OPEN ACCESS

Original research

Estimating individual treatment effects on COPD exacerbations by causal machine learning on randomised controlled trials

Kenneth Verstraete ,[1,2] Iwein Gyselinck ,[1] Helene Huts ,[1,2] Nilakash Das ,[1] Marko Topalovic ,[3] Maarten De Vos ,[2,4] Wim Janssens [1]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/thorax-2022-219382).

[1]Laboratory of Respiratory Diseases and Thoracic Surgery (BREATHE), KU Leuven, Leuven, Belgium
[2]STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium
[3]ArtiQ, Leuven, Belgium
[4]Department of Development and Regeneration, KU Leuven, Leuven, Belgium

Correspondence to
Professor Wim Janssens, Laboratory of Respiratory Diseases and Thoracic Surgery (BREATHE), KU Leuven, Leuven 3000, Belgium; wim.janssens@uzleuven.be

Received 30 June 2022
Accepted 13 March 2023
Published Online First
3 April 2023

Check for updates

© Author(s) (or their employer(s)) 2023. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Verstraete K, Gyselinck I, Huts H, *et al*. *Thorax* 2023;**78**:983–989.

ABSTRACT

Rationale Estimating the causal effect of an intervention at individual level, also called individual treatment effect (ITE), may help in identifying response prior to the intervention.

Objectives We aimed to develop machine learning (ML) models which estimate ITE of an intervention using data from randomised controlled trials and illustrate this approach with prediction of ITE on annual chronic obstructive pulmonary disease (COPD) exacerbation rates.

Methods We used data from 8151 patients with COPD of the Study to Understand Mortality and MorbidITy in COPD (SUMMIT) trial (NCT01313676) to address the ITE of fluticasone furoate/vilanterol (FF/VI) versus control (placebo) on exacerbation rate and developed a novel metric, Q-score, for assessing the power of causal inference models. We then validated the methodology on 5990 subjects from the InforMing the PAthway of COPD Treatment (IMPACT) trial (NCT02164513) to estimate the ITE of FF/umeclidinium/VI (FF/UMEC/VI) versus UMEC/VI on exacerbation rate. We used Causal Forest as causal inference model.

Results In SUMMIT, Causal Forest was optimised on the training set (n=5705) and tested on 2446 subjects (Q-score 0.61). In IMPACT, Causal Forest was optimised on 4193 subjects in the training set and tested on 1797 individuals (Q-score 0.21). In both trials, the quantiles of patients with the strongest ITE consistently demonstrated the largest reductions in observed exacerbations rates (0.54 and 0.53, $p<0.001$). Poor lung function and blood eosinophils, respectively, were the strongest predictors of ITE.

Conclusions This study shows that ML models for causal inference can be used to identify individual response to different COPD treatments and highlight treatment traits. Such models could become clinically useful tools for individual treatment decisions in COPD.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ The effect of a treatment can be heterogeneous over the target population. Identifying the subjects that respond best or worst to the treatment is a problem of causal inference. Machine learning models for causal inference have been proposed, but none have been successfully applied to data sets from large randomised controlled trials.

WHAT THIS STUDY ADDS

⇒ This study is the first to apply causal inference machine learning models to two of the largest randomised controlled trials in chronic obstructive pulmonary disease (COPD) with over 10 000 participants. This study offers a methodology that can be generalised to any randomised controlled trial and provides a Q-score to assess its potential.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ With the proposed methodology, subgroups of responders and non-responders can be identified on an individual basis in data sets from randomised controlled trials. This approach can be used for appropriate patient selection when screening for clinical trials or to guide treatment decisions in daily practice.

INTRODUCTION

Heterogeneity in the effects of clinical interventions is a major issue in healthcare. Treatments are usually proven on group level in randomised controlled trials (RCT) in which the studied treatment is compared against a placebo treatment or the current standard of care. Although most of the RCTs carefully delineate the group of eligible subjects, the primary analysis executed on group level will average individual treatment effects (ITE), thereby neglecting individual variation. Treatment effects on group level are often driven by a small subgroup of subjects who benefit greatly (responders), while other subgroups might experience no effect or even harm from the treatment. To optimise therapeutic decisions and restrict therapy to patients who truly benefit from the intervention, a more individualised approach is required (figure 1).

To understand treatment effects on individual level, we need to demonstrate the causal effect of the treatment on either the benefit, non-response or harm experienced by the subject. For treatment effect predictions, we cannot resort to standard statistical models that only deal with correlation.[1] While such models might be able to learn and predict high correlations between treatment administration and better outcomes in single subjects, they cannot conclude that administering the treatment caused a better outcome as other underlying factors

