Uncovering the heterogeneous effect of inhaled corticosteroids on COPD exacerbations with causal machine learning

Helene Huts[1, 2], Kenneth Verstraete[2], Thomas Strypsteen[1], Iwein Gyselinck[3], Michaël Staes[2,3], Amber Beersaerts[2], Jeroen Berrevoets[4], Wim Janssens[2, 3, *], Maarten De Vos[1, 5, *]

[1] STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium

[2] Laboratory of Respiratory Diseases and Thoracic Surgery (BREATHE), Department of Chronic Diseases and Metabolism (CHROMETA), KU Leuven, Leuven, Belgium

[3] Department of Respiratory Diseases, University Hospitals Leuven, Leuven, Belgium

[4] Department of Informatics, King's College London, Londen, United Kingdom

[5] Woman and Child, Department of Development and Regeneration, KU Leuven, Leuven, Belgium

* Co-senior authors

Corresponding Author

Wim Janssens, MD, PhD.

ORCID iD: http://orcid.org/0000-0003-1830-2982

Department of Chronic Diseases and Metabolism, ON1bis Herestraat 49 - box 706, 3000 Leuven, Belgium

E-mail: wim.janssens@uzleuven.be

ORDIC iDs

Helene Huts: http://orcid.org/0000-0002-0617-1316

Kenneth Verstraete: http://orcid.org/0000-0003-3790-417X

Thomas Strypsteen: http://orcid.org/0000-0001-7485-0321

Iwein Gyselinck: https://orcid.org/0000-0002-4068-7228

Michaël Staes: https://orcid.org/0009-0008-4560-9078

Amber Beersaerts: https://orcid.org/0009-0005-8586-2033

Jeroen Berrevoets: https://orcid.org/0000-0002-0396-5655

Wim Janssens: https://orcid.org/0000-0003-1830-2982

Maarten De Vos: https://orcid.org/0000-0002-3482-5145

Word count

Abstract: 309 words (350 words max)

Main text: 3108 words

Abstract

Background: High blood eosinophils counts can predict an enhanced response to inhaled corticosteroids (ICS) in patients with chronic obstructive pulmonary disease (COPD), but additional insight in treatment effect heterogeneity is needed to optimize clinical disease management. We investigated if causal machine learning models can detect heterogeneity in the effect of two ICS-containing therapies on both time to first

exacerbation and exacerbation rate, and if this approach can identify additional predictors of ICS response.

Methods: Baseline characteristics from patients in the fluticasone furoate/umeclidinium/vilanterol (FF/U/V) and U/V arms of the InforMing the PAthway of COPD Treatment (IMPACT) trial (ClinicalTrials.gov identifier NCT02164513) were used to train a causal survival forest that learned to estimate the effect of fluticasone furoate (an ICS) on time to first exacerbation for each patient. Similarly, a causal forest was trained for exacerbation rate. Results were averaged over 100 train and validation sets in a Monte-Carlo cross-validation approach. The analysis was repeated for a comparison of the FF/V and U/V arms.

Results: This analysis included 4048 FF/U/V, 4034 FF/V and 2025 U/V patients. Significant p-values for AUTOC (area under the targeting operator characteristic) indicated that all models ranked patients well according to individualized treatment and that significant heterogeneity was detected in the effect of ICS on both exacerbation outcomes. Respiratory symptoms, reversibility of airflow limitation and lung function, in addition to eosinophils, were consistently identified as important predictors by all models.

Conclusions: This hypothesis-generating secondary analysis of IMPACT detected significant heterogeneity in the effect of ICS and indicated that, in addition to eosinophils, other clinical patient characteristics also contribute to the prediction of ICS response on both time to first exacerbation and exacerbation rate. The causal machine learning approach in this analysis can identify treatment effect predictors and patterns of response, to enhance insight in COPD disease phenotypes and heterogeneous treatment effects and to improve personalized treatment allocations.

Key words

COPD Exacerbations

Bronchodilators

Corticosteroids

Eosinophils

Machine learning

Predictive Learning Models

Treatment effect heterogeneity

Background

Chronic obstructive pulmonary disease (COPD) is a complex disease that is characterized by a notable heterogeneity in clinical features, in underlying biological mechanisms and in the effect of treatments [1- 4]. Optimal clinical management of the disease requires a precision medicine approach, where treatment decisions are made based on the biological and clinical characteristics of each individual patient [1, 2, 4].

Inhaled corticosteroids (ICS) are effective to prevent COPD exacerbations [5-8]. Predefined [6-8] and post-hoc [9-11] analyses have shown that high blood eosinophil counts predict an enhanced effect of ICS treatment. Yet, the commonly proposed cutoffs – 100 and 300 eosinophils per µL, depending on the clinical context [12] – fall well within the commonly

observed physiological range for patients with COPD [13], making them broadly inclusive. Given the concern over ICS-associated adverse events [14], additional traits or biomarkers may strengthen confidence in initiating or withholding ICS, especially for exacerbating patients in the lower end of this eosinophil spectrum.

Causal machine learning (ML) models aim to estimate the effect of a treatment for each patient, based on their individual baseline characteristics [15]. In this context, individualized treatment effect (ITE) is defined as the difference between a patient's outcome with and without a certain treatment [16]. Causal ML models are flexible models that allow for non-linear interactions between baseline characteristics and a data-driven selection of the most important predictors [15]. This makes these novel models uniquely suited to investigate heterogeneity in the effect of COPD treatments.

The InforMing the PAthway of COPD Treatment (IMPACT; ClinicalTrials.gov: NCT02164513) trial investigated the benefits of a triple therapy of fluticasone furoate (FF, an ICS), umeclidinium (U, a long-acting muscarinic antagonist (LAMA)) and vilanterol (V, a long-acting β_2-agonist (LABA)) compared to dual therapies of FF/V and U/V [6]. In previous analyses of this trial, heterogeneity was detected in the effect of ICS on exacerbation rate with a subgroup analysis for blood eosinophil counts above 150 cells/µL [6], by modelling continuous blood eosinophil counts with fractional polynomials [10] and with a causal forest model that used a wide range of baseline characteristics as input [17]. In addition to exacerbation rate, time to first exacerbation is an outcome that could enhance insights

in the effect of ICS treatment [18]. Recent advancements have extended causal ML strategies to allow for the estimation of ITE for right-censored survival outcomes [19].

In a secondary analysis of the IMPACT trial, we aimed to investigate (1) if a causal ML model can detect heterogeneity in the effect of ICS on time to first exacerbation and (2) if a combination of causal ML models trained for related outcomes –time to first exacerbation and exacerbation rate – can identify a clear profile of strong ICS responders. We first compared the FF/U/V and U/V arms and extended the analysis to evaluate (3) if a pattern of heterogeneous response, similar to that in the FF/U/V versus U/V comparison, can also be observed when comparing the FF/V and U/V arms.

Methods

Study Participants, Outcomes and Baseline Characteristics

IMPACT was a phase III, 52-week, randomized, double-blind, 3-arm parallel group, global multicentre study. Details of the study design, the study protocol, and main results have been published previously [6]. Briefly, the trial was conducted in 37 countries from June 2014 to July 2017 and included 10,355 patients with symptomatic COPD and a history of exacerbations. Moderate exacerbations were defined per protocol as exacerbations that required treatment with antibiotics or systemic glucocorticoids. Severe exacerbations were defined as exacerbations that resulted in hospitalization or death. Patients who received triple therapy of FF/U/V had significantly lower annual rates of moderate or severe exacerbations (primary outcome) and significantly longer times until first

exacerbation (a secondary outcome) compared to patients who received dual therapies of U/V or FF/V.

The secondary analysis of data from IMPACT in this work was approved by the KU Leuven Institutional Privacy and Ethics review committee (SMEC/PRET, ID: G-2022-5402) and followed guidelines from the Predictive Approaches to Treatment effect Heterogeneity (PATH) [20] and Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD+AI) [21] statements.

A set of 17 patient baseline characteristics was used as input features for the models. The selection was made based on patient characteristics that were previously associated with an enhanced effect of ICS [6-10] or increased exacerbation risk [22-27]. Completeness of baseline characteristics was very high. Of the 10,355 patients in the intention-to-treat population, 10,107 patients who had no missing values for any of the selected baseline characteristics were included the analysis. More details on selected characteristics, missing data and preprocessing steps are provided in Supplementary File S1.

Model Development

To optimally use all available data, patients in the FF/U/V and U/V arms were split 100 times in a 70% training and 30% internal validation set to perform Monte-Carlo cross-validation [28]. Using patient baseline characteristics, a causal survival forest [19] learned to estimate the effect of FF on top of U/V for time to first COPD exacerbation for each patient. This effect was expressed as a difference in restricted mean survival time.

Similarly, a causal forest [29] learned to estimate the individualized effect of FF on top of U/V for COPD exacerbation rate. A limit of the causal forest is that it cannot take differences in follow-up time into account when modelling exacerbation rate. To reduce influence of unrealistically high apparent annual rates caused by very short follow-up times, apparent annual rates higher than 15 were set to 15 for 125 patients in the analysis. Both exacerbation outcomes were estimated with a time horizon of 52 weeks, in line with the follow-up time of the IMPACT trial.

Both forest-based causal machine learning models split patients in the train set based on their baseline characteristics to create a decision-tree-like structure that groups patients with similar treatment effects together. The final model is referred to as a "forest" because it is an ensemble of many such trees. To reduce bias and overfitting, each tree is built using a random subset of patients and characteristics, and different patients are used for building the tree structure and for determining the treatment effect in each leaf (a principle called "honesty"). An estimate of individualized treatment effect for an individual patient in the validation set is made by aggregating estimates from all trees. Estimates of treatment effect for each outcome were averaged over 100 internal validation sets to get a final estimate for each patient. R version 4.4.2 (R Foundation for Statistical Computing, Vienna, Austria) and the R *grf* package were used for both models. A detailed description of model development is provided in Supplementary File S1. The same methodology was repeated for FF/V versus U/V.

Model Performance Evaluation

Model performance was evaluated by calculating the area under the targeting operator characteristic (TOC) curve [30]. This is a metric that quantifies observed treatment effect gain relative to the observed average treatment effect. Patient are ranked from highest to lowest estimated effect and a curve is constructed by calculating the average treatment effect for increasing top fractions of ranked patients. The overall average treatment effect is subtracted to ensure that the curve ends at 0 when the last top fraction includes all patients. A higher area under the TOC (AUTOC) curve indicates an overall better ranking of patients from highest to lowest treatment effect (better discrimination). An AUTOC significantly different from zero indicates that significant heterogeneity in treatment effect was detected. To visualize heterogeneity and to evaluate overlap for both modelled outcomes, patients were ranked from highest to lowest estimated effect and split into 5 groups (quintiles). Model calibration was evaluated by comparing the average estimated and observed effects per quintile. A detailed description of evaluation methods is provided in Supplementary File S1.

Results

Study Participants, Outcomes and Baseline Characteristics

This analysis included 4048 FF/U/V patients, 2025 U/V patients and 4034 FF/V patients who had no missing data for any of the baseline characteristics of interest. For patients included in the analysis, average observed outcome values in the FF/U/V versus U/V comparison were significantly in favour of FF/U/V (rate ratio for exacerbation rate: 0.75, $p<0.0001$, hazard ratio for time to first exacerbation: 0.83, $p<0.0001$). The second treatment comparison in this work – FF/V versus U/V – was not part of the original trial

analysis [6] but may also show the heterogeneous effect of ICS. On average, observed exacerbation rates were significantly in favour of FF/V (rate ratio 0.87, p=0.0005), while no effect was observed for time to first exacerbation (hazard ratio 0.98, p=0.6556).

Model Development

The relative importance of patient baseline characteristics was extracted from the models and is shown in Fig. 1. The importance of patient baseline characteristics was determined based on how many times a characteristic was chosen to partition the data in the first splits of the forest model. Blood eosinophil counts were the most important characteristic for all models. Importance was distributed more equally among other characteristics. Some other characteristics important for all models were St. George's Respiratory Questionnaire (SGRQ) symptoms score, forced expiratory volume in 1 second (FEV1) reversibility, and FEV1 percent predicted.

Model Performance Evaluation

Table 1 provides an overview of performance metrics for the final ranking of patients and estimates of treatment effect that were obtained by averaging over 100 internal validation sets. Significant p-values for AUTOC indicated good discrimination performance: all models ranked patients well according to their estimated individualized treatment effect and significant treatment effect heterogeneity was detected. An evaluation of model calibration based on a comparison of average observed and estimated treatment effects per quintile (Fig. 2) showed that average estimated effects for the time-to-first-exacerbation model were close to average observed effects per quintile

for FF/U/V versus U/V, while for FF/V versus U/V the model underestimated the highest and lowest effects. The exacerbation-rate model tended to overestimate the magnitude of treatment effect for both treatment comparisons.

Table 1 Model performance

Treatment comparison	Modelled outcome	AUTOC	
		Mean (95% CI)	**p-value***
FF/U/V-U/V	Time to first exacerbation	11.88 (3.54 to 20.22)	0.0026
	Exacerbation rate	0.39 (0.27 to 0.50)	<0.0001
FF/V-U/V	Time to first exacerbation	22.06 (13.86 to 30.26)	<0.0001
	Exacerbation rate	0.27 (0.14 to 0.41)	<0.0001

The unit of the area under the targeting operator characteristic (AUTOC) curve depends on the modelled outcome. For time to first exacerbation this is a gain in treatment effect in days of restricted mean survival time. For exacerbation rate this is a gain in treatment effect in exacerbation rate. A significant p-value indicates that the AUTOC was significantly different from zero and that significant treatment effect heterogeneity was detected. *FF* Fluticasone Furoate, *U* Umeclidinium, *V* Vilanterol, *CI* Confidence interval. * p-value from a one-sided test

Patients in the FF/U/V versus U/V comparison who were estimated by the models to be in quintile 1 (highest estimated benefit from FF/U/V) showed the largest observed treatment effects in favour of FF/U/V (Fig. 2). For the time-to-first-exacerbation model,

this was a mean increase of 43.69 days until first exacerbation if patients received FF/U/V compared to if they would have received U/V therapy. For the exacerbation rate model, this was a mean reduction of 0.80 exacerbations per year. For both outcomes, the observed effects gradually decreased towards quintile 5 (patients estimated to benefit least from FF/U/V). Patient rankings for both outcomes overlapped most in quintiles 1 and 5 (Fig. 3). Patients in quintile 1 for both outcomes had high blood eosinophil counts, high SGRQ symptoms scores (more severe respiratory symptoms), high FEV1 reversibility and lower FEV1 percent predicted values (worse lung function) compared to patients in quintile 5 for both outcomes (Table 2).

Table 2 Characteristics of patients selected to benefit most or least for both modelled outcomes

Characteristic	Level	FF/U/V versus U/V		FF/V versus U/V	
		Q1Q1 Benefits most from FF/U/V (n = 950)	**Q5Q5 Benefits least from FF/U/V (n = 902)**	**Q1Q1 Benefits most from FF/V (n = 996)**	**Q5Q5 Benefits least from FF/V (n = 797)**
Subject Level Analysis					
Age (years)		65.90 (8.09)	63.70 (8.43)	64.67 (8.33)	65.53 (8.06)
Sex	Male	709 (74.63)	636 (70.51)	715 (71.79)	663 (83.19)
	Female	241 (25.37)	266 (29.49)	281 (28.21)	134 (16.81)
Height (cm)		168.61 (9.60)	168.53 (7.33)	168.71 (9.73)	170.03 (6.32)
Weight (kg)		75.16 (19.08)	78.45 (17.63)	76.84 (19.38)	72.34 (16.76)
GOLD grade	1	5 (0.53)	7 (0.78)	3 (0.30)	1 (0.13)
	2	361 (38.00)	568 (62.97)	436 (43.78)	238 (29.86)
	3	457 (48.11)	298 (33.04)	436 (43.78)	399 (50.06)
	4	127 (13.37)	29 (3.22)	121 (12.15)	159 (19.95)
Previous exacerbation count		1.92 (1.11)	1.73 (0.60)	1.88 (1.00)	1.54 (0.77)
Smoking status	Former	643 (67.68)	594 (65.85)	677 (67.97)	547 (68.63)
	Current	307 (32.32)	308 (34.15)	319 (32.03)	250 (31.37)
Substance Use Analysis					
Number of pack years		46.04 (26.14)	43.49 (25.99)	45.52 (27.43)	51.79 (27.93)
Laboratory Analysis					
Eosinophils (10^9/L)		0.48 [0.38,0.64]	0.09 [0.06,0.14]	0.46 [0.37,0.62]	0.10 [0.06,0.14]
Neutrophils (10^9/L)		4.71 [3.75,5.91]	4.74 [3.80,6.17]	4.82 [3.77,6.04]	4.76 [3.69,6.16]

Pulmonary Function Test Analysis				
FEV1 (L)	1.20 (0.52)	1.54 (0.42)	1.27 (0.52)	1.22 (0.52)
FEV1 reversibility (mL)	147.81 (156.76)	68.67 (129.47)	152.85 (175.15)	64.82 (85.55)
FVC (L)	2.60 (0.85)	2.97 (0.70)	2.68 (0.86)	2.77 (0.78)
FEV1pp (%)	41.61 (14.78)	53.70 (13.65)	43.61 (14.90)	41.32 (16.27)
SGRQ Analysis				
SGRQ Activity score	65.63 (19.33)	53.56 (18.65)	66.98 (19.32)	52.85 (20.56)
SGRQ Impacts score	39.00 (18.37)	25.92 (16.01)	41.01 (19.09)	24.03 (14.77)
SGRQ Symptoms score	70.25 (17.95)	51.78 (16.84)	72.09 (17.14)	52.15 (16.44)

Data are presented as mean (standard deviation), median [interquartile range], and number (%). *Q1Q1* group of patients estimated to be in quintile 1 for both modelled outcomes, *Q5Q5* group of patients estimated to be in quintile 5 for both modelled outcomes, *FF* Fluticasone Furoate, *U* Umeclidinium, *V* Vilanterol, *GOLD* Global Initiative for Chronic Obstructive Lung Disease, *FEV1* Forced Expiratory Volume in 1 second, *FEV1pp* Forced Expiratory Volume in 1 second percent predicted, *FVC* Forced Vital Capacity, *SGRQ* St. George's Respiratory Questionnaire, *n* number of patients.

