Check for updates

ORIGINAL ARTICLE

Multidimensional Cell-Free DNA Fragmentomic Assay for Detection of Early-Stage Lung Cancer

Siwei Wang[1]*, Fanchen Meng[1]*, Ming Li[1]*, Hua Bao[2], Xin Chen[2], Meng Zhu[3], Rui Liu[2], Xiuxiu Xu[2], Shanshan Yang[2], Xue Wu[2], Yang Shao[2,3], Lin Xu[1,4], and Rong Yin[1,4,5,6]

[1]Department of Thoracic Surgery, Jiangsu Key Laboratory of Molecular and Translational Cancer Research, [5]Department of Science and Technology, Nanjing Medical University Affiliated Cancer Hospital, Jiangsu Cancer Hospital, Jiangsu Institute of Cancer Research, Nanjing, China; [2]Geneseeq Research Institute, Nanjing Geneseeq Technology Inc., Nanjing, China; [3]School of Public Health and [4]Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, Jiangsu, China; and [6]Biobank of Lung Cancer, Jiangsu Biobank of Clinical Resources, Nanjing, China

ORCID IDs: 0000-0001-5077-060X (S.W.); 0000-0001-5774-8755 (H.B.); 0000-0003-1871-4477 (X.C.); 0000-0001-5122-1733 (M.Z.); 0000-0002-1642-1243 (R.L.); 0000-0003-4199-7874 (X.X.); 0000-0002-5567-1325 (X.W.); 0000-0003-4585-1792 (Y.S.); 0000-0002-9744-4251 (R.Y.).

Abstract

Rationale: Cell-free DNA (cfDNA) analysis holds promise for early detection of lung cancer and benefits patients with higher survival. However, the detection sensitivity of previous cfDNA-based studies was still low to suffice for clinical use, especially for early-stage tumors.

Objectives: Establish an accurate and affordable approach for early-stage lung cancer detection by integrating cfDNA fragmentomics and machine learning models.

Methods: This study included 350 participants without cancer and 432 participants with cancer. The participants' plasma cfDNA samples were profiled by whole-genome sequencing. Multiple cfDNA features and machine learning models were compared in the training cohort to achieve an optimal model. Model performance was evaluated in three validation cohorts.

Measurements and Main Results: A stacked ensemble model integrating five cfDNA features and five machine learning algorithms constructed in the training cohort (cancer: 113; healthy: 113) outperformed all the models built on individual feature–algorithm combinations. This integrated model yielded superior sensitivities of 91.4% at 95.7% specificity for cohort validation I (area under the curve [AUC], 0.984), 84.7% at 98.6% specificity for validation II (AUC, 0.987), and 92.5% at 94.2% specificity for additional validation (AUC, 0.974), respectively. The model's high performance remained consistent when sequencing depth was down to 0.5× (AUC, 0.966–0.971). Furthermore, our model is sensitive to identifying early pathological features (83.2% sensitivity for stage I, 85.0% sensitivity for <1 cm tumor at the 0.66 cutoff).

Conclusions: We have established a stacked ensemble model using cfDNA fragmentomics features and achieved superior sensitivity for detecting early-stage lung cancer, which could promote early diagnosis and benefit more patients.

Keywords: lung cancer; early detection; cell-free DNA; whole-genome sequencing; machine learning

Lung cancer is the second most common cancer and the leading cause of cancer-related death in the world (1). The survival rates of patients with lung cancer diagnosed at the early localized and late distant stages differ drastically (1), underscoring the importance of early diagnosis for prognosis. Unfortunately, only ∼16% of patients receive a diagnosis at the localized stage (2). Although radiological methods, such as the low-dose computed tomography test, can contribute to a 20% reduction in

(Received in original form September 1, 2021; accepted in final form November 8, 2022)

*These authors contributed equally to this work.

Supported by Key Project of Cutting-edge Clinical Technology of Jiangsu Province grant BE2017759; the Natural Science Fund for Distinguished Young Scholars of Jiangsu Province grant BK20211550; National Natural Science Foundation of China grants 81672295, 81572261, 81802277, 81872378, 81802907, 81702265, and 81602014; and the Project of Jiangsu Provincial Medical Talent grant ZDRCA2016033.

Author Contributions: S.W., Y.S., L.X., and R.Y. participated in conception and design of this study; F.M., M.L., H.B., X.C., R.L., S.Y., and X.W. participated in collection of data; S.W., H.B., X.C., M.Z., R.L., X.X., S.Y., and X.W. provided data analysis and interpretation. S.W., F.M., M.L., and X.C. wrote the manuscript with input from all authors. All authors contributed to manuscript revision and final approval of the manuscript.

Am J Respir Crit Care Med Vol 207, Iss 9, pp 1203–1213, May 1, 2023
Copyright © 2023 by the American Thoracic Society
Originally Published in Press as DOI: 10.1164/rccm.202109-2019OC on November 8, 2022
Internet address: www.atsjournals.org

At a Glance Commentary

Scientific Knowledge on the Subject: Early detection can benefit patients with lung cancer with higher survival rates, but most patients do not receive a diagnosis until metastasis has already occurred. The recent development of cell-free DNA (cfDNA) analysis from liquid biopsy has shown great potential to facilitate the identification of lung cancer. However, improving the detection performance of cfDNA-based assays, especially on early-stage lung cancer, is vital for leveraging their applications.

What This Study Adds to the Field: By integrating five different cfDNA fragmentomic features, the stacked ensemble machine learning model reaches high sensitivity for detecting early-stage lung cancer, exceeding the performance of models built on single features. Its superior performance has also been validated in an external validation dataset and an independent additional dataset. This study demonstrates that the stacked ensemble model is robust in distinguishing lung cancer from healthy subjects using shallow whole-genome sequencing data down to 0.5× coverage depth. Furthermore, its detection ability is consistently high across different disease subtypes, and it can sensitively identify patients with non–small cell lung cancer with very early–stage characteristics. The model in this study has important implications for the current thinking on how to develop accurate and affordable strategies for lung cancer early detection and management in clinical practice.

cancer-related deaths, their usage has been limited because of high false-positive rate, radiation-induced cancer risk, and monetary cost (3–6). A reliable noninvasive approach to detect early-stage lung cancer in an accurate and cost-efficient manner needs to be developed.

The liquid biopsy–based cell-free DNA (cfDNA) analysis has recently opened a new avenue for disease detection. During apoptosis and necrosis, DNA fragments are released into the circulation to form cfDNA (7). They bear the genetic and epigenetic information from the cell and tissue of origin (8). Specifically, a fraction of cfDNA, namely circulating tumor DNA (ctDNA), represents DNA shed from tumor cells (9). Tumor somatic mutations can be detected to distinguish ctDNA and nontumorous cfDNA (10), but the performance of ctDNA mutation calling-based strategy suffers from sensitivity as low as 40% for early-stage lung cancer (11). Alternatively, epigenetic modifications such as DNA methylation and ctDNA fragmentomic signatures, including fragmentation size, have shown diagnostic potential (12–15). However, the sensitivities of the existing methylation or fragmentation size feature-based approach for stage I lung cancer cannot suffice for clinical use, ranging from ∼25% to ∼60% (14, 15). The profile of cfDNA cleavage site motifs represents another class of biomarkers for liquid biopsy in oncology. Recently, researchers have revealed the tumor-associated cfDNA preferred end coordinates in patients with hepatocellular carcinoma, granting a sensitivity of >80% at >90% specificity (16, 17). The utility of cfDNA end motifs, particularly for lung cancer, still needs to be verified.

