Machine learning and deep learning predictive models for long-term prognosis in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis

Luke A Smith, Lauren Oakden-Rayner, Alix Bird, Minyan Zeng, Minh-Son To, Sutapa Mukherjee, Lyle J Palmer

Summary

Background Machine learning and deep learning models have been increasingly used to predict long-term disease progression in patients with chronic obstructive pulmonary disease (COPD). We aimed to summarise the performance of such prognostic models for COPD, compare their relative performances, and identify key research gaps.

Methods We conducted a systematic review and meta-analysis to compare the performance of machine learning and deep learning prognostic models and identify pathways for future research. We searched PubMed, Embase, the Cochrane Library, ProQuest, Scopus, and Web of Science from database inception to April 6, 2023, for studies in English using machine learning or deep learning to predict patient outcomes at least 6 months after initial clinical presentation in those with COPD. We included studies comprising human adults aged 18–90 years and allowed for any input modalities. We reported area under the receiver operator characteristic curve (AUC) with 95% CI for predictions of mortality, exacerbation, and decline in forced expiratory volume in 1 s (FEV_1). We reported the degree of interstudy heterogeneity using Cochran's *Q* test (significant heterogeneity was defined as p≤0·10 or I^2>50%). Reporting quality was assessed using the TRIPOD checklist and a risk-of-bias assessment was done using the PROBAST checklist. This study was registered with PROSPERO (CRD42022323052).

Findings We identified 3620 studies in the initial search. 18 studies were eligible, and, of these, 12 used conventional machine learning and six used deep learning models. Seven models analysed exacerbation risk, with only six reporting AUC and 95% CI on internal validation datasets (pooled AUC 0·77 [95% CI 0·69–0·85]) and there was significant heterogeneity (I^2 97%, p<0·0001). 11 models analysed mortality risk, with only six reporting AUC and 95% CI on internal validation datasets (pooled AUC 0·77 [95% CI 0·74–0·80]) with significant degrees of heterogeneity (I^2 60%, p=0·027). Two studies assessed decline in lung function and were unable to be pooled. Machine learning and deep learning models did not show significant improvement over pre-existing disease severity scores in predicting exacerbations (p=0·24). Three studies directly compared machine learning models against pre-existing severity scores for predicting mortality and pooled performance did not differ (p=0·57). Of the five studies that performed external validation, performance was worse than or equal to regression models. Incorrect handling of missing data, not reporting model uncertainty, and use of datasets that were too small relative to the number of predictive features included provided the largest risks of bias.

Interpretation There is limited evidence that conventional machine learning and deep learning prognostic models demonstrate superior performance to pre-existing disease severity scores. More rigorous adherence to reporting guidelines would reduce the risk of bias in future studies and aid study reproducibility.

Funding None.

Lancet Digit Health 2023; 5: e872–81

Australian Institute for Machine Learning, University of Adelaide, Adelaide, SA, Australia (L A Smith BEng, L Oakden-Rayner PhD, A Bird MBBS, M Zeng MSc, Prof L J Palmer PhD); **School of Public Health, University of Adelaide, Adelaide, SA, Australia** (L A Smith, L Oakden-Rayner, A Bird, M Zeng, Prof L J Palmer); **Health Data and Clinical Trials, Flinders University, Bedford Park, SA, Australia** (M-S To PhD); **South Australia Medical Imaging, Flinders Medical Centre, Bedford Park, SA, Australia** (M-S To); **Department of Respiratory and Sleep Medicine, Southern Adelaide Local Health Network (SALHN), Bedford Park, SA, Australia** (S Mukherjee PhD); **Adelaide Institute for Sleep Health/ Flinders Health and Medical Research Institute, College of Medicine and Public Health, Flinders University, Bedford Park, SA, Australia** (S Mukherjee)

Correspondence to:
Luke Smith, Australian Institute for Machine Learning, University of Adelaide, Adelaide, SA, Australia
luke.a.smith@adelaide.edu.au

Introduction

Chronic obstructive pulmonary disease (COPD) is the third most common cause of death globally.[1] The costs of COPD management are likely to continue to increase worldwide[2,3] and the identification of patients who might have rapid disease progression would be highly beneficial for reducing these costs. These patients could then be given more intensive treatment and follow-up, maximising the efficiency of health service delivery.

Several measures—including lung function tests, questionnaires, exercise capacity, and exacerbation frequency—can be used to assess the risk of poor patient outcomes, with composite prognostic models, such as the body-mass index, airflow obstruction, dyspnoea, and exercise capacity [BODE] index, providing better predictive performance than any single variable.[2] A comprehensive systematic review by Bellou and colleagues[4] in 2019 identified more than 400 published prognostic models for the prediction of clinical outcomes in COPD. Meta-analysis of linear regression models (n=12) and their pooled C-statistics—a measure of predictive performance, equivalent to the area under the receiving operating characteristic curve (AUC)[5]—for predicting mortality ranged between 0·624 and 0·769.

Research in context

Evidence before this study

Chronic obstructive pulmonary disease (COPD) is a leading cause of mortality worldwide with a high degree of heterogeneity in disease progression that conventional prognostic scores have not been able to account for. We searched PubMed, Scopus, and Web of Science to identify reviews of long-term COPD prognostic models published in English between Jan 1, 2016, to March 29, 2022, using search terms related to "COPD", "machine learning", and "prediction". Two systematic reviews of COPD prognostic models reported low rates of methodological rigour and model validation, which were also unfit for clinical use. Neither systematic review compared the performance or merits of machine learning models relative to conventional regression models.
A qualitative review describing the role of artificial intelligence and machine learning models in different clinical contexts for COPD observed a considerable increase in the availability of large datasets with complex data, enabling the broader adoption of artificial intelligence techniques in the clinical context.