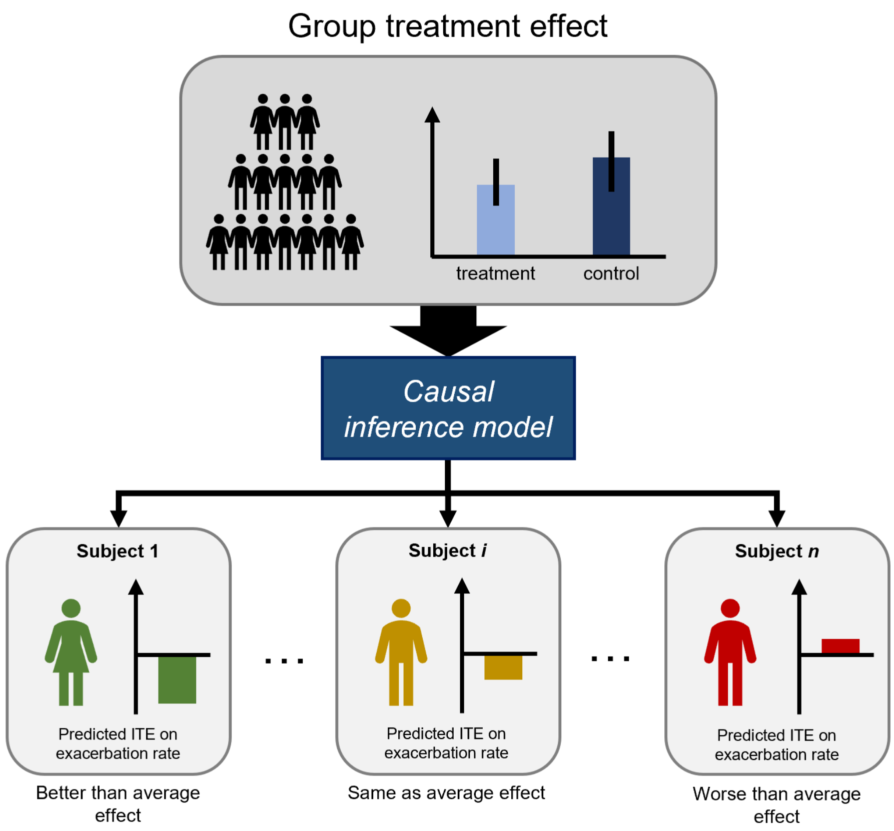
Group treatment effect
treatment
control
Causal inference model
Subject 1
Subject *i*
Subject *n*
Predicted ITE on exacerbation rate
...
Predicted ITE on exacerbation rate
...
Predicted ITE on exacerbation rate
Better than average effect
Same as average effect
Worse than average effect

Figure 1 Estimating individual treatment effect (ITE) for exacerbation rate (lower is better). There is a small overall treatment effect on group level, that is, the mean exacerbation rate in the treatment group was significantly lower compared with the control group. The aim of the ITE model is to rank subjects from strongest to lowest predicted ITE, thereby identifying those with benefit, non-response or harm from the treatment.

may drive the association. Several statistical frameworks have been proposed to deal with causality: graphical models by Pearl,[2] structural equations by Wright,[3] Haavelmo[4] and Heckman and Pinto,[5] and the potential outcome framework by Splawa-Neyman and Dabrowska[6] and Rubin.[7 8] The Neyman-Rubin model is based on the idea of potential outcomes. In the context of treatments, there are two potential outcomes; one with treatment and one without treatment. The ITE of a subject is defined as the difference in both potential outcomes. In reality, for each subject, one potential outcome is observed (factual outcome) while the other is missing (counterfactual outcome), as a subject cannot be simultaneously treated and not treated. Because of this fundamental problem of causal inference, the validation of ITE predictions remains a major obstacle. Data from large RCTs carry the potential to help develop causal inference models, as for every individual receiving the intervention, a matched subject or virtual twin may be present in the control group.

In chronic obstructive pulmonary disease (COPD), it is critical to understand how inhalation therapy (claimed to reduce exacerbations on group level) affects the individual exacerbation rate in subjects hosting a combination of positive and negative individual traits. We, therefore, used the data from the Study to Understand Mortality and MorbidITy in COPD (SUMMIT),[9] one of the largest RCTs in COPD, designed to address the role of combination therapy fluticasone furoate/vilanterol (FF/VI) on mortality and annual (moderate or severe) exacerbation rate. We used machine learning-based causal inference models to learn the causal effect of the treatment on baseline-characterised subjects from the FF/VI and placebo arms and devised a new evaluation metric to measure the performance of causal inference models. We validated the optimal model in an independent test set in SUMMIT. We then validated the methodology itself on the data from the InforMing the PAthway of COPD Treatment (IMPACT) study.[10]

MATERIALS AND METHODS

Study design and population

The SUMMIT study (NCT01313676) was a phase III, multicentre, randomised, double-blind, event-driven, placebo-controlled trial. Subjects (n=16 485) were randomised (1:1:1:1) to receive either inhaled placebo (n=4142), the corticosteroid fluticasone furoate (FF, n=4158), long-acting β_2 agonist vilanterol (VI, n=4146) or the combination FF/VI (n=4144). The study design and main results have been published.[9 11] Annual exacerbation rate was a secondary outcome of the study. Subjects administered FF/VI experienced a significant reduction in exacerbations per year compared with those administered placebo (FF/VI 0.25 vs placebo 0.35).[12]

The IMPACT study (NCT02164513) was a phase III, multicentre, randomised, double-blind, parallel-group, multicentre trial. The 10 355 patients with COPD were randomised (1:2:2) to receive umeclidinium/vilanterol (UMEC/VI, n=2070), FF/VI (n=4134) or FF/UMEC/VI (n=4151). The primary outcome was the annual rate of moderate or severe COPD exacerbations. Triple therapy resulted in a lower rate (0.91) compared with FF/VI (1.07) and UMEC/VI (1.21). Laboratory test biomarkers (chemistry, haematology) were recorded for each participant.