Recent studies have suggested that incorporating a large number of features in multiple dimensions could improve the machine learning model's discrimination ability for early cancer detection (18). For instance, a proof-of-concept study combining cfDNA methylation, fragmentation, end motif, and nucleosome footprint patterns has reached ∼95% sensitivity for hepatocellular carcinoma prediction at 95–98% specificity, outperforming models based on individual features (19). Combining clinical risk factors, protein biomarker concentrations, imaging analysis, etc., with cfDNA fragment size could further boost the prediction power (15). However, it is conceivable that the assay complexity could increase the monetary cost. In addition, the model's prediction power for disease varies depending on the choice of machine learning classification algorithms (20). Leveraging the power of machine learning, researchers have used the stacked ensemble approach to integrate cfDNA genomic features from whole-genome sequencing (WGS) alone and created a highly sensitive model for early-stage colorectal adenocarcinoma detection (21). The validity of this approach is to be explored in the context of lung cancer.

Here, we established a multidimensional stacked ensemble model for robust detection of early-stage non–small cell lung cancer (NSCLC). This model integrated five cfDNA fragmentomic features and five machine learning base models from our comprehensive characterization and has reached superior detection ability using WGS data. We demonstrated that the predictive model is highly sensitive for detecting early NSCLC pathological features. The consistency of its performance at low sequencing depth down to 0.5× is particularly ideal for affordable lung cancer early screening.

Methods

Participant Enrollment

For model construction and internal validation, we enrolled 354 participants in this study cohort of healthy volunteers without cancer ($n = 160$) and previously untreated early-stage NSCLC ($n = 194$), including adenocarcinoma (ADC, $n = 162$) and squamous cell carcinoma (SCC, $n = 32$) from Jiangsu Cancer Hospital, China. After model construction, we performed another validation using plasma samples of 188 participants (healthy control subjects, $n = 70$; untreated NSCLC, $n = 118$ [ADC, $n = 118$])

Data sharing: Qualified researchers who provide methodologically sound, *bona fide* research proposals may direct data requests to the corresponding author to obtain the deidentified data that support the findings of this study to achieve the aims in the approved proposals.

Correspondence and requests for reprints should be addressed to Rong Yin, M.D., Ph.D., Department of Thoracic Surgery, Jiangsu Key Laboratory of Molecular and Translational Cancer Research, Nanjing Medical University Affiliated Cancer Hospital, Jiangsu Cancer Hospital, Jiangsu Institute of Cancer Research, Baiziting 42, Nanjing 210009, China. E-mail: rong_yin@njmu.edu.cn.

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.

from the plasma inventory (Jiangsu Cancer Hospital, China). The clinical information of individual patients with NSCLC and volunteers without cancer is listed in Tables E1–E3 in the online supplement. For further independent validation, we incorporated and tested additional samples, including 120 noncancer and 120 untreated samples spanning stage I to stage IV, from our plasma biobank (Jiangsu Cancer Hospital, China). The clinical information of this additional validation dataset is listed in Tables E4 and E5. More enrollment information is provided in the online supplement (*see* Participant enrollment). Cancer and noncancer cohorts are all sex and age matched. This study was approved by the ethics committee at Jiangsu Cancer Hospital (approval no. JSLMTCR-2017-002) and complied with the Declaration of Helsinki. All participants provided written informed consent.

cfDNA Extraction and WGS

We performed plasma sample collection and cfDNA extraction followed by WGS as described in the online supplement (*see* cfDNA extraction and sequencing). Briefly, the venous blood samples were collected during routine physical checks (healthy volunteers) or preoperatively (patients with cancer). All samples were collected, shipped, and processed uniformly. A total of 5–10 ng of plasma cfDNA per sample was subject to PCR-free WGS library construction with the KAPA Hyper Prep Kit (KAPA Biosystems). The libraries underwent paired-end sequencing on NovaSeq platforms (Illumina). To minimize bias, the sample operating team was blinded to the case or control status of the samples during the whole process.

Bioinformatic Analysis and Modeling

Raw sequencing data processing was performed as described in online supplement (*see* cfDNA fragmentomic features). The libraries in this study have a mean sequencing depth of 11.08× (range, 5.28×–27.85×). To eliminate the potential impact of coverage difference on the prediction power, we down-sampled the libraries to a unified 5× for model construction and evaluation, ensuring the inclusion of all samples. The selected model was further assessed using the WGS data of raw sequencing depths or down-sampling to 4×, 3×, 2×, 1×, and 0.5×.

We extracted five different fragmentomic features from the WGS data for feature selection and model construction. Details on the fragmentomic features are described in the online supplement (*see* cfDNA fragmentomic features). Briefly, we calculated the copy number variation (CNV) according to Wan and colleagues (22). We also created four new features (*see* cfDNA fragmentomic features in the online supplement). The fragmentation size coverage (FSC) and fragmentation size distribution (FSD) features were developed to depict cfDNA fragment size patterns. Two motif features, 6-bp end motif (EDM) and 6-bp breakpoint motif (BPM), can profile the sequences around the end of cfDNA fragments. Of note, the FSC and EDM features were adopted and developed from the previously published cfDNA fragmentation size (DELFI) and 4-bp end motif features (13, 17). We have demonstrated that FSC and EDM outperformed their counterparts in colorectal adenocarcinoma detection (21) and thus set out our study using the optimized features.

The model training was conducted solely in the training dataset, and the validation datasets remained untouched before the model was finalized. The validation I dataset was used for internal validation and determining the cutoff score of 95% specificity. External validation was conducted in the validation II and the independent additional validation datasets.

For model construction, we used each cfDNA fragmentomic feature to build its base model with five base algorithms: generalized linear model (GLM), gradient boosting machine (GBM), random forest, deep learning, and XGBoost. For every cancer and noncancer sample in this study, the algorithms generated a cancer score ranging from 0 to 1. A higher score output by the models represented a higher probability of cancer. The cancer probability scores from all the algorithms were then ensembled into a matrix and analyzed by a GLM algorithm to create the base models for improved robustness and accuracy, as shown by the machine learning community (23). In addition, we used fivefold cross-validation based on the training dataset to optimize model performance and avoid overfitting. For the base models of individual fragmentomic features and the ensemble machine learning model, we assessed their prediction performance based on the area under the curve (AUC) values in the validation cohorts. The details of constructing base models and stacked ensemble models are provided in the online supplement (*see* Model construction).

Model Analytical Validity Assessment

Three patients with cancer (stage I: one; stage II: one; stage III: one) and three healthy participants were chosen from the additional validation dataset to check the model performance within run and between run. For within-run tests, one tube (~10 ml) of venous blood sample was collected from each participant for plasma preparation (batch 1). The plasma of each participant was evenly divided into three parts for three technical replicates. One group of lab technologists handled all the replicates for cfDNA extraction, library preparation, and sequencing. For between-run tests, another round of blood sample collection was performed 3 weeks later on the six participants (batch 2). The plasma preparation, splitting, cfDNA extraction, library preparation, and sequencing steps were processed following the same procedures by a different group of lab technologists. The within-run and between-run tests yielded 36 samples in total. The fragmentomic features were extracted from the 36 samples and analyzed by the predictive model for their cancer scores. The three scores of each participant within the same batch were used to assess the model's repeatability. Batch 1 and batch 2 results of the same participant were compared to evaluate the model's reproducibility.

