Original Paper

Machine Learning Methods for the Diagnosis of Chronic Obstructive Pulmonary Disease in Healthy Subjects: Retrospective Observational Cohort Study

Shigeo Muro[1], MD, PhD; Masato Ishida[2], MSc; Yoshiharu Horie[3], PhD; Wataru Takeuchi[4], MEng; Shunki Nakagawa[4], MSc; Hideyuki Ban[4], PhD; Tohru Nakagawa[5], MD, PhD; Tetsuhisa Kitamura[6], MD, MSc, DPH

[1]Department of Respiratory Medicine, Nara Medical University, Nara, Japan

[2]Department of Respiratory and Immunology, Medical, AstraZeneca KK, Osaka, Japan

[3]Department of Data Science, Medical, AstraZeneca KK, Osaka, Japan

[4]Center for Technology Innovation–Artificial Intelligence, Research & Development Group, Hitachi, Ltd, Tokyo, Japan

[5]Hitachi Health Care Center, Hitachi, Ltd, Ibaraki, Japan

[6]Division of Environmental Medicine and Population Sciences, Department of Social and Environmental Medicine, Graduate School of Medicine, Osaka University, Osaka, Japan

Corresponding Author:
Yoshiharu Horie, PhD
Department of Data Science, Medical
AstraZeneca KK
3-1, Ofuka-cho, Kita-ku
Osaka, 5300011
Japan
Phone: 81 81 6 4802 3600
Email: yoshiharu.horie@astrazeneca.com

Abstract

Background: Airflow limitation is a critical physiological feature in chronic obstructive pulmonary disease (COPD), for which long-term exposure to noxious substances, including tobacco smoke, is an established risk. However, not all long-term smokers develop COPD, meaning that other risk factors exist.

Objective: This study aimed to predict the risk factors for COPD diagnosis using machine learning in an annual medical check-up database.

Methods: In this retrospective observational cohort study (ARTDECO [Analysis of Risk Factors to Detect COPD]), annual medical check-up records for all Hitachi Ltd employees in Japan collected from April 1998 to March 2019 were analyzed. Employees who provided informed consent via an opt-out model were screened and those aged 30 to 75 years without a prior diagnosis of COPD/asthma or a history of cancer were included. The database included clinical measurements (eg, pulmonary function tests) and questionnaire responses. To predict the risk factors for COPD diagnosis within a 3-year period, the Gradient Boosting Decision Tree machine learning (XGBoost) method was applied as a primary approach, with logistic regression as a secondary method. A diagnosis of COPD was made when the ratio of the prebronchodilator forced expiratory volume in 1 second (FEV_1) to prebronchodilator forced vital capacity (FVC) was <0.7 during two consecutive examinations.

Results: Of the 26,101 individuals screened, 1213 met the exclusion criteria, and thus, 24,815 individuals were included in the analysis. The top 10 predictors for COPD diagnosis were FEV_1/FVC, smoking status, allergic symptoms, cough, pack years, hemoglobin A_{1c}, serum albumin, mean corpuscular volume, percent predicted vital capacity, and percent predicted value of FEV_1. The areas under the receiver operating characteristic curves of the XGBoost model and the logistic regression model were 0.956 and 0.943, respectively.

Conclusions: Using a machine learning model in this longitudinal database, we identified a number of parameters as risk factors other than smoking exposure or lung function to support general practitioners and occupational health physicians to predict the development of COPD. Further research to confirm our results is warranted, as our analysis involved a database used only in Japan.

(JMIR Med Inform 2021;9(7):e24796) doi: 10.2196/24796

KEYWORDS

chronic obstructive pulmonary disease; airflow limitation; medical check-up; Gradient Boosting Decision Tree; logistic regression

Introduction

Chronic obstructive pulmonary disease (COPD) is characterized by airflow limitation associated with persistent respiratory symptoms. Most patients with COPD experience exacerbation of symptoms and are at high risk of developing comorbidities such as cardiovascular disease [1].

Long-term exposure to tobacco smoke, vapor, gas, dust, and fumes is an established major risk factor for COPD [2]. However, only a small percentage of smokers develop airflow limitation, while nonsmokers can develop COPD [3]. These inconsistencies indicate that risk factors other than long-term smoking are associated with COPD [4].

The prevalence of COPD has been reported to be 12% to 13% among smokers [5]. However, only 9.4% of patients with airflow limitation have a previous diagnosis of COPD, and European data indicate that up to 80% of COPD cases are undiagnosed [6], suggesting delays in the diagnosis of COPD. The ARCTIC observational cohort study showed that late COPD diagnosis was associated with a higher exacerbation rate and increased comorbidities and costs compared with early diagnosis [7].

To address the issue of undiagnosed COPD, significant risk factors for airflow limitation other than smoking should be identified and evaluated in routine clinical practice. In a cohort study of 9040 individuals from the Japanese general population, concomitant *Chlamydia pneumoniae* and *Mycoplasma pneumoniae* seropositivity was found to be an independent risk factor for airflow limitation [8]. Additionally, Sato et al employed an annual health examination with pulmonary function tests measuring airflow limitation to identify undiagnosed patients with COPD among the Japanese population and found that iron deficiency might be associated with COPD development [9]. However, the follow-up duration of these cohorts was short (<3 years), limiting their ability to identify risk factors for COPD in the general population.

A large questionnaire-based surveillance demonstrated some improvement in diagnostic rates for COPD; however, approximately 60% of eligible participants failed to respond to the questionnaire [10]. While these results suggest that identifying robust and relevant risk factors is likely to improve early diagnosis, the slow progression and heterogeneity of the disease have hindered the identification of such risk factors for COPD development.

The recently reported "Subtype and Stage Inference" machine learning computational model identified subtypes of patients with COPD [11]. Compared with traditional approaches, the advantages of machine learning include the ability to process complex nonlinear relationships between predictors and to provide novel outputs. Therefore, the aim of this study was to apply machine learning methods to predict possible risk factors for the development of airflow limitation, an essential feature of COPD diagnosis, using a Japanese medical check-up database comprising data from a number of healthy subjects to support the early diagnosis of COPD by general practitioners and occupational health physicians.

