ORIGINAL ARTICLE

Machine Learning for Early Lung Cancer Identification Using Routine Clinical and Laboratory Data

Michael K. Gould[1,2], Brian Z. Huang[2], Martin C. Tammemagi[3], Yaron Kinar[4], and Ron Shiff[4]

[1]Department of Health Systems Science, Kaiser Permanente Bernard J. Tyson School of Medicine, Pasadena, California; [2]Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, California; [3]Department of Health Sciences, Brock University, St. Catharines, Ontario, Canada; and [4]Medial EarlySign, Newton, Massachusetts

Abstract

Rationale: Most lung cancers are diagnosed at an advanced stage. Presymptomatic identification of high-risk individuals can prompt earlier intervention and improve long-term outcomes.

Objectives: To develop a model to predict a future diagnosis of lung cancer on the basis of routine clinical and laboratory data by using machine learning.

Methods: We assembled data from 6,505 case patients with non–small cell lung cancer (NSCLC) and 189,597 contemporaneous control subjects and compared the accuracy of a novel machine learning model with a modified version of the well-validated 2012 Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial risk model (mPLCOm2012), by using the area under the receiver operating characteristic curve (AUC), sensitivity, and diagnostic odds ratio (OR) as measures of model performance.

Measurements and Main Results: Among ever-smokers in the test set, a machine learning model was more accurate than the mPLCOm2012 for identifying NSCLC 9–12 months before clinical diagnosis ($P < 0.00001$) and demonstrated an AUC of 0.86, a diagnostic OR of 12.3, and a sensitivity of 40.1% at a predefined specificity of 95%. In comparison, the mPLCOm2012 demonstrated an AUC of 0.79, an OR of 7.4, and a sensitivity of 27.9% at the same specificity. The machine learning model was more accurate than standard eligibility criteria for lung cancer screening and more accurate than the mPLCOm2012 when applied to a screening-eligible population. Influential model variables included known risk factors and novel predictors such as white blood cell and platelet counts.

Conclusions: A machine learning model was more accurate for early diagnosis of NSCLC than either standard eligibility criteria for screening or the mPLCOm2012, demonstrating the potential to help prevent lung cancer deaths through early detection.

Keywords: lung cancer; non–small cell lung carcinoma; early detection of cancer; screening; machine learning

It has long been recognized that the stage of disease is strongly associated with lung cancer survival, with the localized stage of lung cancer having the most favorable prognosis (1). More recently, there has been evidence from randomized controlled trials that lung cancer screening with low-dose computed tomography (LDCT) reduces lung cancer mortality among high-risk current and former smokers (2, 3). Together, these observations support the idea that early detection provides additional opportunities to deliver curative treatment and prevent death from lung cancer.

All else being equal, those at highest risk for lung cancer death have the most to gain from earlier detection (4), although the positive association between lung cancer risk and serious comorbid conditions complicates the picture. Accordingly, a number of risk models have been developed to predict lung cancer incidence and death over varying periods of time. The earliest models were relatively simple and included terms for well-known risk factors such as age, smoking history, and asbestos exposure (5, 6). More recent models have attempted to improve lung cancer discrimination by adding a number of novel predictor variables. One

(Received in original form July 13, 2020; accepted in final form April 6, 2021)

Supported by Medial EarlySign, Inc.

Author Contributions: Conception and design: M.K.G., Y.K., and R.S. Data collection: B.Z.H. Analysis: Y.K. and R.S. Interpretation: M.K.G., B.Z.H., M.C.T., Y.K., and R.S. Manuscript preparation: M.K.G. and Y.K. Review of manuscript for important intellectual content: B.Z.H., M.C.T., and R.S.

Correspondence and requests for reprints should be addressed to Michael K. Gould, M.D., M.S., Department of Health Systems Science, Kaiser Permanente Bernard J. Tyson School of Medicine, 100 South Los Robles Avenue, Pasadena, CA 91101. E-mail: michael.k.gould@kp.org.

This article has a related editorial.

This article has an online supplement, which is accessible from this issue's table of contents at www.atsjournals.org.

Am J Respir Crit Care Med Vol 204, Iss 4, pp 445–453, Aug 15, 2021

Originally Published in Press as DOI: 10.1164/rccm.202007-2791OC on April 6, 2021
Internet address: www.atsjournals.org

At a Glance Commentary

Scientific Knowledge on the Subject: Lung cancer prediction models can help to select high-risk patients for interventions that facilitate early diagnosis. The advantages of prediction using machine learning approaches include the detection of hard-to-discern patterns, the need for fewer restrictive assumptions, and greater flexibility in handling missing data and nonlinear relationships between parameters.

What This Study Adds to the Field: A machine learning model based on routine clinical and laboratory data was more accurate for identifying lung cancer 9–12 months before clinical diagnosis than either standard eligibility criteria for lung cancer screening or a modified version of the well-known 2012 Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial model.

popular model developed to facilitate risk assessment for lung cancer screening is the 2012 Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial model (PLCOm2012) (7). In this model, independent predictors of higher 6-year probability of lung cancer diagnosis included older age, black race, lower educational attainment, a lower body mass index (BMI), chronic obstructive pulmonary disease (COPD), a prior history of cancer, a family history of lung cancer, a current (vs. former) smoking status, a greater smoking intensity and longer smoking duration, and a shorter time since quitting (among former smokers). The model had an area under the receiver operating characteristic curve (AUC) of approximately 0.80 in both the development and the validation data sets, indicating very good discrimination.

Like most existing risk models, the PLCOm2012 was developed by using a standard statistical approach, namely logistic regression. More recently, there has been a groundswell of interest in using artificial intelligence and machine learning approaches to improve models for risk prediction. Potential advantages include the detection of hard-to-discern patterns, the need for fewer restrictive assumptions, and greater flexibility in handling missing data and nonlinear relationships between parameters. Members of our research group had previously used machine learning tools to develop an accurate model that included complete blood count (CBC) values to predict cases of colorectal cancer (8) and showed in a subsequent validation study that the model was able to identify individuals with colorectal cancer 180–360 days before the clinical diagnosis was made. (9) In this paper, we used a similar machine learning approach, also leveraging routine clinical and laboratory values obtained from the electronic health record (EHR), to develop risk models to identify patients with lung cancer before clinical diagnosis.

