Original Research

Machine Learning Prediction of Progression in Forced Expiratory Volume in 1 Second in the COPDGene® Study

Adel Boueiz1,2*, Zhonghui Xu1*, Yale Chang3, Aria Masoomi3, Andrew Gregory1, Sharon M. Lutz4, Dandi Qiao1, James D. Crapo5, Jennifer G. Dy3, Edwin K. Silverman1,2, Peter J. Castaldi1,6, for the COPDGene Investigators.

*Contributed equally

Author Affiliations:
1Channing Division of Network Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts, United States
2Pulmonary and Critical Care Division, Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts, United States
3Department of Electrical and Computer Engineering, Northeastern University, Boston, Massachusetts, United States
4Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, Massachusetts, United States
5Division of Pulmonary Medicine, Department of Medicine, National Jewish Health, Denver, Colorado, United States
6Division of General Medicine and Primary Care, Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts, United States.

Corresponding Author:
Adel Boueiz, MD, MMSc
Channing Division of Network Medicine
Brigham and Women’s Hospital
181 Longwood Avenue
Boston, MA 02115
Phone: 617-525-2111
Email: adel.boueiz@channing.harvard.edu

Citation: Boueiz A, Xu Z, Chang Y, et al. Machine learning prediction of progression in forced expiratory volume in 1 second (FEV1) in the COPDGene® study. Chronic Obstr Pulm Dis. 2022; Published online May 20, 2022. doi: https://doi.org/10.15326/jcopdf.2021.0275

Authors’ email addresses: Adel Boueiz (adel.boueiz@channing.harvard.edu), Zhonghui Xu (zhonghui.xu@channing.harvard.edu), Yale Chang (changyalee@gmail.com), Aria Masoomi (masoomi.a@northeastern.edu), Andrew Gregory (andrew.gregory@channing.harvard.edu), Sharon M. Lutz (sharon.m.lutz@gmail.com), Dandi Qiao (dandi.qiao@channing.harvard.edu), James D. Crapo (CrapoJ@njhealth.org), Jennifer G. Dy (jdy@ece.neu.edu), Edwin K. Silverman (ed.silverman@channing.harvard.edu), Peter J. Castaldi (peter.castaldi@channing.harvard.edu)
Keywords: Chronic obstructive pulmonary disease; Disease progression; 5-year changes in FEV₁; Prediction; Random forest machine learning.

Funding Sources: This work was supported by NHLBI K08 HL141601, R01 HL124233, R01 HL126596, R01 HL147326, U01 HL089897, and U01 HL089856. The COPDGene study (NCT00608764) is also supported by the COPD Foundation through contributions made to an Industry Advisory Board that has included AstraZeneca, Bayer Pharmaceuticals, Boehringer-Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer, and Sunovion.


Note: This article has an online data supplement.
ABBREVIATION LIST

AUC: Area under the curve
BMI: Body mass index
COPD: Chronic Obstructive Pulmonary Disease
COPDGene study: Genetic Epidemiology of COPD study
ΔFEV\textsubscript{1}: Annualized five-year changes in FEV\textsubscript{1}
FEV\textsubscript{1}: Forced expiratory volume in one second
FVC: Forced vital capacity
GOLD: Global Initiative for Chronic Obstructive Lung Disease spirometric grading system
HU: Hounsfield units
IQR: Interquartile range
%LAA-950: Percent of CT scan low attenuation area below -950 HU at end-inspiration
MMRC: Modified Medical Research Council
NHW: Non-Hispanic White
RF: Random forest
RMSE: Root mean squared error
ROC: Receiver operator characteristic
SGRQ: St. George’s Respiratory Questionnaire
ABSTRACT

Background: The heterogeneous nature of COPD complicates the identification of the predictors of disease progression. We aimed to improve the prediction of disease progression in COPD by using machine learning and incorporating a rich dataset of phenotypic features.

Methods: We included 4,496 smokers with available data from their enrollment and 5-year follow-up visits in the Genetic Epidemiology of COPD (COPDGene) study. We constructed linear regression (LR) and supervised random forest (RF) models to predict 5-year progression in FEV1 from 46 baseline features. Using cross-validation, we randomly partitioned participants into training and testing samples. We also validated the results in the COPDGene 10-year follow-up visit.

Results: Predicting the change in FEV1 over time is more challenging than simply predicting the future absolute FEV1 level. For RF, R-squared was 0.15 and the area under the ROC curves for the prediction of subjects in the top quartile of observed progression was 0.71 (testing) and respectively, 0.10 and 0.70 (validation). RF provided slightly better performance than LR. The accuracy was best for GOLD1-2 subjects and it was harder to achieve accurate prediction in advanced stages of the disease. Predictive variables differed in their relative importance as well as for the predictions by GOLD.

Conclusion: RF along with deep phenotyping predicts FEV1 progression with reasonable accuracy. There is significant room for improvement in future models. This prediction model facilitates the identification of smokers at increased risk for rapid disease progression. Such findings may be useful in the selection of patient populations for targeted clinical trials.

INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is the second leading cause of disability, the third leading cause of death, and the only major chronic disease continuing to
increase in mortality. Novel therapies that slow disease progression could result in an improvement in COPD patients’ health status and have a substantial impact on healthcare utilization. The development of such therapies will be aided by improved tools for predicting disease progression, enabling the selection of high-risk groups for targeted treatment.

Predictive models incorporate multiple sources of information to make patient-specific predictions and are widely used in multiple areas of medical practice. Existing models of disease progression in COPD have been limited in the scope of variables assessed. COPD exhibits significant variation in clinical and radiologic presentation as well as disease progression. This disease heterogeneity complicates the identification of the predictors of COPD progression and limits the accuracy of predictive models. Furthermore, COPD often progresses slowly over decades and true disease progression over short time periods can be difficult to detect with existing measurements.

In this study, we aimed to improve the prediction of COPD progression by applying machine learning to a rich dataset of clinical, demographic, patient-reported variables, and imaging features in the COPDGene study. We hypothesized that deep phenotyping at the initial study visit along with random forest modelling, which exploits complex non-linear relationships and interactions among the risk factors, would facilitate the prediction of the rates of disease progression as measured by FEV₁, a key aspect of COPD.

MATERIALS AND METHODS

Study populations

COPDGene Study (NCT00608764, www.copdgene.org). COPDGene is an ongoing multi-institutional longitudinal study to investigate the epidemiologic and genomic characteristics of COPD. COPDGene enrolled self-identified non-Hispanic whites (NHW) and African-Americans (AA) smokers across the full spectrum of disease severity as defined
by the Global Initiative for Chronic Obstructive Lung Disease (GOLD) spirometric staging system. Subjects were aged 45 to 80 years at study enrollment and had at least 10 pack-years of lifetime smoking history. COPDGene collects longitudinal data at five-year intervals; the 10-year study visit is ongoing. Visit 1 and Visit 2 were completed and Visit 3 is ongoing. At each study visit, subjects underwent comprehensive phenotyping, which included spirometry, questionnaire assessment, and inspiratory and expiratory chest CT scans, all of which were done according to a standard procedure with consistent quality control across centers.

**Derivation cohort - COPDGene Study Visit 1 and Visit 2:** We analyzed 4,496 smokers with complete CT scans and relevant covariate data at the baseline visit (Visit 1) and 5-year follow-up visit (Visit 2) in the COPDGene cohort (NCT00608764, www.copdgene.org).

**Temporal validation cohort - COPDGene Study Visit 3:** During the Phase 3 of the COPDGene Study, enrolled subjects returned for their 10-year follow-up visit. At the time of this analysis, 1,833 smokers had completed their 10-year follow-up visit and had available 10-year spirometric and radiologic data. To predict their outcome values at Year 10 (Visit 3), we entered their 5-year (Visit 2) predictor data into the models trained in the derivation cohort. The FEV₁ values for Visit 3 were observed. Our models were trained using only data from Visit 1 and Visit 2, where predictors were at Visit 1 and responses were Visit 2 values or the change in values between Visit 2 and Visit 1. In this setting, cross-validation was used to assess model performance. To provide further “temporal” validation of our models, we tested our already-trained models (no further parameter fitting) by using Visit 2 values for the predictors. This allowed us to compare the predicted Visit 3 values against the observed Visit 3 values to assess the accuracy of each prediction model in the temporal validation cohort.

