Cell-based assay for blood-based diagnostics enhances prediction of lung cancer in patients with lung nodules

Authors

Jason D. Berndt[1,*], Fergal J. Duffy[2,*], Mark D. D'Ascenzo[1,*], Leslie R. Miller[2], Yijun Qi[1], G. Adam Whitney[1], Samuel A. Danziger[2,3], Anil Vachani[4], Pierre P. Massion[5], Stephen A. Deppen[6], Robert J. Lipshutz[1], John D. Aitchison[2,3,7], Jennifer J. Smith[1,**]

*Authors contributed equally to this work; **Author to which correspondence should be addressed

[1] PreCyte, Inc., Seattle, WA, USA, [2] Seattle Children's Research Institute, Seattle, WA, USA, [3]Institute for Systems Biology, Seattle, WA, USA, [4] Pulmonary, Allergy, and Critical Care Division, University of Pennsylvania, Philadelphia, PA, [5] Thoracic Program, Vanderbilt-Ingram Cancer Center, Vanderbilt University, Nashville, Tennessee, [6]Department of Surgery, Tennessee Valley Healthcare System, Veterans Affairs, Nashville, Tennessee; Department of Thoracic Surgery, Vanderbilt University Medical Center, Nashville, Tennessee, [7]Department of Global Health, University of Washington, Seattle, WA.

Funding and Acknowledgements

Research reported in this publication was supported by the National Cancer Institute (NCI) and National Institute of Aging (NIA) of the National Institutes of Health (NIH) under Award Numbers R43CA203455, R44CA203455 and R44AG051282 to PreCyte, the National Institute of Environmental Health Sciences (NIEHS) of the NIH under Award Number P30ES013508 to the University of Pennsylvania, and the NCI of the NIH under award number P30CA068485 and NCI-U01CA152662 to Vanderbilt University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

In memory of Pierre Massion

Abbreviations: AUC, area under the ROC curve; CT, computed tomography; DEG, differentially expressed gene; DMOG, dimethyloxalylglycine; GSEA, gene set enrichment analysis; iCAP, indicator cell assay platform; NPV, negative predictive value; ROC, receiver operator characteristic; MCED, multi-cancer early detection

Abstract 150 words

The iCAP is a tool for blood-based diagnostics that addresses the low signal-to-noise ratio of blood biomarkers by using cells as biosensors. The assay exposes small volumes of patient serum to standardized cells in culture and classifies disease by AI analysis of gene-expression readouts from the cells. It simplifies the complexity of blood into a concise readout in a scalable cell-based assay. We developed the LC-iCAP as a rule-out test for nodule management in CT-based lung cancer screening. The assay achieved an AUC of 0.63 (95% CI 0.50-0.75) in retrospective-blind-temporal validation. When integrated with CT data after validation, it demonstrated potential to reduce unnecessary follow-up procedures by significantly outperforming the Mayo Clinic model with 90% sensitivity, 67% specificity and 95% NPV using an estimated 25% prevalence. Analytical validation established LC-iCAP reproducibility and identified unwanted variation from long-term serum storage suggesting a prospective study design could enhance performance.

Key words:

Lung cancer, AI-based diagnostics, CT screening, blood biomarkers, LC-iCAP, hypoxia response, HIF1A, gene expression, cell-based assay,

Author contribution

JDA, JJS, and RJL and conceived the study, JDB and JJS designed the experiments. AV, PM, and SD selected and provided samples and insights related to clinical need. JDB, LRM and YQ conducted the experimental work and collected the data. Data analysis was performed by FJD, GAW, JDB, JJS, MDD, SAD, SD and YQ. The manuscript was written by JJS. All authors reviewed and approved the final manuscript. JDA, JDB, JJS and RJL supervised the project and provided critical insights throughout the study.

Conflict of interest statement

Competing interests: RJL, JJS, JDB, GAW, MDD, and YQ are current and/or past employees of PreCyte and have equity interests in the company. JDA and RJL serve as board members with equity interests in PreCyte. LRM and SAD also have equity interests in PreCyte. FJD acts as a consultant and/or has performed contracted work for PreCyte. JJS, RJL, JDA, JDB, MDD, GAW, and FJD have filed patent applications related to the compositions, methods, kits, and processes described in this manuscript in the US and in foreign jurisdictions. SD and AV declare no conflicts of interest.

Introduction

Lung cancer is the deadliest cancer in US and worldwide, but early detection can save lives[1]. Therefore, the US has implemented low-dose computed tomography (CT) scans to screen for non-small cell lung cancer (NSCLC) in high-risk populations, a strategy estimated to reduce lung cancer deaths by 20%[2,3]. Nodules are managed based on a calculated malignancy risk from clinical and CT scan data using algorithms such as the Mayo Clinic model[4]. Based on the American College of Chest Physicians (ACCP) recommendations, those with less than 5% risk or greater than 65% risk have clear treatment paths. However, 50-76% of patients screened are assigned an intermediate risk and undergo a diagnostic odyssey often with invasive and costly procedures even though most of these patients do not have cancer[5,6]. Non-invasive tests are needed to more accurately predict malignancy risk in patients with indeterminate pulmonary nodules (IPNs)[3,7]. Specifically, rule-out type tests are needed to identify those with low risk of cancer to save those with benign nodules from invasive and expensive testing. Such a test must have an NPV of at least 95% to reclassify nodules with a 5-65% risk of malignancy to less than 5% to facilitate a meaningful shift in patient management strategies reducing unnecessary interventions and improving clinical outcomes.[8]

There are a few liquid biopsy tests available to patients for nodule management. The data obtained from these liquid biopsies are independent of CT scan data, enabling combinatorial diagnostics. This approach involves combining data from two sufficiently distinct and accurate tests to achieve better performance than either test alone[9]. Patients have access to two rule-out tests from Biodesix[10] and MagArray[11]. After integration with CT data, the test sensitivities are high, ranging from 90-97%, but the specificities are low ranging from 33-44%, which limits the number of patients with benign nodules who have actionable results from the tests. This limitation may be due to the fact that neither test showed significant stand-alone performance in validation studies[10,11], which is not ideal for combinatorial analysis[9,12]. Another test from Veracyte outperforms these blood tests but requires invasive bronchoscopy[13]. A rule-out type blood test that has significant performance independent of CT data and outperforms existing blood tests when combined with CT data would be a significant advance in nodule management. Such a test could raise confidence for both patients and physicians by providing actionable results for more patients and reducing the number of unnecessary invasive follow-up tests.

We are developing a tool called the indicator cell assay platform (iCAP)[14], a novel approach that aims to overcome the low signal-to-noise ratio associated with direct measurement of blood biomarkers by using cultured cells as biosensors. Developing an iCAP involves exposing standardized, cultured cells to a small volume of serum or plasma samples from case and control participants, measuring a global differential gene expression response of the cells, and using machine learning to identify a subset of features for disease classification. Deploying the assay involves measuring only selected genes using a targeted approach like NanoString® or Quantigene® (Fig. 1*A*). The rationale is that through signal transduction, cells can amplify weak signals into strong readouts, enhancing the sensitivity of detection. In addition, cells naturally detect and respond to many types of analytes or combinations thereof, broadening the search space. These blood-based signatures are transformed into a cell-based gene expression readout, measured by well-established next-

generation RNA sequencing (RNA-seq) or targeted transcriptomic approaches. This process effectively condenses vast complexity of blood into a concise readout in a scalable, multi-analyte cell-based assay.

Here, we present development of an iCAP for the early detection of lung cancer (LC-iCAP) with utility for management of nodules identified by CT scan. We developed a model using banked serum from patients with IPNs to distinguish NSCLC from benign nodules and tested it by blind temporal validation. When integrated with CT data, the LC-iCAP demonstrated 90% sensitivity and 67% specificity in blind validation using a cut-off point corresponding to a 95% negative predictive value (NPV) based on a prevalence of 25% in the intended use population.[15,16] The integrated test had significantly better performance than the Mayo Clinic model and the specificity was better than that of the other rule-out blood tests, suggesting actionable results for a greater number of patients with benign nodules using the LC-iCAP. The iCAP has high-throughput scalability and is orthogonal to other diagnostic approaches, suggesting potential for broader applications in multi-cancer early detection (MCED) and combinatorial diagnostics.

Methods

The study design description below follows recommendations of TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis).[17]

Participants and specimen characteristics

The study used archived serum samples collected from adult patients from non-vulnerable populations performed with IRB approval (WCG IRB study 1283522). Banked specimens and clinical data used in this study were from subjects enrolled in the following previously IRB-approved studies: "Molecular Predictors of Lung Cancer behavior," (NCT00898313, Vanderbilt University), "Gene-Environment Interactions in Lung Cancer" (IRB 806390, University of Pennsylvania), and "A Case Control Study of Smokers and Non-Smokers" (IRB 800924, University of Pennsylvania). The study consent forms had provisions allowing use of their samples for future research purposes. Patient identifiable information was not provided to the research team and was not used in this study. All identification numbers used in the manuscript are not known to anyone outside the research group.

Patient attributes and serum characteristics are shown in Figure 2A and Data Files 1 and 4. All samples were collected within 3 months of the CT scan (and diagnostic biopsy, if performed) and prior to any invasive procedure, including surgical lung biopsy. All malignant nodules were diagnosed by follow-up pathological diagnosis; all were non-small cell lung cancer (NSCLC) except for 3 small cell lung cancers. 72% of malignant nodules in the blind test set were stage I and 22% were stage II. Benign nodules were diagnosed by either biopsy with a definitive benign histological diagnosis or by 2 or more years of follow-up with serial imaging. All patients had no known other cancers at time of screening and no previous cancer in the five years preceding the blood draw excluding previous skin cancers treated with surgery only (no radiation or chemotherapy). Serum samples were collected using the protocol recommended by the early detection research network (EDRN)[18], stored at -80°C or below after collection, and unless otherwise stated, thawed once prior to LC-iCAP analysis.

Study design

The study was an observational case versus control study with blind validation following prospective-specimen-collection, retrospective-blinded-evaluation (PRoBE) design.[19] Serum samples were collected at the time of CT scan from cohorts that represent the target population with the future intention of developing liquid biopsies. After outcome status was ascertained by follow-up testing, case and control subjects were selected randomly from the cohorts and their serum was used for developing the LC-iCAP and for blind validation of the assay.