Table 1 Demographic characteristics of the subjects in SUMMIT

	SUMMIT (n=16 485)
Age, years	65±8
Women	4196 (25%)
BMI, kg/m^2	28±6
Smoking status	
Current smoker	7678 (47%)
Smoking history (pack-years)	41±24
Respiratory status	
Postbronchodilator FEV1 (L)	1.7 (0.4)
Predicted postbronchodilator FEV1 (%)	60±6
Exacerbations in 12 months before study	
0	10 021 (61%)
1	4020 (24%)
2+	2444 (15%)
Cardiovascular inclusion criteria	
Manifest disease	
Coronary artery disease	8379 (51%)
Peripheral arterial disease	3145 (19%)
Previous stroke	1595 (10%)
Previous myocardial infarction	2774 (17%)
Diabetes with target organ disease	1503 (9%)
At risk	
Hypercholesterolaemia	8479 (51%)
Hypertension	11 478 (70%)
Diabetes mellitus	3480 (21%)
Peripheral arterial disease	1154 (7%)
Concomitant medications	
Antiplatelet	8517 (52%)
Beta-blocker	5667 (34%)
ACE inhibitor	7655 (46%)
Statin	10 721 (65%)
Long-acting muscarinic antagonist	818 (5%)
Xanthine (including theophylline)	3719 (23%)
Treatment allocation	
Placebo	4111 (25%)
Fluticasone furoate	4135 (25%)
Vilanterol	4118 (25%)
Combination therapy	4121 (25%)

Data are presented as mean ± SD or number (%).
ACE, angiotensin-converting enzyme; BMI, body mass index; FEV1, forced expiratory volume in 1 s; SUMMIT, Study to Understand Mortality and Morbidity in COPD.

Q-score performance metric

The true ITE of an individual is never observed. Hence, ITE predictions can never be validated in the classic way of comparing predictions against ground truth labels. We, therefore, devised our own novel metric that validated our models in a quantile-based and interpretable way called the Q-score. The rationale behind the Q-score is that the observed average treatment effect (ATE) in a subgroup of treated and untreated subjects with similar predicted ITE should be similar to the average predicted ITE in that subgroup. The score compares the predictions of the model against the overall ATE and ranges between $-\infty$ and 1. A Q-score of zero indicates that the model works as well as predicting the ATE for each subject. A strictly positive Q-score means there is heterogeneity in the treatment effect found by the model and that the model predictions are better than predicting ATE for each subject. A negative Q-score indicates that the model did not find any heterogeneity and performs worse than the ATE. More details on the Q-score are found in online supplement.

Methodology

To focus on the treatment effects of FF/VI on annual exacerbation rate, we used the data from the placebo and FF/VI arms in SUMMIT. Parameters with too many missing values (>20%) or considered clinically irrelevant by experienced clinicians were discarded from the analysis. A full list of parameters used in the analyses is provided in online supplement. We split the data set into a training set and independent test set (70:30) by stratifying on the outcome, time on treatment and the number of treated and untreated subjects in both sets. We used the Causal Forest as causal inference model, a specific implementation of the Generalised Random Forest designed to identify heterogeneity in treatment effects.[13] Grid search and threefold cross-validation were used on the training set to optimise model hyperparameters, and the optimal model was then independently validated on the test set. The outputs of the model were predicted ITEs; a predicted ITE <0 meaning a predicted reduction in exacerbation rate with FF/VI compared with the placebo group, for example, an ITE of −0.5 meaning a reduction of 0.5 exacerbations per year. We used the same methodology for IMPACT as for SUMMIT using the UMEC/VI and FF/UMEC/VI arms to assess the added effect of FF on top of UMEC/VI on individual level.

To illustrate the heterogeneity identified by the model and assess whether the model could identify a group of super-responders, we ranked the test set subjects according to their predicted ITE and stratified them into five quantiles (quintiles) that could be easily compared (for both SUMMIT and IMPACT); quintile 1 containing the 20% subjects (from both control and treated arms) predicted to benefit the most, quintile 5 consisting of the 20% subjected predicted to benefit the least. The quantile approach was necessary as validating the estimates on individual level was limited by the fundamental problem of causal inference. For IMPACT, we analysed the time on treatment per quintile and the effect of eosinophil levels on ITE.

Data description and statistical analysis

We performed descriptive statistics on demographic, smoking, spirometric, exacerbation history, cardiovascular risk (only SUMMIT) and concomitant medication parameters of the subjects in SUMMIT and IMPACT. The data are presented as mean±SD, median (Q1–Q3 IQR) or counts and percentages. Missing data were handled during cross-validation using multivariate imputation by chained equations.[14 15] Exacerbation rates between treated and untreated subjects in quantiles were compared with negative binomial analyses using time on treatment as offset variable. The Q-score was used as evaluation metric during cross-validation. Parameter importance on group level was obtained directly from the Causal Forest. Shapley Additive Explanations (SHAP) was used on the final model to generate explanations on individual level to provide a personalised analysis per subject.[16 17] Python 3.8 was used for the entire analysis. EconML[18] by Microsoft Research and Scikit-learn[19] were used for the machine learning models and statsmodels[20] for

