OPEN ACCESS

Original research

Development and validation of PRECISE-X model: predicting first severe exacerbation in COPD

Mohsen Sadatsafavi ,[1] Marc Miravitlles ,[2] Jennifer K Quint ,[3] Valeria Perugini ,[4] Hamid Tavakoli,[5] Joseph Emil Amegadzie,[6] Bernardino Alcazar Navarrete ,[7] on behalf of the Respiratory Effectiveness Group (REG)-COPD working group

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/thorax-2025-223770).

For numbered affiliations see end of article.

Correspondence to
Dr Marc Miravitlles;
marcm@separ.es

Received 20 June 2025
Accepted 2 December 2025

ABSTRACT

Objectives In patients with chronic obstructive pulmonary disease (COPD), severe exacerbations (ECOPDs) impose significant morbidity and mortality. Current guidelines emphasise using ECOPD history to inform preventive treatments but offer limited guidance for risk stratification for the first severe ECOPD.

Methods We developed and validated PRECISE-X using a cohort of newly diagnosed COPD patients from the UK's Clinical Practice Research Datalink (2004–2022), to predict first severe ECOPD over 5 years (primary outcome) and 12 months (secondary outcome). Predictors were selected via clinical expertise and data-driven methods. Internal-external cross-validation was performed across practice regions to evaluate the model's out-of-sample performance in terms of discrimination (c-statistic), calibration and net benefit.

Results The study included 2 19 015 patients (mean age 66.0; 42.4% female). Observed risk of first severe ECOPD was 29.5% at 5 years (4.2% at 1 year). The final model included four mandatory predictors (sex, age, Medical Research Council dyspnoea score and forced expiratory volume in 1 second) and 28 optional predictors. In internal-external cross-validation, the average out-of-sample c-statistic was 0.836 (95% CI 0.827 to 0.846) for 5-year prediction and 0.756 (95% CI 0.746 to 0.766) for 1-year prediction. Calibration across regions was robust, and the model showed positive NB across a wide range of risk thresholds. In a secondary validation assessment among those with available spirometry data with confirmed airflow obstruction, the model was well calibrated and had only a modest decline in discriminatory performance.

Conclusions PRECISE-X accurately predicts the first severe COPD exacerbation using routine clinical data, supporting earlier risk stratification and proactive disease management.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Severe exacerbations of chronic obstructive pulmonary disease (COPD) are associated with high morbidity and mortality. Existing tools primarily focus on patients with a history of exacerbations, offering limited guidance for predicting the first severe event.

WHAT THIS STUDY ADDS

⇒ The PRECISE-X model accurately predicts the risk of a first severe COPD exacerbation using routine clinical data. It demonstrates strong performance across UK regions and enables early identification of high-risk patients.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This model supports proactive, personalised COPD management by identifying at-risk patients earlier. It may also inform the design of clinical trials targeting first exacerbation outcomes and guide treatment decisions in real-world settings.

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is a progressive and heterogeneous condition characterised by persistent respiratory symptoms and airflow limitation. Exacerbations of COPD (ECOPD)—defined by an acute worsening of respiratory symptoms with varying severity and clinical impact—play a critical role in the disease's natural history. While mild and moderate exacerbations can often be managed in outpatient settings, severe exacerbations typically require hospital admission, leading to substantial morbidity and mortality.[1 2] Frequent exacerbations are associated with a steeper decline in lung function, reduced health-related quality of life and increased healthcare resource utilisation.[2] Despite advancements in the pharmacological and non-pharmacological management of COPD,[3] the ability to accurately predict and proactively prevent hospital admissions due to severe ECOPDs remains limited.

The consequences of severe ECOPDs extend far beyond the immediate health crisis. Published data indicate that patients requiring hospital admission for ECOPD face markedly elevated short-term and long-term mortality rates,[4] reflecting both cardiopulmonary complications and broader systemic impact of acute events on disease progression.[5 6] These patients are also at heightened risk of recurrent exacerbations, establishing a cycle of deterioration that undermines clinical stability and accelerates disease progression.[7 8] Given the interplay between exacerbations and disease trajectory, accurate risk stratification at the time of COPD

Check for updates

To cite: Sadatsafavi M, Miravitlles M, Quint JK, *et al*. *Thorax* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/thorax-2025-223770

diagnosis could transform the current reactive care model into proactive, targeted interventions.[9]

Although several tools exist to estimate ECOPD risk, most rely on prior exacerbation history and use specialised biomarkers not routinely available in clinical practice.[10] As such, their value in predicting a patient's first severe ECOPD is limited. There remains a clear need for simple, scalable models that integrate seamlessly into routine clinical workflows. To address this need, we aim to develop a risk prediction model to estimate the likelihood of hospitalisation for a first severe ECOPD at or near diagnosis. Using readily available clinical data, the model will support early risk identification—similar to cardiovascular tools like SCORE and Framingham—by relying on routine variables, with optional predictors if available.[11 12]

MATERIAL AND METHODS

This manuscript adheres to the Transparent Reporting of a multivariable model for Individual Prognosis or Diagnosis (TRIPOD) Statement.[13] We conducted a retrospective observational analysis to develop a risk prediction model for first severe ECOPD following COPD diagnosis, using routinely recorded clinical variables. The model was intended for use at the time of diagnosis in primary care.

Participants

We used the Clinical Practice Research Datalink (CPRD) Aurum (2004–2022),[14 15] containing electronic medical records from 1491 UK general practices. Eligible patients had a new COPD diagnosis based on a validated code-based algorithm (positive predictive value: 86.5%),[16 17] detailed in online supplemental section 1.

