HHS Public Access

Author manuscript

Epidemiology. Author manuscript; available in PMC 2021 May 01.

Published in final edited form as:

Epidemiology. 2020 May ; 31(3): 459–466. doi:10.1097/EDE.0000000000001176.

Validation of Questionnaire-based Case Definitions for Chronic Obstructive Pulmonary Disease

Lydia Feinstein[1,2], **Jesse Wilkerson**[1], **Paivi M Salo**[3], **Nathaniel MacNell**[1], **Matthew F Bridge**[1], **Michael B Fessler**[3], **Peter S Thorne**[4], **Angelico Mendy**[3,4], **Richard D Cohn**[1,5], **Matthew D Curry**[1], **Darryl C Zeldin**[3]

[1]Social & Scientific Systems, Durham, NC.

[2]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC.

[3]Division of Intramural Research, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC.

[4]Department of Occupational and Environmental Health, University of Iowa, Iowa City, Iowa.

[5]Independent consultant, Chapel Hill, NC.

Abstract

Background: Various questionnaire-based definitions of chronic obstructive pulmonary disease (COPD) have been applied using the US-representative National Health and Nutrition Examination Survey (NHANES), but few have been validated against objective lung function data. We validated two prior definitions that incorporated self-reported physician diagnosis, respiratory symptoms, and/or smoking. We also validated a new definition that we developed empirically using gradient boosting, an ensemble machine learning method.

Methods: Data came from 7,996 individuals aged 40–79 years who participated in NHANES 2007–2012 and underwent spirometry. We considered participants "true" COPD cases if their ratio of post-bronchodilator forced expiratory volume in 1 second to forced vital capacity was below 0.7 or the lower limit of normal. We stratified all analyses by smoking history. We developed agradient boosting model for smokers only; predictors assessed (25 total) included sociodemographics, inhalant exposures, clinical variables, and respiratory symptoms.

Results: The spirometry-based COPD prevalence was 26% for smokers and 8% for never smokers. Among smokers, using questionnaire-based definitions resulted in a COPD prevalence ranging from 11%−16%, sensitivity ranging from 18%−35%, and specificity ranging from 88% −92%. The new definition classified participants based on age, bronchodilator use, BMI, smoking pack–years, and occupational organic dust exposure, and resulted in the highest sensitivity (35%)

Corresponding Author: Darryl C Zeldin, MD, Division of Intramural Research, National Institute of Environmental Health Sciences, 111 T.W. Alexander Drive, Bldg 101, A214, Research Triangle Park, NC 27709, Phone: 984-287-3641, zeldin@niehs.nih.gov.

Availability of data and code: All data used in this analysis are publicly available and can be downloaded on the Centers for Disease Control website: https://wwwn.cdc.gov/nchs/nhanes/Default.aspx. SAS code used in this analysis is provided in the Supplemental Digital Content as eAppendix 2.

Conflicts of interest: None declared.

and specificity (92%) among smokers. Among never smokers, the COPD prevalence ranged from 4%−5%, and we attained good specificity (96%) at the expense of sensitivity (9–10%).

Conclusions: Our results can be used to parametrize misclassification assumptions for quantitative bias analysis when pulmonary function data are unavailable.

Keywords

Chronic Obstructive Pulmonary Disease; Pulmonary Emphysema; Chronic Bronchitis; Validation Studies; Machine Learning; Surveys and Questionnaires

INTRODUCTION

Clinical guidelines suggest that chronic obstructive pulmonary disease (COPD) be diagnosed with spirometry and is defined as post-bronchodilator forced expiratory volume in one second to forced vital capacity (FEV_1/FVC) ratio < 0.7.[1] However, in population-based research, large-scale pulmonary function testing is often unavailable, and researchers have had to rely on questionnaire-based definitions to identify COPD.[2–6]

The National Health and Nutrition Examination Survey (NHANES), which was conducted by the U.S. National Center for Health Statistics, is a popular data source for studying COPD.[5,7–12] However, spirometry was only performed in select years (2007–2012) and the subsample with spirometry data often do not overlap with other subsamples with data on key exposures or covariates of interest. Several investigators wanting to use NHANES data to study COPD-related questions have derived their own questionnaire-based case definitions,[5,13,14] but none has been validated.

Misclassification poses a substantial threat to the validity of epidemiologic studies, and researchers are encouraged to perform quantitative bias analysis to gauge how robust their findings are under various assumptions about the extent of misclassification.[15–19] While methods to conduct quantitative bias analyses have existed for many years, these methods rely on the availability of validation data from other studies to parameterize misclassification assumptions. Limited data currently exist in the context of questionnaire-defined COPD.

To increase the availability of validation data for conducting subsequent quantitative bias analyses in population-based studies of COPD that lack pulmonary function testing, we assessed the validity of two questionnaire-based definitions previously applied in NHANES against spirometry data. Because the previously applied COPD definitions were developed using clinical assumptions, we additionally assessed the validity of a new COPD definition that we developed empirically using gradient boosting, an ensemble machine learning method. To increase the utility of our study results across a range of settings, we used nationally representative data, validated both cases and non-cases, and provided classification parameters stratified by a range of covariates.