Methods

Study Design and Population

This was a retrospective observational cohort study to predict the risk factors for COPD diagnosis in healthy individuals. The analysis data set comprised individuals aged ≥30 years who had undertaken more than two medical check-ups, had no history of lung cancer or asthma at the first medical check-up, and could be classified as either having a diagnosis of COPD or as not having COPD. This study was designed according to the *Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Diagnosis* guidelines for prognostic studies [12].

The study protocol was reviewed and approved by the ethics committee of MINS (a nonprofit organization in Tokyo, Japan) and the Research & Development Group and Corporate Hospital Group of Hitachi, Ltd (Tokyo, Japan) prior to the start of data analysis. Individual informed consent was obtained using an opt-out model in agreement with the Institutional Review Board at Hitachi, Ltd. This study was conducted in accordance with the ethical principles of the Declaration of Helsinki.

Data Source

The data source was annual medical check-up data for all Hitachi employees from April 1998 to March 2019. Data were archived in a high-security server that was managed with limited access rights by Hitachi. The annual medical check-up includes clinical measurements and questionnaires to examine the health of employees (Multimedia Appendix 1). Such questionnaires are utilized by Japanese organizations to evaluate their employees' health and give advice about health promotion, such as giving up smoking and exercising regularly based on the second term of the National Health Promotion Movement in the 21st century (Health Japan 21) issued by the Ministry of Health, Labour, and Welfare in Japan [13].

Definition of COPD

COPD was considered according to the lung function status at two consecutive measurements during an annual lung function test when the prebronchodilator (pre-BD) forced expiratory volume in 1 second/forced vital capacity (FEV_1/FVC) was <0.7, as previously employed in a large population-based cohort study [14]. Individuals having a pre-BD FEV_1/FVC ≥0.7 in at least three consecutive annual lung function test measurements were classified as non-COPD. For individuals with more than three records in the non-COPD group, the most recent three records were analyzed. Individuals having less than two lung function tests were excluded from all analyses. Spirometry was calibrated and performed by trained paramedical personnel according to

the American Thoracic Society/European Respiratory Society guidelines [15,16].

Statistical Analysis

Age at COPD Diagnosis

The age distribution for disease diagnosis was evaluated and stratified by smoking status (current smoker, exsmoker, or nonsmoker). The age at COPD diagnosis was defined as the age at the first of two consecutive measurements in which the pre-BD FEV_1/FVC was <0.7.

Risk Factor Prediction Using Machine Learning

Two types of models were constructed for predicting the risk factors for COPD diagnosis within 3 years as follows: a machine learning method (Gradient Boosting Decision Tree machine learning [XGBoost] [17]) and an established statistical method (logistic regression [18]). Individuals who did not meet the study inclusion criteria and/or had lung cancer/asthma were excluded from the analyses. Any individuals with missing data during the 3 years prior to the diagnosis year in the COPD group or during the most recent 3 years in the non-COPD group were excluded from the analyses. Propensity scores were calculated based on age, sex, smoking status, BMI, eosinophil count (EOS), and FEV_1.

Data were randomly divided into a training data set and a test data set at a ratio of 7:3, with the same ratio of COPD to non-COPD individuals. Propensity scoring was used to balance the characteristics of COPD and non-COPD individuals (caliper: 0.2) in the training and test data sets. Next, the training data set was randomly divided 8:2 for model construction (XGBoost and logistic) and evaluation of model performance, respectively. The data split, model construction, and evaluation processes were repeated five times for cross-validation (5-CV approach) [19]. Model parameters, including the depth of the tree and regularization factor, were refined during performance evaluation by the 5-CV approach. Finally, the most optimal model was generated by applying the best parameters confirmed by the 5-CV approach. To evaluate model performance in the unlearned data, the most optimized model was used to evaluate the test data set.

Model construction by logistic regression was performed in a similar way to the XGBoost method. Models were constructed in the training data set (randomly sampled data from the entire data set) and subsequently validated in the test data set after model evaluation.

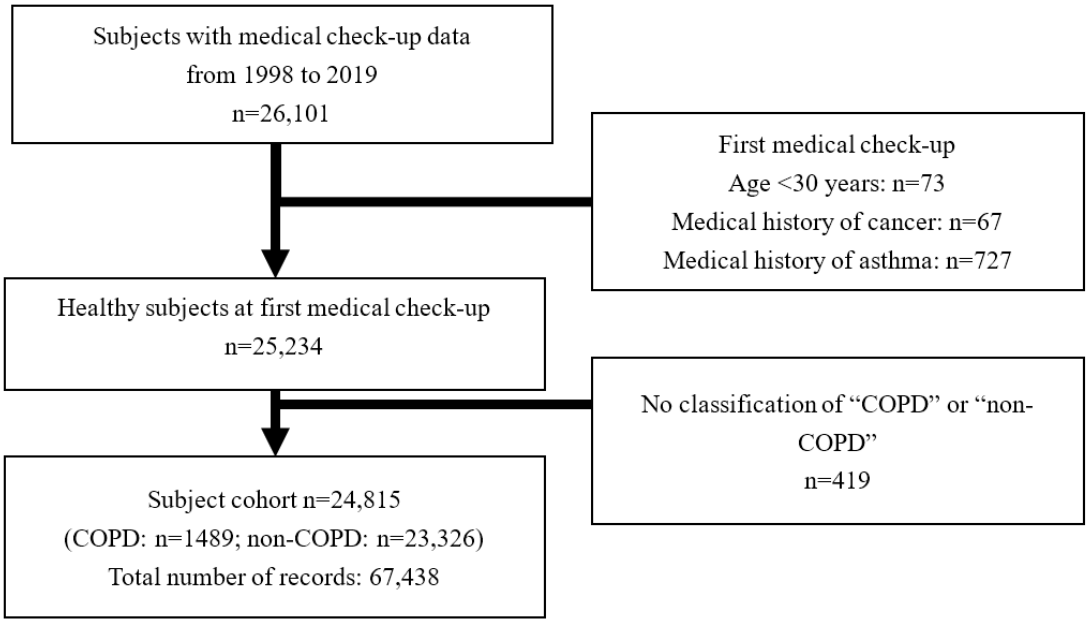
Following model construction, the area under the receiver operating characteristic curve (AUC), positive predictive value, sensitivity, specificity, and F1-measure were calculated for each model to evaluate the performance under both 5-CV and test conditions [20]. The feature importance of the machine learning model was calculated to examine the contribution of each predictor to the model constructed using the Gini impurity method [19]. The feature weight of the logistic regression model was also calculated. All analyses were performed using Python 3.6 software (Python Software Foundation).