Inclusion criteria: patients aged ≥40 years at diagnosis and current or former smokers. Exclusion criteria: (1) any hospitalisation due to COPD occurring before diagnosis or within the first week after diagnosis; and (2) spirometry results within 1 year before or after diagnosis that were incompatible with COPD (defined as forced expiratory volume in 1 second (FEV_1)/forced vital capacity (FVC) ≥0.7). The latter criterion was applied to further improve the accuracy of the case definition (therefore, the positive predictive value of our case definition is likely higher than the original algorithm). We did not exclude patients based on inhaled treatments at diagnosis; instead, treatment history was included as a predictor. Predictions were adjusted for changes in follow-up treatment (see the Statistical Analysis section). The index date—when risk was estimated—was the date of COPD diagnosis. Patients required at least 1 month of follow-up. Variable definitions are detailed in online supplemental sections 1 and 2.

Primary outcome

The primary outcome was the occurrence of a severe ECOPD within 5 years of COPD diagnosis. We used the hospital-based definition of severe exacerbation as evaluated by Whittaker *et al*.[18] We chose severe exacerbations as they are more accurately verifiable and have a substantially greater impact on long-term outcomes compared with moderate events.[19] Specifically, a severe ECOPD was defined as an episode of hospital admission with (International Classification of Diseases ICD)-10 code recorded in hospital records (ICD10 J44.1, J44.0 or J44.9). The ICD-10 codes had to be in the first position or in the second position if the first was a code for lower respiratory tract infection (J22). This validated algorithm has demonstrated a sensitivity of 87.5% for hospitalised exacerbations.[20 21] Because this outcome is objectively ascertained based on inpatient records, its accuracy is unlikely to be influenced by practice region or demographic factors. The secondary outcome was the occurrence of severe ECOPD within 12 months since diagnosis, defined using the same criteria as for the primary outcome.

Predictors

We followed a hybrid hypothesis-free and expert-elicited approach to predictor selection, combining data-driven methods with clinical expertise to prioritise variables routinely available at the point of care. Candidate predictors included demographics, symptom score (Medical Research Council (MRC) Dyspnoea Scale), spirometry measures, socioeconomic status, blood eosinophil count and co-existing or comorbid conditions.

The primary spirometric variable was forced FEV_1. The model was designed to accept either raw or percent-predicted FEV_1, with or without bronchodilation. When percent-predicted values were available, we back-transformed them to raw FEV_1 using regression equations incorporating age, sex and body mass index (BMI; see online supplemental section 3). Exploratory analyses showed raw FEV_1 offered better model fit and fewer missing values. Including age, sex and BMI in the model allowed for proper contextualisation of raw FEV_1, effectively adjusting for individual characteristics. Postbronchodilator values were used when available; otherwise, prebronchodilator values were included, reflecting clinical practice.

For most predictors, we used a look-back window of 12 months. For MRC, the look-back window was 2 years, and a look-forward window of 12 months after the index date was also applied. The latter criterion was applied because for some patients the MRC is assessed after a COPD diagnosis is established. Similarly, BMI was assessed using a 5-year look-back and 1-year look-forward windows. All comorbidity conditions were evaluated in the entire available time window up to the index date. An exception was asthma: given that some COPD patients might be initially diagnosed as asthma, we required records >2 years before diagnosis to minimise misclassified cases later re-coded as COPD.

Following preliminary analyses and expert input, sex, age, FEV_1 (raw or % predicted) and MRC Dyspnoea Score were designated essential due to clinical relevance and availability. These are required to generate a prediction, while optional predictors can vary, supporting flexible use across different clinical settings.

Statistical analysis

We did not calculate a formal sample size, given the large, population-based dataset and expected >20 000 events based on prior studies. We adhered to best practice standards for prediction modelling[22] and for time-to-event outcomes,[23] and we imputed missing data, adjusted for treatment drop-in and conducted internal-external validation to assess model performance.

Imputation of missing predictors

As it is common in electronic medical records, we anticipated a high level of missingness for some predictors. However, methodological guidelines recommend against removing such predictors based on missingness levels alone (or restricting the sample to complete cases, which causes selection bias) and instead recommend proper imputation methods, followed by out-of-sample validation.[24] We followed best practice standards for missing predictor imputation, which recommend different approaches for mandatory and optional predictors.[24] For mandatory

predictors, we used multiple imputation via chained equations, generating five complete datasets. For non-essential predictors, we applied a regression-based approach, allowing these variables to be imputed at the time of model use based on other predictors. The assumption underlying the imputation of missing values is that conditional on the variables included in the imputation model. There are no systematic differences between the actual value of the missing and non-missing predictors.

Adjustment for treatment drop-in

Because many patients change their medications over the 5-year study period, potentially affecting ECOPD risk, it was important for the model to clearly define the risk it computes.[25] The present model reports the expected risk under current treatment (ie, if no change in treatment occurs). We consider this the most relevant prediction, as the 'baseline risk' before any medication changes.

We used Marginal Structural Modelling to adjust for treatment drop-in (ie, the addition of medications to baseline treatment in the future).[25] Follow-up time was divided into adjacent 6-month periods for each patient. In each period, we identified the initiation of any of the following medications: long-acting muscarinic agents (LAMA), long-acting beta agonists (LABA), inhaled corticosteroids (ICS) or azithromycin (when used for more than 30 days). We applied inverse probability-of-treatment weighting to calculate, for each patient, a weight that represents their probability of initiating any of such treatments. Separate logistic regression models were used within each period.

Model development and validation

We followed the most recent practice recommendations, which emphasise using all the data for development rather than splitting data into separate external and internal validation sets.[23] Such recommendations encourage taking advantage of any natural clustering in the data to perform internal-external validation. Accordingly, we used the clustering of the data across the nine health regions in the UK for this purpose.