To assess the model's sensitivity to low ctDNA fraction samples, we retrieved plasma samples from another study, in which early-stage NSCLC tumor tissue and corresponding plasma samples were ultra–deep sequenced using Geneseeq Prime 425-gene panel (Geneseeq Technology Inc.). Tumor-informed somatic mutations in the 425 predefined cancer-related genes targeted by Geneseeq Prime panel were used to identify the maximum variant allelic fraction (max VAF) from the panel-based plasma sequencing. This strategy allowed us to determine max VAF as low as 0.05%, and the max VAF was used to represent the ctDNA fraction (24). In contrast, CNV-based determination from plasma WGS by ichorCNA is inaccurate when the ctDNA fraction is <1% (22). For the selected patients with low ctDNA fraction, we performed plasma WGS and randomly down-sampled the WGS data to the coverage depths of 5×, 4×, 3×, 2×, 1×, and 0.5× for

20 times each. The resulting WGS fractions were applied to the predictive model for their cancer scores. Their cancer/noncancer status was evaluated using the cutoff determined by the validation I cohort. We defined the percentage of the 20 replicates that detected cancer at a certain sequencing depth as detection sensitivity and used it to quantify the model performance.

Statistical Analysis

For statistical analysis, the receiver operating characteristic curves were generated using the pROC package (v. 1.17.0.1). Based on true positive (TP), true negative (TN), false positive (FP), and false negative (FN) of cancer prediction, we calculated the sensitivity (TP/[TP + FN]), specificity (TN/[TN + FP]), and accuracy ([TP + TN]/[TP + FP + TN + FN]), as well as the corresponding 95% confidence interval (CI), using the epiR package (v. 2.0.19) in R (v. 4.0.3). The Fisher's exact test was performed using GraphPad, and the Wilcoxon rank sum and signed rank tests were performed using R.

Results

Participant Disposition and Characteristics

As shown in Figure 1A, the 354 participants from the study cohort were randomly assigned to the training (113 patients [ADC: 96; SCC: 17; stage I: 66; tumor size < 1 cm: 15]; 113 control subjects), and validation I (81 patients [ADC: 66; SCC: 15; stage I: 46; tumor size < 1 cm: 16]; 47 control subjects) datasets. The training dataset was used to train the model. We then used the validation I dataset to evaluate its performance and determine the cutoff for assessing the validation II dataset. It is worth noting that we built the model exclusively in the training dataset. The samples from patients in an independent study were assigned to the validation II cohort (118 patients [ADC: 118; stage I: 85; tumor size < 1 cm: 4]; 70 control subjects) for external validation.

Participants' demographics and characteristics are summarized in Table E3 and are comparable across the three cohorts. The mean ages and sex distribution of subjects with and without cancer are similar in the cohorts. The cancer groups are highlighted by the majority of early-stage diseases (training: stage I, 66/113 [58.4%]; validation I: stage I, 46/81 [56.8%]; validation II: stage I, 85/118 [72.0%]).

Assessment of cfDNA Fragmentomic Features and Machine Learning Algorithms

We performed model selection by evaluating the AUC values of the cfDNA feature and machine learning algorithm combinations in the validation cohorts. As shown in Table E6, we tested the features of FSC, FSD, EDM, BPM, and CNV in all base models (GLM, GBM, deep learning, random forest, and XGBoost) for their AUC. The comparisons of different base models revealed the superior performance of the algorithm-stacked model, as the EDM, BPM, FSC, FSD, and CNV features all yielded higher AUC values in the stacked model than in the single algorithm models. Therefore, we built a stacked ensemble model integrating the five features of the plasma cfDNA fragmentomic features (EDM, BPM, FSC, FSD, and CNV) and five machine learning algorithms (GLM, GBM, random forest, deep learning, and XGBoost) (Figure 1B). The cancer scores from all the individual and stacked models for every participant are listed in Table E7. The resultant stacked ensemble model boosted the prediction power to an AUC of 0.985 (95% CI, 0.983–0.998), outperforming the stacked models using single fragmentomic features (Figure 2A). In addition, the stacked ensemble model's AUC is higher than that of a GLM model built on the fragmentomic features (Table E6). Therefore, we chose the stacked ensemble model as the predictive model for our following evaluation.

We scrutinized the performance of the stacked ensemble model in the two validation datasets separately. The model AUC values are consistently high, at 0.984 (95% CI, 0.966–1.000) and 0.987 (95% CI, 0.970–1.000) in the validation I and II cohorts, respectively (Figure 2B). We chose the cancer score of 0.66 as the cutoff based on the 95.7% specificity in the validation I cohort. This cutoff score also yielded a specificity of 98.6% when applied to the validation II cohort. The resultant sensitivities are 91.4% (95% CI, 83.0–96.5%) in validation I and 84.7% (95% CI, 77.0–90.7%) in validation II. When the two validation cohorts were combined, the model reached 87.4% (95% CI, 82.0–91.7%) sensitivity at 97.4% (95% CI, 92.7–99.5%) specificity (Table 1). We observed that the predicted cancer scores of patients with cancer are significantly higher than those of healthy participants in both validation datasets (Figure 2C). Notably, we plotted the cancer score of all the patients in the validation cohorts based on their stages and observed an upward trend of score distribution from stage I to stage IV (Figure 2D).

To further assess the generalizability of the stacked ensemble model, we tested it in an independent additional validation cohort consisting of 120 patients with NSCLC and 120 participants without cancer, all collected from other retrospective studies (Table E4). The sample demographics of the additional validation cohort, including age and sex, are comparable between subjects with and without cancer and similar to that of the validation datasets (Table E5). The predictive model reached the AUC of 0.974 (95% CI, 0.956–0.991) in the additional validation dataset (Figure 3A). Applying the 0.66 cutoff score, the model reliably distinguished cancer and noncancer samples (Figure 3B and Table E5) and detected subjects with cancer with a 92.5% (95% CI, 86.2–96.5%) sensitivity at 94.2% (95% CI, 88.4–97.6%) specificity (Table 1). Consistently, the cancer score of all the patients in the additional validation dataset also exhibited an upward trend from stage I to stage IV (Figure 3C). In addition, 32 subjects without cancer were detected with benign lung nodules according to computed tomography scans (Table E4). We scrutinized the benign lung nodule status of the subjects without cancer and found the high specificity of our predictive model is independent of the presence of benign nodules (Fisher's exact test).

Analytical Validity Assessments of the Predictive Model

Next, we sought to evaluate the model's stability and robustness at different WGS coverage depths by applying WGS data with varied coverage depths. Using the raw coverage WGS data of the participants (range, 5.28×–27.85×), the stacked ensemble model built on 5× coverage WGS reached the AUC of 0.988 (95% CI, 0.974–1.000) and 0.989 (95% CI, 0.978–0.998) in the two validation cohorts (Figure E1). We also tested the raw coverage WGS data for model construction, and the resultant model yielded the AUC of 0.978 (95% CI, 0.955–1.000) and 0.990 (95% CI, 0.978–0.998) in the two validation cohorts (Figure E1). Hence, the model can perform

