level sensitivity analysis was conducted by randomly selecting 1 examination per patient and recomputing the metrics. This random selection was repeated 100 times, and the mean and

95% interval across selections are reported alongside the examination-level results.

Calibration and Decision Curve Analysis

Predicted probabilities of obstruction were obtained by Platt scaling [31] of the predicted FEV_1 /FVC ratio. The Platt scaler was fit on the test set; the resulting calibration estimate should therefore be considered internal and modestly optimistic. Binary calibration was summarized by the Brier score, the calibration intercept and slope from a logistic recalibration model, and a calibration plot displaying a locally weighted scatterplot smoothing (loess) curve of observed obstruction frequency against predicted probability across the full range of predictions, with observed frequencies by decile of mean predicted probability overlaid as points with 95% CIs [32]. Regression calibration was additionally summarized by the slope and intercept from a linear regression of observed FEV_1 /FVC on model-predicted FEV_1 /FVC. Decision curve analysis was conducted over threshold probabilities of 0.05 to 0.95, comparing model net benefit against treat-all and treat-none reference strategies.

Subgroup Comparisons

An exploratory subgroup performance audit characterized differences across sex, age group, and surname-inferred ethnicity. Age was stratified into 18 to 44, 45 to 54, 55 to 64, 65 to 74, 75 to 84, and >85 years. The 18- to 24-years and 25- to 44-years age groups were pooled into a single 18- to 44-years stratum because airflow obstruction is uncommon in younger adults, and major epidemiological studies of obstructive lung disease similarly restrict or aggregate younger age groups for this reason [4].

In the absence of self-identified race and ethnicity in the dataset, ethnicity was inferred from patient surnames as a proxy using the averaging-of-proportions method proposed by Abdalla et al [33], which combines US Census surname data (2000 and 2010) to assign each surname a probability distribution over broad racial or ethnic categories. Each surname was matched to the census surname lists directly or by closest spelling match and assigned to its most probable category (White, Asian, Black, or Hispanic); surnames absent from the census reference data even after approximate matching were grouped as unknown and reported descriptively but excluded from formal pairwise comparisons. Subgroup results based on this proxy are exploratory.

Within each grouping variable, every nonreference level was compared against a prespecified reference level. Reference levels were the largest stratum within each grouping variable: female for sex, 65 to 74 years for age, and White for ethnicity. For each metric, the difference between the nonreference and reference level was estimated using a paired patient-level bootstrap with 1000 resamples, with 2-sided *P* values derived from the bootstrap distribution of the difference. Equivalence was assessed using two 1-sided tests (TOST) against a prespecified absolute-difference margin of ±0.05, with a pair classified as equivalent when the 95% bootstrap interval for the difference fell entirely within that margin. Within each grouping variable and metric, *P* values were adjusted for multiple comparisons using Holm step-down procedure.

Results

Overall Classification Performance

The held-out test cohort comprised 708 patients (379 female patients and 329 male patients) contributing 3273 examinations, with an obstruction prevalence of 49.4% (1618/3273). For the continuous FEV_1 /FVC prediction, the model achieved a MAE of 0.08 (95% CI 0.07-0.09), a RMSE of 0.10 (95% CI 0.10-0.11), and a Pearson correlation of 0.60 (95% CI 0.54-0.65). For the binary obstruction task at the 0.70 threshold (Table 1), the model achieved a sensitivity of 0.70 (95% CI 0.65-0.74), specificity of 0.72 (95% CI 0.67-0.76), PPV of 0.71 (95% CI 0.65-0.76), NPV of 0.71 (95% CI 0.66-0.76), and F1 score of 0.70 (95% CI 0.66-0.74). Likelihood ratios were LR+ 2.46 (95% CI 2.11-2.88) and LR− 0.42 (95% CI 0.36-0.49). Patient-level estimates closely matched the examination-level estimates (sensitivity 0.69, 95% CI 0.66-0.72; specificity 0.74, 95% CI 0.71-0.78; PPV 0.68, 95% CI 0.65-0.71; NPV 0.75, 95% CI 0.73-0.77; Table 2).

Table 1. Confusion matrix for binary obstruction classification[a].

True class	Model prediction	
	Predicted: no obstruction	Predicted: obstruction
No obstruction (n=1655, 50.6%), n	1186	469
Obstruction (n=1618, 49.4%), n	489	1129
Total (n=3273), n (%)	1675 (51.2)	1598 (48.8)

[a]Cells show counts of test-set examinations. Row percentages indicate the proportion of true-class examinations classified into each predicted class. Column percentages indicate the proportion of predicted-class examinations belonging to each true class.

Table 2. Overall performance on the held-out test cohort.