METHODS

Study Population

We used the 2007–2012 NHANES, which employs a complex, multistage, stratified probability sampling design to select participants representative of the civilian, non-institutionalized US population.[20] The analysis included 7,996 individuals aged 40–79 years with complete respiratory questionnaires, smoking data, and pre-bronchodilator spirometry data (see Figure 1). We limited the analysis to participants ages 40–79 because COPD is extremely rare at ages younger than 40 years[21] and because those >79 years of age were ineligible for spirometry in NHANES. All NHANES protocols were approved by the National Center for Health Statistics Research Ethics Review Board.[22]

Ascertainment of "True" COPD Status

The NHANES spirometry and bronchodilator procedures have been published previously.[23,24] Those with pre-bronchodilator FEV_1/FVC below 0.7[1] or the lower limit of normal for their age, sex, and race/ethnicity[9] were eligible for a second spirometry test after bronchodilator (Albuterol) treatment. Of the 1,605 individuals eligible for post-bronchodilator spirometry, 48% underwent testing (see eTable 1). Participants who were eligible for but did not receive post-bronchodilator spirometry included those who declined to participate, had an incomplete or unreliable exam, or who were excluded for one of the following safety concerns related to the bronchodilator treatment: recent use of a β2-adrenergic bronchodilator; previous adverse reaction to albuterol; current pregnancy or breastfeeding; diagnosed major arrhythmia; elevated blood pressure for age; resting tachycardia; irregular pulse; current use of class 1 antiarrhythmics, an implanted automatic defibrillator, monoamine oxidase inhibitors, anticonvulsant medications for epilepsy, or diuretic therapy without potassium supplementation; current use of tricyclic antidepressants if being treated for heart disease, kidney disease, or a thyroid disorder. When available, post-bronchodilator spirometry values were used, with participants having a FEV_1/FVC below 0.7 or the lower limit of normal being classified as "true" COPD cases.[1] When post-bronchodilator results were unavailable (52%), we used pre-bronchodilator values. Among those with both pre- and post-bronchodilator values, 39% (unweighted) who showed evidence of obstruction at baseline were ultimately classified as COPD negative after bronchodilator treatment.

Questionnaire-based Case Definitions of COPD

We examined two questionnaire-based COPD case definitions previously applied to NHANES data. The first defined COPD as self-reported physician diagnosis of chronic bronchitis or emphysema (Mendy et al.).[5] The second additionally classified COPD as report of chronic cough and phlegm production plus ≥10 pack–years of smoking (Fessler et al.).[14]

We also examined a new COPD case definition that we developed empirically using gradient boosting. Briefly, gradient boosting is an ensemble machine learning method that employs an iterative process to grow decision trees, with each subsequent tree using information from

the prior tree to reduce misclassification.[25,26] The utility of this approach has been demonstrated previously in the epidemiologic literature.[27–29]

Variables considered for the new definition included sociodemographic variables (age, gender, race/ethnicity, educational attainment, poverty to income ratio), body mass index (BMI), inhalant exposure variables (smoking pack–years, secondhand tobacco smoke exposure, and occupational exposure to mineral or organic dusts, exhaust, or other fumes), medical history variables (physician-diagnosed emphysema, physician-diagnosed chronic bronchitis, current physician-diagnosed asthma, hay fever, and use of bronchodilators), and respiratory symptom variables (cough and phlegm for 3 consecutive months for 2 or more years, shortness of breath when walking on stairs or inclines, and impact of wheezing on exercise, sleep, and normal activity). Bronchodilator use included participants who received a Beta-2-adrenergic bronchodilator, anticholinergic bronchodilator, a methylxanthine, or bronchodilator combination prior to pre-bronchodilator spirometry. The variables selected for inclusion in the model were based on known COPD risk factors. Unless otherwise noted above, all variables were defined in accordance with the NHANES analytic guidelines.[30] We compared accuracy (number of correct classifications/total N) between models that included up to four trees and a depth of five. To facilitate interpretation, we ultimately selected the most parsimonious model (i.e., fewest number of variables and branches) once reductions in the error rate leveled off. More details on gradient boosting model are provided in eAppendix 1.

The new COPD definition (depicted visually in Figure 2) ultimately included one tree with a depth of four and five predictor variables: age, bronchodilator use, BMI, smoking pack–years, and occupational organic dust exposure. Participants were classified as COPD cases if they were at least 55 years of age and met any of the following criteria: (1) bronchodilator use, (2) <32 BMI with >36 smoking pack–years, or (3) <32 BMI with >18 years of occupational organic dust exposure.

Statistical Analysis

As COPD risk factors are likely different for those with a history of smoking and those who have never smoked, we stratified all analyses by smoking history, examining current and former smokers separately from never smokers. We were unable to develop a reliable gradient boosting model for never smokers, and thus our new COPD definition was developed only among those with a smoking history. However, we show the classification parameters of this definition among never smokers to determine how it performs compared to the previously applied definitions.

We conducted descriptive analyses to characterize the distribution of COPD predictors in the study population. To assess the validity of each questionnaire-based COPD case definition, we calculated the following classification parameters for predicting COPD status based on spirometry: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy (number of correct classifications/total N). We also calculated COPD prevalence based on each definition.

All analyses appropriately accounted for the NHANES complex survey design. Gradient boosting was conducted in R (Version 3.3.4) using the XGBoost package (Version 0.71.2). All other analyses were conducted in SAS 9.4 (SAS Institute, Inc., Cary, North Carolina).