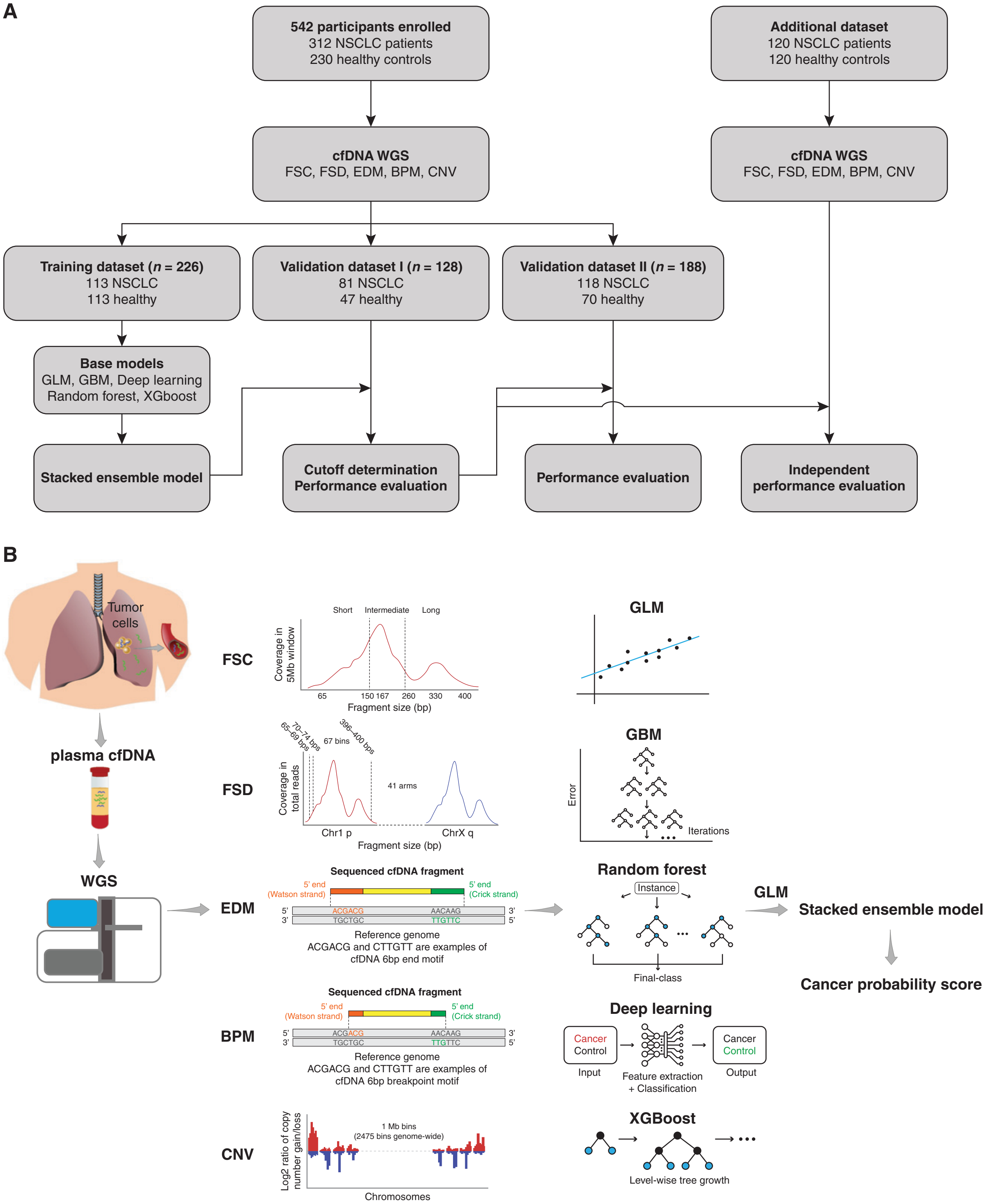

Figure 1. The model illustration and diagnostic performance evaluation in the validation cohorts. (*A*) A total of 542 participants (cancer 312, healthy 230) were included for model construction and validation. Whole-genome sequencing (WGS) of plasma cell-free DNA (cfDNA) was performed, and the five cfDNA features of each subject were profiled. A total of 226 participants (cancer 113, healthy 113) were allocated to training for building the stacked ensemble model from the five base models. One hundred and twenty-eight participants (cancer 81, healthy 47) were allocated to the validation I cohort for assessing the model performance and determining the cutoff score. One hundred and eighty-eight participants (cancer 118, healthy 70) were allocated to an independent validation II cohort for evaluating model performance. An additional

