Online Supplement
Prediction Models for Progression of FEV$_1$ in the COPDGene Study

Adel Boueiz[1,2*], Zhonghui Xu[1*], Yale Chang[3], Aria Masoomi[3], Andrew Gregory[1], Sharon M. Lutz[4], Dandi Qiao[1], James D. Crapo[5], Jennifer G. Dy[3], Edwin K. Silverman, [1,2], Peter J. Castaldi[1,6], for the COPDGene investigators.
Contributed equally

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

Supplemental methods

Study populations

COPDGene Study (NCT00608764, www.copdgene.org). COPDGene is an ongoing multicenter study designed to investigate the genetic and epidemiologic associations of COPD [1]. 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 [2]. 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 [3]. 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) [4], and dyspnea was evaluated using the Modified Medical Research Council (MMRC) dyspnea score [5].

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 [6]. COPD was defined by post-bronchodilator $FEV_1/FVC<0.70$ at baseline visit per the GOLD guidelines [3]. Bronchodilator responsiveness was defined as an increase in FEV_1 or FVC by 200 mL and 12% from baseline. Disease severity was described by GOLD spirometric stage. "GOLD 0" was defined as post-bronchodilator $FEV_1/FVC≥0.70$ at baseline visit and FEV_1 percent predicted ≥80%. Participants with $FEV_1/FVC≥0.70$ but with $FEV_1<80\%$ predicted were considered to have Preserved Ratio Impaired Spirometry (PRISm) [7].

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 [8]. 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 [9,10]. 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 ΔFEV₁

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 ΔFEV₁ can be written in terms of the true measurement and measurement error as follows:

$$Var(Y_2^m - Y_1^m) = Var([Y_2^t + Y_2^e] - [Y_1^t + Y_1^e])$$

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

$$= Var(Y_2^t - Y_1^t) + [Var(Y_2^e) + Var(Y_1^e) - 2Cov(Y_2^e, Y_1^e)] +$$
$$+ [2Cov(Y_2^t, Y_2^e) + 2Cov(Y_1^t, Y_1^e) - 2Cov(Y_2^t, Y_1^e) - 2Cov(Y_2^e, Y_1^t)]$$
$$(1)$$

where Y_i^m denotes FEV₁ measured at Visit i; Y_i^t denotes the true value of FEV₁ at Visit i; and Y_i^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,

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

If we assume further that the measurement error associated with FEV$_1$ at Visit 1 is independent of the measurement error associated with FEV$_1$ at Visit 2, then $Cov(Y_2^e, Y_1^e) = 0$ and we can rewrite equation 2 as follows:

$$Var(Y_2^m - Y_1^m) = Var(Y_2^t - Y_1^t) + Var(Y_2^e) + Var(Y_1^e)$$

$$\frac{Var(Y_2^t - Y_1^t)}{Var(Y_2^m - Y_1^m)} = 1 - \frac{Var(Y_2^e) + Var(Y_1^e)}{Var(Y_2^m - Y_1^m)} \quad (3)$$

From existing literature [11], the coefficient of variation associated with repetitive measurements of FEV$_1$ 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 FEV$_1$. In the COPDGene study,

$$\hat{Var}(Y_i^e) \approx 10,000 \text{ for } i = 1 ; 2$$

$$\hat{Var}(Y_2^m - Y_1^m) = 90,000.$$

Then,

$$\frac{\hat{Var}(Y_2^t - Y_1^t)}{\hat{Var}(Y_2^m - Y_1^m)} = 1 - \frac{\hat{Var}(Y_2^e) + \hat{Var}(Y_1^e)}{\hat{Var}(Y_2^m - Y_1^m)} = 1 - \frac{10000 + 10000}{90000} = \frac{70000}{90000} \approx 0.778$$

$$(4)$$

Based on the assumptions made above, we expect 22.2% of the variance of ΔFEV$_1$ to be due to measurement error in FEV$_1$.

