Society of Medical Education & Research

2026, Volume 6, Issue 1, Page No: 120-133

ISSN: 3108-4826

Journal of Medical Sciences and Interdisciplinary Research

Predictive Performance of Machine Learning and Deep Learning Techniques in Estimating Long-Term Outcomes among Patients with Chronic Obstructive Pulmonary Disease: A Systematic Review and Meta-Analysis

Elena Petrova[1]*, Ivan Georgiev[1], Nikolay Stoyanov[2], Petar Kolev[3]

[1] Department of Pulmonary Medicine and Predictive Analytics, Faculty of Medicine, Sofia University, Sofia, Bulgaria.

[2] Department of Systematic Review and Meta-Analysis in COPD, Faculty of Medicine, University of Plovdiv, Plovdiv, Bulgaria.

[3] Department of Long-Term Outcomes and AI, Faculty of Medicine, University of Ruse, Ruse, Bulgaria.

***E-mail** ✉ elena.petrova@uni-sofia.bg

Abstract

Machine learning and deep learning approaches are increasingly used to forecast long-term disease trajectories in individuals with chronic obstructive pulmonary disease (COPD). This study sought to synthesize the effectiveness of these prognostic tools for COPD, evaluate their performance relative to one another, and pinpoint major areas needing further investigation.We conducted a systematic review and meta-analysis to assess the effectiveness of machine learning- and deep learning-based prognostic models and to outline directions for future studies. We systematically searched PubMed, Embase, the Cochrane Library, ProQuest, Scopus, and Web of Science from their launch until April 6, 2023, targeting English-language publications that used machine learning or deep learning techniques to predict patient outcomes at least 6 months after the first clinical encounter in people diagnosed with COPD. Eligible studies included adults aged 18 to 90 years and accepted any input data. We presented the area under the receiver operating characteristic curve (AUC) along with 95% confidence intervals (CI) for forecasts of death, exacerbations, and reductions in forced expiratory volume in 1 s (FEV1). Heterogeneity across studies was quantified via Cochran's Q test, where notable heterogeneity was indicated by $P \leq 0.10$ or $I^2 > 50\%$. We evaluated reporting standards with the TRIPOD checklist and conducted risk-of-bias evaluations using the PROBAST tool. The review was prospectively registered on PROSPERO (CRD42022323052). The initial literature search yielded 3620 records. After screening, 18 studies met the inclusion standards, including 12 that relied on traditional machine learning and 6 that utilized deep learning. Seven models focused on exacerbation probability, though only six provided AUC values with 95% CI from internal validation sets (combined AUC 0.77 [95% CI: 0.69–0.85]), accompanied by substantial heterogeneity (I^2 97%, $P < 0.0001$). Eleven models addressed mortality risk, with six reporting AUC and 95% CI on internal validation data (pooled AUC 0.77 [95% CI: 0.74–0.80]) and moderate heterogeneity (I^2 60%, $P = 0.027$). Two investigations examined declines in lung function but could not be aggregated. Machine learning and deep learning models did not show meaningful gains compared with established disease severity indices for exacerbation prediction ($P = 0.24$). Three studies offered head-to-head comparisons of machine learning against conventional severity scores for mortality, revealing no difference in pooled results ($P = 0.57$). Among the five studies conducting external validation, model performance was similar to or inferior to standard regression approaches. Key sources of bias included improper management of missing values, omission of model uncertainty estimates, and reliance on datasets undersized relative to the number of predictor variables. Available data provide scant support that conventional machine learning or deep learning prognostic models outperform existing disease severity scoring systems. Stronger compliance with established reporting standards in future work would reduce bias risks and improve replicability.

Keywords: Chronic obstructive pulmonary disease, Machine learning, Deep learning, Prognostic models, Exacerbation, Mortality

Access this article online https://smerpub.com/

Received: 11 December 2025; **Accepted:** 17 March 2026

How to cite this article: Petrova E, Georgiev I, Stoyanov N, Kolev P. Predictive Performance of Machine Learning and Deep Learning Techniques in Estimating Long-Term Outcomes among Patients with Chronic Obstructive Pulmonary Disease: A Systematic Review and Meta-Analysis. J Med Sci Interdiscip Res. 2026;6(1):120-33. https://doi.org/10.51847/VF5CIwlab0

Introduction

Chronic obstructive pulmonary disease (COPD) ranks as the third leading global cause of mortality [1]. The economic burden associated with COPD care is projected to rise further across the world [2, 3], making it valuable to detect individuals likely to experience swift disease

advancement. Such patients could receive intensified management and monitoring, thereby optimizing healthcare resource use.

Multiple indicators—such as pulmonary function assessments, symptom questionnaires, physical endurance measures, and exacerbation history—help gauge the chance of adverse outcomes. Composite tools like the body-mass index, airflow obstruction, dyspnoea, and exercise capacity (BODE) index tend to deliver stronger predictive power than individual factors alone [2]. An earlier broad systematic review by Bellou and colleagues [4] from 2019 cataloged over 400 prognostic models for COPD clinical outcomes. Their meta-analysis of linear regression models (n = 12) produced pooled C-statistics (comparable to AUC [5]) for mortality prediction spanning 0.624 to 0.769. These figures indicate that regression-based models possess only moderate accuracy and limited practical influence on COPD forecasting [6].

COPD represents a major worldwide cause of death and exhibits wide variability in how the condition evolves, which traditional prognostic tools have struggled to capture fully. We queried PubMed, Scopus, and Web of Science for English-language reviews of long-term COPD prognostic models published between January 1, 2016, and March 29, 2022, employing keywords linked to “COPD,” “machine learning,” and “prediction.” Two prior systematic reviews on COPD prognostic models highlighted deficiencies in methodological quality and validation practices, rendering them unsuitable for routine clinical application. Neither review contrasted machine learning performance or advantages against standard regression models. A narrative overview on artificial intelligence and machine learning applications in various COPD clinical settings noted the growth of extensive, multifaceted datasets that support expanded use of AI methods in healthcare.

As far as we are aware, no previous reviews have specifically examined and contrasted prognostic models for COPD outcomes that rely on machine learning and deep learning algorithms. This evaluation establishes a reference point for the capabilities of machine learning models while highlighting existing shortcomings in methodological and reporting practices that must be addressed for continued advancement.