The study involved assay and model development as described in Figure 1*B*. Assay development including establishing standard controls, optimizing experimental parameters and measuring assay reproducibility (Stage 2). Model development included iterative disease modeling with increasing number of samples in Stages 1,3 and 4 (Fig. 2*B*). This process involved optimizing the parameters and hyperparameters of the models through several rounds of training and validation on held-out sets and testing fully parameterized models on a blind test set. During assay development, sample and data quality were evaluated and in the final model configuration, data from samples with technical assay failures and low-quality samples were excluded.

The study used a total 528 serum samples from two sources, Vanderbilt University and University of Pennsylvania in 5 cohorts (Fig. 2*A*). Cohorts 1-4 were used for model development and cohort 5 was used for blind testing of the final model. The blind test set was collected 9 years later and analyzed greater than 1 year later than the other cohorts. For model development, the sample size was selected based on that of other liquid biopsy studies at similar stages of development[11,20]. For blind testing, the sample size was selected to have power to detect significant performance of a model with AUC of ROC >= 0.62[21].

Sample processing batch structure is described in Data files 1 and 4. Each cohort was shipped separately and assayed by LC-iCAP in batches of 20-50 samples. Each experimental batch had ~1:1 ratio of case and control samples, which were roughly balanced for patient gender, age and smoking history. A pair of standard controls were included on each 12-well plate for quality control (QC) consisting of either technical replicates of a reference serum sample from an unaffected male patient (cohorts 1-2), a pair of case and control pooled serum controls (cohort 3), or a pair of DMOG and PBS chemical controls (cohorts 4-5). LC-iCAP gene expression was measured by RNA-seq for cohorts 1-3 for modeling in Stages 1 and 3, and by NanoString Plexset for all cohorts for modeling in Stage 4. For cohort 5, only the DMOG samples were included in the NanoString Plexset analysis, excluding the PBS controls, to fit all samples on a single plate.

Quantification of hemolysis of serum samples

Prior to LC-iCAP analysis, thawed patient serum was evaluated for the breakdown of red blood cells (hemolysis), which has known interference with clinical biochemical tests[22]. Blind samples were visually compared to a reference card[23] by a first scientist and given a score on a gradient of increasing hemolysis from 1-7. In addition, samples were photographed with an iPhone on a white background and later evaluated by a second scientist using the same approach. Sample ratings were averaged and round to the nearest integer. For project stages 3-4, all samples with average scores greater than 3 (> 50 mg/dL of hemoglobin) were omitted from the study. Hemolysis scores are shown in Data files 1 and 4.

Analytical parameters of LC-iCAP assay

Unless otherwise indicated, the following standard protocol was used: 2 x 10^6 lung epithelial cell indicator cells (16HBE) were thawed and plated in a T75 flask in RPMI with 10% FBS (complete medium). After 2 d cells were dissociated with 0.25% trypsin-EDTA for 10 min at 37 °C and plated at 30,000 cells/cm^2 in 12-well Eppendorf moat plates in complete medium. After 24 h, cells were rinsed once with RPMI and incubated for 24 h with in 1 mL RPMI with 5% patient serum. Media was removed, lysis buffer was added, and cells were stored at -80°C for up to two weeks before RNA isolation. For cohorts 1 and 2 used in Stages 1-3, total RNA was isolated manually using RNeasy Mini Kit (Qiagen), for cohorts 3-5, RNA isolation was automated using a MagMax mirVana kit (Invitrogen A27828) on either a KingFisher Flex or a Kingfisher Duo Prime as per the manufacturer's recommendation. RNA was eluted in 100 µL of elution buffer, quantified using a Qubit (Qubit RNA BR Assay Kit), and stored at -80°C. Transcript abundance levels were quantified using either RNA-Seq or NanoString as described below. For cohort 3, cells were passaged an additional time before initiating the experiment.

LC-iCAP proof-of-concept studies in Stage 1

To establish assay feasibility, LC-iCAP RNA-seq data for cohort 1 was used to train a model for lung cancer detection and tested on cohort 2. To identify model features, raw counts from cohort 1 data were used for differential expression analysis and identified 239 differentially expressed genes (DEGs) (Benjamini-Hochberg false discovery rate (FDR) < 0.05). Next, to develop the LC-iCAP classifier, all RNA-seq count data from cohorts 1 and 2 were normalized using the DESeq2 rlog transformation and a series of random forest classifiers (R: randomForest package) were parameterized on cohort 1 using increasing numbers of DEGs in increasing order of FDR as features (5, 10, 20, 25, 50, 75, 100). The 103 samples of cohort 2 were used to test the model performance.

For hierarchical clustering of LC-iCAP data from cohorts 1 and 2, clustering was performed based on the expression of the top 20 differentially expressed in LC-iCAP RNA-seq data from cohort 1. The analysis utilized DEseq2 rlog-transformed counts from the LC-iCAP RNA-seq data, normalized to the mean expression of the benign samples of the same iCAP experimental batch and gene rank was based on median absolute deviation. Three outlier samples were identified and removed (all from the benign class).

RNA-Seq analysis of LC-iCAP RNA

100-650 ng of total RNA per sample was used for automated library preparation and RNA sequencing (RNA-seq) performed by either Covance (cohorts 1-2) or Azenta Life Sciences (formerly Genewiz) (cohort 3). Strand-specific library prep was performed with PolyA selection using TruSeq RNA library Prep Kit (Illumina) with unique dual indices (IDT) and resulting DNA was sequenced on a HiSeq 4000 (Illumina) with paired end 150 bp reads. RNA-seq data were processed using a custom workflow including adapter read trimming using *trimmomatic*[24], genome reference alignment to HG37 (pilot study) or HG38 using *STAR*[25]; and gene-level transcript quantification using R-featureCounts[26]. ERCC spike-ins and genes with mean absolute counts < 10 were removed. Quality was assessed using MultiQC, dupRadar and GATK estimate of library complexity. Where indicated, counts were adjusted to correct GC bias using the FQN[27] or CQN[28] determined by the R-EdgeR package[29]. Differential expression analysis was done using R-DESeq2[30]. GSEA analysis was done using the R-fgsea package in combination with the 50 Hallmark pathway modules from MsigDB[31].

For Stage 1, read duplicates were removed by using –ignoreDups setting in R:featureCounts. For Stages 2-3, the data were normalized for heteroskedasticity by variance stabilizing transformation (VST) using the R-DESeq2 package and for inter-iCAP batch variation using removeBatchEffect from the R-limma package[32]. Outlier samples were identified using robust principal components analysis (ROBPCA) implemented in the R-rrcov package[33].

Development of LC-iCAP standard controls

Pooled serum standards were prepared, consisting of technical replicates of pooled patient serum for each case and control class. Each pool consisted of a mix of serum from 8 different subjects selected from cohorts 1 and 2 based on availability and class separation in hierarchical clustering analysis of LC-iCAP-RNA-seq data (Fig. 3*B*). Each serum pool was made by thawing aliquots of serum from each of 8 subjects, pooling, mixing, aliquoting the serum, and then flash freezing in liquid nitrogen before storing at -80°C.

Dimethyloxalylglycine (DMOG) and PBS were used as a pair of chemical standards to monitor assay performance. To develop the control, we first characterized the response of indicator cells to 6 concentrations of DMOG (0, 0.025, 0.05, 0.1, 0.25, 0.5 mM; Cayman Chemical) by measuring the gene expression readout of the 74-gene development gene set (Data file 2A) using NanoString. Responsive genes were identified as those whose expression fit a linear model as a function of DMOG concentration (p-value < 0.05) (data not shown). DMOG was used at 0.25 mM for monitoring assay performance; for most responsive genes, this condition was in the linear range of the model and had a similar magnitude of responsiveness compared with patient serum. DMOG was resuspended at 50 mM in PBS, aliquoted and stored at -20°C and thawed on ice before use.

Optimization of LC-iCAP experimental parameters and LC-iCAP reproducibility testing

The pooled serum standards described above were used for QC and for reproducibility and optimization studies. 4 technical replicates of each case and control serum pool were analyzed for each LC-iCAP parameter and the number of significantly DEGs (FDR <0.1) for each configuration were compared. Differential expression was measured by analysis of the LC-iCAP development gene panel using NanoString nCounter® technology and/or by RNA-seq using HiSeq4000™ (Illumina). Configurations tested were various serum concentrations (1%, 5%, 10%, and 20%), serum incubation times (6 hours, 24 hours), and 4 cell types (16HBE, A549, MRC5, and Nuli-1). Reproducibility testing included comparison across 3 different 16HBE expansion batches and across different days, and different detection platforms.

Western blot analysis

Samples were processed in the LC-iCAP using standard parameters and protein was isolated and quantified using a BCA kit (Pierce). Equal protein for each sample were loaded on 12- or 15-well NuPage 4-12% Bis-Tris gels with Thermo PageRuler Plus prestained protein ladder and analyzed by western blotting according to LICOR Odyssey recommendations using PVDF membrane (Immobilon-FL) and probing with rabbit anti-HIF1α (D157W) XP® (or rabbit mAB HIF-2α (D9E3)) and mouse anti-beta-actin primary antibodies followed by LICOR NIR secondary antibodies. All primary antibodies were from Cell Signaling Technologies. Membranes were scanned on a LICOR Clx Odyssey imaging system, proteins were quantified using LiCOR Image Studio Lite and data analysis was done in Excel. Western blot images are shown in Figure S5.

Sample and Gene Filters for Comprehensive Modeling Study in Stage 3

Data from 165 samples across cohorts 1-3 were partitioned into training and validation sets, and quality filters were applied, resulting in 137 samples for modeling (Fig. 2B, Data file 1). These data were used for a comprehensive modeling study including 13 different feature selection methods x 8 combinations of 3 optional filters. The three sample filters were: 1) samples with predicted forced expiry volume (FEV) < 50% (based on data showing that low lung function affects LC-iCAP readout (Fig. S6)), 2) samples from never-smokers, and 3) samples from a low-quality RNA-seq batch (identified by QC analysis with assay standards; Fig. S3). The 13 feature selection methods are described in Data file 3 and are based on 3 approaches to identify case versus control differential expression in the training set: 1) Analysis of DEGs across samples from all batches, 2) analysis of DEGs within each individual experimental batch, and 3) GSEA of individual experimental batches.