The COPDGene study design, subject enrollment, and phenotype measurements have been previously reported and additional information is included in the Supplement.
Outcome variables

We constructed models to predict annualized follow-up FEV$_1$ and five-year changes in FEV$_1$ ($\Delta$FEV$_1$). $\Delta$FEV$_1$ (mL/year) was calculated by subtracting the Visit 1 value from the Visit 2 value and dividing by the time between Visit 1 and Visit 2. Negative values represent a lower value of the outcome at Visit 2 (i.e. worsening of the disease over the 5-year period with greater loss of FEV$_1$). From the prediction models of $\Delta$FEV$_1$, we also derived the prediction of Visit 2 FEV$_1$ by adding the predicted five-year change to the observed Visit 1 value.

Feature selection

Candidate predictors consisted of 46 baseline demographic, clinical, physiologic, and imaging variables that were available in the COPDGene population at Visit 1 and had correlation coefficients of less than 0.90 with the other variables. We set the threshold to 0.9 to ensure that only secondary/redundant features are removed, rather than features with potentially complementing information. To confirm this, we reran our experiments with removal of variables with correlation coefficients $\geq$ 0.7 and we compared the performance accuracies.

Training, testing, and validation samples

We trained a prediction model for $\Delta$FEV$_1$ in 4,496 subjects with data from COPDGene Visit 1 and Visit 2 using a nested, 10-fold cross validation (CV) procedure. The inner fold of CV was used for parameter tuning. In the outer fold, our studied derivation cohort was randomly partitioned into ten mutually exclusive subsets (folds) of approximately equal size, using nine folds for training and one fold for testing each time for ten times. This entire procedure was repeated five times to account for the random variability of the partitioning procedure and provide more accurate estimates of the performance. This repeated resampling
procedure created an ensemble of fifty models over which we averaged the predictions, and we then validated the performance of this model using data from COPDGene Visit 3 that had not been used in any aspect of the model training process (temporal validation).

**Random forest supervised machine learning**

Supervised random forest is an ensemble learning method that predicts outcomes by fitting a series of decision trees and aggregating the results across trees. This method can capture non-linear dependencies and has been shown to perform well for a range of tasks. It begins building each tree by randomly selecting participants for the tree with replacement (bootstrap samples). Participants not selected in bootstrapping represent the out-of-bag set. For each bootstrap sample, a decision tree is trained by recursive binary partition of the data until the minimum node size is reached. At each node split, an optimal feature (and its split-point) is identified from a randomly selected subset of features by minimizing a loss measure. The random selection of features reduces the correlation between trees, leading to variance reduction and improved generalization performance. It also allows a moderately informative feature to assert its importance to the prediction. Once an ensemble of trees are grown, the prediction for a new sample is made by aggregating predictions (e.g. averaging for regression and majority vote for classification) from individual trees. In our study, we fixed the number of trees at 500 and tuned the hyperparameters (the bootstrap sampling fraction, the minimal node size and the number of features to use at each split) by minimizing root mean squared error (RMSE) using a nested 10-fold cross-validation within the training data.

**Random forest variable importance and their effects on the prediction**

We calculated variable importance scores as the aggregated increase in the mean squared errors (IncMSE) of predictions estimated with out-of-bag samples when the values of
a given variable are randomly permuted. The larger the increase in prediction error when permuted, the higher the variable importance score (IncMSE), and the more important the variable is to the prediction. Since the “raw” permutation importance has better statistical properties, the importance values were not normalized. Therefore, they cannot be used to compare variable importance across prediction tasks, but they can be used within the same prediction task to rank variables by their contribution to the accuracy of the final model.

**Prediction performance**

We assessed the accuracy of each prediction model using the RMSE and R-squared metrics, indicators of the goodness of fit of a set of predictions to the observed values. For the prediction of ∆FEV₁, we also assessed the ability of the models to correctly identify subjects in the top quartile of disease progression (i.e., greatest decline in FEV₁) as quantified by the AUCs (areas under the receiver operator characteristic (ROC) curves).

**Linear regression**

To compare the performance of random forest to that of a more traditional modelling approach, the same set of predictors in was evaluated in linear regression models.

**Statistical analyses**

We performed a complete case analysis. Descriptive characteristics were reported respectively as percentages and medians with interquartile ranges for categorical and continuous variables. Variables were analyzed using the t-test for normally distributed variables, the Wilcoxon rank sum test for non-normally distributed variables, and chi-square tests for categorical variables. To identify differences in the quality of prediction and variable
importance in subjects with different levels of COPD severity, we also constructed prediction models separately in various GOLD subgroups. All tests of significance were two-tailed with a significance threshold of P-value < 0.05.

RESULTS

Subject characteristics

In total, 4,496 COPDGene subjects (median age: 60; 51% men; 73% NHW) had complete phenotypic data and were included in the analysis. The participant flow diagram is shown in Figure 1.

Characteristics of “rapid FEV<sub>1</sub> progressors” in COPDGene

To describe the characteristics of subjects who were “rapid FEV<sub>1</sub> progressors” and test the null hypothesis that there is no systematic difference in patient characteristics between the two groups, we examined the characteristics of subjects in the top quartile of progression to those in the bottom quartile (Table 1). Compared to subjects in the bottom quartile of ΔFEV<sub>1</sub>, those in the top quartile (“rapid FEV<sub>1</sub> progressors”) had a higher proportion of males with less severe spirometric impairment at baseline but with higher exposure to smoking (pack-years and percent of current smoking), more advanced radiologic disease (total emphysema and gas trapping), more bronchodilator responsiveness, more dyspnea and chronic bronchitis symptoms, and a lower rate of obesity and metabolic syndrome. The many significant P-values support the alternative hypothesis and shed light on the factors that may be associated with or even contribute to the rapid FEV<sub>1</sub> progression. The significant differences between the rapid and slow progressors also underpin the clinical relevance of identifying rapid progressors using a prediction model.
The median change in FEV₁ was -37 (IQR: -66, -9) mL/year (Figure 3). Fifty-seven percent of the studied subjects had a rate of decline in FEV₁ of more than 30 mL/year over the 5-year period and 7% had an increase in FEV₁ of more than 30 mL/year. Rapid FEV₁ progressors had a median change of -91 mL/year compared to 11 mL/year for slow spirometric progressors (Table 1). When assessed according to the severity of airflow limitation, the rate of FEV₁ decline was inversely related to the GOLD grade, with medians of ∆FEV₁ of -46, -38, -31, -16 mL/year for GOLD 1-4, respectively.

**Prediction performance for follow-up FEV₁ and 5-year change in FEV₁**

We constructed the prediction models using a nested cross-validation procedure and we assessed the prediction performance in the COPDGene 10-year follow-up visit. A schematic representation of our model is shown in Figure 2. The list of candidate predictors is provided in Table 2. In the cross-validation testing samples, on average, 89.6% of the variance in follow-up FEV₁ values were explained and the area under the ROC curves for the prediction of subjects in the top quartile of observed disease progression was 0.97 (Table 3 and Figure 4). This high performance was maintained in the temporal validation with an R-squared value of 0.91 and AUC of 0.98 (Table 3). For the prediction of the change in FEV₁ over time (∆FEV₁), the average R-squared value was 0.15 and AUC was 0.71 in the testing samples and respectively, 0.10 and 0.70 in the validation cohort.

The random forest model had slightly better performance for the prediction of ∆FEV₁ compared to linear regression (Table 3). The percentage of variance explained by random forest versus linear regression was 14.7% versus 12.3%. The indirect approach arithmetically transforms the predictions from modeling change in FEV₁ to follow-up FEV₁ predictions, and the best follow-up FEV₁ prediction is achieved via an indirect approach with random forest modeling change in FEV₁. In all cases by all metrics, random forest modeling change in
FEV₁ leads to the best prediction directly in change in FEV₁ and indirectly in follow-up FEV₁. These results demonstrate consistently the superiority of random forest versus linear regression and the merit of modeling change in FEV₁ compared with modeling follow-up FEV₁.

Candidate predictors consisted of variables that were available in the COPDGene population at Visit 1 and had correlation coefficients of less than 0.90 with the other variables. We set the threshold to 0.9 to ensure that only secondary/redundant features are removed, rather than features with potentially complementing information. To confirm this, we reran our experiments with seven variables removed using a correlation criterion of 0.7 (CT-measured total lung volumes at end-inspiration, FEV₁/FVC, GOLD spirometric grade, airway wall thickness, post-bronchodilator FEV₁, sex, and adjusted Perc15 density). We found that by setting the correlation threshold to 0.7, the resulting predictive performance decreased, particularly for the follow-up FEV₁ (median RMSE increased from 269.71 to 278.60 for follow-up FEV₁ and from 46.91 to 47.04 for change in FEV₁).

Setting the number of trees to the default of 500 provided a good compromise between performance and computational efficiency in our datasets, as evidenced by the 10-fold cross-validation loss curves with respect to the number of trees shown in Figure 1S.