Results

Individuals

Data from 26,101 individuals (employees and their families) aged 30 to 75 years, who underwent annual check-ups between April 1998 and March 2019 were included in our analysis. The total number of medical check-up records was 318,568. All 26,101 individuals had lung function test measurements for 3 consecutive years. The medical records for 73 individuals aged <30 years at the first medical check-up, 67 individuals with a history of cancer, and 727 individuals with a history of asthma were excluded, as were data from 419 individuals who had already been diagnosed with COPD (subjects for whom all data points of pre-BD FEV_1/FVC were <0.7 during the observational period) or had not been classified as either COPD or non-COPD (subjects with pre-BD FEV_1/FVC <0.7 without two consecutive measurements). Accordingly, data for 24,815 individuals (corresponding to 67,438 records) were included in the analyses (Figure 1).

Figure 1. Flow diagram of the study. COPD: chronic obstructive pulmonary disease.



Baseline Characteristics

Table 1 shows the baseline characteristics of the COPD and non-COPD groups. Overall, 1489 individuals were considered as having COPD (pre-BD FEV_1/FVC <0.7 at two consecutive measurements during annual lung function tests). In comparison with the non-COPD group, the COPD group had a lower BMI, worse lung function (pre-BD FEV_1, pre-BD percent predicted value of FEV_1 [%FEV_1], and pre-BD FEV_1/FVC), and greater emphysematous change and chronic inflammation as determined by computed tomography. Furthermore, comorbidities, such as arrythmia, duodenal ulcer, colorectal polyp, angina, stomach ulcer, and kidney disease, were more prevalent in the COPD group. Statistically significant differences in hematological parameters (mean corpuscular volume [MCV], mean corpuscular hemoglobin concentration [MCHC], mean corpuscular hemoglobin [MCH], hemoglobin [Hb], and hematocrit [HT] [15]) between the COPD and non-COPD groups were also observed. Inflammatory markers, particularly white blood cell (WBC) count and EOS, were also significantly higher in the COPD group.

Table 1. Subject characteristics stratified by chronic obstructive pulmonary disease status.

Characteristic	Non-COPD[a] (n=23,326)	COPD (n=1489)	*P* value
Age (years), mean (SD)	42 (9.1)	48 (9.3)	<.001
Female, n (%)	3841 (16.5%)	58 (3.9%)	<.001
Smoking status, n (%)			<.001
Current smoker	10,632 (45.6%)	1,021 (68.6%)	
Exsmoker	3534 (15.2%)	202 (13.6%)	
Nonsmoker	9153 (39.3%)	266 (17.9%)	
Unknown/missing	7 (0.0%)	0 (0.0%)	
BMI (kg/m^2), mean (SD)	23 (3.2)	22 (2.7)	<.001
Lung function test, mean (SD)			
Prebronchodilator FEV_1[b]	3.4 (0.7)	3.1 (0.6)	<.001
Prebronchodilator FVC[c]	4.1 (0.8)	4.2 (0.8)	<.001
Prebronchodilator FEV_1/FVC	83.7 (5.4)	74.9 (5.1)	<.001
Comorbidity, n (%)			
Arrythmia	107 (0.5%)	16 (1.1%)	.003
Duodenal ulcer	158 (0.7%)	19 (1.3%)	.02
Colorectal polyp	43 (0.2%)	13 (0.9%)	<.001
Angina	56 (0.2%)	10 (0.7%)	.006
Stomach ulcer	180 (0.8%)	29 (1.9%)	<.001
Kidney disease	77 (0.3%)	12 (0.8%)	.01
Computed tomography finding, n (%)			
Bulla, bleb	108 (0.5%)	31 (2.1%)	<.001
Moderate emphysema	18 (0.1%)	13 (0.9%)	<.001
Mild emphysema	96 (0.4%)	27 (1.8%)	<.001
Calcification of left anterior descending coronary artery	128 (0.5%)	16 (1.1%)	.02
Chronic inflammation	342 (1.5%)	43 (2.9%)	<.001
Laboratory parameters, mean (SD)			
Albumin (U/L)	4.4 (0.2)	4.3 (0.2)	<.001
Alanine aminotransferase (U/L)	209.5 (53.7)	215.2 (54.6)	<.001
Aspartate aminotransferase (U/L)	26.4 (14.8)	24.3 (12.8)	<.001
Blood urea nitrogen (mg/dL)	14.1 (3.2)	14.7 (3.3)	<.001
Cholinesterase (U/L)	320.8 (60.1)	307.8 (58.8)	<.001
Estimated glomerular filtration rate ($mL/min/1.73\ m^2$)	83.6 (14.6)	80.2 (14.1)	<.001
Eosinophil count ($cells/mm^3$)	183.2 (124.5)	195.4 (125.9)	<.001
Gamma-glutamyl transferase (U/L)	42.7 (34.4)	45.8 (34.5)	<.001
Hemoglobin (g/dL)	14.7 (1.4)	14.9 (1.1)	<.001
Hemoglobin A_{1c} (%)	5.3 (0.7)	5.4 (0.7)	<.001
Hematocrit (%)	44.0 (3.6)	44.6 (3.1)	<.001
MCH[d] (pg)	30.5 (1.8)	31.1 (1.7)	<.001
MCHC[e] (g/L)	33.4 (1.0)	33.3 (0.8)	<.001
MCV[f] (fL)	91.3 (4.6)	93.4 (4.4)	<.001

Characteristic	Non-COPD[a] (n=23,326)	COPD (n=1489)	*P* value
WBC[g] count (×10^2 cells/μL)	58.8 (15.0)	62.8 (15.5)	<.001

[a]COPD: chronic obstructive pulmonary disease.