Gene sets for targeted analysis by NanoString

Gene-specific capture probes for NanoString nCounter analysis were synthesized by Integrated DNA Technologies (IDT). The NanoString development gene set was used for LC-iCAP parameter optimization and reproducibility studies. The set consisted of 96 genes including 74 features with case versus control differential expression in cohort 1 selected for detecting the LC-iCAP readout and 7 housekeeping genes for normalization (Data file 2A). The NanoString deployment gene set was used for training and testing the final LC-iCAP models in Stage 4 using a NanoString Plexset readout. The gene set consisted of 95 genes including 85 candidate features for modeling (66 selected in Stage 3 and 19 of the 25 genes from the initial model developed Stage 1 not already in the list, 1 control gene responsive to DMOG and 9 housekeeping genes for normalization (Data file 3D).

NanoString Plexset™ analysis of LC-iCAP RNA

Gene expression analysis of total RNA samples from the LC-iCAP was performed using NanoString Plexet™ technology, a direct detection approach for multiplexed gene expression analysis of up to 96 samples per run that does not involve PCR amplification. For this analysis, the deployment gene set consisting of 87 target genes and 9 candidate housekeeping genes were used as targets (Data file 3). Data were analyzed using an nCounter® Analysis System by the Genomics Resources Center at Fred Hutchinson Cancer Research Center following manufacturers recommendations. First, probe hybridization was performed in solution where gene-specific capture probes and reporter probes attached to fluorescent barcodes were used for detection of each of 96 mRNA molecules in each sample (see Data file 3 for Each reaction used 140 ng of LC-iCAP RNA for patient samples (and 100 ng for DMOG/PBS controls), which were pre-optimized in a calibration experiment to avoid saturating the chip). Next, samples were pooled in groups of 8 and loaded onto an nCounter Prep Station for automated excess probe removal and binding of the probe-target complexes on the surface of the cartridge by streptavidin-biotin linkage to the capture probes. Cartridges were placed in the nCounter Digital Analyzer for data collection, where molecules of RNA were counted by using the target specific “color codes” generated by a string of six fluorescent spots on each reporter probes.

LC-iCAP RNA samples were processed on 9 plates total with up to 96 samples per plate (see Data file 4 for processing configuration). Existing LC-iCAP data were used for cohort 3 and new LC-iCAP data were generated for all other samples using our standard method, except serum samples for cohorts 1-2 were thawed twice before analysis. For quality control, each plate included 8 positive controls composed of *in vitro* transcribed RNA transcripts and corresponding probes, and eight negative controls consisting of probes with no sequence homology to human RNA sequences for lower limit of detection analysis. One positive and one negative control were used for each of the 8 multiplexed samples in each lane. The experiment was done in two batches separated by more than one year using two different code set lot numbers. The first batch consisted of 8 plates containing training and validation set samples (cohorts 1-4) and the second batch consisted of the 9th plate containing cohort 5 blind test samples. A calibration sample used for data normalization was run in the first column of the first Plexset plate for each batch. Although the manufacturer recommends using the same sample for all batches in the study, a different calibration sample was used for each batch due to limited material supply, potentially introducing a Plexset batch effect.

Processing of raw Plexset data was performed using the nSolver analysis package following manufacturer's recommendations (MAN-C0019-08) with specifications as below. Data points falling below the lower limit of detection were floored using the background thresholding procedure with a threshold count value of 20. Next correction for lane and sample technical variation was done by housekeeping (HK) gene normalization, for which the expression value of each gene was divided by the geometric mean of 5 stably expressed housekeeping genes (ABCF1, FCF1, GUSB, POLR2A, and SDHA) in the same sample. Finally, code set calibration was performed to normalize the 8 sets of barcodes using 8 technical replicates of the same LC-iCAP samples, one for each barcode set in each row. Plates 1-8 and plate 9 were processed as two separate calibration batches and nSolver batch normalization was performed. Prior to using Nanostring data for modeling, raw counts were log2 transformed, and outlier removal was performed using ROBPCA[34], for which only data for housekeeping genes for each sample were included in the analysis.

Mitigation of NanoString Plexset miscalibration

Due to the Plexset calibration issue discussed above in the NanoString Plexset analysis Section, which resulted in miscalibration of the training samples (in calibration set 1) with the test set samples (in calibration set 2) and an apparent effect on gene expression in the LC-iCAP (Fig. S7), four different approaches were tested to mitigate this effect. These approaches included: 1) An optional Plexset stability filter to exclude genes with the strongest calibration batch effect between the two calibration batches (plates 1-8 versus plate 9). These genes were defined as those with significant differential expression between DMOG controls on Plexset plates 7-8 (processed with cohort 4) versus those on Plexset plate 9 (processed with cohort 5) (with FDR < 0.1) (Fig. S7, Data File 4). 2) Two different modeling algorithms, including generalized linear modeling (GLM) and RF, each with a different approach to integrating information from different features, which could differently affect the impact of the calibration effect. 3) nSolver batch correction to normalize the two different calibration batches to each other (as described above). 4) The R-normR package of Bioconductor.

Modeling

RF modeling was implemented in the R-caret package[35] using mtry values that were automatically selected for each seed using the default settings. In stage 1, modeling was done by leave-one-out cross-validation with 10 random seeds. In Stage 3, modeling was done by leave-one-out cross-validation with 50 resampling iterations followed by validation on the independent validation set. For each model configuration, 20 random seeds were generated, each using a maximum of 20 gene features selected based on variable importance score. The modeling study was repeated with downsampling of the training set to balance the number case and control samples within each experimental batch, this was done to assess model robustness and possible influence of batch imbalance on model performances. Top models were selected based on performance on the validation set (using either AUC or specificity at a fixed sensitivity of $\geq$ 95%). In Stage 4, GLM modeling was implemented in the R-glmnet package[36]. Nested cross-validation was done using R-nestedcv[37]. For each modeling condition, nested cross-validation was repeated 10 times, each with a unique random seed. Optimal modeling conditions were those with best median AUCs and seeds either at 50th or 75th percentile of performance.

Blind testing:

80 serum samples were selected by Vanderbilt University to meet the selection criteria used for model development with two additional criteria: The selection only included samples of high quality (those with an estimated hemolysis less than 50 mg/dL and storage times of 4 years or less) and samples from patients who were current and former smokers. Patient status and clinical data were blind to researchers at time of prediction. However, information on approximate overall ratio of case to control samples, and nodule size and smoking status were available. Samples were shipped in one batch and processed in random order in 4 LC-iCAP batches and one Plexset batch. LC-iCAP processing included 8 each of DMOG and PBS controls, but only the DMOG controls were included on the Plexset due to space limitations. 9 models using iCAP features and smoking status were tested on the blind validation set. For 3 of these models, a second version of the model including nodule size was tested to measure complementarity with CT scan data. Sample size of the blind test set was selected to have power to identify significant models with AUC of ROC >= 0.62 using an established method[21].

Mayo Clinic model

Pre-test risk of cancer malignancy was calculated for the test set by researchers at Vanderbilt University using Solitary Pulmonary Nodule (SPN) Malignancy Risk Score (Mayo Clinic model)[4]. This model uses 6 clinical risk factors including

nodule spiculation, upper lobe location, smoking status (current or former vs non), nodule diameter, age, extrathoracic cancer diagnosis >= 5 years prior. The test excludes patients with prior lung cancer diagnosis or with history of extrathoracic cancer diagnosed within 5 years of nodule presentation.

Integrated model

A prototype integrated classifier was developed after blind testing by integrating LC-iCAP model M4 with the Mayo Clinic model using an approach similar to that used for the Nodify XL2.[10] This approach is a decision tree model whereby the Mayo model's output is conditionally adjusted by a fixed amount based on a threshold established using the liquid biopsy readout: For LC-iCAP probabilities ≤ 0.45, the Mayo model's output was reduced by a fixed amount and for probabilities > 0.45, the Mayo model's output was used without adjustment. A similar integrated classifier was also developed for LC-iCAP model M3. See Figure S10 for the technical parameters.

The integrated model performance on the blind test set was compared to that of the Mayo model by generating ROC curves for both models and comparing points on each curve corresponding to maximum specificity at ≥ 95% NPV (with 25% disease prevalence) using McNemar's test.[12] Model calibration was not required for this comparison because it used points from the ROC curves corresponding to specific performance metrics rather than the absolute probability estimates. For verification, both models were calibrated to the test set prevalence using logistic regression (R-caret), and ROC curves were regenerated.

Control for error and bias

Key biological resources were authenticated: an aliquot of the 16HBE indicator cells was validated midway through the study externally at IDEXX bioanalytics by CellCheck 16™ Plus, and it passed all three tests. The cell line of origin was confirmed to be correct with 15 of 16 markers of the 16STR profile matching 16HBE from Sigma and no contamination from *mycoplasma spp.* or other species was detected. Serum samples were assayed for hemolysis and storage time and omitted if above established thresholds. Patient gender age and smoking history were approximately balanced between classes. All DNA constructs were sequenced. All chemical resources were from reputable commercial sources.

Controls in data generation included: blinding researchers to disease status; randomizing sample positions; balancing classes within batches; balancing patient attributes between batches; using moat plates to control edge effects in the cell-based assay; developing standard controls and using them to monitor assay performance and batch effects; performing reproducibility studies; measuring RNA integrity before RNA-seq and omitting samples below a threshold 7; and detecting and correcting biases from RNA-seq (GC bias and batch effect) and NanoString Plexset analyses (batch effect).

Modeling practices were used that control for biases and overfitting including: training sets were balanced with approximately equal numbers of case and control samples; all models contained fewer features than samples to prevent overfitting (except for the pilot model, which had 25 features and 12 samples); validation was conducted using independent samples or nested cross-validation, both robust methods that mitigate overestimation of model performance regardless of sample size [40]; different gene expression detection methods were used for feature selection and final model development to avoid effects of platform biases on model performance; final models were evaluated by blind testing with independent samples; and samples for blind testing had temporal independence from the training set for both sample collection and processing, increasing the rigor of the test.