Added value of this study

To the best of our knowledge, there are no other reviews of prognostic models for COPD outcomes using and comparing machine learning and deep learning algorithms. Our assessment provides a baseline for machine learning model performance and highlights the current limitations in both methodology and study reporting quality, which must be overcome to make further improvements in this field.

Implications of all the available evidence

Moving forward, researchers developing machine learning algorithms need to pay close attention to the guidelines provided in PROBAST and TRIPOD to ensure results are meaningfully reported and studies have low risk of bias. Clinical relevance for conventional machine learning models analysing tabular clinical data is currently limited due to the number of variables required for assessment. In contrast, there are increasing opportunities for deep learning assessment of CT scans to be incorporated as an opportunistic assessment within clinical practice once further progress has been made. Based on the evidence presented within our study, we recommend further research into multimodal modelling techniques incorporating imaging and clinical measurements; the development of large, public datasets outside of North America and Europe; and the development of guidelines similar to PROBAST applicable to the deep learning context.

These results reflect that regression models have only a moderate predictive ability and little clinical effect on COPD prognosis.[6]

Machine learning refers to computer algorithms in which rules and correlations between input and output data are automatically learnt from a dataset, allowing for automated inference in a hypothesis-free framework.[7] Deep learning, a subset of machine learning, employs a hierarchical structure to learn more complex structures and relationships within a dataset.[8] Deep learning is particularly powerful when applied to medical images[9] and has shown substantial promise in clinical prognostic tasks using imaging data.[10,11] There has been increasing interest in the use of deep learning to provide new prognostic and diagnostic tools for the identification of patients with COPD who are at risk of worsening outcomes, as existing statistical models have not incorporated information on the structural changes of the lungs in people with COPD, which can be demonstrated with imaging.[12]

In this systematic review and meta-analysis, we aimed to evaluate the performance and quality of published machine learning-based and deep learning-based prognostic models for people with COPD and to compare the performance of these models with previous predictive regression models.[4] For the purpose of evaluating the clinical usefulness of machine learning and deep learning, we focused on model performance regardless of model architecture or selected features.

See Online for appendix

Methods

Search strategy and selection criteria

For this systematic review and meta-analysis, we included studies that used a machine learning or deep learning model to predict mortality risk, exacerbation risk, or decline in lung function within a cohort of adult patients with COPD at least 6 months (or 180 days) after initial clinical presentation. Studies on animals or cohorts of people primarily younger than 18 years or older than 90 years were excluded. Prognostic studies for the development of COPD in patients who do not initially have COPD were excluded, as were prognostic studies that did not include at least 6 months of follow-up. Cohort studies (both prospective and retrospective), case-control studies, randomised controlled trials, and cross-sectional studies were all allowed. We did not limit by country of origin or publication source. We defined machine learning as algorithms (such as random forest analysis and support vector machines) that are more complicated than regression models, which can learn to make decisions based on patterns within data, and we defined deep learning as the use of neural networks with two or more hidden layers. Brief definitions of the machine learning algorithms used in the search terms and selected studies can be found in the appendix (pp 2–5). Full inclusion criteria using the population, intervention, comparison, outcome, and time approach[13] are described in the appendix (pp 1–2).

We searched PubMed, Embase, the Cochrane Library, Scopus, Web of Science, and ProQuest using search terms related to chronic obstructive pulmonary disease, artificial intelligence, machine learning, deep learning, prediction, and prognosis, for studies published in English or with an English translation between database inception and April 6, 2023. Full search queries with all synonymous search terms for each database can be found in the appendix (pp 7–9). Hand searches of included study references and systematic reviews were completed to identify any missed studies.

Studies were uploaded to Covidence and duplicates were removed. LAS and AB independently screened the titles and abstracts of all studies, defaulting to including studies when it was uncertain if they met inclusion criteria. Manuscripts passing initial screening were then reviewed using the full-text version by LAS and AB. Where possible, conflicts were resolved between the two reviewers, otherwise they were resolved by a third, independent reviewer (LJP).

Our protocol was registered on PROSPERO (CRD42022323052) and conducted according to PRISMA guidelines.[14]

Data analysis

Before data collection, we designed a table based on the CHARMS checklist.[15] 11 data fields were extracted for each study (appendix p 10).

Studies were divided into two categories: those that used deep neural networks (the deep learning group) and those that did not (the conventional machine learning group). If a study presented multiple models for the same outcome, we assessed the model with the best performance that also met the inclusion criteria. If multiple similar outcomes were reported (eg, mortality risk at 1 year, 2 years, and 3 years), the outcome most comparable to other studies was included. Data were extracted by LAS and validated by AB and MZ. Disagreements were resolved by LJP.

Each study was assessed using PROBAST[16] to estimate risk of bias (appendix p 10). The checklist comprised 20 questions grouped into four domains (participant selection, predictors, outcome, and analysis). During assessment, each question was answered as yes, probably yes, probably no, no, or no information, with yes indicating a low risk of bias and no indicating a high risk of bias. An answer of no or probably no on one or more questions implied that the study was at a high risk of bias, otherwise the study was coded as being at low risk of bias. We assessed reporting quality using the TRIPOD checklist, a list of 22 items deemed essential for transparent reporting of predictive model development and validation (appendix pp 11–12).[17]