Going forward, investigators creating machine learning algorithms should adhere closely to PROBAST and TRIPOD recommendations to produce transparent, low-bias results that can be reproduced. At present, the clinical utility of conventional machine learning applied to structured clinical data remains constrained by the extensive number of variables involved. Conversely, deep learning analysis of CT imaging holds growing potential for integration into everyday practice as an incidental evaluation once additional advancements occur. Drawing from these findings, we advocate for expanded exploration of multimodal approaches that combine imaging with clinical data; creation of sizable, openly accessible datasets beyond North America and Europe; and formulation of reporting standards akin to PROBAST tailored for deep learning applications.

Machine learning involves computational methods that automatically discover rules and associations between inputs and outputs from data, supporting hypothesis-free predictions [7]. Deep learning, a subfield of machine learning, uses layered architectures to capture intricate patterns and connections within datasets [8]. It excels in medical imaging [9] and has delivered encouraging outcomes in image-based clinical prognosis tasks [10, 11]. Growing attention has focused on deep learning to develop novel prognostic and diagnostic tools for identifying COPD patients at risk of deterioration, given that current statistical models have not integrated details on lung structural alterations visible through imaging [12].

In the present systematic review and meta-analysis, we aimed to appraise the effectiveness and methodological soundness of published machine learning and deep learning prognostic models in COPD populations and to benchmark them against earlier regression-based predictive models [4]. To gauge clinical applicability, we prioritized model performance irrespective of specific architecture or chosen variables.

Materials and Methods

Search strategy and selection criteria

This systematic review and meta-analysis included studies applying machine learning or deep learning models to forecast mortality, exacerbation likelihood, or lung function deterioration in adult COPD patient groups at least 6 months (180 days) after baseline clinical assessment. We excluded animal research and cohorts composed mainly of individuals under 18 or over 90 years old. Studies predicting COPD onset in people without the disease at baseline were omitted, as were those lacking a minimum 6-month follow-up period. We accepted cohort studies (prospective or retrospective),

case-control designs, randomized controlled trials, and cross-sectional analyses. No restrictions were placed on geographic origin or publication venue. Machine learning was characterized as algorithms (for instance, random forests or support vector machines) that exceed basic regression in complexity and learn decision-making from data patterns. Deep learning was defined as neural networks featuring two or more hidden layers. Complete eligibility standards framed by the population, intervention, comparison, outcome, and time framework [13].

We searched PubMed, Embase, the Cochrane Library, Scopus, Web of Science, and ProQuest. The queries incorporated terms connected to chronic obstructive pulmonary disease, artificial intelligence, machine learning, deep learning, prediction, and prognosis. We targeted articles published in English or accompanied by English translations, spanning from the databases' origins up to April 6, 2023. We also manually reviewed reference lists from the included articles and relevant systematic reviews to capture any overlooked publications.

All records were imported into Covidence, where duplicate entries were eliminated. Two reviewers (LAS and AB) independently evaluated titles and abstracts for every study, erring on the side of inclusion when eligibility was unclear. Full-text versions of studies that advanced past the initial stage underwent detailed examination by the same two reviewers. Disagreements were resolved through discussion between LAS and AB whenever feasible; otherwise, a third independent reviewer (LJP) made the final determination.

The study protocol was prospectively registered on PROSPERO (CRD42022323052) and followed the PRISMA guidelines [14].

Data analysis

Before beginning data extraction, we created a standardized extraction form informed by the CHARMS checklist [15]. Eleven specific data elements were collected from each eligible study.

We classified studies into two groups: those employing deep neural networks (the deep learning category) and those that did not (the conventional machine learning category). When a single study reported several models for the same endpoint, we selected the one with the strongest performance that met the inclusion criteria. For studies presenting comparable outcomes across different time horizons (for example, mortality risk at 1, 2, or 3 years), we chose the timeframe that aligned most closely with outcomes in other included research. Data extraction was performed by LAS and independently verified by AB and MZ, with any discrepancies settled by LJP.

Risk of bias for each study was evaluated using the PROBAST tool [16]. This instrument includes 20 items organized into four domains: participant selection, predictors, outcome, and analysis. Each item was rated as yes, probably yes, probably no, no, or no information. A "yes" response signaled low risk of bias, while "no" indicated high risk. Any "no" or "probably no" rating on one or more items resulted in an overall high-risk-of-bias classification for the study; otherwise, it was considered low risk. Reporting transparency was appraised with the TRIPOD checklist, which consists of 22 essential items for clear description of model development and validation [17].

We measured the models' ability to discriminate between patients for outcomes including mortality, exacerbation, or decline in forced expiratory volume in 1 s (FEV1) by extracting AUC values [18] together with their 95% confidence intervals. Hospital admission served as a proxy for exacerbation events. Surrogate indicators for FEV1 decline—such as reduced performance on the 6-minute walk test, shifts in dyspnoea levels, or changes in COPD-related burden assessed via questionnaires or psychophysical scales—were considered, although no qualifying studies were identified. AUC estimates were combined separately for internal validation and external validation datasets for each outcome. Given the expected substantial variability arising from differences in model types, input variables, patient populations, and follow-up durations, we applied a random-effects model for pooling and displayed the findings in forest plots. Pooling was performed without regard to specific model architecture or selected features. Interstudy heterogeneity was quantified using Cochran's Q test [19], with significant heterogeneity defined as $P \leq 0.10$ or $I^2 > 50\%$. This allowed us to determine whether a fixed-effects approach was appropriate. All meta-analyses were conducted with MedCalc statistical software (version 20.116). To benchmark pooled machine learning and deep learning results against conventional regression models, we combined performance metrics from regression-based models for mortality and exacerbation risk previously reported by Bellou and colleagues [4] and used the MedCalc online calculator [20] to compare AUCs. Where feasible, we also performed direct within-study comparisons between deep learning or machine learning

models and established risk scores or linear regression tools (such as the BODE index) using the same dataset.

Results and Discussion

The database search initially retrieved 3620 records. After screening titles and abstracts, followed by full-text reviews for eligibility, 17 studies [21-37] satisfied the inclusion and exclusion criteria (**Figure 1**). Examination of reference lists identified one additional qualifying study [38], bringing the total to 18.