Results

The iCAP is an in vitro cell-based assay platform for blood-based diagnostics that uses indicator cells as biosensors to detect and respond to disease signals in patient serum. The response of cells, discovered by gene expression machine learning analysis, is characteristic of the disease status (Fig. 1*A*;[14]). Here we describe the optimization and validation of the LC-iCAP for management of IPNs identified by CT scan to reduce false positives in lung cancer screening and reduce the number of unnecessary follow-up tests performed for those with benign nodules.

Study summary.

The study had PRoBE design[19] and used archived serum sample that were collected from 5 cohorts of patients with IPNs identified by CT scan (Fig. 2*A*). Nodules were later characterized as malignant or benign and samples were used to develop the LC-iCAP in 4 Stages outlined in Figure 1*B* using an iterative process of model training and validation with successively larger sample sizes to tune model parameters and hyperparameters, followed by an unbiased estimate of the performance of final models using a blind test set (Fig. 2*B*). Integrated into this approach were numerous analyses using both standard controls and patient metadata to pinpoint and mitigate various sources of unwanted variation in the assay readout from preanalytical and analytical sources. For modeling, sample numbers and data partitioning were based on power calculations and on sample and data quality assessments made during the study. Throughout the study, validation sets did not contain samples used for tuning model parameters and the research team was blinded to the final test set until after predictions were made. Both intermediate and final modeling steps followed the standard practice to use fewer features than samples to avoid overfitting [41,42]. Additionally, to reduce potential for error due to systematic biases in detection of gene expression, different detection approaches were used for feature selection (RNA-seq) and for final model testing (NanoString Plexset). The study is described in detail below.

Stage 1: Proof of concept: Training and testing a pilot LC-iCAP model

The aim of this stage was to characterize and initial LC-iCAP readout and develop a first model to establish proof-of-concept of the assay. To characterize an initial LC-iCAP readout, a 12-sample pilot cohort of case and control patient serum samples were acquired from Vanderbilt University (cohort 1; Fig. 2*A,* Data file 1). These samples were analyzed in the LC-iCAP along with reference serum controls. Gene expression was measured by RNA-seq and despite the small sample size, 239 differentially expressed genes were identified (FDR < 0.05; Benjamini-Hochberg). By STRING analysis, these genes had significantly more interactions than expected by chance, and were significantly enriched for HIF1A signaling (KEGG) and response to hypoxia (GO process) (FDR < 0.05), both of which are implicated in lung cancer [43]. Cohort 1 was used to train a Random Forest (RF) classifier for lung cancer prediction using cross-validation with the top 25 DEGs sorted by FDR as features and 10 random seeds.

Next, a set of 103 samples (cohort 2) was acquired and used to test the model (Fig. 2). This cohort was analyzed in 5 LC-iCAP and 2 RNA-seq batches (Data file 1). The pilot model was validated using all samples from cohort 2 (Fig. 2A). The model had significant overall performance (median seed AUC 0.62; 95% CI 0.51-0.73) (Fig. 3*A*). Similar models with 50, 75 or 100 features had comparable performances, but models developed using other data partitions were not significant (data not shown).

LC-iCAP data from cohorts 1 and 2 were analyzed by hierarchical clustering based on the expression of 20 top DEGs in cohort 1 (Fig. 3*B*). This analysis separated the samples into two distinct clusters, each enriched for either case or control classes. The modeling and clustering data suggest that the LC-specific differential expression observed in cohort 1 is generalizable, establishing feasibility of the LC-iCAP. However, they also indicate the presence of noise in the assay, potentially arising from biological diversity among patients, preanalytical variability in serum quality, and/or analytical variability from the cell-based assay or RNA-seq analysis. These three sources of noise were explored in Stages 2-4.

Stage 2: Analytical optimization of cell-based assay and reproducibility testing

Optimization and Standardization of LC-iCAP Assay Parameters:

Assay parameters were optimized with the goal of detecting and controlling for unwanted variation from analytical sources to improve reproducibility and magnitude of differential expression and thus model performance.

First, two sets of assay standard controls were developed to use for these optimization studies and to monitor assay performance. Biological standards were case and control serum pools, each generated by combining aliquots of serum from 8 patients from cohorts 1 and 2 (indicated in Fig. 3*B*). The differential expression readout was quantified using NanoString with a 'development gene set' containing 58 genes with significant case versus control differential expression in cohort 1 (Fig. S2; Data file 2). Analysis of 4 replicates each of case and control pools yielded a baseline differential expression of 55 genes (FDR < 0.1). Chemical standards were PBS and DMOG, a small-molecule agonist of the hypoxia response selected to mimic the activation of the hypoxia response in the assay by case versus control serum (see Stage

1). 24 of the 58 genes were responsive to DMOG and 23 of the 24 responsive genes were coherent with the readout of the pooled serum standards.

For assay optimization, 4 technical replicates of each serum pool standard were assayed under varying assay parameters and case versus control differential expression was compared. We first assessed the impact of multiple freeze-thaw cycles on the differential LC-iCAP readout comparing once-thawed versus twice-thawed serum pools. Both conditions yielded similar numbers of genes with significant differential expression suggesting that twice thawed serum is suitable for optimization studies (Fig. S1*A*). Next, indicator cell types, incubation times and serum concentrations were assessed to select the combination that maximized the differential expression response in the LC-iCAP. Testing parameters were 4 candidate indicator cell types (16HBE, A549, MRC5, and Nuli-1); 2 serum incubation times (6 hours, 24 hours); 4 serum concentrations (1%, 5%, 10%, and 20%); and the effect of Trichostatin A (TCA) addition, an inhibitor of the hypoxic response. The optimal LC-iCAP parameters were found to be a 24 h incubation of either 5% or 10% serum with 16HBE lung epithelial indicator cells in the absence of TCA, which matched the baseline conditions used in Stage 1 (Fig. S1*B-E*). To avoid potential bias from measuring only 58 genes, RNA samples from the TCA and cell type experiments were reanalyzed by RNA-seq, which yielded similar results (data not shown).

In addition to the analytical optimizations of the cell-based assay listed above, we detected and corrected sources of noise arising from RNA-seq and metadata analysis, including LC-iCAP and RNA-seq batch effects, artifactual duplicate reads RNA-seq data, and GC biases in the RNA-seq data. This included analysis of case versus control differential expression in patient metadata and model performance before and after corrections in various combinations. All three corrections improved the number of significant DEGs and model performances (data not shown). Therefore, these corrections were included in data processing in stage 3.

Assessment of Analytical Reproducibility of LC-iCAP Across Varied Experimental Conditions:

Next, we measured the reproducibility of differential expression across a variety of conditions to assess the analytical variability of the LC-iCAP. This was done using the serum pool controls and individual patient serum samples (6 of each class). Reproducibility was measured between three different indicator cell expansion batches, two different LC-iCAP batches run on different days, two different gene expression detection platforms (RNA-seq versus NanoString) and two different NanoString batches. For comparisons with the serum pools, reproducibility of differential expression was measured by fitting a linear regression model to test-versus-baseline data yielding R^2 values from 0.80-0.97 (Fig. S2*A-D*). For comparisons using individual samples, the expression profiles of individual genes across samples were compared between conditions showing that 92-98% of genes had significant correlations between conditions with FDR < 0.1 (Fig. S2*B-D*). For these experiments, reproducibility of differential expression (gene-level variability) was measured instead of the more commonly used measure of gene rank consistency across samples (sample-level variability) because the former is more stringent and has greater relevance in detecting subtle gene expression changes typical of biomarkers [44]. These data show that the LC-iCAP has sufficient analytical reproducibility to detect signal above background across various conditions used for assay development and suggest that variability identified in Stage 1 is from pre-analytical sources.

Validation of Hypoxia Signaling as a Generalizable Marker in LC-iCAP Using Pooled Serum Standards:

To assess the generalizability of the hypoxia response identified in Stage 1, we utilized the pooled serum standards, consisting of 16 samples, 12 of which were distinct from the original cohort 1 (Fig. 3*B*). RNA-seq data from two pooled serum control experiments under standard conditions were merged and analyzed for case versus control differential expression (Datafile 2) and GSEA, which identified hypoxia as the most enriched pathway (adjusted p-value < 0.05, Fig. S4). We further investigated the hypoxia response in the LC-iCAP by analyzing the pooled serum standards using an alternative bronchial epithelial cell line as indicator cells. We identified 61 genes that were significantly differentially expressed in both LC-iCAP RNA-seq experiments with 16HBE and Nuli1 indicator cells and found that the differential expression levels were correlated between the cell types (R 0.77, Fig. 4*A*, *left*). Next, STRING network analysis was done showing that the 47 coherently up-regulated genes had high network connectivity and enrichment of HIF1A/response to hypoxia and other processes (FDR < 0.001, Fig. 4*A, right*). To further elucidate the response mechanism, we compared levels of hypoxia-responsive transcription factors HIF1A and HIF2A between case and control classes in the LC-iCAP using

quantitative western blotting on both serum pools and individual samples. Significantly higher levels of both factors were observed in case sera compared to controls, an effect reduced by the addition of DMOG, a known HIF1A stabilizer (Fig. 4*B*). Collectively, these results suggest that hypoxia signaling is a generalizable marker of lung cancer in the LC-iCAP, underscoring the involvement of HIF1A and HIF2A in the response observed in indicator cells.

Stage 3: Model-based feature selection using LC-iCAP-RNA-seq data

To constrain final models and prevent overfitting, a modeling-based approach was used to identify the optimal genes from LC-iCAP RNA-seq data to use as candidates for final model development in Stage 4 with targeted LC-iCAP NanoString Plexset data. To achieve this downselection of the features, we conducted a large-scale modeling study whereby 104 RF models were trained including all combinations of 13 feature reduction approaches and 8 sample filtering approaches. The models were ranked based on performance on a held-out validation set and candidate features were selected from the top 3 models (Fig. 5).

Feature reduction approaches were based on case versus control differential expression either across all samples or within sample subsets. Sample filters were applied to exclude technical failures and to select subsets of samples based on patient lung function and smoking history, improving sample homogeneity. Meta-analysis of model performances revealed that all sample filters tested enhanced the number of differentially expressed genes (DEGs) and improved model performance. Additionally, feature reduction strategies that selected features from specific sample subsets, rather than the entire training set, produced the best-performing models (Fig. 5 top, Data file 3).