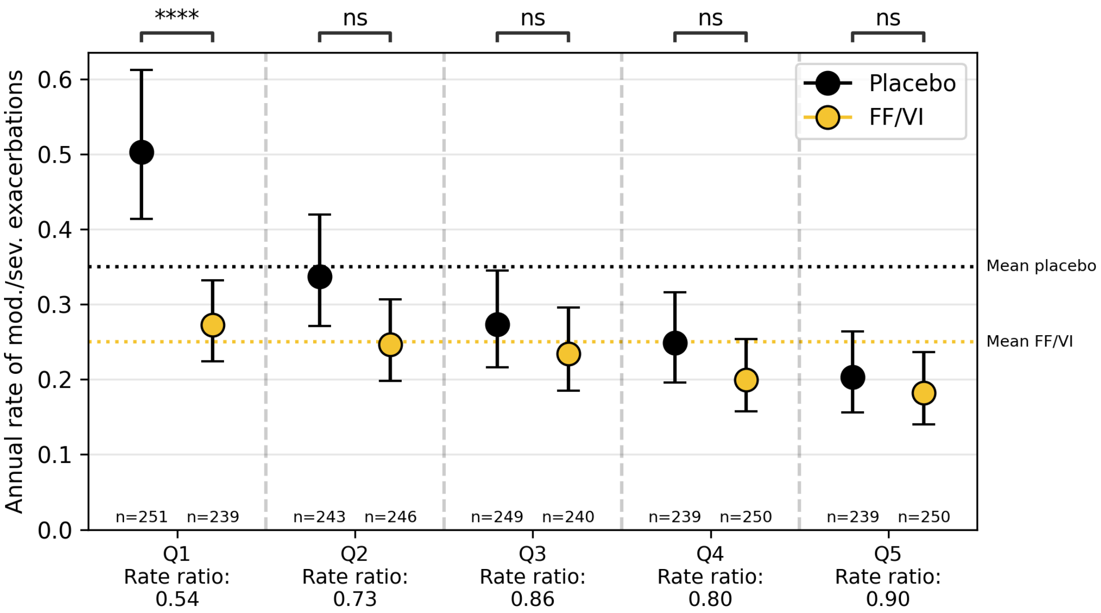

Figure 2 Comparison of annual exacerbation rate between the treated (fluticasone furoate/vilanterol, FF/VI) and untreated (placebo) subjects in the test set in SUMMIT per each quantile in the five quantile discretisation. Negative binomial analysis was used to compare the subjects per quantile. SUMMIT, Study to Understand Mortality and Morbidity in COPD.

the statistical analysis. All p values were two sided and significance level was set at 0.05.

RESULTS

Population characteristics

The characteristics of the participants in SUMMIT are found in table 1. The characteristics of the participants in IMPACT are found in online supplemental table S1.

Development and internal validation results

The training set consisted of 5705 subjects (2857 FF/VI, 2848 placebo). The internal test set consisted of 2446 subjects (1225 FF/VI, 1221 placebo). Based on Q-score, the best model was selected during threefold cross-validation with grid search (parameter grid is provided in the supplement, online supplemental table S3) and then tested on the internal test set. ITE was predicted for all subjects in the test set; the Q-score was 0.61.

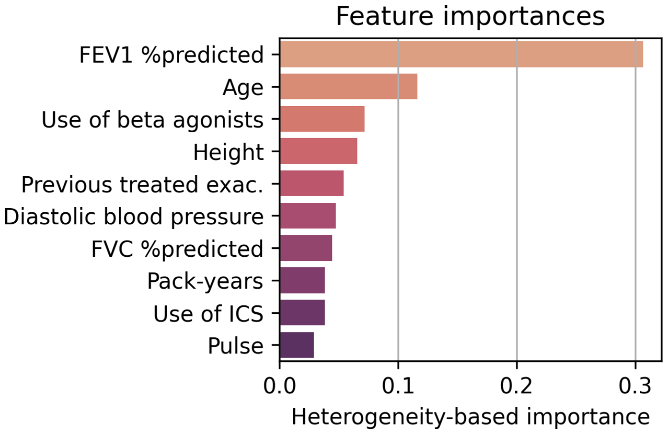

Figure 3 Feature importances of the Causal Forest model in SUMMIT. The importances are a heterogeneity-based measure. FEV1, forced expiratory volume in 1 s; FVC, forced vital capacity; ICS, inhaled corticosteroids.

For interpretation and analysis of the heterogeneity, we describe the five quantile discretisation (figure 2). Two quantile to five quantile discretisation are reported in online supplement. The subjects were sorted on predicted treatment effect in ascending order and split into quantiles. When split into five quantiles, there was a significant reduction in quantile 1 (0.50 placebo vs 0.27 FF/VI, rate ratio 0.54, $p<0.001$), no significant reductions in quantile 2 (0.34 placebo vs 0.25 FF/VI, rate ratio 0.73, $p=0.056$), quantile 3 (0.27 placebo vs 0.23 FF/VI, rate ratio 0.86, $p=0.37$). quantile 4 (0.25 placebo vs 0.20 FF/VI, rate ratio 0.80, $p=0.22$) and quantile 5 (0.20 placebo vs 0.18 FF/VI, rate ratio 0.90, $p=0.56$). To obtain insight into the subject characteristics of the different quantiles, we describe the mean values per quintile in online supplemental table S2. We found no significant differences between the characteristics of treated and non-treated subjects per quantile, indicating that treatment allocation did not determine the quantiles of ITEs. Height, weight, BMI, pack-years and forced expiratory volume in 1 s (FEV1) %predicted increased from quantile 1 to quantile 5, while previous exacerbations, use of β-agonists and use of inhaled corticosteroids (ICS) decreased from quantile 1 to quantile 5.