[b]FEV_1: forced expiratory volume in 1 second.

[c]FVC: forced vital capacity.

[d]MCH: mean corpuscular hemoglobin.

[e]MCHC: mean corpuscular hemoglobin concentration.

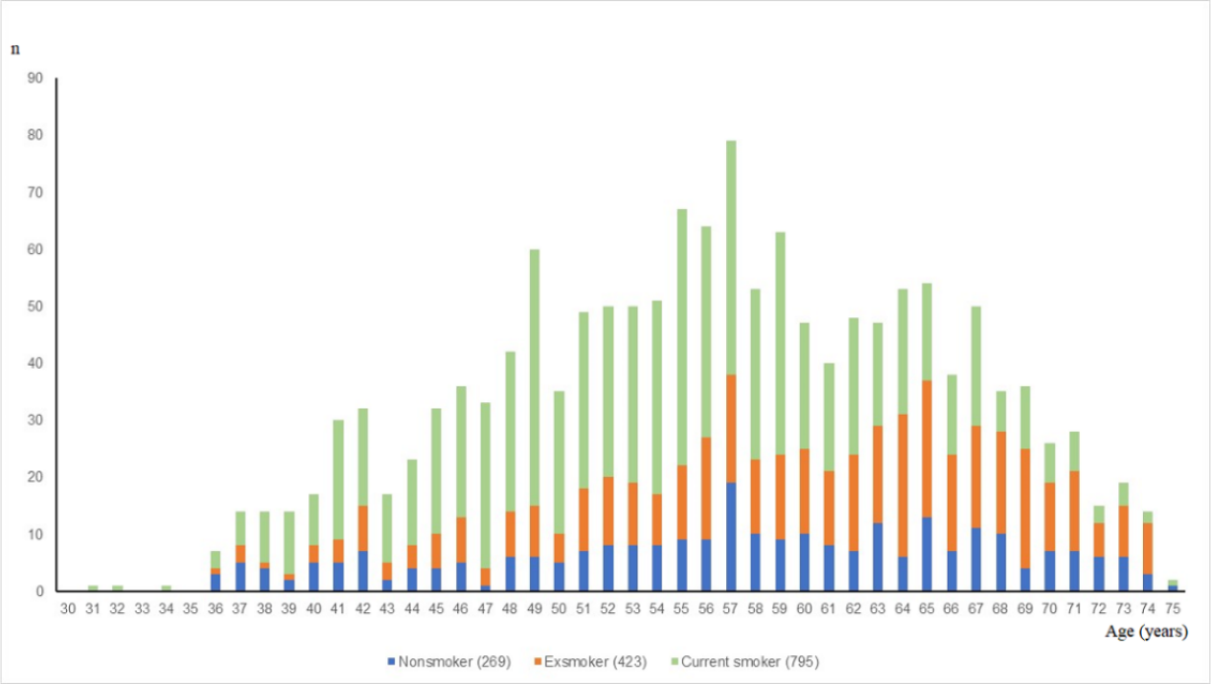
[f]MCV: mean corpuscular volume.

[g]WBC: white blood cell.

Percentage of Individuals With COPD

The overall percentage of individuals with COPD was 6.0% (1489/24,815). According to smoking status, the percentage of individuals with COPD was 8.8% (1021/11,653) among current smokers and 5.4% (202/3736) among exsmokers. Notably, 2.8% (266/9419) of nonsmokers had developed COPD. The peak age at diagnosis of COPD among current smokers and exsmokers was 55 years and 65 years, respectively (Figure 2).

Figure 2. Diagnostic age for chronic obstructive pulmonary disease (COPD) according to smoking status.



Risk Factors for COPD Diagnosis

Overall, 20,265 individuals (COPD: n=954; non-COPD: n=19,311) with 51,432 records (COPD: n=2435; non-COPD: n=48,997) out of 24,815 individuals who met the criteria (Multimedia Appendix 2) were included in the machine learning analysis. Table 2 shows the model performance of the XGBoost and logistic regression models. For both models, the AUC, accuracy, sensitivity, specificity, and F-measure were generally similar between the training and test data sets. The XGBoost model had a higher positive predictive value (0.505) than the logistic regression model (0.441). The AUC was high in the training and test sets for both models (range: 0.892-0.956). Additionally, the accuracy and specificity exceeded 0.883 and 0.879, respectively, for both models.

The most important predictive factors for COPD diagnosis were lung function tests (ie, FEV_1/FVC, percent vital capacity [%VC], and $\%FEV_1$) and smoking status, followed by cough, hematological indices (ie, MCV, MCHC, MCH, Hb, and HT), treatment with antidiabetic drugs, hemoglobin A_{1c}, serum albumin, total protein, and BMI. Other predictive risk factors were EOS, serum alanine aminotransferase, WBC count, and urinary WBC count (Table 3). Logistic regression analysis showed that low FEV_1/FVC and $\%FEV_1$; high %VC; high MCV, MCHC, and Hb; and low HT and MCH were related factors, and that individuals treated with antidiabetic drugs had a higher number of associated risk factors for COPD. Low serum albumin, low total protein, and low BMI were also confirmed as risk factors (Multimedia Appendix 3).