The three top models are shown (Fig. 5 *bottom*), the best of which had an AUC of 0.78 (90% CI 0.63-0.93) on the held-out validation set with sensitivity and specificity of 100% and 60%, respectively. The list of 85 features for final model development in Stage 4 was composed of 66 features selected from each of the 3 top models and an additional 19 genes from the model developed in Stage 1 (Data file 3).

This analysis used cohorts 1-2 from Stage 1 and samples from a new cohort 3. Differential expression analysis revealed that a cohort 3 batch had robust levels of DEGs (>200 DEGs with FDR <0.1) (Data file 2D), and significant upregulation of 'response to hypoxia' GO molecular process in the case versus control condition (p-value < 0.001). With this discovery, there were a total of 3 sample sets in the study with this differential enrichment of the hypoxic response.

Stage 4: Generation of LC-iCAP-Nanostring Plexset data and final model development

This Stage involved generating LC-iCAP Nanostring Plexset data for the 85 candidate features from Stage 3 across cohorts 1-3 and using the data for final LC-iCAP model development and blind testing. Switching from RNA-seq to Nanostring Plexset, a high-throughput platform capable of analyzing the expression of 96 genes across 96 samples per batch, was done to minimize analytical biases in feature selection and to initiate the development of a high-throughput assay configuration suitable for clinical deployment.

First, LC-iCAP Nanostring Plexset data were generated for cohorts 1-3 (Fig. 2*A,* Data file 4). Data were merged, normalized and filtered to exclude low-quality data and samples resulting in 97 samples remaining for modeling (Fig. 6). This included a filter to remove samples with storage exceeding 10 years supported by two meta-analyses of patient LC-iCAP data, which demonstrated a significant effect of storage time on the LC-iCAP readout (Fig. S8).

The 97 samples were used as a training set for model development. Selection of parameters and hyperparameters and estimation of model performances was done with nested cross-validation, an approach that has been demonstrated to generate unbiased performance estimates with small sample sizes[40]. The parameters tested included the sample filters optimized in Stage 3 and inclusion of patient smoking status (current or former) as a covariate feature in the model.

To test a selection of fully parameterized models, a new set of 80-samples from Vanderbilt University (cohort 5) was added to the study for blind testing of (Figs. 2 and 6). All samples of the blind set had pretest risk of malignancy of at least 5% calculated using the Mayo Clinic model and the patient and sample attributes matched those of the other cohorts except, all were from current or former smokers, and all were of high serum quality with storage times between 1-4 years. The sample source was the same as for the training set samples, but the samples for blind testing had temporal independence

from the others, with a collection window that was 9 years later and LC-iCAP processing that was 1 year later, adding more rigor to the validation than concurrent collection and processing.[45] Sample classes were blinded, and researchers were provided only with smoking status and nodule sizes of the patients (for optional inclusion of nodule size in the LC-iCAP model to test for orthogonality to CT data).

Before selecting models for blind testing, we first demonstrated assay reproducibility with the Plexset readout by hierarchical clustering using the standard control data (Fig. S7). This analysis revealed the effects of a Plexset calibration issue between the training and test set described in the Methods section. The method to overcome the miscalibration could not be optimized prior to blind testing because only the calibration batch containing the blind samples was affected; therefore, 4 different normalization approaches were developed (described in Methods and Fig. S9) and top models were tested on the blind set iteratively with different combinations of normalization approaches.

9 models were tested in sequence on the blind set, the last two of which had significant performance with AUCs of 0.64 ad 0.63 (M3 and M4 in Fig. 7). Both significant models used the same combination of 3 approaches to overcome the Plexset miscalibration including using the RF algorithm, nSolver batch correction and the Plexset stability filter together. Case versus control differential expression was compared between the training and test sets and there was significant correlation for not only M3 and M4 genes, with R of 0.80 and 0.73, respectively, but for genes excluded due to Plexset miscalibration suggesting presence of a large number of predictive genes on the Plexset gene panel (Fig. 8). No genes had significant differential expression in both sample sets suggests that the model performance is influenced by multiple features, each contributing small effect sizes.

To explore the potential clinical utility of the LC-iCAP, a prototype clinical version of the assay was developed after blind testing by integrating the LC-iCAP model M4 with the Mayo Clinic model[4] using an approach similar to that used for the Nodify XL2 test offered by Biodesix[10] as described in Figure S10. Performance of this 'iCAP integrated classifier' was compared to that of the Mayo Clinic model by generating ROC curves for each model and comparing maximum specificities at cut points with clinical utility as a rule-out test (corresponding to NPV $\geq$ 95% using an estimated cancer prevalence of 25% in a community pulmonary practice[15,16]) (Fig 9). The performance of the iCAP integrated classifier was significantly better than that of the Mayo Clinic model suggesting clinical utility of the LC-iCAP (McNemar p-value[12] 0.037). A second integrated classifier was developed using LC-iCAP M3, which had a similar performance and significant improvement over the Mayo Clinic model (Fig. 9). Because the models were compared at ROC curve points with specific performance metrics (rather than absolute probability estimates), model calibration was not required and when implemented had no effect on the results (Fig. S11). Significant improvement was not observed using the training set samples for either model, which could be due to longer sample storage times of the training versus test sets, or variability due to low sample sizes. The parameters of the integrated classifier were selected to maximize clinical utility on the blind samples. Thus, while this study validated the performance of the LC-iCAP model, the constants of the iCAP integrated classifier have not yet been validated.

Discussion

Blood biomarkers are needed for the early detection of diseases to improve outcomes, but their very low abundance and high levels of noise present significant technical challenges. We are developing the iCAP, a biosensor assay for blood-based diagnostics to overcome this issue by capitalizing on the cells' evolved ability to detect weak signals in noisy environments. The assay works by using cultured cells as biosensors to detect disease-related molecules in blood and analyzing the gene expression response using machine learning tools to develop disease classifiers. This approach enables using established tools for global gene expression analysis from cultured cells to simplify model development and deployment and avoids inherent noise in gene expression arising from genetic variation between patients[46]. Here, we initiated the development of the LC-iCAP, a blood test for patients with IPNs identified by CT scans to improve malignancy risk assessment and help those with benign nodules avoid invasive biopsies while directing further diagnostic efforts towards those with lung cancer.

Cell-based approaches have brought many benefits to drug discovery that can be applied to blood-based diagnostics using the iCAP.[47–49] Here, the consolidation of disease signals in blood into a cell-based case versus control differential expression readout enabled us to apply statistical enrichment analysis to identify a hypoxia response in the lung cancer-specific

readout, and biochemical analyses to implicate transcription factors HIF1A and HIF2A in mediating the response (Fig. 4). This finding aligns with hypoxia as a well-documented characteristic of lung cancer and other malignancies,[43,50,51] and it suggests that the LC-iCAP readout reflects blood analytes associated with the tumor status.

The cell-based approach also enabled us to develop standardized biological and chemical controls with quantitative multicomponent readouts that we used to optimize assay parameters, perform reproducibility studies and detect and control for sources of unwanted variation. We found that the LC-iCAP consistently demonstrated reproducible disease versus normal differential expression across various conditions and that noise in the assay was primarily from patient heterogeneity and pre-analytical sources. We specifically identified sample storage time and patient lung function as sources of noise, supported by studies demonstrating that sample storage time alters the abundance of blood components,[52,53] and smoking history impacts lung cellular biochemistry.[11,54,55] We used these data to select model parameters, aiming to reduce overfitting by employing a data-driven approach rather than relying solely on performance-based parameterization.

We successfully validated two LC-iCAP models through blind testing, based on the expression of 17 or 36 gene features and patient smoking status, achieving AUCs of 0.63 (95% CI: 0.50–0.75) and 0.64 (95% CI: 0.51–0.76) for the respective models (Fig. 7). Although training and test samples were from Vanderbilt University, the test set samples were temporally independent from the training set samples in collection and processing, enhancing the rigor of the validation process compared to an internal validation approach where one dataset is randomly split into a training and test set.[45]

Several different feature selection approaches were explored during model development, and both validated models used approaches that measured differential expression within subsets of samples rather than the entire training set (Fig. 5). Additionally, while 85% and 77% of gene features with detection above background for models 3 and 4, exhibited consistent differential expression between the training and test sets, none demonstrated significant differential expression in both sets (Fig. 8). These two findings underscore the significant patient-to-patient heterogeneity of early-stage biomarkers and the contribution of numerous features, each with small effect sizes, on model performance. This is consistent with the patient heterogeneity we observed in the LC-iCAP readout and highlights the importance of developing multivariate models for cancer detection.

After blind testing we assess potential clinical utility of the LC-iCAP by integrating it with the Mayo clinic model using a method previously developed for deploying the Nodify XL2 test by Biodesix[10] (Figs. 9 and S10). The 'iCAP integrated classifier' performance was compared to that of the Mayo model at specific cut points yielding ≥ 95% NPV using a 25% prevalence in the intended use population.[15,16] This threshold was selected to maximize clinical utility as a rule-out test to discriminate nodules that have a 5% risk of cancer that can be diverted to surveillance from those with higher risk that require further testing. At this threshold, the iCAP integrated classifier had 67% specificity and 90% sensitivity and significantly better performance than the Mayo model suggesting clinical utility (Fig. 9).

In a clinical setting where a rule-out test is used to guide patient care, specificity at the cut point indicates the percentage of patients with benign nodules who would be directed to surveillance, thus avoiding potentially invasive follow-up procedures. The specificity of the iCAP integrated classifier was 1.5-2X better than that reported for the two other CT-integrated blood tests from Biodesix and Magarray at similar NPV thresholds, (67% versus 44% and 33%[10,11], respectively). This suggests actionable results for a greater number of patients with benign nodules using the LC-iCAP. Sensitivity, which reflects the percentage of patients with malignant nodules who would be correctly directed to follow-up testing, was 90% for the iCAP integrated classifier. Although this is lower than that of the other tests (97% and 94%), the iCAP integrated classifier would have a 95% NPV in the intended use population, attributable to its higher specificity at the cut point, thus not substantially increasing harm.