In the FF/V versus U/V comparison, largest treatment effects in favour of FF/V were observed in quintiles 1 and 2 (patients estimated to benefit most from FF/V, Fig. 2). No clear benefit of FF/V over U/V was observed in other quintiles. For the time-to-first-exacerbation model, the observed benefit in quintile 4 and 5 (patients estimated to benefit least from FF/V) is even significantly in favour of U/V. Overlap in patient rankings was strongest in quintile 1 and 5 (Fig. 3). Patients in quintile 1 for both outcomes had high blood eosinophil counts, high SGRQ symptoms scores, high FEV1 reversibility and a reduced lung function, while patients in quintile 5 for both outcomes had low blood eosinophil counts, low FEV1 reversibility, low SGRQ symptoms scores and a reduced lung

function (Table 2). More detailed performance evaluations and patient characteristics per quintile are provided in Supplementary File S1.

Discussion

Previous analyses have shown that high blood eosinophil counts can be used as a biomarker to predict an enhanced effect of ICS treatment [6-11], but additional insight in treatment effect heterogeneity is needed to optimize clinical disease management. This analysis used causal ML models to investigate heterogeneity in the effect of ICS on COPD exacerbations. Causal ML models are very flexible models that may be uniquely suited for a data-driven investigation of the complex disease dynamics that underly the heterogeneous effect of COPD treatments. Our ML models ranked patients well according to their estimated individualized treatment effect and detected significant heterogeneity in the effect of ICS on both time to first COPD exacerbation and COPD exacerbation rate. Blood eosinophil counts were selected as the most important predictor of treatment effect by all models. Other characteristics that consistently ranked high in importance according to all models were respiratory symptoms, reversibility of airflow limitation and FEV1 percent predicted.

The analysis first focused on investigating heterogeneity in the effect of ICS on top of LABA/LAMA by modelling the FF/U/V arm versus the U/V arm. For both exacerbation outcomes, blood eosinophils counts were selected as most important predictor of treatment effect. Highest eosinophil counts were observed in quintile 1 (patients estimated to benefit most from FF/U/V) and gradually decreased towards quintile 5

(patients estimated to benefit least from FF/U/V). This was in line with the continuous relationship between blood eosinophil counts and ICS effect observed in previous literature [9-10] and the existing hypothesis that patients with eosinophilic inflammation (identified by high blood eosinophil counts) have the highest benefit of ICS treatment [1, 2, 4].

The models also identified other patient characteristics – SGRQ symptoms score, FEV1 reversibility and FEV1 percent predicted – as important features that contribute to the detection of treatment effect heterogeneity. Patients estimated to benefit most from ICS had high SGRQ symptoms scores (more severe respiratory symptoms), high FEV1 reversibility, and reduced lung function values. FEV1 reversibility has been linked to an enhanced ICS effect in a previous subgroup analysis [8], but another subgroup analysis did not confirm this finding [7]. Higher SGRQ symptoms scores and worsening lung function have been linked to increased risk of exacerbations [23, 24, 27].

Previous analyses have associated higher exacerbation counts in the previous year with higher exacerbation risk [23, 24, 26]. Model-based feature importance (Fig. 1) did not indicate the count of exacerbations in the previous year as an important predictor of treatment effect, but higher counts were observed for those patients who were estimated to respond best to ICS (Table 2). IMPACT only included patients who had at least one exacerbation in the previous year. Possibly, in patients with a history of exacerbations and for this specific forest-based model structure, the count of previous exacerbations was less informative compared to other characteristics.

The analysis was repeated for a comparison of the FF/V and U/V arms. In this case, the trade-off between the ICS/LABA in the FF/V arm and the LAMA/LABA in the U/V arm emphasizes traits that predict response to adding ICS versus maximizing bronchodilation. Overall, patients estimated to benefit most from FF/V therapy (quintile 1) showed high blood eosinophil counts, high SGRQ symptoms scores and high FEV1 reversibility. These values fit with a patient profile of eosinophilic inflammation as responder to ICS that was also present in the FF/U/V versus U/V treatment comparison. Patients estimated to benefit least from FF/V therapy (quintile 5) showed low blood eosinophil counts and low FEV1 reversibility, higher SGRQ symptoms scores, a reduced lung function and higher Global Initiative for Chronic Obstructive Lung Disease (GOLD) grades. Interestingly, the model results in Figs. 2 and 3 suggest that in this modelled trade-off between ICS/LABA and LAMA/LABA, patients in quintile 5 do not benefit from ICS/LABA but clearly benefit from LAMA/LABA therapy. A possible explanation is that, in patients who present with reduced lung function but no eosinophilic inflammation, the effect of the LAMA on top of LABA therapy is more beneficial for preventing future exacerbations than adding ICS.

Causal ML models are very suited for investigating complex heterogeneous treatment effects due to their flexibility, but this flexibility also increases the risk of overfitting. This work used a Monte-Carlo cross-validation approach with 100 train and internal validation sets and combined results from two related exacerbation outcomes and two treatment comparisons to increase confidence and extract robust patterns. The causal survival forest was developed specifically for the estimation of individualized treatment effect for right-

censored survival outcomes and is well-suited to model effects on time to first exacerbation. Of the two model types in this analysis, the novel causal survival forest may be the preferred model to use when investigating exacerbations and focusing on a single model because of this suitability. A limit of the causal forest is that it cannot take differences in follow-up time into account when modelling exacerbation rate. Despite this limit, the causal forest succeeded well in ranking patients and rankings largely overlapped with the time to first exacerbation ranking. The similarity in feature importances and patient rankings for both exacerbation outcomes increases confidence in results.

Insights from this analysis are hypothesis generating. Further evidence from prospective studies should be gathered before additions to treatment allocation guidelines are made. In the long run, a well-validated causal ML model could be turned into a tool that supports treatment decisions by estimating a clear and nuanced treatment effect for each individual patient. In addition to time to first exacerbation and exacerbation rate, a model for ICS treatment of patient with COPD could also include other outcomes, like trough FEV1 and changes in SGRQ, and adverse events like pneumonia, which is a notable concern for ICS treatment [14]. The findings in this analysis may be very specific to the inclusion criteria of the IMPACT trial. The development of causal ML strategies that combine data from randomized controlled trials with observational data holds potential to further improve models and enhance the generalizability of conclusions [31].

Conclusions

In data from the large IMPACT randomized trial, we demonstrated a causal machine learning approach that used a selection of baseline characteristics to estimate the effect of ICS treatment on COPD exacerbations for each individual patient. We showed the advantage of combining a novel causal survival forest to model individualized effect on time to first exacerbation with a causal forest for exacerbation rate to identify a robust profile of ICS response. We investigated both exacerbation outcomes in a comparison of two different ICS-containing therapies versus a therapy of only bronchodilators. Models ranked patients well according to their treatment effect and detected significant heterogeneity in the effect of ICS. All models consistently identified blood eosinophil counts as the most important predictor of treatment effect, in conjunction with other baseline features including respiratory symptoms, reversibility of airflow limitation and lung function. This analysis illustrates how causal machine learning models become robust models to identify treatment effect predictors and patterns of response, to enhance insight in complex disease phenotypes and heterogeneous treatment effects and to contribute to a much more personalized allocation of the investigated treatments.

List of abbreviations

AUTOC	Area under the targeting operator characteristic
CI	Confidence interval
COPD	Chronic obstructive pulmonary disease
FEV1	Forced expiratory volume in 1 second
FEV1pp	Forced expiratory volume in 1 second percent predicted
FF	Fluticasone furoate
FVC	Forced vital capacity
GOLD	Global Initiative for Chronic Obstructive Lung Disease
HR	Hazard ratio
ICS	Inhaled corticosteroids
IMPACT	InforMing the PAthway of COPD Treatment
ITE	Individualized treatment effect
LABA	Long-acting β_2-agonist
LAMA	Long-acting muscarinic antagonist
ML	Machine learning
Q	Quintile
RMST	Restricted mean survival time
RR	Rate ratio
SGRQ	St. George's Respiratory Questionnaire
TOC	Targeting operator characteristic
U	Umeclidinium
V	Vilanterol

Declarations

Ethics approval and consent to participate

This study was approved by the KU Leuven Institutional Privacy and Ethics review committee (SMEC/PRET, ID: G-2022-5402). The requirement for informed consent was waived due to the retrospective nature of the study design.

Consent for publication

Not applicable

Availability of data and materials

The anonymized individual participant data of IMPACT (Sponsor ID CTT116855) used in this study was provided upon request by www.ClinicalStudyDataRequest.com with the approval of GlaxoSmithKline Research & Development Ltd. The data are available from www.ClinicalStudyDataRequest.com upon request.

Competing interests

HH is funded by a Strategic Basic Research Fellowship from the Research Foundation – Flanders, Belgium (FWO) and received research funding from Chiesi, KV received travel grants from the Research Foundation – Flanders, Belgium (FWO), IG was funded by a Fundamental Research Fellowship from the Research Foundation – Flanders, Belgium (FWO), received payment for a presentation from Chiesi and a travel grant from Astra Zeneca. WJ received grants or contracts from GSK and Chiesi, consulting fees from GSK, Sanofi, Roche and Astra Zeneca, payment or honoraria for lectures, presentations,

speakers bureaus, manuscript writing or educational events from GSK, Sanofi, Chiesi and Astra Zeneca. MDV received a research grant from the Research Foundation – Flanders, Belgium (FWO) and research funding from Chiesi. The other authors have nothing to disclose.

Funding

This research work was funded by a Strategic Basic Research Fellowship from the Research Foundation – Flanders, Belgium (FWO) for HH (FWO project number 1S30225N); by a research grant from the Research Foundation – Flanders, Belgium (FWO) for MDV and WJ (Research Project: "Artificial Intelligence (AI) for data-driven personalised medicine", FWO project number G0C9623N); by the Horizon Europe 2022 Research and Innovation Program, HORIZON-HLTH-2022-IND-13: "Privacy compliant health data as a service for AI development (PHASE IV AI)", funded by the European Union, under Grant Agreement #101095384; by the KU Leuven Special Research Fund (BOF): "Digital technologies and machine learning models for diagnosis and treatment of chronic respiratory diseases" (BOF project number FKO/24/001); and received funding from the Flemish Government (AI Research Program) and the KU Leuven Chiesi Chair on Phenotyping and individualized therapy in COPD. MDV is affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium.

Authors' contributions

HH, KV, JB, WJ and MDV contributed to the conceptualization and design of this work. HH, KV, TS, IG, MS, AB, WJ and MDV acquired, analysed and/or interpreted the data. HH

drafted the manuscript. All authors critically revised the manuscript for important intellectual content and approved the final version.

Acknowledgements

Not applicable

Authors' information

None

Prior abstract publication/presentation

This work was previously presented as two abstracts at the ERS Congress 2025. Abstracts are due to be published as an online supplement of the European Respiratory Journal but were not published yet at time of submission (3 November 2025).

- Huts H, Verstraete K, Strypsteen S, et al. Investigating individual effects of ICS on COPD exacerbations with causal machine learning. Presented at the European Respiratory Society (ERS) Congress 2025, Amsterdam, 27 Sep 2025 – 01 Oct 2025.
- Huts H, Verstraete K, Strypsteen S, et al. Causal machine learning to investigate heterogeneity in time to first COPD exacerbation for ICS/LABA versus LAMA/LABA. Presented at the European Respiratory Society (ERS) Congress 2025, Amsterdam, 27 Sep 2025 – 01 Oct 2025.

References

1. Agustí A, Bafadhel M, Beasley R, et al. Precision medicine in airway diseases: moving to clinical practice. Eur Respir J. 2017;50(4):1701655. doi: 10.1183/13993003.01655-2017.
2. Agusti A, Bel E, Thomas M, et al. Treatable traits: toward precision medicine of chronic airway diseases. Eur Respir J. 2016;47(2):410–419. doi: 10.1183/13993003.01359-2015.
3. Barnes PJ. Inflammatory endotypes in COPD, Allergy. 2019;74(7):1249–1256. doi: 10.1111/all.13760.
4. David B, Bafadhel M, Koenderman L, De Soyza A. Eosinophilic inflammation in COPD: from an inflammatory marker to a treatable trait. Thorax. 2021;76(2):188–195. doi: 10.1136/thoraxjnl-2020-215167.
5. Ferguson GR, Rabe KF, Martinez FJ, et al. Triple therapy with budesonide/glycopyrrolate/formoterol fumarate with co-suspension delivery technology versus dual therapies in chronic obstructive pulmonary disease (KRONOS): a double-blind, parallel-group, multicentre, phase 3 randomised controlled trial. Lancet Respir Med. 2018;6(10):747–758. doi: 10.1016/S2213-2600(18)30327-8.
6. Lipson DA, Barnhart F, Brealey N, et al. Once-daily single-inhaler triple versus dual therapy in patients with COPD. N Engl J Med. 2018;378(18):1671–1680. doi: 10.1056/NEJMoa1713901.
7. Papi A, Vestbo J, Fabbri L, et al. Extrafine inhaled triple therapy versus dual bronchodilator therapy in chronic obstructive pulmonary disease (TRIBUTE): a

double-blind, parallel group, randomised controlled trial. Lancet. 2018;391(10125):1076–1084. doi: 10.1016/S0140-6736(18)30206-X.

8. Rabe KF, Martinez FJ, Ferguson GT, et al. Triple Inhaled Therapy at Two Glucocorticoid Doses in Moderate-to-Very-Severe COPD. N Engl J Med. 2020;383(1):35–48. doi: 10.1056/NEJMoa1916046.
9. Bafadhel M, Peterson S, De Blas MA, et al. Predictors of exacerbation risk and response to budesonide in patients with chronic obstructive pulmonary disease: a post-hoc analysis of three randomised trials. Lancet Respir Med. 2018;6(2):117–126. doi: 10.1016/S2213-2600(18)30006-7.
10. Pascoe S, Barnes N, Brusselle G, et al. Blood eosinophils and treatment response with triple and dual combination therapy in chronic obstructive pulmonary disease: analysis of the IMPACT trial. Lancet Respir Med. 2019;7(9):745–756. doi: 10.1016/S2213-2600(19)30190-0.
11. Pascoe S, Locantore N, Dransfield MT, Barnes NC, Pavord ID. Blood eosinophil counts, exacerbations, and response to the addition of inhaled fluticasone furoate to vilanterol in patients with chronic obstructive pulmonary disease: a secondary analysis of data from two parallel randomised controlled trials. Lancet Respir Med. 2015;3(6):435–442. doi: 10.1016/S2213-2600(15)00106-X.
12. GOLD: 2025 Global Strategy for Prevention, Diagnosis and Management of COPD. www.goldcopd.org. Accessed June 22 2025.
13. Kolsum U, Southworth T, Jackson N, Singh D. Blood eosinophil counts in COPD patients compared to controls. Eur Respir J. 2019;54(4):1900633. doi: 10.1183/13993003.00633-2019.

14. Heerfordt CK, Eklöf J, Sivapalan P, et al. Inhaled Corticosteroids in Patients with Chronic Obstructive Pulmonary Disease and Risk of Acquiring Streptococcus Pneumoniae Infection. A Multiregional Epidemiological Study. Int J Chron Obstruct Pulmon Dis. 2023;18:373–384. doi: 10.2147/COPD.S386518.
15. Feuerriegel S, Frauen D, Melnychuk V, et al. Causal machine learning for predicting treatment outcomes. Nat Med. 2024;30(4):958–968. doi: 10.1038/s41591-024-02902-1.
16. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688-701. doi: 10.1037/h0037350.
17. Verstraete K, Gyselinck I, Huts H, et al. Estimating individual treatment effects on COPD exacerbations by causal machine learning on randomised controlled trials. Thorax. 2023;78(10):983–989. doi: 10.1136/thorax-2022-219382.
18. Gyselinck I, Huts H, De Vos M, Janssens W. Time T-to understand treatment response. Lancet Respir Med. 2025;13(6):482-484. doi: 10.1016/S2213-2600(25)00118-3.
19. Cui Y, Kosorok MR, Sverdrup E, Wager S, Zhu R. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2023;85(2):179–211. doi: 10.1093/jrsssb/qkac001.
20. Kent DM, Paulus JK, van Klaveren D, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Ann Intern med. 2020;172(1):35–45. doi: 10.7326/M18-3667.

21. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi: 10.1136/bmj-2023-078378.
22. Bhowmik A, Seemungal TA, Sapsford RJ, Wedzicha JA. Relation of sputum inflammatory markers to symptoms and lung function changes in COPD exacerbations. Thorax. 2000;55(2):114–120. doi: 10.1136/thorax.55.2.114.
23. Hurst JR, Vestbo J, Anzueto A, et al. Susceptibility to exacerbation in chronic obstructive pulmonary disease. N Engl J Med. 2010;363(12):1128–1138. doi: 10.1056/NEJMoa0909883.
24. Jenkins CR, Celli B, Anderson JA, et al. Seasonality and determinants of moderate and severe COPD exacerbations in the TORCH study. Eur Respir J. 2012;39(1):38-45. doi: 10.1183/09031936.00194610.
25. Lonergan M, Dicker AJ, Crichton ML, et al. Blood neutrophil counts are associated with exacerbation frequency and mortality in COPD. Respir Res. 2020;21(1):166. doi: 10.1186/s12931-020-01436-7.
26. Make BJ, Eriksson G, Calverley PM, et al. A score to predict short-term risk of COPD exacerbations (SCOPEX). Int J Chron Obstruct Pulmon Dis. 2015;10:201–209. doi: 10.2147/COPD.S69589.
27. Seemungal TA, Donaldson GC, Paul EA, Bestall JC, Jeffries DJ, Wedzicha JA. Effect of exacerbation on quality of life in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 1998;157(5 Pt 1):1418–1422. doi: 10.1164/ajrccm.157.5.9709032.

28. Xu Q-S, Liang Y-Z. Monte Carlo cross validation. Chemometrics and Intelligent Laboratory Systems. 2001;56(1):1–11. doi: 10.1016/S0169-7439(00)00122-2.
29. Athey S, Tibshirani J, Wager S. Generalized random forests. The Annals of Statistics. 2019;47(2):1148–1178. doi: 10.1214/18-AOS1709.
30. Yadlowsky S, Fleming S, Shah N, Brunskill E, Wager S. Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects. J Am Stat Assoc. 2024;120(549):38–51. doi: 10.1080/01621459.2024.2393466.
31. Colnet B, Mayer I, Chen G, et al. Causal Inference Methods for Combining Randomized Trials and Observational Studies: A Review. Statistical science. 2024;39(1):165–191. doi: 10.1214/23-STS889.

Supplementary Information

File name: Supplementary File S1

File format: Portable document format (pdf) .pdf

Title of the data: Supplementary File S1

Description of the data: A structured pdf file with additional details on the methodology used in this study, and tables and figures with additional results.

Figure legends

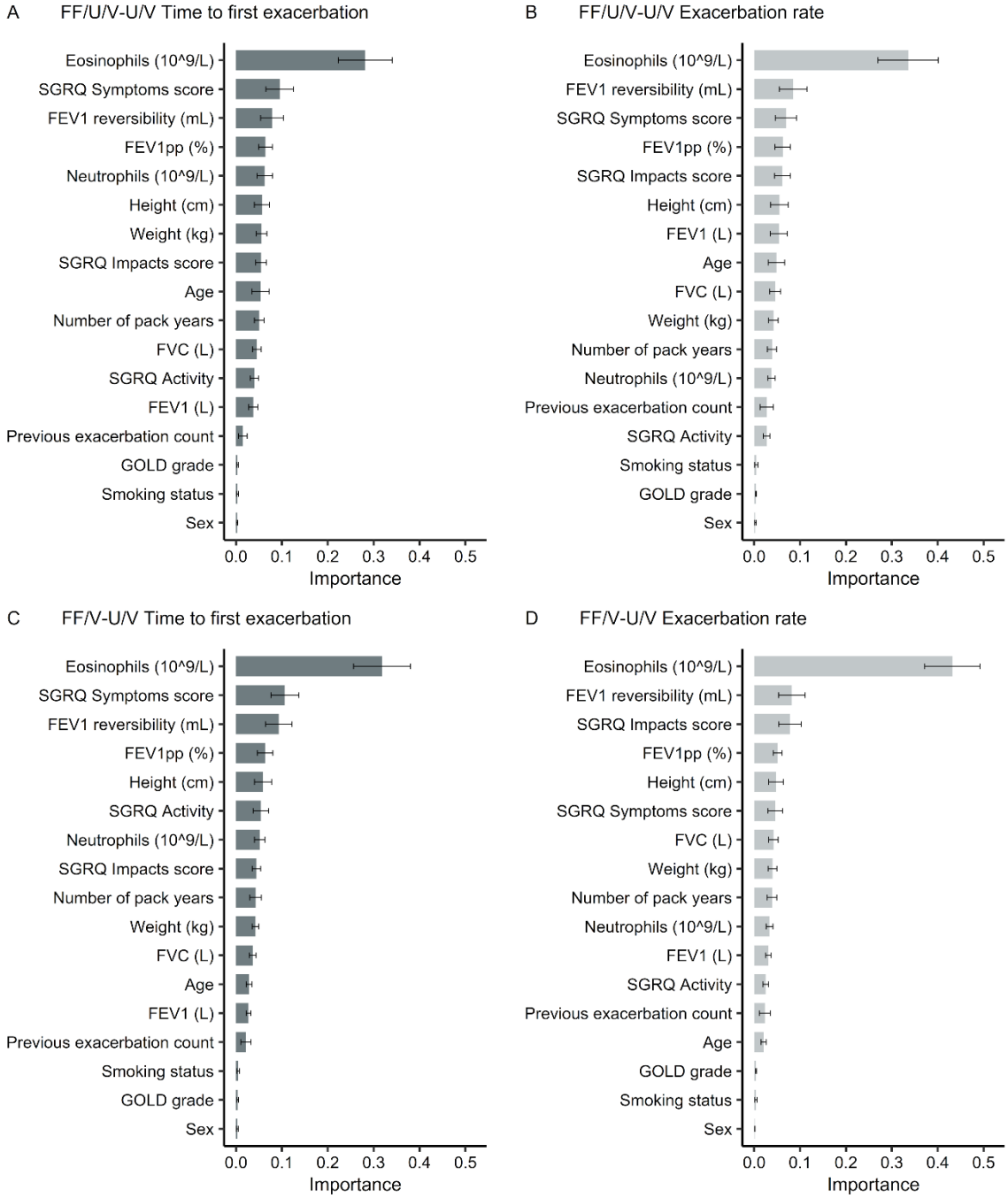


Fig. 1 Model-based feature importance.

Importance is plotted on a relative scale: for each model, the sum of all importances is equal to 1. In **A** and **B**, a triple therapy of fluticasone furoate (FF), umeclidinium (U) and vilanterol (V) was compared against a dual therapy of U/V. **C** and **D** compared dual

585 therapy of FF/V against U/V. For **A** and **C**, models estimated individualized treatment
586 effect for time to first exacerbation. For **B** and **D**, models estimated individualized
587 treatment effect for exacerbation rate. Error bars indicate the standard deviation over 100
588 training sets. *SGRQ* St. George's Respiratory Questionnaire, *FEV1* forced expiratory volume
589 in 1 second, *FEV1pp* forced expiratory volume in 1 second percent predicted, *FVC* forced
590 vital capacity, *GOLD* Global Initiative for Chronic Obstructive Lung Disease
591

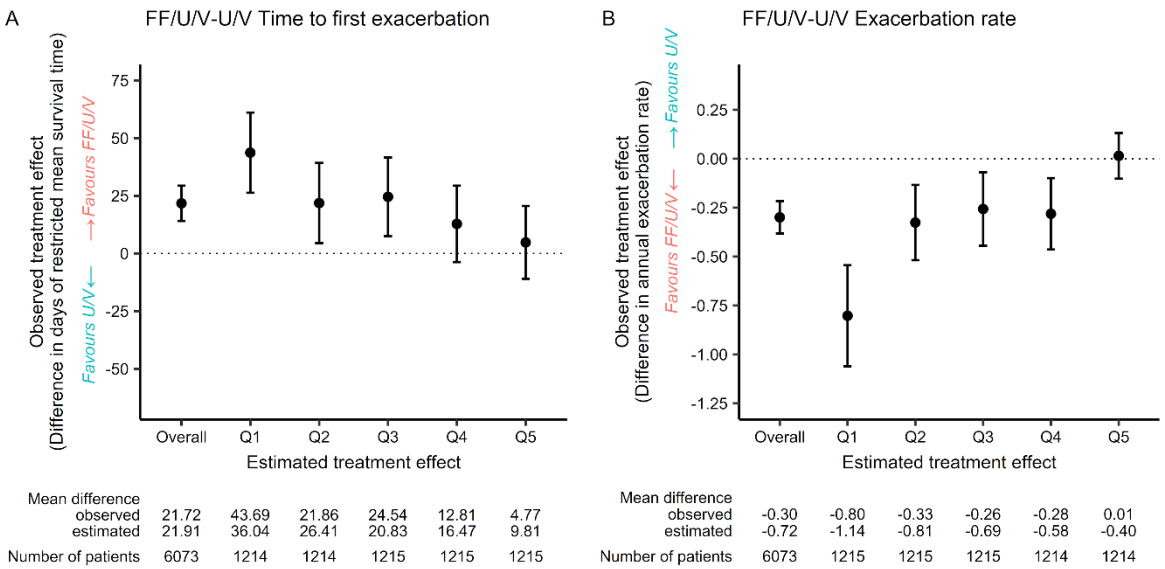
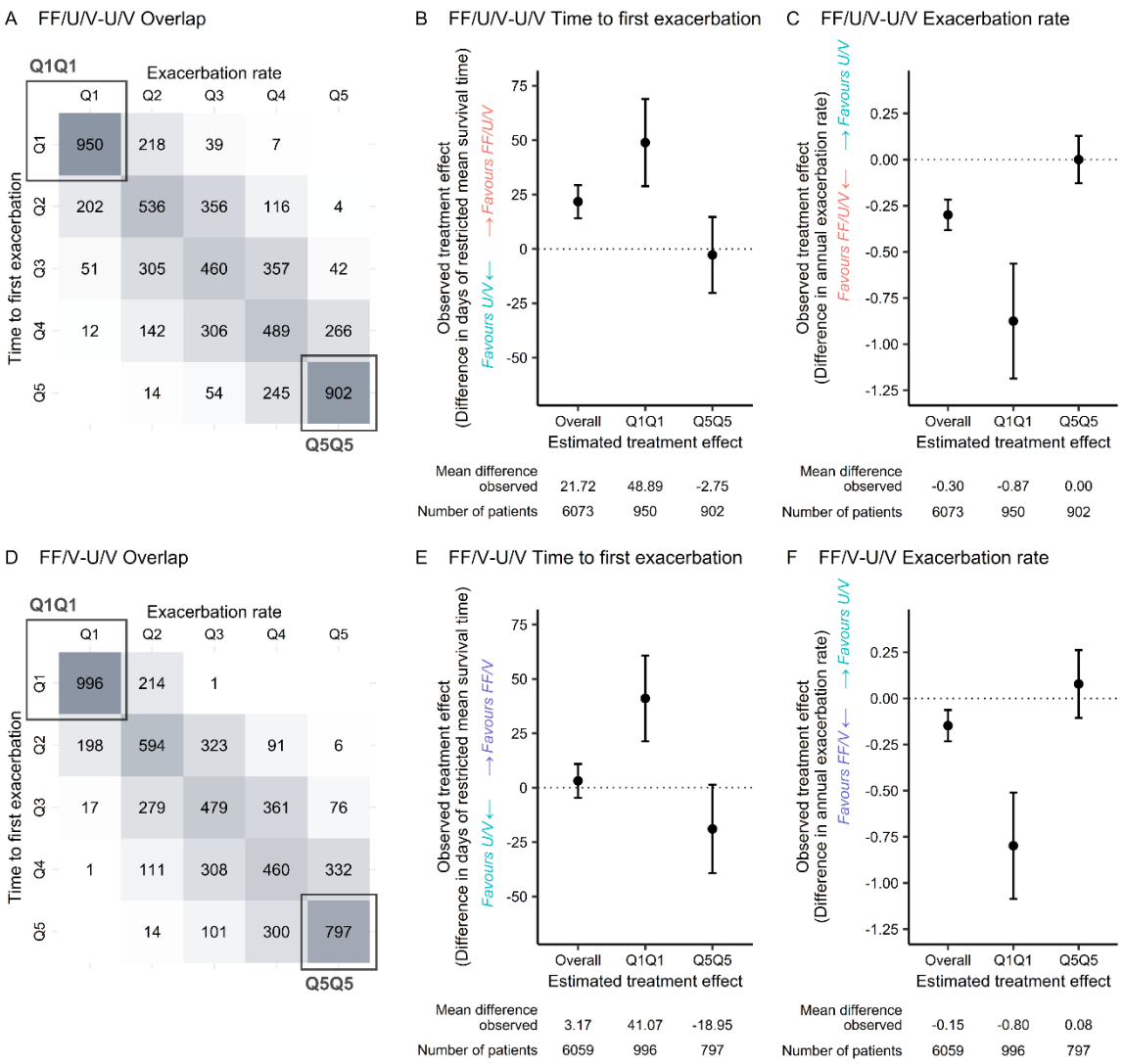


Fig. 2 Observed treatment effects per quintile of estimated treatment effect.

In **A** and **B**, a triple therapy of fluticasone furoate (FF), umeclidinium (U) and vilanterol (V) was compared against dual therapy of U/V. **C** and **D** compared dual therapy of FF/V against U/V. Patients were ranked according to their estimated individualized treatment effect (ITE) and split into quintiles. Quintile 1 (Q1) contains patients estimated to benefit most from an inhaled corticosteroid-containing therapy (either FF/U/V or FF/V), while patients in quintile 5 (Q5) were estimated to benefit least from this therapy. For **A** and **C**, models estimated ITE for time to first exacerbation. Observed treatment effects are

601 expressed as a difference in restricted mean survival time. For **B** and **D**, models estimated
602 ITE for exacerbation rate. Observed treatment effects are expressed as a difference in
603 mean annual exacerbation rate between the inhaled corticosteroid-containing therapy
604 and U/V therapy. Error bars indicate 95% confidence intervals.
605



606

607 **Fig. 3 Overlap in patient rankings for exacerbation rate and time to first exacerbation.**

608 In **A, B** and **C** a triple therapy of fluticasone furoate (FF), umeclidinium (U) and vilanterol

609 (V) was compared against a dual therapy of U/V. **D, E** and **F** compared dual therapy of

610 FF/V against U/V. Patients were ranked according to their estimated individual treatment

611 effect and split into five quintiles. **A** and **D** illustrate overlap in patient rankings for both

612 exacerbation outcomes with a confusion matrix. **B, C, E** and **F** visualize observed

613 treatment effect for patients estimated to benefit most from inhaled corticosteroid-

614 containing therapy (either FF/U/V or FF/V) according to both rankings (Q1Q1), and for

615 patients estimated to benefit least from this therapy according to both rankings (Q5Q5).

Observed treatment effects are expressed as the difference in restricted mean survival time and the difference in mean annual exacerbation rate between the inhaled corticosteroid-containing therapy and U/V therapy. Error bars indicate 95% confidence intervals.

Uncovering the heterogeneous effect of inhaled corticosteroids on COPD exacerbations with causal machine learning

Helene Huts, Kenneth Verstraete, Thomas Strypsteen, Iwein Gyselinck, Michaël Staes, Amber Beersaerts, Jeroen Berrevoets, Wim Janssens*, Maarten De Vos*

* Co-senior authors

Supplemental Material

(1758 words, 7 references, 8 tables, 6 figures)

Supplementary Methods

1. **Study Participants, Outcomes and Baseline Characteristics**
2. **Model Development**
    - **2.1. Causal machine learning models**
    - **2.2. Monte-Carlo cross validation**
3. **Model Performance Evaluation**
    - **3.1. Area Under the Targeting Operator Characteristic (AUTOC)**
    - **3.2. Quintile visualizations**
    - **3.3. Data description and statistical analysis**

Supplementary Results

1. **Model Performance Evaluation**

Supplementary Tables

1. **Table S1:** Patient baseline characteristics selected as initial model input features
2. **Table S2:** Extended model performance
3. **Table S3:** Intervals of estimated individualized treatment effects
4. **Table S4:** Patient characteristics per quintile for a ranking made by the FF/U/V versus U/V time to first exacerbation model
5. **Table S5:** Patient characteristics per quintile for a ranking made by the FF/U/V versus U/V exacerbation rate model
6. **Table S6:** Patient characteristics per quintile for a ranking made by the FF/V versus U/V time to first exacerbation model

7. **Table S7:** Patient characteristics per quintile for a ranking made by the FF/V versus U/ exacerbation rate model
8. **Table S8:** Characteristic of patients selected to benefit most or least for both modelled outcomes, using a larger range of interest

Supplementary Figures

1. **Figure S1:** Monte-Carlo cross-validation
2. **Figure S2:** Area under the Targeting Operator Characteristic (AUTOC) curves for rankings based on estimated individualized treatment effect
3. **Figure S3:** Observed rate and hazard ratios per quintile of estimated treatment effect
4. **Figure S4:** Observed mean exacerbation rates and restricted mean survival times per quintile of estimated treatment effect
5. **Figure S5:** Observed cumulative incidence curves per quintile of estimated treatment effect
6. **Figure S6:** Overlap in patient rankings for exacerbation rate and time to first exacerbation, using a larger range of interest.