Internal-external cross-validation involved leaving out one region at a time, fitting the model on the remaining eight regions and validating it on the external, left-out region. The final model was based on using the entire dataset. As the model is tested in a sample that is not used for its training, this approach provides the same credibility as an external cross-validation while training the final model on the full data.

We used semiparametric survival (Cox proportional hazards) modelling with least absolute shrinkage and selection operator (LASSO) for prediction. We chose the Cox model due to the flexibility of this platform in separating baseline hazard from predictor effects, thus enabling the model to be adjusted based on outcome prevalence in other settings (as is often required when transporting the model into a new population with different outcome prevalence[26]). Also, by shrinking predictor coefficients, LASSO prevents model overfitting that would negatively affect its generalisability (especially given the multitude of predictors). The tuning parameter of LASSO was automatically selected based on minimising cross-validated mean-squared error of predictions.

After a preliminary analysis verified the large effective sample size, we included an initial set of 32 predictors, comprising four essential and 28 optional predictors. One-way interaction terms among essential predictors, as well as selected interactions between essential and optional predictors, were included (eg, asthma history by sex) based on clinical judgement. In line with best practice standards, we also considered the missingness of a predictor as a separate binary predictor.[24] No rescaling or standardisation of predictors was performed.

Metrics of model performance were the c-statistic (also known as the area under the receiver operating characteristic curve—the closer to one, the better), mean calibration (the difference between the average observed and predicted risk—the closer to zero, the better) and calibration slope (capturing if the calibration plot follows the identity line—the closer to one, the better). We also evaluated the clinical utility of the model in terms of its net benefit (NB).[27] Unlike statistical metrics of model performance, whose relevance to the practical usefulness of a model is not always clear, NB is a utility measure: if a model has a higher NB over not using it, it is expected to provide clinical utility. The NB of a model needs to be evaluated at thresholds of interest for classifying patients into low-risk versus high-risk categories. To the best of our knowledge, such thresholds have not been determined for severe ECOPDs (an example of an established threshold is a 10% 10-year risk of cardiovascular outcomes for initiating statin therapy).[28] As such, we evaluated NB at 0.1–0.9 thresholds (with 0.1 increments). C-statistic, calibration plots and NB were calculated as time-dependent metrics for the 5-year time horizon (primary outcome) and 1-year time horizon (secondary outcome).

Because the internal-external cross-validation was repeated nine times (each time leaving one region out), this approach generated nine sets of model performance metrics. We used random-effects meta-analysis to pool these results. We assessed whether the variability across regions was large enough to warrant local adjustments of the model for each jurisdiction. The final model was fitted on the entire dataset.

Secondary validation

While we used a validated case definition with high predictive value for COPD, to examine the validity of predictions among those with spirometrically confirmed COPD, we performed a dedicated validation study in the subset of patients with available spirometry data where $FEV_1/FVC < 0.7$.

RESULTS

A total of 219 015 patients were included in the final analytical sample. Table 1 presents the demographic characteristics and the risk factor profile of the sample. Of the total, 42.4% were female. The average age at the time of diagnosis was 66.0 years, with average FEV_1(%) of 64.3 of percent predicted and an FEV_1/FVC ratio of 0.53.

On average, each patient contributed 1.80 years of follow-up time, during which 23 205 severe ECOPD events were recorded. Figure 1 provides the Kaplan-Meier curve for survival (not experiencing severe ECOPD). The observed probability of experiencing a severe ECOPD by year 5 was 29.5% (95% CI 29.1% to 29.8%). This probability was 4.2% (95% CI 4.1% to 4.3%) by 12 months.

The final model included the four essential predictors: age, sex, FEV1 (raw or percent predicted) and MRC score, and the following optional predictors: smoking status, history of all-cause hospital admissions and emergency department visits, BMI, FVC, socioeconomic status (Townsend Score), blood eosinophil count and history of several co-existing conditions (asthma, anxiety, hypertension, heart failure, cerebrovascular disease, ischaemic heart disease, gastro-oesophageal reflux disease, sleep apnoea, chronic kidney disease, bronchiectasis, osteoporosis, pneumonia and active rhinitis), as well as inhaler medications (ICS, LABA, LAMA, short-active anti-muscarinic agent, short-acting

Table 1 Baseline characteristics: Predicting ECOPD among patients newly diagnosed with COPD