Secondary analyses

We conducted several secondary analyses to inform future bias analyses and studies conducted in specific sub-populations (see eTables 2–11). Specifically, we examined classification parameters stratified by:

1. Former *vs.* current smoking
2. Gender (women *vs.* men)
3. Age (40–59 *vs.* 69+ years of age)
4. Race/ethnicity (non-Hispanic white *vs.* non-Hispanic black *vs.* Hispanic)
5. With *vs.* without asthma
6. Fixed ratio *vs.* lower limit of normal cut-point for obstruction

We then examined classification parameters limiting to participants with moderate-to-severe COPD (see eTable 8), defined in accordance with guidelines as having a post-bronchodilator FEV_1/FVC below 0.7 or the lower limit of normal with a FEV_1 value less than 80% predicted.[1] As FEV_1 predicted values are only defined for non-Hispanic white, non-Hispanic black, and Mexican American participants, this sensitivity analysis was additionally limited to this subgroup.

To assess the extent to which missing post-bronchodilator data may have affected our results, we corrected for potential misclassification using the re-weighting approach described by Tilert et al (see eTable 9).[11] For additional context, we also examined classification parameters using pre-bronchodilator values only (see eTable 10) and excluding those who did not undergo post-bronchodilator testing despite being eligible (see eTable 11).

RESULTS

Descriptive Analysis

Figure 1 shows unweighted frequencies of the study population, which included 3,940 current or former smokers and 4,056 never smokers. Among smokers, the prevalence of spirometry-defined COPD was 26%, while among never smokers the prevalence was 8%. Table 1 shows the distribution of participant characteristics by smoking and COPD status.

Case Definition Classification Parameters

Table 2 shows the COPD prevalence and classification parameters for each case definition. Among current and previous smokers, the COPD prevalence ranged from 11%−16%. In this population, the Mendy case definition tied for the highest specificity (92%) but had the lowest prevalence (11%) and sensitivity (18%). Our new case definition (gradient boosting) resulted in the highest sensitivity (35%) while still maintaining a high specificity (92%). Our new definition also performed better in terms of positive predictive value (59%) relative to

the previous case definitions, had similar negative predictive value (81% relative to 77% −78%), and resulted in better overall accuracy (78% relative to 72%−73%).

All three case definitions performed similarly among never-smokers. In this group, the prevalence of COPD ranged from 4%−5%, sensitivity ranged from 9%−10%, specificity was 96%, PPV ranged from 14%−18%, NPV was 93%, and overall accuracy ranged from 89% −90%.

DISCUSSION

Overall, our study provides guidance to researchers studying COPD and related conditions in the absence of pulmonary function testing. We provide estimates stratified by smoking status and within levels of a range of covariates, which can be used to parametrize future bias analyses.

All three of the definitions we validated resulted in a high specificity (range: 88%−96%), but a relatively low sensitivity (range: 9%−35%). There is a direct tradeoff between sensitivity and specificity, and the ideal tradeoff between the two has to be evaluated against the research or clinical objective in question, as well as characteristics of the disease.[31] For example, compromising specificity for sensitivity is necessary in the context of common outcomes and screening tools intended for clinical application. Several validated COPD screening questionnaires exist, including the COPD Diagnostic Questionnaire,[13] the COPD Population Screener,[32] the Lung Function Questionnaire[33], and the Salzburg COPD-screening questionnaire.[34] These questionnaires were primarily designed to identify those at risk for COPD who would benefit from spirometry to confirm the diagnosis. As such, these clinical questionnaires tend to have a high sensitivity for COPD identification at the expense of a lower specificity in order to confidently rule out those without the disease.

Although compromising specificity for sensitivity is acceptable and often necessary when administering screening tools, such criteria may not be as optimal in other scenarios. For example, a high specificity may be most important when examining a primary outcome, particularly in the context of a rare outcome such as COPD,[31] as in the paper by Mendy et al. that looked at the association between house dust endotoxin and COPD.[5] In our analysis, this case definition (affirmative response to physician-diagnosed chronic bronchitis or emphysema) tied for the highest specificity among both smokers (92%) and never smokers (96%). As expected, this relatively restrictive definition also had a low sensitivity (18% among smokers and 9% among never smokers), suggesting that there may be a high proportion of false negatives among non-cases. Indeed, under-diagnosis of COPD is widely documented in the literature,[35–37] including in NHANES.[10] The extent to which underdiagnosis may bias an analysis likely depends on the application, and a previously published sensitivity analysis found that relative risk estimates were robust to changes in false-negative probabilities compared to changes in false-positive probabilities.[2]

Among smokers, the new definition that we developed using gradient boosting resulted in a higher sensitivity than the previously used definitions while matching the highest specificity. Notably, the new definition does not rely on self-reports of chronic bronchitis or

emphysema, as our model suggested that these variables were weaker predictors of COPD status than age, bronchodilator use, BMI, smoking pack–years, and number of years of occupational exposure to organic dust.

