CANCER GENOMICS & PROTEOMICS *16*: 229-244 (2019)
doi:10.21873/cgp.20128

Diagnosis of Non-small Cell Lung Cancer for Early Stage Asymptomatic Patients

CHERYLLE GOEBEL[1], CHRISTOPHER L. LOUDEN[2], ROBERT MCKENNA Jr.[3], OSITA ONUGHA[3], ANDREW WACHTEL[4] and THOMAS LONG[5]

[1]*Goebel Consulting Inc, Research and Development, Mountain View, CA, U.S.A.;*
[2]*Louden Consulting, Statistics, San Antonio, TX, U.S.A.;*
[3]*Providence Saint John's Health Center/John Wayne Cancer Institute, Thoracic Surgery, Santa Monica, CA, U.S.A.;*
[4]*Cedar-Sinai Medical Center, Pulmonary Medicine, Los Angeles, CA, U.S.A.;*
[5]*Lung Cancer Proteomics LLC, Executive Board, Hebron, IN, U.S.A.*

Abstract. *Background/Aim: In 2016 in the United States, 7 of 10 patients were estimated to die following lung cancer diagnosis. This is due to a lack of a reliable screening method that detects early-stage lung cancer. Our aim is to accurately detect early stage lung cancer using algorithms and protein biomarkers. Patients and Methods: A total of 1,479 human plasma samples were processed using a multiplex immunoassay platform. 82 biomarkers and 6 algorithms were explored. There were 351 NSCLC samples (90.3% Stage I, 2.3% Stage II, and 7.4% Stage III/IV). Results: We identified 33 protein biomarkers and developed a classifier using Random Forest. Our test detected early-stage Non-Small Cell Lung Cancer (NSCLC) with a 90% accuracy, 80% sensitivity, and 95% specificity in the validation set using the 33 markers. Conclusion: A specific, non-invasive, early-detection test, in combination with low-dose computed tomography, could increase survival rates and reduce false positives from screenings.*

According to the American Cancer Society, on a global scale, lung cancer is the leading cause of cancer-related incidence and death at 2.09 million cases and 1.76 million deaths in 2018 (1). By 2016, in the US, an estimated 538,243 living individuals were diagnosed with lung and bronchus cancer (2). An additional 228,150 new cases with an estimated 142,670 deaths are expected in 2019 (3).

Correspondence to: Cherylle Goebel, Goebel Consulting Inc, Research and Development, 780 Montague Expressway, Suite 703, San Jose, CA 95131, U.S.A. Tel: +1 4084936627, e-mail: cherylleg@lcproteomics.com

Key Words: Early stage lung cancer, biomarkers, proteomics, immunoassay, detection, diagnosis, non-small cell lung cancer.

Lung cancer originates in the lungs, but can metastasize to other organs in the body. It is classified based on the histological profile of the tumor cell and predominantly falls into two major categories: i) small cell lung cancer (SCLC, 13%) and ii) non-small cell lung cancer (NSCLC, 84%) (3). Detection at Stage I or II for NSCLC can offer good prognosis.

Symptoms and detection. Current methods of detecting lung cancer include a chest x-ray (CXR), computed tomography (CT) scan, magnetic resonance imaging (MRI), positron emission tomography (PET) scan, sputum analysis, and lung biopsy. Despite the advancement in technology and the extensive cancer research, 57% of lung cancer patients are diagnosed only after a tumor has metastasized to a different location. Under these circumstances, there is little chance of a cure, and the 5-year survival rate is less than 6% (2). Late diagnosis of lung cancer can be attributed to: i) primarily the lack of symptoms at early-stage lung cancer (4) ii) misdiagnosis of the disease since early symptoms (persistent cough, shortness of breath, chest pain, wheezing, and hemoptysis) are often misinterpreted (5) iii) the lack of proven benefit for lung cancer screening until recently (6), and iv) cost effect and limited access to state-of-the-art detection methods in indigent populations (7).

It is evident that the sooner lung cancer is diagnosed, the better the prognosis for the patient. However, only 16% of patients were diagnosed when the disease was still restricted to the lungs and even for these, only 57.4% survived 5 years (8). Based on the 2016 Cancer Statistics Review by SEER (8), the 5-year survival rate decreased to 30.8% and 5.2% for patients with regional and distant stage lung cancer, respectively. Only 19.4% of all diagnosed lung cancer patients from 2009-2015 survived 5 years.

NLST clinical trial. In 2010, a US clinical trial, sponsored by the National Lung Screening Trial (NLST), showed a 20%

decrease in lung cancer deaths for high-risk patients who were screened with low-dose spiral computed tomography (LDCT scans) compared to standard chest x-ray (CXR). The NLST patient population consisted of 53,454 current or former smokers with a ≥30 pack-year smoking history between 55-74 years of age and no history or symptoms of lung cancer. The results of the NLST study (9, 10) were:
i) patient adherence rate to screening was greater than 90%,
ii) the rate of positive screen tests was 24.2% with LDCT and 6.9% with radiography,
iii) false positive rates (lung mass) were 96.4% for the LDCT and 94.5% for the CXR screened group,
iv) the mortality rate from lung cancer was 20% lower with LDCT compared to the CXR screened group
v) the rate of death from any cause was 6.7% lower with LDCT compared to the CXR screened group.

We had a few caveats concerning the NLST study. First, due to the substantial pack history of the NLST participants, it was not known whether these findings were relevant to individuals who were non-smokers or those who smoked less. Second, screening using LDCT and radiography showed a high rate of false positives which can lead to unnecessary lung biopsy, surgery, other follow-up diagnostic tests, and stress. Additional factors associated with LDCT and radiography screening that warrants further exploration include: i) the economic implications of over diagnosis, ii) the radiation exposure from multiple scans may lead to development of other cancers, and iii) the access to high-quality screening in certain settings.

Based on the NLST study, there is clearly a need for an accurate, early-stage, non-invasive test to further stratify risk patients with small, indeterminate nodules primarily identified using CT scanning. Our study aimed to identify potential biomarkers and an algorithm to predict early stage NSCLC with high accuracy using human plasma samples for use as an adjunctive test for lung cancer screening.

Patients and Methods

The patients from our selected population for the Lung Cancer Detector Test-1 (LCDT1) were 40+ years old, long-term smokers, and had been diagnosed with indeterminate nodules in the lungs.

Study population. This study consisted of 1,479 subjects (2,958 samples) distributed in the following cohorts: i) asthma sufferers, ii) non-smokers, iii) smokers, iv) NSCLC, and v) other cancers. Asthma (also includes other respiratory diseases, *e.g.* chronic obstructive pulmonary disease (COPD)) was included in the study population as symptoms share similarities with early signs of lung cancer. Other cancers (*e.g.* breast, prostate, ovarian, pancreatic, and colorectal cancer) were included to evaluate specificity to lung cancer. Table I summarizes the main criteria used to select samples for each cohort. Samples originated from Bulgaria, Romania, Russia, Ukraine and the United States and consisted of a mixture of Caucasian, African-American, and Hispanic populations.

Table I. *Baseline criteria for selecting NSCLC and other samples.*

Samples	Gender	Age	Cancer stage	Smoking status
NSCLC	M/F	NA	IA/B, IIA/B	Non-smoker, smoker
Non-smoker	M/F	≥25 y/o	NA	Non-smoker
Smoker	M/F	≥40 y/o	NA	Smoker
Asthma	M/F	NA	NA	Non-smoker, smoker
Other cancers	M/F	NA	All stages	Non-smoker, smoker

All sample groups consisted of a mixture of Caucasian, African-American and Hispanic. The NSCLC group mainly consisted of patients diagnosed with Stage I and II NSCLC regardless of age and smoking status. 13 out of 160 NSCLC samples were stage III and IV. The Non-Smoker group consisted of healthy individuals that were 25 y/o or older, non-smokers with no diagnosis of any cancer. The Smoker group included individuals 40 y/o, smoked 1 pack per day for 10 years, and no diagnosis of any cancer. The Asthma group included individuals diagnosed with asthma (and with other respiratory diseases, *e.g.* COPD) but no lung cancer, regardless of age and smoking status. Other cancers included patients diagnosed with breast, ovarian, prostate, pancreatic, or colon-rectal cancer.

Training set. The training set consisted of 554 Subjects (1,108 samples) ran in duplicates to evaluate 82 biomarkers and 6 multivariate analysis methods and to train the selected algorithm. There were: i) 160 NSCLC which served as a positive control, ii) 140 healthy non-smokers which served as our negative control, iii) 33 asthma sufferers, iv) 131 high-risk smokers, and v) 90 other cancers. Protein biomarker concentrations, gender, and race were included in the analysis. Subjects ranged from 25-94 years old and were distributed between female and male cohorts as shown in Table II. Patient samples consisted predominantly of Caucasians and some African-Americans and Hispanics. Most NSCLC samples were Stage I (Table III). All samples were randomized and cohorts were distributed evenly across the total plates of the study using R and Python.

Validation set. To verify the performance of the selected biomarkers and the final algorithm, a blind and independent sample set of 925 subjects (1,850 samples) was processed.