Methods

We performed a retrospective study of case patients with non–small cell lung cancer (NSCLC) and control subjects without lung cancer to develop (train) and validate (test) a model to predict a future diagnosis of lung cancer on the basis of features (variables) obtained as part of usual clinical care, including sociodemographic characteristics, smoking history, inpatient and outpatient diagnoses, and routine laboratory test results such as CBC values.

Setting

Kaiser Permanente Southern California (KPSC) is a large, integrated healthcare system that provides comprehensive care to a diverse population of over 4.6 million members in southern California (10). At KPSC, approximately 1,000 new lung cancer cases are diagnosed and/or treated by clinicians at 15 KPSC hospitals each year. Of note, lung cancer screening with LDCT was not available until late 2013. The study was approved by the KPSC Institutional Review Board.

Study Population

We identified case patients and control subjects by using data from the KPSC Cancer Registry and Research Data Warehouse. Case patients included KPSC members with lung cancer diagnosed between 2008 and 2015 who were 45–90 years old on the date of diagnosis. Control subjects included KPSC members without a current or prior diagnosis of lung cancer who were 45–90 years old on an assigned index (pseudodiagnosis) date of July 1. On the basis of the findings of prior research, we required the presence of at least one CBC in the 12 months before the indexed date of diagnosis (for case patients) or pseudodiagnosis (for control subjects). We excluded case patients and control subjects who had not been continuously enrolled for at least 12 months before the most recent outpatient CBC preceding the index date. To assemble the final set of control subjects without lung cancer, we randomly sampled (without replacement) 5% of nonexcluded individuals on each index date.

Subsequently, we randomly sampled 60% of the case and control samples to create a training set while reserving the other 40% of the sample for testing (validation). The testing samples were used only for reporting the performance of the various models and were not used for development or fine-tuning. For this analysis, we trained the model by using a mixed sample of ever- and never-smokers but conservatively elected to present results from a test sample that was restricted to the clinically relevant target population of current and former smokers. Additional details regarding cohort assembly are provided in the online supplement.

Data Collection

From the KPSC tumor registry, we collected histologic data and the American Joint Commission on Cancer stage. Patients with nonbronchogenic histologic types were excluded, and bronchogenic carcinoma categories were collapsed into small cell lung cancer and four types of NSCLC (adenocarcinoma, squamous cell carcinoma, large cell carcinoma, and unspecified NSCLC). This analysis was limited to patients with NSCLC, which is more amenable to early diagnosis and cure than small cell lung cancer.

From the data warehouse, we collected membership information (enrollment), demographic characteristics (age, sex, race, ethnicity, marital status, and geocoded information about income and education at the level of the census block), BMI data, smoking information (including smoking status, smoking intensity, smoking duration, pack-years, and quit-years), comorbid conditions, spirometric results, laboratory test results (including CBC values), hospitalizations, and vital status data.

Spirometric results included prebronchodilator values for FEV_1% predicted and the FEV_1/FVC ratio. For

comorbid conditions, hospitalizations, and laboratory test results, we included data from all encounters that had occurred within 5 years before the index date. The burden of comorbid conditions was summarized by using the Deyo adaptation of the Charlson Comorbidity Index (11). Laboratory test results were restricted to outpatient encounters. Laboratory test results of interest included CBC, coagulation, serum electrolyte, and hepatic and kidney function test results. In total, 834 different features were included in the models (*see* Table E1 in the online supplement).

Outcomes

The primary outcome (case definition) was a diagnosis of NSCLC, with prespecified secondary analyses being used for different NSCLC subtypes, including adenocarcinoma and squamous cell carcinoma. Lung cancer cases were identified from the KPSC tumor registry as specified above. The diagnosis date was recorded in the tumor registry by trained staff as per instructions in the Surveillance, Epidemiology, and End Results Program coding manual (12).

Analysis

For each individual, we selected all dates on which a CBC sample was available before the index date. To minimize information leakage, we shifted the index date for each control subject to a randomly selected date within the year before the July 1 pseudodiagnosis date. The number of samples for case patients and control subjects was matched by calendar year in the model development process. For each sample, a feature vector was generated to represent the patient's or subject's demographic characteristics, smoking history, BMI, comorbid conditions, hospitalizations, and diagnoses. The vector also captured both the static (e.g., average, minimum, and maximum values) and dynamic (trending) properties of laboratory test and spirometric results.

Classification method. For model development, we used Extreme Gradient Boosting (XGBoost) (13), a state-of-the-art, machine learning algorithm. The XGBoost algorithm iteratively builds an ensemble of decision trees. Its main advantages over classical logistic regression–based risk models are the ability to handle missing values and to capture nonlinear relationships between the model features and the outcome, as well as having higher order interactions between features. We used the C implementation of XGBoost and tested the performance with and without data imputation. For comparison, we developed additional models by using logistic regression with elastic-net regularization on the same set of model features. The models' hyperparameters were tuned by using Bayesian optimization (for XGBoost) and grid searching (for logistic regression) with cross-validation on the training set. The hyperparameters of the models are provided in the online supplement.

Data imputation. Missing values for smoking duration and time since quitting were imputed using linear regression. For smoking intensity, we imputed missing values by taking the median value after stratifying by current or former smoking status. For imputation of other model variables, we took the median value after stratifying by age and sex.

Comparison models. We compared the accuracy of the models to a modified version of the well-validated PLCOm2012 (mPLCOm2012) (7). The mPLCOm2012 included all of the original variables except for family history, which was documented infrequently and difficult to extract from the EHR. To optimize the performance of the mPLCOm2012 in our study population, we reparameterized the model by using the training set data from case patients with NSCLC and control subjects. In an approach similar to that of the PLCOm2012, we modeled smoking intensity (cigarettes per day) as a nonlinear function by using the generalized additive model framework (14).

Measuring model performance. We used the AUC as an overall measure of discrimination. In addition, we calculated sensitivity and the diagnostic odds ratio (OR) at prespecified values of specificity. We reported the diagnostic OR instead of positive or negative predictive values, which depend on disease prevalence. We used bootstrapping to estimate confidence intervals (CIs) and compared AUCs by using the nonparametric Delong method (15).