While the data are promising, this is study has limitations. Our data suggests that our use of archived samples with variable storage times up to 10 years had a negative effect on model development and performance (Fig. S8). In addition, due to a technical issue necessitating concurrent batch correction and model testing, nine models were tested on the blind test set. Finally, although temporal validation is rigorous, external validation is required for testing generalizability of the model to new collection sites. We plan to conduct a multi-site prospective study with a larger sample size to further improve the

performance and robustness of the model, followed by a clinical utility study as recommended in the American Thoracic Society policy statement.[8]

Because of its broad search space, multicomponent readout and cost-effective scalability, the iCAP could have utility as a next-generation platform for cancer screening including multi-cancer early detection (MCED). This could involve either using an array of indicator cells or tuning the LC-iCAP to detect multiple cancer types. Supporting this, HIF1A in the LC-iCAP readout has central roles in general tumor biology, including response to inflammation, adaptation to hypoxia, and stimulation of growth of certain cancers. [50,51] Notably, hypoxic tumors are more likely to metastasize and are less likely to respond to treatment[51] and to our knowledge, blood biomarkers of tumor hypoxia have not yet been identified. Future studies include exploring use of fluorescent reporters and single cell analysis to further simplify and amplify the LC-iCAP readout, as well as scaling all analytical steps to 96-well configurations using tools already in use for diagnostics such as QuantiGene Plex.

To achieve the Cancer Moonshot initiative's goal of accurate early detection with minimal overdiagnosis and missed cases, relying on only one test is unlikely to yield optimal clinical utility. Multimodal approaches, which combine multiple orthogonal tests each with independent performance, can outperform individual tests in isolation.[9,12] The iCAP is a non-conventional approach that is complementary to other more traditional tests and thus well-suited for combinatorial diagnostics. Through collaborative efforts, just as the power of combining treatments has revolutionized therapeutics, the integration of diverse diagnostic modalities holds the promise of transforming diagnostics as well.

References

1. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **71**, 209–249 (2021).
2. National Lung Screening Trial Team. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening | NEJM. *N Engl J Med* 395–409 (2011) doi:10.1056/NEJMoa1102873.
3. US Preventive Services Task Force *et al.* Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **325**, 962 (2021).
4. Swensen, S. J. The Probability of Malignancy in Solitary Pulmonary Nodules: Application to Small Radiologically Indeterminate Nodules. *Arch. Intern. Med.* **157**, 849 (1997).
5. Lokhandwala, T. *et al.* Costs of Diagnostic Assessment for Lung Cancer: A Medicare Claims Analysis. *Clin. Lung Cancer* **18**, e27–e34 (2017).
6. Massion, P. P. & Walker, R. C. Indeterminate Pulmonary Nodules: Risk for Having or for Developing Lung Cancer? *Cancer Prev. Res. (Phila. Pa.)* **7**, 1173–1178 (2014).
7. Rendle, K. A. *et al.* Rates of Downstream Procedures and Complications Associated With Lung Cancer Screening in Routine Clinical Practice. *Ann. Intern. Med.* **177**, 18–28 (2024).

8. Mazzone, P. J. *et al.* Evaluating Molecular Biomarkers for the Early Detection of Lung Cancer: When Is a Biomarker Ready for Clinical Use? An Official American Thoracic Society Policy Statement. *Am. J. Respir. Crit. Care Med.* **196**, e15–e29 (2017).
9. Sniatynski, M. J. *et al.* Ranks underlie outcome of combining classifiers: Quantitative roles for diversity and accuracy. *Patterns* **3**, 100415 (2021).
10. Silvestri, G. A. *et al.* Assessment of Plasma Proteomics Biomarker's Ability to Distinguish Benign From Malignant Lung Nodules. *Chest* **154**, 491–500 (2018).
11. N. Trivedi, N. *et al.* Risk assessment for indeterminate pulmonary nodules using a novel, plasma-protein based biomarker assay. *Biomed. Res. Clin. Pract.* **3**, (2018).
12. Kuncheva, L. I. *Combining Pattern Classifiers: Methods and Algorithms*. (J. Wiley, Hoboken, NJ, 2004).
13. Silvestri, G. A. *et al.* A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. *N. Engl. J. Med.* **373**, 243–251 (2015).
14. Danziger, S. A. *et al.* An indicator cell assay for blood-based diagnostics. *PLOS ONE* **12**, e0178608 (2017).
15. Gould, M. K. *et al.* Recent Trends in the Identification of Incidental Pulmonary Nodules. *Am. J. Respir. Crit. Care Med.* **192**, 1208–1214 (2015).
16. Tanner, N. T. *et al.* Management of Pulmonary Nodules by Community Pulmonologists: A Multicenter Observational Study. *Chest* **148**, 1405–1414 (2015).
17. Moons, K. G. M. *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann. Intern. Med.* **162**, W1-73 (2015).
18. Tuck, M. K. *et al.* Standard Operating Procedures for Serum and Plasma Collection: Early Detection Research Network Consensus Statement Standard Operating Procedure Integration Working Group. *J. Proteome Res.* **8**, 113–117 (2009).
19. Pepe, M. S., Feng, Z., Janes, H., Bossuyt, P. M. & Potter, J. D. Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design. *JNCI J. Natl. Cancer Inst.* **100**, 1432–1438 (2008).
20. Li, X. -j. *et al.* A Blood-Based Proteomic Classifier for the Molecular Characterization of Pulmonary Nodules. *Sci. Transl. Med.* **5**, 207ra142-207ra142 (2013).

21. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).

22. Lippi, G. *et al.* Haemolysis: an overview of the leading cause of unsuitable specimens in clinical laboratories. *Clin. Chem. Lab. Med.* **46**, 764–772 (2008).

23. A Quick-Reference Tool for Hemolysis Statis | CDC. *Centers for Disease Control and Prevention* https://www.cdc.gov/ncezid/dvbd/stories/research-lab-diagnostics/hemolysis-palette.html (2023).

24. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

25. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

26. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

27. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics* **12**, 480 (2011).

28. Hansen, K. D., Irizarry, R. A. & WU, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).

29. Chen, Y., McCarthy, D., Baldoni, P., Robinson, M. & Smyth, G. edgeR: differential analysis of sequence read count data User's Guide.

30. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

31. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).

32. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

33. Hubert, M., Rousseeuw, P. J. & Vanden Branden, K. ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics* **47**, 64–79 (2005).

34. Croux, C., Filzmoser, P. & Oliveira, M. R. Algorithms for Projection–Pursuit robust principal component analysis. *Chemom. Intell. Lab. Syst.* **87**, 218–225 (2007).

35. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **28**, 1–26 (2008).

36. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).

37. Lewis, M. J. *et al.* nestedcv: an R package for fast implementation of nested cross-validation with embedded feature selection designed for transcriptomics and high-dimensional data. *Bioinforma. Adv.* **3**, vbad048 (2023).

38. Platt, J. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Adv Large Margin Classif* **10**, (2000).

39. Böken, B. On the appropriateness of Platt scaling in classifier calibration. *Inf. Syst.* **95**, 101641 (2021).

40. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. *PLOS ONE* **14**, e0224365 (2019).

41. Olivier, M., Asmis, R., Hawkins, G. A., Howard, T. D. & Cox, L. A. The Need for Multi-Omics Biomarker Signatures in Precision Medicine. *Int. J. Mol. Sci.* **20**, 4781 (2019).

42. Song, W. *et al.* A Comprehensive Evaluation of Cross-Omics Blood-Based Biomarkers for Neuropsychiatric Disorders. *J. Pers. Med.* **11**, 1247 (2021).

43. Shimoda, L. A. & Semenza, G. L. HIF and the Lung. *Am. J. Respir. Crit. Care Med.* **183**, 152–156 (2011).

44. Ballouz, S. & Gillis, J. AuPairWise: A Method to Estimate RNA-Seq Replicability through Co-expression. *PLOS Comput. Biol.* **12**, e1004868 (2016).

45. Altman, D. G., Vergouwe, Y., Royston, P. & Moons, K. G. M. Prognosis and prognostic research: validating a prognostic model. *BMJ* **338**, b605 (2009).

46. DeBoever, C. *et al.* Large-Scale Profiling Reveals the Influence of Genetic Variation on Gene Expression in Human Induced Pluripotent Stem Cells. *Cell Stem Cell* **20**, 533-546.e7 (2017).

47. Blay, V., Tolani, B., Ho, S. P. & Arkin, M. R. High-Throughput Screening: today's biochemical and cell-based approaches. *Drug Discov. Today* **25**, 1807–1821 (2020).

48. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).

49. Wei, F., Wang, S. & Gou, X. A review for cell-based screening methods in drug discovery. *Biophys. Rep.* **7**, 504–516 (2021).

50. Iommarini, L., Porcelli, A. M., Gasparre, G. & Kurelac, I. Non-Canonical Mechanisms Regulating Hypoxia-Inducible Factor 1 Alpha in Cancer. *Front. Oncol.* **7**, 286 (2017).

51. Weidemann, A. & Johnson, R. S. Biology of HIF-1α. *Cell Death Differ.* **15**, 621–627 (2008).

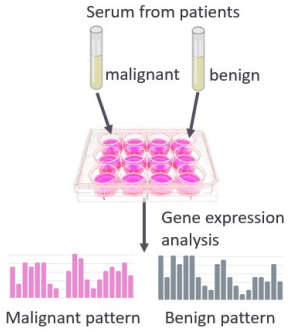
52. Haslacher, H. *et al.* The effect of storage temperature fluctuations on the stability of biochemical analytes in blood serum. *Clin. Chem. Lab. Med.* **55**, 974–983 (2017).

53. Wagner-Golbs, A. *et al.* Effects of Long-Term Storage at −80 °C on the Human Plasma Metabolome. *Metabolites* **9**, 99 (2019).

54. Fahrmann, J. F. *et al.* Blood-Based Biomarker Panel for Personalized Lung Cancer Risk Assessment. *J. Clin. Oncol.* **40**, 876–883 (2022).

55. Mao, L. *et al.* Clonal genetic alterations in the lungs of current and former smokers. *J. Natl. Cancer Inst.* **89**, 857–862 (1997).

A



B

Stage 1	Establish LC-iCAP proof of concept
Stage 2	Optimize assay parameters and test analytical reproducibility
Stage 3	Optimize model parameters and select model features
Stage 4	Transition to HTP readout Train and blind test final model

Figure 1. *A*, The iCAP for blood-based diagnostics. Standardized cells are exposed to serum from patients. Gene expression readout of cells is used to develop machine learning-based models to predict disease. *B,* Stages of LC-iCAP development. Sources of unwanted variation from patient biological diversity and serum quality were detected and mitigated in stages 3-4. HTP, high throughput.