Sample collection and handling. Human plasma samples were obtained from five blood banks: Asterand, BioReclamation, BioSource, Geneticist, and Proteogenex. Clinical information, such as age, gender, pathology and stage, race, origin, smoking status, and sample collection dates, was obtained. Sample were collected and processed according the respective blood bank's protocols. Plasma samples were transported on dry ice overnight to our storage site in Michigan City, Indiana, USA. Vials were inspected visually for damage upon receipt and were stored at –80˚C until analysis.

Selection of biomarkers. Eighty-two commercially available protein analytes were used for the initial screening. The list was narrowed down to 33 biomarkers with a diagnostic potential for early stage lung cancer: The 33 biomarkers were CA-125, CEA, CYFRA21-1, EGFR/HER1/ErBB1, Gro-Pan, HGF, IL-10, IL-12p70, IL-16, IL-2, IL-4, IL-5, IL-7, IL-8, IL-9, Leptin, LIF, MCP-1, MIF, MIG, MMP-7, MMP9, MPO, NSE, PDGF-BB, Rantes, Resistin, sFasL, SAA, sCD40-ligand, sICAM-1, TNFRI, and sTNFRII.

Table II. *Age range and number of samples per cohort.*

	Training set (554 subjects)				Validation set (925 subjects)				
	Age range	Median age	No. of female	No. of male	Age range	Median age	No. of female	No. of male	Total
Asthma	29-83	59	25	8	19-83	63	54	20	107
Non-smoker	25-77	55	70	70	45-80	55	90	88	318
NSCLC	42-94	66	78	82	42-94	66	84	107	351
Smoker	44-79	53	66	65	25-77	53	82	86	299
Other Cancers	29-83	60	56	34	26-89	63	198	116	404
Total	25-94	59	295	259	19-94	63	508	417	1479

Our process in selecting these biomarkers included: i) a statistical importance as measured by the decrease in Gini impurity, ii) analysis of ratio of biomarker distribution between healthy and diseased states, iii) the biomarker's overall patterns observed for specific cohorts, iv) a known biological relevance of these markers for NSCLC via a physio-pathological pathway and/or through literature, and v) the biomarkers' performance in multiplexed system using the selected developed algorithm model.

Multiplexed immunoassay procedure. This study used a custom-made multiplexed immunoassay developed by Millipore (Billerica, MA, USA) to measure the concentration of selected biomarkers in human plasma samples. The reagent kits were designed on magnetic beads using a capture sandwich immunoassay format. The assay was performed according to the manufacturer's protocol. Briefly, samples were thawed on ice or at 4˚C and were visually inspected for turbidity, hemolysis, lipaemia, and other signs of degradation. The plates were read using the FlexMap 3D (Luminex Technologies, Austin, TX, CA), which measures the fluorescence of the beads and of the bound SA-PE. The Bio-Plex Manager 6.1 (Bio-Rad Laboratories, Hercules, CA, USA) was used for data acquisition using a 5PL logistic curve to obtain analyte concentrations. Sample processing was performed by Eve Technologies Corporation (Calgary, Alberta, Canada).

Multivariate classification and analysis. The multivariate classification methods that were evaluated independently included the Support Vector Machine (SVM), Random Forest (RF), Penalized Regression (LASSO and Ridge Regression), Adaptive Boosting (AdaBoost), and decision rules found by Genetic Algorithms. These algorithms considered two independent measurements of 33 biomarkers from a single subject, their gender and smoking status, and classified each measurement as positive or negative for NSCLC. If any of the measurement for a plasma sample was classified as a subject with NSCLC by the algorithm, then the subject was considered positive for NSCLC. The algorithm that was selected had the best performance, as measured by its sensitivity and specificity, as well as the highest level of stability, as measured by bootstrap estimates of the algorithm's performance.

The penalized logistic regression model. When developing predictive models based on many variables, such as the concentration of numerous biomarkers for NSCLC, logistic regression may fail due to non-convergence, (11) or it may be that regression coefficients for the parameter estimates have a large variance. By reducing the number of parameters, such as the number of biomarkers in the model, one can reduce the variance of the regression coefficients.

Table III. *Breakdown of NSCLC stages.*

NSCLC stage	Training (78F/82M)	Validation (84F/107M)	Total	Percentage
IA/B	147	170	317	90.3%
IIA/B	0	8	8	2.3%
IIIA/B	6	6	12	3.4%
IVA	7	7	14	4.0%
Total	160	191	351	100.0%

F: Female; M: male.

One way to effectively reduce the number of parameters in the model is to constrain the L_p-norm or L_∞-norm of the parameter vector, β, of the model by some positive values, usually denoted as t. A form of penalized regression that allows such constraints is the least absolute shrinkage and selection operator (LASSO) model. The LASSO model adds the constraint that the L_1-norm of the parameter vector, β, is no greater than some given value, t (*i.e.*, p=1). Ridge regression is another form of penalized regression that adds the constraint that the L_2-norm of the parameter vector, β, is no greater than some given value, t (*e.g.* t=2).

The naïve Bayes classifier. The set of Bayes classifiers are a set of classifiers based on Bayes' theorem:

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

All classifiers of this type seek to find the probability (P) that an observation belongs to a class (A or B) given the data for that observation. The class with the highest probability is the one to which each new observation is assigned. Theoretically, Bayes classifiers have the lowest error rates amongst the set of classifiers. In practice, however, this does not always occur due to violations of the assumptions made about the data when applying a Bayes classifier.

Table IV. *Median concentrations and p-Values between NSCLC and healthy controls.*

		Lung cancer (160)		Healthy (140)		
Biomarker	Units	Median	Q1, Q3	Median	Q1, Q33	*p*-Value
CA125	U/ml	6.9	4.5, 14.4	4.3	3.6, 5.2	<0.001
CCL20/MIP-3α	pg/ml	157.6	112.8, 240.7	175.9	104.2, 236.6	0.087
CEA	ng/ml	3.1	2, 6.1	1.9	1.5, 2.4	0.001
CXCL11/I-TAC	pg/ml	71.2	41.9, 154.5	53.0	34.2, 87.9	0.288
CXCL9/MIG	ng/ml	2.1	1.3, 4.2	1.2	0.7, 2.2	0.052
CYFRA 21-1	ng/ml	2.8	1.8, 4.9	3.5	1.2, 4.7	<0.001
GM-CSF	pg/ml	42.7	27.7, 79.2	41.1	29.8, 57	0.019
Granzyme-β	pg/ml	65.5	40.2, 100.2	62.3	39.7, 98.9	0.188
GRO Pan	pg/ml	0.7	0.3, 2.3	0.3	0.2, 0.8	0.248
HGF	pg/ml	409.9	281.9, 566.4	280.2	192, 409.7	0.003
IFN-α2	pg/ml	115.9	85.6, 188.3	123.0	88.7, 187.9	0.662
IFN-β	ng/ml	28.5	17.7, 39.9	27.5	15.9, 35.7	0.089
IFN-γ	pg/ml	21.5	13.1, 31.2	17.5	12.7, 31.9	0.462
IL-10	pg/ml	18.0	13.1, 29.4	15.8	10.4, 22.8	0.006
IL-12P40	pg/ml	36.3	22.4, 49.3	39.5	27.6, 51.4	0.048
IL-12p70	pg/ml	14.9	12.9, 20.2	12.2	9.8, 25.8	0.735
IL-13	pg/ml	106.7	33.5, 219.8	104.9	81.8, 345.1	0.035
IL-15	pg/ml	15.5	10.8, 20.1	40.1	13, 71.1	0.072
IL-16	ng/ml	0.8	0.4, 1.4	0.9	0.3, 1.6	0.298
IL-17A	pg/ml	130.6	87.5, 207.8	118.5	84.5, 185.7	0.672
IL-17F	pg/ml	249.9	164.2, 327	265.8	163.6, 349.6	0.981
IL-1α	pg/ml	172.2	90, 345.7	160.6	111.2, 306.4	0.024
IL-1β	pg/ml	13.6	11.1, 22.2	13.8	10.4, 33.7	0.013
IL-1Rα	pg/ml	69.4	33.3, 190.9	97.5	36.1, 251.5	0.270
IL-2	pg/ml	18.3	16.2, 39.8	21.5	15.2, 29.8	0.095
IL-20	pg/ml	271.3	199.4, 387.7	232.8	165.2, 300.4	0.410
IL-21	pg/ml	294.2	199.3, 510.4	316.8	175.8, 478.5	0.900
IL-22	ng/ml	2.0	1.4, 3.0	2.2	1.4, 3.0	0.789
IL-23	ng/ml	24.5	14.7, 39.6	23.3	16.0, 39.8	0.990
IL-27	ng/ml	4.6	3.1, 7.5	4.5	2.7, 7.6	0.372
IL-3	pg/ml	10.0	9.9, 11.1	11.3	10.2, 51.2	0.454
IL-31	ng/ml	2.3	1.4, 4.1	2.4	1.4, 4.2	0.470
IL-4	pg/ml	70.6	30.5, 446.7	202.1	35.3, 224.1	0.562
IL-5	pg/ml	67.1	34.6, 154.4	104.5	63.9, 149.7	0.183
IL-6	pg/ml	22.1	14.4, 37.3	24.2	21.4, 66.1	<0.001
IL-7	pg/ml	27.4	14.8, 42.2	23.1	12.9, 121.9	0.336
IL-8	pg/ml	35.4	15.7, 178.3	17.1	14.3, 40.3	0.068
IL-9	pg/ml	15.6	13.1, 24.2	15.1	12, 65.2	0.024
IP-10	ng/ml	0.8	0.5, 1.1	0.6	0.4, 0.9	0.736
Leptin	ng/ml	21.9	9.5, 37.0	19.9	5.6, 54.0	0.912
LIF	pg/ml	58.4	32, 98.7	32.3	26.8, 55.6	0.185
MCP-1	pg/ml	553.4	405, 723.3	343.1	245.4, 476.1	<0.001
MCP-3	pg/ml	103.9	64.9, 132.7	124.7	57.3, 355.5	0.276
M-CSF	ng/ml	4.5	2.3, 14.4	5.0	4.7, 11.1	0.029
MIF	pg/ml	398.1	217.4, 1152.9	277.3	176.5, 473.2	0.638
MIP-1α	pg/ml	45.5	29.7, 50.8	42.2	33.5, 59.1	0.129
MIP-1β	pg/ml	70.4	45.4, 89.2	67.0	46.4, 81.3	0.946
MMP-7	ng/ml	4.0	2.6, 5.5	2.2	1.5, 3.1	<0.001
MMP-9	ng/ml	51.1	30.4, 90.8	37.4	13.4, 61.5	0.055
MPO	mg/ml	0.1	0.07, 0.3	0.1	0.07, 0.2	0.060
NGF	pg/ml	90.1	43.4, 146.1	80.8	43.2, 136.5	0.858
NSE	ng/ml	6.9	4.4, 10.9	6.6	4.5, 9.3	0.004
OPG	pg/ml	693.5	515.6, 883.7	492.4	380.8, 598.3	<0.001
PAI-1-(total)	ng/ml	73.2	56.5, 101.8	54.9	40.8, 80.4	0.733
PDGF-AB/BB	ng/ml	45.7	32.1, 78.5	38.8	30.3, 53.1	0.022
PLGF	pg/ml	76.0	43.4, 116.9	73.7	36.5, 95.8	0.006