Because the model uses longitudinal data over variable periods of time, discrimination depends strongly on the time of scoring relative to the cancer diagnosis date. Therefore, we measured the model performance in different time windows relative to outcome date (e.g., 90–180 d or 270–365 d before the date of diagnosis). The most interesting and clinically useful time windows are farther out from the outcome date to allow for possible stage shifts and improved survival, but these windows are close enough so that there will be visible evidence of cancer on a LDCT scan.

Subgroup analyses. In the primary analysis, we evaluated the model's performance among ever-smokers of 55–80 years of age. In subgroup analyses, we stratified the analysis by lung cancer stage and histology and restricted the analysis to patients who had complete smoking information and met the 2013 U.S. Preventive Services Task Force (USPSTF) eligibility criteria for lung cancer screening.

Calibration. To facilitate use, we recalibrated the score by using isotonic regression (16) on cross-validation samples and by evaluating the probability of lung cancer being diagnosed within 1 year in each bin. The number of control samples was adjusted to account for the 5% subsampling used for cohort assembly. We then tested the predicted probabilities on the validation set and used the Spiegelhalter z test to evaluate the calibration of the model.

Feature importance. To determine which features contributed most to model predictions, we used Shapley additive explanation (SHAP) values (17) on the XGBoost predictions. Each SHAP value measures how much each feature contributes, either positively or negatively, to a single lung cancer risk level assigned by the model. We used an open implementation of the SHAP values approach for both calculation and visualization (17).

Results

Among 9,498 members with a first diagnosis of lung cancer between 2008 and 2015, we excluded 67 who were not between 45 and 90 years old on the date of diagnosis, 1,615 who did not have a CBC within 12 months of diagnosis, and 577 who had not had 12 months of continuous membership before the date of the qualifying CBC, leaving a total of 6,505 case patients with non–small cell carcinoma who were included in this analysis and 708 case patients with small cell carcinoma who were excluded from this analysis (Figure 1).

Similarly, among 7,778,974 potential control subjects without a prior diagnosis of lung cancer, we excluded patients who were not between 45 and 90 years of age in any given year between 2008 and 2015, had not had a CBC by the index date in any given year, or had not had 12 months of continuous membership before the qualifying CBC or 24 months of

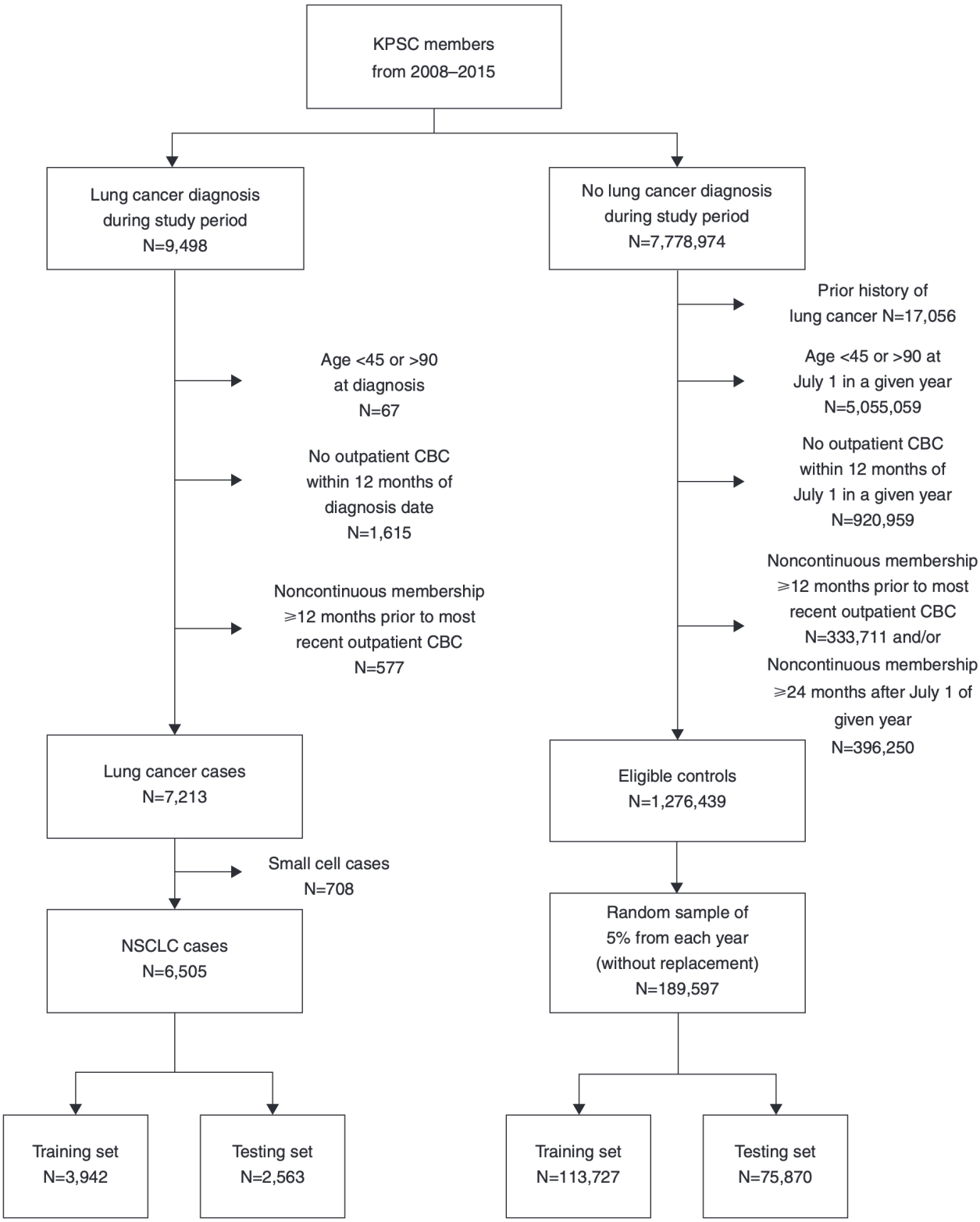

Figure 1. Assembly of case patients and control subjects for the training and test sets. CBC = complete blood count; KPSC = Kaiser Permanente Southern California; NSCLC = non–small cell lung cancer.

continuous membership after the index date in any given year. We randomly sampled 5% of potential control subjects from each year, resulting in a total of 189,597 control subjects. Data from case patients and control subjects were subsequently divided into a training set that included 3,942 case patients and 113,727 control subjects and a test set that included 2,563 case patients and 75,870 control subjects (Figure 1).