Variable*	Male (N=1 26 064)	Female (N=92 951)	Total (N=2 19 015)
Age at diagnosis of COPD, mean (SD)	66.3 (11.2)	65.6 (11. 6)	66.0 (11.4)
Current smoker, N (%)	70 428 (55.9)	57 407 (61.8)	127 835 (58.3)
BMI (kg/m^2), mean (SD)	26.9 (7.9)	26.8 (8.2)	26.9 (8.0)
% missing	57.8	56.8	57.4
Blood eosinophil count (X10^9/L), mean (SD)	0.123 (0.291)	0.099 (0.213)	0.113 (0.261)
% missing	18.8	16.0	17.6
FEV1, mean (SD)	1.94 (0.75)	1.43 (0.63)	1.73 (0.75)
% missing	70.5	71.3	70.8
FEV1 percent predicted, mean (SD)	63.0 (18.5)	66.2 (18.3)	64.3 (18.5)
% missing	84.0	84.5	84.3
FEV1/FVC ratio, mean (SD)	0.53 (0.12)	0.54 (0.09)	0.53 (0.11)
% missing	45.8	54.1	49.3
FVC, mean (SD)	3.09 (0.98)	2.19 (0.72)	2.72 (0.98)
% missing	42.9	47.0	44.6
MRC scale, N (%)			
1	16 977 (16.8%)	10 290 (13.9%)	27 267 (15.6%)
2	34 251 (33.8%)	25 157 (34.0%)	59 408 (33.9%)
3	26 417 (26.1%)	20 272 (27.4%)	46 689 (26.6%)
4	18 340 (18.1%)	13 688 (18.5%)	32 028 (18.3%)
5	5341 (5.3%)	4504 (6.1%)	9845 (5.6%)
% missing	19.6	20.5	20.0
Socioeconomic status, N (%)			
1	18 420 (14.6)	12 071 (13.0)	30 491 (14.0)
2	22 378 (17.8)	15 881 (17.1)	38 259 (17.5)
3	23 831 (18.9)	16 912 (18.2)	40 743 (18.6)
4	27 544 (21.9)	20 809 (22.4)	48 353 (22.1)
5	33 595 (26.7)	27 049 (29.2)	60 644 (27.8)
% missing	0.2	0.2	0.2
History of all-cause admissions, N (%)	21 451 (17.0)	14 984 (16.1)	36 435 (16.6)
History of all-cause emergency department visits, N (%)	17 305 (13.7)	12 003 (12.9)	29 308 (13.4)
Co-existing conditions, N (%)			
Active rhinitis	3301 (2.6)	2368 (2.5)	5669 (2.6)
Anxiety	59 094 (46.9)	53 150 (57.2)	112 244 (51.2)
Asthma	26 592 (21.1)	25 865 (27.8)	52 457 (24.0)
Arrhythmia	2819 (2.2)	1590 (1.7)	4409 (2.0)
Bronchiectasis	2630 (2.1)	2005 (2.2)	4635 (2.1)
Chronic kidney disease	40 512 (32.1)	36 477 (39.2)	76 989 (35.2)
Diabetes	34 779 (27.6)	22 117 (23.8)	56 896 (26.0)
Gastro-oesophageal reflux disease	1991 (1.6)	1821 (2.0)	3812 (1.7)
Heart failure	8008 (6.4)	3643 (3.9)	11 651 (5.3)
Hypertension	21 127 (16.8)	14 958 (16.1)	36 085 (16.5)
Ischaemic heart disease	24 282 (19.3)	9812 (10.6)	34 094 (15.6)
Osteoporosis	12 540 (9.9)	13 060 (14.1)	25 600 (11.7)
Pneumonia	1572 (1.2)	1686 (1.8)	3258 (1.5)
Sleep apnoea	2528 (2.0)	731 (0.8)	3259 (1.5)
Stroke	17 482 (13.9)	9357 (10.1)	26 839 (12.3)
Medication use†			
SABA	71 927 (57.1)	59 763 (64.3)	13,1690 (60.1)
SAMA	9834 (7.8)	7460 (8.0)	17 294 (7.9)
LABA	30 600 (24.3)	26 388 (28.4)	56 988 (26.0)
LAMA	22 378 (17.8)	16 510 (17.8)	38 888 (17.8)
ICS	40 158 (31.9)	35 405 (38.1)	75 563 (34.5)
Mucolytics	3711 (2.9)	2745 (3.0)	6456 (2.9)
Region, N(%)			
North East	5087 (4.0)	4653 (5.0)	9740 (4.4)
North West	27 929 (22.2)	22 687 (24.4)	50 616 (23.1)
Yorkshire and The Humber	4416 (3.5)	3430 (3.7)	7846 (3.6)
East Midlands	2843 (2.3)	1882 (2.0)	4725 (2.2)
West Midlands	21 940 (17.4)	15 606 (16.8)	37 546 (17.1)
East of England	5395 (4.3)	3718 (4.0)	9113 (4.2)
London	17 530 (13.9)	12 070 (13.0)	29 600 (13.5)
South East	23 802 (18.9)	16 610 (17.9)	40 412 (18.5)
South West	17 122 (13.6)	12 295 (13.2)	29 417 (13.4)

*Variables without % missing value indicator had no missing values. For co-existing conditions, lack of diagnostic code was interpreted as lack of the corresponding condition.
†Defined by ingredients. For example, an individual on a single-inhaler ICS+LABA therapy would satisfy both ICS and LABA use.
BMI, body mass index; COPD, chronic obstructive pulmonary disease; ECOPD, exacerbations of chronic obstructive pulmonary disease; FEV1, forced expiratory volume at one second; FVC, forced vital capacity; ICS, inhaled corticosteroid; LABA, long-acting beta agonist; LAMA, long-acting muscarinic agent; MRC, Medical Research Council; SABA, short-acting beta-agonist; SAMA, short-active anti-muscarinic agent.

beta-agonist) and the use of mucolytics. Detailed model structure and regression coefficients are provided in the online supplemental section 4.

Figure 2 provides the results of the meta-analyses for c-statistic (left panel), calibration in the large (middle panel) and calibration

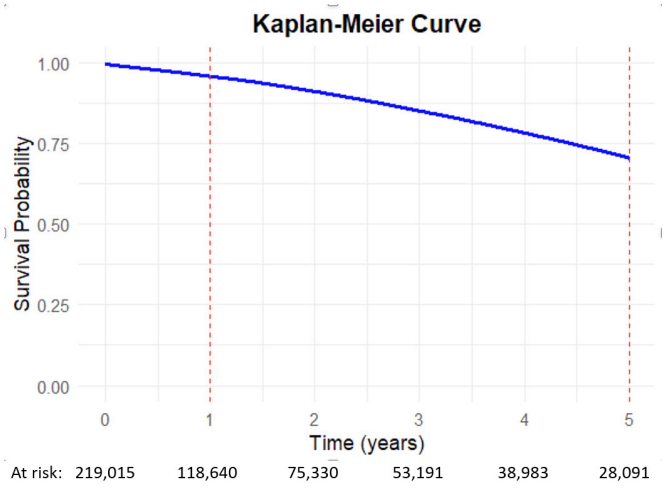

Figure 1 Kaplan-Meier curve (probability of being event-free) for severe exacerbations of chronic obstructive pulmonary disease.