A

Description of samples	Class*	Sample count per cohort			
		cohort 1	cohort 2	cohort 3	cohort 5**
Serum from patients with non-calcified IPNs identified by CT scan between 5-30 mm in diameter (95% > 7 mm)	Case: patients with malignant nodules	6	53	50	40
	Control: patients with benign nodules	6	50	50	40
	subtotal	12	103	100	80
	total	295			

B

	Stage 1	Stage 3	Stage 4
Training set: for learning while tuning model parameters & hyperparameters	cohort 1 12 samples	cohorts 1-3 109 samples	cohorts 1-3 97 samples
Validation set: a held-out set for estimating model performances while training	cohort 2 103 samples	cohorts 1-2 28 samples	none
Test set: for unbiased estimation of performance of the final fully-specified models	none	none	cohort 5 79 samples

Figure 2. *A*, Summary of sample characteristics and cohorts used for LC-iCAP development. *IPNs were classed as malignant or benign after CT scan by either diagnostic biopsy or resection or by 2 or more years of serial imaging. **Cohort 5 subjects were current and former smokers by design. Cohort 4 is not shown as it was replaced with cohort 5 due to low sample quality (see Methods). *B*, Modeling data partitions for 3 stages. IPN, indeterminate pulmonary nodule; CT, low-dose computed tomography. Patents have no other known cancer at the time of CT scan. All samples shown were from Vanderbilt University.

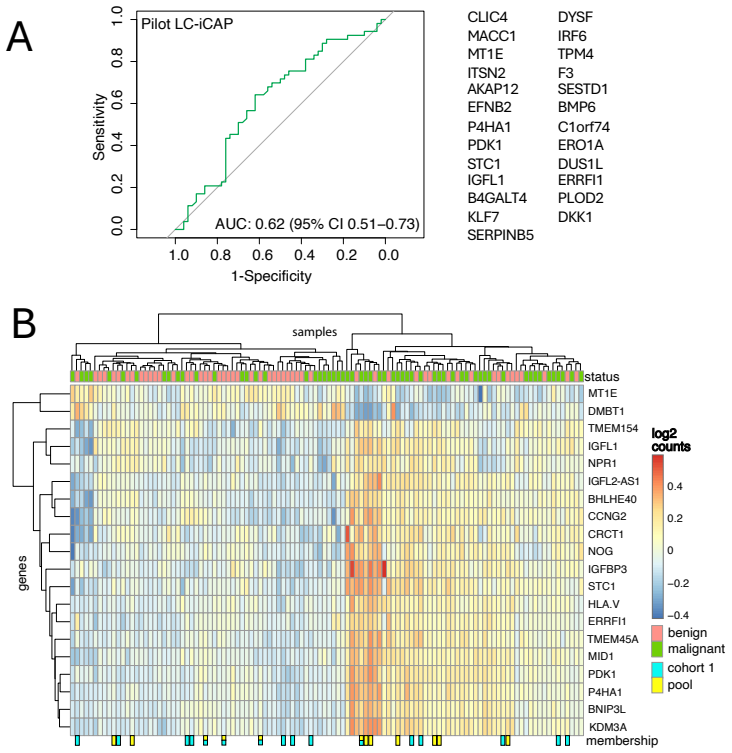


Figure 3. Proof of concept of LC-iCAP. *A*, ROC curve showing performance of a pilot LC-iCAP model trained on a 12-samples training set (cohort 1) and tested on 103-sample validation set (cohort 2). The model is based on 25 gene expression features (*right*). *B*, Hierarchical clustering of cohorts 1 and 2 based on LC-iCAP gene expression shows grouping of samples by class. Gene expression values are log counts that have been normalized to the mean expression of the benign samples within the same experimental batch. The 20 genes used for clustering were the top 20 differentially expressed genes in LC-iCAP data of cohort 1 based on median absolute deviation. Dendograms show gene clusters (*left*) and two distinct sample clusters (*top*). Only 112 of 115 samples are shown as 3 outliers were removed. Samples used to make serum pools (biological standards) and the 12 samples in cohort 1 are marked.

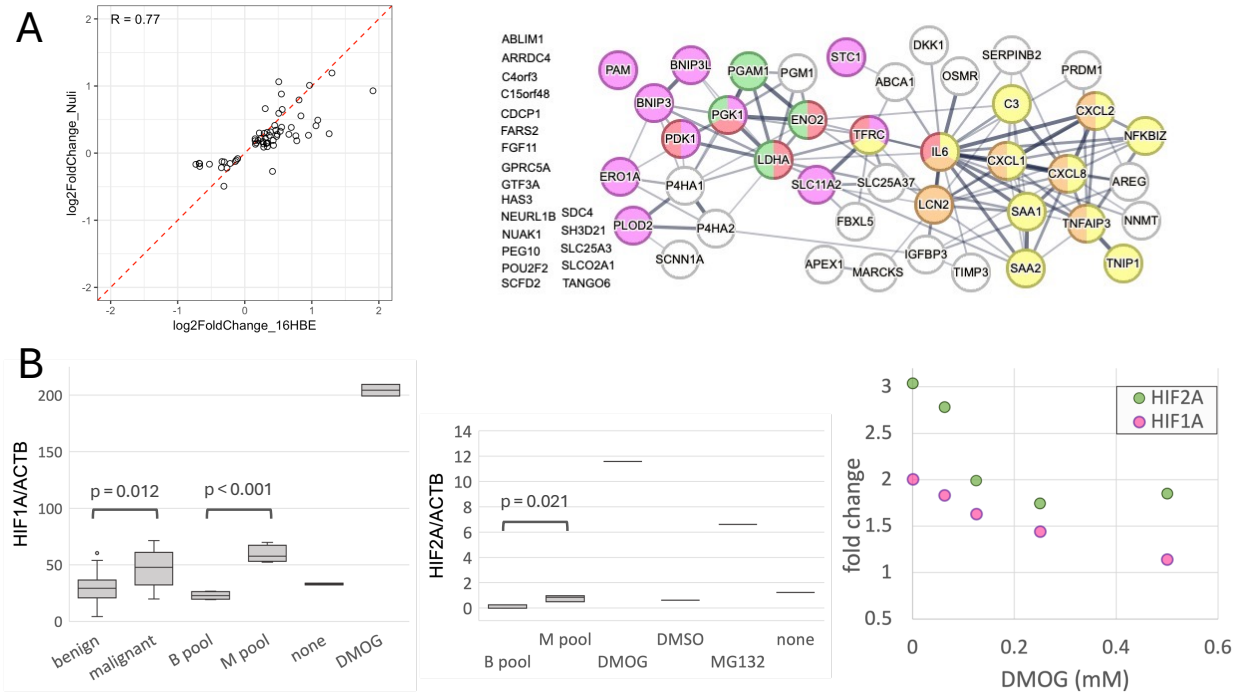


Figure 4. The LC-iCAP response to case versus control serum is enriched for HIF1A-mediated hypoxia signaling. ***A.*** *Left,* comparison of case versus control differential expression in LC-iCAP RNAseq data between two different bronchial epithelial indicator cell lines (16HBE and NuLi-1) using technical replicates of pooled serum controls. The 61 genes with differential expression in both cell types are shown (FDR < 0.1). *Right,* STRING network analysis of the upregulated genes show significantly enriched connectivity (p-value <1E-16) and enrichment for HIF1A/response to hypoxia (*red/pink*), glycolysis (*green*), and IL-17 signaling/inflammation (*orange/yellow*) (FDR < 0.001). Edge thickness represent strength of interactions and unconnected nodes are listed at the left. ***B.*** Results of western blot analysis showing upregulation of HIF1A and HIF2A transcription factors in the LC-iCAP in response to case versus control serum. *Left and middle,* box plots showing levels of HIF1A or HIF2A normalized to actin across 4 replicates of pooled serum controls (M pool and B pool) and 28 different individual serum samples including 14 of each class (malignant and benign). Individual samples included the 12 samples used to make the pools. Positive controls were one or two replicates each of DMOG or Mg132 versus DMSO or no stimulus. *Right*, one replicate of each serum pool was analyzed in the LC-iCAP with increasing concentrations of DMOG showing that DMOG dampens case versus control differential expression for both factors. Western blot images are in Figure S5.

109-sample training set and 28-sample validation set from cohorts 1-3

104 RF MODEL CONFIGURATIONS: Trained 20 seeds for each using 50 x LOOCV and tested on the validation set

8 SAMPLE FILTERS: Combinations of 3 optional omissions:	% significant models
Patients with low lung function	35%
Samples from failed RNAseq batch	29%
Patients who have never smoked	27%
No filter	15%
13 FEATURE FILTERS OF 3 TYPES:	
DEGs across all batches	0%
DEGs within individual batches	27%
Leading edge genes from GSEA of individual batches	27%

3 TOP MODELS SELECTED

Model + sample filter	Gene feature type	Train / validation sample number	Train / validation AUC (90% CI)	% significant models (with down sampling)**
M3-1		14* / 26	0.81 / 0.81 (0.66-0.95)	45% (25%)
M3-2		70 / 25	0.65 / 0.71 (0.54-0.89)	90% (60%)
M3-3		80 / 27	0.86 / 0.78 (0.63-0.93)	100% (70%)

*training set only included cohort 3 samples.**modeling was repeated with down sampling to balance classes within each experimental batch for training.

66 features used for final model development in stage 4

Figure 5. Stage 3 model-based feature selection using LC-iCAP RNAseq data for final model development in stage 4. 104 model configurations were trained and tested. % significant models is the percent of models having at least one seed with significant performance in validation. 66 features from multiple seeds of the three top models were use for stage 4. DEG, differentially expressed gene; LOOCV, leave one out cross validation; GSEA, gene set enrichment analysis.