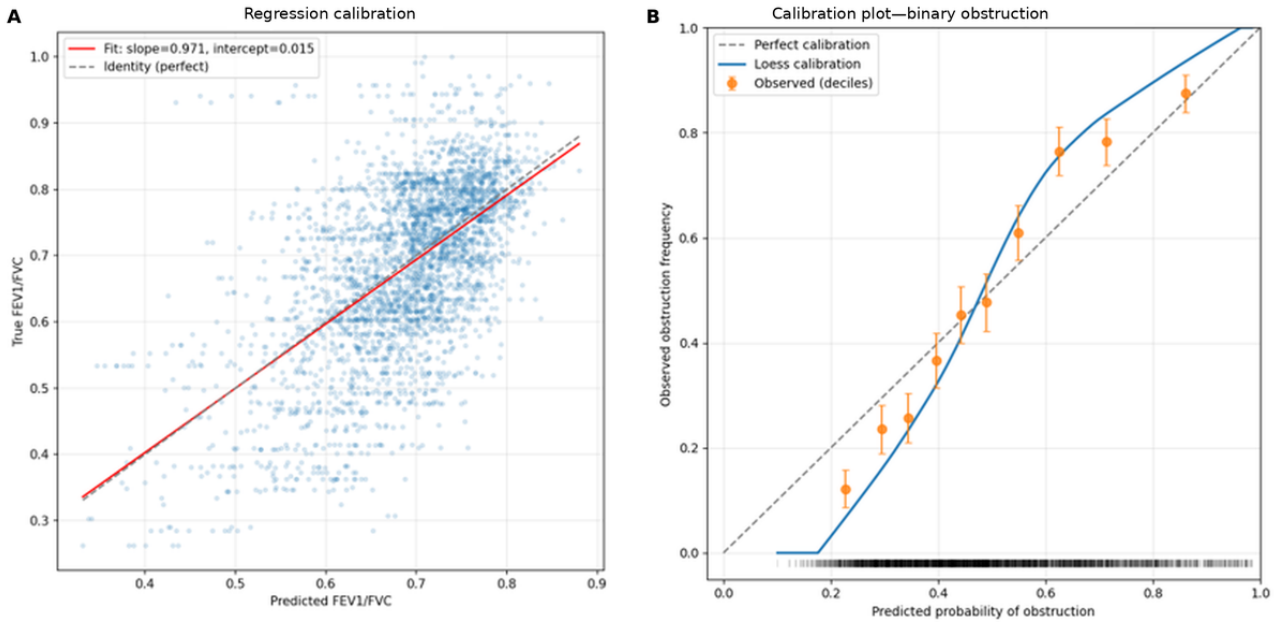
Metric	Examination level, estimate (95% CI)	Patient level, estimate (95% CI)
Sensitivity	0.70 (0.65-0.74)	0.69 (0.66-0.72)
Specificity	0.72 (0.67-0.76)	0.74 (0.71-0.78)
PPV[a]	0.71 (0.65-0.76)	0.68 (0.65-0.71)
NPV[b]	0.71 (0.66-0.76)	0.75 (0.73-0.77)
F_1-score	0.70 (0.66-0.74)	0.69 (0.66-0.71)
LR+[c]	2.46 (2.11-2.88)	2.69 (2.36-3.08)
LR–[d]	0.42 (0.36-0.49)	0.42 (0.38-0.46)
MAE[e]	0.08 (0.07-0.09)	0.08 (0.07-0.08)
RMSE[f]	0.10 (0.10-0.11)	0.10 (0.10-0.10)
Pearson *r*	0.60 (0.54-0.65)	0.60 (0.55-0.64)

[a]PPV: positive predictive value.
[b]NPV: negative predictive value.
[c]LR+: positive likelihood ratio.
[d]LR–: negative likelihood ratio.
[e]MAE: mean absolute error.
[f]RMSE: root mean squared error.

Calibration and Decision Curve Analysis

Regression calibration was excellent, with a slope of 0.97 and an intercept of 0.015 estimated from the linear regression of observed vs predicted FEV_1 /FVC values (Figure 1A). Binary probability calibration showed mild miscalibration (Brier score=0.195; calibration intercept=0.04; calibration slope=1.41; Figure 1B), with the loess calibration curve indicating slight overprediction at low predicted probabilities and underprediction at high predicted probabilities.