The discriminative ability of prognostic models to predict mortality, exacerbation, or decline in forced expiratory volume after 1 s (FEV_1) was quantified using AUC[18] and associated 95% CIs. Hospitalisation was used as a surrogate measure for exacerbation. Decline in 6 min walk test performance, changes in dyspnoea, or changes in COPD burden, as measured by questionnaires or psychophysical scales, were used as surrogate markers for decline in FEV_1, though no suitable studies were found. AUCs were pooled for each outcome, with internal validation and external validation pooled separately. We anticipated a high degree of inter-study heterogeneity among the included studies due to different models, predictors, cohorts, and outcome timeframes. Therefore, we used a random effects model to pool AUCs across studies and presented the results in forest plots. AUCs were pooled independently of model architecture or features. We reported the degree of interstudy heterogeneity using the Cochran's *Q* test[19] (where significant heterogeneity was defined as $p \leq 0 \cdot 10$ or $I^2 > 50\%$) to assess whether a fixed effects model could have been used. Meta-analyses were done using the MedCalc (20.116) statistical software. To compare pooled results of machine learning and deep learning studies against existing regression models, we pooled the model

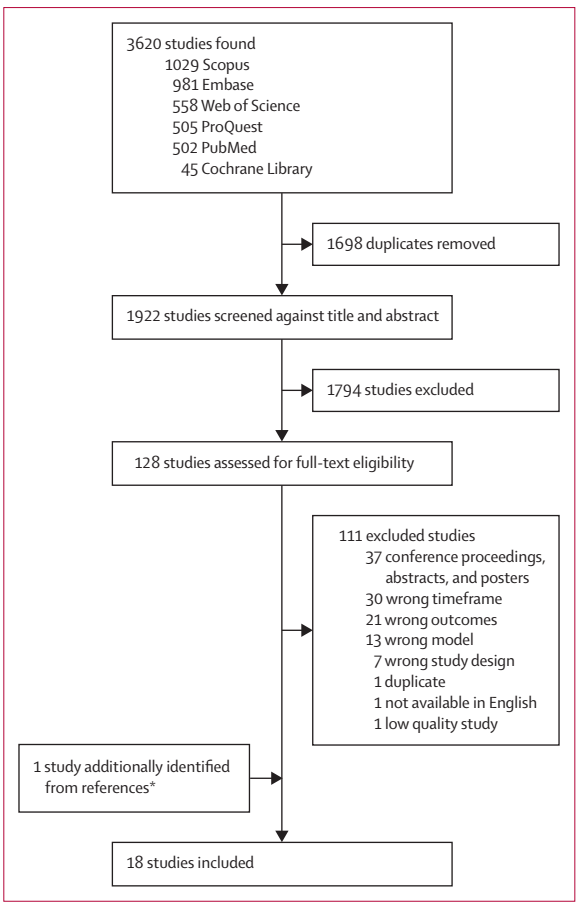

Figure 1: **Study selection**
*To ensure all literature was covered, additional studies referenced in the bibliographies of eligible articles were screened against the eligibility requirements by LAS, and then included into a second round of the systematic review.

performances of regression models for mortality and exacerbation risk reported by Bellou and colleagues,[4] and used the MedCalc online calculator[20] to compare between AUCs. Additionally, where possible, we directly compared the performance of deep learning and machine learning models against the performance of conventional risk scores or linear regression models (such as the BODE index) reported on the same dataset within the study.

Role of the funding source

There was no funding source for this study.

Results

A total of 3620 studies were identified in the initial search. Following title and abstract screening, and full-text eligibility assessments, 17 studies[21–37] met the inclusion and exclusion criteria (figure 1). We searched reference lists and found an additional eligible study[38] for a total of 18 studies. Studies that were considered for inclusion but rejected are described in the appendix (p 13). The basic cohort demographics for studies in each group are shown in the appendix (pp 14–18).

Two studies predicted decline in lung function,[26,35] seven studies predicted exacerbation risk,[21,23,28,29,32,33,37] and

	Deep learning (n=6)		Machine learning (n=12)		Combined (n=18)	
	Studies reported	Pooled value*	Studies reported	Pooled value*	Studies reported	Pooled value*
Age	4 (67%)	67·1 (64·4–71·6)	11 (92%)	63·5 (59·7–67·7)	15 (83%)	67 (60·9–68·6)
Male sex	4 (67%)	57% (50–60)	10 (83%)	65% (53–83)	14 (78%)	57% (52–69)
Location	6 (100%)	..	12 (100%)	..	18 (100%)	..
Ethnicity†	2 (33%)	13% (11–16)	4 (33%)	32% (32–33)	6 (33%)	18% (13–31)
Internal Validation	6 (100%)	..	12 (100%)	..	18 (100%)	..
External Validation	4 (67%)	..	1 (8%)	..	5 (28%)	..
COPDGene (int)	3 (50%)	..	2 (17%)	..	5 (28%)	..
ECLIPSE (ext)	2 (33%)	..	1 (8%)	..	3 (17%)	..
AUC	4 (67%)	..	10 (83%)	..	14 (78%)	..
Confidence Intervals	4 (67%)	..	5 (42%)	..	9 (50%)	..

Data are n (%), unless specified. *Median (IQR) are used for reporting pooled statistics. †Median values for proportion of Black or African-American participants reported in US studies.