We are aware of at least two prior studies that validated questionnaire-defined COPD.[2] In the Nurses' Health Study, Barr et al. validated self-reported COPD (cases only) against medical records and reported positive predictive values ranging from 78%−86%, higher than we observed in this study. This may be explained by the fact that the authors only validated cases and by differences in the study populations, with the Nurses' Health Study population likely being more educated on medical topics and thus more likely to accurately self-report disease status than the general U.S. population. These differences suggest that relying on self-reported disease status may be more reliable in certain settings, whereas using an alternate definition such as the one we developed using gradient boosting may be more reliable in other settings.

The other previously published study, which was conducted in Sweden among a nationally representative general population sample, found results more consistent with what we saw in our US-representative study, but with even lower sensitivity.[38] Using self-reported physician diagnosis, the authors found a sensitivity of 5.6% and a specificity of 99.7%. Using a definition based on questionnaire-assessed chronic bronchitis symptoms, the authors found a sensitivity of 4.6% and a specificity of 97.9%. The lower sensitivities and higher specificities reported by the authors compared to our study may be driven by the fact that the authors did not stratify by smoking status (their results are most comparable to what we observed among never smokers) and that they relied solely on pre-bronchodilator spirometry.

Our results indicate that accurately classifying COPD among never smokers may require a different case definition than among smokers, possibly due to the heterogeneous set of risk factors that may predispose these individuals to COPD.[39–42] The varying etiologies may also have contributed to our inability to estimate a reliable gradient boosting model in this population. Given that, globally, 25%−45% of individuals with COPD never smoked,[43] it is important that more research is done to develop optimal criteria to identify COPD among never smokers in population-based studies when spirometry is not readily available.

We used spirometry testing as the gold standard for the diagnosis of COPD given that current guidelines support pulmonary function testing as an integral aspect of diagnosing COPD.[1] In the clinical setting, spirometry is typically performed in conjunction with a thorough clinical examination. Spirometry is estimated to have a sensitivity of 92% when used in the primary care setting.[44] Regardless, we cannot exclude the possibility that some patients with COPD were not identified in our population due to the absence of a complete clinical evaluation. Additionally, our gold standard definition may capture chronic obstructive lung diseases other than emphysema and chronic bronchitis, such as bronchiectasis. Alternative imaging modalities such as computed tomography and magnetic resonance imaging are becoming better recognized and may offer more precise diagnostic capabilities.[45]

Although COPD is generally a disease associated with aging, there may be a low prevalence among young adults.[46] The classification parameters demonstrated in the present study for 40–79 year olds may not translate to younger populations. Likewise, because NHANES limited spirometry testing to those <80 years of age, our classification should be used with caution for populations 80 years of age and older.

A large proportion (52%) of eligible participants did not undergo post-bronchodilator spirometry. Using data from NHANES 2007–2010, Tilert et al. showed that COPD prevalence reduced 33% when using post- rather than pre-bronchodilator spirometry.[11] Replicating the methods the authors used to account for the misclassification, we found that our results remained relatively unchanged. This is reassuring given the high number of participants who are likely to have a medical contraindication to bronchodilator treatment in a population-based setting. While bronchodilator treatment can help distinguish COPD from other airway diseases (e.g., asthma), its utility may be limited in a population-based setting given the extent of contraindications to the treatment (see safety exclusions described in the methods section) and willingness of participants to undergo the treatment for the sake of the study.

Despite these limitations, our study also had several strengths, including using data from a US representative study with a detailed respiratory questionnaire, extensive subgroup analyses, and validation data for both cases and non-cases. We were also able to make use of previously collected data rather than employing a design that required validating self-reported disease against medical records, another common validation study design. Although such studies make an important contribution to the literature, in the case of COPD, medical record documentation of disease has been shown to lead to even more extreme underestimation of disease prevalence than questionnaires that rely on self-reported disease status.[47] Gradient boosting, the machine learning approach we used to develop our new case definition, was also advantageous in that it is non-parametric and thus does not require assumptions about the functional form of predictor variables (e.g., normality). It also inherently accounts for any underlying interactions between predictors. Gradient boosting and other ensemble machine learning approaches have been used to successfully predict many clinical outcomes of interest, including cardiovascular events, fetal growth, and mortality.[26–28]

NHANES continues to be widely utilized in COPD research. Examining the validity of simple case definitions built from easily assessed questionnaire items can be used to inform future population-based studies with limited resources for clinical-based assessments. Our validation results suggest that misclassification based on these definitions may be considerable, and our results add to the limited resources available for researchers to conduct sensitivity analyses.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Source of Funding:

This work was supported, in part, by the Intramural Research Program of the National Institutes of Health, National Institute of Environmental Health Sciences (Z01-ES-025041), and through a contract to Social & Scientific Systems funded by the National Institute of Environmental Health Sciences (HHSN273201600002I).