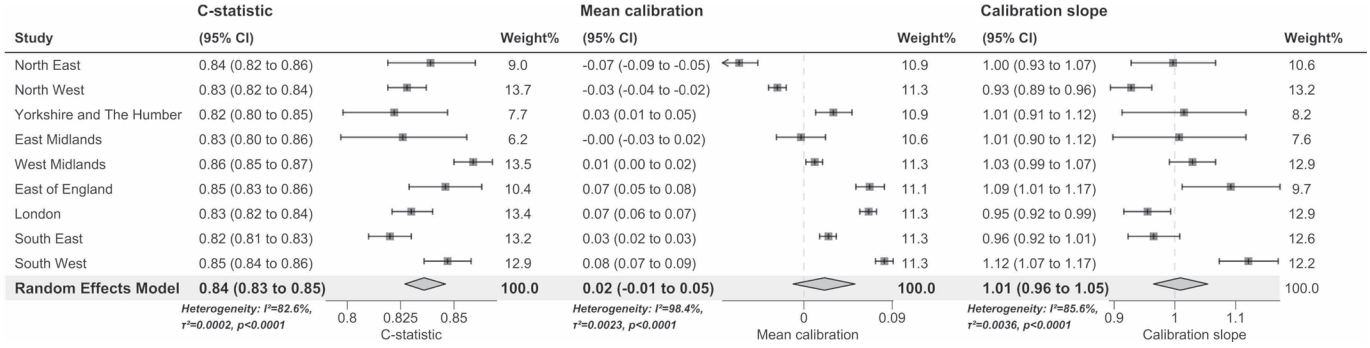

Figure 2 Meta-analysis of the out-of-sample performance of the model for 5-year prediction. Left panel: c-statistic; middle panel: mean calibration; right panel: calibration slope.

slope (right panel) of the full model for the primary endpoint (5-year risk prediction). The pooled c-statistic for 5-year prediction was 0.836 (95% CI 0.827 to 0.846). The pooled mean calibration error was not statistically significantly different from 0 (point estimate: 0.021 (95% CI −0.011 to 0.052). The calibration slope was 1.009 (95% CI 0.965 to 1.054), which was not statistically significantly different from 1.0.

In its simplest form, containing only essential predictors, the model still had a c-statistic of 0.792 (95% CI 0.784 to 0.800) for the primary endpoint and 0.705 (95%CI 0.697 to 0.714) for the secondary outcome.

Online supplemental section 5 reports results for 1-year prediction, with a random-effects pooled c-statistic of 0.756 (95%CI 0.746 to 0.766). Mean calibration error and calibration slope were, respectively, 0.001 and 1.009 (not significantly different from 0 and 1, respectively).

The incremental NB of the model (at 0.1 increments in risk threshold) is provided in figure 3. The performance of the model was evaluated for all nine regions and each of the nine thresholds, resulting in 81 region-threshold combinations. For both 5-year and 12-month predictions, the NB was negative in only one of the 81 region-threshold combinations.

Receiver operating characteristics and calibration plots for the final model on both outcomes are available in online supplemental section 6.

Online supplemental section 7 provides the results of the secondary validation study among those with recorded FEV_1/FVC<0.7. The model was very well calibrated in this subsample, and its discriminatory performance was only modestly lower than in the full sample (pooled c-statistic at 5 years and 12 months of 0.817 (95% CI 0.808 to 0.827) and 0.721 (95%CI 0.705 to 0.737)).

DISCUSSION

We used a large and representative primary care database from the UK to develop and validate a novel prediction model for first severe ECOPD after COPD diagnosis. With easy to obtain demographic and clinical variables, such as age, sex, MRC Dyspnoea Scale and FEV_1, this prediction model provided excellent out-of-sample accuracy for 5-year prediction of this outcome and very good accuracy for 12-month predictions. The out-of-sample calibration was also robust. Critically, while discrimination and calibration varied across regions, the model showed potential for delivering clinical utility (positive net NB) across a wide spectrum of risk thresholds. The model was well calibrated when validation was restricted to those with FEV_1/FVC<0.7.

A history of exacerbations is the strongest known predictor of future COPD exacerbations,[29] and models in established COPD populations consistently emphasise its importance.[30–32] However, this information is often unavailable at the time of diagnosis. As such, assessing the risk of a first hospitalised ECOPD becomes critical. Although some models predict severe ECOPD based on clinical characteristics, few have been specifically developed or validated for use at diagnosis, when prior exacerbation history is lacking.[33 34]

Identifying patients at high risk of severe ECOPD at the time of diagnosis can support timely risk reduction strategies,[35] especially given the negative impact of early events on disease progression.[36 37] By focusing on diagnosis-time data from real-world primary care, our model addresses a key gap and offers insights that general practitioners can readily apply in everyday clinical settings.

A major strength of our analysis is the use of internal-external cross-validation, which fully used the dataset while enabling