Table 1: **Study statistics reporting frequencies and pooled values**

	Outcome	Sample size	AUC (95% CI)	Other performance indicators
Gonzalez et al (2018)[21]	≥1 exacerbation in 1-year period; ≥1 exacerbation in 3-year period	Training 6016 (COPDGene); internal validation 1000; external validation 1672 (ECLIPSE)	Internal validation 0·64 (0·60–0·68); external validation 0·54 (0·51–0·57)	..
Singla et al (2021)[23]	≥1 exacerbation in a 5-year period	Training 10 300 (5-fold cross validation)	0·73 (0·71–0·75)*	AUPRC 0·42 (SD 0·02), recall 0·47 (SD 0·04), accuracy 80·83%
Gonzalez et al (2018)[21]	3-year all-cause mortality	Training 5740 (COPDGene); internal validation 1000; external validation 1672 (ECLIPSE)	Internal validation 0·72 (0·63–0·81) External validation 0·60 (0·56–0·64)	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 (2020)[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 (2022)[24]	5-year all-cause mortality	Training 3475; internal validation (hold-out) 315; internal validation (temporal) 394; external validation (VHSMC) 416; external validation (AMC) 337	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)	..
Singla et al (2021)[23]	5-year all-cause mortality	Training 10 300 (5-fold cross validation)	0·62 (no CI calculated)	HR 1·54 (1·09–2·17), p<0·0001
Tang et al (2018)[38]	1-year all-cause mortality	Training 10 850; internal validation 4650	..	Accuracy 78·89% (no CI)
Yun et al (2021)[25]	3-year and 5-year all-cause mortality	Training 344 (5-fold cross validation); external validation 102	Internal validation 0·80 (0·72–0·88); external 0·72 (0·57–0·86)	..

AUCs reported with recalculated 95% CI. HRs are reported as HR (95% CI). AMC=Asan Medical Centre dataset. AUC=area under receiver operator characteristic curve. AUPRC=area under precision recall curve. HR=hazard ratio. VHSMC=Veteran Health Service Medical Centre dataset. *Number of cases not reported, so reported 95% CI used instead of calculated 95% CI. †Median follow-up for internal validation was 7·95 years and median follow-up for external validation was 2·90 years.

Table 2: **Model performance for deep learning**

11 studies predicted mortality.[21–25,27,30,31,34,36,38] Two of the included studies predicted both exacerbations and mortality.[21,23]

Summaries of study statistics reporting frequencies and median values, pooled across deep learning studies, machine learning studies, and all studies are shown in table 1. Overall, age, sex, and location of recruitment were consistently reported in studies, but ethnicity demographics were not. All studies performed internal validation, either using a hold-out validation set or k-fold cross-validation. In comparison, external validation was only performed in four deep learning studies[21,22,24,25] and one machine learning study.[30] Of these, three studies[21,22,30] were developed using the COPDGene cohort[39] and externally validated on the ECLIPSE cohort,[40] and the other two were both developed on South Korean cohorts and externally validated on separate South Korean[24] and Malaysian[25] cohorts. Cohort demographics for COPDGene and ECLIPSE are shown in the appendix (p 19).

Six studies applied deep learning to the task of COPD prognosis,[21–25,38] whereas 12 studies used machine learning.[26–37] Model development for deep learning and conventional machine learning models is summarised in the appendix (pp 20–23). Model performance for deep learning is summarised in table 2 and model performance for machine learning is summarised in table 3.

The pooled AUC for exacerbation prediction across machine learning and deep learning regimes on the internal validation datasets was 0·77 (95% CI 0·69–0·85; figure 2). Examination of heterogeneity as measured using Cochran's *Q* test showed significant heterogeneity (I^2 97%, p<0·0001). Whereas there were not enough deep learning studies predicting exacerbation to be pooled independently, there were enough studies using machine learning, and their pooled AUC was 0·80 (0·72–0·87, I^2 97%, p<0·0001). One study[21] performed external validation on exacerbation prediction and saw a drop in performance from AUC 0·64 (95% CI 0·58–0·71) to 0·55 (0·51–0·57).

The pooled AUC for mortality prediction on the internal validation datasets was 0·77 (95% CI 0·74–0·80), with significant heterogeneity (I^2 60%, p=0·027; figure 3). Individually, pooled AUC was 0·78 (0·73–0·84, I^2 22%, p=0·28) for deep learning studies and 0·77 (0·73–0·80, I^2 80%, p=0·0070) for machine learning studies. Comparison between machine learning and deep learning approaches revealed no significant differences in pooled AUCs for mortality (p=0·44). Four studies[21,24,25,30] performed external validation on mortality prediction, with pooled AUC 0·67 (0·62–0·72, I^2 79%, p=0·0023). The pooled internal validation performance for these studies was significantly higher (p=0·0049) at AUC 0·76 (0·72–0·80, I^2 42%, p=0·16).

	Outcome	Sample size	AUC (95% CI)	Other performance indicators
Boueiz et al (2022)[26]	Change in FEV_1 after 5-year period; new FEV_1 after 5-year period	Training 4496 (10-fold cross validation); internal validation (temporal) 1833	Internal validation (cross validation) 0·71 (0·69–0·72); internal validation (temporal) 0·70*	..
Sharma et al (2021)[35]	Decline in FEV_1 ≥30 mL per year over 3-year period	Training 42 (5-fold cross validation)	0·81 (0·66–0·96)	Accuracy 84·6%, sensitivity 0·875, specificity 0·800
Kor et al (2022)[28]	≥1 exacerbation in 6-month period (for patients who have not previously exacerbated)	Training 407; internal validation 102	0·83 (0·74–0·92)	Sensitivity 79·41%, specificity 77·94%, PPV 64·29%, NPV 88·33%, F1 score 71·05%, accuracy 78·43%
Le (2021)[29]	≥1 exacerbation in 1-year period	Training 190 117 (10-fold cross validation)	0·67 (0·67–0·68)	..
Nguyen (2017)[33]	≥1 hospitalisation in 180-day period	Training 81; internal validation 27	0·88 (0·81–0·95)	Accuracy 88%, sensitivity 0·83, specificity 0·93
Zeng et al (2022)[37]	≥1 exacerbation in 1-year period	Training 36 047; internal validation 7529	0·866 (0·83–0·90)	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 (2023)[32]	≥1 hospitalisation in 3-year period	Training 328; internal validation 141†	0·78 (0·69–0·87)	Accuracy 78%, F1 score 71%
Esteban et al (2011)[27]	5-year all-cause mortality	Training 611; internal validation 348	0·74 (0·68–0·80)	..
Moll et al (2020)[30]	All-cause mortality risk‡	Training 1974; internal validation 658; external validation 1268	Internal validation 0·731 (0·682–0·780); external validation 0·688 (0·655–0·721)	..
Morales et al (2018)[31]	3-year all-cause mortality	Training 163 587; internal validation 40 895†	0·80 (0·79–0·80)	..
Pinto-Plata et al (2019)[34]	3-year all-cause mortality	Training 90; internal validation 30 (100% mortality)	..	Accuracy 85%, sensitivity 81%, specificity 89%
Tang et al (2021)[36]	1-year all-cause mortality	Training 10 850; internal validation 4650	..	Accuracy 64·6%