The characteristics of the training and test samples appeared to be similar (Table E2). Overall, case patients and control subjects appeared to be similar in terms of sex, marital status, education, and income, but case patients were older, were more likely to be white, and were more likely to be current smokers (Table 1). Among ever-smoking case patients and control subjects with available information, case patients appeared to have a greater number of pack-years, packs per day, and years of smoking than control subjects and appeared to have fewer years since quitting (among former smokers). There was less missing information about smoking behavior for case patients than for control subjects.

Table 1. Sociodemographic Characteristics of Case Patients and Control Subjects

	Case Patients (*Total of 6,505*)	Control Subjects (*Total of 189,597*)
Age, yr, *n* (%)		
45.0–54.0	360 (5.5)	55,886 (29.5)
55.0–64.0	1,185 (18.2)	57,309 (30.2)
65.0–74.0	2,371 (36.4)	46,115 (24.3)
75.0–84.0	2,147 (33.0)	24,898 (13.1)
85.0–90.0	442 (6.8)	5,389 (2.8)
Sex, M, *n* (%)	3,274 (50.3)	80,458 (42.4)
Race or ethnicity, *n* (%)		
White	4,113 (63.2)	92,086 (48.6)
Black	865 (13.3)	20,789 (11.0)
Hispanic	796 (12.2)	51,180 (27.0)
Asian	632 (9.7)	20,944 (11.0)
Other	99 (1.5)	4,598 (2.4)
Marital status, *n* (%)		
Married or living as married	3,763 (57.8)	113,518 (59.9)
Widowed	1,226 (18.8)	18,229 (9.6)
Divorced	744 (11.4)	15,359 (8.1)
Never married	533 (8.2)	22,721 (12.0)
Separated	59 (0.9)	1,931 (1.0)
Other	1 (0)	100 (0.1)
Unknown	179 (2.8)	17,739 (9.4)
Education, *n* (%)		
Less than high school	457 (7.0)	16,181 (8.5)
Some high school	41 (0.6)	1,260 (0.7)
High school graduate	2,079 (32.0)	57,463 (30.3)
Some college, no degree	2,017 (31.0)	56,327 (29.7)
Associate's degree	1 (0)	61 (0)
Bachelor's degree	1,669 (25.7)	50,175 (26.5)
Graduate degree	155 (2.4)	4,969 (2.6)
Unknown	86 (1.3)	3,161 (1.7)
Smoking status, *n* (%)		
Former	3,896 (59.9)	58,838 (31.0)
Current	1,353 (20.8)	12,825 (6.8)
Never	1,194 (18.4)	115,511 (60.9)
Passive	28 (0.4)	1,031 (0.5)
Unknown	34 (0.5)	1,392 (0.7)
Smoking behavior		
Ever-smokers, *n* (%)	5,249 (80.6)	71,663 (37.8)
Pack-years, mean (SD)	39.4 (35.5)	21.5 (23.9)
Packs/d, mean (SD)	1.1 (0.8)	0.9 (0.7)
Years of smoking, mean (SD)	35.5 (14.6)	22 (14.3)
Quit-years (among former smokers), mean (SD)	18.4 (14.3)	23.4 (15.2)
Missing smoking information		
Pack-years, *n* (%)	1,917 (36.5)	39,736 (55.4)
Packs/d, *n* (%)	1,666 (31.7)	36,956 (51.6)
Years of smoking, *n* (%)	1,649 (31.4)	34,491 (48.1)
Quit-years, *n* (%)	602 (15.5)	12,599 (21.4)
Household income, U.S. $, mean (SD)	65,259 (28,382)	67,388 (29,102)

Model Accuracy

We calculated the AUC, sensitivity, and diagnostic OR of the various models among ever-smokers in the test set in different time windows before clinical diagnosis (Table 2, Figure 2). The best performing model was the XGBoost model without imputation, hereafter referred to as the Medial EarlySign (MES) model. In the time window of 9–12 months before diagnosis, the AUC of the MES model was 0.856 (95% CI, 0.841 – 0.871). At a specificity of 95%, the MES model sensitivity was 40.1% (95% CI, 35.6 – 44.0%) and the diagnostic OR was 12.7 (95% CI, 10.5–14.9). In comparison, the AUC of the mPLCOm2012 was 0.791 (95% CI, 0.771 – 0.810: *P* value for comparison <0.00001). This value was similar to the reported AUC of 0.80 (95% CI, 0.78 – 0.81) in the PLCOm2012 validation set. The sensitivity (27.9%; 95% CI, 24.1 – 32.1%) and diagnostic OR (7.4; 95% CI, 6.0 – 9.0) of the mPLCOm2012 were also lower than those of the MES model. A logistic regression model with the same features as the nonlinear MES model had an intermediate AUC of 0.840 (95% CI, 0.824 – 0.858); at a specificity of 95%, the sensitivity of this model was 39.4% (95% CI, 35.4 – 44.7%), and the diagnostic OR was 12.3 (95% CI, 10.4 – 15.3). Both the AUC and the sensitivity of the MES model appeared to be especially advantageous over the mPLCOm2012 when using data from more proximate time windows (Figure E1).

Model Accuracy in the USPSTF Screening–Eligible Population

Within the USPSTF screening–eligible population of case patients and control subjects with complete smoking information, the MES model had an AUC of 0.807 (95% CI, 0.767–0.845) in the 9- to 12-month time window, whereas the mPLCOm2012 had an AUC of 0.712 (95% CI, 0.662 – 0.752) (Figure 2).