Global feature importances in SUMMIT

The global feature importances could be directly computed from the Causal Forest model and were based on how much treatment effect heterogeneity they created. The most important parameter on group level was FEV1% predicted, followed by age, use of β-agonists, height, number of previous treated COPD exacerbations in the previous year, diastolic blood pressure, FVC% predicted, pack-years, use of ICS and pulse rate (figure 3).

Personalised analysis using SHAP in SUMMIT

Two examples of personalised analyses for SUMMIT are found in online supplemental. Analysis 1 describes a subject who was predicted to be a non-responder (ITE close to 0), while analysis 2 describes a subject with a predicted ITE of −0.43, that is, a super-responder predicted to experience 0.43 exacerbations

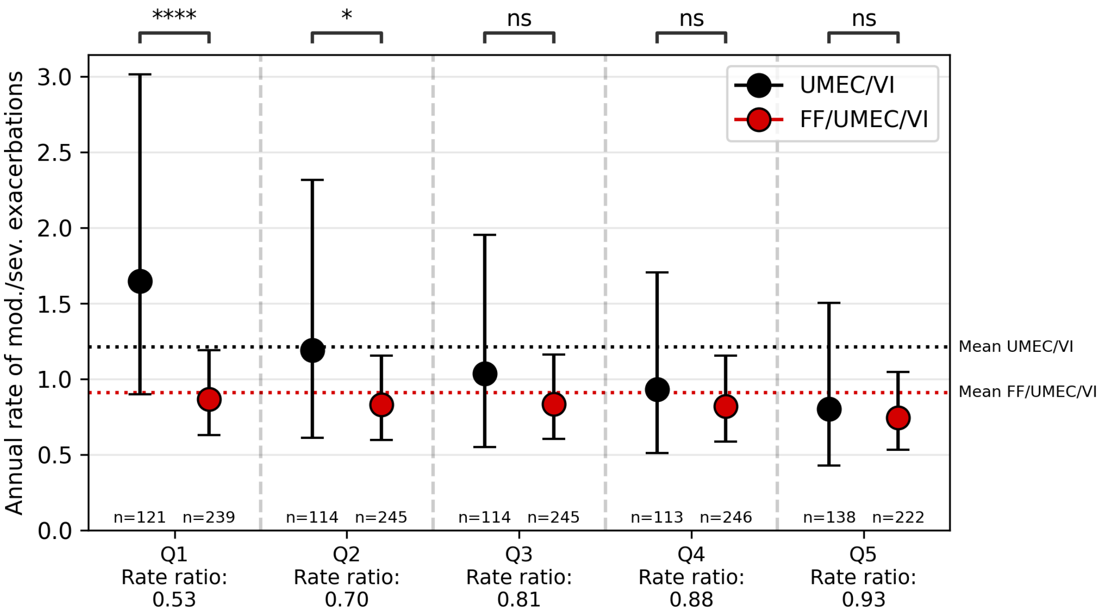

Figure 4 Comparison of annual exacerbation rate between the triple therapy (fluticasone furoate/umeclidinium/vilanterol, FF/UMEC/VI) arm and LABA/LAMA arm (umeclidinium/vilanterol, UMEC/VI) subjects in the test set in IMPACT per each quantile in the five quantile discretisation. Negative binomial analysis was used to compare the subjects per quantile. IMPACT, InforMing the PAthway of COPD Treatment.

per year less with the FF/VI treatment. In both analyses, it is highlighted which parameters were most important for each prediction.

Validation in IMPACT

The training set consisted of 4193 subjects (2792 FF/UMEC/VI, 1401 UMEC/VI) and the test set of 1797 subjects (1197 FF/UMEC/VI, 600 UMEC/VI). The Q-score on the test set in IMPACT was 0.21. Analogous to SUMMIT, the five quantile discretisation is illustrated in figure 4 and the full analysis is found in online supplemental. The subject characteristics per quantile are described in online supplemental table S3.

As in SUMMIT, the global feature importances were directly extracted from the optimal Causal Forest model. The dominant parameters were eosinophils/leucocytes (%) and eosinophils, followed by years smoked, FEV1 reversibility, FEV1/FVC, systolic blood pressure, diastolic blood pressure, total COPD assessment test score, neutrophils and FEV1 %predicted (figure 5).

As we noticed a selective dropout for the UMEC/VI arm, we analysed the on-treatment time per quintile (figure 6). For each quintile, the FF/UMEC/VI subjects had a median time of 1 with IQR 0.02. For the UMEC/VI subjects, the on-treatment time decreased per quintile. In quintile 1, the median time was 0.99 with IQR 0.62, in quintile 2, the median time was 1 with IQR 0.43, and in quintile 3, the median time was 1 with IQR 0.24. In both quintiles 4 and 5, the median time was 1 with IQR 0.02.