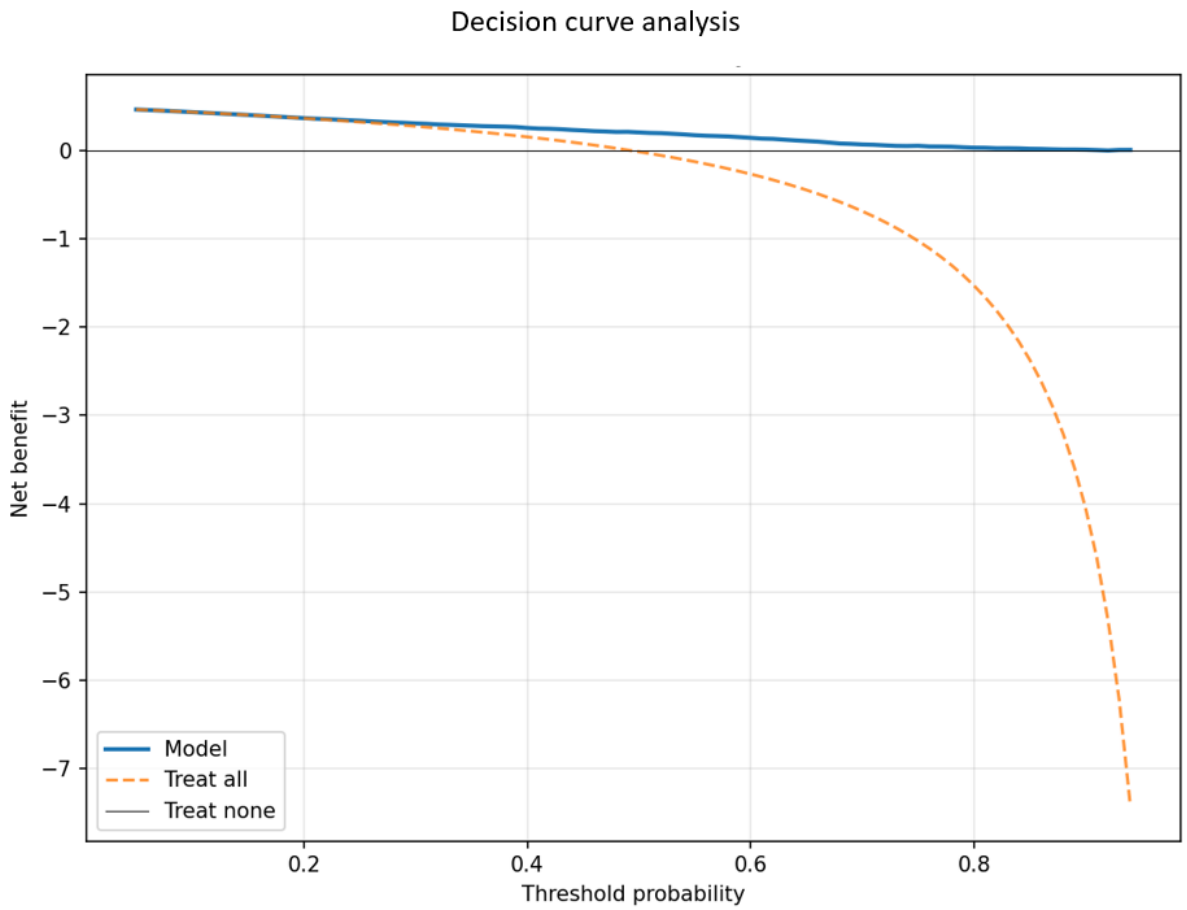
Figure 1. Calibration of the forced expiratory volume in 1 second (FEV_1)/forced vital capacity (FVC) and binary obstruction predictions on the held-out test cohort. (A) Regression calibration: scatter of observed against predicted FEV_1 /FVC, with the identity line shown for reference. (B) Binary obstruction calibration: observed obstruction frequency plotted against predicted probability of obstruction. The solid line is a loess-smoothed calibration curve fitted across the full range of predicted probabilities; points show observed frequencies by decile of mean predicted probability with 95% CIs. The dashed identity line indicates perfect calibration.



Decision curve analysis over threshold probabilities of 0.05 to 0.95 showed positive net benefit relative to both treat-all and treat-none strategies across the threshold range of 0.08 to 0.94 (Figure 2). However, the incremental net benefit over the treat-all strategy was minimal at threshold probabilities below approximately 0.27, and above approximately 0.86 the model curve approached the treat-none strategy. At a net-benefit margin of 0.02, improvement over both reference strategies was

concentrated in the intermediate threshold range of approximately 0.27 to 0.86.

Figure 2. Decision curve analysis on the held-out test cohort.



Subgroup Performance

Overview

Subgroup performance was characterized across sex, age group, and surname-inferred ethnicity using paired patient-level bootstrap with 2-sided *P* values and Holm correction within each grouping variable and metric. Equivalence at a ±0.05 margin was assessed using TOST. The full pairwise comparison matrix is provided in Multimedia Appendix 2. Of the 54 comparisons, 13 reached statistical significance after Holm correction. No comparison met the ±0.05 equivalence criterion, which requires the 95% CI of the difference to fall entirely within that margin.