**Analysis of signal to noise ratio for 5-year change in FEV₁**

Changes in spirometric measures are more commonly used endpoints in COPD clinical trials. Predicting future FEV₁ values is not the same as predicting the changes of FEV₁ over the same period, since the ∆FEV₁ over a fixed time period generally contributes a relatively small amount to the overall variance of FEV₁ at a given time point. Given the often gradual rate of progression of COPD, five-years is a relatively short observation period, and one of the
concerns is that the signal to noise ratio in our progression variables is insufficient for reliable prediction. To determine the signal-to-noise characteristics of our progression variables, we calculated the expected signal-to-noise ratio using previously published values of measurement error for FEV1. An important parameter in these calculations is the extent of correlation in errors between the two study measurements. Since empiric data were unavailable, we assumed independence between these errors; therefore, these estimates likely represent a lower bound on the proportion of noise in these measures. We estimated that measurement error accounted for at least 22% of the variance of ΔFEV1 (calculations are included in the supplement). Thus, the theoretical upper bound for prediction performance of ΔFEV1 was 78%.

**Important predictors and their effects on prediction**

Figure 5 shows the ranking of the top-20 predictors based on their importance scores in the random forest models. Several of the known COPD disease progression risk factors were present as top-ranked risk factors in our models and other new predictors were identified. The most important variables for FEV1 progression included baseline spirometry, CT-measured total lung volume, bronchodilator responsiveness, gas trapping, total emphysema, and smoking exposure. Variables like the number of COPD exacerbations in the prior year, selected comorbidities, and dyspnea scores were of less importance.

**Prediction of COPD progression stratified by GOLD grade**

To determine whether progression was determined by different variables at different GOLD spirometric grades, we examined the performance of random forest prediction models for pre-specified subgroups of smokers stratified by GOLD grade (n= 4,496 (Overall), 499 (PRISm), 2,116 (GOLD 0), 1,318 (GOLD 1-2), and 563 (GOLD 3-4)). We observed significant differences in predictive performance across these subgroups. The model performance
accuracy was best for GOLD 1-2 subjects and it was harder to achieve accurate prediction in advanced stages of the disease. The area under the ROC curves for the prediction of subjects in the top quartile of disease progression was 0.66 (GOLD 0), 0.73 (GOLD 1-2), and 0.58 (GOLD 3-4). The predictors of disease progression were also different by GOLD grade (Figure 5). For instance, bronchodilator responsiveness seems to be less important and emphysema and airway disease more important in the prediction of \( \Delta FEV_1 \) in subjects at more advanced stages of the disease.

**Effects of accounting of smoking status in both baseline and follow-up visits on the prediction performance**

At Visit 1, 47% of the studied subjects were current smokers and 53% were former smokers. At Visit 2, 37% of the studied subjects were current smokers and 63% were former smokers. At Visit 3, 30.6% of the studied subjects were current smokers and 69.4% were former smokers. In terms of change of the smoking status between visits, 35% remained current smokers at Visit 1 and Visit 2 and 50.7% remained former smokers at Visit 1 and Visit 2. 11.9% were current smokers at Visit 1 and former smokers at Visit 2 and 2.2% were former smokers at Visit 1 and current smokers at Visit 2. 27.9% of studied subjects remained current smokers at Visit 2 and Visit 3 and 63% remained former smokers at Visit 2 and Visit 3. 6.4% were current smokers at Visit 2 and former smokers at Visit 3 and 2.7% were former smokers at Visit 2 and current smokers at Visit 3. We reran our prediction models adding the smoking status variable at Visit 2 in the derivation cohort (and Visit 3 smoking status for the temporal cohort). No major effect on the prediction performance was noted as shown in Table 1S.

**DISCUSSION**
This current study showed that the prediction of change in FEV1, which is more relevant for disease progression, is more challenging than predicting absolute FEV1 level. Our prediction models for ∆FEV1 represent the current state of the art for prediction of prospective change in FEV1. But there is significant room for improvement in future models. The most important predictive variables came from a wide range of clinical, spirometric, and imaging features. Baseline spirometry, CT-measured total lung volumes, and bronchodilator responsiveness dominated the prediction. In addition, the predictive performance and the relative importance of predictors differed by GOLD grade.

Several screening tools are available to identify patients with undiagnosed COPD and to predict outcomes in patients with COPD1,8,9,20-25. While Zafari et al. and Chen et al. developed and validated risk models to accurately predict lung function trajectory8,9, our study is the first to apply advanced machine learning methods, use an extensive set of phenotypic measurements and comorbidities, predict not only the follow-up values but also the more relevant “change” variables, and identify the relative importance of the predictors at various stages of the disease. With respect to the outcomes evaluated in these two papers, our predictive models gave similar performance for the prediction of future values of FEV1. Our study added the prediction of prospective changes in FEV1 that were not reported in these previously published studies. Predicting the change over time is more challenging than simply predicting the future value, since the change typically represents a small proportion of the overall variance in a given pair of FEV1 measurements separated by five years or less. However, it is important to assess the ability of models to predict prospective changes since this is an important outcome for clinical trials.

Given the superiority of non-linear models compared with linear model with regards to exploiting complex relationships and interactions among the risk factors26, we chose random forest as our primary model due to its flexibility and generalizability, and the fact that the
interpretation of decision trees are more natural to clinicians than some of the other black-box models. Despite hundreds of trees, the ensemble method (bagging) and the base learner (decision tree) in random forest are easier to understand and interpret than many other black-box models with more sophisticated ensemble methods (e.g. boosting) or base learners (e.g. kernels, neural networks)\textsuperscript{27}. The similar performance of cross-validation and temporal validation attests to the generalizability of our models rather than overfitting, which would result in poor temporal validation performance compared to cross-validation performance. The sharp performance gap between predicting follow-up FEV\textsubscript{1} and (rate of) change in FEV\textsubscript{1} seems unintuitive at first glance. To explain this in other terms, imagine that a predictive model for change in height was developed for a cohort of adults. A model that predicted “height five years from baseline” by simply substituting the baseline height value would be very accurate, since there is little to no change in adult height over that timeframe. While FEV\textsubscript{1} does change over a five year timeframe, the absolute amount of change is usually small relative to baseline FEV\textsubscript{1} volumes. This is predicting the total FEV\textsubscript{1} in five years is a much easier (but less clinically relevant) problem than predicting the change in FEV\textsubscript{1} over five years. The key rationale is that five years is a short time period in terms of COPD progression, leading to a high correlation of FEV\textsubscript{1} values between two visits (therefore high prediction performance with follow-up FEV\textsubscript{1}) and a low signal-to-noise ratio in the FEV\textsubscript{1} 5-year progression measurements (hence poor prediction performance with change in FEV\textsubscript{1}). Despite this, there may still be merit in modeling the change in FEV\textsubscript{1} even with a short 5-year period, as we found a modest improvement in predicting follow-up FEV\textsubscript{1} using models built to predict change in FEV\textsubscript{1} that can then be transformed to follow-up FEV\textsubscript{1} (median RMSE: 258.87 and 231.38 for follow-up FEV\textsubscript{1} at Visit 2 and Visit 3, respectively). This improvement could be attributed to the change in FEV\textsubscript{1} models taking into account the uneven time lapse between visits.
Random forests offers superior prediction of disease progression relative to linear regression, and this improved performance stems from the ability of these models to more efficiently capture non-linear interactions between predictors. The predictive accuracy of our models may potentially be further improved by including additional predictors (such as DLCO, pulmonary vascular measures, and relevant molecular biomarkers) and exploring other machine learning algorithms (such as deep learning). At present, these models are not ready for clinical use but could be useful in the design of COPD clinical trials to enrich the study populations by patients who are most likely to experience rapid disease progression and benefit from therapeutic interventions. For clinical use, better performing models that have been more extensively validated in multiple additional and relevant target populations are necessary.