Analysis of eosinophils in IMPACT

Using the SHAP methodology, the influence of eosinophil levels on ITE (SHAP values) for each individual was calculated. Figure 7 illustrates the SHAP values in function of eosinophil levels for each individual in the test set of IMPACT. A grey line

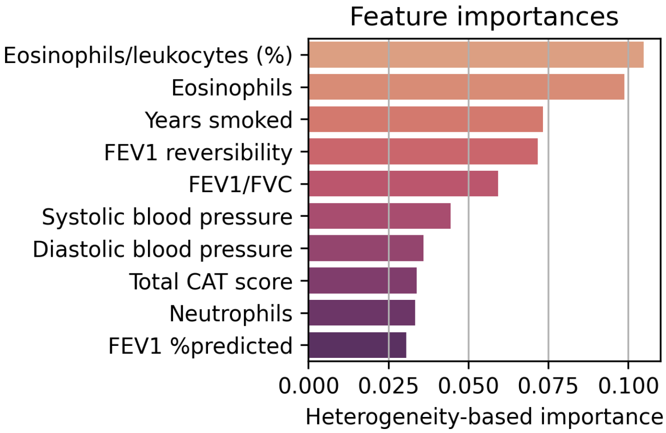

Figure 5 Feature importances of the Causal Forest model in IMPACT. The importances are a heterogeneity-based measure. CAT, COPD assessment test; FEV1, forced expiratory volume in 1 s; FVC, forced vital capacity; IMPACT, InforMing the PAthway of COPD Treatment.

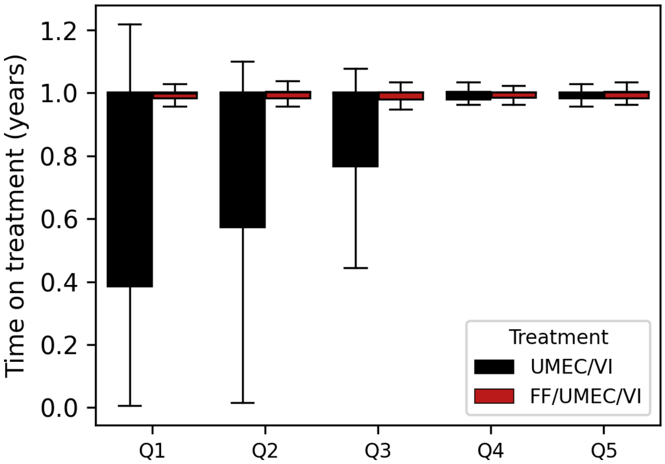

Figure 6 The on-treatment times per quintile for the LABA/LAMA (umeclidinium/vilanterol, UMEC/VI) and triple therapy (fluticasone furoate/umeclidinium/vilanterol, FF/UMEC/VI) arms in IMPACT. IMPACT, InforMing the PAthway of COPD Treatment.

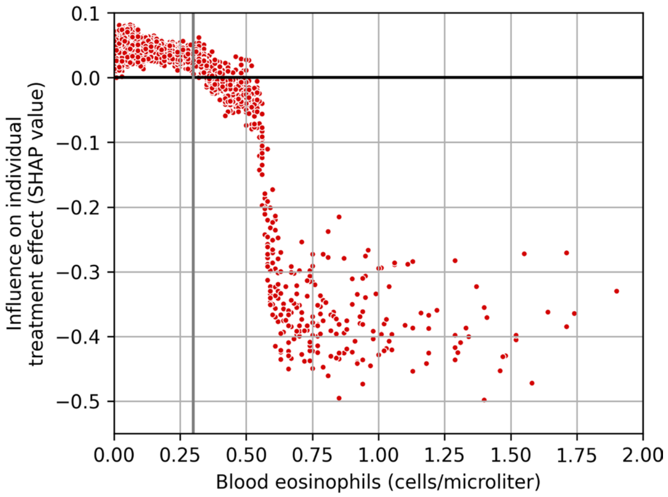

Figure 7 Influence of eosinophils (SHAP values) on the individual treatment effect predictions in the test set of IMPACT. Positive SHAP values are indicating a harm and negative values are indicating a benefit of the intervention. The 300 cells/µL cut-off is indicated by the grey line. IMPACT, InforMing the PAthway of COPD Treatment; SHAP, Shapley Additive Explanations.

indicates the clinical cut-off of 300 cells/µL. A positive SHAP value suggests the eosinophil level to have an increasing influence on exacerbation rate (worse). A negative SHAP value points towards a decreasing influence on exacerbation rate (better).

DISCUSSION

We used the data from subjects of the placebo and FF/VI arm in the SUMMIT study in COPD to develop a methodology and machine learning model that estimates the ITE of FF/VI on exacerbations based on baseline characteristics. Compared with classic prediction models, two potential outcomes are estimated in causal inference, either implicitly or explicitly, and their difference is the ITE or the actual prediction of the model. One of the major hurdles to overcome is the inability to truly validate ITE predictions since the true effect of any individual is never observed. As metrics from literature, such as the concordance-statistic-for benefit,[21] Qini curve and coefficient[22] and Alaa's influence function-based loss function[23] had several shortcomings or were difficult to interpret for clinicians, we designed our own metric: the Q-score. This score quantifies the fit between predicted ITE and observed treatment effect for all possible quantile splits. We identified ±20% of the patient population as strong responders to the FF/VI intervention with rate ratio 0.54 compared with the overall rate ratio of 0.71. FEV1 %predicted was found to be the highly dominant parameter for individual FF/VI effect estimations. Age, use of β-agonists, height and history of treated exacerbations were other important parameters, although substantially lower in magnitude compared with FEV1 %predicted (figure 3).