Table 2. Comparison of performance of the Gradient Boosting Decision Tree machine learning (XGBoost) and logistic regression models.

Variable	XGBoost[a] model		Logistic regression model	
	Training, mean (SE)	Test, mean	Training, mean (SE)	Test, mean
Positive predictive value	0.505 (0.099)	0.362	0.441 (0.110)	0.285
AUC[b]	0.956 (0.015)	0.898	0.943 (0.022)	0.892
Accuracy	0.917 (0.032)	0.918	0.884 (0.049)	0.883
Sensitivity	0.845 (0.021)	0.877	0.874 (0.039)	0.901
Specificity	0.960 (0.016)	0.919	0.946 (0.025)	0.882
F-measure	0.370 (0.107)	0.513	0.306 (0.110)	0.434

[a]XGBoost: Gradient Boosting Decision Tree machine learning.

[b]AUC: area under the receiver operating characteristic curve.

Table 3. Importance of each predictor in the XGBoost model.

Variable	Importance value
Forced expiratory volume in 1 second/forced vital capacity	0.2824
Smoking status	0.0329
Allergic symptoms (yes/no)	0.0303
Symptom-cough (yes/no)	0.0294
Smoking-pack year	0.0222
Hemoglobin A_{1c}	0.0197
Albumin	0.0195
Mean corpuscular volume	0.0177
%Vital capacity	0.0165
%Forced expiratory volume in 1 second	0.0164
Treatment with an antidiabetic drug (yes/no)	0.0162
Allergic disease (yes/no)	0.0146
Hematocrit	0.0144
Urinary red blood cells	0.0143
Hemoglobin	0.0138
Age	0.0128
Smoking duration	0.0127
High density lipoprotein cholesterol	0.0123
Mean corpuscular hemoglobin concentration	0.0122
Total protein	0.0118
BMI	0.0118
Number of eosinophils	0.0115
Mean corpuscular hemoglobin	0.0114
Serum white blood cells	0.0111
Fasting blood sugar	0.0110
Serum alanine aminotransferase	0.0108
Pulse rate	0.0108
Forced expiratory volume in 1 second	0.0107
Urinary white blood cells	0.0104
Diastolic blood pressure	0.0103

For future utilization of risk factors for disease assessment in daily clinical practice, the machine learning process was validated using a questionnaire to predict risk factors for COPD development (Multimedia Appendix 1). Of 30 variables, 25 were clinical parameters that overlapped between the two methods. The top 30 risk factors also included the following five questions: “I am regularly doing exercise,” “I have chest compression and pain,” “Average sleeping time in the past 1 month,” “I have breakfast every day,” and “Body fat ratio” (Multimedia Appendix 4). Among these, logistic regression analysis showed that insufficient sleeping time and not having breakfast every day were risk factors for COPD (Multimedia Appendix 3).

Discussion

This study applied a machine learning method, a powerful tool to analyze large quantities of complex data, to predict risk factors for COPD. This is the first study to investigate more than 300,000 records from working-age adults in Japan utilizing an annual medical check-up database. This system allows healthy employees to track their health conditions over time by clinical measurements and questionnaires. We found that the most significant predictor of COPD diagnosis was the absolute value of FEV_1/FVC, indicating that low FEV_1 in early adulthood is an important factor in the development of COPD. Childhood asthma is associated with impaired lung function, lower lung function in adulthood, and higher risk of COPD even for nonsmoker participants, as previously reported by Martinez et

al [21]. In our speculation, some part of the nonsmoker COPD population might have had a history of childhood asthma, increasing susceptibility to passive smoke exposure or airway pollution and resulting in the early diagnosis of COPD in nonsmokers compared with exsmokers in the study. Smoking status had the second highest impact on disease diagnosis. Among individuals with a smoking history, the peak age of COPD diagnosis was older in exsmokers than in current smokers. This finding suggests that smoking cessation delays the diagnosis of COPD, consistent with a previous study in which smoking cessation was reported to affect the natural history of COPD [22].

Erythrocyte indices (MCV and MCHC) might also be available as potential predictors of COPD diagnosis in addition to lung function measurements. These data are supported by a previous report in which continuous smoking had a significant effect on hematological parameters compared with nonsmoking, and it may be associated with an increased risk of COPD [23]. The increased levels of MCV and MCHC in individuals with COPD support a previous finding that impaired lung function has a strong association with ischemic heart disease [24]. Conversely, the presence of an allergic disease appeared to have a preventive effect on airflow limitation, which is in contrast with observations from the Tasmanian Longitudinal Health Study in which the presence of allergic diseases was an early predictor of lung trajectories toward COPD [25]. However, the Hokkaido cohort study showed that subjects with multiple asthma-like features had slower lung function decline [26]. From the findings of our observational study in Japan, we can speculate that early diagnosis and intervention for allergic diseases may have less impact on lung function and that regular and frequent medical intervention could lead to an overall increase in life expectancy among patients who can readily access appropriate treatment by respiratory specialists.

Furthermore, individuals with decreased levels of serum albumin and total protein, as well as lower hemoglobin A_{1c} and BMI may be at risk of developing cachexia, a common condition among patients with COPD [27]. With respect to other identified risk factors, a retrospective cross-sectional study showed an association between EOS and airflow limitation in patients with COPD [28]. Given that increased alanine aminotransferase levels have been observed in patients with obstructive sleep apnea [29], individuals at risk of developing COPD might be exposed to intermittent hypoxia, indicating that a reduced sleeping time, as determined in the study questionnaire, might also represent a risk factor for COPD. Even minor changes in hematological parameters might be attributable to hypoxic conditions, leading to sleep disruption. Additionally, frequently missing breakfast might accelerate malnutrition in the COPD group. Furthermore, significantly higher prevalence rates of chronic neck and lower back pain in patients with COPD compared with healthy individuals were observed in a population-based study, although the findings were not confirmed by logistic regression analysis [30], and the link between COPD and back pain remains unknown. The observation of increased WBC counts in patients with COPD compared with healthy controls [31] suggests that systemic inflammation may be involved in the pathogenesis of COPD [32].