Table IV. *Continued*

Table IV. *Continued*

Biomarker	Units	Lung cancer (160)		Healthy (140)		*p*-Value
		Median	Q1, Q3	Median	Q1, Q33	
RANKL	ng/ml	1.1	0.6, 2.0	1.3	0.7, 1.9	0.069
RANTES	ng/ml	45.3	20.9, 119.4	29.1	17.1, 50.5	0.086
Resistin	ng/ml	17.8	13.6, 27.8	13.2	10.3, 17.1	0.197
SAA	mg/ml	20.0	7.8, 65.4	9.0	4.3, 20.8	0.095
sCD40L	pg/ml	143.0	82.3, 391.6	228.2	119.1, 675.4	0.476
SCF	pg/ml	77.8	53.5, 92.6	59.7	44.5, 75.6	0.829
SDF-1α-β	ng/ml	0.7	0.5, 1.0	0.8	0.6, 1.0	0.468
sEGFR	pg/ml	522.3	399.2, 742.2	510.3	379.6, 684.3	0.760
sE-Selectin	ng/ml	56.2	41.6, 69.5	71.2	55.3, 87.7	0.020
sFasL	pg/ml	65.4	53.9, 72.5	60.5	51.2, 82.6	0.143
sHer2	ng/ml	5.1	4.3, 5.9	5.0	4.4, 6.2	0.094
sICAM-1	mg/ml	0.8	0.5, 1.1	0.4	0.3, 0.6	0.183
sIL-2Rα	ng/ml	1.1	0.8, 1.5	0.8	0.6, 1.1	0.004
sTNFRII	ng/ml	8.7	6.6, 12.9	6.3	5.5, 8.0	<0.001
sVCAM-1	mg/ml	1.1	0.9, 1.2	1.0	0.9, 1.2	0.478
TGF-α	pg/ml	18.0	12.0, 24.0	33.2	21.8, 61.7	0.069
TGF-β1	ng/ml	2.0	1.3, 2.9	2.0	1.3, 2.5	0.158
TSP-1	mg/ml	5.7	2.9, 8.2	3.5	1.7, 6.7	0.401
TSP-2	ng/ml	3.0	2.2, 4.2	2.7	1.9, 3.6	0.019
TNF-α	pg/ml	22.7	17.1, 30	17.5	13.9, 22	<0.001
TNF-β	ng/ml	0.3	0.1, 0.6	1.1	0.7, 1.16	0.767
TNFRI	ng/ml	1.4	1.0, 2.2	1.0	0.8, 1.2	0.575
TPO	ng/ml	0.7	0.5, 1.2	2.3	1.3, 12.0	0.174
TRAIL	pg/ml	128.3	82.5, 159.6	110.0	84.1, 137.7	0.016
VEGF-A	pg/ml	157.7	98.2, 373.6	175.2	82.3, 283.7	0.718
VEGF-C	ng/ml	0.7	0.5, 1.1	0.8	0.5, 1.1	0.502

p-Values are from *t*-test and are unadjusted for multiple comparisons.

The naïve Bayes classifier is one example of a Bayes classifier. It simplifies the calculations of the probabilities used in classification by assuming that each class is independent of the other classes given the data. Naïve Bayes classifiers are used in many prominent anti-spam filters due to the ease of implantation and speed of classification but have the drawback that the assumptions required are rarely met in practice. Tools for implementing naïve Bayes classifiers as discussed herein are available for the statistical software computing language and environment, R. For example, the R package "e1071," version 1.5-25, includes tools for creating, processing and using naïve Bayes classifiers.

Neural nets. One way to think of a neural net is as a weighted directed graph where the edges and their weights represent the influence each vertex has on the others to which it is connected. There are two parts to a neural net, the input layer (formed by the data) and the output layer (the values, which in this case are the classes to be predicted). Between the input layer and the output layer is a network of hidden vertices. There may be, depending on the way the neural net is designed, several vertices between the input layer and the output layer.

Neural nets are widely used in artificial intelligence and data mining, but there is the danger that the models the neural nets produce will over-fit the data (*i.e*. the model will fit the current data very well but will not fit future data well).

Tools for implementing neural nets as discussed herein are available for the statistical software computing language and environment, R. For example, the R package "e1071," version 1.5-25, includes tools for creating, processing, and using neural nets.

k-Nearest neighbor classifiers. The nearest neighbor classifiers are a subset of memory-based classifiers. These are classifiers that must "remember" what is in the training set in order to classify a new observation. Nearest neighbor classifiers do not require a model to be fit. To create a k-nearest neighbor (knn) classifier, the following steps are taken:

i) Calculate the distance from the observation to be classified to each observation in the training set. The distance can be calculated using any valid metric, though Euclidian and Mahalanobis distances are often used.

ii) Count the number of observations amongst the k nearest observations that belong to each group. The group that has the highest count is the group to which the new observation is assigned. Nearest neighbor algorithms have problems dealing with categorical data due to the requirement that a distance needs to be calculated between two points but that can be overcome by defining a distance arbitrarily between any two groups. This class of algorithm is also sensitive to changes in scale and metric. With these issues in mind, nearest neighbor algorithms can be very powerful, especially in large data sets.

Table V. *Biomarker importance list*.

Biomarker	Importance	Biomarker	Importance	Biomarker	Importance
MCP-1	15.138	CCL20/MIP-3α	7.680	LIF	4.726
MMP-7	14.794	HGF	7.410	IL-1β	4.642
sCD40L	12.935	PAI-1 total	7.171	GM-CSF	4.375
sE-Selectin	12.787	CXCL11/ITAC	7.057	IL-12P70	4.294
TNFRI	12.292	MIP-1β	6.838	IL-6	4.233
CA125	12.235	Thrombospondin-1	6.658	IL-10	3.840
sICAM-1	12.194	SCF	6.574	IFNy	3.754
CEA	11.837	NSE	6.457	IL-12P40	3.542
OPG	11.115	SDF-1α/B	6.398	IFNα2	3.540
Leptin	11.048	PLGF	6.174	TGF-α	3.448
SAA	10.923	NGF	6.143	IL-3	2.847
MMP-9	9.998	sEGFR	6.021	IL-16	2.517
IL-8	9.453	IFNβ	5.919	TPO	2.431
CXCL9-MIG	9.296	TNFα	5.914	IL-5	2.318
MIF	9.273	VEGF-C	5.858	IL-20	2.080
GRO-pan	9.187	sIL-2Ra	5.767	MIP-1α	2.017
sTNFRII	9.110	IL-17A	5.756	Gender	1.484
IP-10	9.016	IL-1Rα	5.740	IL-9	1.335
MPO	9.015	RANKL	5.706	IL-2	1.227
THBS2	8.858	sFasL	5.694	IL-15	1.028
RANTES	8.754	CYFRA21-1	5.672	MCP-3	1.003
sHer2	8.530	IL-23	5.587	IL-7	0.889
sVCAM-1	8.292	Granzyme-B	5.579	IL-4	0.581
Resistin	8.241	IL-31	5.400	IL-13	0.385
TRAIL	8.049	IL-22	5.180	TNFβ	0.302
PDGF-AB/BB	7.947	IL-17F	5.042	Race	0.000

IL-1a, IL-21, IL-27, M-CSF, TGFb1 and VEGF-A were removed due to missing data.