Rapid decline in lung function has previously been associated with a range of factors such as smoking exposure, bronchodilator reversibility, higher baseline FEV\textsubscript{1}, higher baseline FVC, exacerbations in the prior year, low BMI, African American race, female sex, emphysema, upper lobe emphysema predominance, and CT-detected small airway abnormalities\textsuperscript{5,6,8,28-33}. Our study detected several of these known COPD disease progression risk factors and identified other new predictors for FEV\textsubscript{1} decline. Our study is the first to our knowledge to demonstrate that the patterns of predictors vary by GOLD spirometric grade. The intriguing variations in the importance of different risk factors depending on the studied subgroup may help inform further exploration of predictive risk factors and future development of new risk prediction algorithms. Compared to subjects in the bottom quartile of \(\Delta\)FEV\textsubscript{1}, those in the top quartile (“rapid FEV\textsubscript{1} progressors”) had less severe spirometric impairment and more advanced radiologic disease (total emphysema and gas trapping) at baseline. It is possible that the association of less severe spirometric impairment at baseline with more rapid FEV\textsubscript{1} progression is an artifact related to the inability to lose sufficient FEV\textsubscript{1} at the same rate compared to when disease is more severe (a physiologic floor in FEV\textsubscript{1} which, once reached,
results in diminished FEV\textsubscript{1} response to additional cigarette exposure). It is also possible that the association between more severe emphysema with more rapid FEV\textsubscript{1} decline may represent a “winner’s curse”. However, it is important to note that baseline FEV\textsubscript{1} were accounted for in our analyses as these variables were among the predictors in the prediction models. In addition, the fact that our cross-validation and temporal validation performances are similar argues against the presence of large winner’s curse effects.

The relative unimportance of certain traditional risk factors such as COPD exacerbations in the prior year, selected comorbidities, race, and sex in our machine learning predictive models may be consistent with the disparate results from previous studies. For example, although some publications have suggested a significant excess loss of FEV\textsubscript{1} for each COPD exacerbation \cite{29,34,35}, others have reported minimal \cite{6} or no relationship \cite{36}. Such discrepancy may also result from differences in methodology between studies as well as differences sample size, study duration, study population, and variable definitions. The relative unimportance of certain traditional risk factors in our models may also indicate that, while these risk factors may attain statistical significance in some models, they do not provide much additional predictive value after considering more important risk factors.

Dimensionality and collinearity are important factors to consider in building and interpreting prediction models. While our data has a reasonable dimensionality in respect to the sample size, random forest performs well with high dimensional data\cite{37}. Collinearity is more of a challenge for interpreting the feature relevance ratings than the prediction performance. It is worth noting that the permutation-based feature importance scores we utilized in this study capture the marginal importance of a feature; additional approaches for capturing conditional/partial feature importance in the presence of associated features have been proposed\cite{18}. However, there is a heuristic component to these diverse feature importance scoring techniques, and there is currently no consensus or clear theoretical underpinning for
them. It has been argued that there is a marginal-partial feature importance dimension, and the researcher must determine where he/she falls on this dimension based on his/her perspective on variable importance and the research question under consideration³⁸.

The random forest's tunability of the number of trees hyperparameter has not been thoroughly investigated until recent years. For mean squared error loss in regression (and other loss functions in classification), it has been theoretically proven that increasing the number of trees does not lead to overfitting and that setting it to a computationally feasible large number is more favored than tuning the hyperparameter³⁹. Setting the number of trees to the default of 500 provided a good compromise between performance and computational efficiency in our datasets.

This study has a number of strengths. Analyses were performed within a well-characterized cohort that included subjects at all stages of disease severity. In addition, by focusing on prediction rather than the study of individual risk factors, our results provide useful context regarding the relative importance of specific predictors. By constructing models in subjects stratified by GOLD spirometric grade, we demonstrated that patterns of optimal predictors vary by specific disease outcome and GOLD grade. Validation of our findings in the temporal cohort represents another strength of our paper.

Our study also has limitations. We only used two measurements of lung function separated by approximately 5 years. The large sample size available helped to overcome some of the inherent challenges in low signal-to-noise ratio with studies of COPD progression over a relatively short period of time. However, with longer follow-up and more measurements in future studies, we will be better able to isolate measurement noise from real disease progression which will result in greater predictive accuracy. Our analysis was based on subjects who had completed their second study visit, and it is possible that patients who were lost to follow-up differed from those available for analysis. Many of the patients with airflow obstruction were
receiving therapy for their disease. Although no existing pharmacotherapy has been conclusively shown to affect the rates of disease progression, this still may have influenced our results. However, we chose not to include pharmacotherapy data in these analyses in order to reduce biases likely present in patient-reported pharmaco-epidemiologic data\textsuperscript{40,41}. It is recognized that as the number of potential risk factors increases, the complexity of the models can cause overfitting. We addressed this by appropriate hyperparameter tuning and by evaluating the performance of our predictive models in cross-validation and in the temporal cohort. Lastly, because COPDGene is one of the few available studies with deeply phenotyped subjects at all stages of disease severity, extensive clinical, spirometric, and imaging features, and follow-up data, there is currently no other appropriate replication cohort for the analyses performed, and lack of validation in an independent set of subjects limits the generalizability of our findings. It will be important for future investigations to validate these findings in independent large cohorts of similarly well-characterized smokers with the same or greater length of follow-up time.

Random forest machine learning in conjunction with deep phenotyping improves the prediction accuracy of COPD progression. The present study improves our ability to identify patients at risk for rapid disease progression, and these models may be useful for the development of targeted disease-modifying therapies.
FIGURE LEGENDS

**Figure 1.** Participants’ flow diagram and general framework of the study.

**Figure 2.** Random forest modeling framework.

**Figure 3.** Histograms of change in FEV$_1$ ($\Delta$FEV$_1$; mL/year) between Visit 1 and Visit 2 and between Visit 2 and Visit 3.

**Figure 4.** Receiver operator characteristic (ROC) curves of the performance of the random forest follow-up FEV$_1$ and 5-year change in FEV$_1$ models in correctly identifying subjects in the top quartiles of spirometric progression in the COPDGene Visit 1 / Visit 2 cross-validation testing samples. Solid lines represent the average performance, and colored dots represent the performance in each of the sampling iterations.
**Figure 5.** Heatmaps of the top-20 predictors of Visit 2 FEV$_1$ (mL) (A) and change in FEV$_1$ (mL/year) (B). The x-axis contains the group assignments (All, PRISm, GOLD0, GOLD 1-2, and GOLD 3-4). The y-axis includes the top-20 predictors ranked by their importance scores in the predictive models built in the “All” group (decreasing order with the best predictors on top). Darker shades of blue indicate a higher rank of the predictor. White cells indicate variables that do not fall within the top-20 ranks. The sample sizes were (n= 4,496 (All), 499 (PRISm), 2,116 (GOLD 0), 1,318 (GOLD 1-2), and 563 (GOLD 3-4)). Abbreviations: BMI: Body mass index; FEV$_1$: Forced expiratory volume in 1 second; FEF25-75: Forced expiratory flow at 25–75% of forced vital capacity (FVC); Bronchodilator responsiveness (%) FEV$_1$: Percentage of subjects with post-bronchodilator increase in FEV$_1$ of at least 12% from baseline; Bronchodilator responsiveness (%) FVC: Percentage of subjects with post-bronchodilator increase in FVC of at least 12% from baseline; GOLD: Global Initiative for Chronic Obstructive Lung Disease; SGRQ: St. George’s Respiratory Questionnaire; MMRC: Modified Medical Research Council; Adjusted Perc15 density: Cut off value in Hounsfield units (HU) below which 15% of all voxels are distributed on a lung CT scan (per convention, adjusted Perc15 density values are reported as the HU + 1000); Gas trapping (%): Percentage of lung voxels with a density less than -856 HU at end exhalation; Pi10: Square root of the wall area of a hypothetical airway of a 10-mm internal perimeter; % Segmental airway wall thickness: Percentage of the wall relative to the total bronchial area for the segmental airways.
Acknowledgements: The authors wish to thank the thousands of patients who participated in the COPDGene study over the last 10 years.

Author Contributions: Drs. Boueiz and Castaldi had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. COPDGene investigators were instrumental in the design and implementation of the COPDGene study and collected and analyzed data cited in this article. All authors have reviewed, approved and endorsed all content and conclusions of this article.

Declaration of Interest: The COPDGene study is funded by National Heart, Lung, and Blood Institute grants U01 HL089897 and U01 HL089856. The COPDGene study (NCT00608764) is also supported by the COPD Foundation through contributions made to an Industry Advisory Board that has included AstraZeneca, Bayer Pharmaceuticals, Boehringer-Ingelheim,
Genentech, GlaxoSmithKline, Novartis, Pfizer, and Sunovion. While some individual authors of this manuscript were employed by one of the listed funders at the time the work of this study was conducted, these employment relationships did not constitute undue influence by funders. These funders have had no official role in the collection, management, analysis and interpretation of the data or design and conduct of the study. All authors have completed a Conflict of Interest form, disclosing any real or apparent financial relationships including receiving royalties, honoraria or fees for consulting, lectures, speakers’ bureaus, continuing education, medical advisory boards or expert testimony; receipt of grants; travel reimbursement; direct employment compensation. These disclosure forms have been filed with the Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation Editorial Office and are available for review, upon request at COPDC@njhealth.org.
REFERENCES


Table 1. Characteristics of the rapid spirometric progressors.