Our results also indicate that smoking cessation should be prioritized for the prevention of COPD and that smokers with sleep disturbances, back pain, and/or low BMI and malnutrition may be at increased risk of developing COPD and should be considered as candidates for lifestyle intervention therapy. Furthermore, the five key questions included in our questionnaire should be validated in future investigations and potentially implemented in daily practice as part of an annual medical check-up to prevent COPD.

The positive predictive value of the XGBoost model was comparable to that of a self-scored persistent airflow obstruction screening questionnaire in the Japanese population previously reported by Samukawa et al [33]. However, our models showed more accuracy because the sensitivity and specificity of our models achieved higher figures, and the AUC reached over 0.9 compared with that of the questionnaire, which ranged from 0.595 to 0.612. The AUCs of the XGBoost and logistic regression models were similar, while the most important factor related to COPD diagnosis was FEV_1 in both models. However, some variables differed in importance in each model. Kuhn et al reported that machine learning approaches can incorporate high-order nonlinear interactions among predictors that cannot be addressed by traditional modeling approaches (eg, logistic regression models) [34]. However, machine learning methods cannot elucidate whether a causal relationship exists between the identified variable and the disease. Thus, the association between risk factors detected using a machine learning model and COPD requires validation in future prospective studies.

A strength of this study was the use of longitudinal lung function test data from healthy individuals from April 1998 to March 2019. In general, medical checkup data are not linked to medical records, meaning that profiles of lung function tests over time could not be investigated. However, it was possible to evaluate longitudinal lung function tests because the database included data from individuals from a point in time when they were healthy until they had developed COPD. Additionally, data from healthy individuals were included, allowing lung function test results from when they were diagnosed with COPD to be investigated. Finally, both clinical measurements and questionnaire variables were included in the database, thereby increasing the potential to identify several different risk factors for COPD.

The limitations of this study include the definition of COPD diagnosis by airflow limitation with pulmonary function tests. Instead of post-BD spirometry data as suggested by the ATS/European respiratory guidelines, we employed pre-BD spirometry data for the diagnosis of COPD since no post-BD spirometry was performed in the annual medical check-up. The precise diagnosis of COPD cannot always be demonstrated by airflow limitation alone; however, we believe that the diagnostic approach was reasonable from a clinical perspective as airflow limitation has been reported to be a poor prognostic factor in the general population [35]. Low lung function values (FEV_1/FVC <0.7) might be observed at a single time point in some individuals for no discernable reason. Therefore, we

considered COPD as lung function of FEV_1/FVC <0.7 on two consecutive occasions. In terms of differentiation between asthma and COPD, we cannot exclude the possibility of misclassification of asthma as COPD in some patients since reversibility tests were not performed in the annual medical check-up, but participants with a medical history of asthma were excluded. Additionally, a database from a single organization was analyzed in this study; thus, the results might include bias based on the type of industry or the organizational structure of the company, limiting the generalizability of the findings. To obtain more generalizable findings, studies using other databases are necessary. Finally, some unknown confounders may have remained; therefore, we plan to perform model validation by analyzing other databases. Well-controlled prospective studies should be conducted to confirm the predictive factors for COPD diagnosis.

In conclusion, our machine learning method applied to longitudinal medical check-up data, including general questionnaires and laboratory parameters, identified hematological, nutritional, and inflammatory parameters as potential risk factors for COPD. These parameters, along with lung function and smoking status, may be useful in identifying at-risk individuals and may lead to an earlier diagnosis.

Acknowledgments

The authors thank ND Smith, Y Baba, and K Dohi of EMC Japan (Osaka, Japan) for critical reading and native checking of the manuscript. We also thank Tricia Newell and Clare Cox of Edanz Evidence Generation for providing editing support. This study was funded by AstraZeneca KK.

Authors' Contributions

SM and TK contributed to data interpretation and reviewed the manuscript. YH planned the analyses, contributed to data interpretation, and drafted the manuscript. MI designed the study and drafted the manuscript. SN, WT, and HB conducted the analyses. TN corrected and provided the data, and reviewed the analysis including data cleansing and preprocessing. All authors take full responsibility for the content and editorial decisions, and approved the final version.

Conflicts of Interest

SM has received honoraria from AstraZeneca KK, Boehringer Ingelheim Japan, GlaxoSmithKline KK, Novartis Pharma KK, Meiji Seika Pharma Co, Ltd, Kyorin Pharmaceutical Co, Ltd, Otsuka Pharmaceutical Co, Ltd, Teijin Pharma Ltd, CHEST MI, Inc, Daiichi Sankyo Co, Ltd, Chugai Pharmaceutical Co, Ltd, Sanofi KK, Actelion Pharmaceuticals Japan Ltd, and Olympus Corporation. MI and YH are employees of AstraZeneca KK. WT, SN, HB, and TN are employees of Hitachi, Ltd.

Multimedia Appendix 1

Clinical assessments and questions used in the analysis, and list of variables.
[DOCX File , 56 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Numbers of records and individuals included in the machine learning model.
[DOCX File , 42 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Association between chronic obstructive pulmonary disease and the top 30 variables of importance based on logistic regression.
[DOCX File , 45 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Importance of each predictor in the XGBoost model (including questionnaire items).
[DOCX File , 44 KB-Multimedia Appendix 4]

References

1. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. Am J Respir Crit Care Med 2017 Mar 01;195(5):557-582. [doi: 10.1164/rccm.201701-0218PP] [Medline: 28128970]
2. de Marco R, Accordini S, Marcon A, Cerveri I, Antó JM, Gislason T, European Community Respiratory Health Survey (ECRHS). Risk factors for chronic obstructive pulmonary disease in a European cohort of young adults. Am J Respir Crit Care Med 2011 Apr 01;183(7):891-897. [doi: 10.1164/rccm.201007-1125OC] [Medline: 20935112]
3. Fletcher C, Peto R. The natural history of chronic airflow obstruction. Br Med J 1977 Jun 25;1(6077):1645-1648 [FREE Full text] [doi: 10.1136/bmj.1.6077.1645] [Medline: 871704]

