Identifying biomarkers of mild-stage emphysema in COPD patients via interpretable machine learning

L Egerod[1,2], CL Bager[3], DJ Leeming[2], A Azim[4], B Angermann[5], MA Karsdal[3], CB Nanthakumar[6], JC Yates[7], A Platt[4], GW Clarke[5,8], R Gupta[9], JMB Sand[2]

[1]Department of Biomedical Sciences, University of Copenhagen, Copenhagen, Denmark; [2]Hepatic and Pulmonary Research, Nordic Bioscience, Herlev, Denmark; [3]Nordic Bioscience, Herlev, Denmark; [4]Translational Science and Experimental Medicine, Research and Early Development, Respiratory and Immunology, AstraZeneca, Cambridge, United Kingdom; [5]Translational Science and Experimental Medicine, Research and Early Development, Respiratory and Immunology, AstraZeneca, Gothenburg, Sweden; [6]Respiratory Biology, GSK-HQ London, UK; [7]GSK R&D, Durham, NC, USA; [8]Peter Gorer Department of Immunobiology, School of Immunology and Microbial Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom; [9]Independent researcher

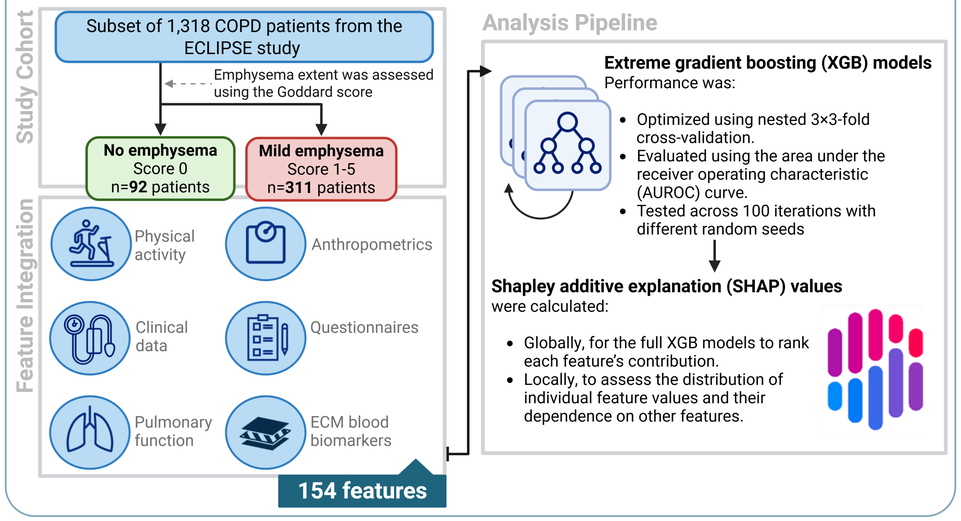
BACKGROUND

Emphysema results from alveolar damage—causing abnormal extracellular matrix (ECM) remodeling and impaired lung function. Early detection in mild, often asymptomatic stages is key to timely intervention. Although computed tomography (CT) scans are the most accurate detection method, pathological changes may occur before emphysema becomes visible, highlighting the need for non-invasive early detection approaches.

This study aimed to develop a proof-of-concept machine learning (ML) pipeline to identify patients with mild-stage emphysema using circulating biomarkers.

METHODS

The analyses were based on data from the **ECLIPSE study**, which enrolled 2,164 COPD patients between 2005 and 2010. For this study, participants needed high-quality baseline CT scans and available serum samples from the 3-month visit, resulting in **1,318 COPD patients** included in the analysis.



AUROC: area under the receiving operating characteristic; CT: computed tomography; ECM: extracellular matrix; ELA-HNE: neutrophil elastase-degraded elastin; ML: machine learning; PRO-C6: type VI collagen formation; SHAP: shapley additive explanation; XGB: extreme gradient boosting.