Performance by Sex

Performance estimates by sex are reported in Table 3. Sensitivity was higher among male patients (0.74, 95% CI 0.68-0.79) than among female patients (0.65, 95% CI 0.59-0.71), with the pairwise difference (absolute difference +0.08, 95% CI−0.00 to +0.17) borderline nonsignificant after Holm correction (Holm *P*=.052). PPV was significantly higher in male patients (absolute difference +0.13, 95% CI +0.03 to +0.23, Holm *P*=.02), consistent with the higher obstruction prevalence in this subgroup (898/1583, 56.7% vs 720/1690, 42.6%). Differences in specificity, NPV, and likelihood ratios did not reach statistical significance.

Table 3. Performance by sex and surname-inferred ethnicity.

Subgroup	Examinations, N	Prevalence, n (%)	Sensitivity (95% CI)	Specificity (95% CI)	PPV[a] (95% CI)	NPV[b] (95% CI)	LR+[c] (95% CI)	LR−[d] (95% CI)
Sex								
Female	1690	720 (42.6)	0.65 (0.59-0.71)	0.72 (0.67-0.77)	0.64 (0.54-0.72)	0.73 (0.67-0.80)	2.33 (1.92-2.93)	0.48 (0.39-0.58)
Male	1583	898 (56.7)	0.74 (0.68-0.79)	0.71 (0.65-0.77)	0.77 (0.70-0.83)	0.67 (0.60-0.74)	2.54 (2.07-3.17)	0.37 (0.30-0.46)
Surname-inferred ethnicity								
White (reference)	2054	1232 (60)	0.75 (0.70-0.79)	0.67 (0.62-0.73)	0.77 (0.72-0.82)	0.64 (0.57-0.71)	2.26 (1.93-2.75)	0.38 (0.31-0.46)
Asian	616	191 (31)	0.43 (0.29-0.56)	0.81 (0.74-0.88)	0.51 (0.31-0.67)	0.76 (0.65-0.86)	2.29 (1.38-3.70)	0.70 (0.54-0.89)
Hispanic	172	45 (26.2)	0.68 (0.43-0.91)	0.64 (0.50-0.83)	0.40 (0.16-0.70)	0.85 (0.69-0.97)	1.90 (1.11-4.16)	0.50 (0.13-0.92)
Black	107	51 (47.7)	0.82 (0.52-1.00)	0.82 (0.65-0.97)	0.81 (0.36-0.97)	0.84 (0.60-1.00)	4.61 (2.06-15.95)	0.22 (0.00-0.60)
Unknown	318	99 (31.1)	0.57 (0.40-0.74)	0.73 (0.63-0.82)	0.48 (0.29-0.68)	0.79 (0.67-0.89)	2.07 (1.34-3.31)	0.60 (0.36-0.84)

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]LR+: positive likelihood ratio.

[d]LR−: negative likelihood ratio.

Performance by Age Group

Performance estimates by age group are reported in Table 4. Sensitivity was significantly lower in the 18-44 year stratum (0.30, 95% CI 0.15-0.46) than in the 65-to-74-year reference (0.76, 95% CI 0.69-0.82; absolute difference −0.46; Holm P<.001) and in the 45-to-54-year stratum (0.46, 95% CI 0.25-0.64; absolute difference −0.32; Holm P=.008). Specificity was significantly higher in both younger strata (18 to 44 years: absolute difference +0.32; Holm P<.001; 45 to 54 years: absolute difference +0.26; Holm P<.001), and NPV was significantly higher (18 to 44 years: absolute difference +0.22; Holm P=.03; 45 to 54 years: absolute difference +0.20; Holm P=.03). LR− was significantly worse in the 18-to-44-year stratum (absolute difference +0.36; Holm P=.01). Performance in the 55-to-64-year, 75-to-84-year, and >85-year age groups did not differ significantly from the 65-to-74-year reference group on any metric after Holm correction.

Table 4. Performance by age group.