All AUCs reported with recalculated 95% CI. HRs are reported as HR (95% CI). AUC=area under receiver operator curve. HR=hazard ratio. PPV=positive predictive value. NPV=negative predictive value. FEV_1=forced expiratory volume in 1 s. *Regression model has no event counts so 95% CIs could not be calculated; these are values reported in the original study. †Training and validation populations estimated from total population and reported train-validation split percentages. ‡Median follow-up for internal validation 6·4 years; median follow-up for external validation 7·2 years.

Table 3: **Model performance for conventional machine learning**

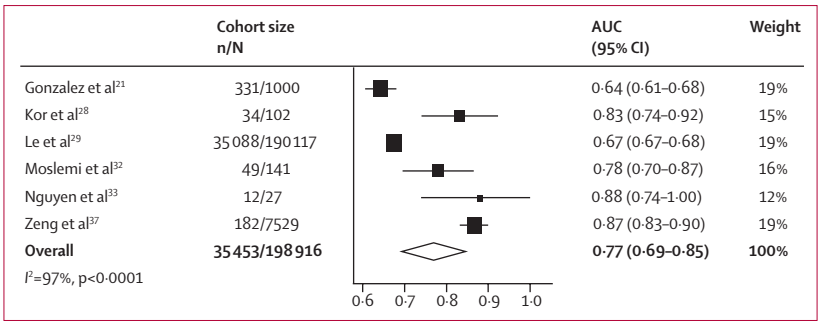

Figure 2: **Random effects forest plot for exacerbation prediction models**
n refers to cases of exacerbation and N refers to group size. p values from Cochran's Q test. AUC=area under the receiver operator characteristic curve.

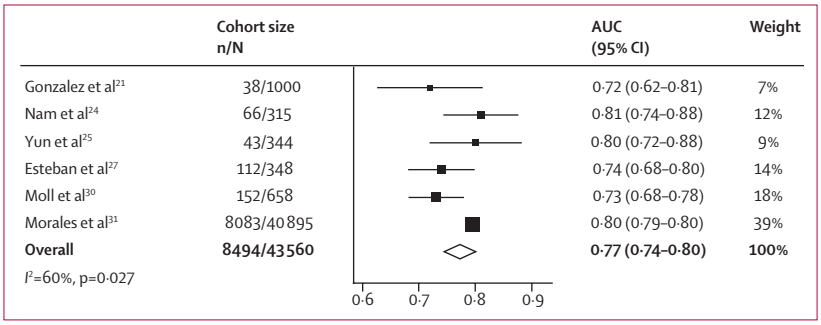

Figure 3: **Random effects forest plot for mortality prediction models**
n refers to cases of exacerbation and N refers to group size. p values from Cochran's Q test. AUC=area under the receiver operator characteristic curve.

As only two studies reported a decline in lung function, no pooled AUC could be calculated. Decline in lung function greater than 30 mL over a 5-year period was predicted with AUC 0·81 (95% CI 0·66–0·96) on a cohort of 42 patients using 5-fold cross validation,[35] and the overall decline in lung volume was predicted with AUC 0·71 (0·69–0·72) on a cohort of 4496 patients using ten-fold nested cross validation (table 3).[26]

The pooled prognostic performance of regression models previously reported by Bellou and colleagues[4] was 0·66 (95% CI 0·49–0·82) for exacerbation prediction and 0·69 (0·68–0·71) for mortality prediction. For prognosis of mortality, both machine learning (p=0·0001) and deep learning (p<0·0001) studies had significantly better predictive performance. For exacerbation prognosis, machine learning studies had a numerically higher pooled estimate than Bellou and colleagues,[4] but the difference was not significant (p=0·14 for pooling across machine learning studies and p=0·24 for pooling across all studies).

A small number of studies allowed direct comparison of the performance of deep learning or machine learning models to more conventional risk scores or regression models on the same dataset. Three deep learning studies[21,23,24] compared mortality prognosis to the BODE index, two of which were conducted using the same population (COPDGene).[21,23] These studies reported marginal but non-significant improvement in AUC for mortality using deep learning versus the BODE index. There were seven machine learning studies that also reported conventional indices (BODE index, and age, dyspnoea, airflow obstruction index) or regression models using the same predictors as the machine learning models in the same dataset.[26,27,30,31,35–37] Five of these studies[26,27,30,35,37] reported marginally improved performance in the machine learning model, one reported no difference in performance,[31] and one reported marginally worse performance in the machine learning model.[36] Of the five studies that reported improved performance from machine learning, one study reported the improved performance to be significant.[30] Three of the seven studies[27,30,31] reported AUCs and shared an outcome (mortality) and were amenable to meta-analysis. Meta-analysis indicated no significant difference in pooled AUC between the machine learning (pooled AUC 0·77) and conventional regression models (pooled AUC 0·75; p=0·57).