REFERENCES

1.   Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDGene) study design. *COPD.* 2010;7(1):32-43.

2.   Vogelmeier CF, Criner GJ, Martinez FJ, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report: GOLD Executive Summary. *Arch Bronconeumol.* 2017;53(3):128-149.

3.   Vestbo J, Hurd SS, Agusti AG, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med.* 2013;187(4):347-365.

4.   Jones PW, Quirk FH, Baveystock CM, Littlejohns P. A self-complete measure of health status for chronic airflow limitation. The St. George's Respiratory Questionnaire. *Am Rev Respir Dis.* 1992;145(6):1321-1327.

5.   Mahler DA, Wells CK. Evaluation of clinical methods for rating dyspnea. *Chest.* 1988;93(3):580-586.

6.   Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. *Am J Respir Crit Care Med.* 1999;159(1):179-187.

7.   Wan ES, Castaldi PJ, Cho MH, et al. Epidemiology, genetics, and subtyping of preserved ratio impaired spirometry (PRISm) in COPDGene. *Respir Res.* 2014;15:89.

8.   Coxson HO, Rogers RM, Whittall KP, et al. A quantification of the lung surface area in emphysema using computed tomography. *Am J Respir Crit Care Med.* 1999;159(3):851-856.

9.   Parr DG, Sevenoaks M, Deng C, Stoel BC, Stockley RA. Detection of emphysema progression in alpha 1-antitrypsin deficiency using CT densitometry; methodological advances. *Respir Res.* 2008;9:21.

10.   Quanjer PH, Stanojevic S, Cole TJ, et al. Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. *Eur Respir J.* 2012;40(6):1324-1343.

11.   Tweeddale PM, Alexander F, McHardy GJ. Short term variability in FEV1 and bronchodilator responsiveness in patients with obstructive ventilatory defects. *Thorax.* 1987;42(7):487-490.

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

	Random forest		Linear regression	
	COPDGene Visit 1 / Visit 2 testing	**COPDGene Visit 2 / Visit 3 temporal validation**	**COPDGene Visit 1 / Visit 2 testing**	**COPDGene Visit 2 / Visit 3 temporal validation**
RMSE				
Follow-up FEV_1	270.039 [258.728, 276.634]	229.864	269.756 [260.626, 276.281]	226.883
Change in FEV_1 (mL/year)	**47.074** [45.630, 48.758]	**50.992**	47.996 [46.183, 49.368]	51.370
Follow-up FEV_1 (indirect)	**258.990** [249.593, 267.809]	**223.180**	263.273 [253.742, 270.256]	224.712
R-squared				
Follow-up FEV_1	0.896 [0.890, 0.904]	0.920	0.896 [0.889, 0.903]	0.922
Change in FEV_1 (mL/year)	**0.146** [0.123, 0.173]	**0.0903**	0.123 [0.100, 0.144]	0.0768
Follow-up FEV_1 (indirect)	**0.904** [0.896, 0.912]	**0.924**	0.900 [0.894, 0.909]	0.923
AUC				
Follow-up FEV_1	0.974 [0.970, 0.979]	0.979	0.974 [0.970, 0.979]	0.979
Change in FEV_1 (mL/year)	**0.706** [0.691, 0.727]	**0.692**	0.701 [0.687, 0.714]	0.676
Follow-up FEV_1 (indirect)	**0.977** [0.973, 0.982]	0.979	0.975 [0.972, 0.980]	**0.979**

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 FEV_1 and change in FEV_1 (mL/year), we also considered an indirect model on follow-up FEV_1 where the prediction from modeling change in FEV_1 (mL/year) is arithmetically converted to prediction of follow-up FEV_1. The prediction performance for change in FEV_1 (mL/year) is shaded with grey color and the best performance in predicting follow-up FEV_1 and change in FEV_1 (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; FEV_1: Forced expiratory volume in one second; RMSE: Root mean square error.

FIGURE LEGEND

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.

Follow-up FEV1   Change in FEV1