Accuracy by Lung Cancer Stage and Histology

We evaluated model accuracy for different stages and histologic types of NSCLC (Table 3). For both the MES model and the mPLCOm2012, the AUC and sensitivity were higher for identifying squamous cell

Table 2. Accuracy of XGBoost 1, XGBoost 2, mPLCOm2012, and LR Models in the Test Set in Different Populations and Time Windows

Population by Time Window by Model	AUC	5% FPR		10% FPR	
		Sensitivity (%)	Odds Ratio	Sensitivity (%)	Odds Ratio
Ever-smokers					
90–180 d					
XGBoost 1	**0.870 (0.856–0.886)**	**48.9 (45.1–53.7)**	**18.2 (15.6–22.0)**	**64.0 (60.9–68.3)**	**16.0 (14.0–19.4)**
XGBoost 2	0.867 (0.850–0.883)	47.3 (43.2–51.7)	17.1 (14.5–20.4)	63.4 (59.0–66.4)	15.6 (12.9–17.8)
mPLCOm2012	0.800 (0.782–0.820)	32.2 (28.4–36.1)	9.0 (7.5–10.7)	47.4 (43.3–51.8)	8.1 (6.8–9.6)
LR	0.861 (0.846–0.878)	46.0 (43.0–52.0)	16.2 (14.3–20.6)	60.7 (56.9–64.8)	13.9 (11.9–16.6)
180–270 d					
XGBoost 1	**0.862 (0.845–0.878)**	42.1 (38.3–47.8)	13.8 (11.8–17.4)	**59.9 (53.7–63.7)**	**13.4 (10.4–15.8)**
XGBoost 2	0.861 (0.846–0.877)	44.3 (40.4–48.6)	15.1 (12.9–17.9)	58.7 (54.3–63.3)	12.8 (10.7–15.5)
mPLCOm2012	0.796 (0.777–0.814)	30.1 (25.7–33.4)	8.1 (6.5–9.5)	47.4 (42.5–52.0)	8.1 (6.6–9.7)
LR	0.853 (0.838–0.869)	**44.7 (39.8–49.2)**	**15.3 (12.5–18.4)**	57.0 (52.5–61.6)	11.9 (9.9–14.4)
270–365 d					
XGBoost 1	**0.856 (0.840–0.872)**	**40.3 (35.4–43.7)**	**12.8 (10.4–14.7)**	**58.4 (53.8–61.8)**	**12.6 (10.5–14.6)**
XGBoost 2	0.854 (0.840–0.869)	38.8 (34.6–43.8)	12.0 (10.1–14.8)	54.0 (50.2–59.5)	10.6 (9.1–13.2)
mPLCOm2012	0.791 (0.771–0.810)	27.9 (24.1–32.1)	7.3 (6.0–8.9)	41.0 (37.0–46.5)	6.2 (5.3–7.8)
LR	0.840 (0.824–0.858)	39.4 (35.4–44.7)	12.3 (10.4–15.3)	53.7 (49.8–58.8)	10.4 (8.9–12.9)
USPSTF screening–eligible					
90–180 d					
XGBoost 1	0.850 (0.829–0.879)	**40.0 (34.3–52.7)**	**12.7 (9.9–21.1)**	**56.2 (48.9–63.8)**	**11.6 (8.6–15.8)**
XGBoost 2	**0.852 (0.822–0.874)**	39.5 (32.6–47.7)	12.4 (9.2–17.4)	55.1 (46.5–62.7)	11.1 (7.8–15.1)
mPLCOm2012	0.745 (0.709–0.780)	21.4 (14.7–27.3)	5.1 (3.3–7.0)	30.1 (22.2–37.6)	3.8 (2.6–5.4)
LR	0.828 (0.795–0.856)	37.3 (29.0–45.2)	11.3 (7.8–15.7)	50.3 (42.0–56.6)	9.1 (6.5–11.7)
180–270 d					
XGBoost 1	**0.819 (0.787–0.848)**	**31.0 (22.2–38.4)**	**8.5 (5.4–11.8)**	**44.9 (37.0–56.3)**	**7.3 (5.3–11.6)**
XGBoost 2	0.815 (0.787–0.849)	29.1 (21.3–38.0)	7.8 (5.1–11.6)	44.2 (37.6–57.4)	7.1 (5.4–12.1)
mPLCOm2012	0.735 (0.694–0.773)	16.8 (10.6–23.3)	3.8 (2.3–5.7)	25.5 (17.8–32.1)	3.1 (1.9–4.2)
LR	0.800 (0.762–0.832)	29.7 (22.2–38.6)	8.0 (5.4–11.9)	36.7 (27.8–45.9)	5.2 (3.5–7.6)
270–365 d					
XGBoost 1	**0.807 (0.767–0.845)**	**37.8 (25.2–45.2)**	**11.5 (6.4–15.6)**	**49.0 (38.9–58.0)**	**8.6 (5.7–12.4)**
XGBoost 2	0.798 (0.765–0.833)	32.2 (21.3–38.2)	9.0 (5.1–11.7)	43.4 (35.4–52.2)	6.9 (4.9–9.8)
mPLCOm2012	0.712 (0.662–0.752)	21.9 (12.4–28.1)	5.3 (2.7–7.3)	27.7 (19.7–35.9)	3.4 (2.2–5.0)
LR	0.802 (0.767–0.837)	**37.8 (30.1–47.0)**	**11.5 (8.2–16.9)**	46.9 (40.0–55.4)	7.9 (6.0–11.2)

Definition of abbreviations: AUC = area under the receiver operating characteristic curve; FPR = false-positive rate; LR = logistic regression; mPLCOm2012 = modified version of the well-validated 2012 Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial model; USPSTF = U.S. Preventive Services Task Force; XGBoost 1 = XGBoost without imputation; XGBoost 2 = XGBoost with imputation; XGBoost = Extreme Gradient Boosting.

$P < 0.001$ for AUC comparisons between XGBoost models and mPLCOm2012 in both populations and in all time windows. The best value(s) within each group are indicated with bold typeface.

carcinoma than for identifying adenocarcinoma ($P < 0.0001$). In addition, for both models, the AUC and sensitivity were higher for identifying patients with stage 0–II NSCLC than for identifying patients with stage III–IV NSCLC ($P < 0.05$).

Comparison of the MES Model with USPSTF Eligibility Criteria for Lung Cancer Screening

In the population of ever-smokers, the specificity of the USPSTF eligibility criteria for identifying lung cancer was 91.4% (95% CI, 91.1 – 91.8%), and the sensitivity was 26.6% (95% CI, 24.2–29.3%). At the same value of specificity, the MES model sensitivity was 53.0% (95% CI, 48.7 – 57.3%), and the mPLCOm2012 sensitivity was 38.6% (95% CI, 34.2 – 42.7%).