Tools for implementing k-nearest neighbor classifiers as discussed herein are available for the statistical software computing language and environment, R. For example, the R package "e1071," version 1.5-25, includes tools for creating, processing and using k-nearest neighbor classifiers.

Random forests. Classification trees are typically noisy. Random forests attempt to reduce this noise by taking the average of many trees. The result is a classifier whose error has reduced variance compared to a classification tree.

To grow a forest:

i) Use the algorithm

For b=1 to B, where B is the number of trees to be grown in the forest.

ii) Draw a bootstrap sample.

iii) Grow a classification tree, T_b on the bootstrap sample.

iv) Output the set $\{T_b\}_1^B$. This set is the random forest.

To classify a new observation using the random forest, classify the new observation using each classification tree in the random forest. The class to which the new observation is classified most often amongst the classification trees is the class to which the random forest classifies the new observation. Random forests reduce many of the problems found in classification trees but at the price of interpretability.

Tools for implementing random forests as discussed herein are available for the statistical software computing language and environment, R. For example, the R package "randomForest," version 4.6-2, includes tools for creating, processing, and using random forests.

Results

Biomarker characteristics. The median plasma concentrations for all biomarkers were used to represent the central tendency of the plasma concentrations to provide resistance to bias due to skewed distributions and outliers as shown on Table IV and Figure 1. The *p*-Values using the T-test, unadjusted for multiple comparisons, are statistically significant at a 0.05 level for 25 of 82 biomarkers (Table IV). From a univariate perspective, this indicates there are many biomarkers that discriminated NSCLC from other pathologies to a degree.

The biomarkers were ordered based on importance in the Random Forest model used to distinguish between NSCLC and non-NSCLC samples (Table V). A biomarker's importance was defined as the decrease in Gini impurity (12) between a model, including the biomarker, and one without it. Thirty-three plasma proteins were selected based on statistical and biological significance. We conclude that gender is marginally significant, as otherwise shown in previous studies done by Izbicka *et al*. (13), and race is not an important factor.

Table VI. *10-Fold cross-validation for 6 multivariate classification algorithms and 33 biomarkers.*

	Accuracy (CI)	Sensitivity (CI)	Specificity (CI)	PPV (CI)	NPV (CI)
RF	0.899 (0.851-0.935)	0.982 (0.921-0.998)	0.865 (0.802-0.914)	0.747 (0.640-0.835)	0.992 (0.963-0.999)
AdaBoost	0.884 (0.834-0.923)	0.947 (0.866-0.985)	0.858 (0.794-0.901)	0.730 (0.621-0.821)	0.956 (0.937-0.993)
Lasso	0.869 (0.816-0.910)	0.912 (0.818-0.968)	0.851 (0.785-0.902)	0.712 (0.602-0.806)	0.960 (0.915-0.985)
RR	0.869 (0.816-0.910)	0.895 (0.796-0.955)	0.858 (0.794-0.901)	0.718 (0.607-0.813)	0.956 (0.937-0.993)
GA	0.798 (0.738-0.849)	0.790 (0.671-0.879)	0.801 (0.730-0.861)	0.616 (0.502-0.723)	0.904 (0.843-0.946)
SVM	0.864 (0.811-0.906)	0.877 (0.774-0.943)	0.858 (0.794-0.901)	0.714 (0.601-0.810)	0.945 (0.896-0.975)

NPV: Negative predictive value; PPV: positive predictive value; CI: 95% confidence interval; SVM: support vector machine; RF: random forest; RR: ridge regression; GA: genetic algorithms. All values for model performance are calculated using a 10-fold cross validation and are by subject results.

Multivariate analysis. To calculate the performance of the model, we used 10-fold cross validation. The data was divided into 10 partitions, using nine partitions to train and evaluate the model on the remaining partition. The process is repeated until all partitions are used to evaluate the model. The average of all results from each step represents an unbiased estimate of the model's performance.

The results of the algorithm comparison indicated that the Random Forest was the best algorithm, in conjunction with the selected 33 biomarkers, to use for developing the LCDT1. This model has an accuracy of 89.9%, sensitivity of 98.2%, and specificity of 86.5% for NSCLC (Table VI) in the Training Set.

AUC/ROC curves. The ROC Curve (Figure 2) for the 33 biomarkers have an AUC of 0.963. When other non-NSCLC cancers were removed from analysis, the AUC improved to 0.974. This indicates potential for clinical use.

Table VII. *Validation performance using the 33 biomarkers and developed classifier.*

	A, NS, all cancers, S (925 subjects)		NS, NSCLC, S (537 subjects)	
	Value	95% CI: LCL, HCL	Value	95% CI: LCL, HCL
Accuracy	0.820	0.794, 0.843	0.900	0.873, 0.924
Sensitivity	0.803	0.742, 0.855	0.803	0.742, 0.855
Specificity	0.823	0.742, 0.850	0.954	0.928, 0.972
PPV	0.539	0.481, 0.597	0.904	0.853, 0.942
NPV	0.942	0.922, 0.958	0.899	0.865, 0.927

LCL: Lower confidence limit; HCL: higher confidence limit; A: Asthma; NS: non-smoker; S: smoker; NSCLC: non-small cell lung cancer. All cancers include breast, ovarian, prostate, pancreatic, colon-rectal cancer, and non-small cell lung cancer. PPV: positive predictive value; NPV: negative predictive value.

Validation performance. An independent sample set of 925 (N=1,850) subjects was processed using the selected 33 biomarkers and developed classifier. The validation results yielded 82% accuracy, 80.3% sensitivity, and 82.3% specificity (Table VII). The cohorts consisted of 74 asthma sufferers, 178 healthy non-smokers, 191 NSCLC, 168 high-risk smokers, and 314 other cancer patients. The majority of NSCLC samples were Stage I (Table III).

The performance of the algorithm improved when the asthma and non-NSCLC cancers were removed from the data set. A total of 537 samples (178 non-smokers, 191 NSCLC, and 168 smokers) were included in this data set. The results yielded an increase in accuracy to 90% and specificity to 95.4% with the sensitivity consistent at 80.3% (Table VII).

Significant markers. Biomarkers that were upregulated by more than 50% using median concentration in lung cancer patients compared to healthy non-smokers included IL-7 (530%), IL-10 (272%), SAA (268%), MMP-9 (251%), IL-8 (247%), Gro (226%), MIG (226%), Rantes (191%), TNFRI (185%), and Resistin (150%). sCD40L and IL-5 showed a 56-57% down-regulation in NSCLC patients compared to healthy non-smokers (Figure 1).

The markers selected in our panel were a mixture of markers that may be significant for NSCLC, Non-Smokers (healthy), Smokers (high-risk, no cancer), and Asthma sufferers.

Assay precision. Inter-assay precision was determined using a low and high-quality control processed in duplicates over 24 plates, performed over 4 days by 4 different operators using 4 Luminex platforms. Biomarker inter-assay precision was between 2.8-5.5% and 3.8-7.9% for all analytes using MFI and concentration CV, respectively. Total plate %CV for each plate ran for all panels were between 3.4-6.5% and 5-11.4% for all analytes (using blanks, standards, QC and samples) using MFI and concentration CV, respectively. No cross-reactivity between analytes in each panel was observed.