4. Stang P, Lydick E, Silberman C, Kempel A, Keating ET. The prevalence of COPD: using smoking rates to estimate disease frequency in the general population. Chest 2000 May;117(5 Suppl 2):354S-359S. [doi: 10.1378/chest.117.5_suppl_2.354s] [Medline: 10843976]
5. Fukuchi Y, Nishimura M, Ichinose M, Adachi M, Nagai A, Kuriyama T, et al. COPD in Japan: the Nippon COPD Epidemiology study. Respirology 2004 Nov;9(4):458-465. [doi: 10.1111/j.1440-1843.2004.00637.x] [Medline: 15612956]
6. Soriano JB, Zielinski J, Price D. Screening for and early detection of chronic obstructive pulmonary disease. The Lancet 2009 Aug 29;374(9691):721-732. [doi: 10.1016/S0140-6736(09)61290-3] [Medline: 19716965]
7. Larsson K, Janson C, Ställberg B, Lisspers K, Olsson P, Kostikas K, et al. Impact of COPD diagnosis timing on clinical and economic outcomes: the ARCTIC observational cohort study. Int J Chron Obstruct Pulmon Dis 2019;14:995-1008 [FREE Full text] [doi: 10.2147/COPD.S195382] [Medline: 31190785]
8. Muro S, Tabara Y, Matsumoto H, Setoh K, Kawaguchi T, Takahashi M, Nagahama Study Group. Relationship Among Chlamydia and Mycoplasma Pneumoniae Seropositivity, IKZF1 Genotype and Chronic Obstructive Pulmonary Disease in A General Japanese Population: The Nagahama Study. Medicine (Baltimore) 2016 Apr;95(15):e3371 [FREE Full text] [doi: 10.1097/MD.0000000000003371] [Medline: 27082601]
9. Sato K, Shibata Y, Inoue S, Igarashi A, Tokairin Y, Yamauchi K, et al. Impact of cigarette smoking on decline in forced expiratory volume in 1s relative to severity of airflow obstruction in a Japanese general population: The Yamagata-Takahata study. Respir Investig 2018 Mar;56(2):120-127. [doi: 10.1016/j.resinv.2017.11.011] [Medline: 29548649]
10. Jordan RE, Adab P, Sitch A, Enocson A, Blissett D, Jowett S, et al. Targeted case finding for chronic obstructive pulmonary disease versus routine practice in primary care (TargetCOPD): a cluster-randomised controlled trial. The Lancet Respiratory Medicine 2016 Sep;4(9):720-730. [doi: 10.1016/S2213-2600(16)30149-7] [Medline: 27444687]
11. Young AL, Bragman FJS, Rangelov B, Han MK, Galbán CJ, Lynch DA, COPDGene Investigators. Disease Progression Modeling in Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med 2020 Feb 01;201(3):294-302 [FREE Full text] [doi: 10.1164/rccm.201908-1600OC] [Medline: 31657634]
12. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015 Jan 06;162(1):55-63. [doi: 10.7326/M14-0697] [Medline: 25560714]
13. Sugiyama K, Tomata Y, Takemi Y, Tsushita K, Nakamura M, Hashimoto S, et al. Awareness and health consciousness regarding the national health plan "Health Japan 21" (2nd edition) among the Japanese population in 2013 and 2014. Nihon Koshu Eisei Zasshi 2016;63(8):424-431 [FREE Full text] [doi: 10.11236/jph.63.8_424] [Medline: 27681283]
14. Terzikhan N, Verhamme KMC, Hofman A, Stricker BH, Brusselle GG, Lahousse L. Prevalence and incidence of COPD in smokers and non-smokers: the Rotterdam Study. Eur J Epidemiol 2016 Aug;31(8):785-792 [FREE Full text] [doi: 10.1007/s10654-016-0132-z] [Medline: 26946425]
15. Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, ATS/ERS Task Force. Standardisation of spirometry. Eur Respir J 2005 Aug;26(2):319-338 [FREE Full text] [doi: 10.1183/09031936.05.00034805] [Medline: 16055882]
16. Celli BR, MacNee W, ATS/ERS Task Force. Standards for the diagnosis and treatment of patients with COPD: a summary of the ATS/ERS position paper. Eur Respir J 2004 Jun;23(6):932-946 [FREE Full text] [doi: 10.1183/09031936.04.00014304] [Medline: 15219010]
17. Chen T, Guestrin C. XGBoost : A Scalable Tree Boosting System. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
18. Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, et al. Strong rules for discarding predictors in lasso-type problems. J R Stat Soc Series B Stat Methodol 2012 Mar;74(2):245-266 [FREE Full text] [doi: 10.1111/j.1467-9868.2011.01004.x] [Medline: 25506256]
19. James G, Witten D, Hastie T, Tibshirani R. Introduction. In: An Introduction to Statistical Learning. Springer Texts in Statistics, vol 103. New York, USA: Springer; 2013.
20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [Medline: 3203132]
21. Martinez FJ, Han MK, Allinson JP, Barr RG, Boucher RC, Calverley PMA, et al. At the Root: Defining and Halting Progression of Early Chronic Obstructive Pulmonary Disease. Am J Respir Crit Care Med 2018 Jun 15;197(12):1540-1551 [FREE Full text] [doi: 10.1164/rccm.201710-2028PP] [Medline: 29406779]
22. Bai J, Chen X, Liu S, Yu L, Xu J. Smoking cessation affects the natural history of COPD. Int J Chron Obstruct Pulmon Dis 2017;12:3323-3328 [FREE Full text] [doi: 10.2147/COPD.S150243] [Medline: 29180862]
23. Malenica M, Prnjavorac B, Bego T, Dujic T, Semiz S, Skrbo S, et al. Effect of Cigarette Smoking on Haematological Parameters in Healthy Population. Med Arch 2017 Apr;71(2):132-136 [FREE Full text] [doi: 10.5455/medarh.2017.71.132-136] [Medline: 28790546]
24. Eriksson B, Lindberg A, Müllerova H, Rönmark E, Lundbäck B. Association of heart diseases with COPD and restrictive lung function--results from a population survey. Respir Med 2013 Jan;107(1):98-106 [FREE Full text] [doi: 10.1016/j.rmed.2012.09.011] [Medline: 23127573]