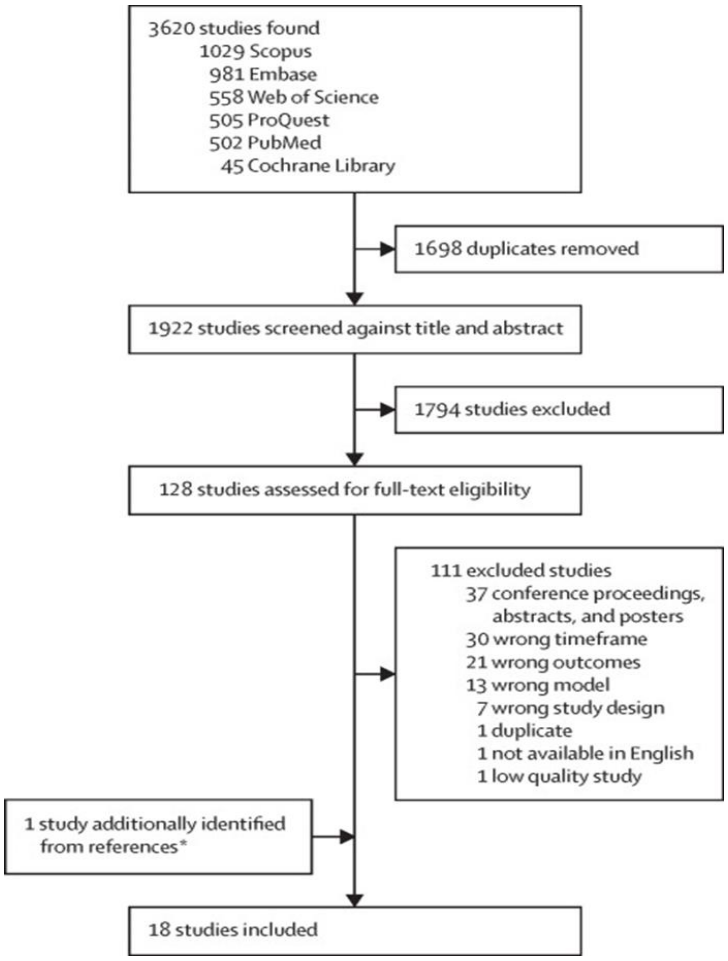

Figure 1. Selection of studies. *To achieve thorough coverage of all relevant publications, LAS reviewed any further articles cited within the reference sections of the qualifying papers, checked them against the predefined eligibility standards, and integrated qualifying ones into an additional cycle of the systematic review.

Two investigations examined predictions of worsening lung function [26, 35]. Seven papers addressed the likelihood of disease exacerbations [21, 23, 28, 29, 32, 33, 37], while eleven focused on the risk of death [21-27, 30, 31, 34, 36, 38]. Two papers evaluated both exacerbation and mortality risks simultaneously [21, 23]. **Table 1** displays combined overviews of key study features—such as reported counts and median figures—separately for the deep learning group, the conventional machine learning group, and the entire collection of papers. Participant age, gender distribution, and site of enrollment were routinely described across the research, but details on participants' ethnic backgrounds were largely absent. Every included study carried out internal validation, most often via a reserved test portion or k-fold cross-validation techniques. External validation, however, was conducted in only four deep learning papers [21, 22, 24, 25] and a single conventional machine learning paper [30]. Of those five, three [21, 22, 30] were built on data from the COPDGene cohort [39] and tested

on the ECLIPSE cohort [40]. The remaining pair originated from South Korean patient populations and were externally assessed on different South Korean [24] and Malaysian [25] groups.

Table 1. Distribution of reported study characteristics and pooled summary estimates.

Characteristic	Machine learning (n=12)		Deep learning (n=6)		Combined analysis (n=18)	
	Studies reported	Pooled value*	Studies reported	Pooled value*	Studies reported	Pooled value*
Age	11 (92%)	63.5 (59.7–67.7)	4 (67%)	67.1 (64.4–71.6)	15 (83%)	67 (60.9–68.6)
Male sex	10 (83%)	65% (53–83)	4 (67%)	57% (50–60)	14 (78%)	57% (52–69)
Location	12 (100%)	..	6 (100%)	..	18 (100%)	..
Ethnicity†	4 (33%)	32% (32–33)	2 (33%)	13% (11–16)	6 (33%)	18% (13–31)
Internal validation	12 (100%)	..	6 (100%)	..	18 (100%)	..
External validation	1 (8%)	..	4 (67%)	..	5 (28%)	..
COPDGene (internal dataset)	2 (17%)	..	3 (50%)	..	5 (28%)	..
ECLIPSE (external dataset)	1 (8%)	..	2 (33%)	..	3 (17%)	..
AUC	10 (83%)	..	4 (67%)	..	14 (78%)	..
Confidence intervals	5 (42%)	..	4 (67%)	..	9 (50%)	..

Data are presented as n (%) unless otherwise specified.

* Pooled estimates are expressed as median values with interquartile ranges (IQR).

† For studies conducted in the United States, median values represent the proportion of participants identified as Black or African-American.

Six investigations employed deep learning techniques for COPD outcome forecasting [21-25, 38], while the remaining twelve relied on conventional machine learning approaches [26-37].

Table 2 presents performance metrics for the deep learning models, and **Table 3** presents those for the machine learning models.

Table 2. Model performance for deep learning.

Reference	Outcome	AUC (95% CI)	Sample size	Other performance indicators
González *et al.* [21]	≥ 1 exacerbation in 1-year period; ≥ 1 exacerbation in 3-year period	Internal validation 0.64 (0.60–0.68); external validation 0.54 (0.51–0.57)	Training 6016 (COPDGene); internal validation 1000; external validation 1672 (ECLIPSE)	..
Singla *et al.* [23]	≥ 1 exacerbation in 5 years	0.73 (0.71–0.75)*	Training 10 300 (5-fold cross-validation)	AUPRC 0.42 (SD: 0.02), recall 0.47 (SD: 0.04), accuracy 80.83%
González *et al.* [21]	3-year all-cause mortality	Internal validation 0.72 (0.63–0.81); external validation 0.60 (0.56–0.64)	Training 5740 (COPDGene); internal validation 1000; external validation 1672 (ECLIPSE)	Internal validation HR: 2.69 (1.19–6.05), P = 0.017; external validation HR: 1.64 (1.19–2.26), P = 0.003
Humphries *et al.* [22]	Mortality risk‡	..	Training 2407; internal validation 7143; external validation 1962	Internal validation HR (5 strata): 1.5 (1.0–2.2), 1.6 (1.1–2.4), 2.4 (1.6–3.5), 2.7 (1.8–4.2), 2.9 (1.7–4.9); external validation HR (5 strata): 1.5 (1.0–2.2), 1.7