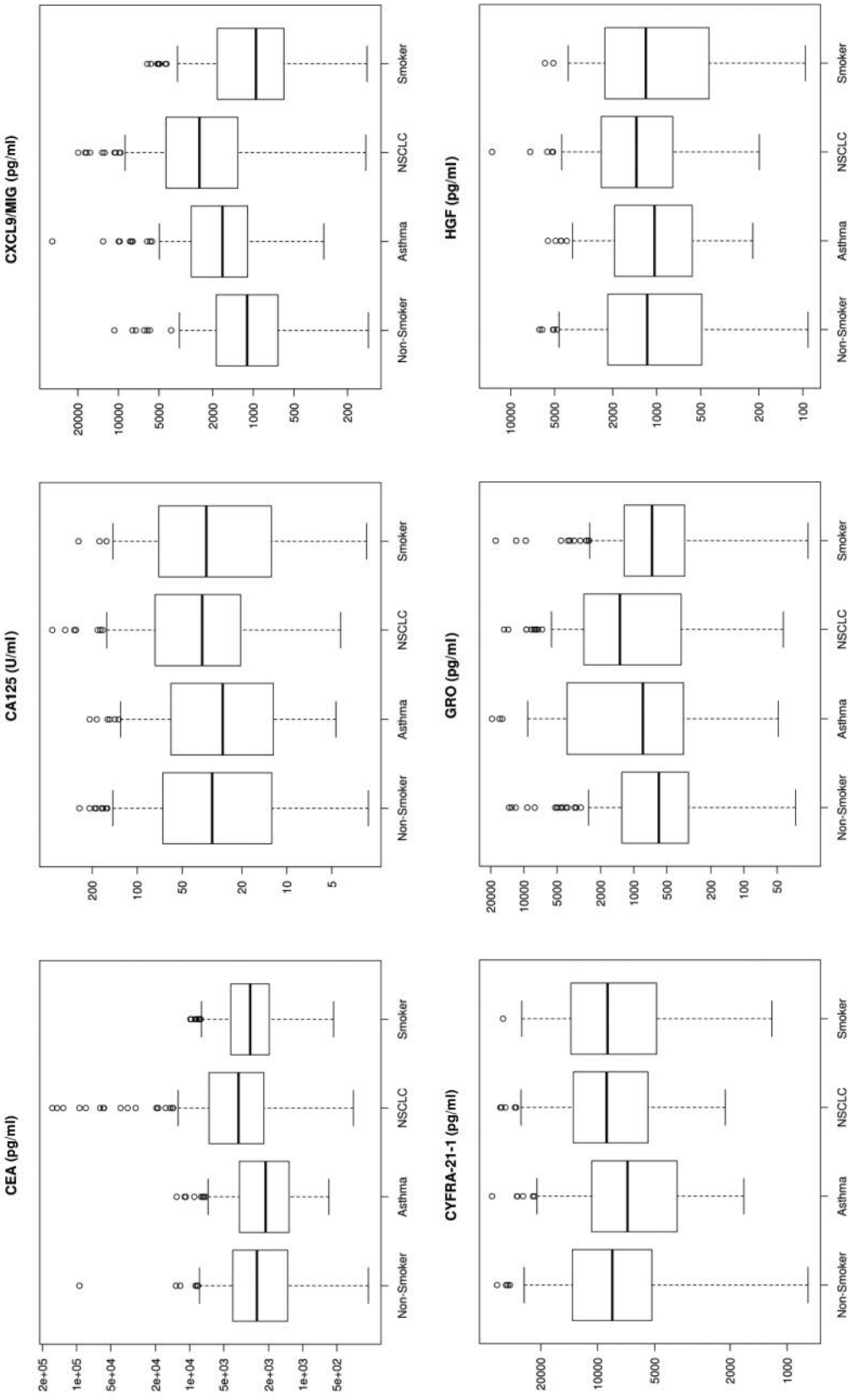

Figure 1. *Continued*

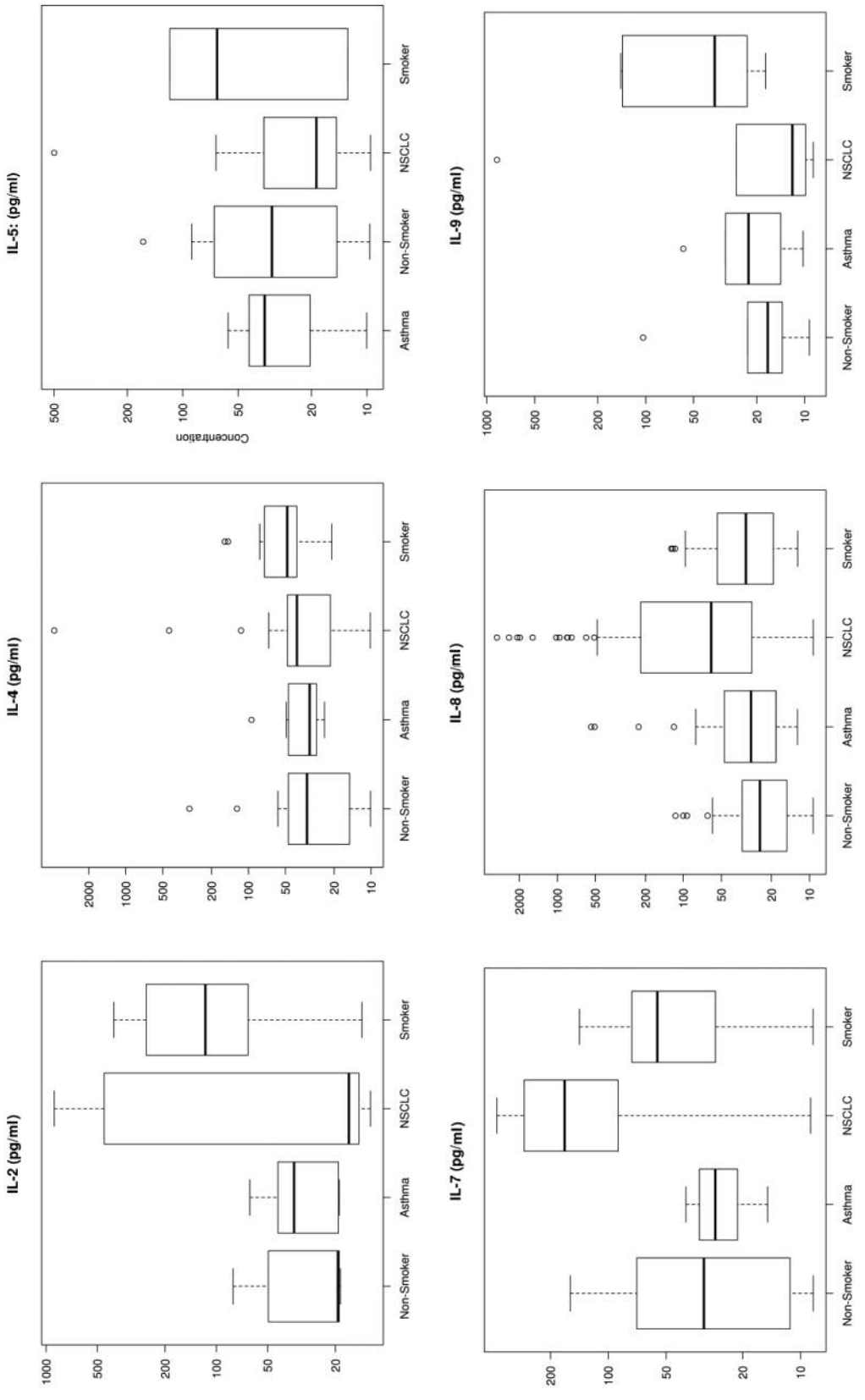

Figure 1. *Continued*

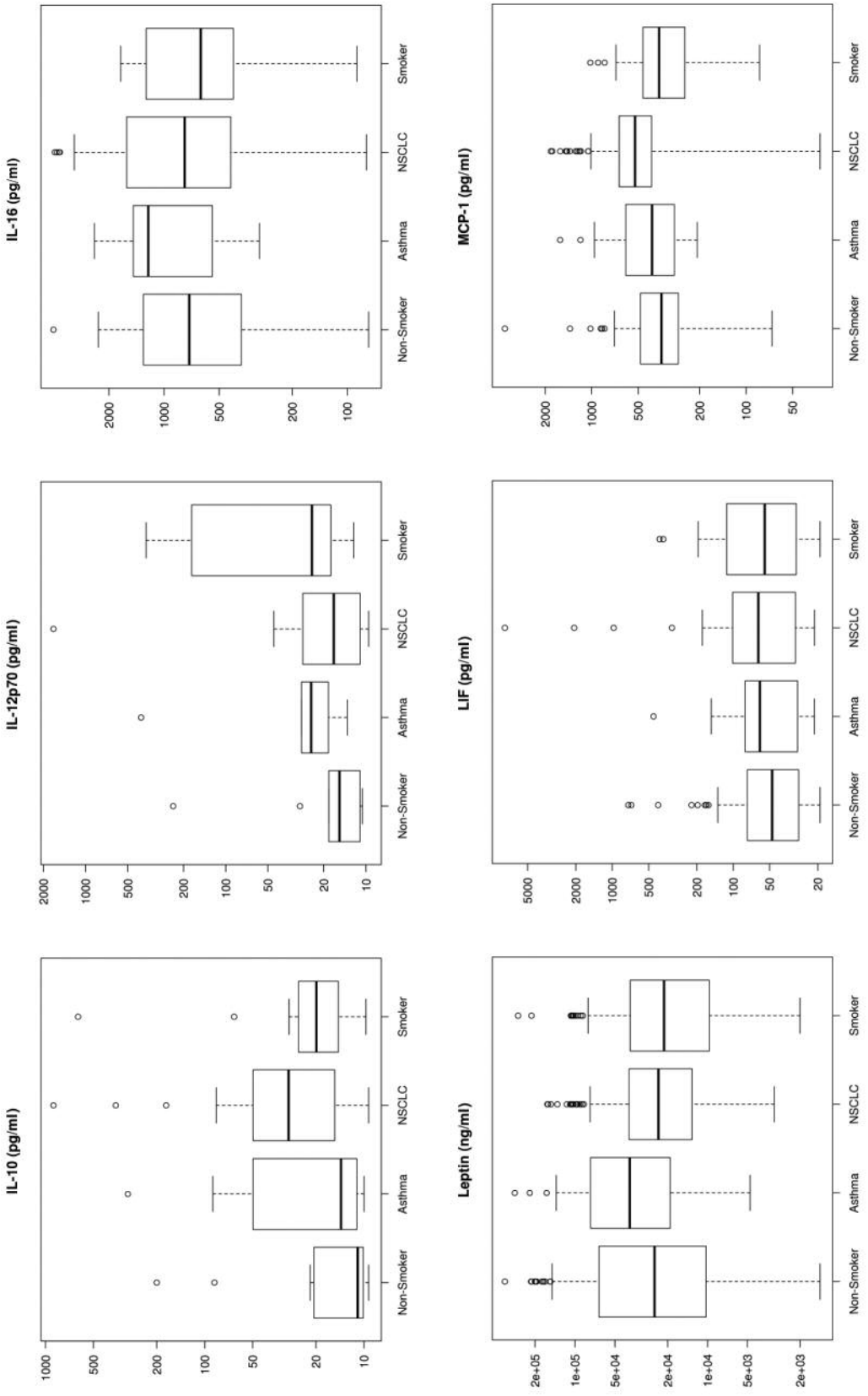

Figure 1. *Continued*

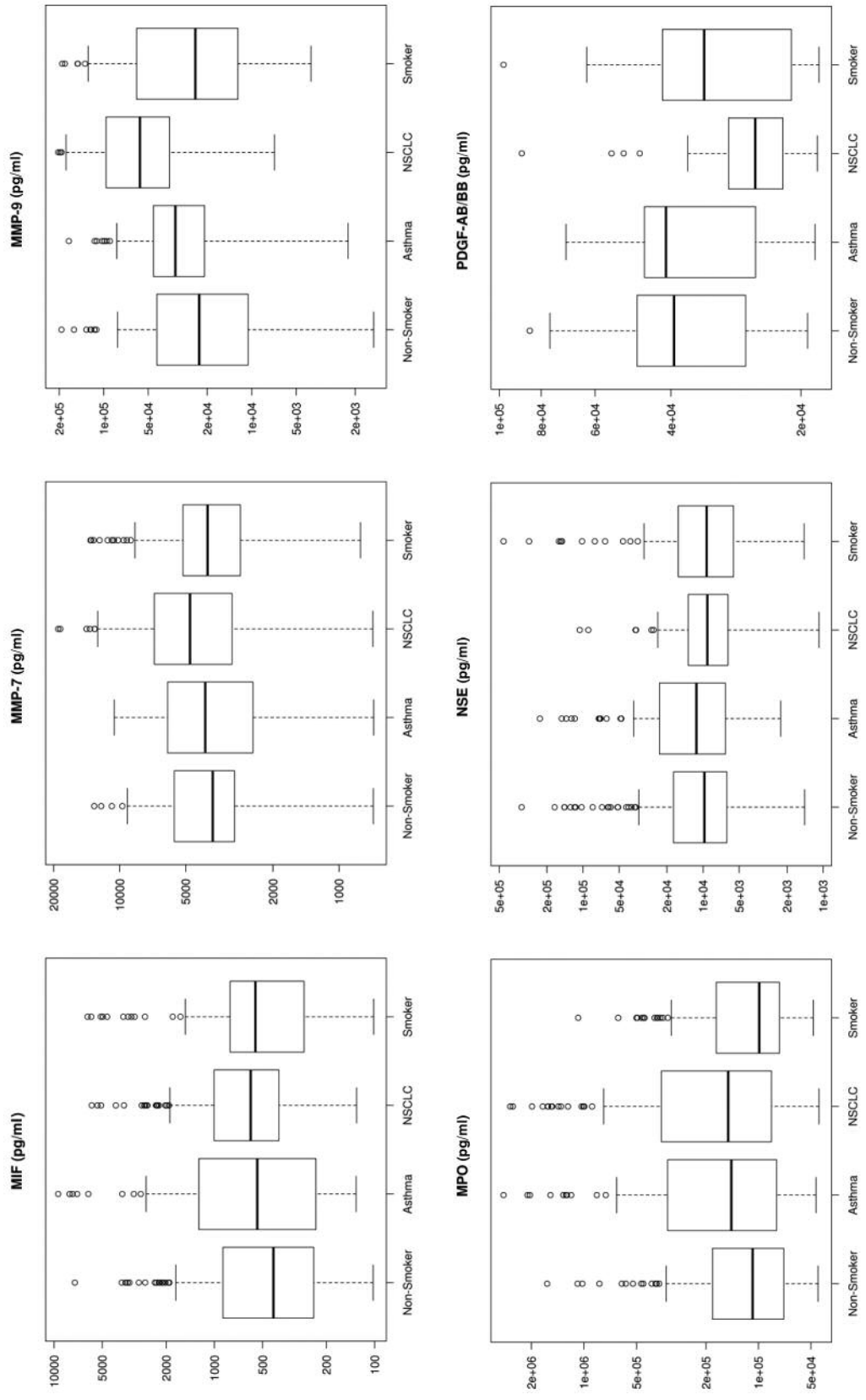

Figure 1. *Continued*

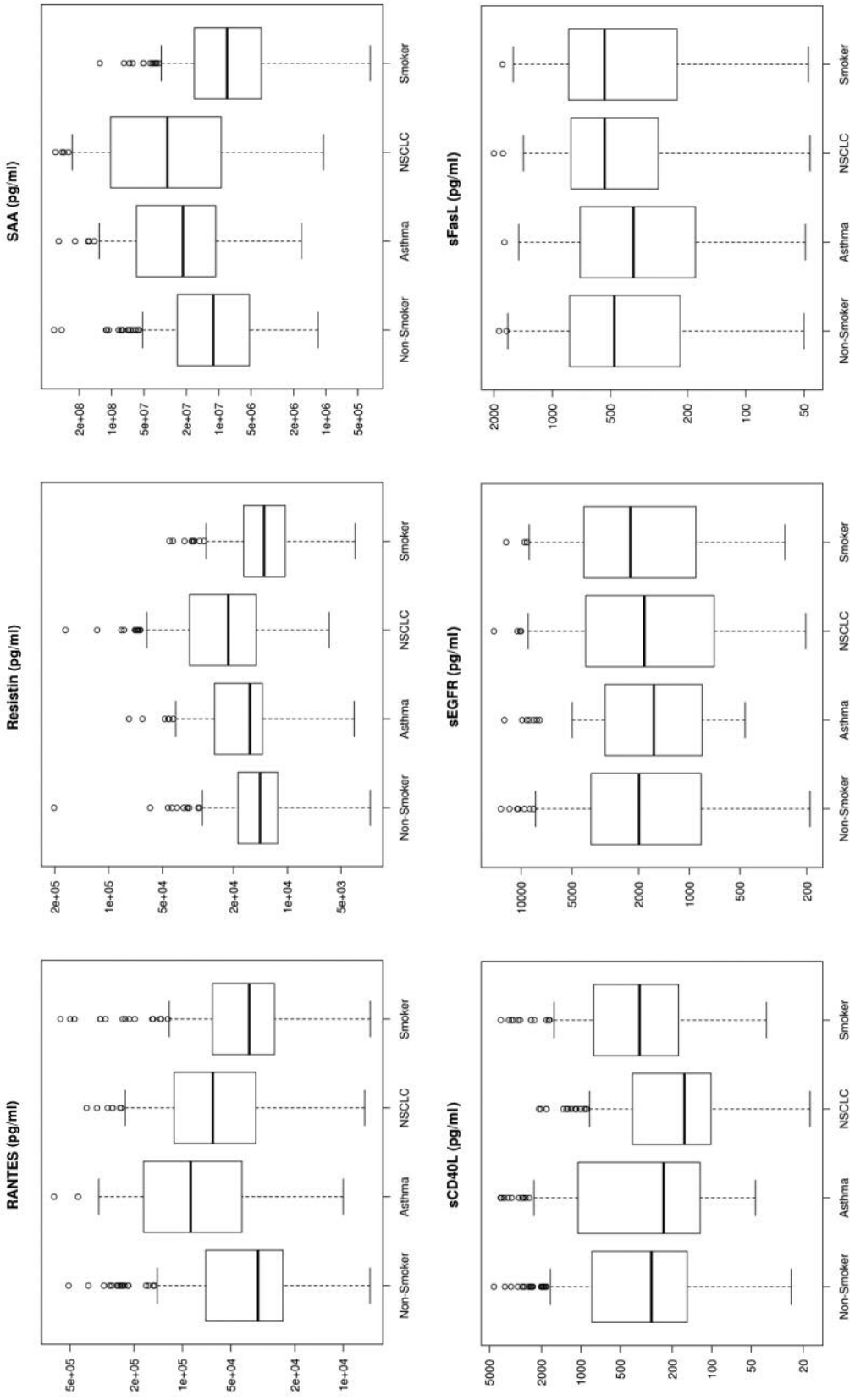

Figure 1. *Continued*

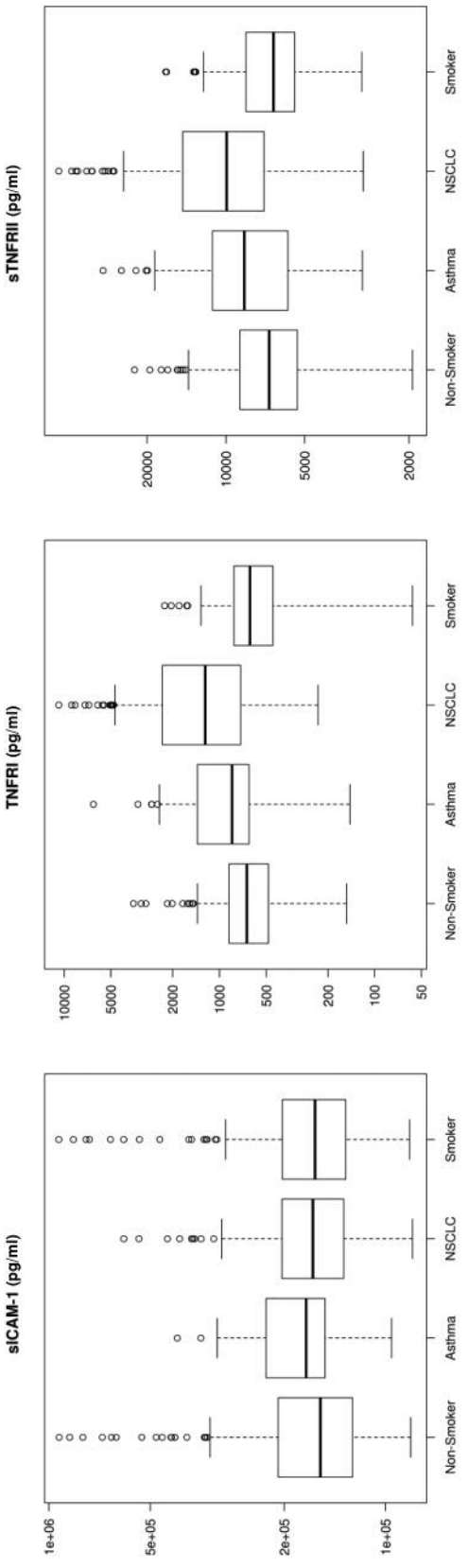

Figure 1. *Box plots for each of the 33 analytes. These plots show the distribution of each pathology cohort's biomarker concentration. Each box represents the interquartile range (from the 25th percentile to the 75th). The bar in the middle of the box is the median. The tails on either end of the box (if present) reach out to the minimum and maximum. If there are values greater than 1.5 times the interquartile range plus the 75th percentile (or less than the 25th percentile minus 1.5 times the interquartile range), they are represented by individual points.*

Discussion

Multiplex assays have been evaluated as a potential diagnostic tool for lung cancer in numerous studies (13-19). Several biomarkers that can statistically discriminate lung cancer from healthy populations have been identified, and our results correlate with these findings. Examples for such biomarkers include (but are not limited to): i) CEA, ii) CYFRA21-1, iii) NSE, iv) VEGF-C, v) MMP7, and vi) MMP9. Independent lung tumor studies by Okamura *et al*., and Wieskopf *et al*., have shown a sensitivity range of 43% and 59% with specificity of 89% and 94%, respectively, for CYFRA21-1 in detecting LC (14, 15). CEA has been shown to have a sensitivity of 69% and a specificity of 68% for LC (14). For NSE, an SCLC detection sensitivity as high as 74% has been reported (16). Tamura *et al*., (17) in 2004 showed elevated serum levels for VEGF-C, MMP-9, and VEGF in lung cancer patients with lymph node metastasis. These markers had individual sensitivities of 85%, 63%, 80%, and specificities of 68%, 75%, and 59%, respectively. The performance of a combination of these 3 markers was at 83% sensitivity and 76% specificity (17, 18). Other studies have shown that Matrix Metallo-Proteinases expression is associated with lung cancer tissue growth (19) and MMP-9 as a prognostic indicator of relapse in patients with lung adenocarcinoma (20). Our results are consistent with previous literature reports which indicate an increased expression in MMP-7 (127%) and MMP-9 (251%) in lung cancer patients (Figure 1).