Adherence to TRIPOD guidelines was good, with only one study[23] reporting less than 70% of the items (appendix p 24). 14 (78%) of 18 studies did not present or provide access to the full prediction model (item 15a) and nine (50%) of 18 studies did not report CIs with the prediction model's performance measures (item 16). All included studies were determined to be at high risk of bias according to PROBAST (appendix p 25). The analysis domain indicated a high risk of bias among all studies. Poor handling of missing data was the most frequent risk of bias (item 4.4): two studies[36,38] had no missing data, three[23,28,29] did not report handling, one[37] used a model that inherently accounts for missing data, one[34] used minimum-value imputation, and the rest used complete case analysis without justifying the missing completely at random assumption. Nine (50%) of 18 studies did not report CIs for any of their model outcomes (item 4.7). The ratio of positive outcome events to the number of predictive variables using in prognostic models, also known as the event per variable (EPV) ratio, was used to assess item 4.1, with an EPV of less than 10 indicating a high risk of bias.[41] Six (50%) of 12[28,33–37] machine learning studies had an EPV of less than 10 in the training cohort, with a median number of positive events of 82 (IQR 33–1560) and median number of predictive variables in the final model of 154 (IQR 45–412). The remaining six machine learning studies had EPV greater than 20, with median positive events 1645 (IQR 244–4075) and median predictive variables 11·5 (IQR 6–39).

Discussion

We conducted a systematic review and meta-analysis of prognostic studies for COPD that used machine learning and deep learning and found a total of 18 studies. Outcomes investigated included mortality, exacerbation, and decline in lung function. There is some evidence to suggest that both machine learning and deep learning

algorithms provide better predictive performance than existing regression models independent of architecture or features, particularly for the task of mortality prediction. However, this evidence does not consider the high risk of bias present in all studies or results on external validation. Poor handling of missing data, small sample sizes, and unreported CIs were the main contributors to a high risk of bias.

The substantial inter-study heterogeneity observed within each clinical outcome made performance comparisons challenging due to variance in outcome timeframes, cohort demographics, predictive variables, and model architectures. However, we also noted that, within the scope of training deep learning models, it is likely that high heterogeneity is inevitable, as even random parameter initialisation can affect model convergence.[42] Therefore, pooling of highly heterogeneous results was not unreasonable provided we were mindful of the possible sources of variance.

The pooled performance of machine learning and deep learning models on internal validation sets was greater than the pooled performance of regression models on external validation sets. However, of the five models that were externally validated (one prognosed exacerbation and four prognosed mortality), external performance was worse than internal performance by an average of 0·09 AUC, and worse than or equal to regression models. External validation would be needed to make a definitive comparison, and the absence of external validation is an ongoing methodological issue with machine learning and deep learning studies.[4,43,44]

Almost all studies had high risk of bias from handling missing data inappropriately (eg, complete case analysis without justifying missing completely at random assumptions). PROBAST guidelines recommend using multiple imputation to minimise bias while preserving the overall distribution of the data;[41] however, finding an equivalent technique for imputing missing imaging data is an area of ongoing research.[45]

When considering the risk of bias due to overfitting, an EPV of more than 20 suggests a low risk of bias, whereas an EPV of less than 10 suggests a high risk of bias.[41] Within the six studies with an EPV of less than 10, studies had both small datasets with few outcome events, while also training models with many predictive variables. Multivariate feature selection to reduce the number of features should be used to reduce the risk of overfitting if cohort size cannot be increased.[41]

Datasets were consistently larger in deep learning studies than in machine learning studies, with median training populations of 6016 (IQR 2407–10 300) people in deep learning studies and 1292·5 (183–29 747) people in machine learning studies. In general, deep learning models have better performance and generalisability when trained on larger datasets. Despite deep learning having better performance when trained on larger datasets, we qualitatively observed an inverse relationship between deep learning mortality model performance and number of patients within the training cohort, suggesting that deep learning models were overfitted to training data. Regularisation techniques are methods for reducing overfitting in deep learning with minimal impact on performance on training data. They are varied and reasonably effective at improving generalisability, and include dropout,[46] batch normalisation,[47] and data augmentation.[48] We note that the concepts of overfitting and regularisation are poorly defined for deep learning models, as it is known that models can perfectly fit (or what would conventionally be considered as overfitting) to training data and still maintain accurate test performance and generalise well to external datasets.[49] It is currently unclear how to assess the risk of overfitting in deep learning models for predictive tasks apart from external validation and guidelines similar to PROBAST are needed for the deep learning context.

Although age and sex demographics were reported in more than 75% of studies and location of recruitment was identifiable in all studies, ethnicity of study populations was poorly reported. For complete analysis and comparison of studies, full reporting of cohort demographics is required, as there are substantial racial and sex-based disparities in COPD susceptibility.[50,51] Self-identified race has been recently shown to be a strong potential confounder of image-based deep learning analyses.[52] As ethnicity is not reported in the ECLIPSE cohort, a common validation cohort for large-scale COPD studies, there is a need for additional, large-scale studies that properly report the full cohort demographics. All studies, except for three, were developed in North America and western Europe (either the USA, Canada, UK, or Spain). The other three studies were developed in South Korea[24,25] and Taiwan.[28] Therefore, caution is required when extrapolating the results to low-income and middle-income countries, where the effects of COPD are the most severe.[53]

Because of the poor generalisability to external validation sets and high risk of bias, we conclude that there is limited evidence that machine learning or deep learning models have better performance than existing regression models, despite their increased complexity. However, future research using machine learning and deep learning models could lead to greater generalisability and risk of bias. In particular, the development of risk-of-bias guidelines, such as PROBAST, targeted to the development and validation of deep learning models that address topics such as missing data, generalisation, dataset size, and reporting standards would be useful to standardise study design and improve quality of evidence.