REFERENCES

1. Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD). 2017. Report.
2. Barr RG, Herbstman J, Speizer FE, Camargo CA, Jr. Validation of self-reported chronic obstructive pulmonary disease in a cohort study of nurses. Am J Epidemiol 2002;155(10):965–71. [PubMed: 11994237]
3. Liu Y, Croft JB, Anderson LA, Wheaton AG, Presley-Cantrell LR, Ford ES. The association of chronic obstructive pulmonary disease, disability, engagement in social activities, and mortality among US adults aged 70 years or older, 1994–2006. Int J Chron Obstruct Pulmon Dis 2014;9:75–83. [PubMed: 24477269]
4. Martinez CH, Richardson CR, Han MK, Cigolle CT. Chronic obstructive pulmonary disease, cognitive impairment, and development of disability: the health and retirement study. Ann Am Thorac Soc 2014;11(9):1362–70. [PubMed: 25285360]
5. Mendy A, Salo PM, Cohn RD, Wilkerson J, Zeldin DC, Thorne PS. House Dust Endotoxin Association with Chronic Bronchitis and Emphysema. Environ Health Perspect 2018;126(3):037007. [PubMed: 29578323]
6. Cheruvu VK, Odhiambo LA, Mowls DS, Zullo MD, Gudina AT. Health-related quality of life in current smokers with COPD: factors associated with current smoking and new insights into sex differences. Int J Chron Obstruct Pulmon Dis 2016;11:2211–2219. [PubMed: 27695308]
7. Ford ES, Mannino DM, Wheaton AG, Giles WH, Presley-Cantrell L, Croft JB. Trends in the prevalence of obstructive and restrictive lung function among adults in the United States: findings from the National Health and Nutrition Examination surveys from 1988–1994 to 2007–2010. Chest 2013;143(5):1395–406. [PubMed: 23715520]
8. Fragoso CA, Concato J, McAvay G, Yaggi HK, Van Ness PH, Gill TM. Staging the severity of chronic obstructive pulmonary disease in older persons based on spirometric Z-scores. J Am Geriatr Soc 2011;59(10):1847–54. [PubMed: 22091498]
9. 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):179–87. [PubMed: 9872837]
10. Martinez CH, Mannino DM, Jaimes FA, Curtis JL, Han MK, Hansel NN, Diaz AA. Undiagnosed Obstructive Lung Disease in the United States. Associated Factors and Long-term Mortality. Ann Am Thorac Soc 2015;12(12):1788–95. [PubMed: 26524488]
11. Tilert T, Dillon C, Paulose-Ram R, Hnizdo E, Doney B. Estimating the U.S. prevalence of chronic obstructive pulmonary disease using pre- and post-bronchodilator spirometry: the National Health and Nutrition Examination Survey (NHANES) 2007–2010. Respir Res 2013;14:103. [PubMed: 24107140]
12. Tilert T, Paulose-Ram R, Howard D, Butler J, Lee S, Wang MQ. Prevalence and factors associated with self-reported chronic obstructive pulmonary disease among adults aged 40–79: the National Health and Nutrition Examination Survey (NHANES) 2007–2012. EC Pulmonol Respir Med 2018;7(9):650–662. [PubMed: 30294723]
13. Busse PJ, Cohn RD, Salo PM, Zeldin DC. Characteristics of allergic sensitization among asthmatic adults older than 55 years: results from the National Health and Nutrition Examination Survey, 2005–2006. Ann Allergy Asthma Immunol 2013;110(4):247–52. [PubMed: 23535087]
14. Fessler MB, Carnes MU, Salo PM, Wilkerson J, Cohn RD, King D, Hoppin JA, Sandler DP, Travlos G, London SJ, Thorne PS, Zeldin DC. House Dust Endotoxin and Peripheral Leukocyte Counts: Results from Two Large Epidemiologic Studies. Environ Health Perspect 2017;125(5):057010. [PubMed: 28599265]
15. Fox MP. Creating a demand for bias analysis in epidemiological research. J Epidemiol Community Health 2009;63(2):91. [PubMed: 19141660]

16. Greenland S Basic methods for sensitivity analysis of biases. Int J Epidemiol 1996;25(6):1107–16. [PubMed: 9027513]
17. Jurek AM, Lash TL, Maldonado G. Specifying exposure classification parameters for sensitivity analysis: family breast cancer history. Clin Epidemiol 2009;1:109–17. [PubMed: 20865092]
18. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol 2014;43(6):1969–85. [PubMed: 25080530]
19. Lash TL, Olshan AF. EPIDEMIOLOGY Announces the "Validation Study" Submission Category. Epidemiology 2016;27(5):613–4. [PubMed: 27388372]
20. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. https://wwwn.cdc.gov/Nchs/Nhanes/2011-2012/SPX_G.htm Accessed 08/13/2018, 2018.
21. Ntritsos G, Franek J, Belbasis L, Christou MA, Markozannes G, Altman P, Fogel R, Sayre T, Ntzani EE, Evangelou E. Gender-specific estimates of COPD prevalence: a systematic review and meta-analysis. Int J Chron Obstruct Pulmon Dis 2018;13:1507–1514. [PubMed: 29785100]
22. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). NCHS Research Ethics Review Board (ERB) Approval. https://www.cdc.gov/nchs/nhanes/irba98.htm, 2019.
23. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey (NHANES): Respiratory Health Spirometry Procedures Manual. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2011.
24. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey (NHANES): Respiratory Health Bronchodilator Procedures Manual. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2008–2012.
25. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann. Statist 2001;29(5):1189–1232.
26. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O, written on behalf of AMEB-DCTCG. Predictive analytics with gradient boosting in clinical medicine. Ann Transl Med 2019;7(7):152. [PubMed: 31157273]
27. Naimi AI, Platt RW, Larkin JC. Machine Learning for Fetal Growth Prediction. Epidemiology 2018;29(2):290–298. [PubMed: 29199998]
28. Nanayakkara S, Fogarty S, Tremeer M, Ross K, Richards B, Bergmeir C, Xu S, Stub D, Smith K, Tacey M, Liew D, Pilcher D, Kaye DM. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. PLoS Med 2018;15(11):e1002709. [PubMed: 30500816]
29. Setodji CM, McCaffrey DF, Burgette LF, Almirall D, Griffin BA. The Right Tool for the Job: Choosing Between Covariate-balancing and Generalized Boosted Model Propensity Scores. Epidemiology 2017;28(6):802–811. [PubMed: 28817469]
30. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey. https://wwwn.cdc.gov/nchs/nhanes/analyticguidelines.aspx#analytic-guidelines.
31. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol 2012;65(3):343–349 e2. [PubMed: 22197520]
32. Martinez FJ, Raczek AE, Seifer FD, Conoscenti CS, Curtice TG, D'Eletto T, Cote C, Hawkins C, Phillips AL, Group C-PCW. Development and initial validation of a self-scored COPD Population Screener Questionnaire (COPD-PS). COPD 2008;5(2):85–95. [PubMed: 18415807]
33. Hanania NA, Mannino DM, Yawn BP, Mapel DW, Martinez FJ, Donohue JF, Kosinski M, Rendas-Baum R, Mintz M, Samuels S, Jhingran P, Dalal AA. Predicting risk of airflow obstruction in primary care: Validation of the lung function questionnaire (LFQ). Respir Med 2010;104(8):1160–70. [PubMed: 20226647]
34. Weiss G, Steinacher I, Lamprecht B, Kaiser B, Mikes R, Sator L, Hartl S, Wagner H, Studnicka M. Development and validation of the Salzburg COPD-screening questionnaire (SCSQ): a