Naturally, attempts have been made to improve the diagnostic value of potential biomarkers using different combinations (of markers and algorithms), aiming to develop a product that would complement current gold standard diagnostic methods for lung cancer. Such an example is Paula's Test (21), which uses 4 biomarkers (CEA, CYFRA21-1, CA125, and NY-ESO-1) in blood serum and has a sensitivity of 77% and a specificity of 80% using a validation set (N=150) that includes NSCLC and healthy cohorts. Another study performed by Somalogic (22, 23) used 8 tissue homogenates and 1,326 serum NSCLC samples, which resulted in the development of a 12-protein panel (cadherin-1, CD30 ligand, endostatin, HSP90α, LRIG3, MIP-4, pleiotrophin, PRKCI, RGM-C, SCF-sR, sL-selectin, and YES). Somalogic's panel could distinguish NSCLC (Stage I-III) from controls, concluding with a sensitivity of 89% and a specificity of 83% in their blind study. In our study, the LCDT1 used 33 biomarkers in blood plasma achieving a sensitivity of 80.3% and a specificity of 82.3% using a validation set of 925 subjects consisting of NSCLC, asthma sufferers, smokers, non-smokers, and other cancer patients. Notably, specificity increased to 95.4% when samples were restricted to NSCLC from healthy controls (*e.g*. smoker, non-smoker) (Table VII).

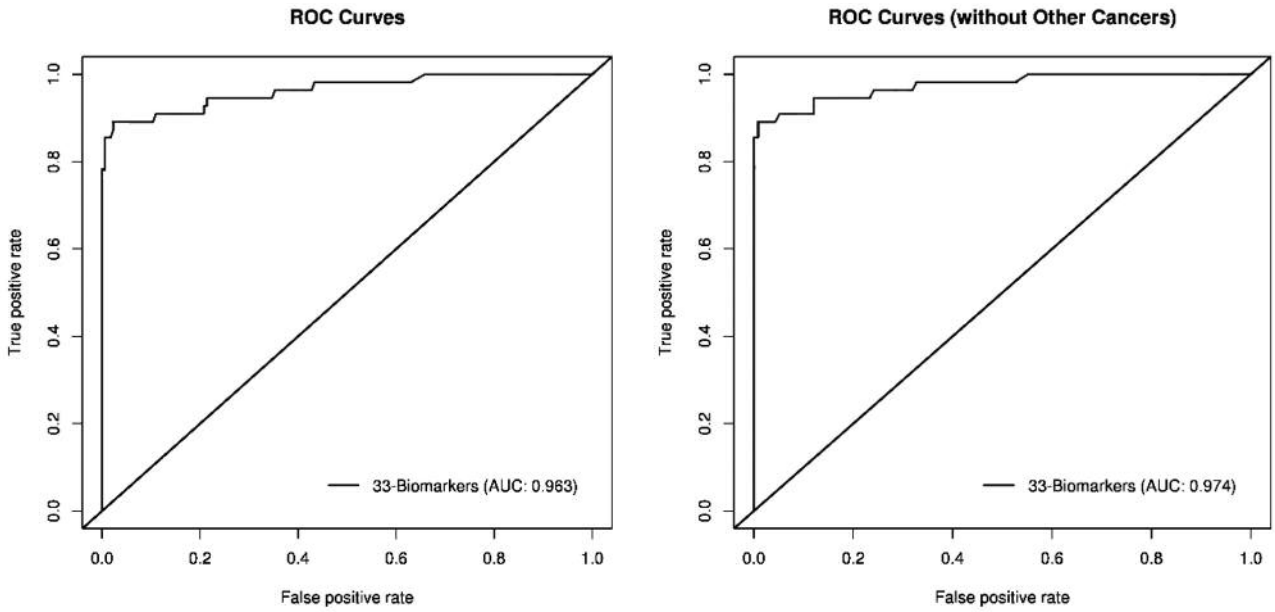

Figure 2. *AUC/ROC graphs.*

The enhanced performance in the algorithm when non-NSCLC cancers and asthma were removed may indicate that i) the number of non-NSCLC cancer and asthma samples were insufficient to create a model that distinguished the differences between these samples, and/or ii) other cancers and asthma may have a masking effect on the NSCLC population affecting the results. This phenomenon requires further investigation.