Age group (years)	Exams, N	Prevalence, n (%)	Sensitivity (95% CI)	Specificity (95% CI)	PPV[a] (95% CI)	NPV[b] (95% CI)	LR+[c] (95% CI)	LR−[d] (95% CI)
18-44	303	55 (18.2)	0.30 (0.15-0.46)	0.92 (0.88-0.96)	0.46 (0.21-0.73)	0.86 (0.77-0.93)	3.88 (1.70-9.53)	0.76 (0.58-0.93)
45-54	310	71 (22.9)	0.46 (0.25-0.64)	0.86 (0.79-0.92)	0.49 (0.22-0.72)	0.85 (0.76-0.92)	3.33 (1.65-6.29)	0.63 (0.42-0.88)
55-64	622	255 (41)	0.68 (0.57-0.78)	0.74 (0.65-0.82)	0.65 (0.49-0.78)	0.77 (0.68-0.85)	2.62 (1.87-3.87)	0.43 (0.29-0.61)
65-74 (reference)	924	536 (58)	0.76 (0.69-0.82)	0.60 (0.51-0.69)	0.72 (0.62-0.81)	0.64 (0.52-0.76)	1.90 (1.51-2.52)	0.40 (0.28-0.55)
75-84	850	535 (62.9)	0.73 (0.67-0.80)	0.63 (0.54-0.71)	0.77 (0.68-0.84)	0.58 (0.46-0.69)	1.97 (1.59-2.56)	0.43 (0.31-0.55)
>85	264	166 (62.9)	0.65 (0.49-0.87)	0.49 (0.34-0.66)	0.68 (0.48-0.85)	0.45 (0.23-0.79)	1.26 (0.85-1.99)	0.73 (0.29-1.26)

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]LR+: positive likelihood ratio.

[d]LR−: negative likelihood ratio.

Performance by Surname-Inferred Ethnicity

Performance estimates by surname-inferred ethnicity are reported in Table 3. Asian patients (n=616 examinations) showed a multimetric pattern of conservative prediction relative to White patients: sensitivity was significantly lower (0.43 vs 0.75; absolute difference −0.32; Holm P<.001), specificity significantly higher (0.81 vs 0.67; absolute difference +0.14; Holm P=.006), PPV significantly lower (0.51 vs 0.77; absolute difference −0.26; Holm P=.01), and LR− significantly worse (absolute difference +0.33; Holm P<.001). NPV did not differ significantly. Hispanic patients (n=172 examinations) showed significantly higher NPV than White patients (0.85 vs 0.64; absolute difference +0.21; Holm P=.048), consistent with lower obstruction prevalence in this subgroup (45/172, 26.2% vs 1232/2054, 60%). Sensitivity, specificity, PPV, and likelihood ratios did not differ significantly. Black vs White comparisons (n=107 examinations in the Black subgroup) were statistically inconclusive on all metrics, with wide CIs reflecting the small subgroup size.

Discussion

Principal Findings

This study provides early exploratory evidence that a deep learning model can extract signals predictive of FEV_1 /FVC and airflow obstruction from routine chest radiographs in a North American cohort. The model showed moderate discrimination, well-calibrated regression predictions, mild miscalibration in the binary task, and positive decision curve net benefit, with improvements over treat-all and treat-none reference strategies concentrated in the intermediate threshold range of approximately 0.27 to 0.86.

Performance was not uniform across demographic strata. The patient-level cluster bootstrap with Holm correction identified meaningfully reduced sensitivity in Asian patients relative to White patients (absolute difference −0.32; Holm P<.001), paired with higher specificity (+0.14; Holm P=.006), lower PPV (−0.26; Holm P=.01), and worse LR− (+0.33; Holm P<.001). This pattern is consistent with the model operating more conservatively in this subgroup, where the observed obstruction prevalence was also lower. Substantial age-related variation was documented: sensitivity was significantly lower in the 18- to 44-years (−0.46; Holm P<.001) and 45- to 54-years (−0.32; Holm P=.008) age groups relative to the 65 to 74 years reference, paired with significantly higher specificity in both younger strata. This again is a conservative pattern in subgroups with lower observed obstruction prevalence. A statistically significant difference in NPV between Hispanic and White patients (+0.21; Holm P=.048) was also identified; given NPV's prevalence dependence and the lower obstruction prevalence in the Hispanic subgroup, this finding likely reflects prevalence rather than differential discrimination. Sex differences in sensitivity were borderline nonsignificant after correction (+0.08; Holm P=.052); PPV was significantly higher in male patients (+0.13; Holm P=.02), consistent with higher obstruction prevalence in the male subgroup.