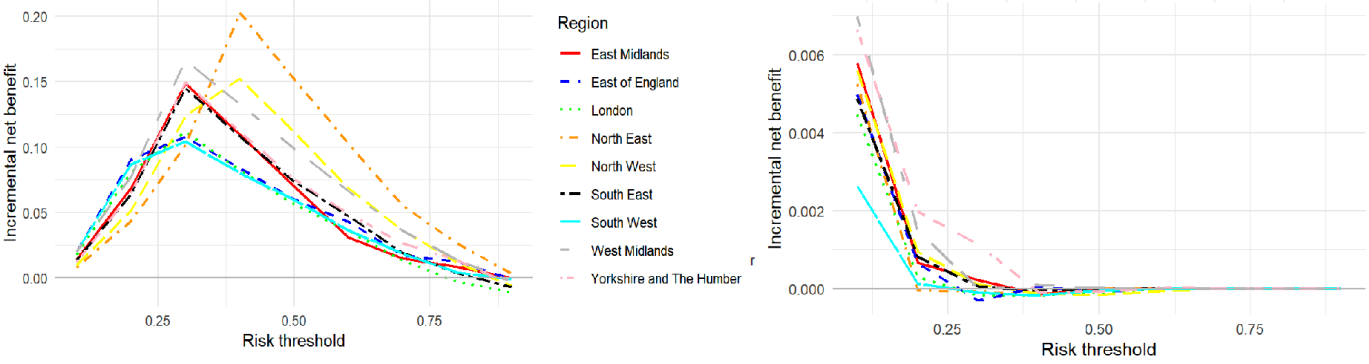

Figure 3 Incremental net benefit of the model over default strategies (treating no one or treating all) for 5-year (left) and 1-year (right) prediction.

robust out-of-sample testing. This method—now a best practice in prediction modelling—helps prevent overfitting and improves generalisability. As a result, our estimates of discrimination and calibration are more reliable than those from conventional split-sample validation. We also used rigorous methods to address missing data and treatment drop-in. Finally, the model's flexible structure—requiring only four core predictors but allowing up to 32—supports broad implementation, whether embedded in electronic records or used in resource-limited settings.

However, several limitations should be acknowledged. Some predictors (including spirometry values) had a high degree of missingness, which is common in electronic medical records. For example, only 28% of the development sample of the widely used QRISK3 model (for 10-year risk of cardiovascular disease) had non-missing essential predictors.[38] Current best practices advise against excluding potentially important predictors solely due to missingness. Selection bias by focusing on complete cases is likely to have a more severe impact on the validity of results than the violation of the assumption of missingness at random (conditional on the value of other predictors).[24] We also acknowledge that some predictors did not have the desired level of granularity. For example, reliable pack-years calculation in CPRD is not currently feasible. This might have affected the predictive power of the model. While missing or low-resolution data may reduce predictive power, the true measure of performance lies in out-of-sample validation, where our model remained robust.

While the model demonstrated strong predictive accuracy and calibration across UK primary care regions, its generalisability to other healthcare systems, populations and disease patterns remains uncertain. Independent external validation—especially in non-UK settings—is needed to confirm both performance and clinical utility. Additionally, we did not assess all possible predictors or their complex interactions. Incorporating biomarkers, socioenvironmental variables[39] or genomic data may further improve risk stratification. Lastly, we did not account for competing risks such as mortality, which is especially relevant in older, multimorbid populations. Future research using joint or multi-state models could provide a more comprehensive understanding of COPD progression and outcomes.

Another key observation is the variability in calibration and predictive accuracy across time horizons and regions. The model showed more consistent calibration over a 1-year prediction window compared with longer-term forecasts (judging by the heterogeneity of c-statistic across regions). This likely reflects the nature of COPD, where short-term risks are more directly influenced by current clinical and demographic factors, while longer-term outcomes are also shaped by differences in healthcare access, socioeconomic status and environmental exposures across regions. Despite regional disparities, our NB analyses indicate that the model retains meaningful clinical utility across different thresholds and geographies. However, independent external validation is essential before applying the model in new populations, particularly outside the UK, to ensure accuracy in settings with different risk profiles.

The strong predictive performance of PRECISE-X for identifying patients at risk of a first severe ECOPD has important clinical implications. Like the Framingham and Systematic Risk Score Evaluation (SCORE) tools used in cardiovascular prevention,[40] PRECISE-X could support early intervention strategies in COPD. By identifying high-risk individuals at diagnosis, clinicians could more effectively tailor preventive measures—such as pharmacotherapy, smoking cessation, pulmonary rehabilitation or vaccination. Proactively targeting those at greatest risk may help reduce healthcare burden, improve patient outcomes and slow disease progression.

Beyond direct patient care, this model also has important research applications. It provides a framework for designing clinical trials that prioritise severe ECOPD as a primary outcome—unlike most current COPD trials, where such events are secondary. By using our risk stratification tool during recruitment, trials can identify high-risk individuals more likely to benefit from novel therapies. This enrichment strategy can improve trial efficiency and increase the likelihood of detecting meaningful treatment effects.

In summary, our study shows that the risk of a first severe exacerbation in newly diagnosed COPD patients can be accurately predicted using routinely collected clinical data—even without prior exacerbation history. The model's consistent performance across regions and risk thresholds highlights its potential to support a proactive, prevention-focused approach to COPD care. Integrating the tool into electronic health records could enable early identification of high-risk patients and timely interventions. It may also enhance randomised trial design by facilitating efficient recruitment of high-risk individuals. Further validation and implementation in diverse settings could improve outcomes and COPD care delivery.

Author affiliations
[1]Respiratory Evaluation Sciences Program, Faculty of Pharmaceutical Sciences, University of British Columbia, Vancouver, British Columbia, Canada
[2]Respiratory Department, Vall d'Hebron University Hospital / Vall d'Hebron Research Institute (VHIR), Vall d'Hebron Barcelona Hospital Campus. CIBER de Enfermedades Respiratorias (CIBERES), Barcelona, Spain
[3]School of Public Health, Imperial College London, London, UK
[4]Respiratory Effectiveness Group, Cambridgeshire, UK
[5]Meta Health Informatics, Vancouver, British Columbia, Canada
[6]Epidemiology and Intelligence Services and Faculty of Pharmaceutical Sciences, British Columbia Centre for Disease Control and The University of British Columbia, Vancouver, British Columbia, Canada
[7]Respiratory Department, Hospital Universitario Virgen de las Nieves, ibs.Granada, Universidad de Granada, Granada, Spain

Collaborators Respiratory Effectiveness Group (REG)-COPD working group: Nicolas Roche - nicolas.roche@aphp.fr; Joan B Soriano - jbsoriano2@gmail.com; Omar Usmani - o.usmani@imperial.ac.uk; Therese Lapperre - Therese.Lapperre@uza.be; Chin Kook Rhee - chinkook77@gmail.com; Matevz Harlander - matevz.harlander@gmail.com; Alan Kaplan - for4kids@gmail.com.