<table>
<thead>
<tr>
<th></th>
<th>Top quartile progressors (n = 1,124)</th>
<th>Bottom quartile progressors (n = 1,124)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>58.3 [51.9, 65.0]</td>
<td>59.7 [52.3, 66.6]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Male (%)</td>
<td>66.0%</td>
<td>51.2%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Non-Hispanic whites (%)</td>
<td>72.3%</td>
<td>69.6%</td>
<td>0.098</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>173.0 [166.7, 180.0]</td>
<td>170.0 [162.6, 177.5]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>BMI</td>
<td>27.7 [24.5, 31.7]</td>
<td>29.3 [25.6, 33.9]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Pack-years of smoking</td>
<td>40.5 [29.9, 57.0]</td>
<td>38.0 [25.7, 52.5]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Current smoking (%)</td>
<td>56.1%</td>
<td>45.2%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Total emphysema (%LAA-950)</td>
<td>2.7 [0.7, 7.2]</td>
<td>1.7 [0.5, 5.0]</td>
<td>0.003</td>
</tr>
<tr>
<td>U/L ratio</td>
<td>0.38 [0.00, 0.91]</td>
<td>0.41 [0.00, 0.92]</td>
<td>0.43</td>
</tr>
<tr>
<td>Airway wall thickening (%)</td>
<td>49.6 [44.3, 55.2]</td>
<td>50.1 [44.6, 56.3]</td>
<td>0.09</td>
</tr>
<tr>
<td>Pi10</td>
<td>2.1 [1.8, 2.5]</td>
<td>2.2 [1.9, 2.6]</td>
<td>0.01</td>
</tr>
<tr>
<td>Gas trapping (%)</td>
<td>15.9 [7.6, 30.3]</td>
<td>12.3 [5.5, 25.0]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>FEV1 (percent predicted)</td>
<td>86.4 [68.9, 99.7]</td>
<td>78.9 [60.9, 91.5]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>FEV1/FVC</td>
<td>0.71 [0.61, 0.79]</td>
<td>0.72 [0.62, 0.79]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Pre/Post- bronchodilator FEV1 (% change)</td>
<td>5.5 [1.8, 11.0]</td>
<td>2.8 [-1.4, 7.9]</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Pre/Post- bronchodilator FVC (% change)</td>
<td>3.1 [-1.0, 8.8]</td>
<td>1.0 [-3.6, 6.7]</td>
<td>0.04</td>
</tr>
<tr>
<td>MMRC dyspnea score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>563 (50.1%)</td>
<td>529 (47.1%)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>171 (15.2%)</td>
<td>177 (15.7%)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>143 (12.7%)</td>
<td>137 (12.2%)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>166 (14.8%)</td>
<td>199 (17.7%)</td>
<td>0.37</td>
</tr>
<tr>
<td>4</td>
<td>81 (7.2%)</td>
<td>82 (7.3%)</td>
<td></td>
</tr>
<tr>
<td>SGRQ score</td>
<td>31.8 ± 25.9</td>
<td>29.3 ± 24.5</td>
<td>0.03</td>
</tr>
<tr>
<td>Chronic bronchitis (%)</td>
<td>21.1%</td>
<td>16.4%</td>
<td>0.004</td>
</tr>
<tr>
<td>Metabolic syndrome (%)</td>
<td>16.0%</td>
<td>20.3%</td>
<td>0.01</td>
</tr>
<tr>
<td>Obesity (%)</td>
<td>35.1%</td>
<td>45.0%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>GOLD: PRISm</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>PRISm</td>
<td>75 (6.7%)</td>
<td>546 (48.6%)</td>
<td>138 (12.3%)</td>
</tr>
<tr>
<td></td>
<td>194 (17.3%)</td>
<td>468 (41.6%)</td>
<td>65 (5.8%)</td>
</tr>
</tbody>
</table>

5-year change in FEV₁ (mL/year)

-91.0 [-117.0, -77.0] 11.0 [0.0, 32.0] <0.001

All values are from Visit 1.

BMI: Body mass index; FEV₁: Forced expiratory volume in 1 second; FVC: Forced vital capacity; MMRC: Modified Medical Research Council; SGRQ: St. George’s Respiratory Questionnaire; Exacerbation frequency: Percent of subjects reporting at least one COPD exacerbation in the previous year; Metabolic syndrome: 3 of 4: BMI ≥ 30 (measured), diabetes mellitus, hypertension, and high cholesterol (all self-report); Obesity: BMI ≥ 30; GOLD: Global Initiative for Chronic Obstructive Lung Disease; PRISm: Preserved Ratio Impaired Spirometry.

Emphysema is defined as percent of CT low attenuation area below -950 Hounsfield units (HU) at end-inspiration using Thirona software (% LAA-950); U/L ratio: Ratio of %LAA-950 in upper lung third to %LAA-950 in lower lung third; Airway wall area percent is the percentage of the wall area compared with the total bronchial area for segmental airways; Pi10: Square root of the wall area of a hypothetical airway of 10-mm internal perimeter. “Change between Visit 1 and Visit 2 per year” variables are defined as (Value at Visit 2 - Value at Visit 1) / Time between Visit 1 and Visit 2 in years.

Variables are expressed as mean and standard deviation for continuous normally distributed variables, median and interquartile range (25th to 75th percentile) for continuous non-normally distributed variables, and percentages for categorical variables. P-values are obtained using t-test for the continuous normally distributed variables, Wilcoxon rank sum test for the continuous non-normally distributed variables, and chi-square test for the proportions. P-values < 0.05 are bolded and italicized.
Table 2. Variables included in the prediction algorithms.

<table>
<thead>
<tr>
<th>Demographics:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Age at study enrollment</td>
</tr>
<tr>
<td>• Sex</td>
</tr>
<tr>
<td>• Race</td>
</tr>
<tr>
<td>• Body mass index (BMI)</td>
</tr>
<tr>
<td>• Height</td>
</tr>
<tr>
<td>• Pack-years of smoking</td>
</tr>
<tr>
<td>• Current smoking</td>
</tr>
<tr>
<td>• Age at smoking initiation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Family history:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Family history of COPD, chronic bronchitis, or emphysema</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Functional measures:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Modified Medical Research Council (MMRC) dyspnea scale</td>
</tr>
<tr>
<td>• St. George's Respiratory Questionnaire (SGRQ)</td>
</tr>
<tr>
<td>• 6-minute walk distance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COPD characteristics:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Chronic bronchitis (Chronic cough and phlegm for ≥ 3 months/year for at least 2 consecutive years)</td>
</tr>
<tr>
<td>• Blue Bloater (Chronic bronchitis, BMI &gt; 25, Resting oxygen saturation &lt; 90%)</td>
</tr>
<tr>
<td>• Pink puffer (Emphysema &gt; 10%, BMI ≤ 20, Resting oxygen saturation ≥ 90%)</td>
</tr>
<tr>
<td>• Number of COPD exacerbations over the prior year (Number of self-reported acute worsening of respiratory symptoms that required the use of antibiotics and/or systemic steroids in the previous year)</td>
</tr>
</tbody>
</table>
- History of severe COPD exacerbation (Self-report of COPD exacerbation requiring an emergency department visit or hospital admission)
- Need for courses of systemic steroids
- Poor exercise capacity (6-minute walk distance < 500 feet)
- Hypoxemia (Resting oxygen saturation ≤ 88%)
- Severe early-onset COPD (Age < 55 years, FEV₁ < 50% predicted)

Comorbidities:

- Diabetes mellitus (Self-report)
- Hypertension (Self-report)
- Dyslipidemia (Self-report)
- Pneumothorax (Self-report)
- Gastro-esophageal reflux disease (Self-report)
- Osteoporosis (Self-report)
- Coronary artery disease (Self-report of heart attack, coronary artery disease, angina, angioplasty, or coronary artery bypass graft)
- Congestive heart failure (Self-report)
- Peripheral vascular disease (Self-report)
- Metabolic syndrome (3 of 4: BMI ≥ 30 (measured), self-reported diabetes mellitus, hypertension, and high cholesterol)
- Physician diagnosis of asthma before age 40 (Self-report)
- Asthma/COPD overlap (Self-report)
- Obstructive sleep apnea (Self-report)

Spirometry:

- Post-bronchodilator FEV₁
- Post-bronchodilator FVC
- FEV₁/FVC
- Post-bronchodilator FEF₂₅₋₇₅
- Pre/Post- bronchodilator FEV₁ (% change)
- Pre/Post- bronchodilator FVC (% change)
- GOLD

**Radiology:**
- Total emphysema (%LAA-950)
- Emphysema distribution (Upper over lower lung third %LAA-950 ratio)
- Gas trapping (Percentage of low attenuation area less than -856HU at end-expiration)
- CT-measured total lung volumes at end-inspiration
- Airway wall thickness (Obtained along the center line of the lumen, in the middle third of the airway segment, for one segmental airway of each lung lobe)
- Pi10 (Square root of the wall area of a hypothetical airway of 10-mm internal perimeter)
Table 3. Prediction performance of random forest and linear regression in the cross-validation testing samples and temporal validation cohort.