questionnaire development and validation study. NPJ Prim Care Respir Med 2017;27(1):4. [PubMed: 28127061]

35. Almagro P, Soriano JB. Underdiagnosis in COPD: a battle worth fighting. Lancet Respir Med 2017;5(5):367–368. [PubMed: 28389226]
36. Hill K, Goldstein RS, Guyatt GH, Blouin M, Tan WC, Davis LL, Heels-Ansdell DM, Erak M, Bragaglia PJ, Tamari IE, Hodder R, Stanbrook MB. Prevalence and underdiagnosis of chronic obstructive pulmonary disease among patients at risk in primary care. CMAJ 2010;182(7):673–8. [PubMed: 20371646]
37. Lamprecht B, Soriano JB, Studnicka M, Kaiser B, Vanfleteren LE, Gnatiuc L, Burney P, Miravitlles M, Garcia-Rio F, Akbari K, Ancochea J, Menezes AM, Perez-Padilla R, Montes de Oca M, Torres-Duque CA, Caballero A, Gonzalez-Garcia M, Buist S, Bold Collaborative Research Group tEPISTtPT, the PSG. Determinants of underdiagnosis of COPD in national and international surveys. Chest 2015;148(4):971–985. [PubMed: 25950276]
38. Murgia N, Brisman J, Claesson A, Muzi G, Olin AC, Toren K. Validity of a questionnaire-based diagnosis of chronic obstructive pulmonary disease in a general population-based study. BMC Pulm Med 2014;14:49. [PubMed: 24650114]
39. Barker DJ, Godfrey KM, Fall C, Osmond C, Winter PD, Shaheen SO. Relation of birth weight and childhood respiratory infection to adult lung function and death from chronic obstructive airways disease. BMJ 1991;303(6804):671–5. [PubMed: 1912913]
40. Lamprecht B, McBurnie MA, Vollmer WM, Gudmundsson G, Welte T, Nizankowska-Mogilnicka E, Studnicka M, Bateman E, Anto JM, Burney P, Mannino DM, Buist SA, Group BCR. COPD in never smokers: results from the population-based burden of obstructive lung disease study. Chest 2011;139(4):752–763. [PubMed: 20884729]
41. Prescott E, Lange P, Vestbo J. Socioeconomic status, lung function and admission to hospital for COPD: results from the Copenhagen City Heart Study. Eur Respir J 1999;13(5):1109–14. [PubMed: 10414412]
42. Fuller-Thomson E, Chisholm RS, Brennenstuhl S. COPD in a Population-Based Sample of Never-Smokers: Interactions among Sex, Gender, and Race. Int J Chronic Dis 2016;2016:5862026. [PubMed: 28054032]
43. Salvi SS, Barnes PJ. Chronic obstructive pulmonary disease in non-smokers. Lancet 2009;374(9691):733–43. [PubMed: 19716966]
44. Schneider A, Gindner L, Tilemann L, Schermer T, Dinant GJ, Meyer FJ, Szecsenyi J. Diagnostic accuracy of spirometry in primary care. BMC Pulm Med 2009;9:31. [PubMed: 19591673]
45. Coxson HO, Leipsic J, Parraga G, Sin DD. Using pulmonary imaging to move chronic obstructive pulmonary disease beyond FEV1. Am J Respir Crit Care Med 2014;190(2):135–44. [PubMed: 24873985]
46. de Marco R, Accordini S, Cerveri I, Corsico A, Sunyer J, Neukirch F, Kunzli N, Leynaert B, Janson C, Gislason T, Vermeire P, Svanes C, Anto JM, Burney P, European Community Respiratory Health Survey Study G. An international survey of chronic obstructive pulmonary disease in young adults according to GOLD stages. Thorax 2004;59(2):120–5. [PubMed: 14760151]
47. Borlee F, Yzermans CJ, Krop E, Aalders B, Rooijackers J, Zock JP, van Dijk CE, Maassen CB, Schellevis F, Heederik D, Smit LA. Spirometry, questionnaire and electronic medical record based COPD in a population survey: Comparing prevalence, level of agreement and associations with potential risk factors. PLoS One 2017;12(3):e0171494. [PubMed: 28273094]