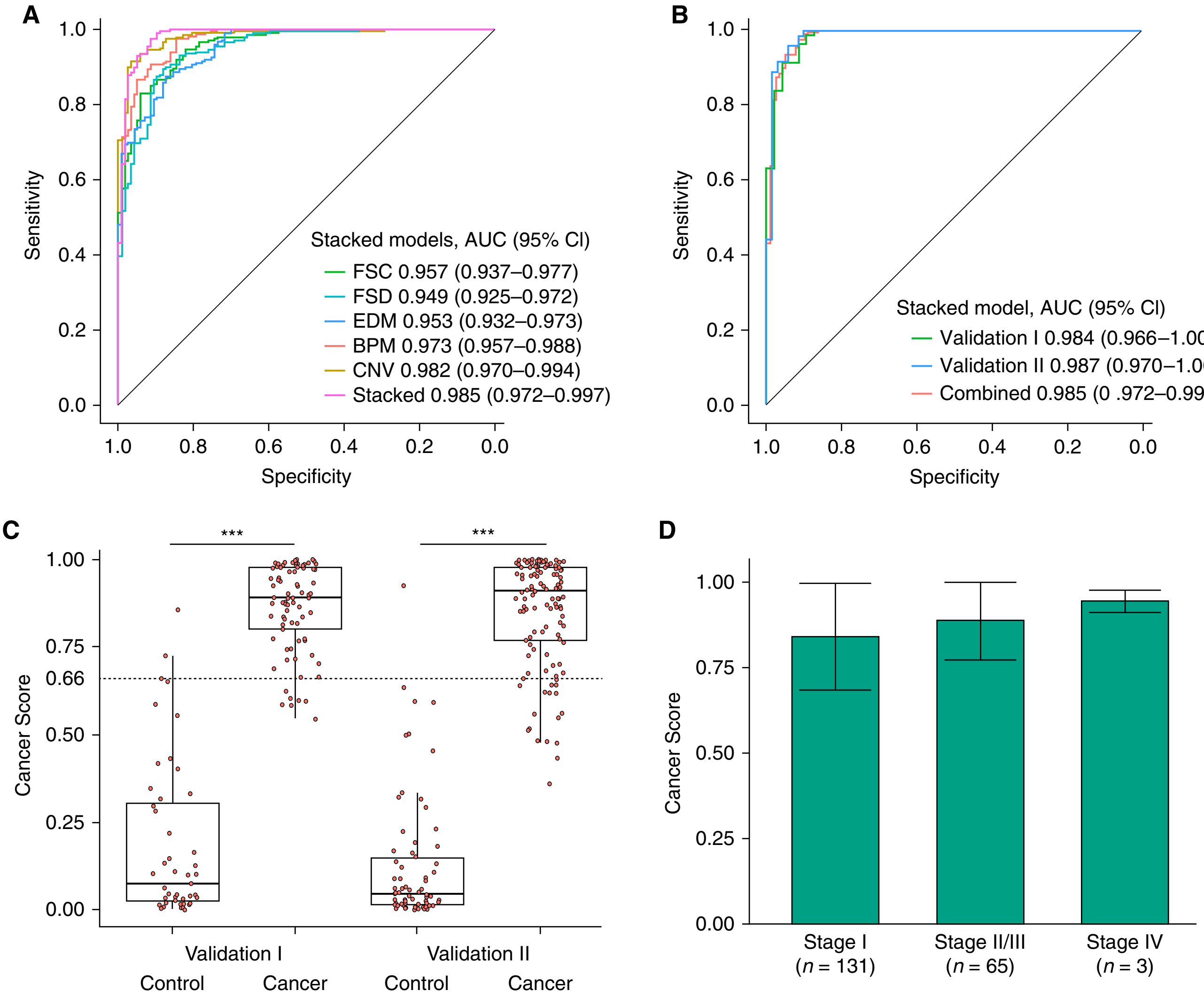

Figure 2. Development and evaluation of the predictive model in the validation cohorts. (*A*) Receiver operating characteristic (ROC) curve evaluating the performance of different stacking models in distinguishing patients with early lung cancer from healthy subjects in the combined validation cohorts. (*B*) ROC curve evaluating the performance of the stacked ensemble model in the combined validation cohort and its validation I and II cohorts separately. (*C*) The boxplots showing the distribution of cancer scores in the patient and control groups of the validation cohorts. The 95% specificity cutoff score for the internal validation I set is 0.66, and a *t* test was performed for the comparison between cancer and control subsets (***$P < 0.001$). (*D*) Distribution of cancer scores from patients grouped by cancer stage in the validation cohorts. The bar plot shows the mean value and SD of each stage group. The case numbers in the groups are indicated. AUC = area under the curve; BPM = breakpoint motif; CI = confidence interval; CNV = copy number variation; EDM = end motif; FSC = fragmentation size coverage; FSD = fragmentation size distribution.

consistently using either raw or 5× sequencing depth WGS data. In the meantime, we assessed its robustness at lower coverage depth by gradually down-sampling. Upon reducing WGS coverages to 4×, 3×, 2×, 1×, and 0.5×, we found their AUC values remained high in both validations I (⩾0.966) and II (⩾0.971) datasets (Table E8). Although a slight decrease in AUC was observed when the coverage depth was down, the 1× WGS data can still yield clinically usable results (validation I: 91.4% sensitivity at 93.6% specificity; validation II: 83.9% sensitivity at 97.1% specificity) (Figure E2).

Figure 1. (*Continued*). 240 participants (cancer 120, healthy 120) were from independent studies and included for further external validation. (*B*) Schematic diagram of the stacked ensemble model construction and cancer probability score determination. Plasma cfDNA was extracted from the participant and subject to WGS. The sequencing reads were mapped to a human reference genome to determine the FSC, FSD, EDM, BPM, and CNV features. The genome-wide feature profiles were then applied to the five machine learning algorithms, with the resultant matrix processed by a second-layer GLM algorithm to form the stacked ensemble model and calculate the participant's cancer probability score. BPM = breakpoint motif; CNV = copy number variation; EDM = end motif; FSC = fragmentation size coverage; FSD = fragmentation size distribution; GBM = gradient boosting machine; GLM = generalized linear model; NSCLC = non–small cell lung cancer.

Table 1. Diagnostic Performance of the Predictive Model in the Validation Cohorts and Additional Cohort

	Actual	
Validation I Cohort	**Cancer**	**Healthy**
Predicted		
Cancer	74	2
Healthy	7	45
Sensitivity (95% CI)	91.4% (83.0–96.5%)	
Specificity (95% CI)	95.7% (85.5–99.5%)	
Accuracy (95% CI)	93.0% (87.1–96.7%)	
	Actual	
Validation II Cohort	**Cancer**	**Healthy**
Predicted		
Cancer	100	1
Healthy	18	69
Sensitivity (95% CI)	84.7% (77.0–90.7%)	
Specificity (95% CI)	98.6% (92.3–100.0%)	
Accuracy (95% CI)	89.9% (84.7–93.8%)	
	Actual	
Combined Validation Cohorts	**Cancer**	**Healthy**
Predicted		
Cancer	174	3
Healthy	25	114
Sensitivity (95% CI)	87.4% (82.0–91.7%)	
Specificity (95% CI)	97.4% (92.7–99.5%)	
Accuracy (95% CI)	91.1% (87.4–94.0%)	
	Actual	
Additional Validation Cohorts	**Cancer**	**Healthy**
Predicted		
Cancer	113	9
Healthy	7	111
Sensitivity (95% CI)	92.5% (86.2–96.5%)	
Specificity (95% CI)	94.2% (88.4–97.6%)	
Accuracy (95% CI)	93.3% (89.4–96.1%)	

Definition of abbreviation: CI = confidence interval.
The sensitivity and accuracy were calculated with the 95% CI at the 95.7% specificity of validation I.

Using within-run and between-run replicates from three patients with cancer of the additional validation cohort (stage I: one; stage II: one; stage III: one) and three healthy control subjects, we evaluated the model's repeatability and reproducibility (Table E9). As shown in Figure 4A, the predicted cancer scores of three replicates for each condition showed 100% accuracy in determining cancer and noncancer samples. Furthermore, we observed high consistency for both within-run tests (the three replicates of each testing condition) and between-run tests (batch 1 and batch 2 of the same participants).

We then set out to test the robustness of the predictive model in cancer samples with low ctDNA fractions using the 0.66 cutoff score. The lowest ctDNA fraction of the analyte and sequencing depth for consistent detection were assessed. The max VAFs of the three selected patients are 1.66%, 0.13%, and 0.05% (Table E10). The plasma cfDNA samples of these three patients were sequenced by WGS and randomly down-sampled for prediction. For patient 1 (VAF 1.66%), the model remained at 100% sensitivity when down-sampled to 0.5× coverage. For patient 2 (VAF, 0.13%) and patient 3 (VAF, 0.05%), the model sensitivity is 100% at 3× coverage or higher. Notably, at the 2× coverage, our model consistently maintained a sensitivity of ⩾95.0%, which is the well-accepted probability by the field for determining the detection limit, on all three patients. Even with the lowest VAF (0.05%) and sequencing coverage (0.5×), our model showed a 75.0% (15/20) sensitivity in identifying cancer (Figure 4B).

Performance of the Predictive Model in Different Cancer Sample Subgroups

We further examined the model performance in different lung cancer subgroups using the validation datasets, whereas the subgroups had no statistically significant difference between different categories (Fisher's exact test, P values are all >0.05). As shown in Table 2 (95.7% specificity using the 0.66 cutoff score), assay detection sensitivity is consistently high across different subgroups of patients with cancer. Specifically, our assay reliably identified both SCC and ADC (sensitivity, 93.3%; 95% CI, 68.1–99.8% and sensitivity, 87.0%; 95% CI, 81.2–91.5%, respectively). The model is suitable for detecting early pathological features (stage I: 83.2%; 95% CI, 75.7–89.2% and <1 cm tumor: 85.0%; 95% CI, 62.1–96.8%). Furthermore, the model showed comparably high sensitivities for identifying subjects with lung cancer regardless of gender, age, tumor location, lymph node metastasis, focal number by computed tomography scan, and risky behavior such as cigarette smoking (Table 2).

Discussion

In this study, we aimed to improve the early detection of lung cancer and established a stacked ensemble machine learning model integrating multiple cfDNA fragmentomic features to achieve high sensitivity in differentiating subjects with early-stage NSCLC and subjects without cancer.