RESULTS

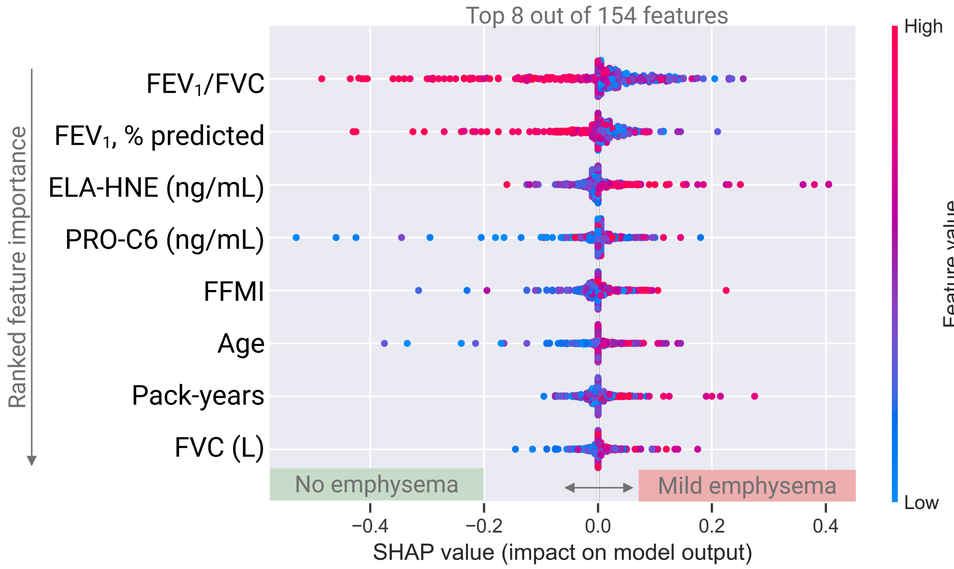
Model performance

XGB models integrating 154 features achieved an AUROC of **0.69** (95%CI: **0.65-0.73**)

Top features of mild emphysema

Low **FEV_1/FVC ratio** and **% predicted FEV_1**.

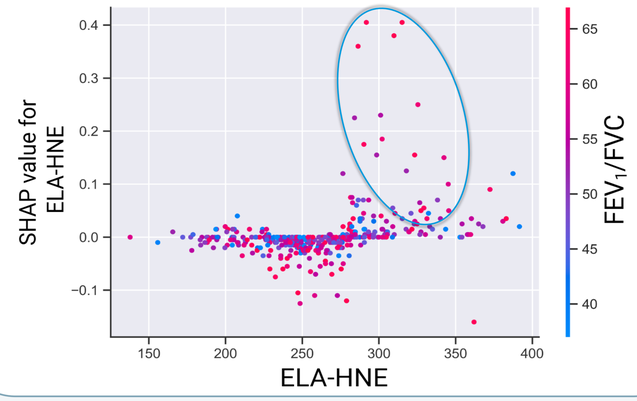
High levels of neutrophil elastase-degraded elastin (**ELA-HNE**), type VI collagen formation (**PRO-C6**), and **fat-free mass index**.



SHAP *"Beeswarm"* plot explained: The *y*-axis displays the feature names in order of importance, from top to bottom. The *x*-axis represents the SHAP value, indicating the degree of change in log-odds. Each point represents an individual patient (i.e., a row of data) from the original dataset. The color of each point corresponds to the value of the associated feature, with red indicating high values and blue indicating low values.

ECM remodeling in mild-stage patients without a decline in lung function

Some patients with **high levels of ELA-HNE** and a **high FEV1/FVC ratio** were classified as mild emphysema cases by the XGB models.



CONCLUSION

ML demonstrates potential for early-stage emphysema diagnosis through biomarker-driven methods.

Quantifying fragments of ECM remodeling—driven by immune cell activity and collagen formation—could potentially serve as early diagnostic biomarkers in patients without lung function decline.

Contact: Line Egerod, lelu@nordicbio.com
Funding: The ECLIPSE study (*NCT00292552*) was funded by GSK