Contributors MM and BAN conceived the study question. MS, BAN, MM and VP developed the original study protocol. VP was responsible for project management. VP and MS procured the data. JKQ provided expert advice on data elements and shared analysis code from previous studies. MS developed the statistical analysis plan. JEA created the study cohort. HT conducted statistical analyses. JKQ and MS supervised cohort creation and data analysis. MS and BAN wrote the first version of the manuscript. All authors critically commented on the manuscript and approved the final version. MS is the data guardian for this study. MM is the guarantor for this study.

Funding This study was funded by an unrestricted grant from AstraZeneca UK Limited (grant n: ESR-22-21856).

Disclaimer This manuscript has been deposited as a preprint on SSRN and can be accessed at https://ssrn.com/abstract=5167969.

Competing interests MS has received funding from GSK, AstraZeneca and Boehringer Ingelheim directly into his research account at the University of British Columbia for unrelated projects. He has received honoraria from GSK, AstraZeneca and Boehringer Ingelheim for activities unrelated to this work. MM has received speaker fees from AstraZeneca, Boehringer Ingelheim, Chiesi, Cipla, GlaxoSmithKline, Glenmark Pharmaceuticals, Menarini, Kamada, Takeda, Zambon, Tabuk Pharmaceuticals, CSL Behring, Zambon, Specialty Therapeutics, Sanofi/Regeneron, Grifols and Novartis, consulting fees from AstraZeneca, Atriva Therapeutics, Boehringer Ingelheim, BEAM Therapeutics, GondolaBio, Chiesi, GlaxoSmithKline, CSL Behring, Ferrer, Korrobio, Menarini, Mereo Biopharma, Spin Therapeutics, Specialty Therapeutics, Palobiofarma SL, Takeda, Novartis, Novo Nordisk, Sanofi/Regeneron, Zambon, Zentiva and Grifols and research grants from Grifols. JKQ has been supported by institutional research grants from the Medical Research Council, NIHR,

Health Data Research, GSK, BI, AZ, Insmed, Sanofi and received personal fees for advisory board participation, consultancy or speaking fees from GlaxoSmithKline, BI, Sanofi, Chiesi, AstraZeneca. BAN has received fees from GSK, grants, personal fees and non-financial support from AstraZeneca, personal fees and non-financial support from Boehringer Ingelheim, personal fees and non-financial support from Chiesi, grants, personal fees and non-financial support from Laboratorios Menarini, personal fees from Bial, personal fees from Zambon, personal fees from Gilead, personal fees from MSD, and personal fees from Sanofi. VP, HT and JEA have no conflict of interest.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information. The dataset used in this study was obtained under licence from Clinical Practice Research Datalink (CPRD) Aurum database. The study protocol and data analysis code are available as online supplemental materials.

ORCID iDs
Mohsen Sadatsafavi https://orcid.org/0000-0002-0419-7862
Marc Miravitlles https://orcid.org/0000-0002-9850-9520
Jennifer K Quint https://orcid.org/0000-0003-0149-4869
Valeria Perugini https://orcid.org/0000-0002-9516-4862
Bernardino Alcazar Navarrete https://orcid.org/0000-0003-2356-9366