				(1.1–2.5), 2.9 (2.0–4.3), 5.3 (3.6–7.7), 99.7 (6.3–14.8)
Nam *et al*. [24]	5-year all-cause mortality	Internal validation (hold-out) 0.81 (0.74–0.88); internal validation (temporal) 0.76 (0.70–0.82); external validation (VHSMC) 0.72 (0.66–0.78); external validation (AMC) 0.83 (0.78–0.88)	Training 3475; internal validation (hold-out) 315; internal validation (temporal) 394; external validation (VHSMC) 416; external validation (AMC) 337	..
Singla *et al*. [23]	5-year all-cause mortality	0.62 (no CI calculated)	Training 10 300 (5-fold cross-validation)	HR: 1.54 (1.09–2.17), P < 0.0001
Tang *et al*. [38]	1-year all-cause mortality	..	Training 10 850; internal validation 4650	Accuracy 78.89% (no CI)
Yun *et al*. [25]	3-year and 5-year all-cause mortality	Internal validation 0.80 (0.72–0.88); external 0.72 (0.57–0.86)	Training 344 (5-fold cross-validation); external validation 102	..

AUC values are shown together with newly computed 95% confidence intervals. Hazard ratios are presented as HR (95% CI). AMC refers to the Asan Medical Centre dataset. AUC stands for area under the receiver operating characteristic curve. AUPRC denotes area under the precision-recall curve. HR indicates hazard ratio. VHSMC stands for the Veterans Health Service Medical Centre dataset.

*The total number of cases was not provided in the original report; therefore, the originally published 95% CI was retained rather than deriving a new one.

† The median duration of follow-up was 7.95 years for the internal validation set and 2.90 years for the external validation set.

Table 3. Model performance for conventional machine learning.

Reference	Outcome	AUC (95% CI)	Sample size	Other performance indicators
Boueiz *et al*. [26]	Change in FEV_1 after 5-year period; new FEV_1 after 5-year period	Internal validation (cross-validation) 0.71 (0.69–0.72); internal validation (temporal) 0.70*	Training 4496 (10-fold cross-validation); internal validation (temporal) 1833	..
Sharma *et al*. [35]	Decline in $FEV_1 \geq 30$ mL per year over 3 years	0.81 (0.66–0.96)	Training 42 (5-fold cross-validation)	Accuracy 84.6%, sensitivity 0.875, specificity 0.800
Kor *et al*. [28]	≥ 1 exacerbation in 6 months (for patients who have not previously exacerbated)	0.83 (0.74–0.92)	Training 407; internal validation 102	Sensitivity 79.41%, specificity 77.94%, PPV 64.29%, NPV 88.33%, F1 score 71.05%, accuracy 78.43%
Le [29]	≥ 1 exacerbation in 1 year	0.67 (0.67–0.68)	Training 190 117 (10-fold cross-validation)	..

Nguyen [33]	≥ 1 hospitalization in 180-day period	0.88 (0.81–0.95)	Training 81; internal validation 27	Accuracy 88%, sensitivity 0.83, specificity 0.93
Zeng *et al.* [37]	≥ 1 exacerbation in 1-year period	0.866 (0.83–0.90)	Training 36 047; internal validation 7529	Accuracy 90.3% (89.6–91.0), sensitivity 56.6 (49.2–64.2), specificity 91.2 (90.5–91.8), PPV 13.7 (11.2–16.2), NPV 98.8 (98.6–99.1)
Moslemi *et al.* [32]	≥ 1 hospitalization in 3 years	0.78 (0.69–0.87)	Training 328; internal validation 141†	Accuracy 78%, F1 score 71%
Esteban *et al.* [27]	5-year all-cause mortality	0.74 (0.68–0.80)	Training 611; internal validation 348	..
Moll *et al.* [30]	All-cause mortality risk‡	Internal validation 0.731 (0.682–0.780); external validation 0.688 (0.655–0.721)	Training 1974; internal validation 658; external validation 1268	..
Morales *et al.* [31]	3-year all-cause mortality	0.80 (0.79–0.80)	Training 163 587; internal validation 40 895†	..
Pinto-Plata *et al.* [34]	3-year all-cause mortality	..	Training 90; internal validation 30 (100% mortality)	Accuracy 85%, sensitivity 81%, specificity 89%
Tang *et al.* [36]	1-year all-cause mortality	..	Training 10 850; internal validation 4650	Accuracy 64.6%

All reported AUC values include newly computed 95% confidence intervals. Hazard ratios are displayed as HR (95% CI). AUC stands for area under the receiver operating characteristic curve. HR represents the hazard ratio. PPV indicates positive predictive value. NPV denotes negative predictive value. FEV1 refers to forced expiratory volume in 1 s.

* The regression model lacked event counts, preventing calculation of 95% CIs; the values shown are those originally published in the study.

† Training and validation group sizes were derived from the overall cohort size and the reported percentages allocated to each split.

‡ Median follow-up duration was 6.4 years for internal validation and 7.2 years for external validation.

The combined AUC for forecasting exacerbations using both machine learning and deep learning models on internal validation data reached 0.77 (95% CI: 0.69–0.85); (**Figure 2**). Assessment of variability via Cochran's Q test revealed considerable heterogeneity (I^2 97%, P < 0.0001). Although the number of deep learning studies addressing exacerbation risk was insufficient for separate pooling, machine learning studies could be aggregated, yielding a pooled AUC of 0.80 (95% CI: 0.72–0.87, I^2: 97%, P < 0.0001). A single investigation [21] conducted external validation for exacerbation prediction and recorded a decline in accuracy from an AUC of 0.64 (95% CI: 0.58–0.71) internally to 0.55 (95% CI: 0.51–0.57) externally.