Calibration

The MES model appeared to be well-calibrated on 55- to 80-year-old ever-smokers in the validation sample, although it seemed to underestimate the probability of cancer for patients in the highest risk decile (Figure E2), and the Spiegelhalter z test P value did not indicate statistical significance ($P = 0.075$).

Feature Importance

By using SHAP analysis (Figure 3), we found that the 10 most informative features in the model were smoking duration, age, pack-years, BMI, white blood cell (WBC) count, history of COPD, time since quitting smoking, Hispanic ethnicity, trends over time in values for HDL (high-density lipoprotein), and red cell distribution width values. Intuitively, a longer smoking duration, older age, higher number of pack-years, higher WBC count, and history of COPD all contributed to a greater risk of cancer, whereas a higher BMI, longer time since quitting, and Hispanic ethnicity contributed to a lower risk.

Discussion

In this large population-based sample of case patients with NSCLC and control subjects, we used machine learning to develop a set of models with good accuracy for identifying lung cancer as early as 9–12 months before clinical diagnosis. Indeed, the MES models were more accurate than the mPLCOm2012 at

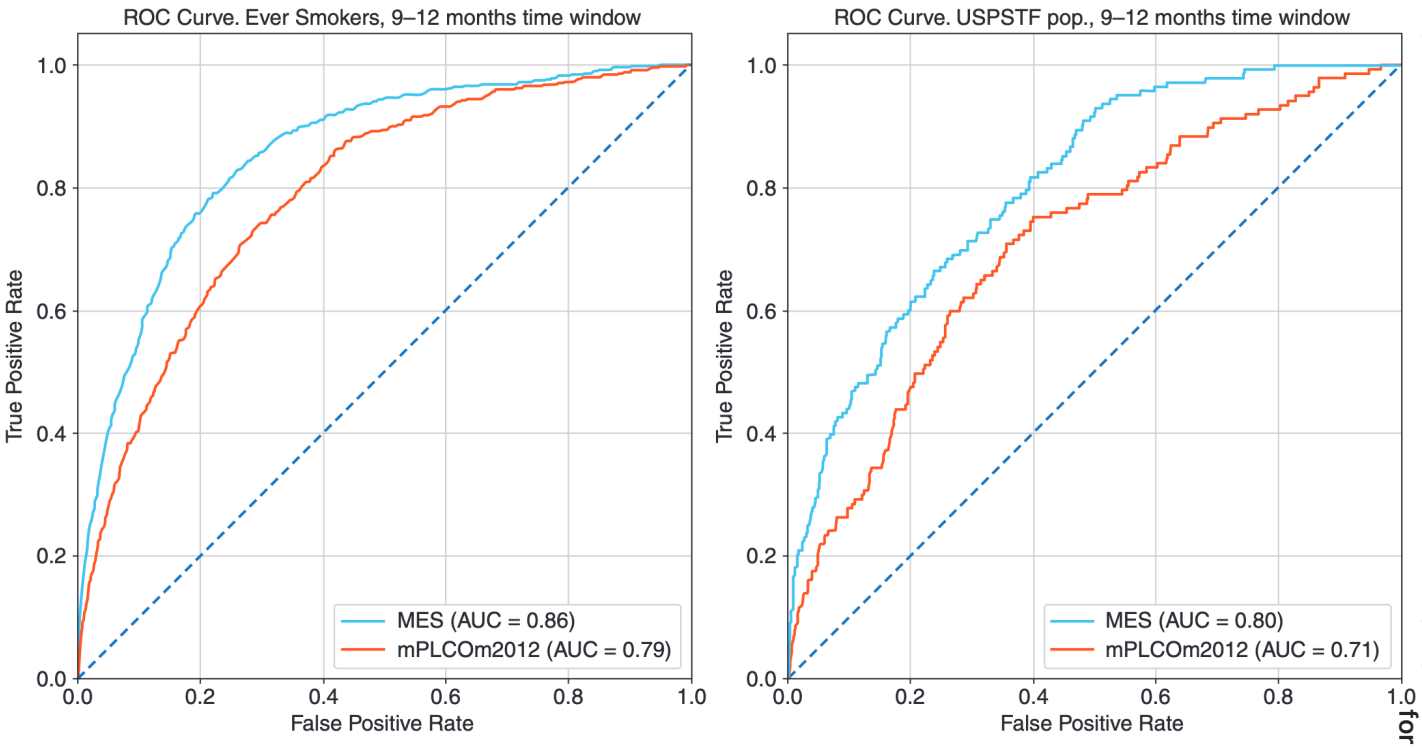

Figure 2. ROC curves for MES model and mPLCOm2012 9–12 months before clinical diagnosis. (Left panel) Ever-smoker population. (Right panel) USPSTF screening–eligible population. Dashed line indicates line of identity. AUC = area under the ROC curve; MES = Medial EarlySign; mPLCOm2012 = modified version of the 2012 Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial model; pop. = population; ROC = receiver operating characteristic; USPSTF = U.S. Preventive Services Task Force.

Table 3. Model Performance in the Test Set for Different Cancer Stages and Histologic Subtypes for Ever-Smokers Ages 55–80 Years within the 9- to 12-Month Time Window

Type by Stage by Model	AUC	Sensitivity (%) at 95% Specificity	Diagnostic Odds Ratio
NSCLC			
0–II			
XGBoost 1	0.880 (0.858–0.901)	43.2 (35.5–49.4)	14.4 (10.5–18.6)
mPLCOm2012	0.817 (0.785–0.847)	33.5 (26.3–41.4)	9.6 (6.8–13.4)
III–IV			
XGBoost 1	0.846 (0.828–0.863)	38.4 (33.8–44.6)	11.8 (9.7–15.3)
mPLCOm2012	0.782 (0.758–0.803)	24.7 (20.4–29.8)	6.2 (4.9–8.1)
Adenocarcinoma			
0–IV			
XGBoost 1	0.835 (0.814–0.858)	33.6 (27.5–39.7)	9.6 (7.2–12.5)
mPLCOm2012	0.756 (0.728–0.782)	21.1 (15.5–25.7)	5.1 (3.5–6.6)
Squamous cell carcinoma			
0–IV			
XGBoost 1	0.888 (0.865–0.914)	47.3 (39.3–55.9)	17.1 (12.3–24.1)
mPLCOm2012	0.845 (0.819–0.872)	36.6 (30.1–44.7)	11.0 (8.2–15.4)

