Supplemental (Online) Material:

For NHANES 2007-08, 2009-2010 and 2011-2012 spirometry datasets, the test results for the most frequently used spirometric parameters are included in the files SPX_E, SPX_F and SPX_G, respectively, together with instructions and analytic recommendations. Data on smoking habits were available for adult participants in the files named SMQ_E, SMQ_F and SMQ_G, respectively; serum concentrations of cotinine (a nicotine metabolite) and urinary concentrations of NNAL [4-(methylnitrosamino)-1-(3-pyridyl)-1-butanonol, a metabolite of a tobacco-specific nitrosamine] were available in the files COTNAL_E, COTNAL_F and COTNAL_G, respectively. In the SPXRAW_E, SPXRAW_F and SPXRAW_G datasets, the variable SPXRAW was made available as a comma-delimited string of numeric values between -64 and 192. These values represent the raw instantaneous flows measured during a forced expiratory maneuver and expressed as changes in exhaled volume (in milliliters) over equal, standard 0.01-second time intervals. In order to compute a timed, cumulative volume estimate, sequential volume-change values are summed over the time period desired. For example, to calculate the FEV_1, consecutive volume-change values are summed from the start of expiration up to and including the measured value at one second (including the 100^{th} value). The total number of data points in a specific comma-delimited string vary from one spirometry curve to another, depending on the duration of a particular expiratory maneuver. The variable SPXPTS in the datasets provided the total number of data points for a particular data curve.

A number of additional variables are available in the NHANES datasets for each individual curve: SPAPOS (spirometry testing position - standing or seated); SPAPLAT (whether an exhalation plateau was achieved during the maneuver); SPAACC (if the individual curve was acceptable by American Thoracic Society or ATS criteria - A, B or C grades); and SPAQEFF (the effort rating for the individual curve, A through F grades). The latter two variables are based on ATS data collection standards: A, exceeds the minimum ATS criteria (3 acceptable and 2 reproducible curves); B, meets ATS criteria (3 acceptable and 2 reproducible curves); C, potentially usable value, but does not meet ATS standards, and estimates are usually based on 2 curve results with values within 200 mL of each other; D, questionable results, use with caution; F: results not valid (none of the spirometry curves were in the latter category).

In all three datasets, survey participants of age 6 to 79 years, who were deemed eligible for spirometry were included. All tests were done using Ohio 822/827 dry-rolling seal volume spirometers (https://wwwn.cdc.gov/nchs/data/nhanes/2011-2012/manuals/spirometry_procedures_manual.pdf). The normal predicted values used were those derived on the NHANES III cohort by Hankinson et al.[15] Per the NHANES protocol, participants eligible for spirometry performed an initial or 'baseline' 1^{st} test *spirometry examination. If certain criteria were met (i.e., FEV_1/FVC < LLN by NHANES III predicted values), participants then underwent a 2nd spirometry examination, after inhaling a β2-adrenergic bronchodilator. Multiple individual spirometry curves were thus obtained, during both the 1st and the 2nd test spirometry examinations. Given some unintended under- or over-sampling in various NHANES cohorts, in order to accurately reflect the U.S. population characteristics, the recommended NHANES full-sample 2-Year MEC Examination Weight (WTMEC2YR) values were used to analyze the NHANES spirometry data; as such, we used weighted measurements in the analytic reports.*

Some of the prior models for PFT normal values used regular linear regression (standard least squares method) by gender and race or ethnicity group, relying on predictive variables such as age, height and occasionally weight. In this set of analyses, both regular regression and generalized ('regularized' or 'penalized' personality) regression models were performed, employing optimization techniques such as ridge regression, lasso, elastic net and double lasso methods (with or without adaptive features), and using either native values, or logarithmic, gamma, Weibull or SHASH (sinh-arcsinh)[16] transformations of the response variables. Among the distributions fit of the response variables, the logarithmic transformation provided an acceptable trade-off between ease of implementation and adjustment to normality, so we used it in the presented models, which were regular regression models. In various models investigated, we assessed the AEX-FV as Y variable versus the following inputs (which on univariate analyses were independent predictors of the Y variable): gender; race or ethnicity group; age, height, weight and their interactions; forced expiratory time; and effort grade. For internal validation of all models developed, we randomly partitioned the combined NHANES cohorts into a training set (70%) and a validation set (30%). We dropped from the model the variables that contributed to the validation R^2 (which the model tried to maximize) by less than 0.02, albeit significant on multivariate analyses. Significant interactions on multivariate analyses were also ignored if their estimates or beta coefficients were very small (<0.001). In each model, we generated variable importance reports, which compute indices of factor contribution that are independent of the model type and fitting method used, estimating variability in the predicted response based on a range of variation for each factor. As such, if the variation in one factor causes a high variability of the response, then that effect is important relative to the model. We used two approaches: (1) the independent resampled inputs method, in which for each factor, Monte Carlo samples are obtained by resampling its set of observed values (this option works best when the factors are uncorrelated and their values are not represented by a uniform distribution); and (2) the dependent resampled inputs method, in which factor values are constructed from observed combinations using a k-nearest neighbors' approach in order to account for

Ioachimescu OC, *et al. J Investig Med* 2022; 70:1247–1257. doi: 10.1136/jim-2021-002057

correlation (this option treats observed variance and covariance as representative of the covariance structure for the model's likely correlated factors). Factors that contributed by less than 5% to the total effect by both methods were excluded. The maximum likelihood estimation method was preferred to ridge, lasso, elastic net or double lasso optimized regression methods if the R^2 did not increase by more than 0.02. Similarly, neural networks (machine learning algorithms) were not included if they did not lead to an improvement in R^2 by more than 0.02 over the generalized regression models.

Ioachimescu OC, *et al. J Investig Med* 2022; 70:1247–1257. doi: 10.1136/jim-2021-002057