From a clinical perspective, machine learning models with many input variables are not useful to practitioners who, for example, see one patient and must gather 100 variables to make a prognosis. In comparison, conventional risk scores (eg, the BODE index) and deep learning models using CT scans have more potential as clinical tools. The increasing use of CTs for differential

diagnosis, for screening of lung cancer, and as a prerequisite to surgical interventions[2] provides more opportunities for CT-based deep learning models to opportunistically prognose patients with no additional effort from the practitioner, provided the model had undergone sufficient integrated testing.[54]

Furthermore, the successful development of deep learning models relies on large volumes of high-quality data and intelligently designed deep learning algorithms. With regards to imaging data, publicly available, longitudinal cohort studies, such as COPDGene and ECLIPSE, have collected high-quality data that can be used as a universal baseline for testing and validating models. However, the datasets are still small relative to the size of other imaging datasets used within computer science (eg, ImageNet, with over 14 million images). Further research to develop deep learning models that properly handle CT scans is also needed. Of the studies included, only Singla and colleagues[23] developed a model using three-dimensional volumes instead of two-dimensional slices. The incorporation of multiple modalities is also a promising research direction. We observed that Moslemi and colleagues,[32] Nam and colleagues,[24] and Singla and colleagues[23] each report that combining imaging and clinical features in a final prognostic model provides better performance than either modality alone. Additionally, recent developments with large language models provide new opportunities for research.[55] These models have demonstrated impressive capabilities in being able to extract information from free text and might be able to be adapted to new domains with only a small amount of model updating through prompting;[56] however, these models remain largely untested for medical tasks.

Finally, imaging deep learning models might also assist in identifying disease phenotypes corresponding to severity, demonstrated by Humphries and colleagues,[22] who developed a model to classify the Fleischner grade for emphysema severity in lung CTs that was able to better differentiate between patient mortality risk than the original Fleischner score graded by humans.

Despite the potential of machine learning, disease prognosis is difficult and there is no guarantee that machine learning or deep learning models will be able to sufficiently predict patient outcomes. It is also important to mention that the clinical utility of deep learning and machine learning models cannot be demonstrated through internal and external validation studies alone.[57] Impact studies, interventional studies, and silent trials[54] are needed to evaluate whether these models can improve patient outcomes and hospital resource management. To the best of our knowledge, such studies have not occurred for COPD outcomes and are unlikely to occur until sufficient improvements have been made to demonstrate clinical equipoise.

The previous study by Bellou and colleagues[4] assessed 408 prognostic regression models and found only seven (1·7%) models were at low risk of bias according to PROBAST guidelines when reporting analysis, primarily due to the relatively small training dataset sizes with insufficient feature selection or external validation and poor handling of missing data. Our study exclusively examined 18 non-regression machine learning models and deep learning models and, similarly, reported a high risk of bias in the analysis of all studies for the same primary reasons, although we additionally emphasise better reporting of CIs for model outcomes.

Our study is the first comprehensive review of non-regression-based machine learning and deep learning models applied to COPD outcome prognosis. The major strengths of our systematic review and meta-analysis were the comprehensive literature search of multiple databases following PRISMA guidelines, the independent screening and data extraction, and the detailed quality and risk-of-bias assessment following TRIPOD and PROBAST checklists. The major limitation of this study was the heterogeneity between studies, restricting meaningful inference of the performance of models included or regarding the best approaches to machine learning and deep learning modelling in COPD outcome prediction. The use of recalculated 95% CIs, although essential for comparison, was another limitation of our study. In general, the calculated 95% CIs matched those that were reported, except for in three studies.[21,24,25] For Yun and colleagues,[25] the reported 95% CI was significantly narrower than what was calculated or reported in other studies. Since this study had a training cohort of 344 participants, the smallest of all the included deep learning studies, we believed that it was suitable to use the calculated 95% CI. In the case of González and colleagues[21] and Nam and colleagues,[24] the calculated 95% CIs were narrower than those originally reported, resulting in optimistic CIs for pooled results involving these studies. Wider CIs would decrease the significance of any differences found between groupings, and, therefore, would not change the findings of this study.

Our systematic review analysed 18 deep learning and machine learning studies for COPD mortality, exacerbation, and functional decline prognosis and found limited evidence of them outperforming previously reported regression models due to poor generalisability on external validation sets based on performance alone. All studies were at high risk of bias, with small events per variable ratios, poor handling of missing values, and unreported uncertainty metrics being the main factors. Future research should emphasise adherence to PROBAST guidelines where possible. Despite the limited evidence, we recommend that work on deep learning methods continues, with emphasis on models with efficient three-dimensional analysis capacity and multimodal feature analysis.

Contributors

LAS contributed to study conception, design, collection and analysis of data, and draft writing. AB and MZ contributed to the study design, data collection, and critical revision of the manuscript. LJP and LO-R

contributed to the study design, data analysis, critical revision of the manuscript, and supervision of LAS. SM and M-ST contributed to the study design and critical revision of the manuscript. AB and MZ independently accessed and verified all extracted data. All authors had full access to all the data in the study and had final responsibility for the decision to submit to publication.

Declaration of interests

LAS and AB are supported by GlaxoSmithKline. LO-R received stocks from Sirona Medical Imaging as a consultant on artificial intelligence development and implementation. All other authors declare no competing interests.