The combinations of biomarkers, algorithms, sample types, and platforms displayed varying results. We found that the number of proteins (too few), as well as selectivity of these protein markers (too general), in a panel can cause the test to become generic (or applicable) to many cancers rather than to a specific one. This effect can be attributed to the pleiotropic nature of many of these proteins and overlapping signal response pathways of the immune system. A combination of markers that are individually significant, or those that may even exhibit a pattern between cohorts, for asthma sufferers (or COPD), NSCLC, smokers, and non-smokers, may augment the discrimination effects of the algorithm.

From an overall set of 351 NSCLC samples, 90.3% were Stage I, 2.3% were Stage II, and 7.4% were Stage III/IV (Table III). Our validation results illustrate the algorithm's ability to detect NSCLC, *i.e.*, especially Stage I, with an 82% (95%CI: 79-84) accuracy, 80.3% (95%CI: 74-85) sensitivity, 82.3% (95%CI: 74-84) specificity, 53.9% (95%CI: 48-60) PPV, and 94.2% (95%CI: 92-96) NPV. The International Early Lung Cancer Action Program Investigators (24) showed that LC patients who were diagnosed at Stage I and underwent a surgical resection (*e.g.* lobectomy, wedge resection, segmentectomy, and bilobectomy) one-month post diagnosis had a 92% (95%CI: 88-95) 10-year survival rate.

Currently, Medicare covers the use of LDCT for lung cancer screening using the eligibility criteria for the NLST to define high-risk individuals. Ma *et al.*, estimated that at least 8.6 million Americans qualified as high risk for lung cancer and were recommended to receive annual screening with low-dose CT scans in 2010 (25). A study by Brenner *et al.*, have estimated that in individuals between the ages of 50-70 who undergo an annual CT screen for lung cancer, a total of 1,080 (230 males and 850 females) of 100,000 screened will develop a radiation-related cancer (26, 27). About 2% (95% uncertainty limits, 1%-3%) of the 1.4 million cancers diagnosed in the United States in 2010 could be related to CT scans (27, 28). Hence, a simple blood-based test could significantly reduce radiation-related cancers in the future.