Definition of abbreviations: AUC = area under the receiver operating characteristic curve; mPLCOm2012 = modified version of the 2012 Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial model; NSCLC = non–small cell lung cancer; XGBoost 1 = XGBoost without imputation; XGBoost = Extreme Gradient Boosting.
$P < 0.0001$ for all AUC comparisons between the XGBoost 1 model and mPLCOm2012 in all stages and subtypes. For both models, $P < 0.05$ for the comparison between early stages (0–II) and late stages (III–IV), and $P < 0.0001$ for comparison of adenocarcinoma and squamous cell carcinoma.

all time periods. Thus, the MES model is potentially useful as a tool for identifying individuals for early lung cancer detection with LDCT.

During this time window, we believe that case patients were unlikely to have had symptoms or overt signs of lung cancer. In a meta-analysis, the median time from onset of symptoms to lung cancer diagnosis was reported to be 41–143 days (18), corresponding to the apparent 5-month period between symptom documentation and diagnosis in our data set (Figure E3). Indeed, the 9- to 12-month time window may represent a "sweet spot" at which patients are likely to have an identifiable nodule on LDCT scans.

We envision at least two possible approaches for implementation in clinical practice. First, individual providers and patients could use the model to estimate risk and inform personalized decision-making about lung cancer screening in the clinic. Second, health systems could run the model at the population level to identify high-risk patients for subsequent outreach.

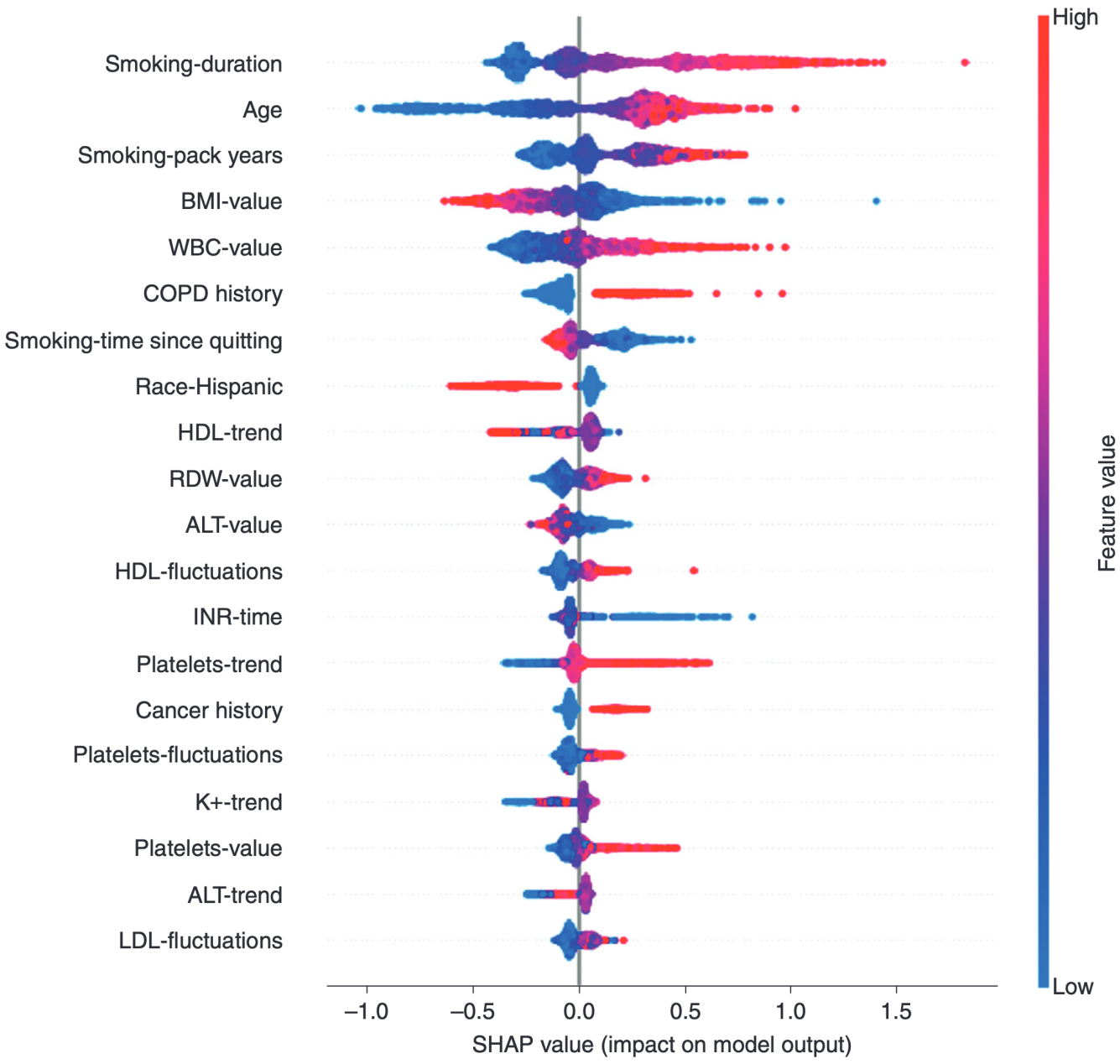

Figure 3. Shapley additive explanation (SHAP) summary plot of 20 feature clusters, derived by aggregating related values of a particular feature (e.g., the average, minimum, and maximum). Each dot corresponds to the SHAP value of the feature cluster for the lung cancer risk score of a given case patient or control subject at a certain point in time. A feature's SHAP value (*x*-axis) represents the contribution of the specific feature to the risk score, with positive values indicating a contribution that increases the risk score and negative values indicating a contribution that lowers the score. The location of the dot on the *x*-axis represents its SHAP value, whereas its color represents the cluster's value (the actual value of the feature that is represented in the cluster), with red representing higher values (for features measured along a continuum) or affirmative responses (for binary features). The dots are piled up vertically to show their density. The feature clusters are sorted by their mean absolute SHAP values. ALT = alanine aminotransferase; BMI = body mass index; COPD = chronic obstructive pulmonary disease; HDL = high-density lipoprotein; INR = international normalized ratio; LDL = low-density lipoprotein; RDW = red cell distribution width; WBC = white blood cell.