<table>
<thead>
<tr>
<th></th>
<th>Random forest</th>
<th></th>
<th></th>
<th>Linear regression</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COPDGene</td>
<td>COPDGene</td>
<td>COPDGene</td>
<td></td>
<td>COPDGene</td>
</tr>
<tr>
<td></td>
<td>Visit 1 / Visit 2 testing</td>
<td>Visit 2 / Visit 3 temporal validation</td>
<td>Visit 1 / Visit 2 testing</td>
<td></td>
<td>Visit 2 / Visit 3 temporal validation</td>
</tr>
<tr>
<td>RMSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow-up FEV₁</td>
<td>269.711 [259.252, 276.476]</td>
<td>236.742</td>
<td>270.166 [260.796, 276.134]</td>
<td></td>
<td>234.978</td>
</tr>
<tr>
<td>Change in FEV₁ (mL/year)</td>
<td>46.913 [45.647, 48.795]</td>
<td>52.289</td>
<td>48.003 [46.187, 49.262]</td>
<td></td>
<td>52.819</td>
</tr>
<tr>
<td>Follow-up FEV₁ (indirect)</td>
<td>258.872 [249.820, 268.046]</td>
<td>231.377</td>
<td>263.583 [253.860, 270.307]</td>
<td></td>
<td>233.926</td>
</tr>
<tr>
<td>R-squared</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow-up FEV₁</td>
<td>0.896 [0.890, 0.903]</td>
<td>0.913</td>
<td>0.896 [0.889, 0.903]</td>
<td></td>
<td>0.915</td>
</tr>
<tr>
<td>Change in FEV₁ (mL/year)</td>
<td>0.147 [0.126, 0.173]</td>
<td>0.104</td>
<td>0.123 [0.097, 0.147]</td>
<td></td>
<td>0.0857</td>
</tr>
<tr>
<td>Follow-up FEV₁ (indirect)</td>
<td>0.904 [0.895, 0.912]</td>
<td>0.917</td>
<td>0.900 [0.894, 0.909]</td>
<td></td>
<td>0.915</td>
</tr>
<tr>
<td>AUC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow-up FEV₁</td>
<td>0.974 [0.970, 0.979]</td>
<td>0.975</td>
<td>0.974 [0.970, 0.979]</td>
<td></td>
<td>0.975</td>
</tr>
<tr>
<td>Change in FEV₁ (mL/year)</td>
<td>0.706 [0.688, 0.724]</td>
<td>0.704</td>
<td>0.698 [0.682, 0.715]</td>
<td></td>
<td>0.685</td>
</tr>
<tr>
<td>Follow-up FEV₁ (indirect)</td>
<td>0.977 [0.973, 0.982]</td>
<td>0.976</td>
<td>0.975 [0.972, 0.980]</td>
<td></td>
<td>0.976</td>
</tr>
</tbody>
</table>

The derivation cohort (COPDGene Study Visit 1 and Visit 2) was randomly partitioned into training and testing samples using 10-fold cross validation. This procedure was repeated five times to account for the random variability of the partitioning procedure. This repeated resampling procedure created an ensemble of fifty models over which we averaged the predictions, and we then validated the performance of this model using data from COPDGene Visit 3 (temporal validation). To predict the outcome values at Year 10 (Visit 3), we entered the subjects’ 5-year (Visit 2) predictor data into the models trained in the derivation cohort. Besides directly modeling follow-up FEV₁ and change in FEV₁ (mL/year), we also considered an indirect model on follow-up FEV₁ where the prediction from modeling change in FEV₁ (mL/year)
is arithmetically converted to prediction of follow-up FEV₁. The prediction performance for change in FEV₁ (mL/year) is shaded with grey color and the best performance in predicting follow-up FEV₁ and change in FEV₁ (mL/year) is highlighted with bold font.

Variables are expressed as median and interquartile range (IQR) (25th to 75th percentile) when applicable.

AUC: Area under the ROC curve for prediction of subjects in the top quartile of COPD progression; FEV₁: Forced expiratory volume in one second; RMSE: Root mean square error.
ACKNOWLEDGEMENTS

COPDGene Investigators - Core Units:

Administrative Center: James D. Crapo, MD (PI); Edwin K. Silverman, MD, PhD (PI); Barry J. Make, MD; Elizabeth A. Regan, MD, PhD

Genetic Analysis Center: Terri H. Beaty, PhD; Peter J. Castaldi, MD, MSc; Michael H. Cho, MD, MPH; Dawn L. DeMeo, MD, MPH; Adel Boueiz, MD, MMSc; Marilyn G. Foreman, MD, MS; Auyon Ghosh, MD; Lystra P. Hayden, MD, MMSc; Craig P. Hersh, MD, MPH; Jacqueline Hetmanski, MS; Brian D. Hobbs, MD, MMSc; John E. Hokanson, MPH, PhD; Wonji Kim, PhD; Nan Laird, PhD; Christoph Lange, PhD; Sharon M. Lutz, PhD; Merry-Lynn McDonald, PhD; Dmitry Prokopenko, PhD; Matthew Moll, MD, MPH; Jarrett Morrow, PhD; Dandi Qiao, PhD; Elizabeth A. Regan, MD, PhD; Aabida Saferali, PhD; Phuwanat Sakornsakolpat, MD; Edwin K. Silverman, MD, PhD; Emily S. Wan, MD; Jeong Yun, MD, MPH

Imaging Center: Juan Pablo Centeno; Jean-Paul Charbonnier, PhD; Harvey O. Coxson, PhD; Craig J. Galban, PhD; MeiLan K. Han, MD, MS; Eric A. Hoffman, Stephen Humphries, PhD; Francine L. Jacobson, MD, MPH; Philip F. Judy, PhD; Ella A. Kazerooni, MD; Alex Kluiber; David A. Lynch, MB; Pietro Nardelli, PhD; John D. Newell, Jr., MD; Aleena Notary; Andrea Oh, MD; Elizabeth A. Regan, MD, PhD; James C. Ross, PhD; Raul San Jose Estepar, PhD; Joyce Schroeder, MD; Jered Sieren; Berend C. Stoel, PhD; Juerg Tschirren, PhD; Edwin Van Beek, MD, PhD; Bram
van Ginneken, PhD; Eva van Rikxoort, PhD; Gonzalo Vegas SanchezFerrero, PhD; Lucas Veitel; George R. Washko, MD; Carla G. Wilson, MS

*PFT QA Center, Salt Lake City, UT*: Robert Jensen, PhD

*Data Coordinating Center and Biostatistics, National Jewish Health, Denver, CO*: Douglas Everett, PhD; Jim Crooks, PhD; Katherine Pratte, PhD; Matt Strand, PhD; Carla G. Wilson, MS

*Epidemiology Core, University of Colorado Anschutz Medical Campus, Aurora, CO*: John E. Hokanson, MPH, PhD; Erin Austin, PhD; Gregory Kinney, MPH, PhD; Sharon M. Lutz, PhD; Kendra A. Young, PhD

*Version Date: March 26, 2021*

*Mortality Adjudication Core*: Surya P. Bhatt, MD; Jessica Bon, MD; Alejandro A. Diaz, MD, MPH; MeiLan K. Han, MD, MS; Barry Make, MD; Susan Murray, ScD; Elizabeth Regan, MD; Xavier Soler, MD; Carla G. Wilson, MS

*Biomarker Core*: Russell P. Bowler, MD, PhD; Katerina Kecharis, PhD; Farnoush BanaeiKashani, PhD

*COPDGene Investigators - Clinical Centers:*

*Ann Arbor VA*: Jeffrey L. Curtis, MD; Perry G. Pernicano, MD

Copyright *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation* ©2022
Published online May 20, 2022  doi: [https://doi.org/10.15326/jcopdf.2021.0275](https://doi.org/10.15326/jcopdf.2021.0275)
Baylor College of Medicine, Houston, TX: Nicola Hanania, MD, MS; Mustafa Atik, MD; Aladin Boriek, PhD; Kalpatha Guntupalli, MD; Elizabeth Guy, MD; Amit Parulekar, MD