Limitations

The study has several limitations that should be acknowledged. The retrospective single-center design means that findings may not fully generalize without further external validation, which we have not yet performed. All performance estimates were derived from a single random training, validation, or test partition; although the patient-level cluster bootstrap quantifies metric uncertainty conditional on this split, performance may vary under alternative partitions. Ethnicity was inferred from patient surnames using a census-derived proxy; this approach does not capture self-identified race, ethnicity, or social context and carries some risk of misclassification. Subgroup sample sizes vary across strata; in particular, the Black subgroup (n=107 examinations) was small, and the Black vs White comparison was statistically inconclusive on all metrics. Equivalence at a ±0.05 margin (TOST) was not established for any subgroup pair, reflecting the precision limit of the test cohort rather than signaling broad demographic disparities. Turning to calibration, binary probability calibration showed mild miscalibration (slope=1.41), with predicted probabilities slightly compressed toward the middle of the range. Regression calibration of the underlying continuous prediction was excellent (slope=0.97; intercept=0.015), suggesting that the binary miscalibration arises from the thresholding step rather than from systematic bias in the regression. The 0.70 threshold was additionally applied to prebronchodilator spirometry rather than the postbronchodilator measurement in the GOLD definition; because prebronchodilator testing identifies more obstruction, the model may overidentify obstruction relative to a postbronchodilator reference standard [30].

Future Directions

Several directions for future work emerge naturally from these findings. Multisite external validation in independent North American and international cohorts is the highest-priority next step, both to assess generalizability and to provide an opportunity to recalibrate binary predictions in cohorts with representative disease prevalence. Prospective evaluation with self-identified demographic data would address the surname-proxy limitation and enable subgroup-specific net benefit to be formally characterized. Additionally, integrating clinical metadata such as smoking history and comorbidities could further improve model performance. The present study used a single ConvNeXt-Base architecture, which does not establish optimal predictive performance for this task. Systematic comparisons of alternative architectures, as well as ensemble and stacking approaches, may yield meaningful differences in performance and represent a natural direction for future investigation.

Conclusions

This study adds to a growing body of evidence that chest radiographs carry signals predictive of pulmonary function that can be extracted by deep learning models. In a North American cohort, the model achieved moderate discrimination, well-calibrated regression predictions, and positive decision curve net benefit, but performance varied across demographic strata, particularly in Asian patients and younger age groups. Multisite external validation and subgroup-specific performance

verification are needed to establish whether these findings generalize beyond the present cohort.

Acknowledgments

The authors would like to thank the radiology and pulmonary function technologists at the participating institution for assistance with data collection. The authors used ChatGPT (OpenAI) to assist with drafting portions of the manuscript. The authors reviewed and revised the content and take full responsibility for the final manuscript.

Data Availability

The data supporting this study consist of linked clinical and imaging records from Trillium Health Partners and cannot be made publicly available due to patient privacy regulations and institutional data governance policies. Researchers interested in replication may contact the corresponding author to discuss access through the Trillium Health Partners Research Ethics Board. The analysis code and trained model weights are available from the corresponding author upon request. The model was implemented in Python using PyTorch and the Hugging Face Transformers library; the ConvNeXt-Base architecture is publicly available on Hugging Face [29].

Funding

This research received no external funding.

Conflicts of Interest

None declared.

Multimedia Appendix 1

TRIPOD+AI checklist.
[PDF File (Adobe PDF File), 316 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Pairwise subgroup comparisons with Holm-corrected *P* values and outcomes of two one-sided tests for equivalence.
[PDF File (Adobe PDF File), 267 KB-Multimedia Appendix 2]

References

1. Chronic respiratory disease is third leading cause of death globally with air pollution killing 1.3 million people. Institute for Health Metrics and Evaluation. Apr 25, 2023. URL: https://www.healthdata.org/news-release/chronic-respiratory-disease-third-leading-cause-death-globally-air-pollution-killing-13 [accessed 2025-10-19]
2. Chronic obstructive pulmonary disease (COPD). World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd) [accessed 2025-10-19]
3. Wang Y, Han R, Ding X, Feng W, Gao R, Ma A. Chronic obstructive pulmonary disease across three decades: trends, inequalities, and projections from the Global Burden of Disease Study 2021. Front Med (Lausanne). Mar 24, 2025;12:1564878. [FREE Full text] [doi: 10.3389/fmed.2025.1564878] [Medline: 40196348]
4. Burney P, Patel J, Minelli C, Gnatiuc L, Amaral AF, Kocabaş A, et al. Prevalence and population-attributable risk for chronic airflow obstruction in a large multinational study. Am J Respir Crit Care Med. Jun 01, 2021;203(11):1353-1365. [FREE Full text] [doi: 10.1164/rccm.202005-1990OC] [Medline: 33171069]
5. Chen S, Kuhn M, Prettner K, Yu F, Yang T, Bärnighausen T, et al. The global economic burden of chronic obstructive pulmonary disease for 204 countries and territories in 2020-50: a health-augmented macroeconomic modelling study. Lancet Glob Health. Aug 2023;11(8):e1183-e1193. [FREE Full text] [doi: 10.1016/S2214-109X(23)00217-6] [Medline: 37474226]
6. Lin CH, Cheng SL, Chen CZ, Chen CH, Lin SH, Wang HC. Current progress of COPD early detection: key points and novel strategies. Int J Chron Obstruct Pulmon Dis. Jul 19, 2023;18:1511-1524. [FREE Full text] [doi: 10.2147/COPD.S413969] [Medline: 37489241]
7. Ho T, Cusack RP, Chaudhary N, Satia I, Kurmi OP. Under- and over-diagnosis of COPD: a global perspective. Breathe (Sheff). Mar 2019;15(1):24-35. [FREE Full text] [doi: 10.1183/20734735.0346-2018] [Medline: 30838057]
8. Global strategy for prevention, diagnosis and management of COPD: 2025 report. Global Initiative for Chronic Obstructive Lung Disease. URL: https://goldcopd.org/2025-gold-report/ [accessed 2025-10-19]
9. Perret JL, Vicendese D, Simons K, Jarvis DL, Lowe AJ, Lodge CJ, et al. Ten-year prediction model for post-bronchodilator airflow obstruction and early detection of COPD: development and validation in two middle-aged population-based cohorts. BMJ Open Respir Res. Dec 2021;8(1):e001138. [FREE Full text] [doi: 10.1136/bmjresp-2021-001138] [Medline: 34857526]
10. Schneider A, Gindner L, Tilemann L, Schermer T, Dinant GJ, Meyer FJ, et al. Diagnostic accuracy of spirometry in primary care. BMC Pulm Med. Jul 10, 2009;9:31. [FREE Full text] [doi: 10.1186/1471-2466-9-31] [Medline: 19591673]