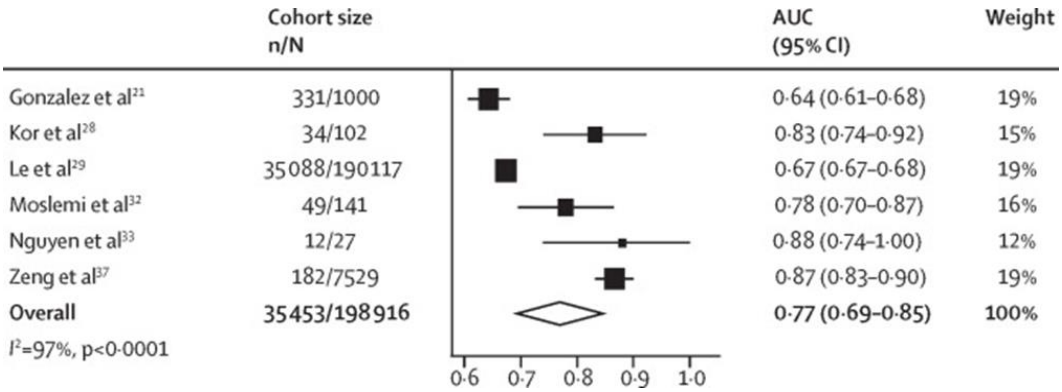

Figure 2. Random-effects forest plot showing models for predicting exacerbation risk.

n indicates the number of exacerbation events, and N denotes the total sample size. P-values were obtained from Cochran's Q test. AUC stands for area under the receiver operating characteristic curve.

The combined AUC for mortality prediction using internal validation datasets was 0.77 (95% CI: 0.74–0.80), accompanied by notable heterogeneity (I^2: 60%, P = 0.027); (**Figure 3**). When examined separately, deep learning studies produced a pooled AUC of 0.78 (95% CI: 0.73–0.84, I^2: 22%, P = 0.28), while machine learning studies yielded 0.77 (95% CI: 0.73–0.80, I^2: 80%, P = 0.0070). Statistical comparison of the pooled AUCs between machine learning and deep learning methods for mortality showed no meaningful difference (P = 0.44). Four studies [21, 24, 25, 30] performed external validation for mortality prediction, yielding a pooled AUC of 0.67 (95% CI: 0.62–0.72, I^2: 79%, P = 0.0023). For these same studies, the pooled internal validation performance was substantially better (P = 0.0049), with an AUC of 0.76 (95% CI: 0.72–0.80, I^2: 42%, P = 0.16).

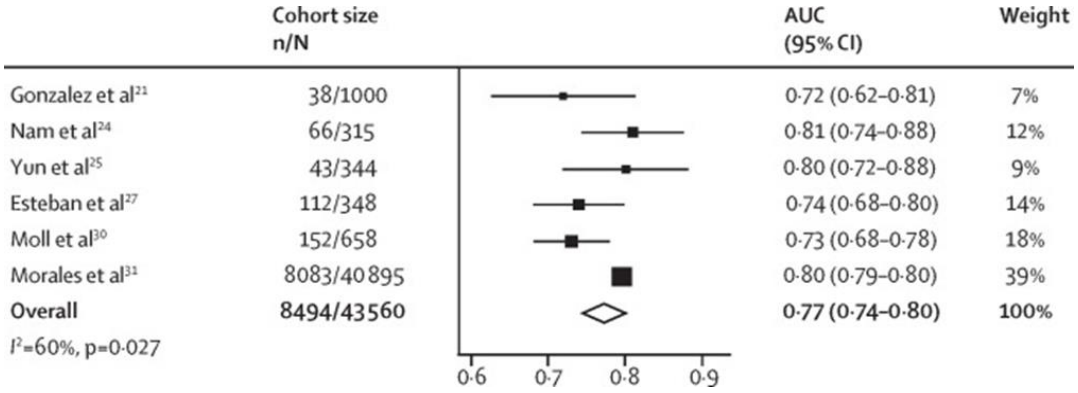

Figure 3. Random-effects forest plot displaying models for mortality risk prediction

n denotes the count of events (exacerbations), and N indicates the overall group size. P values derive from Cochran's Q test. AUC stands for area under the receiver operating characteristic curve.

Since only two investigations provided data on lung function decline, pooling of AUC values was not feasible. One study predicted a drop in lung function exceeding 30 mL across 5 years, achieving an AUC of 0.81 (95% CI: 0.66–0.96) in a group of 42 patients via 5-fold cross-validation [35]. Another study forecasted the general reduction in lung volume with an AUC of 0.71 (95% CI: 0.69–0.72) in a cohort of 4496 patients using ten-fold nested cross-validation (**Table 3**) [26].

The aggregated predictive accuracy of regression models earlier summarized by Bellou and colleagues [4] stood at 0.66 (95% CI: 0.49–0.82) for exacerbation forecasting and 0.69 (95% CI: 0.68–0.71) for mortality forecasting. For mortality prognosis, both machine learning (P = 0.0001) and deep learning (P < 0.0001) approaches demonstrated markedly superior performance compared with these regression models. For exacerbation prognosis, machine learning studies produced a numerically greater pooled estimate than that reported by Bellou and colleagues [4]. Yet, the contrast did not reach statistical significance (P = 0.14 when pooling only machine learning studies and P = 0.24 when pooling all studies).

A limited subset of studies enabled head-to-head evaluation of deep learning or machine learning models against traditional risk scores or regression models within identical datasets. Three deep learning investigations [21, 23, 24] compared mortality prediction against the BODE index, with two drawing from the same patient population (COPDGene) [21, 23]. These papers noted slight, non-significant gains in AUC for mortality using deep learning relative to the BODE index. Seven machine learning studies additionally supplied data on standard indices (such as the BODE index or the age, dyspnoea, airflow obstruction index) or regression models built with the same input variables on the same data [26, 27, 30, 31, 35-37]. Five of these [26, 27, 30, 35, 37] described better results with the machine learning model, one found equivalent performance [31], and one observed marginally inferior results for the machine learning approach [36]. Among the five showing improvement, only one indicated a statistically significant advantage [30]. Three of the seven studies [27, 30, 31] reported AUCs and shared the mortality

outcome, allowing meta-analysis. This analysis found no meaningful difference in pooled AUC between machine learning models (pooled AUC 0.77) and conventional regression models (pooled AUC 0.75; P = 0.57).