When analysing each quantile separately, we found that exacerbation rates between treated and untreated subjects in quintile 1 (the 20% subjects predicted to benefit the most) were reduced by 46%. In particular, the observed exacerbation rate in this group was reduced from 0.50 to 0.27 by the treatment, higher than the average reduction of 0.35 to 0.25 (29% reduction) in the SUMMIT trial itself. In quintile 2, a similar rate ratio as the overall rate ratio was observed (0.73, p=0.056) while in quintile 3, 4 and 5, we observed no significant effect and rate ratio was increased. These data indicate that our methodology was able to identify responders and vice versa, non-responders in SUMMIT.

The five most important features from the Causal Forest model were FEV1% predicted, age, use of β-agonists, height and number of previous COPD exacerbations. We used SHAP to generate prediction explanations per individual in order to improve clinical interpretation. It also allowed us to examine on group level how each parameter influenced predictions over all subjects in the data set. Lower lung function was found to be the main determinant of treatment response, as patients in the first quantile had a 10% point reduction of FEV1. It not only aligns with previous findings that FEV1 is a main determinant of exacerbation frequency but also indicates that those subjects with low FEV1 are more amendable to FF/VI intervention.[24]

The analysis on SUMMIT had several limitations. First, because RCTs restrict the intervention to subjects within stringent inclusion criteria, it reduces the generalisability of our prediction model to a real-life population. Second, the SUMMIT trial was an event-driven study. Consequently, each of the participants had a different number of follow-up visits, potentially introducing bias in the analyses. Third, SUMMIT did not focus on frequent exacerbators, and subsequently, observed exacerbation rates were low (0.25 in FF/VI and 0.35 in placebo arm). Last, SUMMIT was focused on cardiovascular parameters that are associated with cardiovascular death but provide less prediction power for future exacerbations. Clinical features that are known to be important predictors for exacerbation rates and determinants of ICS effects (white blood cell differentiation—eosinophils)[25] were not part of the SUMMIT data set.

Because of the limitations in SUMMIT, we validated the methodology on the data from the IMPACT trial focused on subjects with a history of exacerbations and with annual exacerbation rate as primary outcome. Blood was collected from the subjects, so blood parameters such as eosinophils were included. Contrary to SUMMIT, the control arm used in our analysis in IMPACT was not treated with placebo but with LABA/LAMA. As such, the ITE in IMPACT was defined as the added effect of ICS. Whereas SUMMIT was event driven and each subject had a different follow-up time, the follow-up time in IMPACT was fixed to 52 weeks.

The same model as in SUMMIT, Causal Forest, was applied to IMPACT. The triple therapy arm was two times the size of the LABA/LAMA arm, but this posed no problem for the model. The Causal Forest model confirmed that eosinophil levels were important predictors for treatment effects and was also able to identify subgroups of responders and non-responders in the trial. Less heterogeneity was detected compared with SUMMIT; the only quintile with significant treatment effects was the group with elevated blood eosinophil levels (online supplemental table S3), the other four quantiles showed no significant added effect of ICS. Completely in line with the literature, the model independently found the subgroup of eosinophil-based responders. Post hoc analysis with SHAP showed that approximately 300 eosinophils/µL is indeed the cut-off where many subjects seemed to experience benefit (figure 7). However, a large proportion in the responder group had <300 eosinophils/µL. The advantage of the machine learning model is the ability to identify subjects below that threshold as responders and subjects above that threshold as non-responders based on the other characteristics. The on-treatment time for the subjects in the triple therapy arm was the same in all quintiles. Remarkably, for subjects in the LABA/LAMA arm, the on-treatment time was lowest in quintile 1 and increased over the next two quintiles. In quintiles 4 and 5, the subjects with the least effects of FF and the lowest exacerbations rates, the on-treatment time was the same as their counterparts in the triple therapy arm. This finding may be explained by the fact that patients on LABA/LAMA who were experiencing more frequent exacerbations (and most likely presenting with high blood eosinophils) were taken of study medication for open-label triple therapy. To some extent, these subjects with reduced time in the trial were driving the overall treatment effect of the IMPACT trial.

To the best of our knowledge, this is the first study to explore causal inference models on large phase III RCTs in COPD. We proposed a new metric, the Q-score, for evaluating causal inference models based on an intuitive notion that is easily interpretable for clinical practice. In line with what was observed during the quantile analysis, there is more heterogeneity to be found in SUMMIT compared with IMPACT, as illustrated by the higher Q-score (SUMMIT V.0.61 vs IMPACT V.0.21) and narrower CIs. Together, our approach may represent a full generic methodology for developing causal inference models on RCT for any domain. Our models showed potential in finding subgroups in the target populations that might or might not benefit from a treatment. Apart from its impact on clinical practice, such models may affect clinical trial design by an optimised patient selection at study entry.

CONCLUSION

We described a general machine learning approach that can be used to identify individual treatment response using data from RCTs. We illustrated this methodology in the SUMMIT trial on the FF/VI treatment for exacerbations. We then validated the methodology on the IMPACT trial on the FF/UMEC/VI treatment for exacerbations. Trained on data from different treatments, such models may become useful tools for patient selection in clinical trials and personalised medicine in COPD.