11. Gershon AS, Victor JC, Guan J, Aaron SD, To T. Pulmonary function testing in the diagnosis of asthma: a population study. Chest. May 2012;141(5):1190-1196. [doi: 10.1378/chest.11-0831] [Medline: 22030804]
12. GBGH adding full pulmonary function testing to enhance access for local patients. Georgian Bay General Hospital. URL: https://gbgh.on.ca/gbgh-adding-full-pulmonary-function-testing/ [accessed 2025-10-19]
13. Periard MA, Chaloner P. Diagnostic X-ray imaging quality assurance: an overview. Can J Med Radiat Technol. Oct 1996;27(4):171-177. [FREE Full text]
14. Koo HJ, Lee SM, Seo JB, Lee SM, Kim N, Oh SY, et al. Prediction of pulmonary function in patients with chronic obstructive pulmonary disease: correlation with quantitative CT parameters. Korean J Radiol. Apr 2019;20(4):683-692. [FREE Full text] [doi: 10.3348/kjr.2018.0391] [Medline: 30887750]
15. Ostridge K, Williams NP, Kim V, Harden S, Bourne S, Clarke SC, et al. Relationship of CT-quantified emphysema, small airways disease and bronchial wall dimensions with physiological, inflammatory and infective measures in COPD. Respir Res. Feb 20, 2018;19(1):31. [FREE Full text] [doi: 10.1186/s12931-018-0734-y] [Medline: 29458372]
16. Hendrix W, Hendrix N, Scholten ET, Mourits M, Trap-de Jong J, Schalekamp S, et al. Deep learning for the detection of benign and malignant pulmonary nodules in non-screening chest CT scans. Commun Med (Lond). Oct 27, 2023;3(1):156. [FREE Full text] [doi: 10.1038/s43856-023-00388-5] [Medline: 37891360]
17. Wang Y, Li Y, Chen W, Zhang C, Liang L, Huang R, et al. Deep learning for spirometry quality assurance with spirometric indices and curves. Respir Res. Apr 21, 2022;23(1):98. [FREE Full text] [doi: 10.1186/s12931-022-02014-9] [Medline: 35448995]
18. Bodduluri S, Nakhmani A, Reinhardt JM, Wilson CG, McDonald ML, Rudraraju R, et al. Deep neural network analyses of spirometry for structural phenotyping of chronic obstructive pulmonary disease. JCI Insight. Jul 09, 2020;5(13):e132781. [FREE Full text] [doi: 10.1172/jci.insight.132781] [Medline: 32554922]
19. Geng K, Shi Z, Zhao X, Ali A, Wang J, Leader J, et al. BeyondCT: a deep learning model for predicting pulmonary function from chest CT scans. arXiv. Preprint posted online on August 10, 2024. [doi: 10.48550/arXiv.2408.05645]
20. Guo W, Li M, Li Y, Fan X, Wu L. Differentiating emphysema from emphysema-dominated COPD patients with CT imaging feature and machine learning. Int J Chron Obstruct Pulmon Dis. Jul 25, 2025;20:2615-2628. [FREE Full text] [doi: 10.2147/COPD.S527914] [Medline: 40734725]
21. Wu Y, Xia S, Liang Z, Chen R, Qi S. Artificial intelligence in COPD CT images: identification, staging, and quantitation. Respir Res. Aug 22, 2024;25(1):319. [FREE Full text] [doi: 10.1186/s12931-024-02913-z] [Medline: 39174978]
22. Park H, Yun J, Lee SM, Hwang HJ, Seo JB, Jung YJ, et al. Deep learning-based approach to predict pulmonary function at chest CT. Radiology. Apr 2023;307(2):e221488. [doi: 10.1148/radiol.221488] [Medline: 36786699]
23. Yoshida A, Kai C, Futamura H, Oochi K, Kondo S, Sato I, et al. Spirometry test values can be estimated from a single chest radiograph. Front Med (Lausanne). Mar 6, 2024;11:1335958. [FREE Full text] [doi: 10.3389/fmed.2024.1335958] [Medline: 38510449]
24. Silverberg M. How radiologists overcome barriers to provide imaging in low to middle income countries. Radiological Society of North America News. Jul 11, 2024. URL: https://www.rsna.org/news/2024/july/imaging-in-lmics [accessed 2025-10-19]
25. Ueda D, Matsumoto T, Yamamoto A, Walston SL, Mitsuyama Y, Takita H, et al. A deep learning-based model to estimate pulmonary function from chest x-rays: multi-institutional model development and validation study in Japan. Lancet Digit Health. Aug 2024;6(8):e580-e588. [FREE Full text] [doi: 10.1016/S2589-7500(24)00113-4] [Medline: 38981834]
26. Collins GS, Moons KG, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. Apr 16, 2024;385:e078378. [FREE Full text] [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
27. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. Presented at: CVPR 2022; Jun 18-24, 2022; New Orleans, LA. [doi: 10.1109/cvpr52688.2022.01167]
28. Hosseinzadeh Taher MR, Haghighi F, Gotway MB, Liang J. Large-scale benchmarking and boosting transfer learning for medical image analysis. Med Image Anal. May 2025;102:103487. [doi: 10.1016/j.media.2025.103487] [Medline: 40117988]
29. facebook / convnext-base-384-22k-1k. Hugging Face. URL: https://huggingface.co/facebook/convnext-base-384-22k-1k [accessed 2025-10-19]
30. Singh D, Stockley R, Anzueto A, Agusti A, Bourbeau J, Celli BR, et al. GOLD Science Committee recommendations for the use of pre- and post-bronchodilator spirometry for the diagnosis of COPD. Eur Respir J. Feb 06, 2025;65(2):2401603. [FREE Full text] [doi: 10.1183/13993003.01603-2024] [Medline: 39638416]
31. Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola AJ, Bartlett PL, Schölkopf B, Schuurmans D, editors. Advances in Large Margin Classifiers. Cambridge, MA. MIT Press; 1999:61-74.
32. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. Dec 16, 2019;17(1):230. [FREE Full text] [doi: 10.1186/s12916-019-1466-7] [Medline: 31842878]