Compliance with TRIPOD recommendations was generally strong, as only one study [23] covered fewer than 70% of the required items. Fourteen out of 18 studies (78%) failed to supply or link to the complete prediction model (item 15a), and nine out of 18 (50%) omitted confidence intervals around the model's performance metrics (item 16). According to the PROBAST evaluation, every included study carried a high overall risk of bias. The analysis domain was a source of high risk of bias in all cases. Inappropriate management of missing data represented the most common issue (item 4.4): two studies [36, 38] had no missing values, three [23, 28, 29] provided no details on handling methods, one [37] applied a model that naturally accommodates missing data, one [34] relied on minimum-value imputation, and the others performed complete-case analysis without confirming the missing-completely-at-random assumption. Nine of the 18 studies (50%) did not report confidence intervals for any model outcomes (item 4.7). The events-per-variable (EPV) ratio—calculated as the number of positive outcome events divided by the number of predictor variables—was examined for item 4.1, with an EPV below 10 signaling high risk of bias [41]. Six of the 12 machine learning studies (50%) [28, 33–37] had an EPV under 10 in the training set, showing a median of 82 positive events (IQR 33–1560) and a median of 154 predictor variables in the final model (IQR 45–412). The other six machine learning studies had EPV values above 20, with a median of 1645 positive events (IQR 244–4075) and a median of 11.5 predictor variables (IQR: 6–39).

We conducted a systematic review and meta-analysis of machine learning and deep learning studies focused on COPD prognosis, identifying 18 eligible investigations. The endpoints examined included mortality, exacerbation events, and deterioration in lung function. Findings indicate that both machine learning and deep learning methods may deliver greater predictive accuracy than traditional regression models, regardless of specific architecture or chosen variables—especially when forecasting mortality. Nevertheless, this conclusion must be tempered by the universally high risk of bias across studies and the results observed during external validation. Primary factors driving high risk of bias included inadequate management of missing data, limited sample sizes, and failure to report confidence intervals.

Marked heterogeneity between studies for each clinical endpoint complicated direct performance comparisons, stemming from variations in outcome time windows, patient characteristics, input features, and modeling techniques. Even so, we recognize that substantial heterogeneity is to be expected when developing deep learning models, since factors such as random weight initialization alone can influence training outcomes [42]. Consequently, aggregating results despite high heterogeneity remains acceptable when potential sources of variation are acknowledged.

Pooled results for machine learning and deep learning models on internal validation data surpassed the pooled performance of regression models from external validation sets. Yet, among the five models subjected to external validation (one for exacerbation and four for mortality), external performance declined by roughly 0.09 AUC on average compared with internal results. It proved inferior to or comparable with regression models. Robust conclusions about superiority require thorough external validation, which remains a widespread methodological shortcoming in machine learning and deep learning research [4, 43, 44].

Nearly every study exhibited high risk of bias arising from unsuitable handling of missing data (for instance, relying on complete-case analysis without verifying missing-completely-at-random assumptions). PROBAST guidelines advocate multiple imputation to limit bias while maintaining data distribution [41]; however, comparable strategies for imputing absent imaging information remain an active research topic [45]. Regarding overfitting risk, an EPV exceeding 20 generally denotes low risk of bias, whereas an EPV below 10 signals high risk [41]. The six studies with EPV under 10 featured both modest datasets with sparse outcome events and models incorporating numerous predictor variables. When expanding cohort size is not possible, applying multivariate feature selection to trim the number of inputs can help mitigate overfitting [41].

Deep learning studies typically utilized larger datasets than machine learning studies, with median training cohort sizes of 6016 (IQR: 2407–10,300) versus 1292.5 (IQR: 183–29,747), respectively. In broad terms, deep learning models tend to achieve stronger performance and better generalizability with bigger training sets. Despite this pattern, we observed a qualitative inverse association between deep learning mortality model

accuracy and training cohort size, suggesting potential overfitting to the training data. Regularization methods can curb overfitting in deep learning with little detriment to training performance; common options include dropout [46], batch normalization [47], and data augmentation [48]. Traditional notions of overfitting and regularization do not translate straightforwardly to deep learning, as these models can achieve near-perfect fits to training data (what would classically be viewed as overfitting) while still delivering strong test performance and external generalizability [49]. At present, evaluating overfitting risk in deep learning for prognostic applications remains difficult beyond external validation, underscoring the need for specialized guidelines analogous to PROBAST in the deep learning setting.

Although more than 75% of the studies documented age and sex characteristics, and the recruitment location was clear in every case, reporting of participants' ethnicity remained inadequate. For thorough evaluation and comparison across investigations, complete disclosure of cohort demographics is essential. This is particularly important given the well-documented racial and sex-related differences in COPD risk and progression [50, 51]. Self-reported race has lately been identified as a potentially significant confounder in image-based deep learning assessments [52]. Since ethnicity data are unavailable for the ECLIPSE cohort—a frequently used validation set in major COPD research—there is a clear need for new, sizable studies that fully describe participant demographics. Except for three papers, all studies originated from North America or Western Europe (specifically the USA, Canada, UK, or Spain). The remaining three were conducted in South Korea [24, 25] and Taiwan [28]. Consequently, applying these findings to low- and middle-income countries—where COPD exerts its greatest impact—requires considerable caution [53].

Because of modest external generalizability and a pervasive high risk of bias, current evidence does not strongly support the notion that machine learning or deep learning models outperform conventional regression approaches, even with their greater sophistication. That said, continued development in these areas holds promise for enhanced generalizability and reduced bias in the future. In particular, creating dedicated risk-of-bias tools—modeled after PROBAST—specifically for deep learning model development and validation would help address issues such as missing data, generalizability, dataset scale, and reporting practices, thereby standardizing methodology and elevating evidence quality.

From a practical clinical standpoint, machine learning tools demanding numerous input variables offer limited value to clinicians who must assess individual patients rapidly and cannot realistically collect dozens of data points per case. In contrast, established risk scores (for example, the BODE index) and deep learning models applied to CT imaging show greater promise for real-world use. The expanding role of CT scans in differential diagnosis, lung cancer screening, and preoperative evaluation [2] creates natural opportunities for CT-based deep learning tools to provide prognostic insights opportunistically, without imposing extra workload on clinicians—assuming the models have been rigorously validated in integrated clinical settings [54].