Compared with USPSTF criteria, which had a specificity of 91.4% in our sample, the main MES model (XGBoost without data imputation) had better sensitivity (53.0% vs. 26.6%) at this same cut point for specificity. In addition, within the USPSTF screening–eligible subpopulation, the MES model was more accurate than the mPLCOm2012 for identifying lung cancer 9–12 months before clinical diagnosis (with an AUC of 0.81 vs. 0.71). Thus, another potential use of the MES model is to identify a pool of higher-risk candidates within the USPSTF screening–eligible population for early detection.

Notably, the MES model was somewhat more accurate for identifying patients with

stage 0–II lung cancer than patients with more advanced disease, and it was substantially more accurate for identifying squamous cell carcinoma than for identifying adenocarcinoma, as was similarly true for the mPLCOm2012. This may reflect stronger associations between squamous cell carcinoma and influential variables in both models, such as age, pack-years, quit-years, and COPD. The models have several other variables in common, including BMI, race, current smoking status, and a prior history of cancer. Several of these variables also appear in other "legacy" models of lung cancer risk, especially older age and higher amounts of exposure to tobacco smoke (5, 6). The higher accuracy of the MES model likely reflects the additional contribution of other unique variables, primarily laboratory values, including the WBC count, platelet count, and red cell distribution width, highlighting the untapped value of routine laboratory test results as potential biomarkers of lung cancer risk.

Our study has several limitations. Although we demonstrated that the MES model had good accuracy for identifying undiagnosed lung cancer up to 9–12 months before the clinical diagnosis, the model was most accurate 0–3 months before the diagnosis, a time period in which laboratory testing may be prompted by lung cancer suspicion. Missing or inaccurate information about smoking behavior also presents a challenge. More complete information about pack-years and quit-years would likely result in better model performance for both the MES model and the mPLCOm2012 that we used for comparison.

The "black box" of machine learning is another limitation. To maximize transparency, we reported SHAP values and provided a list of features that were highly influential. Furthermore, we showed that a logistic regression model using the same set of 834 features was only slightly less accurate than the machine learning MES model, indicating that the improved discrimination was probably due to the large number of model features. Indeed, further validation on additional populations is required to understand which of the models generalize best, in particular as model accuracy might depend on the missingness pattern of the data. With appropriate computing resources, it would be feasible to implement the MES model in many healthcare systems by using routinely available data from EHRs and the pulmonary function laboratory.

In summary, we used machine learning to develop a novel lung cancer risk model based on routine clinical information and laboratory test results. The resulting MES model was more accurate for identifying undiagnosed lung cancer than either standard eligibility criteria for lung cancer screening or the mPLCOm2012. On the basis of clinical characteristics and laboratory testing performed 9–12 months before a clinical diagnosis of cancer, the MES model was able to identify lung cancer with a sensitivity and specificity of 40.3% and 95%, respectively, with a positive test result indicating a 13-fold elevation in the odds of lung cancer. With further validation and refinement, the MES model has the potential to help prevent lung cancer deaths through enabling earlier diagnosis. ■

Author disclosures are available with the text of this article at www.atsjournals.org.

References

1. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, *et al.* International Association for the Study of Lung Cancer Staging and Prognostic Factors Committee, Advisory Boards, and Participating Institutions. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. *J Thorac Oncol* 2016;11:39–51.
2. De Koning H, Van Der Aalst C, Ten Haaf K, Oudkerk M. PL02.05 effects of volume CT lung cancer screening: mortality results of the NELSON randomised-controlled population based trial [abstract]. *J Thorac Oncol* 2018;13:S185.
3. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, *et al.* National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011;365:395–409.
4. Bach PB, Gould MK. When the average applies to no one: personalized decision making about potential benefits of lung cancer screening. *Ann Intern Med* 2012;157:571–573.
5. Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, *et al.* Variations in lung cancer risk among smokers. *J Natl Cancer Inst* 2003;95: 470–478.
6. Spitz MR, Etzel CJ, Dong Q, Amos CI, Wei Q, Wu X, *et al.* An expanded risk prediction model for lung cancer. *Cancer Prev Res (Phila)* 2008;1: 250–254.
7. Tammemägi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, *et al.* Selection criteria for lung-cancer screening. *N Engl J Med* 2013; 368:728–736.
8. Kinar Y, Kalkstein N, Akiva P, Levin B, Half EE, Goldshtein I, *et al.* Development and validation of a predictive model for detection of colorectal cancer in primary care by analysis of complete blood counts: a binational retrospective study. *J Am Med Inform Assoc* 2016;23: 879–890.
9. Hornbrook MC, Goshen R, Choman E, O'Keeffe-Rosetti M, Kinar Y, Liles EG, *et al.* Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data. *Dig Dis Sci* 2017;62: 2719–2727.
10. Koebnick C, Langer-Gould AM, Gould MK, Chao CR, Iyer RL, Smith N, *et al.* Sociodemographic characteristics of members of a large, integrated health care system: comparison with US Census Bureau data. *Perm J* 2012;16:37–41.
11. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J Clin Epidemiol* 1992;45: 613–619.
12. Adamo M, Dickie L, Ruhl J. SEER program coding and staging manual 2018. Bethesda, MD: National Cancer Institute; 2018.
13. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco, CA: Association for Computing Machinery; 2016:785–794.
14. Hastie TJ, Tibshirani RJ. Generalized additive models. Murray Hill, NJ: AT&T Lab; 1990.
15. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–845.
16. Robertson T, Wright F, Dykstra R. Order restricted statistical inference. New York, NY: Wiley; 1988.
17. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018;2:749–760.
18. Jacobsen MM, Silverstein SC, Quinn M, Waterston LB, Thomas CA, Benneyan JC, *et al.* Timeliness of access to lung cancer diagnosis and treatment: a scoping literature review. *Lung Cancer* 2017;112:156–164.