25. Bui DS, Lodge CJ, Burgess JA, Lowe AJ, Perret J, Bui MQ, et al. Childhood predictors of lung function trajectories and future COPD risk: a prospective cohort study from the first to the sixth decade of life. Lancet Respir Med 2018 Jul;6(7):535-544. [doi: 10.1016/S2213-2600(18)30100-0] [Medline: 29628376]
26. Suzuki M, Makita H, Konno S, Shimizu K, Kimura H, Kimura H, Hokkaido COPD Cohort Study Investigators. Asthma-like Features and Clinical Course of Chronic Obstructive Pulmonary Disease. An Analysis from the Hokkaido COPD Cohort Study. Am J Respir Crit Care Med 2016 Dec 01;194(11):1358-1365. [doi: 10.1164/rccm.201602-0353OC] [Medline: 27224255]
27. von Haehling S, Anker MS, Anker SD. Prevalence and clinical impact of cachexia in chronic illness in Europe, USA, and Japan: facts and numbers update 2016. J Cachexia Sarcopenia Muscle 2016 Dec;7(5):507-509 [FREE Full text] [doi: 10.1002/jcsm.12167] [Medline: 27891294]
28. Huang W, Huang C, Wu P, Chen C, Cheng Y, Chen H, et al. The association between airflow limitation and blood eosinophil levels with treatment outcomes in patients with chronic obstructive pulmonary disease and prolonged mechanical ventilation. Sci Rep 2019 Sep 17;9(1):13420 [FREE Full text] [doi: 10.1038/s41598-019-49918-z] [Medline: 31530874]
29. Chin K, Nakamura T, Takahashi K, Sumi K, Ogawa Y, Masuzaki H, et al. Effects of obstructive sleep apnea syndrome on serum aminotransferase levels in obese patients. Am J Med 2003 Apr 01;114(5):370-376. [doi: 10.1016/s0002-9343(02)01570-x] [Medline: 12714126]
30. de Miguel-Díez J, López-de-Andrés A, Hernandez-Barrera V, Jimenez-Trujillo I, Del Barrio JL, Puente-Maestu L, et al. Prevalence of Pain in COPD Patients and Associated Factors: Report From a Population-based Study. Clin J Pain 2018 Sep;34(9):787-794. [doi: 10.1097/AJP.0000000000000598] [Medline: 29485534]
31. Biljak VR, Pancirov D, Cepelak I, Popović-Grle S, Stjepanović G, Grubišić T. Platelet count, mean platelet volume and smoking status in stable chronic obstructive pulmonary disease. Platelets 2011;22(6):466-470. [doi: 10.3109/09537104.2011.573887] [Medline: 21506665]
32. Agustí A, Edwards LD, Rennard SI, MacNee W, Tal-Singer R, Miller BE, Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) Investigators. Persistent systemic inflammation is associated with poor clinical outcomes in COPD: a novel phenotype. PLoS One 2012;7(5):e37483 [FREE Full text] [doi: 10.1371/journal.pone.0037483] [Medline: 22624038]
33. Samukawa T, Matsumoto K, Tsukuya G, Koriyama C, Fukuyama S, Uchida A, et al. Development of a self-scored persistent airflow obstruction screening questionnaire in a general Japanese population: the Hisayama study. Int J Chron Obstruct Pulmon Dis 2017;12:1469-1481 [FREE Full text] [doi: 10.2147/COPD.S130453] [Medline: 28553099]
34. Kuhn S, Egert B, Neumann S, Steinbeck C. Building blocks for automated elucidation of metabolites: machine learning methods for NMR prediction. BMC Bioinformatics 2008 Sep 25;9:400 [FREE Full text] [doi: 10.1186/1471-2105-9-400] [Medline: 18817546]
35. Akkermans RP, Biermans M, Robberts B, ter Riet G, Jacobs A, van Weel C, et al. COPD prognosis in relation to diagnostic criteria for airflow obstruction in smokers. Eur Respir J 2014 Jan;43(1):54-63 [FREE Full text] [doi: 10.1183/09031936.00158212] [Medline: 23563262]

Abbreviations

5-CV: five times for cross-validation
AUC: area under the receiver operating characteristic curve
BD: bronchodilator
COPD: chronic obstructive pulmonary disease
EOS: eosinophil count
FEV1: forced expiratory volume in 1 second
FVC: forced vital capacity
Hb: hemoglobin
HT: hematocrit
MCHC: mean corpuscular hemoglobin concentration
MCV: mean corpuscular volume
WBC: white blood cell
XGBoost: Gradient Boosting Decision Tree machine learning method

Edited by G Eysenbach; submitted 05.10.20; peer-reviewed by C Gandhi; comments to author 28.10.20; revised version received 17.11.20; accepted 11.04.21; published 06.07.21

Please cite as:
Muro S, Ishida M, Horie Y, Takeuchi W, Nakagawa S, Ban H, Nakagawa T, Kitamura T
Machine Learning Methods for the Diagnosis of Chronic Obstructive Pulmonary Disease in Healthy Subjects: Retrospective Observational Cohort Study
JMIR Med Inform 2021;9(7):e24796
URL: https://medinform.jmir.org/2021/7/e24796
doi: 10.2196/24796
PMID: 34255684