Our multidimensional assay incorporates the advantages of cfDNA fragmentomic features into the stacked ensemble machine learning model. The cleavage and fragmentation process of cfDNA is nonrandom. Certain genomic regions have shown preferred cleavage patterns, termed preferred end sites, which are associated with conditions such as tissue sources and disease status due to chromatin accessibility, nuclease activities, etc. (12, 16, 17, 25). As the tumor-associated preferred end sites are more pervasive and readily detectable than mutations, the fragmentomic features have

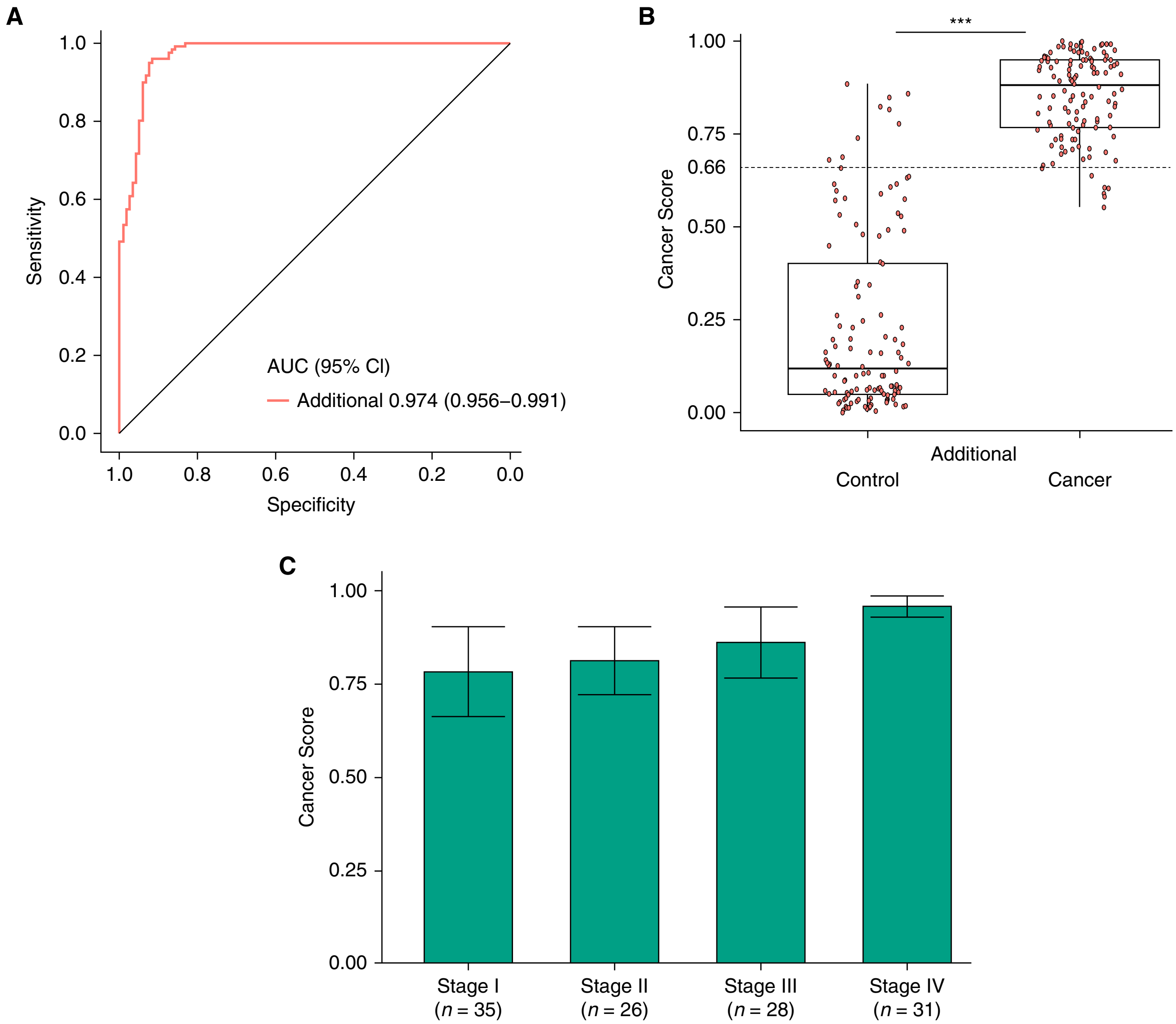

Figure 3. Evaluation of the stacked ensemble predictive model in the independent additional validation cohort. (*A*) Receiver operating characteristic curve evaluating the performance of the predictive model in distinguishing patients with early lung cancer from healthy subjects in the additional validation cohort. (*B*) The boxplots showing the distribution of cancer scores in the patient and control groups of the additional validation cohort. The 95% specificity cutoff score based on the validation I set is 0.66, and a *t* test was performed for the comparison between cancer and control subsets (*$0.01 < P < 0.05$, **$0.001 < P < 0.01$, and ***$P < 0.001$). (*C*) Distribution of cancer scores from patients grouped by cancer stage in the additional validation cohort. The bar plot shows the mean value and standard deviation of each stage group. The case numbers in the groups are indicated. AUC = area under the curve; CI = confidence interval.

become an emerging class of ctDNA signatures (12). The superior performance of our model also underscored the promising role of cfDNA fragmentomics in cancer detection. As mentioned earlier, cfDNA methylation and fragmentation patterns have also been used for identifying patients with lung cancer, but their sensitivities for stage I lung cancer ranging from ∼25% to ∼60% are unsatisfying for clinical use (12–15). By combining pathological features, computed tomography imaging, and serum protein biomarker, the new DELFI model has improved its detection performance (15). However, it still experienced a detection bias against the patients with early-stage disease (sensitivity, 91% of stage I/II vs. 96% of stage III/IV, at 80% specificity), likely due to the high proportion of later-stage samples used for its model construction. We have comprehensively evaluated the existing cfDNA features and integrated the better ones into our predictive model. The ensemble machine learning models have shown advantages in model accuracy, reproducibility, and interpretability for bioinformatics applications, and the stacking strategy can optimally integrate the predictions made by single base models (26). This study confirmed the superior prediction power of the stacked models to the single base models. Compared with the existing DELFI model, our model achieved higher sensitivities of 83.2% for patients with stage I disease, respectively, at the specificity of 95.7%. More importantly, our strategy ensured that all the next-generation sequencing features are accessible from one WGS testing and eased patients' burden of