Multiplex immunoassays have been evaluated as a potential diagnostic tool for lung cancer in numerous studies. However, the variability of commercially available protein assays due to the complex nature of proteins (*i.e.*, structure, stability, interactions) has made the translation process of immunoassays into diagnostic tools challenging. A final product would necessitate the development of a consistent reagent kit using an optimal combination of biomarkers coupled with an algorithm that has been thoroughly validated and revalidated for functionality and clinical utility. At present, we have simplified our model by decreasing the number of variables to a subset of 21 biomarkers (from the 33) with promising results (Figure 1).

Conflicts of Interest

Thomas Long serves as the Chief Executive Officer and Chief Financial Officer for Lung Cancer Proteomics LLC. Cherylle Goebel and Chris Louden are consultants for Lung Cancer Proteomics LLC.

Authors' Contributions

CG designed the study and performed result analysis. CLL provided statistical analysis and algorithm development. CG and CLL wrote the manuscript with contribution from the others. All authors read and approved the final manuscript.

Acknowledgements

This study was sponsored by Lung Cancer Proteomics LLC previously known as Cancer Prevention and Cure LLC.

References

1 American Cancer Society. Global Cancer Facts & Figures 4th Edition. Atlanta: American Cancer Society; pp. 25-28, 2018. Available at: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/global-cancer-facts-and-figures/global-cancer-facts-and-figures-4th-edition.pdf. Last accessed on 8th May 2019.

2 SEER 18 2009-2015: All Races, Both Sexes by SEER Summary Stage 2000. Available at: https://seer.cancer.gov/statfacts/html/lungb.html. Last accessed on 8th May 2019.

3 American Cancer Society. Facts & Figures 2019: Table 1. Estimated Number* of New Cancer Cases and Deaths by Sex, US, 2019. American Cancer Society. Atlanta, Ga; p 4, p. 18-19, 2019. Available at: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf. Last accessed on 8th May 2019.

4 Birring SS and Peake MD: Symptoms and the early diagnosis of lung cancer. Thorax *60(4)*: 268-269, 2005. PMID: 15790977. DOI:10.1136/thx.2004.032698

5 Ellis PM and Vandermeer R: Delays in the diagnosis of lung cancer. J Thorac Dis *3(3)*: 183-188, 2011. PMID: 22263086. DOI:10.3978/j.issn.2072-1439.2011.01.01

6 Ten Haaf K, Jeon J, Tammemägi MC, Han SS, Kong CY, Plevritis SK, Feuer EJ, de Koning HJ, Steyerberg EW and Meza R: Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study. PLoS Med *14(4)*: e1002277, 2017. PMID: 28376113. DOI: 10.1371/journal. pmed.1002277

7 Pyenson BS and Tomicki SM: Lung cancer screening: A cost-effective public health imperative. Am J Public Health 108(10): 1292-1293, 2018. PMID: 30207779. DOI: 10.2105/ AJPH.2018.304659

8 SEER Cancer Statistics Review, 1975-2016: Table 15.14 Non-Small Cell Cancer of the Lung and Bronchus (Invasive). National Cancer Institute. Bethesda, MD, 2018. Available at: https://seer.cancer.gov/csr/1975_2016/results_single/sect_15_table.14.pdf. Last accessed on 8th May 2018.

9 National Lung Screening Trial Research Team: The National Lung Screening Trial: overview and study design. Radiology *258*: 243-253, 2011. DOI: 10.1148/radiol.10091808

10 National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM and Sicks JD: Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med *365(5)*: 395-409, 2011. PMID: 21714641. DOI: 10.1056/NEJMoa1102873

11 Louppe G, Wehenkel L, Sutera A and Geurts P: Understanding variable importances in forests of randomized trees. In Advances in Neural Information Processing Systems 26 (NIPS 2013). 1987-2016: Neural Information Processing Systems Foundation, Inc. Available at: https://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized-trees.pdf. Last accessed on 8th May 2019.

12 Dupuy A and Simon RM: Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst *99(2)*: 147-157, 2007. PMID: 17227998. DOI: 10.1093/jnci/djk018

13 Izbicka E, Streeper RT, Michalek JE, Louden CL, Diaz A 3rd and Campos DR: Plasma biomarkers distinguish non-small cell lung cancer from asthma and differ in men and women. Cancer Genomics Proteomics *9*: 27-35, 2012. PMID: 22210046.

14 Okamura K, Takayama K, Izumi M, Harada T, Furuyama K and Nakanishi Y: Diagnostic value of CEA and CYFRA 21-1 tumor markers in primary lung cancer. Lung Cancer *80(1)*: 45-49, 2013. PMID: 23352032. DOI: 10.1016/j.lungcan.2013.01.002

15 Wieskopf B, Demangeat C, Purohit A, Stenger R, Gries P, Kreisman H and Quoix E: Cyfra 21-1 as a biologic marker of non-small cell lung cancer. Evaluation of sensitivity, specificity, and prognostic role. Chest *108(1)*: 163-69, 1995. PMID: 7541742.

16 Ferrigno D, Buccheri G and Giordano C: Neuron-specific enolase is an effective tumour marker in non-small cell lung cancer (NSCLC). Lung Cancer *41(3)*: 311-320, 2003. PMID: 12928122.

17 Tamura M, Oda M, Matsumoto I, Tsunezuka Y, Kawakami K, Ohta Y and Watanabe G: The combination assay with circulating vascular endothelial growth factor (VEGF)-C, matrix metalloproteinase-9, and VEGF for diagnosing lymph node metastasis in patients with non-small cell lung cancer. Ann Surg Oncol *11(10)*: 928-933, 2004. PMID: 15383417. DOI: 10.1245/ASO.2004.01.013

18 Ahn JM and Cho JY: Current serum lung cancer biomarkers. J Mol Biomark Diagn S4:001, 2013. DOI:10.4172/2155-9929.s4-001

19 Safranek J, Pesta M, Holubec L, Kulda V, Dreslerova J, Vrzalova J, Topolca n O, Pesek M, Finek J and Treska V: Expression of MMP-7, MMP-9, TIMP-1 and TI.MP-2 mRNA in lung tissue of patients with non-small cell lung cancer (NSCLC) and benign pulmonary disease. Anticancer Res *29(7)*: 2513-2517, 2009. PMID: 19596921.

20 Lee CY, Shim HS, Lee S, Lee JG, Kim DJ and Chung KY: Prognostic effect of matrix metalloproteinase-9 in patients with resected Non-small cell lung cancer. J Cardiothorac Surg *10(1)*: 44, 2015. PMID: 25888323. DOI: 10.1186/s13019-015-0248-3

21 Doseeva V, Colpitts T, Gao G, Woodcock J and Knezevic V: Performance of a multiplexed dual analyte immunoassay for the early detection of non-small cell lung cancer. J Transl Med *13*: 55, 2015. PMID: 25880432. DOI: 10.1186/s12967-015-0419-y

22 Ostroff RM, Bigbee WL, Franklin W, Gold L, Mehan M, Miller YE, Pass HI, Rom WN, Siegfried JM, Stewart A, Walker JJ,

Weissfeld JL, Williams S, Zichi D and Brody EN: Unlocking biomarker discovery: large scale application of aptamer proteomic technology for early detection of lung cancer. PLoS One *5*: e15003, 2010. PMID: 21170350. DOI: 10.1371/journal.pone.0015003.

23 Mehan MR, Ayers D, Thirstrup D, Xiong W, Ostroff RM, Brody EN, Walker JJ, Gold L, Jarvis TC, Janjic N, Baird GS and Wilcox SK: Protein signature of lung cancer tissues. PLoS One *7(4)*: e35157, 2012. PMID: 22509397. DOI: 10.1371/journal.pone.0035157

24 International Early Lung Cancer Action Program Investigators, Henschke CI, Yankelevitz DF, Libby DM, Pasmantier MW, Smith JP and Miettinen OS: Survival of patients with stage I lung cancer detected on CT screening. N Engl J Med *355(17)*: 1763-1771, 2006. PMID: 17065637. DOI: 10.1056/NEJMoa060476

25 Ma J, Ward EM, Smith R and Jemal A: Annual number of lung cancer deaths potentially avertable by screening in the United States. Cancer *119(7)*: 1381-1385, 2013. PMID: 23440730. DOI: 10.1002/cncr.27813

26 Brenner DJ: Radiation risks potentially associated with low-dose CT screening of adult smokers for lung cancer. Radiology *231(2)*: 440-445, 2004. PMID: 15128988. DOI: 10.1148/radiol.2312030880

27 Linet MS, Slovis TL, Miller DL, Kleinerman R, Lee C, Rajaraman P and Berrington de Gonzalez: Cancer risks associated with external radiation from diagnostic imaging procedures. CA Cancer J Clin *62(2)*: 75-100, 2012. PMID: 22307864. DOI: 10.3322/caac.21132

28 Berrington de Gonzalez A, Mahesh M, Kim KP, Bhargavan M, Lewis R, Mettler F and Land C: Projected cancer risks from computed tomographic scans performed in the United States in 2007. Arch Intern Med *169*: 2071-2077, 2009. PMID: 20008689. DOI: 10.1001/archinternmed.2009.440

Received February 19, 2019
Revised May 9, 2019
Accepted May 13, 2019