Brigham and Women’s Hospital, Boston, MA: Dawn L. DeMeo, MD, MPH; Craig Hersh, MD, MPH; Francine L. Jacobson, MD, MPH; George Washko, MD

Columbia University, New York, NY: R. Graham Barr, MD, DrPH; John Austin, MD; Belinda D’Souza, MD; Byron Thomashow, MD

Duke University Medical Center, Durham, NC: Neil MacIntyre, Jr., MD; H. Page McAdams, MD; Lacey Washington, MD

HealthPartners Research Institute, Minneapolis, MN: Charlene McEvoy, MD, MPH; Joseph Tashjian, MD

Johns Hopkins University, Baltimore, MD: Robert Wise, MD; Robert Brown, MD; Nadia N. Hansel, MD, MPH; Karen Horton, MD; Allison Lambert, MD, MHS; Nirupama Putcha, MD, MHS

Lundquist Institute for Biomedical Innovation at Harbor UCLA Medical Center, Torrance, CA: Richard Casaburi, PhD, MD; Alessandra Adami, PhD; Matthew Budoff, MD; Hans Fischer, MD; Janos Porszasz, MD, PhD; Harry Rossiter, PhD; William Stringer, MD
Michael E. DeBakey VAMC, Houston, TX: Amir Sharafkhaneh, MD, PhD; Charlie Lan, DO

Minneapolis VA: Christine Wendt, MD; Brian Bell, MD; Ken M. Kunisaki, MD, MS

Morehouse School of Medicine, Atlanta, GA: Eric L. Flenaugh, MD; Hirut Gebrekristos, PhD; Mario Ponce, MD; Silanath Terpenning, MD; Gloria Westney, MD, MS

National Jewish Health, Denver, CO: Russell Bowler, MD, PhD; David A. Lynch, MB
Reliant Medical Group, Worcester, MA: Richard Rosiello, MD; David Pace, MD

Temple University, Philadelphia, PA: Gerard Criner, MD; David Ciccolella, MD; Francis Cordova, MD; Chandra Dass, MD; Gilbert D’Alonzo, DO; Parag Desai, MD; Michael Jacobs, PharmD; Steven Kelsen, MD, PhD; Victor Kim, MD; A. James Mamary, MD; Nathaniel Marchetti, DO; Aditi Satti, MD; Kartik Shenoy, MD; Robert M. Steiner, MD; Alex Swift, MD; Irene Swift, MD; Maria Elena Vega-Sanchez, MD

University of Alabama, Birmingham, AL: Mark Dransfield, MD; William Bailey, MD; Surya P. Bhatt, MD; Anand Iyer, MD; Hrudaya Nath, MD; J. Michael Wells, MD

University of California, San Diego, CA: Douglas Conrad, MD; Xavier Soler, MD, PhD; Andrew Yen, MD
University of Iowa, Iowa City, IA: Alejandro P. Comellas, MD; Karin F. Hoth, PhD; John Newell, Jr., MD; Brad Thompson, MD

University of Michigan, Ann Arbor, MI: MeiLan K. Han, MD MS; Ella Kazerooni, MD MS; Wassim Labaki, MD MS; Craig Galban, PhD; Dharshan Vummidi, MD

University of Minnesota, Minneapolis, MN: Joanne Billings, MD; Abbie Begnaud, MD; Tadashi Allen, MD

University of Pittsburgh, Pittsburgh, PA: Frank Sciurba, MD; Jessica Bon, MD; Divay Chandra, MD, MSc; Joel Weissfeld, MD, MPH

University of Texas Health, San Antonio, San Antonio, TX: Antonio Anzueto, MD; Sandra Adams, MD; Diego Maselli-Caceres, MD; Mario E. Ruiz, MD; Harjinder Singh
Figure 1

NHW and AA smokers (age: 45-80 years old; ≥10 pack-years of lifetime smoking history) in the COPDGene cohort at Visit 1 (n = 10,263)

Subjects with Visit 2 data (n = 6,756)

Excluding subjects with unavailable 5-year spirometric data (n = 5,622)

Excluding subjects with any missing data for the studied candidate predictors: (Derivation cohort, n = 4,496)

Subjects with available Visit 3 data: (Temporal validation cohort, n = 1,833)
Figure 2
Figure 3

Change in FEV1 between Visit 1 and Visit 2

Change in FEV1 between Visit 2 and Visit 3

Copyright Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation ©2022
Published online May 20, 2022    doi: https://doi.org/10.15326/jcopdf.2021.0275
Figure 4

Follow-up FEV$_1$ (mL)

5-year change in FEV$_1$(mL/year)
**Figure 5**

Follow-up $\text{FEV}_1$ (mL)  

5-year change in $\text{FEV}_1$(mL/year)
Prediction Models for Progression of FEV₁ in the COPDGene Study

SUPPLEMENT

Adel Boueziz¹²*, Zhonghui Xu¹*, Yale Chang³, Aria Masoomi³, Andrew Gregory¹, Sharon M. Lutz⁴, Dandi Qiao¹, James D. Crapo⁵, Jennifer G. Dy³, Edwin K. Silverman¹², Peter J. Castaldi¹⁶, for the COPDGene investigators.

*Contributed equally

¹Channing Division of Network Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA; ²Pulmonary and Critical Care Division, Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA; ³Department of Electrical and Computer Engineering, Northeastern University, Boston, MA; ⁴Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, MA; ⁵Division of Pulmonary Medicine, Department of Medicine, National Jewish Health, Denver, CO; ⁶Division of General Medicine and Primary Care, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA.
Supplemental methods

**Study populations**

**COPDGene Study** (NCT00608764, [www.copdgene.org](http://www.copdgene.org)). COPDGene is an ongoing multicenter study designed to investigate the genetic and epidemiologic associations of COPD. COPDGene enrolled self-identified non-Hispanic whites (NHW) and African-Americans (AA) smokers across the full spectrum of disease severity as defined by the Global Initiative for Chronic Obstructive Lung Disease (GOLD) spirometric staging system. Subjects were aged 45 to 80 years at study enrollment and had at least 10 pack-years of lifetime smoking history. They were recruited at 21 U.S. clinical centers. Exclusion criteria included pregnancy, history of other lung diseases except asthma, prior lobectomy or lung volume reduction surgery (LVRS), active cancer, or known or suspected lung cancer. Subjects who underwent LVRS or lung transplant between visits and subjects who had more than 1-liter increase of FEV$_1$ between visits were also excluded from the analysis. Written, informed consents were obtained for all participants. The study and consent forms were approved by the Partners Human Research Committee (number 2007P000554/BWH).

**Demographic and clinical data**

Data on demographics, smoking burden, respiratory morbidity, exacerbations, and comorbidities used in this analysis were recorded at the baseline visit (Visit 1). History of COPD exacerbations in the previous year was defined as acute worsening of respiratory symptoms that required the use of antibiotics and/or systemic steroids. Severe exacerbation was defined as a COPD exacerbation requiring an emergency department visit or hospital admission. Respiratory disease-related health impairment and quality of life were assessed using the St George’s Respiratory Questionnaire (SGRQ) and dyspnea was evaluated using the Modified Medical Research Council (MMRC) dyspnea score.
Spirometric measurements

At both visits, spirometry was performed before and after administration of 180 mcg of inhaled albuterol (ndd Easy-One spirometer, Andover, MA). Percent predicted values were calculated using Hankinson NHANES reference equations. COPD was defined by post-bronchodilator FEV₁/FVC<0.70 at baseline visit per the GOLD guidelines. Bronchodilator responsiveness was defined as an increase in FEV₁ or FVC by 200 mL and 12% from baseline. Disease severity was described by GOLD spirometric stage. “GOLD 0” was defined as post-bronchodilator FEV₁/FVC≥0.70 at baseline visit and FEV₁ percent predicted ≥80%. Participants with FEV₁/FVC≥0.70 but with FEV₁<80% predicted were considered to have Preserved Ratio Impaired Spirometry (PRISm).