Data sharing

The datasets generated or analysed during the current study are available from the corresponding author on reasonable request.

Acknowledgments

LAS, AB, and MZ were supported by the Australian Government Research Training Program scholarship. This work was not supported by a specific grant from any funding agencies in the public, commercial, or not-for-profit sectors. We thank Mary Filsell for her advice and guidance regarding strategies for literature searches.

Editorial note: The Lancet Group takes a neutral position with respect to territorial claims in published maps and institutional affiliations.

References

1 WHO. Chronic obstructive pulmonary disease (COPD). https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd) (accessed April 17, 2023).
2 Global Initiative for Chronic Obstructive Lung Disease. Global strategy for prevention, diagnosis and management of COPD: 2023 report. https://goldcopd.org/2023-gold-report-2/ (accessed April 13, 2022).
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:** 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:** 406–10.
7 Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015; **349:** 255–60.
8 Chauhan NK, Singh K. A Review on conventional machine learning *vs* deep learning. International Conference on Computing, Power and Communication Technologies (GUCON); Sept 28–29, 2018.
9 Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017; **42:** 60–88.
10 Tufail AB, Ma Y-K, Kaabar MKA, 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, 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:** 500–07.
12 Washko GR. The role and potential of imaging in COPD. *Med Clin North Am* 2012; **96:** 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:** 167–71.
14 Page MJ, McKenzie JE, Bossuyt PM, 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, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014; **11:** e1001744.
16 Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019; **170:** 51–58.
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:** 251–59.
18 Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit* 1997; **30:** 1145–59.
19 Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003; **327:** 557–60.
20 MedCalc Software. Comparison of AUC of independent ROC curves. https://www.medcalc.org/calc/comparison_of_independentROCtest.php (accessed April 21, 2023).
21 González G, Ash SY, Vegas-Sánchez-Ferrero G, et al. Disease staging and prognosis in smokers using deep learning in chest computed tomography. *Am J Respir Crit Care Med* 2018; **197:** 193–203.
22 Humphries SM, Notary AM, Centeno JP, et al. Deep learning enables automatic classification of emphysema pattern at CT. *Radiology* 2020; **294:** 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:** 1168–81.
24 Nam JG, Kang H-R, Lee SM, et al. Deep learning prediction of survival in patients with chronic obstructive pulmonary disease using chest radiographs. *Radiology* 2022; **305:** 199–208.
25 Yun J, Cho YH, Lee SM, et al. Deep radiomics-based survival prediction in patients with chronic obstructive pulmonary disease. *Sci Rep* 2021; **11:** 15144.
26 Boueiz A, Xu Z, Chang Y, et al. Machine learning prediction of progression in forced expiratory volume in 1 second in the copdgene study. *Chronic Obstr Pulm Dis* 2022; **9:** 349–65.
27 Esteban C, Arostegui I, Moraza J, et al. Development of a decision tree to assess the severity and prognosis of stable COPD. *Eur Respir J* 2011; **38:** 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:** 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, University of Maryland, 2021.
30 Moll M, Qiao D, Regan EA, et al. Machine learning and prediction of all-cause mortality in COPD. *Chest* 2020; **158:** 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:** 150–55.
32 Moslemi A, Makimoto K, Tan WC, et al. Quantitative CT lung imaging and machine learning improves prediction of emergency room visits and hospitalizations in COPD. *Acad Radiol* 2023; **30:** 707–16.
33 Nguyen TT. Using random forest model for risk prediction of hospitalization and rehospitalization associated with chronic obstructive pulmonary disease. PhD thesis, University of Minnesota, 2017.
34 Pinto-Plata V, Casanova C, Divo M, 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. Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging; March 18, 2021.
36 Tang C, Plasek JM, Shi X, et al. Estimating time to progression of chronic obstructive pulmonary disease with tolerance. *IEEE J Biomed Health Inform* 2021; **25:** 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:** 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. 2018 IEEE International Conference on Bioinformatics and Biomedicine; Dec 3–6, 2018.

39 Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDGene) study design. *COPD* 2010; **7:** 32–43.

40 Vestbo J, Anderson W, Coxson HO, et al. Evaluation of COPD longitudinally to identify predictive surrogate end-points (ECLIPSE). *Eur Respir J* 2008; **31:** 869–73.

41 Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med* 2019; **170:** W1–33.

42 Fellicious C, Weissgerber T, Granitzer M. Effects of random seeds on the accuracy of convolutional neural networks. International Conference on Machine Learning, Optimization, and Data Science; July 19–23, 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:** 25–34.

44 Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013; **10:** e1001381.

45 Jang J-H, Choi J, Roh HW, et al. Deep learning approach for imputation of missing values in actigraphy data: algorithm development study. *JMIR Mhealth Uhealth* 2020; **8:** 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:** 1929–58.

47 Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. 32nd International Conference on Machine Learning; July 6–11, 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. 32nd Conference on Neural Information Processing Systems; Dec 2–8, 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:** 1110–16.

51 Mamary AJ, Stewart JI, Kinney GL, 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:** 177–84.

52 Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health* 2022; **4:** e406–14.

53 López-Campos JL, Tan W, Soriano JB. Global burden of COPD. *Respirology* 2016; **21:** 14–23.

54 McCradden MD, Anderson JAA, A Stephenson E, et al. A research ethics framework for the clinical translation of healthcare machine learning. *Am J Bioeth* 2022; **22:** 8–22.

55 Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023; **620:** 172–80.

56 Wei J, Bosma M, Vincent, et al. Finetuned language models are zero-shot learners. *arXiv* 2021; published online Feb 8. https://doi.org/10.48550/arXiv.2109.01652 (preprint).

57 Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *NPJ Digit Med* 2018; **1:** 40.