REFERENCES

1 Waeijen-Smit K, Crutsen M, Keene S, *et al*. Global mortality and readmission rates following COPD exacerbation-related hospitalisation: a meta-analysis of 65 945 individual patients. *ERJ Open Res* 2024;10:00838-2023.
2 Soler-Cataluña JJ, Martínez-García MA, Román Sánchez P, *et al*. Severe acute exacerbations and mortality in patients with chronic obstructive pulmonary disease. *Thorax* 2005;60:925–31.
3 Aaron SD. Management and prevention of exacerbations of COPD. *BMJ* 2014;349:g5237.
4 Suissa S, Dell'Aniello S, Ernst P. Long-term natural history of chronic obstructive pulmonary disease: severe exacerbations and mortality. *Thorax* 2012;67:957–63.
5 Vogelmeier CF, Rhodes K, Garbe E, *et al*. Elucidating the risk of cardiopulmonary consequences of an exacerbation of COPD: results of the EXACOS-CV study in Germany. *BMJ Open Resp Res* 2024;11:e002153.
6 Hurst JR, Gale CP, Global Working Group on Cardiopulmonary Risk. MACE in COPD: addressing cardiopulmonary risk. *Lancet Respir Med* 2024;12:345–8.
7 Halpin DM, Miravitlles M, Metzdorf N, *et al*. Impact and prevention of severe exacerbations of COPD: a review of the evidence. *Int J Chron Obstruct Pulmon Dis* 2017;12:2891–908.
8 García-Polo C, Alcázar-Navarrete B, Ruiz-Iturriaga LA, *et al*. Factors associated with high healthcare resource utilisation among COPD patients. *Respir Med* 2012;106:1734–42.
9 Agustí A, Celli BR, Criner GJ, *et al*. Global Initiative for Chronic Obstructive Lung Disease 2023 Report: GOLD Executive Summary. *Arch Bronconeumol* 2023;59:232–48.
10 Bellou V, Belbasis L, Konstantinidis AK, *et al*. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. *BMJ* 2019;367:l5358.
11 SCORE2-OP working group and ESC Cardiovascular risk collaboration. SCORE2-OP risk prediction algorithms: estimating incident cardiovascular event risk in older persons in four geographical risk regions. *Eur Heart J* 2021;42:2455–67.
12 D'Agostino RB Sr, Grundy S, Sullivan LM, *et al*. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA* 2001;286:180–7.
13 TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:q902.
14 Wolf A, Dedman D, Campbell J, *et al*. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol* 2019;48:1740–1740g.
15 Clinical Practice Research Datalink. CPRD aurum May 2022 (version 2022.05.001) [data set]. Clinical practice research datalink. 2022. Available: https://www.cprd.com/cprd-aurum-may-2022-dataset [accessed 06 Jan 2025]
16 Requena G, Wolf A, Williams R, *et al*. Feasibility of using Clinical Practice Research Datalink data to identify patients with chronic obstructive pulmonary disease to enrol into real-world trials. *Pharmacoepidemiol Drug Saf* 2021;30:472–81.
17 Quint JK, Müllerova H, DiSantostefano RL, *et al*. Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). *BMJ Open* 2014;4:e005540.
18 Whittaker H, Rothnie KJ, Quint JK. Exploring the impact of varying definitions of exacerbations of chronic obstructive pulmonary disease in routinely collected electronic medical records. *PLoS One* 2023;18:e0292876.
19 Rothnie KJ, Müllerová H, Smeeth L, *et al*. Natural History of Chronic Obstructive Pulmonary Disease Exacerbations in a General Practice-based Population with Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* 2018;198:464–71.
20 Rothnie KJ, Müllerová H, Thomas SL, *et al*. Recording of hospitalizations for acute exacerbations of COPD in UK electronic health care records. *Clin Epidemiol* 2016;8:771–82.
21 Rothnie KJ, Müllerová H, Hurst JR, *et al*. Validation of the Recording of Acute Exacerbations of COPD in UK Primary Care Electronic Healthcare Records. *PLoS One* 2016;11:e0151357.
22 Collins GS, Dhiman P, Ma J, *et al*. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* 2024;384:e074819.
23 McLernon DJ, Giardiello D, Van Calster B, *et al*. Assessing Performance and Clinical Usefulness in Prediction Models With Survival Outcomes: Practical Guidance for Cox Proportional Hazards Models. *Ann Intern Med* 2023;176:105–14.
24 Sisk R, Sperrin M, Peek N, *et al*. Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study. *Stat Methods Med Res* 2023;32:1461–77.
25 Sperrin M, Martin GP, Pate A, *et al*. Using marginal structural models to adjust for treatment drop-in when developing clinical prediction models. *Stat Med* 2018;37:4142–54.
26 Steyerberg EW, Borsboom GJJM, van Houwelingen HC, *et al*. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23:2567–86.
27 Sadatsafavi M, Adibi A, Puhan M, *et al*. Moving beyond AUC: decision curve analysis for quantifying net benefit of risk prediction models. *Eur Respir J* 2021;58:2101186:58:.
28 Lloyd-Jones DM, Braun LT, Ndumele CE, *et al*. Use of Risk Assessment Tools to Guide Decision-Making in the Primary Prevention of Atherosclerotic Cardiovascular Disease: A Special Report From the American Heart Association and American College of Cardiology. *J Am Coll Cardiol* 2019;73:3153–67.
29 Hurst JR, Vestbo J, Anzueto A, *et al*. Susceptibility to exacerbation in chronic obstructive pulmonary disease. *N Engl J Med* 2010;363:1128–38.
30 Bertens LCM, Reitsma JB, Moons KGM, *et al*. Development and validation of a model to predict the risk of exacerbations in chronic obstructive pulmonary disease. *Int J Chron Obstruct Pulmon Dis* 2013;8:493–9.
31 Adibi A, Sin DD, Safari A, *et al*. The Acute COPD Exacerbation Prediction Tool (ACCEPT): a modelling study. *Lancet Respir Med* 2020;8:1013–21.
32 Safari A, Adibi A, Sin DD, *et al*. ACCEPT 2·0: Recalibrating and externally validating the Acute COPD exacerbation prediction tool (ACCEPT). *EClinicalMedicine* 2022;51:101574.
33 Garcia-Aymerich J, Serra Pons I, Mannino DM, *et al*. Lung function impairment, COPD hospitalisations and subsequent mortality. *Thorax* 2011;66:585–90.
34 Nuñez A, Marras V, Harlander M, *et al*. Clinical and spirometric variables are better predictors of COPD exacerbations than routine blood biomarkers. *Respir Med* 2020;171:106091.
35 Anzueto A, Miravitlles M. Chronic Obstructive Pulmonary Disease Exacerbations: A Need for Action. *Am J Med* 2018;131:15–22.
36 Sadatsafavi M, Xie H, Etminan M, *et al*. The association between previous and future severe exacerbations of chronic obstructive pulmonary disease: Updating the literature using robust statistical methodology. *PLoS One* 2018;13:e0191243.
37 Miravitlles M, Calle M, Alvarez-Gutierrez F, *et al*. Exacerbations, hospital admissions and impaired health status in chronic obstructive pulmonary disease. *Qual Life Res* 2006;15:471–80.
38 Hippisley-Cox J, Coupland CAC, Bafadhel M, *et al*. Development and validation of a new algorithm for improved cardiovascular risk prediction. *Nat Med* 2024;30:1440–7.
39 Fernández-Villar A, Cimas Hernando JE, Figueira Gonçalves JM, *et al*. Continuity of Care in Chronic Obstructive Pulmonary Disease Exacerbations: Challenges and Priorities. *Arch Bronconeumol* 2024;60:327–9.
40 Lloyd-Jones DM, Wilson PWF, Larson MG, *et al*. Framingham risk score and prediction of lifetime risk for coronary heart disease. *Am J Cardiol* 2004;94:20–4.