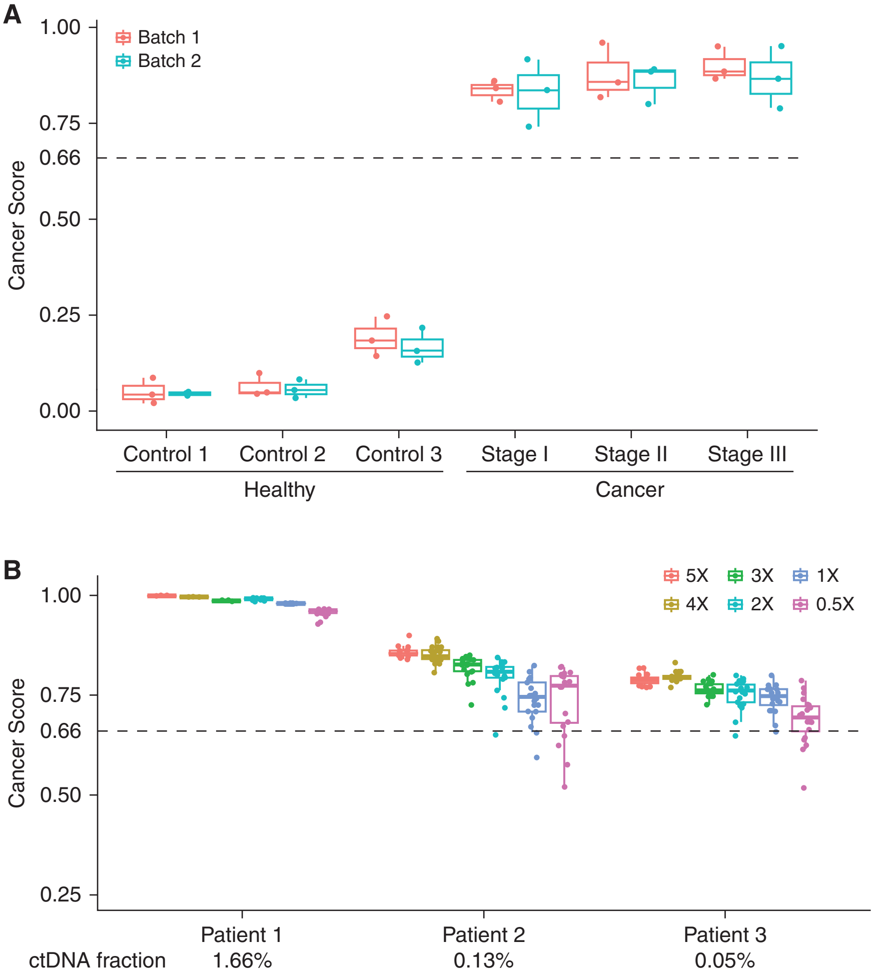

Figure 4. Analytical validity assessments of the stacked ensemble predictive model. (*A*) Boxplots showing within-run and between-run tests to evaluate the repeatability and reproducibility of the predictive model. Three healthy control subjects and three patients with cancer (stage I, II, and III, respectively) were processed with three replicates for each condition and had blood draws twice, with a 3-week interval between batch 1 and batch 2. The 95% specificity cutoff score based on the validation I set is 0.66. (*B*) Three cancer samples with low variant allele frequency (VAF) were subject to down-sampling *in silico*. The patients' identification and their VAF values are indicated. Each sample was down-sampled from 5× to 0.5× with 20 repeats. The 95% specificity cutoff score based on the validation I set is 0.66. ctDNA = circulating tumor DNA.

taking multiple tests. In the future, we plan to include additional clinical information to improve the model performance (27).

Our model can facilitate the development of lung cancer early detection in clinical practice. First, its performance is robust using shallow WGS data. Even when the coverage depth was down to 0.5×, the model retained 91.4% sensitivity at 93.6% specificity in the validation I cohort and 78.0% sensitivity at 94.3% specificity in the validation II cohort. The assay robustness with low coverage depth can help reduce the sequencing cost. Consistent with the better detection performance in the later stages of the existing fragmentomic assays (15, 28), the predicted scores of patients with cancer exhibited an upward trend from stage I to IV in our model (Figures 2D and 3C). More importantly, the stacked ensemble model has outperformed the previous models in distinguishing subjects with and without cancer and thus achieved a stably high detection ability across different disease stages (Table 2). Indeed, the sensitivity of our model for very early–stage (stage I) and small-size (<1 cm) tumors reached 83.2% and 85.0%, respectively, indicating its superior detection ability for early-stage characteristics. Although early diagnosis is a key to less morbid treatment and favorable prognosis, most patients with lung cancer are still receiving diagnoses at the metastatic stage (29, 30). Therefore, our strategy could be particularly useful in promoting lung cancer early detection.

This study has several limitations. This model was built on patients with early-stage NSCLC in the Chinese population. NSCLC accounts for about 85% of all lung cancer (31), and this study did not include small-cell lung carcinoma samples. In this study, we have demonstrated the application of cfDNA fragmentomics for high-sensitivity lung cancer detection, but the underlying mechanism is still undefined in the field. We are currently pursuing a comprehensive investigation to select variables and explore new features to leverage our model. Also, the study size is relatively small, so the sensitivity of several small subgroups, such as patients

Table 2. Diagnostic Sensitivities of the Predictive Model in Different Lung Cancer Patient Subgroups of the Validation I and II Cohorts and Their Combination

	Validation I		Validation II		Combined Validation	
Cohort	**Sensitivity (95% CI)**	**TP/Total**	**Sensitivity (95% CI)**	**TP/Total**	**Sensitivity (95% CI)**	**TP/Total**
Histology						
ADC	90.9% (81.3–96.6%)	60/66	84.7% (77.0–90.7%)	100/118	87.0% (81.2–91.5%)	160/184
SCC	93.3% (68.1–99.8%)	14/15	N/A	N/A	93.3% (68.1–99.8%)	14/15
Stage						
I	91.3% (79.2–97.6%)	42/46	78.8% (68.6–86.9%)	67/85	83.2% (75.7–89.2%)	109/131
II/III	90.6% (75.0–98.0%)	29/32	100.0% (89.4–100.0%)	33/33	95.4% (87.1–99.0%)	62/65
IV	100.0% (29.2–100.0%)	3/3	N/A	N/A	100.0% (29.2–100.0%)	3/3
Tumor size						
<1 cm	81.2% (54.4–96.0%)	13/16	100.0% (39.8–100.0%)	4/4	85.0% (62.1–96.8%)	17/20
⩾1 cm	93.9% (85.0–98.3%)	61/65	84.2% (76.2–90.4%)	96/114	87.7% (82.0–92.1%)	157/179
Differentiation level						
Low	83.3% (35.9–99.6%)	5/6	77.8% (40.0–97.2%)	7/9	80.0% (51.9–95.7%)	12/15
Low–medium	95.7% (78.1–99.9%)	22/23	86.0% (73.3–94.2%)	43/50	89.0% (79.5–95.1%)	65/73
Medium	91.3% (72.0–98.9%)	21/23	96.4% (81.7–99.9%)	27/28	94.1% (83.8–98.8%)	48/51
Medium–high	87.5% (67.6–97.3%)	21/24	64.7% (38.3–85.8%)	11/17	78.0% (62.4–89.4%)	32/41
High	100.0% (29.2–100.0%)	3/3	50.0% (1.3–98.7%)	1/2	80.0% (28.4–99.5%)	4/5
Focality						
Unifocal	92.4% (83.2–97.5%)	61/66	85.7% (77.8–91.6%)	96/112	88.2% (82.5–92.5%)	157/178
Multifocal	86.7% (59.5–98.3%)	13/15	66.7% (22.3–95.7%)	4/6	81.0% (58.1–94.6%)	17/21
Sex						
Female	94.4% (81.3–99.3%)	34/36	76.7% (64.0–86.6%)	46/60	83.3% (74.4–90.2%)	80/96
Male	88.9% (75.9–96.3%)	40/45	93.1% (83.3–98.1%)	54/58	91.3% (84.1–95.9%)	94/103
Age						
⩽65	92.5% (81.8–97.9%)	49/53	94.0% (85.4–98.3%)	58/67	86.6% (76.0–93.7%)	107/120
>65	89.3% (71.8–97.7%)	25/28	96.1% (86.5–99.5%)	42/51	82.4% (69.1–91.6%)	67/79
Location						
Left	84.6% (65.1–95.6%)	22/26	80.4% (66.9–90.2%)	41/51	81.8% (71.4–89.7%)	63/77
Right	94.5% (84.9–98.9%)	52/55	88.1% (77.8–94.7%)	59/67	91.0% (84.4–95.4%)	111/122
Lymph node metastasis						
Yes	90.5% (69.6–98.8%)	19/21	100.0% (82.4–100.0%)	19/19	95.0% (83.1–99.4%)	38/40
No	91.7% (81.6–97.2%)	55/60	81.8% (72.8–88.9%)	81/99	85.5% (79.1–90.6%)	136/159
Smoking						
Yes	91.7% (61.5–99.8%)	11/12	91.7% (73.0–99.0%)	22/24	91.7% (77.5–98.3%)	33/36
No	91.3% (82.0–96.7%)	63/69	83.0% (73.8–89.9%)	78/94	86.5% (80.3–91.3%)	141/163