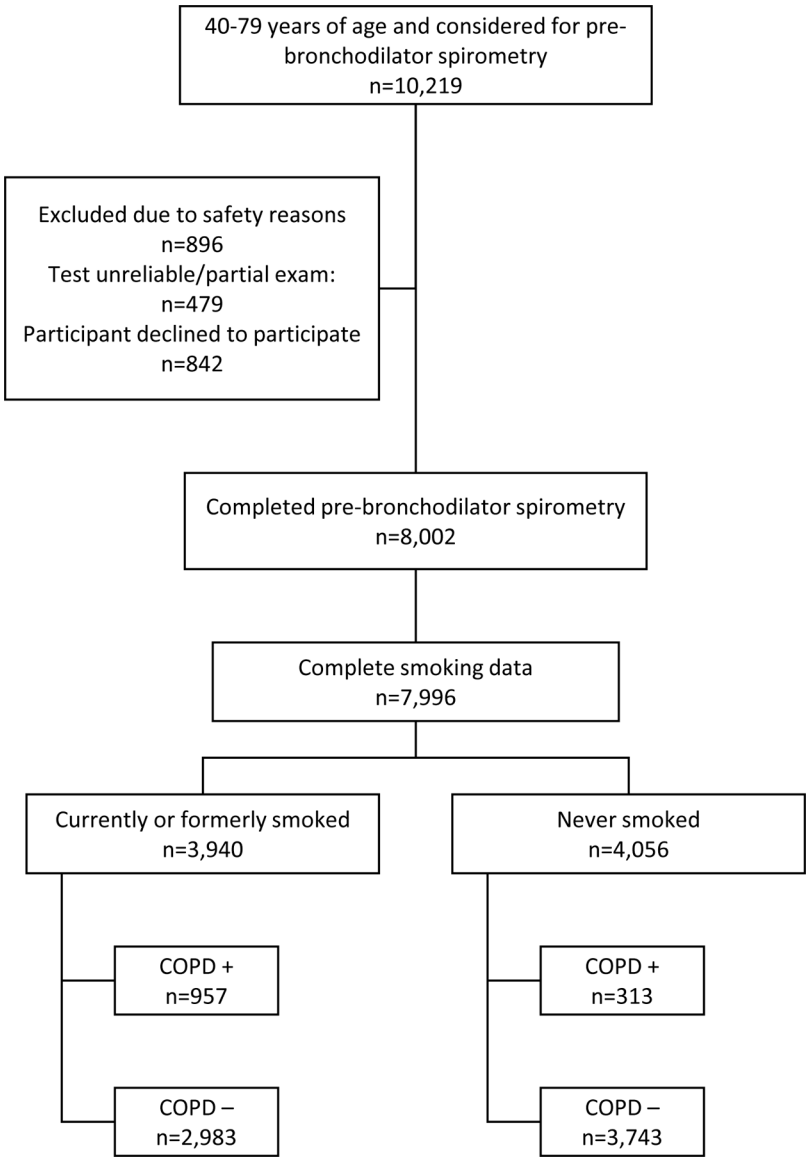


Figure 1.
Study sample selection.

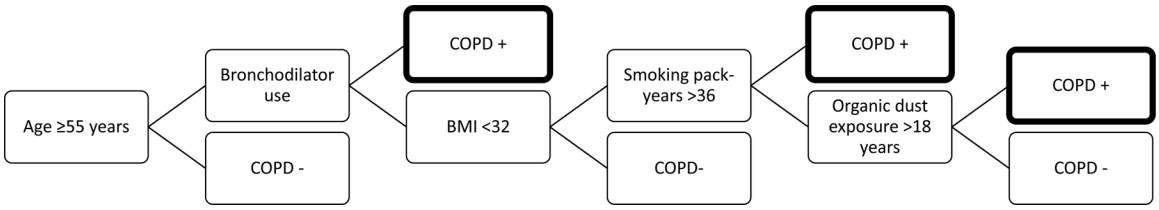


Figure 2.
COPD case definition for former and current smokers that was developed empirically using gradient boosting.

Table 1.

Distribution of participant characteristics by smoking and COPD status (weighted percentage or median with corresponding IQR).