CT measurements

Using 3D Thirona software (www.thirona.eu), emphysema was quantified as the percentage of lung voxels with attenuation lower than -950 HU at maximal inspiration (%LAA-950) at Visit 1 and Visit 2. The ratio of lung upper third to lower third emphysema %LAA-950 was used to evaluate the apico-basal emphysema distribution (ratio950). The Hounsfield units at the 15th percentile of the CT density histogram at end-inspiration using Thirona software corrected for the variations in depth of inspiration (Adjusted Perc15) were used in the analyses of longitudinal changes in emphysema, as this may be a more robust measure of emphysema progression. Airway disease was assessed using VIDA software (www.vidadiagnostics.com) as gas trapping (percentage of low attenuation units less than -856HU at end-expiration), airway wall thickness (obtained along the center line of the lumen, in the middle third of the airway segment, for one segmental airway of each lung lobe; the mean value across all lobes was used for analysis), and Pi10 (the square root of the wall area of a hypothetical airway of 10-mm internal perimeter).
Variance due to measurement error of ∆FEV₁

Consider the measured difference in FEV₁ from Visit 1 to Visit 2. Assuming the measured outcome FEV₁ is comprised of the true value of FEV₁ and measurement error, the variance of ∆FEV₁ can be written in terms of the true measurement and measurement error as follows:

\[
Var(Y_{2m}^m - Y_{1m}^m) = Var([Y_2^t + Y_2^e] - [Y_1^t + Y_1^e])
\]

\[
= [Var(Y_2^t) + Var(Y_1^t) - 2Cov(Y_2^t, Y_1^t)] + [Var(Y_2^e) + Var(Y_1^e) - 2Cov(Y_2^e, Y_1^e)] + \\
[2Cov(Y_2^t, Y_2^e) + 2Cov(Y_1^t, Y_1^e) - 2Cov(Y_2^t, Y_1^e) - 2Cov(Y_2^e, Y_1^t)]
\]

\[
= Var(Y_2^t - Y_1^t) + [Var(Y_2^e) + Var(Y_1^e) - 2Cov(Y_2^e, Y_1^e)] + \\
[2Cov(Y_2^t, Y_2^e) + 2Cov(Y_1^t, Y_1^e) - 2Cov(Y_2^t, Y_1^e) - 2Cov(Y_2^e, Y_1^t)]
\]

(1)

where \( Y_{im}^m \) denotes FEV₁ measured at Visit \( i \); \( Y_{im}^t \) denotes the true value of FEV₁ at Visit \( i \); and \( Y_{im}^e \) denotes the measurement error associated with the measured value of FEV₁ at Visit \( i \) for \( i = 1 ; 2 \). Equation 1 assumes that the measured outcome FEV₁ is comprised of the true value of FEV₁ and measurement error such that \( i = 1 ; 2 \).

Assuming the true measurement is independent of the measurement error, then the covariance between the true measurement and measurement error is zero (i.e. \( Cov(Y_j^e, Y_k^t) = 0 \) for \( j, k = 1 ; 2 \)). Then,

\[
Var(Y_{2m}^m - Y_{1m}^m) = Var(Y_2^t - Y_1^t) + [Var(Y_2^e) + Var(Y_1^e) - 2Cov(Y_2^e, Y_1^e)]
\]

(2)

If we assume further that the measurement error associated with FEV₁ at Visit 1 is independent of the measurement error associated with FEV₁ at Visit 2, then \( Cov(Y_2^e, Y_1^e) = 0 \) and we can rewrite equation 2 as follows:

\[
Var(Y_{2m}^m - Y_{1m}^m) = Var(Y_2^t - Y_1^t) + Var(Y_2^e) + Var(Y_1^e)
\]

\[
\frac{Var(Y_2^t - Y_1^t)}{Var(Y_{2m}^m - Y_{1m}^m)} = 1 - \frac{Var(Y_2^e) + Var(Y_1^e)}{Var(Y_{2m}^m - Y_{1m}^m)}
\]

(3)
From existing literature, the coefficient of variation associated with repetitive measurements of FEV₁ over a short period of time in patients with obstructive lung disease was shown to be ~0.04-0.3% over a wide range of FEV₁. In the COPDGene study,

\[
\begin{align*}
\text{Var}(Y_i^e) & \approx 10,000 \text{ for } i = 1; 2 \\
\text{Var}(Y_2^m - Y_1^m) & = 90,000.
\end{align*}
\]

Then,

\[
\frac{\text{Var}(Y_2^m - Y_1^m)}{\text{Var}(Y_2^m - Y_1^m)} = 1 - \frac{\text{Var}(Y_2^e) + \text{Var}(Y_1^e)}{\text{Var}(Y_2^m - Y_1^m)} = 1 - \frac{10000 + 10000}{90000} = \frac{70000}{90000} \approx 0.778
\]

Based on the assumptions made above, we expect 22.2% of the variance of ΔFEV₁ to be due to measurement error in FEV₁.
FIGURE LEGENDS

**Figure 1S.** 10-fold cross-validation loss curves with respect to the number of trees. Setting the number of trees to the default of 500 for our analysis provided a good compromise between performance and computational efficiency in our datasets.
REFERENCES


Table 1S. Secondary analysis of the prediction performance of random forest and linear regression accounting for the change in smoking status between visits.

<table>
<thead>
<tr>
<th></th>
<th>Random forest</th>
<th></th>
<th>Linear regression</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COPDGene Visit 1 / Visit 2 testing</td>
<td>COPDGene Visit 2 / Visit 3 temporal validation</td>
<td>COPDGene Visit 1 / Visit 2 testing</td>
<td>COPDGene Visit 2 / Visit 3 temporal validation</td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow-up FEV1</td>
<td>270.039 [258.728, 276.634]</td>
<td>229.864</td>
<td>269.756 [260.626, 276.281]</td>
<td>226.883</td>
</tr>
<tr>
<td>Change in FEV1 (mL/year)</td>
<td><strong>47.074</strong> [45.630, 48.758]</td>
<td><strong>50.992</strong></td>
<td>47.996 [46.183, 49.368]</td>
<td>51.370</td>
</tr>
<tr>
<td>Follow-up FEV1 (indirect)</td>
<td><strong>258.990</strong> [249.593, 267.809]</td>
<td><strong>223.180</strong></td>
<td>263.273 [253.742, 270.256]</td>
<td>224.712</td>
</tr>
<tr>
<td><strong>R-squared</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow-up FEV1</td>
<td>0.896 [0.890, 0.904]</td>
<td>0.920</td>
<td>0.896 [0.889, 0.903]</td>
<td>0.922</td>
</tr>
<tr>
<td>Change in FEV1 (mL/year)</td>
<td><strong>0.146</strong> [0.123, 0.173]</td>
<td><strong>0.0903</strong></td>
<td>0.123 [0.100, 0.144]</td>
<td>0.0768</td>
</tr>
<tr>
<td>Follow-up FEV1 (indirect)</td>
<td><strong>0.904</strong> [0.896, 0.912]</td>
<td><strong>0.924</strong></td>
<td>0.900 [0.894, 0.909]</td>
<td>0.923</td>
</tr>
<tr>
<td><strong>AUC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow-up FEV1</td>
<td>0.974 [0.970, 0.979]</td>
<td>0.979</td>
<td>0.974 [0.970, 0.979]</td>
<td>0.979</td>
</tr>
<tr>
<td>Change in FEV1 (mL/year)</td>
<td><strong>0.706</strong> [0.691, 0.727]</td>
<td><strong>0.692</strong></td>
<td>0.701 [0.687, 0.714]</td>
<td>0.676</td>
</tr>
<tr>
<td>Follow-up FEV1 (indirect)</td>
<td><strong>0.977</strong> [0.973, 0.982]</td>
<td>0.979</td>
<td>0.975 [0.972, 0.980]</td>
<td><strong>0.979</strong></td>
</tr>
</tbody>
</table>

The derivation cohort (COPDGene Study Visit 1 and Visit 2) was randomly partitioned into training and testing samples using 10-fold cross validation. This procedure was repeated five times to account for the random variability of the partitioning procedure. This repeated resampling procedure created an ensemble of fifty models over which we averaged the predictions, and we then validated the performance of this model using data from COPDGene Visit 3 (temporal validation). To predict the outcome values at Year 10 (Visit 3), we entered the subjects’ 5-year (Visit 2) predictor data into the models trained in the derivation cohort. Besides directly modeling follow-up FEV1 and change in FEV1 (mL/year), we also considered an indirect model on follow-up FEV1 where the prediction from modeling change in FEV1 (mL/year) is arithmetically converted to prediction of follow-up FEV1. The prediction performance for change in FEV1 (mL/year) is shaded with grey color and the best performance in predicting follow-up FEV1 and change in FEV1 (mL/year) is highlighted with bold font. As compared to Table 3 in the main manuscript, this table reports the results of prediction modeling after adding to the list of predictors the smoking status variable at Visit 2 in the derivation cohort (and Visit 3 smoking status for the temporal cohort). No major effect on the prediction performance was noted compared to the results in Table 3. Variables are expressed as median and interquartile range (IQR) (25th to 75th percentile) when applicable.

AUC: Area under the ROC curve for prediction of subjects in the top quartile of COPD progression; FEV1: Forced expiratory volume in one second; RMSE: Root mean square error.
Figure 1S