Moreover, effective deep learning model development depends on access to extensive high-quality datasets and well-crafted algorithmic designs. For imaging, established public longitudinal cohorts such as COPDGene and ECLIPSE supply valuable high-quality data suitable as benchmarks for model testing and validation. Nevertheless, these collections remain modest in scale compared with massive imaging repositories used in computer science (such as ImageNet, which contains over 14 million images). Additional work is required to advance deep learning techniques capable of properly processing CT scans. Among the reviewed studies, only Singla and colleagues [23] utilized three-dimensional volumetric data rather than two-dimensional slices. Exploring multimodal integration represents another encouraging avenue. Moslemi and colleagues [32], Nam and colleagues [24], and Singla and colleagues [23] each demonstrated that merging imaging and clinical information in a unified prognostic model yielded superior results compared with using either data type in isolation. Furthermore, emerging large language models open fresh research possibilities [55]. These systems have shown remarkable skill at parsing unstructured text and can often be adapted to new fields with minimal fine-tuning via prompting [56], although their application to medical prediction tasks remains largely unexplored.

Finally, deep learning models focused on imaging may also help uncover disease subtypes linked to severity levels. For instance, Humphries and colleagues [22] created a tool to automatically assign Fleischner grades for emphysema severity on lung CT scans; this approach

distinguished mortality risk more effectively than manual human grading using the original Fleischner criteria.

Despite these possibilities, predicting disease course remains inherently challenging, and there is no assurance that machine learning or deep learning will achieve reliable patient-level forecasts. Clinical usefulness cannot be established solely through internal and external validation studies [57]. Dedicated impact studies, interventional trials, and silent trials [54] are necessary to determine whether such models can meaningfully enhance patient outcomes and optimize healthcare resource allocation. To our knowledge, no such evaluations have yet been performed for COPD-related endpoints, and they are improbable until stronger evidence establishes clinical equipoise.

The earlier review by Bellou and colleagues [4] examined 408 regression-based prognostic models. It determined that only seven (1.7%) carried low risk of bias under PROBAST criteria, mainly because of small training sets, inadequate feature selection or external validation, and suboptimal missing data management. Our analysis, limited to 18 non-regression machine learning and deep learning models, similarly identified high risk of bias in the analysis domain across all studies for comparable core reasons, while additionally highlighting deficiencies in confidence interval reporting for model performance.

This work represents the first in-depth systematic review of non-regression machine learning and deep learning approaches for COPD outcome prediction. Key strengths include an exhaustive search across multiple databases conducted in line with PRISMA standards, dual independent processes for study screening and data extraction, and rigorous quality and bias appraisals guided by TRIPOD and PROBAST instruments. The primary limitation was the considerable heterogeneity among studies, which hindered firm conclusions about overall model performance or identification of optimal modeling strategies for COPD prognosis. The need to recalculate 95% CIs—while vital for fair comparison—introduced another constraint. In most cases, the recalculated intervals closely aligned with originally reported values, except in three studies [21, 24, 25]. For Yun and colleagues [25], the published 95% CI was notably narrower than our calculation and those seen in comparable research. Given that this study had the smallest training cohort (344 participants) among deep learning investigations, we deemed use of the recalculated CI appropriate. For González and colleagues [21] and Nam and colleagues [24], the recalculated CIs were narrower than those originally published, leading to somewhat optimistic pooled estimates involving these papers. Broader CIs would likely reduce the statistical significance of observed differences between groups and therefore would not alter the overall conclusions of this review.

Conclusion

In summary, our systematic review evaluated 18 machine learning and deep learning studies addressing COPD mortality, exacerbation, and functional decline, revealing only limited evidence that these models surpass previously reported regression tools based on performance on external validation sets. All studies exhibited high risk of bias, driven chiefly by low events-per-variable ratios, suboptimal missing data handling, and absent uncertainty measures. Future investigations should prioritize alignment with PROBAST principles wherever applicable. Despite these constraints, we encourage continued development of deep learning techniques, particularly those emphasizing efficient three-dimensional image processing and multimodal data integration.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. World Health Organization. Chronic obstructive pulmonary disease (COPD) [Internet]. 2023 [cited 2023 Apr 17]. Available from: https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd)
2. Global Initiative for Chronic Obstructive Lung Disease. Global strategy for prevention, diagnosis and management of COPD: 2023 report [Internet]. 2022 [cited 2022 Apr 13]. Available from: https://goldcopd.org/2023-gold-report-2/
3. Zafari Z, Li S, Eakin MN, Bellanger M, Reed RM. Projecting long-term health and economic burden of

COPD in the United States. Chest. 2021;159(4):1400-10.
4. Bellou V, Belbasis L, Konstantinidis AK, Tzoulaki I, Evangelou E. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. BMJ. 2019;367:l5358.
5. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med Res Methodol. 2012;12:82.
6. Lane ND, Gillespie SM, Steer J, Bourke SC. Uptake of clinical prognostic tools in COPD exacerbations requiring hospitalisation. COPD. 2021;18(4):406-10.
7. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science. 2015;349(6245):255-60.
8. Chauhan NK, Singh K. A review on conventional machine learning vs deep learning. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON); 2018; Greater Noida, India. IEEE; 2018.
9. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88.
10. Tufail AB, Ma YK, Kaabar MKA, Martínez F, Junejo AR, Ullah I, et al. Deep learning in cancer diagnosis and prognosis prediction: a minireview on challenges, recent trends, and future directions. Comput Math Methods Med. 2021;2021:9025470.
11. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. Eur Heart J. 2017;38(7):500-7.
12. Washko GR. The role and potential of imaging in COPD. Med Clin North Am. 2012;96(4):729-43.
13. Riva JJ, Malik KM, Burnie SJ, Endicott AR, Busse JW. What is your research question? An introduction to the PICOT format for clinicians. J Can Chiropr Assoc. 2012;56(3):167-71.
14. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
15. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11(10):e1001744.
16. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51-8.
17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br J Cancer. 2015;112(2):251-9.
18. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30(7):1145-59.
19. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557-60.
20. MedCalc Software. Comparison of AUC of independent ROC curves [Internet]. [cited 2023 Apr 21]. Available from: https://www.medcalc.org/calc/comparison_of_independentROCtest.php
21. González G, Ash SY, Vegas-Sánchez-Ferrero G, Onieva Onieva J, Rahaghi FN, Ross JC, et al. Disease staging and prognosis in smokers using deep learning in chest computed tomography. Am J Respir Crit Care Med. 2018;197(2):193-203.
22. Humphries SM, Notary AM, Centeno JP, Strand MJ, Crapo JD, Silverman EK, et al. Deep learning enables automatic classification of emphysema pattern at CT. Radiology. 2020;294(2):434-44.
23. Singla S, Gong M, Riley C, Sciurba F, Batmanghelich K. Improving clinical disease subtyping and future events prediction through a chest CT-based deep learning approach. Med Phys. 2021;48(3):1168-81.
24. Nam JG, Kang HR, Lee SM, Kim H, Choi H, Kim S, et al. Deep learning prediction of survival in patients with chronic obstructive pulmonary disease using chest radiographs. Radiology. 2022;305(1):199-208.
25. Yun J, Cho YH, Lee SM, Park J, Kim N, Seo JB. Deep radiomics-based survival prediction in patients with chronic obstructive pulmonary disease. Sci Rep. 2021;11:15144.