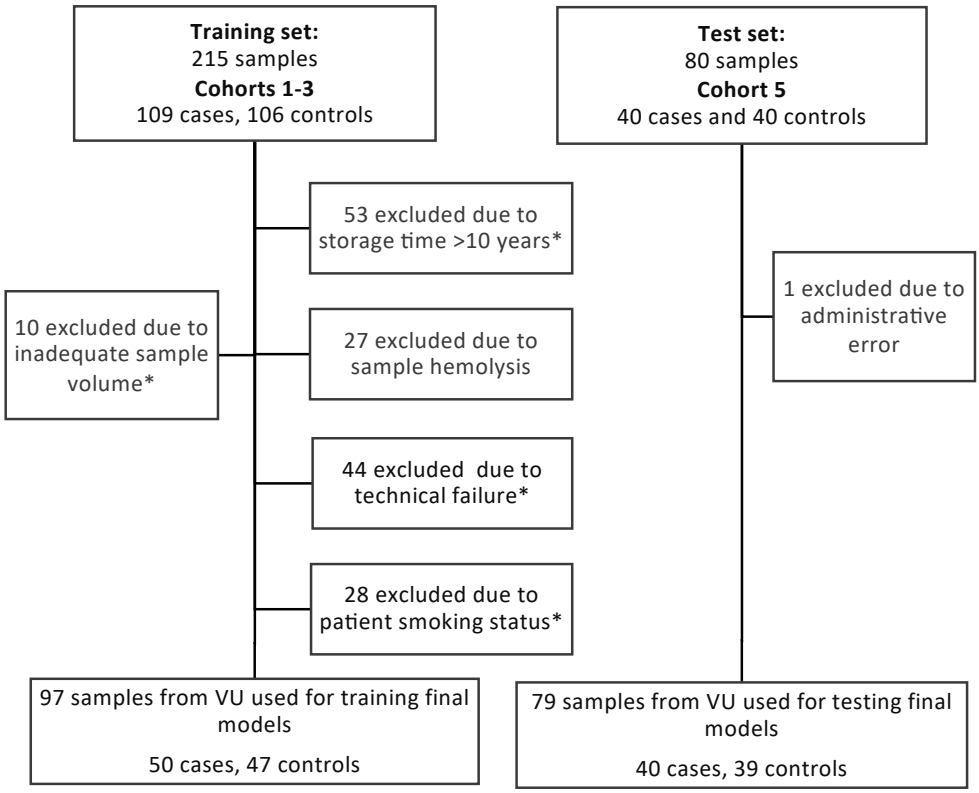


Figure 6. Retrospective sample flow diagram for final LC-iCAP models in stage 4. *Left*, The training set is made up of 3 cohorts. *Right,* a new cohort was acquired after completion of data collection for blind validation. The new samples were collected ~10 years after training samples. Sample numbers are not cumulative, as a sample may belong to multiple exclusion groups. *Samples omitted from final classifier training were used for LC-iCAP development and/or feature selection. All samples are from Vanderbilt University.

Table 1. Participant demographics

	Train, N=97				Test, N=79			
Characteristic	**N**	**benign**, N = 47	**malignant**, N=50	**p-value**	**N**	**benign**, N = 39	**malignant**, N= 40	**p-value**
Age, Mean (SD)	97	63 (9)	64 (8)	0.50[1]	79	64 (8)	65 (7)	0.72[1]
sex, n (%)	97			0.59[2]	79			0.24[2]
Female	38	19 (40)	19 (38)		28	12	16	
Male	59	28 (60)	31 (62)		51	27	24	
smoking status, n (%)	97			0.02[2]	79			0.08[2]
current	43	25 (53)	18 (36)		35	20	15	
former	54	22 (47)	32 (64)		44	19	25	
pack years	97	53 (28)	54 (34)	0.90[1]	79	60 (35)	62 (28)	0.741
nodule size, Mean (SD)	97	12.24 (5.97)	17.82 (4.35)	<0.001[1]	79	12.07 (5.33)	17.87 (6.53)	<0.001[1]
storage years, Mean (SD)	97	8.10 (1.59)	6.49 (2.18)	<0.001[1]	79	3.44 (0.72)	3.65 (0.62)	0.16

[1] Welch Two Sample t-test [2] Person's Chi-squared test

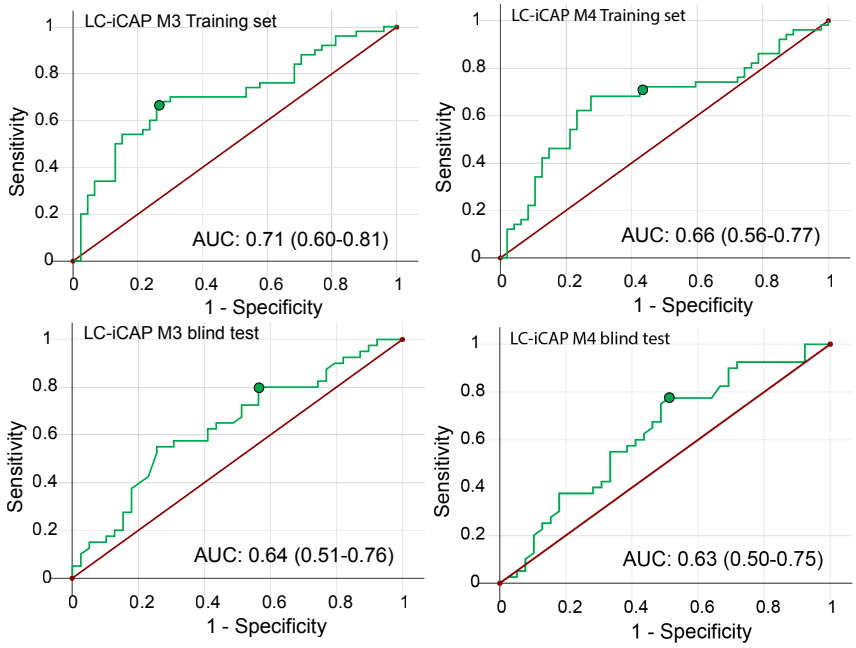


Figure 7. ROC curves of M3 and M4 on the training and blind test sets. AUC 95% CI is shown in brackets. LC-iCAP integrated classifiers were developed using the cut points shown on the blind test ROC (*green circle*) (0.49 for M3 and 0.45 for M4).

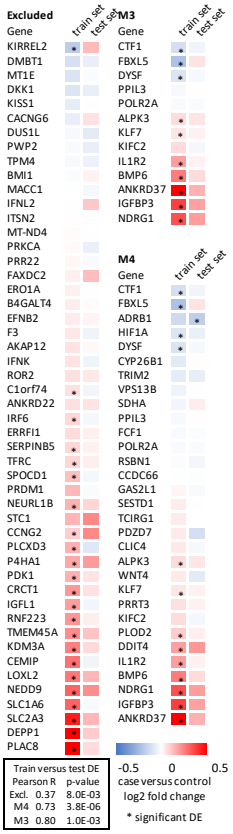


Figure 8. Heatmap of LC-iCAP Plexset data showing case versus control differential expression for genes of M3 and M4 and genes not used in either model. Significant differential expression is indicated with asterisks (Mann Whitney U test p-values <0.05). There was significant correlation between training and test sets across genes in models M3 and M4 and excluded genes (Pearson correlations and p-values are shown). Only genes detected above background are shown. These data suggest that the LC-iCAP is informed by multiple features, each with small effect sizes.

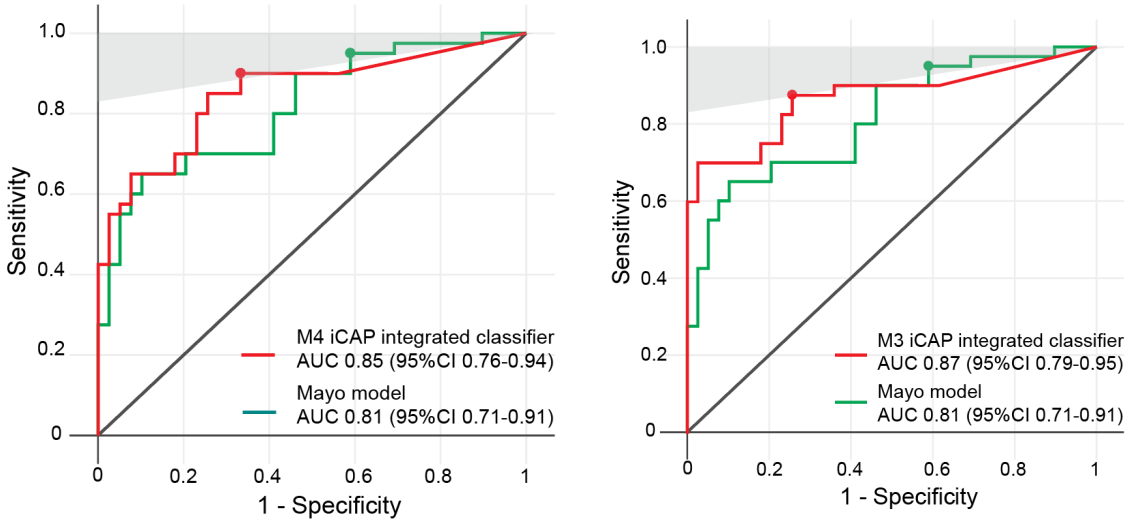


Figure 9. The iCAP-integrated classifiers outperforms the Mayo Clinic model on the test set. ROC curves for the Mayo Clinic model and the iCAP-integrated classifier (which combines LC-iCAP with the Mayo Clinic model as defined in Fig. 10) are shown for each LC-iCAP classifier M3 and M4. The region of the graphs corresponding to 95% NPV or greater was determined with an estimated 25% prevalence in the expected clinical population (shaded in gray). The model performances were compared at cut points that maximize specificity and have >=95% NPV, which were selected to yield clinical utility as rule-out tests (marked with colored nodes). Both of the integrated classifiers had significantly higher accuracies compared to the Mayo model (liberal one-tailed McNemar's test p-value = 0.037 and 0.006 for M4 and M3 integrated classifier, respectively). At the cut points, the models had similar NPVs (94.7% for M3 integrated, 95.2% for M4 integrated and 96.1% for the Mayo model), corresponding to a 4-5% cancer risk for negative results. However, the integrated classifiers had better specificities (74.4% for M3 integrated, 66.7% for M4 integrated, and 41.0% for the Mayo model). AUC, area under the receiver operating characteristic (ROC) curve; CI, confidence interval; NPV, negative predictive value.