	Smokers			Never smokers		
	Overall	**COPD −**	**COPD +**	**Overall**	**COPD −**	**COPD +**
Weighted median (IQR)						
Age (years)	54 [47–63]	52 [46–60]	59 [52–67]	52 [46–61]	52 [45–60]	60 [50–68]
Family income to poverty ratio	3 [2–5]	3 [2–5]	3 [2–5]	4 [2–5]	4 [2–5]	4 [2–5]
Body mass index (kg/m^2)	28 [25–32]	28 [25–33]	27 [24–31]	29 [25–33]	29 [25–33]	27 [24–30]
Pack-years of smoking	16 [5–33]	14 [4–28]	26 [10–45]	NA	NA	NA
Occupational exposures						
Years exposed to mineral dusts (dust from rock, sand, concrete, coal, asbestos, silica or soil)	0 [0–8]	0 [0–6]	0 [0–15]	0 [0–0]	0 [0–0]	0 [0–0]
Years exposed to organic dusts (dust from baking flours, grains, wood, cotton, plants or animals)	0 [0–2]	0 [0–1]	0 [0–9]	0 [0–0]	0 [0–0]	0 [0–0]
Years exposed to exhaust fumes from trucks, buses, heavy machinery or diesel engines	0 [0–5]	0 [0–3]	0 [0–14]	0 [0–0]	0 [0–0]	0 [0–1]
Years exposed to other fumes gases, vapors, or fumes	0 [0–8]	0 [0–6]	0 [0–13]	0 [0–1]	0 [0–1]	0 [0–2]
Weighted percentage						
Male	55	52	62	42	42	50
Hispanic ethnicity	9	11	3	11	12	6
Educational attainment						
Did not graduate high school	21	21	21	13	13	11
High school graduate/GED	25	24	29	21	22	18
Some college/associates degree	31	31	30	27	27	25
College graduate or higher	23	24	19	39	39	45
Self-rated health						
Excellent or very good	39	40	33	52	52	55
Good	42	42	43	35	35	30
Fair or poor	19	18	23	13	13	15
Any second-hand smoke exposure	32	30	38	9	9	9
Medical history						
Physician-diagnosed emphysema: ever	4	2	10	0	0	1
Physician-diagnosed chronic bronchitis: current	4	3	7	2	1	6
Physician-diagnosed asthma: ever	14	11	21	12	11	23

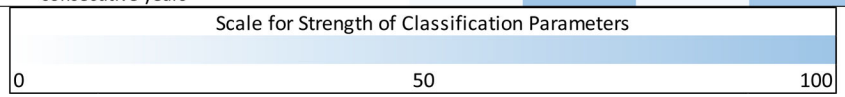
	Smokers			Never smokers		
	Overall	**COPD −**	**COPD +**	**Overall**	**COPD −**	**COPD +**
Hay fever: past year	21	21	22	21	20	27
Respiratory symptoms						
Chronic cough	13	10	20	5	5	4
Chronic phlegm	10	7	18	3	3	3
Shortness of breath when walking on level ground or up a slight hill	35	30	47	23	23	29
Wheezing in chest in the past year	19	15	28	8	8	15
Wheezing disturbs sleep	8	7	13	4	4	6
Wheeze during exercise	8	6	13	4	3	9
Bronchodilator use	7	4	16	3	3	9
Wheezing limits physical activity	7	5	11	4	3	8

Abbreviations: COPD, chronic obstructive pulmonary disease; IQR, interquartile range

Table 2.

Prevalence of COPD and classification parameters for questionnaire-based definitions.

Approach	Case Definition	Prevalence (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	Accuracy [a] (%)
Former and current smokers (n=3,940)							
Gradient boosting	• Age ≥55, and any one of • Bronchodilator use, or • BMI <32 and >36 pack-year history, or • BMI <32 and >18 years organic dust exposure	15 (13-17)	35 (29-41)	92 (90-93)	59 (53-64)	81 (78-83)	78 (75-80)
Mendy et al.[5]	• Affirmative response to physician-diagnosed chronic bronchitis or emphysema	11 (9-12)	18 (14-22)	92 (90-93)	42 (35-49)	77 (75-78)	73 (71-75)
Fessler et al.[14]	• Affirmative response to physician-diagnosed chronic bronchitis or emphysema, or • ≥10 pack-year history of smoking and affirmative response to coughing up phlegm on most days for ≥3 consecutive months for ≥2 consecutive years	16 (14-18)	27 (23-32)	88 (86-90)	44 (38-49)	78 (76-80)	72 (71-74)
Never smokers (n=4,056)							
Gradient boosting	• Age ≥55, and any one of • Bronchodilator use, or • BMI <32 and >36 pack-year history, or • BMI <32 and >18 years organic dust exposure	4 (3-5)	10 (4-16)	96 (95-97)	18 (10-26)	93 (92-94)	90 (88-91)
Mendy et al.[5]	• Affirmative response to physician-diagnosed chronic bronchitis or emphysema	5 (4-6)	9 (4-13)	96 (95-97)	14 (7-21)	93 (92-94)	89 (88-90)
Fessler et al.[14]	• Affirmative response to physician-diagnosed chronic bronchitis or emphysema, or • ≥10 pack-year history of smoking and affirmative response to coughing up phlegm on most days for ≥3 consecutive months for ≥2 consecutive years	5 (4-6)	9 (4-13)	96 (95-97)	14 (7-21)	93 (92-94)	89 (88-90)



Abbreviations: COPD, chronic obstructive pulmonary disease; BMI, body mass index; PPV, positive predictive value; NPV, negative predictive value

[a] Accuracy is the proportion of true classifications (true positives + true negatives) among all classifications (true positives + true negatives + false positives + false negatives)