26. Boueiz A, Xu Z, Chang Y, Cho MH, Castaldi PJ, Silverman EK, et al. Machine learning prediction of progression in forced expiratory volume in 1 second in the COPDGene study. Chronic Obstr Pulm Dis. 2022;9(3):349-65.
27. Esteban C, Arostegui I, Moraza J, Aburto M, Quintana JM, García-Lomas M, et al. Development of a decision tree to assess the severity and prognosis of stable COPD. Eur Respir J. 2011;38(6):1294-300.
28. Kor CT, Li YR, Lin PR, Lin SH, Wang BY, Lin CH. Explainable machine learning model for predicting first-time acute exacerbation in patients with chronic obstructive pulmonary disease. J Pers Med. 2022;12(2):228.
29. Le TT. Use of machine learning to predict COPD treatments and exacerbations in medicare older adults: a comparison of multiple approaches [PhD thesis]. College Park (MD): University of Maryland; 2021.
30. Moll M, Qiao D, Regan EA, Hunninghake GM, Make BJ, Tal-Singer R, et al. Machine learning and prediction of all-cause mortality in COPD. Chest. 2020;158(3):952-64.
31. Morales DR, Flynn R, Zhang J, Trucco E, Quint JK, Zutis K. External validation of ADO, DOSE, COTE and CODEX at predicting death in primary care patients with COPD using standard and machine learning approaches. Respir Med. 2018;138:150-5.
32. Moslemi A, Makimoto K, Tan WC, Hogg JC, Kirby M, Coxson HO, et al. Quantitative CT lung imaging and machine learning improves prediction of emergency room visits and hospitalizations in COPD. Acad Radiol. 2023;30(4):707-16.
33. Nguyen TT. Using random forest model for risk prediction of hospitalization and rehospitalization associated with chronic obstructive pulmonary disease [PhD thesis]. Minneapolis (MN): University of Minnesota; 2017.
34. Pinto-Plata V, Casanova C, Divo M, Marin JM, de Torres JP, Polverino F, et al. Plasma metabolomics and clinical predictors of survival differences in COPD patients. Respir Res. 2019;20:219.
35. Sharma M, Westcott A, McCormack DG, Parraga G. Hyperpolarized gas magnetic resonance imaging texture analysis and machine learning to explain accelerated lung function decline in ex-smokers with and without COPD. In: Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging; 2021 Mar 18; Online. SPIE; 2021.
36. Tang C, Plasek JM, Shi X, Zhang H, Bates DW, Zhou L. Estimating time to progression of chronic obstructive pulmonary disease with tolerance. IEEE J Biomed Health Inform. 2021;25(1):175-80.
37. Zeng S, Arjomandi M, Tong Y, Liao ZC, Luo G. Developing a machine learning model to predict severe chronic obstructive pulmonary disease exacerbations: retrospective cohort study. J Med Internet Res. 2022;24(1):e28953.
38. Tang C, Plasek JM, Zhang H, Xiong Y, Bates DW, Zhou L. A deep learning approach to handling temporal variation in chronic obstructive pulmonary disease progression. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018; Madrid, Spain. IEEE; 2018.
39. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, et al. Genetic epidemiology of COPD (COPDGene) study design. COPD. 2010;7(1):32-43.
40. Vestbo J, Anderson W, Coxson HO, Crim C, Dawber F, Edwards L, et al. Evaluation of COPD longitudinally to identify predictive surrogate endpoints (ECLIPSE). Eur Respir J. 2008;31(4):869-73.
41. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med. 2019;170(1):W1-33.
42. Fellicious C, Weissgerber T, Granitzer M. Effects of random seeds on the accuracy of convolutional neural networks. In: International Conference on Machine Learning, Optimization, and Data Science (MOD); 2020; Siena, Italy. Springer; 2020.
43. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015;68(1):25-34.
44. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381.
45. Jang JH, Choi J, Roh HW, Lee S, Cho YS, Kim S, et al. Deep learning approach for imputation of missing values in actigraphy data: algorithm development study. JMIR Mhealth Uhealth. 2020;8(7):e16113.

46. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. JMLR. 2014;15(1):1929-58.
47. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: 32nd International Conference on Machine Learning (ICML); 2015; Lille, France. PMLR; 2015.
48. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. 2019;6:60.
49. Belkin M, Hsu D, Mitra P. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In: 32nd Conference on Neural Information Processing Systems (NeurIPS); 2018; Montréal, Canada. 2018.
50. Dransfield MT, Davis JJ, Gerald LB, Bailey WC. Racial and gender differences in susceptibility to tobacco smoke among patients with chronic obstructive pulmonary disease. Respir Med. 2006;100(6):1110-6.
51. Mamary AJ, Stewart JI, Kinney GL, Hokanson JE, Shenoy K, Dransfield MT, et al. Race and gender disparities are evident in COPD underdiagnoses across all severities of measured airflow obstruction. Chronic Obstr Pulm Dis (Miami). 2018;5(3):177-84.
52. Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen LC, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health. 2022;4(6):e406-14.
53. López-Campos JL, Tan W, Soriano JB. Global burden of COPD. Respirology. 2016;21(1):14-23.
54. McCradden MD, Anderson JAA, Stephenson EA, Drysdale E, Erdman L, Goldenberg A, et al. A research ethics framework for the clinical translation of healthcare machine learning. Am J Bioeth. 2022;22(5):8-22.
55. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
56. Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, et al. Fine-tuned language models are zero-shot learners [preprint]. arXiv. 2021 [cited 2023 Apr 21]. doi:10.48550/arXiv.2109.01652
57. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med. 2018;1:40.