Definition of abbreviations: ADC = adenocarcinoma; CI = confidence interval; SCC = squamous cell carcinoma; TP = true positive.
The sensitivities (%) were calculated with the 95% confidence interval at the 95.7% specificity of validation I. The numbers of TP predictions and total cases are indicated for each subgroup in the particular cohorts.

with SCC, and low- and high-differentiation tumors, may be underestimated or overestimated. We are expanding the size and diversity of our study to achieve more consolidated conclusions. In addition, we sought to identify the limit of our model for cancer detection by directly using real patient blood samples with low ctDNA fractions. Such analysis confirmed the detection of samples with VAF 0.05%, representing the lowest limit of reliable ctDNA fraction detection by our current technology. We acknowledge that using standard analytical validity would help accurately determine the assay limit of detection. However, the existing commercial cfDNA reference standards are artificially prepared, genomic variant–based references. There are also references derived from mixed human cell lines, which are artificially fragmented to resemble human plasma cfDNA but result in fragment size and end motifs different from the real human plasma cfDNA. Thus, these standards are unsuitable for representing the *bona fide* human blood cfDNA and evaluating the stacked ensemble multidimensional fragmentomic model.

Taken together, we have established a stacked ensemble model integrating five fragmentomic features from plasma cfDNA WGS data. Our model exhibited high sensitivity in distinguishing patients with early-stage NSCLC from control subjects without cancer. Furthermore, we demonstrated the consistency and robustness of the assay by testing its performance across different WGS coverage depths. Together with its superior detection power for very early–stage, small-size tumors in patients with NSCLC, our multidimensional model has provided an accurate and affordable approach to promoting early detection of NSCLC and improving the outcomes of patients. ■

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

Acknowledgment: The authors thank all the patients, volunteers, and their families; the investigators; and the site personnel who participated in this study.

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA Cancer J Clin* 2021;71:7–33.
2. Avanzini S, Kurtz DM, Chabon JJ, Moding EJ, Hori SS, Gambhir SS, *et al.* A mathematical model of ctDNA shedding predicts tumor detection size. *Sci Adv* 2020;6:eabc4308.
3. American Cancer Society. Can lung cancer be found early? 2021 [accessed 2021 Mar 9]. Available from: https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/detection.html.
4. Church TR, Black WC, Aberle DR, Berg CD, Clingan KL, Duan F, *et al.*; National Lung Screening Trial Research Team. Results of initial low-dose computed tomographic screening for lung cancer. *N Engl J Med* 2013;368:1980–1991.
5. 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.
6. Berrington de González A, Kim KP, Berg CD. Low-dose lung computed tomography screening before age 55: estimates of the mortality reduction required to outweigh the radiation-induced cancer risk. *J Med Screen* 2008;15:153–158.
7. Stroun M, Maurice P, Vasioukhin V, Lyautey J, Lederrey C, Lefort F, *et al.* The origin and mechanism of circulating DNA. *Ann N Y Acad Sci* 2000; 906:161–168.
8. Sun K, Jiang P, Chan KC, Wong J, Cheng YK, Liang RH, *et al.* Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc Natl Acad Sci USA* 2015;112:E5503–E5512.
9. Fece de la Cruz F, Corcoran RB. Methylation in cell-free DNA for early cancer detection. *Ann Oncol* 2018;29:1351–1353.
10. Benesova L, Belsanova B, Suchanek S, Kopeckova M, Minarikova P, Lipska L, *et al.* Mutation-based detection and monitoring of cell-free tumor DNA in peripheral blood of cancer patients. *Anal Biochem* 2013; 433:227–234.
11. Chabon JJ, Hamilton EG, Kurtz DM, Esfahani MS, Moding EJ, Stehr H, *et al.* Integrating genomic features for non-invasive early lung cancer detection. *Nature* 2020;580:245–251.
12. Lo YMD, Han DSC, Jiang P, Chiu RWK. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. *Science* 2021;372:eaaw3616.
13. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* 2019;570:385–389.
14. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV; CCGA Consortium. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol* 2020;31:745–759.
15. Mathios D, Johansen JS, Cristiano S, Medina JE, Phallen J, Larsen KR, *et al.* Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat Commun* 2021;12:5060.
16. Jiang P, Sun K, Tong YK, Cheng SH, Cheng THT, Heung MMS, *et al.* Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. *Proc Natl Acad Sci USA* 2018;115:E10925–E10933.
17. Jiang P, Sun K, Peng W, Cheng SH, Ni M, Yeung PC, *et al.* Plasma DNA end-motif profiling as a fragmentomic marker in cancer, pregnancy, and transplantation. *Cancer Discov* 2020;10:664–673.
18. Gould MK, Huang BZ, Tammemagi MC, Kinar Y, Shiff R. Machine learning for early lung cancer identification using routine clinical and laboratory data. *Am J Respir Crit Care Med* 2021;204:445–453.
19. Chen L, Abou-Alfa GK, Zheng B, Liu JF, Bai J, Du LT, *et al.*; PreCar Team. Genome-scale profiling of circulating cell-free DNA signatures for early detection of hepatocellular carcinoma in cirrhotic patients. *Cell Res* 2021;31:589–592.
20. Mouliere F, Smith CG, Heider K, Su J, van der Pol Y, Thompson M, *et al.* Fragmentation patterns and personalized sequencing of cell-free DNA in urine and plasma of glioma patients. *EMBO Mol Med* 2021;13: e12881.
21. Ma X, Chen Y, Tang W, Bao H, Mo S, Liu R, *et al.* Multi-dimensional fragmentomic assay for ultrasensitive early detection of colorectal advanced adenoma and adenocarcinoma. *J Hematol Oncol* 2021;14: 175.
22. Wan N, Weinberg D, Liu TY, Niehaus K, Ariazi EA, Delubac D, *et al.* Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. *BMC Cancer* 2019;19:832.
23. Zhang C, Ma Y. *Ensemble machine learning: methods and applications.* Boston, MA: Springer; 2012.
24. Nakamura Y, Okamoto W, Kato T, Esaki T, Kato K, Komatsu Y, *et al.* Circulating tumor DNA-guided treatment with pertuzumab plus trastuzumab for HER2-amplified metastatic colorectal cancer: a phase 2 trial. *Nat Med* 2021;27:1899–1903.
25. Chan KC, Jiang P, Sun K, Cheng YK, Tong YK, Cheng SH, *et al.* Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. *Proc Natl Acad Sci USA* 2016;113:E8159–E8168.
26. Cao Y, Geddes TA, Yang JYH, Yang PY. Ensemble deep learning in bioinformatics. *Nat Mach Intell* 2020;2:500–508.
27. Liu QX, Zhou D, Han TC, Lu X, Hou B, Li MY, *et al.* A noninvasive multianalytical approach for lung cancer diagnosis of patients with pulmonary nodules. *Adv Sci (Weinh)* 2021;8:2100104.
28. Klein EA, Richards D, Cohn A, Tummala M, Lapham R, Cosgrove D, *et al.* Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann Oncol* 2021;32:1167–1177.
29. American Cancer Society. Lung cancer survival rates; 2021 [last revised 2021 Jan 29; accessed 2021 Mar 9]. Available from: https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/survival-rates.html.
30. Blandin Knight S, Crosbie PA, Balata H, Chudziak J, Hussell T, Dive C. Progress and prospects of early detection in lung cancer. *Open Biol* 2017;7:170070.
31. Zappa C, Mousa SA. Non-small cell lung cancer: current treatment and future advances. *Transl Lung Cancer Res* 2016;5:288–300.