Supplementary Methods

Study Participants, Outcomes and Baseline Characteristics

The analysis in this work used data from the InforMing the PAthway of COPD Treatment (IMPACT) trial[1]. Based on selection criteria, an initial set of 17 patient characteristics, measured during the trial at baseline, was selected as input features for the causal machine learning models in the analysis. Completeness of baseline characteristics was very high. Of the 10,355 patients in the intention-to-treat population, 10,107 patients for whom all characteristics were available were included in the analysis. **Table S1** provides an overview of selected characteristics and their completeness. Both the causal forest and causal survival forest that will be used in this analysis do not require feature normalization but do require numerical features. Features were therefore not normalized but binary features were recoded to numerical values:

- Smoking status
    - "Former Smoker": 0
    - "Current Smoker": 1
- Sex
    - "M": 0
    - "F": 1

Model Development

<u>Causal machine learning models</u>

This analysis investigated the fluticasone furoate/umeclidinium/vilanterol (FF/U/V) arm versus the U/V arm, and the FF/U arm versus the U/V arm. Individualized treatment effect was estimated with a causal survival forest[2] for time to first exacerbation. The horizon

parameter for the causal survival forest was set to 365 days, in line with the follow-up time of the IMPACT trial. All other model parameters were set to default. The outcome variable used to train this model was the time to first moderate or severe COPD exacerbation (dataset code: AVAL TTFMSVOT). The causal survival forest is well-suited for working with survival outcomes. During modelling it also takes censoring of the outcome (dataset code: CNSR TTFMSVOT) into account.

Individualized treatment effect for exacerbation rate was estimated with a causal forest[3]. All model parameters were set to default. The outcome variable used to train this model was the annual rate of moderate or severe COPD exacerbations (dataset code: RTMDSVOT). Values for exacerbation rate in RTMDSVOT were divided by 1000 to obtain the correct scale of annual exacerbation rate. A limit of the causal forest is that it cannot take differences in follow-up time into account when modelling exacerbation rate. To reduce influence of unrealistically high apparent annual rates caused by very short follow-up times, apparent annual rates higher than 15 were set to 15 for 125 patients in the analysis.

For both outcomes, the values measured at 52 weeks (end of follow-up of the IMPACT trial) were used for training. The R implementation from the *grf* package was used for both causal models.

Monte-Carlo cross-validation

This analysis used a Monte-Carlo cross-validation[4] approach to optimally use all available data**. Figure S1** illustrates this approach. This figure was inspired by a similar figure in the supplementary material of work by Goligher et al[5]. Patients in the FF/U/V and U/V arms of

the IMPACT trial were split in a 70% training and 30% internal validation set. This split was stratified to ensure that the ratio of FF/U/V to U/V patients and the observed rate and hazard ratios were similar for the training and the internal validation set. The split was repeated 100 times. Each time patients for training and internal validation set were randomly selected from the full dataset. Each patient was part of approximately 30 internal validation sets. For each patient, all internal validation set predictions were averaged to get a final estimate of their individualized treatment effect. These final estimates were used for all quintile visualizations of model results. The same methodology was repeated for FF/V versus U/V.

Model Performance Evaluation

Area under the targeting operator characteristic (AUTOC)

A detailed description of the area under the targeting operator characteristic (AUTOC) has been published previously[6]. Briefly, the targeting operator characteristic (TOC) is a curve that is constructed by ranking patients according to a prioritization rule – in this analysis their individual treatment effect as estimated by a causal machine learning model. The curve quantifies treatment effect gain relative to the average treatment effect for increasing top fractions of ranked patients. Because the TOC curve depends on the ability of the model to correctly rank patients according to their individual treatment effect, the area under the TOC curve is a measure of discrimination. A larger area under the curve indicates better discrimination. Confidence intervals for the TOC curve can be acquired through a half-sample bootstrapping procedure. Heterogeneity was detected by the model if the area under the TOC (AUTOC) curve is significantly different from 0. The analysis in this work used the R implementation of the AUTOC metric that is part of the R *grf* package.

Performance of the final models was evaluated by calculating an average AUTOC for individualized treatment effect (ITE) estimations in the 100 internal validation sets. Because of the cross-validation structure of training and internal validation sets, hypothesis tests for detecting heterogeneity were not independent. Instead of a mean value, a median, two-sided (H_0: $\mathrm{AUTOC} = 0$, H_A: $\mathrm{AUTOC} \neq 0$) p-value was reported to evaluate the presence of heterogeneity. This is a strategy for cross-validation p-values suggested in the documentation of the *grf* package and by Chernozhukov et al[7].

From all ITE estimations on the 100 internal validation sets, an average estimated ITE was calculated for each patient in the dataset. For each outcome and each treatment comparison, all patients were ranked according to their average estimated ITE. An AUTOC was calculated for this final ranking. This procedure is similar to out-of-bag estimation on the full dataset. A one-sided p-value (H_A: $\mathrm{AUTOC} > 0$) was reported, as suggested on the documentation of the *grf* package.

Quintile visualizations

Heterogeneity in treatment effect was visualized by splitting each ranking of all patients into five groups (quintiles). Quintile 1 contained the 20% of patients (from both arms) that was estimated to benefit most from the treatment that contained inhaled corticosteroids (ICS), while quintile 5 contained the 20% of patient estimated to benefit least from ICS-containing treatment. This illustration could have used any number of groups, but five groups were chosen to not make visualizations overly complex.

Hazard ratios were calculated using a Cox proportional hazards model with covariates for treatment group, sex, exacerbation history, smoking status at screening, geographical region and post-bronchodilator forced expiratory volume in 1 second (FEV1) percent predicted at screening. For the evaluation of calibration, the observed average treatment effect per quintile was calculated as a difference in restricted mean survival time for treatment and control arm, with covariates for sex, exacerbation history, smoking status at screening, geographical region and post-bronchodilator FEV1 percent predicted at screening. The truncation time point was set to 365 days, which corresponds with the follow-up time of the trial.

Mean exacerbation rates and rate ratios were calculated using a negative binomial generalized linear model with the logarithm of the on-treatment at risk duration as offset and covariates for treatment group, sex, smoking status at screening, geographical region, and post-bronchodilator forced expiratory volume in 1 second percent predicted at screening.

Data Description and Statistical Analysis

Most of the calculations in this analysis were performed with R version 4.4.2. The *grf* package was used for the causal survival forest, causal forest and to calculate the area under the targeting operator characteristic. The *MASS* (version 7.3-61) and *emmeans* (version 1.10.5) packages were used to calculate negative binomial generalized linear models, mean exacerbation rates, rate ratios and p-values. Hazard ratios and p-values from log-rank tests were calculated with the *survival* package (version 3.7-0), and restricted mean survival times were calculated with the *survRM2* package (version 1.0-4). Summary statistics of patient

characteristics were obtained with Python version 3.10 and the Python package *TableOne* (version 0.8.0).

Supplementary Results

Model performance evaluation

Table S2 summarizes model performance for 100 training sets, 100 internal validation sets and for the final averaged estimates of treatment effect. Curves that correspond with AUTOC values for final average estimates are shown in **figure S2.** Patient were ranked by their individualized treatment effect as estimated by the models and split into 5 quintiles. Intervals of estimated individualized treatment effect per quintile are listed in **Table S3.** **Figure 2** in the main text illustrates observed treatment effect per quintile on an absolute scale: as the difference between a patient's mean outcome if they receive ICS-containing therapy compared to if they would have received a therapy that did not contain ICS. Alternatively, observed treatment effect per quintile can be reported on a relative scale: as rate and hazard ratios **(figure S3),** or as the patient's mean outcome for each treatment arm **(figure S4).** For time to first exacerbation, observed treatment effect can also be visualized with cumulative incidence curves **(figure S5). Tables S4, S5, S6 and S7** list patient characteristics per quintile for each outcome and treatment comparison. **Table 2** and **figure 3** in the main text investigate responder characteristics and observed treatment effects for patients who are in quintile 1 or 5 for both modelled outcomes. **Table S8** and **figure S6** show results using a larger range of interest: for patients in quintile 1 or 2 and quintile 4 or 5 for both modelled outcomes.

Supplementary References

1. Lipson DA, Barnhart F, Brealey N, *et al.* Once-daily single-inhaler triple versus dual therapy in patients with COPD. *N Engl J Med,* 2018;378:1671–1680.
2. Cui Y, Kosorok MR, Sverdrup E, *et al.* Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology,* 2023;85:179–211.
3. Athey S, Tibshirani J, Wager S. Generalized random forests. *The Annals of Statistics,* 2019;47:1148–1178.
4. Xu Q-S, Liang Y-Z. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems,* 2001;56:1–11.
5. Goligher EC, Lawler PR, Jensen TP, *et al.* Heterogeneous Treatment Effects of Therapeutic-Dose Heparin in Patients Hospitalized for COVID-19. *JAMA,* 2023;329:1066-1077.
6. Yadlowsky S, Fleming S, Shah N, *et al.* Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects*. J Am Stat Assoc,* 2024;120:38–51.
7. Chernozhukov V, Demirer M, Duflo E, *et al.* Fisher-Schultz Lecture: Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments, with an Application to Immunization in India [preprint]. arXiv; 2023 [accessed 2025 June 4]. Available from: https://arxiv.org/pdf/1712.04802.

Table S1 Patient baseline characteristics selected as model input features

Characteristic	Level	Missing values (number)	Completeness (%)	Overall (n = 10,355)	FF/U/V (n = 4151)	FF/V (n = 4134)	U/V (n = 2070)
Subject Level Analysis							
Age (years)		4	99.96	65.33 (8.26)	65.38 (8.24)	65.31 (8.29)	65.27 (8.25)
Sex	Male	0	100	6870 (66.34)	2766 (66.63)	2748 (66.47)	1356 (65.51)
	Female			3485 (33.66)	1385 (33.37)	1386 (33.53)	714 (34.49)
Height (cm)		0	100	167.49 (9.25)	167.40 (9.16)	167.59 (9.29)	167.47 (9.37)
Weight (kg)		3	99.97	75.01 (19.41)	74.86 (19.52)	75.24 (19.59)	74.86 (18.84)
GOLD grade	1	8	99.92	22 (0.21)	10 (0.24)	8 (0.19)	4 (0.19)
	2			3719 (35.94)	1535 (37.03)	1455 (35.20)	729 (35.23)
	3			4982 (48.15)	1934 (46.66)	2031 (49.14)	1017 (49.15)
	4			1624 (15.70)	666 (16.07)	639 (15.46)	319 (15.42)
Previous exacerbation count		0	100	1.71 (0.86)	1.72 (0.86)	1.69 (0.83)	1.73 (0.90)
Smoking status	Former	0	100	6768 (65.36)	2715 (65.41)	2711 (65.58)	1342 (64.83)
	Current			3587 (34.64)	1436 (34.59)	1423 (34.42)	728 (35.17)
Substance Use Analysis							
Number of pack years		0	100	46.64 (26.64)	46.66 (26.73)	46.43 (26.15)	47.02 (27.41)
Laboratory Analysis							
Blood eosinophils (10^9/L)		22	99.79	0.17 [0.09,0.27]	0.16 [0.09,0.27]	0.17 [0.10,0.27]	0.17 [0.10,0.28]
Blood neutrophils (10^9/L)		22	99.79	4.93 [3.89,6.21]	4.92 [3.85,6.19]	4.92 [3.90,6.18]	4.95 [3.94,6.27]
Pulmonary Function Test Analysis							
FEV1 (L)		22	99.79	1.17 (0.48)	1.17 (0.48)	1.17 (0.48)	1.17 (0.47)
FEV1 reversibility (mL)		108	98.96	105.01 (137.31)	106.79 (147.60)	104.89 (131.83)	101.67 (126.22)
FVC (L)		22	99.79	2.54 (0.80)	2.53 (0.80)	2.55 (0.81)	2.53 (0.79)
FEV1pp (%)		27	99.74	41.85 (14.79)	41.88 (15.06)	41.77 (14.56)	41.95 (14.70)
SGRQ Analysis							
SGRQ Activity score		105	98.99	65.62 (19.47)	65.92 (19.31)	65.34 (19.66)	65.60 (19.43)
SGRQ Impacts score		105	98.99	37.22 (19.35)	37.30 (19.36)	37.43 (19.51)	36.66 (19.00)
SGRQ Symptoms score		105	98.99	65.94 (18.47)	65.92 (18.52)	66.33 (18.26)	65.24 (18.79)

Characteristics are reported for the intention-to-treat population (10,355 patients) and stratified by treatment arm. Values overall and for each treatment arm are reported as mean (standard deviation), median [interquartile range], and number (%).

FF, Fluticasone Furoate; U, Umeclidinium; V, Vilanterol; GOLD, Global Initiative for Chronic Obstructive Lung Disease; FEV1, Forced Expiratory Volume in 1 second; FEV1pp, Forced Expiratory Volume in 1 second percent predicted; FVC, Forced Vital Capacity; SGRQ, St. George's Respiratory Questionnaire; n, number of patients.

Table S2 Extended model performance

Treatment comparison	Modelled outcome	100 OOB training sets		100 internal validation sets		Final ranking of all patients	
		AUTOC		AUTOC		AUTOC	
		Mean (95% CI)	p-value[a]	Mean (95% CI)	p-value[a]	Mean (95% CI)	p-value[b]
FF/U/V-U/V	Time to first exacerbation	7.09 (-2.13 to 16.31)	0.0840	11.43 (-2.87 to 25.74)	0.1222	11.88 (3.54 to 20.22)	0.0026
	Exacerbation rate	0.33 (0.18 to 0.47)	<0.0001	0.37 (0.15 to 0.59)	0.0005	0.39 (0.27 to 0.50)	<0.0001
FF/V-U/V	Time to first exacerbation	17.20 (7.75 to 26.64)	0.0004	21.47 (6.66 to 36.27)	0.0050	22.06 (13.86 to 30.26)	<0.0001
	Exacerbation rate	0.23 (0.04 to 0.41)	0.0171	0.29 (0.02 to 0.56)	0.0302	0.27 (0.14 to 0.41)	<0.0001

FF, Fluticasone Furoate; U, Umeclidinium; V, Vilanterol; AUTOC, Area under the targeting operator characteristic, CI, Confidence interval.

[a] Median p-value from a two-sided test; [b] p-value from a one-sided test

Table S3 Intervals of estimated individualized treatment effects

Treatment comparison	Modelled outcome	Interval of estimated individualized treatment effect					
		Overall	**Quintile 1**	**Quintile 2**	**Quintile 3**	**Quintile 4**	**Quintile 5**
FF/U/V – U/V	Time to first exacerbation	57.94 to -5.96	57.94 to 30.08	30.08 to 23.28	23.28 to 18.64	18.64 to 14.07	14.07 to -5.96
	Exacerbation rate	-1.84 to -0.18	-1.84 to -0.90	-0.90 to -0.75	-0.75 to -0.64	-0.64 to -0.51	-0.51 to -0.18
FF/V-U/V	Time to first exacerbation	51.52 to -24.92	51.52 to 16.10	16.07 to 5.78	5.78 to -0.22	-0.22 to -6.42	-6.43 to -24.92
	Exacerbation rate	-1.88 to 0.10	-1.88 to -0.63	0.63 to -0.38	-0.38 to -0.26	-0.26 to -0.16	-0.16 to 0.10

Quintile 1 contains patients estimated to benefit most from an inhaled corticosteroid-containing therapy (either FF/U/V or FF/V), while patients in quintile 5 were estimated to benefit least from this therapy. For time to first exacerbation, estimated treatment effects are expressed as a difference in days of restricted mean survival time. For exacerbation rate, estimated treatment effects are expressed as a difference in mean annual exacerbation rate between the inhaled corticosteroid-containing therapy and U/V therapy.

FF, Fluticasone Furoate; U, Umeclidinium; V, Vilanterol.

Table S4 Patient characteristics per quintile for a ranking made by the FF/U/V versus U/V time to first exacerbation model.

Characteristic	Level	Quintile 1 Benefits most from FF/U/V		Quintile 2		Quintile 3		Quintile 4		Quintile 5 Benefits least from FF/U/V	
		U/V (n = 414)	FF/U/V (n = 800)	U/V (n = 413)	FF/U/V (n = 801)	U/V (n = 414)	FF/U/V (n = 801)	U/V (n = 395)	FF/U/V (n = 820)	U/V (n = 389)	FF/U/V (n = 826)
Subject Level Analysis											
Age (years)		65.01 (8.31)	65.58 (8.09)	66.28 (8.33)	66.63 (8.14)	66.55 (7.93)	66.42 (7.99)	65.05 (8.51)	64.73 (8.25)	63.30 (7.84)	63.75 (8.40)
Sex	Male	305 (73.67)	589 (73.62)	265 (64.16)	498 (62.17)	241 (58.21)	484 (60.42)	252 (63.80)	551 (67.20)	263 (67.61)	576 (69.73)
	Female	109 (26.33)	211 (26.38)	148 (35.84)	303 (37.83)	173 (41.79)	317 (39.58)	143 (36.20)	269 (32.80)	126 (32.39)	250 (30.27)
Height (cm)		168.84 (9.82)	168.82 (9.61)	166.79 (10.56)	166.01 (9.88)	166.60 (9.22)	166.25 (9.68)	167.30 (8.75)	167.42 (8.66)	167.68 (7.99)	168.33 (7.40)
Weight (kg)		75.72 (19.52)	76.21 (19.95)	73.59 (20.83)	72.54 (20.62)	73.34 (19.52)	72.31 (19.39)	74.26 (16.58)	75.11 (18.63)	77.00 (17.19)	77.99 (18.40)
GOLD grade	1	2 (0.48)	4 (0.50)							2 (0.51)	5 (0.61)
	2	152 (36.71)	338 (42.25)	117 (28.33)	242 (30.21)	99 (23.91)	199 (24.84)	130 (32.91)	262 (31.95)	215 (55.27)	458 (55.45)
	3	214 (51.69)	365 (45.62)	216 (52.30)	427 (53.31)	234 (56.52)	427 (53.31)	188 (47.59)	379 (46.22)	143 (36.76)	292 (35.35)
	4	46 (11.11)	93 (11.62)	80 (19.37)	132 (16.48)	81 (19.57)	175 (21.85)	77 (19.49)	179 (21.83)	29 (7.46)	71 (8.60)
Previous exacerbation count		1.90 (1.11)	1.89 (1.02)	1.61 (0.84)	1.65 (0.85)	1.68 (0.91)	1.68 (0.87)	1.74 (0.86)	1.69 (0.89)	1.75 (0.69)	1.71 (0.65)
Smoking status	Former	288 (69.57)	524 (65.50)	257 (62.23)	528 (65.92)	265 (64.01)	518 (64.67)	250 (63.29)	525 (64.02)	251 (64.52)	558 (67.55)
	Current	126 (30.43)	276 (34.50)	156 (37.77)	273 (34.08)	149 (35.99)	283 (35.33)	145 (36.71)	295 (35.98)	138 (35.48)	268 (32.45)
Substance Use Analysis											
Number of pack years		47.02 (26.59)	45.69 (25.75)	50.32 (30.30)	48.60 (26.87)	48.86 (26.61)	50.88 (29.31)	45.82 (23.98)	46.48 (26.47)	42.59 (27.58)	42.13 (24.58)
Laboratory Analysis											
Eosinophils (10^9/L)		0.41 [0.31,0.57]	0.43 [0.32,0.58]	0.24 [0.17,0.30]	0.23 [0.17,0.28]	0.14 [0.09,0.20]	0.14 [0.09,0.19]	0.12 [0.07,0.16]	0.11 [0.07,0.16]	0.10 [0.05,0.14]	0.09 [0.06,0.14]
Neutrophils (10^9/L)		4.64 [3.71,5.80]	4.65 [3.60,5.82]	5.16 [4.02,6.33]	4.93 [3.95,6.07]	4.89 [3.94,6.33]	5.09 [4.02,6.26]	5.08 [4.04,6.42]	5.03 [3.86,6.28]	5.01 [4.02,6.51]	4.87 [3.82,6.38]
Pulmonary Function Test Analysis											
FEV1 (L)		1.24 (0.52)	1.24 (0.52)	1.07 (0.44)	1.06 (0.43)	1.02 (0.42)	1.01 (0.43)	1.10 (0.40)	1.10 (0.42)	1.43 (0.45)	1.43 (0.46)
FEV1 reversibility (mL)		144.81 (147.72)	140.36 (171.72)	92.28 (117.70)	104.73 (122.64)	97.58 (115.49)	109.06 (170.86)	98.00 (123.69)	101.85 (129.61)	75.60 (113.92)	79.54 (132.45)
FVC (L)		2.64 (0.86)	2.63 (0.83)	2.39 (0.76)	2.36 (0.77)	2.30 (0.74)	2.30 (0.75)	2.48 (0.71)	2.49 (0.77)	2.85 (0.72)	2.87 (0.74)
FEV1pp (%)		42.49 (14.10)	42.63 (14.75)	39.06 (13.49)	39.21 (13.07)	37.96 (13.26)	37.49 (13.88)	39.91 (14.00)	39.52 (14.27)	50.55 (15.16)	50.15 (15.50)
SGRQ Analysis											
SGRQ Activity score		62.96 (20.32)	65.31 (19.29)	67.55 (19.64)	67.95 (18.17)	70.83 (17.47)	69.95 (18.45)	67.97 (17.78)	67.58 (18.42)	58.26 (19.52)	59.07 (20.18)
SGRQ Impacts score		36.42 (18.71)	38.71 (18.51)	37.03 (19.70)	38.35 (19.11)	40.38 (18.92)	40.10 (19.28)	38.39 (18.09)	38.39 (19.42)	30.82 (18.37)	31.51 (19.39)
SGRQ Symptoms score		69.22 (19.18)	70.68 (17.88)	67.14 (18.28)	67.44 (18.76)	68.65 (18.36)	70.01 (17.52)	66.15 (17.19)	66.55 (17.03)	54.28 (16.83)	55.21 (17.33)
COPD Exacerbation Summary Analysis											
On treatment at risk duration (years)		0.78 (0.35)	0.89 (0.27)	0.78 (0.36)	0.87 (0.28)	0.80 (0.34)	0.89 (0.26)	0.84 (0.31)	0.90 (0.26)	0.91 (0.25)	0.92 (0.23)

Time to event analysis											
Observed first event	0	173 (41.79)	404 (50.50)	192 (46.49)	394 (49.19)	195 (47.10)	402 (50.19)	201 (50.89)	430 (52.44)	246 (63.24)	502 (60.77)
	1	241 (58.21)	396 (49.50)	221 (53.51)	407 (50.81)	219 (52.90)	399 (49.81)	194 (49.11)	390 (47.56)	143 (36.76)	324 (39.23)
Time to first event/censor (days)		179.65 (149.22)	225.38 (143.97)	189.80 (149.13)	216.81 (140.81)	199.77 (148.46)	229.97 (139.53)	221.17 (142.08)	234.73 (138.53)	254.90 (140.40)	260.79 (131.89)
Exacerbation - Andersen-Gill Analysis											
Number of observed events	0	173 (41.79)	404 (50.50)	192 (46.49)	394 (49.19)	195 (47.10)	402 (50.19)	201 (50.89)	430 (52.44)	246 (63.24)	502 (60.77)
	1	122 (29.47)	225 (28.12)	124 (30.02)	211 (26.34)	120 (28.99)	220 (27.47)	126 (31.90)	224 (27.32)	86 (22.11)	211 (25.54)
	2	52 (12.56)	89 (11.12)	42 (10.17)	108 (13.48)	44 (10.63)	110 (13.73)	30 (7.59)	85 (10.37)	31 (7.97)	73 (8.84)
	3	25 (6.04)	48 (6.00)	34 (8.23)	62 (7.74)	30 (7.25)	43 (5.37)	27 (6.84)	45 (5.49)	12 (3.08)	26 (3.15)
	4	19 (4.59)	16 (2.00)	11 (2.66)	13 (1.62)	14 (3.38)	16 (2.00)	7 (1.77)	19 (2.32)	9 (2.31)	10 (1.21)
	5	12 (2.90)	7 (0.88)	5 (1.21)	8 (1.00)	3 (0.72)	3 (0.37)	2 (0.51)	10 (1.22)	2 (0.51)	1 (0.12)
	6	7 (1.69)	5 (0.62)	2 (0.48)	2 (0.25)	6 (1.45)	3 (0.37)		5 (0.61)	3 (0.77)	2 (0.24)
	7	2 (0.48)	3 (0.38)	2 (0.48)	1 (0.12)		2 (0.25)	1 (0.25)	1 (0.12)		1 (0.12)
	8			1 (0.24)	1 (0.12)	1 (0.24)	2 (0.25)		1 (0.12)		
	9	1 (0.24)	3 (0.38)								
	10				1 (0.12)	1 (0.24)		1 (0.25)			
	11	1 (0.24)									

Values are reported as mean (standard deviation), median [interquartile range], and number (%).

FF, Fluticasone Furoate; U, Umeclidinium; V, Vilanterol; GOLD, Global Initiative for Chronic Obstructive Lung Disease; FEV1, Forced Expiratory Volume in 1 second; FEV1pp, Forced Expiratory Volume in 1 second percent predicted; FVC, Forced Vital Capacity; SGRQ, St. George's Respiratory Questionnaire; n, number of patients.

Table S5 Patient characteristics per quintile for a ranking made by the FF/U/V versus U/V exacerbation rate model.

Characteristic	Level	Quintile 1 Benefits most from FF/U/V		Quintile 2		Quintile 3		Quintile 4		Quintile 5 Benefits least from FF/U/V	
		U/V (n = 399)	FF/U/V (n = 816)	U/V (n = 420)	FF/U/V (n = 795)	U/V (n = 394)	FF/U/V (n = 821)	U/V (n = 407)	FF/U/V (n = 807)	U/V (n = 405)	FF/U/V (n = 809)
Subject Level Analysis											
Age (years)		65.64 (7.88)	65.85 (7.77)	65.69 (8.03)	66.28 (7.42)	65.65 (8.24)	65.37 (7.96)	65.44 (8.27)	65.44 (8.93)	63.89 (8.75)	64.10 (8.88)
Sex	Male	295 (73.93)	612 (75.00)	268 (63.81)	483 (60.75)	225 (57.11)	501 (61.02)	263 (64.62)	520 (64.44)	275 (67.90)	582 (71.94)
	Female	104 (26.07)	204 (25.00)	152 (36.19)	312 (39.25)	169 (42.89)	320 (38.98)	144 (35.38)	287 (35.56)	130 (32.10)	227 (28.06)
Height (cm)		169.26 (10.03)	168.62 (9.64)	166.60 (9.99)	166.34 (9.82)	166.22 (9.54)	166.08 (9.24)	167.04 (8.69)	166.93 (8.93)	168.11 (8.08)	168.88 (7.56)
Weight (kg)		76.75 (19.66)	75.22 (19.90)	71.63 (19.13)	73.05 (19.99)	73.34 (19.47)	73.20 (19.94)	74.42 (18.32)	73.69 (18.48)	77.77 (17.07)	79.08 (18.62)
GOLD grade	1	2 (0.50)	3 (0.37)				1 (0.12)			2 (0.49)	5 (0.62)
	2	116 (29.07)	272 (33.33)	89 (21.19)	178 (22.39)	102 (25.89)	217 (26.43)	161 (39.56)	344 (42.63)	245 (60.49)	488 (60.32)
	3	203 (50.88)	373 (45.71)	225 (53.57)	397 (49.94)	213 (54.06)	440 (53.59)	207 (50.86)	386 (47.83)	147 (36.30)	294 (36.34)
	4	78 (19.55)	168 (20.59)	106 (25.24)	220 (27.67)	79 (20.05)	163 (19.85)	39 (9.58)	77 (9.54)	11 (2.72)	22 (2.72)
Previous exacerbation count		1.89 (1.17)	1.87 (1.04)	1.66 (0.87)	1.70 (0.95)	1.62 (0.81)	1.67 (0.90)	1.78 (0.91)	1.67 (0.75)	1.73 (0.63)	1.71 (0.61)
Smoking status	Former	268 (67.17)	540 (66.18)	268 (63.81)	542 (68.18)	253 (64.21)	518 (63.09)	266 (65.36)	516 (63.94)	256 (63.21)	537 (66.38)
	Current	131 (32.83)	276 (33.82)	152 (36.19)	253 (31.82)	141 (35.79)	303 (36.91)	141 (34.64)	291 (36.06)	149 (36.79)	272 (33.62)
Substance Use Analysis											
Number of pack years		48.91 (28.35)	47.15 (25.35)	48.23 (26.81)	48.75 (26.79)	47.78 (27.39)	46.89 (27.09)	45.84 (26.88)	46.07 (28.89)	44.16 (26.47)	44.81 (25.55)
Laboratory Analysis											
Eosinophils (10^9/L)		0.43 [0.32,0.60]	0.43 [0.30,0.59]	0.22 [0.12,0.29]	0.19 [0.12,0.27]	0.15 [0.09,0.22]	0.15 [0.09,0.21]	0.13 [0.08,0.19]	0.13 [0.08,0.18]	0.10 [0.06,0.15]	0.10 [0.07,0.15]
Neutrophils (10^9/L)		5.13 [4.07,6.40]	4.95 [3.87,6.22]	5.19 [4.03,6.37]	5.23 [4.18,6.36]	4.83 [3.85,5.93]	4.92 [3.89,6.20]	4.69 [3.75,6.14]	4.80 [3.70,6.13]	4.96 [4.05,6.41]	4.59 [3.62,5.98]
Pulmonary Function Test Analysis											
FEV1 (L)		1.13 (0.50)	1.12 (0.50)	0.99 (0.45)	0.96 (0.43)	1.03 (0.39)	1.03 (0.42)	1.20 (0.43)	1.22 (0.41)	1.50 (0.41)	1.51 (0.42)
FEV1 reversibility (mL)		144.14 (145.67)	144.73 (154.30)	113.00 (114.59)	116.77 (144.29)	95.61 (108.84)	108.31 (119.22)	92.41 (132.36)	92.53 (159.18)	65.01 (114.47)	71.99 (150.53)
FVC (L)		2.53 (0.84)	2.50 (0.83)	2.32 (0.79)	2.23 (0.77)	2.34 (0.73)	2.35 (0.75)	2.57 (0.70)	2.61 (0.73)	2.89 (0.73)	2.96 (0.70)
FEV1pp (%)		38.92 (14.69)	38.69 (14.99)	35.83 (13.30)	35.01 (13.05)	38.37 (12.59)	38.15 (13.02)	43.86 (13.52)	44.81 (13.87)	52.67 (12.95)	52.53 (13.36)
SGRQ Analysis											
SGRQ Activity score		67.79 (19.99)	69.54 (18.99)	70.89 (19.55)	72.01 (17.83)	70.63 (17.29)	70.11 (17.41)	64.11 (17.52)	64.24 (18.25)	54.46 (17.98)	53.79 (18.15)
SGRQ Impacts score		40.13 (19.36)	42.45 (18.69)	41.69 (19.57)	42.62 (18.58)	39.90 (18.55)	40.83 (19.11)	35.23 (18.10)	35.47 (18.90)	26.32 (14.99)	25.52 (15.90)
SGRQ Symptoms score		70.95 (18.20)	71.23 (17.68)	70.56 (17.29)	71.39 (16.28)	66.97 (17.51)	68.48 (16.91)	62.99 (17.64)	64.77 (17.85)	54.53 (18.38)	53.69 (18.12)
COPD Exacerbation Summary Analysis											
On treatment at risk duration (years)		0.75 (0.37)	0.88 (0.28)	0.80 (0.34)	0.88 (0.28)	0.81 (0.34)	0.89 (0.26)	0.84 (0.32)	0.91 (0.25)	0.91 (0.24)	0.93 (0.22)

Time to event analysis											
Observed first event	0	155 (38.85)	389 (47.67)	180 (42.86)	358 (45.03)	185 (46.95)	429 (52.25)	219 (53.81)	469 (58.12)	268 (66.17)	487 (60.20)
	1	244 (61.15)	427 (52.33)	240 (57.14)	437 (54.97)	209 (53.05)	392 (47.75)	188 (46.19)	338 (41.88)	137 (33.83)	322 (39.80)
Time to first event/censor (days)		167.73 (147.39)	215.92 (143.66)	187.32 (146.34)	207.06 (141.43)	196.51 (144.12)	234.95 (138.11)	224.91 (147.38)	251.57 (136.82)	265.24 (136.83)	258.77 (131.38)
Exacerbation - Andersen-Gill Analysis											
Number of observed events	0	155 (38.85)	389 (47.67)	180 (42.86)	358 (45.03)	185 (46.95)	429 (52.25)	219 (53.81)	469 (58.12)	268 (66.17)	487 (60.20)
	1	130 (32.58)	241 (29.53)	129 (30.71)	219 (27.55)	123 (31.22)	214 (26.07)	110 (27.03)	206 (25.53)	86 (21.23)	211 (26.08)
	2	46 (11.53)	94 (11.52)	55 (13.10)	132 (16.60)	36 (9.14)	104 (12.67)	33 (8.11)	67 (8.30)	29 (7.16)	68 (8.41)
	3	32 (8.02)	54 (6.62)	26 (6.19)	52 (6.54)	32 (8.12)	45 (5.48)	28 (6.88)	43 (5.33)	10 (2.47)	30 (3.71)
	4	18 (4.51)	18 (2.21)	16 (3.81)	20 (2.52)	8 (2.03)	13 (1.58)	12 (2.95)	12 (1.49)	6 (1.48)	11 (1.36)
	5	9 (2.26)	9 (1.10)	8 (1.90)	6 (0.75)	3 (0.76)	7 (0.85)	2 (0.49)	6 (0.74)	2 (0.49)	1 (0.12)
	6	6 (1.50)	4 (0.49)	2 (0.48)	5 (0.63)	4 (1.02)	3 (0.37)	3 (0.74)	4 (0.50)	3 (0.74)	1 (0.12)
	7	1 (0.25)	3 (0.37)	2 (0.48)	2 (0.25)	2 (0.51)	3 (0.37)				
	8		1 (0.12)	1 (0.24)	1 (0.13)	1 (0.25)	2 (0.24)				
	9	1 (0.25)	3 (0.37)								
	10			1 (0.24)			1 (0.12)			1 (0.25)	
	11	1 (0.25)									

Values are reported as mean (standard deviation), median [interquartile range], and number (%).

FF, Fluticasone Furoate; U, Umeclidinium; V, Vilanterol; GOLD, Global Initiative for Chronic Obstructive Lung Disease; FEV1, Forced Expiratory Volume in 1 second; FEV1pp, Forced Expiratory Volume in 1 second percent predicted; FVC, Forced Vital Capacity; SGRQ, St. George's Respiratory Questionnaire; n, number of patients.

Table S6 Patient characteristics per quintile for a ranking made by the FF/V versus U/V time to first exacerbation model.

Characteristic	Level	Quintile 1 Benefits most from FF/V		Quintile 2		Quintile 3		Quintile 4		Quintile 5 Benefits least from FF/V	
		U/V (n= 400)	FF/V (n = 811)	U/V (n = 416)	FF/V (n = 796)	U/V (n = 429)	FF/V (n = 783)	U/V (n = 400)	FF/V (n = 812)	U/V (n = 380)	FF/V (n = 832)
Subject Level Analysis											
Age (years)		64.09 (8.63)	64.94 (8.25)	66.01 (8.15)	66.09 (8.47)	65.62 (8.49)	65.47 (8.63)	65.41 (8.04)	65.37 (8.14)	65.13 (7.85)	64.89 (7.89)
Sex	Male	274 (68.50)	548 (67.57)	249 (59.86)	448 (56.28)	239 (55.71)	428 (54.66)	253 (63.25)	555 (68.35)	311 (81.84)	706 (84.86)
	Female	126 (31.50)	263 (32.43)	167 (40.14)	348 (43.72)	190 (44.29)	355 (45.34)	147 (36.75)	257 (31.65)	69 (18.16)	126 (15.14)
Height (cm)		167.82 (9.77)	167.83 (9.98)	165.74 (10.83)	165.20 (10.58)	165.58 (9.91)	165.47 (9.34)	167.74 (8.13)	168.42 (8.39)	170.71 (6.33)	170.75 (6.55)
Weight (kg)		76.06 (19.12)	76.90 (19.93)	76.37 (21.22)	74.66 (20.20)	73.57 (19.26)	75.01 (19.47)	74.89 (17.81)	75.82 (20.30)	72.85 (16.13)	73.18 (17.09)
GOLD grade	1	2 (0.50)	1 (0.12)		2 (0.25)	1 (0.23)	3 (0.38)	1 (0.25)			1 (0.12)
	2	174 (43.50)	375 (46.24)	153 (36.78)	294 (36.93)	174 (40.56)	296 (37.80)	117 (29.25)	251 (30.91)	95 (25.00)	195 (23.44)
	3	189 (47.25)	341 (42.05)	199 (47.84)	405 (50.88)	203 (47.32)	389 (49.68)	212 (53.00)	429 (52.83)	192 (50.53)	423 (50.84)
	4	35 (8.75)	94 (11.59)	64 (15.38)	95 (11.93)	51 (11.89)	95 (12.13)	70 (17.50)	132 (16.26)	93 (24.47)	213 (25.60)
Previous exacerbation count		1.86 (1.05)	1.90 (0.93)	1.82 (0.94)	1.75 (0.81)	1.76 (0.86)	1.68 (0.78)	1.66 (0.87)	1.62 (0.81)	1.56 (0.70)	1.53 (0.80)
Smoking status	Former	260 (65.00)	547 (67.45)	275 (66.11)	518 (65.08)	289 (67.37)	493 (62.96)	235 (58.75)	531 (65.39)	252 (66.32)	563 (67.67)
	Current	140 (35.00)	264 (32.55)	141 (33.89)	278 (34.92)	140 (32.63)	290 (37.04)	165 (41.25)	281 (34.61)	128 (33.68)	269 (32.33)
Substance Use Analysis											
Number of pack years		44.98 (27.24)	43.56 (25.79)	46.78 (27.29)	46.22 (25.65)	45.70 (28.26)	46.09 (25.16)	47.01 (25.37)	47.38 (27.46)	50.73 (27.52)	49.43 (26.14)
Laboratory Analysis											
Eosinophils (10^9/L)		0.42 [0.34,0.59]	0.42 [0.32,0.58]	0.24 [0.15,0.29]	0.22 [0.15,0.28]	0.14 [0.08,0.20]	0.15 [0.09,0.21]	0.12 [0.07,0.17]	0.12 [0.08,0.17]	0.10 [0.06,0.15]	0.10 [0.06,0.15]
Neutrophils (10^9/L)		4.74 [3.70,6.02]	4.71 [3.68,5.97]	5.01 [4.10,6.35]	5.04 [4.02,6.19]	4.90 [3.94,6.16]	4.95 [4.00,6.26]	5.03 [4.08,6.42]	4.98 [3.94,6.34]	4.96 [3.93,6.35]	4.89 [3.81,6.16]
Pulmonary Function Test Analysis											
FEV1 (L)		1.29 (0.51)	1.25 (0.51)	1.12 (0.42)	1.14 (0.45)	1.18 (0.46)	1.15 (0.45)	1.12 (0.44)	1.15 (0.47)	1.15 (0.50)	1.15 (0.49)
FEV1 reversibility (mL)		143.18 (156.06)	148.82 (177.24)	106.15 (121.28)	116.12 (123.93)	100.21 (136.73)	104.44 (117.91)	86.92 (105.16)	82.13 (125.86)	72.08 (89.28)	74.48 (86.23)
FVC (L)		2.65 (0.84)	2.62 (0.84)	2.41 (0.76)	2.40 (0.78)	2.43 (0.77)	2.43 (0.77)	2.44 (0.74)	2.52 (0.82)	2.74 (0.75)	2.73 (0.78)
FEV1pp (%)		44.69 (14.24)	43.80 (14.61)	42.00 (14.23)	42.71 (13.55)	43.77 (14.29)	43.12 (14.28)	40.11 (13.87)	40.37 (14.08)	38.68 (16.07)	38.48 (15.35)
SGRQ Analysis											
SGRQ Activity score		66.77 (18.54)	67.03 (19.19)	67.43 (20.08)	68.11 (18.63)	67.93 (18.36)	68.20 (18.26)	66.33 (18.35)	65.99 (18.08)	58.85 (20.66)	57.72 (21.79)
SGRQ Impacts score		40.64 (18.63)	40.50 (19.21)	36.89 (19.72)	40.47 (19.53)	38.87 (18.75)	39.80 (19.41)	36.22 (18.71)	37.95 (19.06)	30.21 (17.65)	28.99 (17.91)
SGRQ Symptoms score		71.51 (17.78)	72.46 (16.52)	67.13 (19.63)	69.26 (18.75)	67.06 (18.16)	68.39 (17.89)	63.67 (17.49)	65.29 (17.32)	56.00 (17.34)	56.46 (16.68)
COPD Exacerbation Summary Analysis											
On treatment at risk duration (years)		0.79 (0.35)	0.87 (0.29)	0.78 (0.36)	0.85 (0.31)	0.84 (0.32)	0.83 (0.32)	0.85 (0.31)	0.82 (0.33)	0.85 (0.31)	0.82 (0.32)

Time to event analysis											
Observed first event	0	167 (41.75)	403 (49.69)	188 (45.19)	409 (51.38)	216 (50.35)	402 (51.34)	216 (54.00)	381 (46.92)	220 (57.89)	436 (52.40)
	1	233 (58.25)	408 (50.31)	228 (54.81)	387 (48.62)	213 (49.65)	381 (48.66)	184 (46.00)	431 (53.08)	160 (42.11)	396 (47.60)
Time to first event/censor (days)		179.62 (150.98)	212.57 (146.28)	190.42 (146.48)	217.01 (144.29)	215.60 (146.79)	207.77 (148.17)	224.21 (146.50)	197.72 (145.47)	233.55 (144.12)	210.84 (144.40)
Exacerbation - Andersen-Gill Analysis											
Number of observed events	0	167 (41.75)	403 (49.69)	188 (45.19)	409 (51.38)	216 (50.35)	402 (51.34)	216 (54.00)	381 (46.92)	220 (57.89)	436 (52.40)
	1	115 (28.75)	213 (26.26)	135 (32.45)	219 (27.51)	116 (27.04)	206 (26.31)	110 (27.50)	234 (28.82)	102 (26.84)	237 (28.49)
	2	51 (12.75)	116 (14.30)	41 (9.86)	106 (13.32)	45 (10.49)	85 (10.86)	37 (9.25)	98 (12.07)	25 (6.58)	93 (11.18)
	3	29 (7.25)	46 (5.67)	30 (7.21)	41 (5.15)	29 (6.76)	45 (5.75)	20 (5.00)	52 (6.40)	20 (5.26)	40 (4.81)
	4	17 (4.25)	20 (2.47)	13 (3.12)	14 (1.76)	10 (2.33)	33 (4.21)	10 (2.50)	30 (3.69)	10 (2.63)	19 (2.28)
	5	10 (2.50)	7 (0.86)	7 (1.68)	3 (0.38)	3 (0.70)	8 (1.02)	3 (0.75)	8 (0.99)	1 (0.26)	5 (0.60)
	6	7 (1.75)	3 (0.37)	1 (0.24)	3 (0.38)	6 (1.40)	3 (0.38)	3 (0.75)	4 (0.49)	1 (0.26)	1 (0.12)
	7	2 (0.50)	3 (0.37)			3 (0.70)	1 (0.13)		4 (0.49)		
	8			1 (0.24)					1 (0.12)	1 (0.26)	
	9	1 (0.25)									1 (0.12)
	10				1 (0.13)	1 (0.23)		1 (0.25)			
	11	1 (0.25)									

Values are reported as mean (standard deviation), median [interquartile range], and number (%).

FF, Fluticasone Furoate; U, Umeclidinium; V, Vilanterol; GOLD, Global Initiative for Chronic Obstructive Lung Disease; FEV1, Forced Expiratory Volume in 1 second; FEV1pp, Forced Expiratory Volume in 1 second percent predicted; FVC, Forced Vital Capacity; SGRQ, St. George's Respiratory Questionnaire; n, number of patients.

Table S7 Patient characteristics per quintile for a ranking made by the FF/V versus U/V exacerbation rate model.

Characteristic	Level	Quintile 1 Benefits most from FF/V		Quintile 2		Quintile 3		Quintile 4		Quintile 5 Benefits least from FF/V	
		U/V (n = 407)	FF/V (n = 805)	U/V (n = 401)	FF/V (n = 811)	U/V (n = 419)	FF/V (n = 793)	U/V (n = 409)	FF/V (n = 803)	U/V (n = 389)	FF/V (n = 822)
Subject Level Analysis											
Age (years)		64.59 (8.49)	64.93 (8.06)	64.36 (8.00)	65.01 (8.41)	65.22 (7.79)	65.49 (8.18)	65.70 (8.53)	65.51 (8.56)	66.50 (8.34)	65.78 (8.18)
Sex	M	302 (74.20)	600 (74.53)	281 (70.07)	523 (64.49)	243 (58.00)	462 (58.26)	232 (56.72)	495 (61.64)	268 (68.89)	605 (73.60)
	F	105 (25.80)	205 (25.47)	120 (29.93)	288 (35.51)	176 (42.00)	331 (41.74)	177 (43.28)	308 (38.36)	121 (31.11)	217 (26.40)
Height (cm)		169.53 (9.85)	169.51 (9.64)	168.62 (9.81)	167.72 (10.06)	166.55 (9.96)	165.96 (9.47)	165.22 (8.81)	166.67 (9.27)	167.33 (7.44)	167.98 (7.40)
Weight (kg)		78.59 (20.47)	77.40 (19.73)	77.98 (18.61)	78.67 (19.56)	75.51 (18.31)	74.21 (19.49)	70.66 (18.10)	73.39 (19.24)	70.94 (17.21)	71.89 (18.44)
GOLD grade	1	2 (0.49)	1 (0.12)		1 (0.12)		4 (0.50)	2 (0.49)			1 (0.12)
	2	149 (36.61)	320 (39.75)	130 (32.42)	281 (34.65)	154 (36.75)	281 (35.44)	137 (33.50)	262 (32.63)	143 (36.76)	267 (32.48)
	3	203 (49.88)	358 (44.47)	212 (52.87)	434 (53.51)	204 (48.69)	398 (50.19)	198 (48.41)	395 (49.19)	178 (45.76)	402 (48.91)
	4	53 (13.02)	126 (15.65)	59 (14.71)	95 (11.71)	61 (14.56)	110 (13.87)	72 (17.60)	146 (18.18)	68 (17.48)	152 (18.49)
Previous exacerbation count		1.86 (1.13)	1.83 (0.93)	1.77 (0.87)	1.76 (0.87)	1.78 (0.92)	1.70 (0.80)	1.68 (0.82)	1.64 (0.80)	1.58 (0.67)	1.55 (0.75)
Smoking status	Former	272 (66.83)	541 (67.20)	247 (61.60)	519 (64.00)	275 (65.63)	504 (63.56)	259 (63.33)	524 (65.26)	258 (66.32)	564 (68.61)
	Current	135 (33.17)	264 (32.80)	154 (38.40)	292 (36.00)	144 (34.37)	289 (36.44)	150 (36.67)	279 (34.74)	131 (33.68)	258 (31.39)
Substance Use Analysis											
Number of pack years		47.74 (28.52)	46.61 (26.91)	46.73 (27.55)	45.60 (26.63)	45.16 (27.50)	42.83 (23.42)	42.87 (23.40)	46.83 (25.82)	52.74 (27.98)	50.77 (27.01)
Laboratory Analysis											
Eosinophils (10^9/L)		0.44 [0.37,0.60]	0.44 [0.35,0.59]	0.24 [0.16,0.29]	0.23 [0.15,0.28]	0.15 [0.10,0.21]	0.15 [0.09,0.21]	0.12 [0.07,0.18]	0.12 [0.08,0.17]	0.10 [0.05,0.14]	0.10 [0.06,0.14]
Neutrophils (10^9/L)		5.09 [3.98,6.34]	4.92 [3.90,6.09]	5.03 [4.05,6.42]	5.15 [4.11,6.47]	4.91 [3.98,6.14]	5.05 [3.96,6.26]	4.91 [3.95,6.36]	4.79 [3.87,6.12]	4.79 [3.71,6.15]	4.68 [3.64,5.93]
Pulmonary Function Test Analysis											
FEV1 (L)		1.26 (0.52)	1.24 (0.51)	1.16 (0.41)	1.15 (0.44)	1.15 (0.44)	1.12 (0.44)	1.11 (0.46)	1.14 (0.47)	1.19 (0.51)	1.19 (0.49)
FEV1 reversibility (mL)		143.00 (151.84)	151.22 (172.26)	133.29 (127.53)	134.24 (139.31)	97.28 (120.06)	102.33 (114.50)	68.78 (113.39)	73.44 (109.56)	66.94 (91.14)	64.28 (87.16)
FVC (L)		2.66 (0.84)	2.68 (0.84)	2.56 (0.76)	2.52 (0.80)	2.45 (0.76)	2.42 (0.80)	2.41 (0.76)	2.49 (0.80)	2.57 (0.78)	2.61 (0.78)
FEV1pp (%)		42.55 (14.63)	42.13 (15.04)	40.45 (13.30)	40.95 (13.25)	42.15 (14.28)	41.61 (14.00)	41.80 (14.89)	41.66 (14.65)	42.62 (16.23)	41.97 (15.56)
SGRQ Analysis											
SGRQ Activity score		66.30 (20.30)	67.33 (19.90)	71.51 (19.24)	71.86 (18.42)	68.52 (17.59)	68.67 (17.27)	65.02 (17.99)	63.83 (18.26)	56.13 (18.81)	55.23 (20.02)
SGRQ Impacts score		40.26 (19.71)	41.07 (19.75)	43.62 (19.33)	45.49 (19.45)	39.34 (18.00)	41.24 (18.23)	34.36 (17.99)	34.86 (18.08)	25.27 (14.14)	24.96 (14.93)
SGRQ Symptoms score		70.26 (19.19)	70.73 (18.00)	70.28 (17.84)	73.24 (16.43)	67.93 (16.47)	69.54 (16.10)	63.30 (17.56)	64.08 (17.09)	53.78 (17.80)	54.14 (17.10)
COPD Exacerbation Summary Analysis											
On treatment at risk duration (years)		0.76 (0.36)	0.85 (0.31)	0.82 (0.33)	0.86 (0.29)	0.83 (0.32)	0.82 (0.33)	0.84 (0.31)	0.81 (0.34)	0.85 (0.31)	0.85 (0.30)

Time to event analysis											
Observed first event	0	161 (39.56)	399 (49.57)	180 (44.89)	390 (48.09)	221 (52.74)	400 (50.44)	218 (53.30)	399 (49.69)	227 (58.35)	443 (53.89)
	1	246 (60.44)	406 (50.43)	221 (55.11)	421 (51.91)	198 (47.26)	393 (49.56)	191 (46.70)	404 (50.31)	162 (41.65)	379 (46.11)
Time to first event/censor (days)		173.94 (149.83)	206.51 (147.35)	190.90 (146.61)	211.65 (144.00)	218.26 (144.83)	208.84 (145.21)	224.59 (145.62)	197.02 (148.17)	234.78 (146.51)	221.51 (143.45)
Exacerbation - Andersen-Gill Analysis											
Number of observed events	0	161 (39.56)	399 (49.57)	180 (44.89)	390 (48.09)	221 (52.74)	400 (50.44)	218 (53.30)	399 (49.69)	227 (58.35)	443 (53.89)
	1	129 (31.70)	205 (25.47)	115 (28.68)	235 (28.98)	119 (28.40)	232 (29.26)	117 (28.61)	222 (27.65)	98 (25.19)	215 (26.16)
	2	52 (12.78)	120 (14.91)	46 (11.47)	106 (13.07)	34 (8.11)	81 (10.21)	36 (8.80)	97 (12.08)	31 (7.97)	94 (11.44)
	3	32 (7.86)	49 (6.09)	31 (7.73)	45 (5.55)	26 (6.21)	37 (4.67)	22 (5.38)	50 (6.23)	17 (4.37)	43 (5.23)
	4	16 (3.93)	17 (2.11)	14 (3.49)	24 (2.96)	10 (2.39)	33 (4.16)	9 (2.20)	25 (3.11)	11 (2.83)	17 (2.07)
	5	8 (1.97)	8 (0.99)	9 (2.24)	4 (0.49)	3 (0.72)	6 (0.76)	2 (0.49)	4 (0.50)	2 (0.51)	9 (1.09)
	6	5 (1.23)	4 (0.50)	4 (1.00)	5 (0.62)	4 (0.95)	2 (0.25)	3 (0.73)	3 (0.37)	2 (0.51)	
	7	2 (0.49)	2 (0.25)		2 (0.25)	2 (0.48)	2 (0.25)		2 (0.25)	1 (0.26)	
	8			1 (0.25)				1 (0.24)	1 (0.12)		
	9	1 (0.25)									1 (0.12)
	10		1 (0.12)	1 (0.25)				1 (0.24)			
	11	1 (0.25)									

Values are reported as mean (standard deviation), median [interquartile range], and number (%).

FF, Fluticasone Furoate; U, Umeclidinium; V, Vilanterol; GOLD, Global Initiative for Chronic Obstructive Lung Disease; FEV1, Forced Expiratory Volume in 1 second; FEV1pp, Forced Expiratory Volume in 1 second percent predicted; FVC, Forced Vital Capacity; SGRQ, St. George's Respiratory Questionnaire; n, number of patients.

Table S8 Characteristics of patients selected to benefit most or least for both modelled outcomes, using a larger range of interest

Characteristic	Level	FF/U/V versus U/V		FF/V versus U/V	
		Q1 and/or Q2 Benefits most from FF/U/V (n = 1906)	Q4 and/or Q5 Benefits least from FF/U/V (n = 1902)	Q1 and/or Q2 Benefits most from FF/V (n = 2002)	Q4 and/or Q5 Benefits least from FF/V (n = 1889)
Subject Level Analysis					
Age (years)		66.04 (7.89)	64.32 (8.61)	64.91 (8.31)	65.41 (8.08)
Sex	Male	1330 (69.78)	1301 (68.40)	1360 (67.93)	1395 (73.85)
	Female	576 (30.22)	601 (31.60)	642 (32.07)	494 (26.15)
Height (cm)		167.74 (9.94)	167.87 (7.97)	168.03 (10.02)	168.79 (7.51)
Weight (kg)		74.20 (20.25)	76.48 (18.01)	77.52 (19.80)	73.11 (18.01)
GOLD grade	1	5 (0.26)	7 (0.37)	4 (0.20)	2 (0.11)
	2	619 (32.48)	991 (52.10)	794 (39.66)	571 (30.23)
	3	960 (50.37)	780 (41.01)	956 (47.75)	949 (50.24)
	4	322 (16.89)	124 (6.52)	248 (12.39)	367 (19.43)
Previous exacerbation count		1.79 (1.01)	1.72 (0.72)	1.84 (0.95)	1.58 (0.77)
Smoking status	Former	1264 (66.32)	1240 (65.19)	1309 (65.38)	1251 (66.23)
	Current	642 (33.68)	662 (34.81)	693 (34.62)	638 (33.77)
Substance Use Analysis					
Number of pack years		47.69 (26.76)	44.02 (26.08)	45.68 (26.74)	48.89 (26.73)
Laboratory Analysis					
Eosinophils (10^9/L)		0.35 [0.25,0.50]	0.10 [0.06,0.15]	0.34 [0.26,0.48]	0.11 [0.06,0.15]
Neutrophils (10^9/L)		4.91 [3.85,6.09]	4.74 [3.76,6.16]	4.92 [3.90,6.14]	4.83 [3.81,6.22]
Pulmonary Function Test Analysis					
FEV1 (L)		1.12 (0.49)	1.38 (0.43)	1.22 (0.49)	1.18 (0.49)
FEV1 reversibility (mL)		132.54 (151.42)	83.18 (133.94)	140.02 (154.61)	69.58 (94.40)
FVC (L)		2.49 (0.83)	2.81 (0.72)	2.59 (0.83)	2.63 (0.79)
FEV1pp (%)		39.64 (14.23)	48.99 (14.25)	42.66 (14.21)	40.88 (15.29)
SGRQ Analysis					
SGRQ Activity score		67.47 (19.28)	59.21 (18.89)	67.62 (19.55)	58.76 (19.89)
SGRQ Impacts score		39.71 (18.80)	30.84 (17.88)	40.98 (19.57)	29.64 (17.53)
SGRQ Symptoms score		69.91 (17.93)	57.70 (17.97)	71.03 (18.02)	57.64 (17.46)

Values are reported as mean (standard deviation), median [interquartile range], and number (%). Q1 and/or Q2, group of patients estimated to be in quintile 1 and/or quintile 2 for both modelled outcomes; Q4 and/or Q5, group of patients estimated to be in quintile 4 and/or quintile 5 for both modelled outcomes; FF, Fluticasone Furoate; U, Umeclidinium; V, Vilanterol; GOLD, Global Initiative for Chronic Obstructive Lung Disease; FEV1, Forced Expiratory Volume in 1 second; FEV1pp, Forced Expiratory Volume in 1 second percent predicted; FVC, Forced Vital Capacity; SGRQ, St. George's Respiratory Questionnaire; n, number of patients.

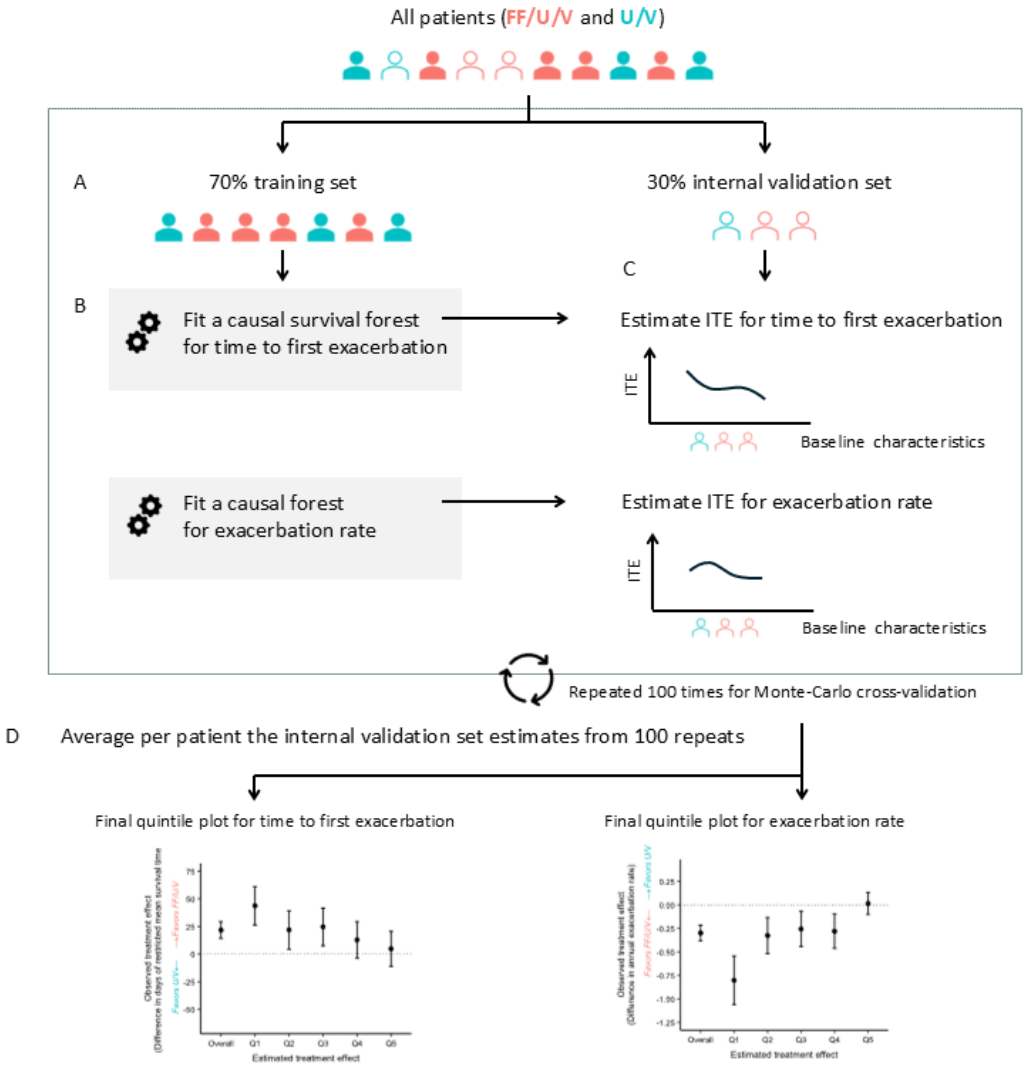


Figure S1 Monte-Carlo cross-validation. **A** Patients in the fluticasone furoate/umeclidinium/vilanterol (FF/U/V) and U/V arms are split in a 70% training set and a 30% internal validation set. **B** A causal survival forest is trained estimate individualized treatment effect (ITE) for time to first exacerbation, while a causal forest is trained to estimate ITE for exacerbation rate. **C** Both models estimate individualized treatment effects on the internal validation set. For the Monte-Carlo cross-validation approach, **A, B** and **C** are repeated 100 times. **D** Internal validation set estimations of treatment effect are averaged to get a final treatment effect estimation for each patient. The same methodology was repeated for FF/V versus U/V. Figure inspired by a similar figure in the supplementary material of work by Goligher et al[E4].

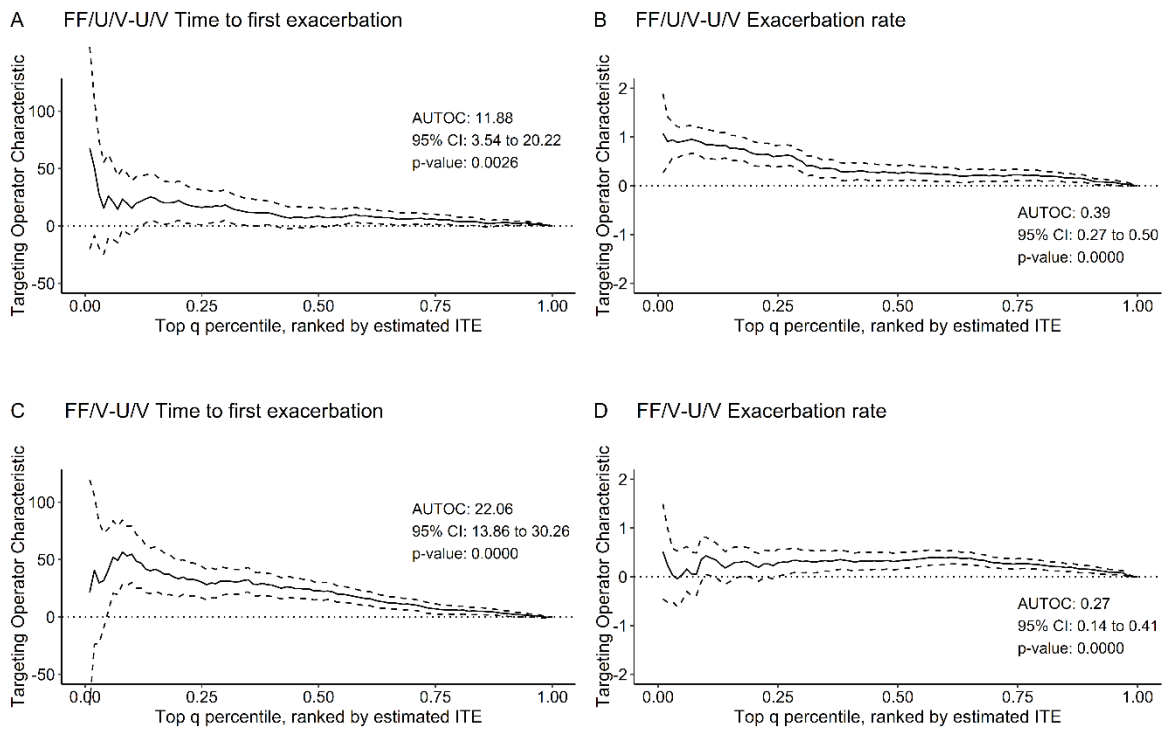


Figure S2 Area under the Targeting Operator Characteristic (AUTOC) curves for rankings based on estimated individualized treatment effect (ITE). In **A** and **B,** a triple therapy of fluticasone furoate (FF), umeclidinium (U) and vilanterol (V) was compared against a dual therapy of U/V. **C** and **D** compared dual therapy of FF/V against U/V. For **A** and **C,** models estimated individualized treatment effect for time to first exacerbation. For **B** and **D,** models estimated individualized treatment effect for exacerbation rate. Dashed lines indicate 95% confidence intervals (CI). P-values are from a one-sided test.

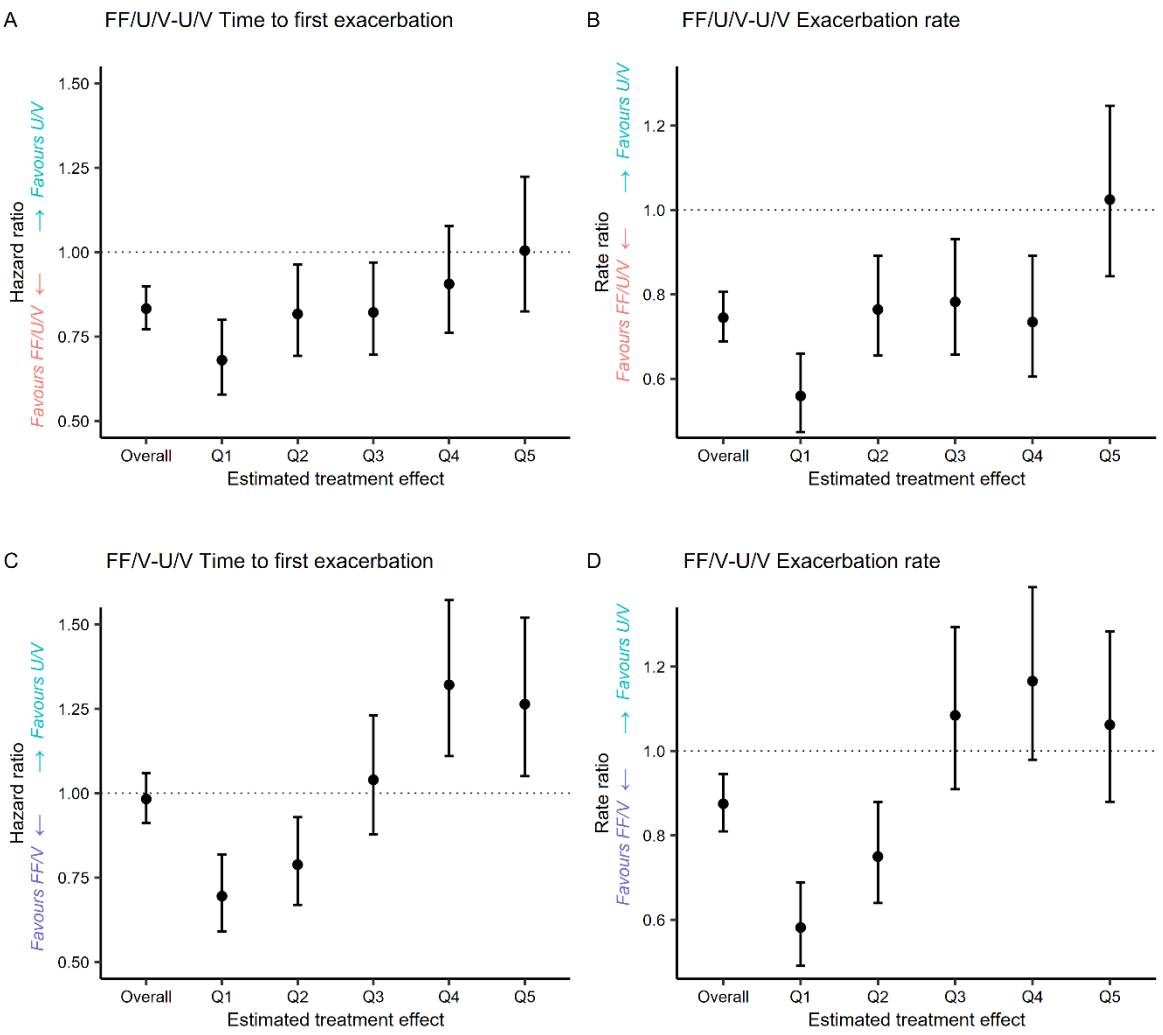


Figure S3 Observed rate and hazard ratios per quintile of estimated treatment effect. In **A** and **B,** a triple therapy of fluticasone furoate (FF), umeclidinium (U) and vilanterol (V) was compared against a dual therapy of U/V. **C** and **D** compared dual therapy of FF/V against U/V. Patients were ranked according to their estimated individualized treatment effect (ITE) and split into quintiles. Quintile 1 (Q1) contains patients estimated to benefit most from an inhaled corticosteroid-containing therapy (either FF/U/V or FF/V), while patients in quintile 5 (Q5) were estimated to benefit least from this therapy. For **A** and **C,** models estimated ITE for time to first exacerbation. For **B** and **D,** models estimated ITE for exacerbation rate. Error bars indicate 95% confidence intervals.

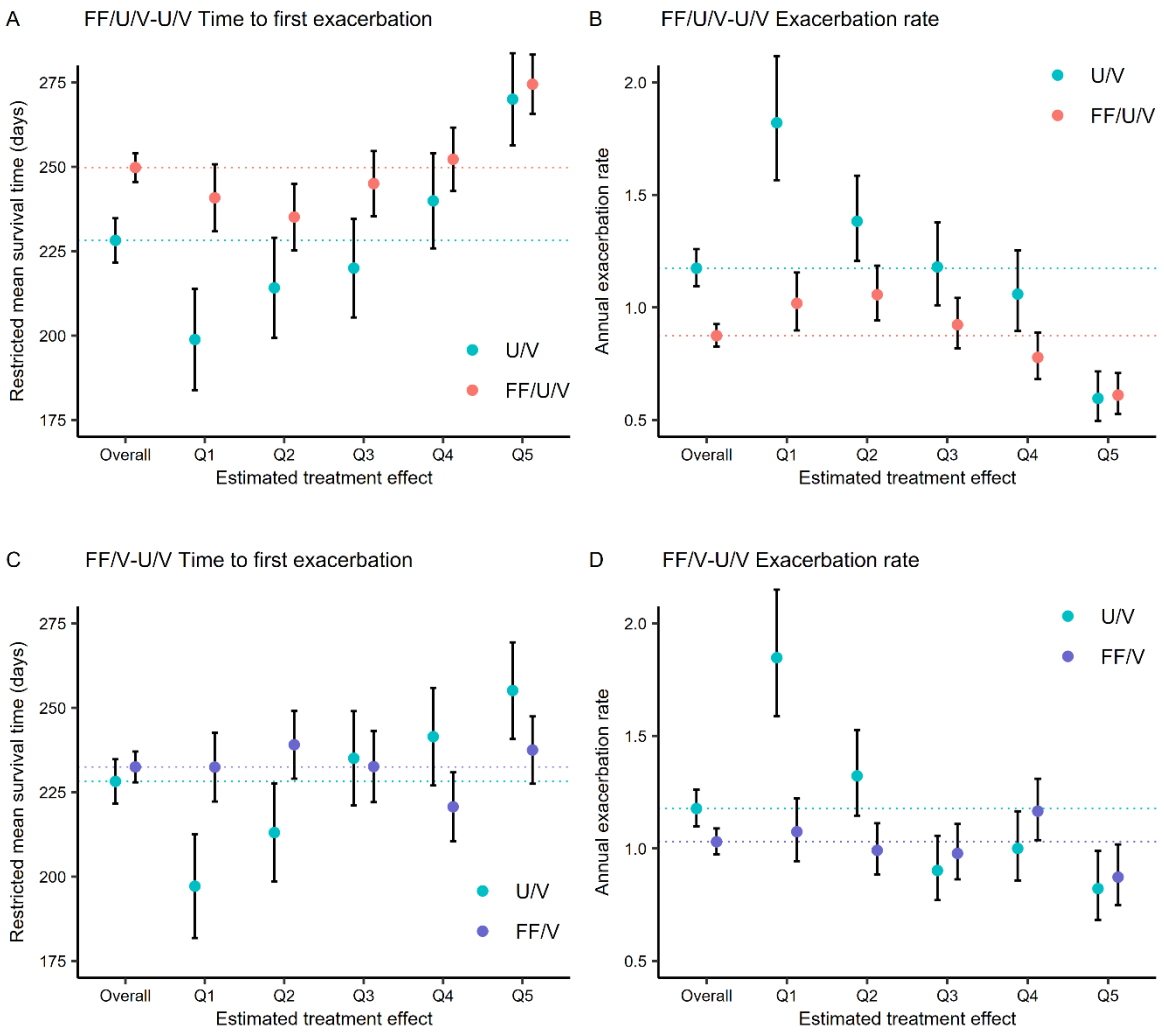


Figure S4 Observed mean exacerbation rates and restricted mean survival times per quintile of estimated treatment effect. In **A** and **B,** a triple therapy of fluticasone furoate (FF), umeclidinium (U) and vilanterol (V) was compared against a dual therapy of U/V. **C** and **D** compared dual therapy of FF/V against U/V. Patients were ranked according to their estimated individualized treatment effect and split into quintiles. Quintile 1 (Q1) contains patients estimated to benefit most from an inhaled corticosteroid-containing therapy (either FF/U/V or FF/V), while patients in quintile 5 (Q5) were estimated to benefit least from this therapy. For **A** and **C,** models estimated individualized treatment effect for time to first exacerbation. For **B** and **D,** models estimated individualized treatment effect for exacerbation rate. Error bars indicate 95% confidence intervals.

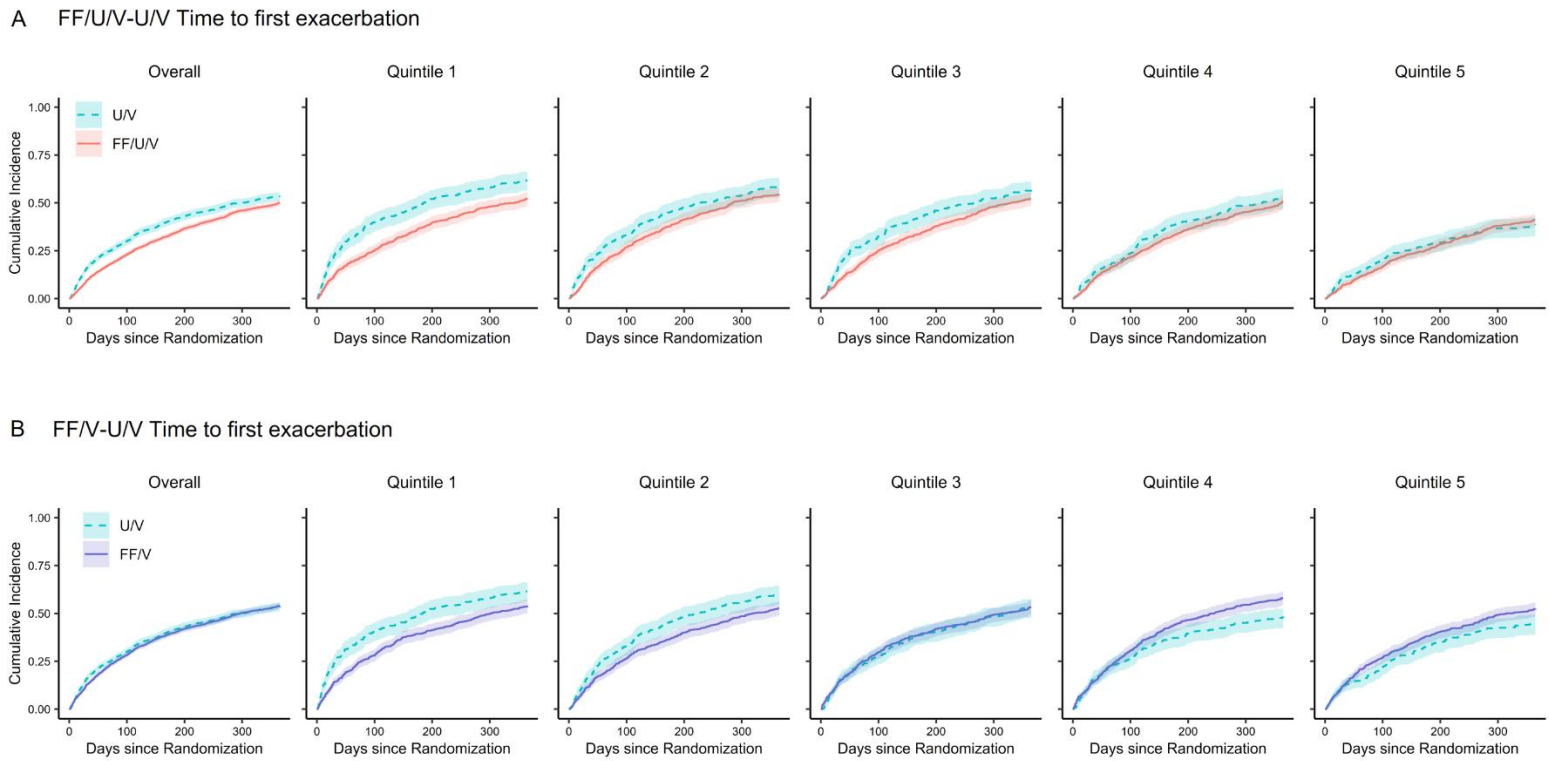


Figure S5 Observed cumulative incidence curves per quintile of estimated treatment effect. In **A,** a triple therapy of fluticasone furoate (FF), umeclidinium (U) and vilanterol (V) was compared against a dual therapy of U/V. **B** compared dual therapy of FF/V against U/V. Patients were ranked according to their estimated individualized treatment effect (ITE) and split into quintiles. Quintile 1 (Q1) contains patients estimated to benefit most from an inhaled corticosteroid-containing therapy (either FF/U/V or FF/V), while patients in quintile 5 (Q5) were estimated to benefit least from this therapy. Both **A** and **B** are results from models that estimated ITE for time to first exacerbation. Shaded area indicates 95% confidence intervals.

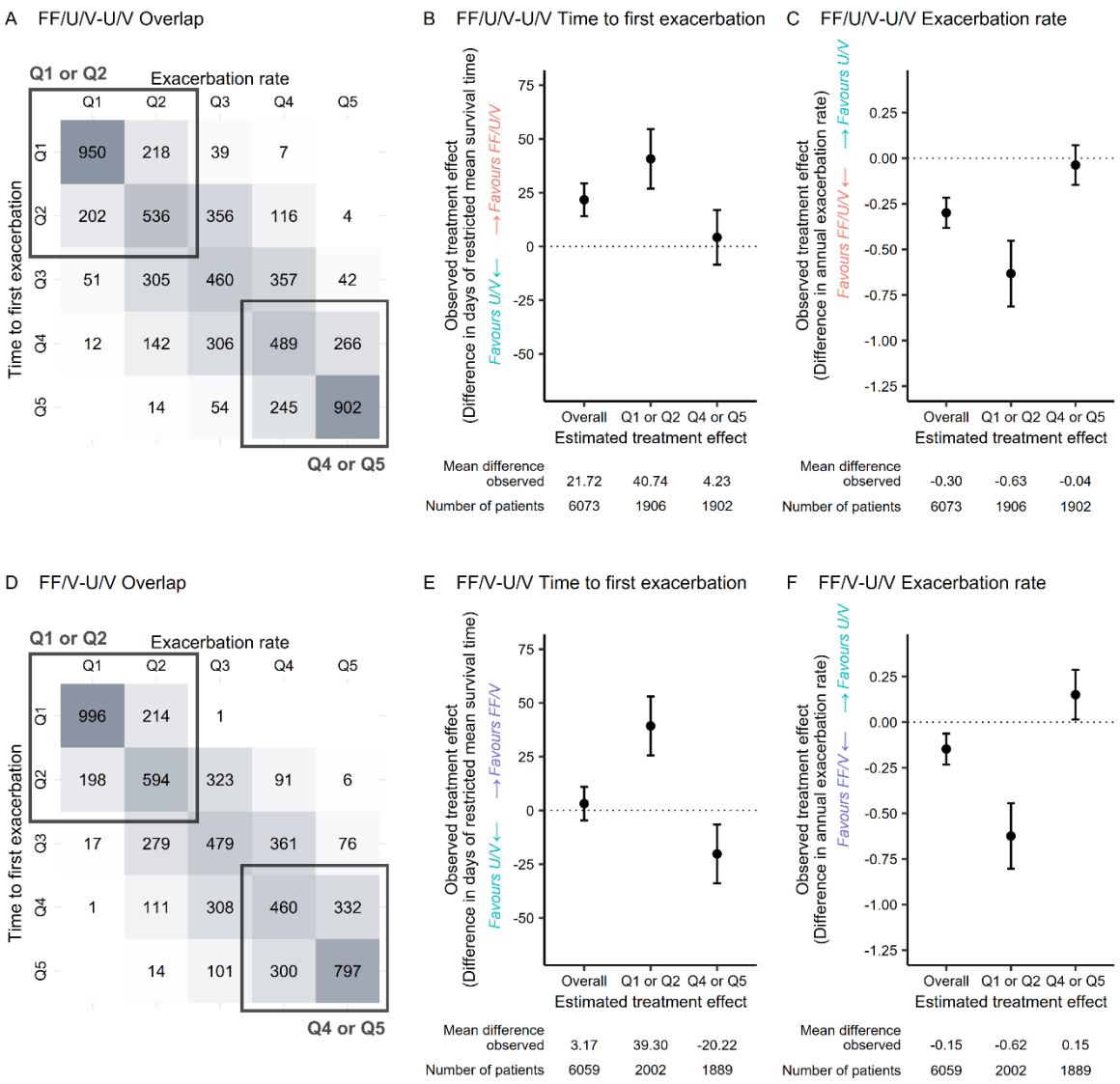


Figure S6 Overlap in patient rankings for time to first exacerbation and exacerbation rate, using a larger range of interest. In **A, B** and **C** a triple therapy of fluticasone furoate (FF), umeclidinium (U) and vilanterol (V) was compared against a dual therapy of U/V. **D, E** and **F** compared dual therapy of FF/V against U/V. Patients were ranked according to their estimated individual treatment effect and split into five quintiles. **A** and **D** illustrate overlap in patient rankings for both exacerbation outcomes with a confusion matrix. **B, C, E** and **F** visualize observed treatment effect for patients estimated to benefit most from inhaled corticosteroid-containing therapy (either FF/U/V or FF/V) according to both rankings (Q1 or

Q2), and for patients estimated to benefit least from this therapy according to both rankings (Q4 or Q5). Observed treatment effects are expressed as the difference in observed restricted mean survival time and the difference in observed mean annual exacerbation rate between the inhaled corticosteroid-containing therapy and U/V therapy. Error bars indicate 95% confidence intervals.

TRIPOD+AI

Version: 11-January-2024

For "Uncovering the heterogeneous effect of inhaled corticosteroids on COPD exacerbations with causal machine learning"
Huts H, Verstraete K, Strypsteen T, Gyselinck I, Staes M, Beersaerts A, Berrevoets J, Janssens W*, De Vos M* * Co-senior authors

Section/Topic	Item	Development / evaluation[1]	Checklist item	Reported on page
TITLE				
Title	1	D;E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	1
ABSTRACT				
Abstract	2	D;E	See TRIPOD+AI for Abstracts checklist	2, 3
INTRODUCTION				
Background	3a	D;E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	4-6
	3b	D;E	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)	4-6
	3c	D;E	Describe any known health inequalities between sociodemographic groups	4-6
Objectives	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	4-6
METHODS				
Data	5a	D;E	Describe the sources of data separately for the development and evaluation datasets (e.g., randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	5-8
	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	5-8
Participants	6a	D;E	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres	5-8
	6b	D;E	Describe the eligibility criteria for study participants	5-8
	6c	D;E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	5-8
Data preparation	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	7, 8 + supplement 4, 5
Outcome	8a	D;E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	5-8
	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	5-8
	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	5-8
Predictors	9a	D	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building	7
	9b	D;E	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	7 + supplement 11, 12
	9c	D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	NA
Sample size	10	D;E	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	6-8 + supplement 4-6, 24
Missing data	11	D;E	Describe how missing data were handled. Provide reasons for omitting any data	7 + supplement 4, 11, 12
Analytical methods	12a	D	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	7-9 + supplement 4-6, 24
	12b	D	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation).	supplement 4
	12c	D	Specify the type of model, rationale[2], all model-building steps, including any hyperparameter tuning, and method for internal validation	7-9 + supplement 4-8
	12d	D;E	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations[3]	NA
	12e	D;E	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	9 + supplement 6-8
	12f	E	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	NA
	12g	E	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application programming interface)	7-9 + supplement 4-8
Class imbalance	13	D;E	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	NA
Fairness	14	D;E	Describe any approaches that were used to address model fairness and their rationale	NA
Model output	15	D	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	7, 8 + supplement 4-6

[1] D=items relevant only to the development of a prediction model; E=items relating solely to the evaluation of a prediction model; D;E=items applicable to both the development and evaluation of a prediction model
[2] Separately for all model building approaches.
[3] TRIPOD-Cluster is a checklist of reporting recommendations for studies developing or validating models that explicitly account for clustering or explore heterogeneity in model performance (eg, at different hospitals or centres). Debray et al, BMJ 2023; 380: e071018 [DOI: 10.1136/bmj-2022-071018]

Training versus evaluation	16	D;E	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	7, 8 + supplement 5-6
Ethical approval	17	D;E	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent	7, 20
OPEN SCIENCE				
Funding	18a	D;E	Give the source of funding and the role of the funders for the present study	21
Conflicts of interest	18b	D;E	Declare any conflicts of interest and financial disclosures for all authors	20, 21
Protocol	18c	D;E	Indicate where the study protocol can be accessed or state that a protocol was not prepared	5, 6
Registration	18d	D;E	Provide registration information for the study, including register name and registration number, or state that the study was not registered	3, 5, 6
Data sharing	18e	D;E	Provide details of the availability of the study data	20
Code sharing	18f	D;E	Provide details of the availability of the analytical code[4]	Proprietary
PATIENT & PUBLIC INVOLVEMENT				
Patient & Public Involvement	19	D;E	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	No involvement
RESULTS				
Participants	20a	D;E	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	6-8
	20b	D;E	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups.	5-8 + supplement 4-6, 11, 12, 15-23
	20c	E	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome).	6-9 + supplement 4-7, 11,12
Model development	21	D;E	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)	6-9 + supplement 4-7
Model specification	22	D	Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or re-use (e.g., freely available, proprietary)[5]	Proprietary
Model performance	23a	D;E	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation.	10-13 + supplement 9, 13-14, 25-30
	23b	D;E	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD Cluster for additional details[3].	NA
Model updating	24	E	Report the results from any model updating, including the updated model and subsequent performance	NA
DISCUSSION				
Interpretation	25	D;E	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	14-18
Limitations	26	D;E	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability	14-18
Usability of the model in the context of current care	27a	D	Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model	NA
	27b	D	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	NA
	27c	D;E	Discuss any next steps for future research, with a specific view to applicability and generalizability of the model	16-18

From: Collins GS, Moons KGM, Dhiman P, et al. *BMJ* 2024;385:e078378. doi:10.1136/bmj-2023-078378

[4] This relates to the analysis code, for example, any data cleaning, feature engineering, model building, evaluation.
[5] This relates to the code to implement the model to get estimates of risk for a new individual.