Twitter Kenneth Verstraete @K_Verstraete

Contributors KV performed the analyses and drafted the manuscript. KV, IG, WJ interpreted the analyses clinically. KV, IG, HH, ND, MT and MDV assessed the methodology critically. KV, IG, HH, ND, MT, MDV and WJ critically revised the content. KV, IG, HH, ND, MT, MDV and WJ gave final approval of the manuscript. WJ acts as guarantor of the overall content and accepts full responsibility for the work and the conduct of the study, had access to the data, and controlled the decision to publish.

Funding This research received funding from the Flemish Government under the 'Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen' program, from AstraZeneca KU Leuven Chair in Respiratory Diseases (LSASZ7-O2010) and FWO Research Project:'Artificial Intelligence (AI) for data-driven personalised medicine' (G0C9623N). IG is funded by Research Foundation—Flanders (FWO, 11N3922N). WJ is supported as senior investigator of the Research Foundation—Flanders (FWO, 1800715N|1800720N).

Competing interests KV has nothing to disclose. IG receives personal funding from Research Foundation Flanders (FWO). HH has nothing to disclose. ND has nothing to disclose. MT is CEO and co-founder of ArtiQ but received no payments related to the manuscript. MDV received funding from the AI in Flanders project. WJ received grants from AstraZeneca and Chiesi and obtained fees from AstraZeneca, Chiesi and GlaxoSmithKline. He is chairman of Board of Flemish Society for TBC prevention and board member of ArtiQ.

Patient consent for publication Not applicable.

Ethics approval Approval was provided by the Institutional Privacy and Ethic review committee (SMEC/PRET) (ID: G-2020-2044).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The anonymised individual participant data set of SUMMIT (Sponsor ID HZC113782) and IMPACT (Sponsor ID CTT116855) were provided upon request by clinicalstudydatarequest.com with the approval of GlaxoSmithKline Research & Development Ltd. Protocols, reporting and analysis plans and clinical study reports are available for both studies.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Kenneth Verstraete http://orcid.org/0000-0003-3790-417X
Iwein Gyselinck http://orcid.org/0000-0002-4068-7228
Helene Huts http://orcid.org/0000-0002-0617-1316
Nilakash Das http://orcid.org/0000-0002-8960-7188
Marko Topalovic http://orcid.org/0000-0001-6101-755X
Maarten De Vos http://orcid.org/0000-0002-3482-5145
Wim Janssens http://orcid.org/0000-0003-1830-2982

REFERENCES

1 Holland PW. Statistics and causal inference. *J Am Stat Assoc* 1986;81:945–60.
2 Pearl J. Causal diagrams for empirical research. *Biometrika* 1995;82:669–88.
3 Wright S. The method of path coefficients. *Ann Math Statist* 1934;5:161–215.
4 Haavelmo T. The statistical implications of a system of simultaneous equations. *Econometrica* 1943;11:1.
5 Heckman J, Pinto R. Causal analysis after Haavelmo. *Econ Theory* 2015;31:115–51.
6 Splawa-Neyman J, Dabrowska DM, Speed TP. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist Sci* 1990;5:465–72.
7 Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974;66:688–701.
8 Rubin DB. *Matched sampling for causal effects*. Cambridge University Press, 2006: 1–489.
9 Vestbo J, Anderson J, Brook RD, *et al*. The study to understand mortality and morbidity in COPD (Summit) study protocol. *Eur Respir J* 2013;41:1017–22.
10 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;378:1671–80.
11 Vestbo J, Anderson JA, Brook RD, *et al*. Fluticasone furoate and vilanterol and survival in chronic obstructive pulmonary disease with heightened cardiovascular risk (summit): a double-blind randomised controlled trial. *Lancet* 2016;387:1817–26.
12 Martinez FJ, Vestbo J, Anderson JA, *et al*. Effect of fluticasone furoate and vilanterol on exacerbations of chronic obstructive pulmonary disease in patients with moderate airflow obstruction. *Am J Respir Crit Care Med* 2017;195:881–8.
13 Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Statist* 2019;47:1179–203.
14 Raghunathan T, Lepkowski J, Van Hoewyk J, *et al*. A multivariate technique for multiply imputing missing values using A sequence of regression models. *Surv Methodol* 2001;27:85–96.
15 van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* 2007;16:219–42.
16 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*. 2017: 4766–75.
17 Lundberg SM, Nair B, Vavilala MS, *et al*. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018;2:749–60.
18 Battocchi K, Dillon E, Syrgkanis V, *et al*. EconML: a python package for ML-based heterogeneous treatment effects estimation; 2019.
19 Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;12:2825–30.
20 Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. In: *Proceedings of the 9th Python in Science Conference*. 2010: 92–6.
21 van Klaveren D, Steyerberg EW, Serruys PW, *et al*. The proposed "concordance-statistic for benefit" provided a useful metric when modeling heterogeneous treatment effects. *J Clin Epidemiol* 2018;94:59–68.
22 Radcliffe NJ. Using control groups to target on predicted lift: building and assessing uplift model. *Direct Mark Anal J* 2007:14–21.
23 Alaa AM, van der Schaar M. Validating causal inference models via influence functions. In: *Proceedings of the 36th International Conference on Machine Learning*. 2019: 191–201.
24 Hurst JR, Vestbo J, Anzueto A, *et al*. Susceptibility to exacerbation in chronic obstructive pulmonary disease. *N Engl J Med* 2010;363:1128–38.
25 Barnes NC, Sharma R, Lettis S, *et al*. Blood eosinophils as a marker of response to inhaled corticosteroids in COPD. *Eur Respir J* 2016;47:1374–82.