33. Abdalla M, Abdalla S, Maurer LR, Ortega G, Abdalla M. American Black authorship has decreased across all clinical specialties despite an increasing number of Black physicians between 1990 and 2020 in the USA. J Racial Ethn Health Disparities. Apr 2024;11(2):710-718. [doi: 10.1007/s40615-023-01554-0] [Medline: 36877380]

Abbreviations

COPD: chronic obstructive pulmonary disease
CT: computed tomography
DICOM: Digital Imaging and Communications in Medicine
FEV1: forced expiratory volume in 1 second
FVC: forced vital capacity
GOLD: Global Initiative for Chronic Obstructive Lung Disease
LMIC: low- and middle-income country
LR–: negative likelihood ratio
LR+: positive likelihood ratio
MAE: mean absolute error
NPV: negative predictive value
PPV: positive predictive value
RMSE: root mean squared error
TOST: two 1-sided tests
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Plus Artificial Intelligence

Edited by Y Huo; submitted 13.Nov.2025; peer-reviewed by D Vicendese, T Lee; comments to author 17.Mar.2026; revised version received 27.Jun.2026; accepted 29.Jun.2026; published 05.Aug.2026

Please cite as:
Nashnoush E, D'Couto H, Fine B, Celi LA, Abdalla M
Deep Learning Estimation of Forced Expiratory Volume in One Second/Forced Vital Capacity and Obstructive Lung Disease Classification From Chest Radiographs With Subgroup Performance Analysis in a North American Cohort: Retrospective Cohort Study
JMIR AI 2026;5:e87770
URL: https://ai.jmir.org/2026/1/e87770
doi: 10.2196/87770
PMID: