Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# VarSee™ (Lung), a Blood-Based Multi-Protein Biomarker Panel for the Early Detection of Lung Cancer: A Prospective Pilot Diagnostic Accuracy Study

**Jahnavi Kulkarni**
BioMarkIQ Scientific Technology Private Limited
**Kamal Biswas**
BioMarkIQ Scientific Technology Private Limited
**Kalyani Gore**
BioMarkIQ Scientific Technology Private Limited
**Ashish Dhande**
Dhande Dhande Pathlab Diagnostics Pvt. Ltd
**Atharva Joshi**
BioMarkIQ Scientific Technology Private Limited
**Manasi Khasnis**
BioMarkIQ Scientific Technology Private Limited
**Snahlata Singh**
snahlata.singh@biomarkiq.com

BioMarkIQ Scientific Technology Private Limited

**Research Article**

**Keywords:** Lung cancer, Early detection, Serum protein biomarkers, Liquid biopsy, Chemiluminescence immunoassay, Machine learning, Random forest, Diagnostic accuracy, Neutrophil-to-lymphocyte ratio, STARD

**Posted Date:** November 1st, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11291035/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** Competing interest reported. J.K., K.B., K.G., A.J., M.K. and S.S. are employed by BioMarkIQ Scientific Technologies Pvt. Ltd., which develops and markets the VarSee™ (Lung) platform evaluated in this article. BioMarkIQ holds a pending Indian patent application (Application No. 202421079476, "Machine Learning Method for Curating Biomarker Panel, for Screening of Cancer(s) and System Therefor") covering the biomarker-panel curation method described in the Methods.

## Abstract

# Background

Most non-small-cell lung cancers are diagnosed at an advanced stage because early disease is largely asymptomatic. Low-dose computed tomography (LDCT), the recommended screening modality for high-risk individuals, is constrained by radiation exposure, limited accessibility and a high false-positive rate. We evaluated VarSee™(Lung), a serum multi-protein biomarker panel coupled to a machine-learning classifier, as a blood-based adjunct for lung-cancer detection.

# Methods

In a consecutive, prospective, single-centre, case-control pilot study with biomarker measurement performed retrospectively on banked serum samples, seven serum proteins (CA125, CA19-9, carcinoembryonic antigen [CEA], cytokeratin-19 fragment [CYFRA 21 − 1], human epididymis protein 4 [HE4], neuron-specific enolase [NSE] and progastrin-releasing peptide [ProGRP]), quantified by chemiluminescence immunoassay (CLIA), together with the neutrophil-to-lymphocyte ratio (NLR) and four clinical covariates, were measured in 100 participants (50 with histologically confirmed lung cancer and 50 non-cancer controls) and entered into a calibrated random-forest classifier (the VarSee™ (Lung) algorithm) to generate a diagnostic score. Diagnostic performance was assessed against a histopathological/CT reference standard using stratified five-fold cross-validation, reported in accordance with STARD 2015 and, for the multivariable model, TRIPOD.

# Results

VarSee™ (Lung) distinguished lung-cancer cases from controls with a sensitivity of 100% (95% CI 92.9–100), a specificity of 96.0% (95% CI 86.5–98.9), a positive predictive value of 96.2%, a negative predictive value of 100%, an accuracy of 98.0% (95% CI 93.0-99.4) and an area under the ROC curve of 0.997 ((95% CI 0.990–1.000). In this small cohort, sensitivity was 100% at every tumour stage (I-IV). The NLR was the dominant predictor (impurity importance 0.501), followed by CA19-9 (0.116) and CYFRA 21 − 1 (0.066); accuracy and AUC were consistent across five candidate classification algorithms (accuracy 0.950–0.980; AUC 0.995–0.997).

# Conclusions

In this pilot, a blood-based seven-protein panel integrated with the NLR discriminated lung cancer from cancer-free controls with high sensitivity and specificity, supporting further evaluation as a radiation-free adjunct for early detection. Given the small, single-centre design and cancer-free (rather than benign-

disease-matched) controls, these findings are hypothesis-generating and require confirmation in a larger, multi-centre study with external validation before any clinical or screening application is considered.

# Trial registration:

Clinical Trials Registry - India, CTRI/2026/03/105171. Retrospectively registered (enrolment began prior to registration).

## Background

Lung cancer remains one of the deadliest malignancies worldwide, accounting for approximately 18.7% of all cancer-related deaths [1]. The high case-fatality rate is driven largely by late detection, where a substantial majority of non-small-cell lung cancers (NSCLC), the predominant histological group, are diagnosed at Stage III or IV [2], reflecting the paucisymptomatic or entirely asymptomatic nature of early disease.

The National Comprehensive Cancer Network (NCCN) defines high risk on the basis of a 20-pack-year smoking history, occupational exposure and age above 50 years, and additionally recognises second-hand smoke, prior malignancy and pre-existing lung disease as risk factors [3]. A deteriorating air-quality index and emerging literature linking ambient air pollution to lung carcinogenesis place a substantial global population at risk [4].

Low-dose computed tomography (LDCT) is the only guideline-recommended screening modality for high-risk individuals, but repeated exposure to ionising radiation, limited accessibility outside tertiary centres, and a high false-positive rate restrict its use for population-scale, repeated screening [5, 6]. Beyond these clinical limitations, LDCT-based screening demands substantial capital infrastructure including dedicated CT scanners and trained radiologists, together rendering it largely unscalable in low- and middle-income countries (LMICs) and other resource-constrained settings in the screening segment. A radiation-free, non-invasive, and infrastructure-light alternative is therefore an unmet need, for which a multi-analyte, machine-learning-based blood biomarker algorithm represents a promising non-invasive candidate.

Tumour development perturbs several biological axes at once; epithelial turnover, neuroendocrine signalling, angiogenesis, metabolic reprogramming, extracellular matrix invasion, and systemic inflammation, and these perturbations are reflected in the peripheral blood as altered biomarker concentrations [7]. Circulating tumour cells (CTCs) and cell-free DNA (cfDNA) have been used to predict recurrence and treatment response, but neither is reliably enriched in the earliest disease stages, which limits their diagnostic utility for detection as distinct from monitoring [8–13]. In contrast, a growing body of evidence indicates that certain proteins are measurably elevated in blood during early-stage cancer and can be exploited for early detection, although no single circulating protein reports on more than one

or two of the biological axes above, which is why single-marker assays have generally shown early-stage sensitivities too low to support a screening claim on their own [14–16].

We hypothesised that a combination of serum protein biomarkers, spanning complementary biological axes and integrated by a purpose-built algorithm, could discriminate lung cancer from non-cancer controls with high sensitivity and specificity. The seven-protein panel evaluated here is CA125 (MUC16), CA19-9, CEA, CYFRA 21 – 1, HE4 (WFDC2), NSE and ProGRP, assembled through a structured, machine-learning-guided pipeline that screened the published literature and prioritised markers that perform well in combination rather than in isolation, as detailed in the methods section. Systemic inflammation, an axis no single protein in the panel reports on, is captured at no incremental cost by the neutrophil-to-lymphocyte ratio (NLR) derived from the same blood draw's complete blood count. We report a prospective pilot study, with biomarker measurement performed retrospectively on banked serum samples, evaluating the diagnostic accuracy of the resulting VarSee™ (Lung) score in 100 participants against a histopathological and radiological reference standard, following the STARD 2015 and TRIPOD reporting guidelines for diagnostic accuracy and multivariable prediction model studies, respectively [17–18].

## Methods

# Study design and setting

This was a consecutive, prospective, single-centre, case-control pilot diagnostic accuracy study. Patients were recruited at Shlok Superspeciality Hospital, Gujarat, India. The study was approved by the hospital's institutional ethics committee (ECR/1718/Inst/GJ/2022, protocol number SSC-BM-01-2024), Gujarat, India, and retrospectively registered with the Clinical Trials Registry, India (CTRI/2026/03/105171), as enrolment began before the registry entry was filed. Serum biomarker analysis was performed at the Dhande Pathlab Diagnostics Pvt. Ltd. laboratory, Pune, Maharashtra, India, on an NABL-accredited platform. Reporting follows the STARD 2015 statement for studies of diagnostic accuracy and, for the multivariable classifier, the TRIPOD statement for prediction-model studies; completed checklists are provided as Additional file 1 (STARD) and Additional file 2 (TRIPOD), and the participant flow diagram is shown in Fig. 2.

## Biomarker panel: discovery pipeline and biological rationale

The seven-protein VarSee™ (Lung) panel is the output of a structured, multi-stage discovery pipeline combining automated literature mining, natural-language-processing (NLP)-assisted triage, manual expert curation, and quantitative prioritisation, conducted prior to and independently of the pilot study reported here. A tailored Boolean literature search retrieved an initial corpus of 312,658 records, which an NLP pipeline (spaCy, NLTK, a fine-tuned BERT classifier) triaged to a master set of 146,380 records. Two-tier manual curation - a title/abstract review followed by full-text analysis requiring a control-versus-

cancer comparison, a reported area under the receiver-operating-characteristic curve (AUC) for a specific protein biomarker, and a blood/serum/plasma sample yielded 738 fully curated articles, of which 389 met all analytical criteria for structured data extraction.

Each biomarker study observation was assigned a composite score integrating diagnostic performance, study robustness, biomarker characteristics, and clinical relevance (weighted combination of AUC, mean sensitivity/specificity, a logistic transform of the reported p-value, normalised cohort size, biomarker type, expression direction, cancer stage, clinical implication, and sample/cohort completeness). Aggregated scores, combined with orthogonal validation against independent resources (KM Plotter, Web of Science/Scopus, FDA premarket records, NCCN/ESMO/ASCO guidelines, the Clinical Trials Registry- India, EDRN, the Human Protein Atlas and OncoMX), yielded a shortlist of 29 candidate biomarkers, technically annotated for assay-level feasibility and narrowed to 12 biomarkers with the strongest cross-study performance and robustness. The panel was restricted to the seven analytes available on an automated chemiluminescence-immunoassay (CLIA) platform to ensure reproducibility, minimal inter-operator and inter-batch variation, and clinical scalability on installed diagnostic-laboratory infrastructure. Selection favoured complementarity across biological axes over marginal stand-alone rank, so that the panel spans distinct, largely non-redundant mechanistic pathways. The principal biological role of each analyte and of the NLR is summarised in Table 1. They are provided for the biological context and are not findings of the current pilot study.

Candidate biomarkers identified through automated literature mining and NLP-assisted triage underwent two-tier manual curation, composite quantitative scoring cross-validated against orthogonal external resources, technical/assay feasibility annotation, and platform restriction, yielding the final seven-analyte panel. This discovery pipeline was conducted prior to and independently of the pilot diagnostic accuracy study reported here.

Table 1

**Biological axis and principal mechanistic role of the seven VarSee™ (Lung) panel serum biomarkers and the neutrophil-to-lymphocyte ratio (NLR) in lung tumourigenesis**

| Analyte | Biological axis | Principal mechanistic role |
|---|---|---|
| HE4 (WFDC2) | Epithelial turnover & immune modulation | Anti-apoptotic; promotes cell-cycle progression, epithelial-mesenchymal transition, angiogenic signalling and an immunosuppressive tumour microenvironment. |
| ProGRP | Neuroendocrine signalling | Precursor of gastrin-releasing peptide; mitogenic via the GRP receptor, promotes angiogenesis and EMT-driven invasion; carries most of its signal in neuroendocrine/small-cell histology. |
| CA19-9 | Glycosylation & adhesion biology | Sialylated Lewis(a) glycoprotein; alters cell-cell and cell-matrix adhesion, correlates with histological grade and impairs immune recognition via altered glycosylation. |
| CYFRA 21 – 1 | Epithelial turnover & tumour burden | Soluble cytokeratin-19 fragment released on epithelial-cell lysis; reflects tumour turnover and can influence invasion and metastasis; tracks therapy response. |
| NSE (ENO2) | Metabolic reprogramming (Warburg effect) | Glycolytic enzyme expressed by neuroendocrine cells and neuroendocrine-featured NSCLC; sustains glycolytic flux, tumour growth and motility. |
| CEA | Cell adhesion, invasion & immune evasion | Immunoglobulin-superfamily glycoprotein; promotes tumour-cell aggregation and invasion, inhibits cytotoxic T-/NK-cell activation, and supports chronic microenvironment inflammation. |
| CA125 (MUC16) | Extracellular-matrix invasion & angiogenesis | Heavily glycosylated mucin; facilitates cell adhesion, metastatic seeding, immune evasion and new-vessel formation. |
| NLR | Systemic inflammation | Neutrophil-to-lymphocyte ratio from the routine complete blood count; captures the tumour-associated systemic inflammatory response not reported by any single protein in the panel. |

Table 1. *Mechanistic roles are summarised from the biomarker discovery and prioritisation pipeline used for panel discovery and are provided for biological context.*

# Participants

Cases and controls were enrolled prospectively at Shlok Superspeciality Hospital: cases were identified and approached at the time of a new, treatment-naive lung cancer diagnosis, and controls were identified and approached at the time of volunteer presentation, with no history of malignancy. Participants were screened against the eligibility criteria, and written informed consent to participate was obtained from each participant at the time of enrolment, before any sample was collected. Recruitment started on 01-Aug-2024 and went till 31-Dec-2024.

Following enrolment, a venous blood sample was drawn from each participant and serum was separated and banked at – 80°C under controlled conditions. Biomarker measurement; the seven-analyte chemiluminescence immunoassay panel together with the complete-blood-count-derived neutrophil-to-lymphocyte ratio was not performed at the time of collection; the banked serum samples were retrieved and assayed more than one year after enrolment, with laboratory personnel blinded throughout to each participant's case or control status and to the reference-standard diagnosis.

This sequence of prospective identification and enrolment of participants and prospective collection and banking of specimens, followed by retrospective, blinded measurement of the biomarker panel on the banked specimens corresponds to a PRoBE (Prospective-specimen collection, Retrospective-Blinded-Evaluation) design, the architecture recommended for the pivotal evaluation of a candidate biomarker's diagnostic accuracy [24]. This design avoids a key source of bias seen in purely retrospective biomarker studies, in which cases and controls are assembled from archived specimens of varying age, collection protocol and storage history after the diagnosis is already known. Every specimen analysed here was collected and banked under an identical protocol, at the time of prospectively ascertained diagnosis or volunteer presentation, independently of and before any biomarker result existed.

Patients were eligible if aged over 18 years, of either sex, newly diagnosed with lung cancer confirmed by tissue biopsy and computed tomography (CT), and had not received anticancer treatment or undergone surgery prior to enrolment. Cases were excluded if of paediatric age or with a history of any prior malignancy. Non-cancer controls were eligible if aged over 18 years, of either sex, with no history of malignancy. The reference standard was histopathological examination of tissue biopsy together with CT, including tumour staging and histological subtyping. Cases and controls were verified by different reference standards; histopathological and radiological confirmation for cases, versus clinical history review alone for controls, a two-gate (differential verification) design. Participant flow through screening, enrolment, index testing, reference-standard assessment, and analysis is shown in Fig. 2. No analyte value was missing for any participant, so no median imputation was applied, and no sample was returned as non-reportable for a missing complete blood count.

## Index test: sample collection, handling, quality control and biomarker measurement

Following serum separation, samples were stored at – 80°C and transported to the BioMarkIQ laboratory under controlled conditions. Haemolysed or turbid samples were excluded on physical quality control. Thawed serum samples were centrifuged at 4000 rpm for 10 minutes and loaded onto a MAGLUMI X3 fully automated CLIA analyser (SNIBE Co., Ltd., Shenzhen, China), with calibration and quality-control checks performed before each analytical run at Dhande Pathlab Diagnostics Pvt. Ltd. Measured biomarker concentrations were exported from the analyser platform and, together with the participant's complete-blood-count-derived NLR and clinical covariates (age, sex, smoking and alcohol history), entered into the VarSee™ (Lung) analysis pipeline (the index test) for automated score generation.

Laboratory personnel generating index-test results were not informed of the reference-standard diagnosis. Because biomarker measurement occurred more than one year after reference-standard determination, clinicians and pathologists establishing the reference-standard diagnosis had no knowledge of, and could not have been influenced by, the VarSee™ (Lung) index-test result, which did not yet exist at that time.

## The VarSee™ (Lung) algorithm and score generation

The model architecture, hyperparameters and the 0.50 probability threshold were specified before analysis and were not tuned on the cohort. The classifier was developed on the present 100-participant cohort, and all performance reported here is therefore internal. Following development, the artefact was locked and version-controlled at release, with a fixed random seed, so that identical inputs yield identical outputs in subsequent use. The model returns a calibrated probability of lung cancer; a positive VarSee™ (Lung) result is declared at a pre-specified probability threshold of 0.50, mapped for reporting onto a 0-100 VarSee™ (Lung) score. Missing analyte values are median-imputed; the NLR is a required input and is never imputed, so a sample lacking a valid complete blood count is returned as non-reportable rather than scored. Model discrimination was estimated internally using stratified five-fold cross-validation, with performance reported on out-of-fold predictions (Fig. 3).

To assess whether diagnostic performance depended on the specific learning algorithm chosen, four alternative classifiers - logistic regression, an uncalibrated random forest, gradient boosting, and a support-vector machine with radial basis function kernel (SVM-RBF) were trained and evaluated under an identical stratified five-fold cross-validation scheme on the same 100-participant cohort.

# Statistical analysis

Continuous variables are presented as mean ± standard deviation (SD) or, for biomarker concentrations, as medians, given right-skewed distributions on inspection of log-transformed histograms. Between-group differences in baseline characteristics were assessed using the two-sided Mann–Whitney U test for age and Fisher's exact test for categorical variables. Age is summarised as median and interquartile range in addition to mean and standard deviation, because the control age distribution was left-skewed. Diagnostic performance of the VarSee™ (Lung) (sensitivity, specificity, positive and negative predictive values, accuracy, and AUC) was estimated under stratified five-fold cross-validation and summarised on out-of-fold predictions, with 95% confidence intervals computed using the Wilson score method for binomial proportions and, for the area under the ROC curve, by bootstrap resampling of the out-of-fold predictions over 2,000 replicates. Feature contributions were characterised using random-forest impurity-based (mean decrease in impurity) importance across the twelve model inputs. No formal sample-size calculation was performed for this pilot; the target enrolment of 50 cases and 50 controls was set pragmatically to obtain an initial estimate of diagnostic accuracy ahead of a fully powered validation study. Analyses were performed in Python 3 using scikit-learn and SciPy.

## Results

# Participant characteristics

One hundred participants were analysed: 50 with histologically confirmed lung cancer and 50 non-cancer controls (Table 2). Cases were older than controls (mean 62.6 ± 7.4 vs. 57.8 ± 12.5 years), with a comparable sex distribution (cases 27 male/23 female; controls 28 male/22 female)One hundred participants were analysed: 50 with histologically confirmed lung cancer and 50 non-cancer controls (Table 2). The two arms were comparable in age (median 63.5 vs. 60.5 years) and sex distribution (cases 27 male/23 female; controls 28 male/22 female). Smoking (54% vs. 44%) and alcohol use (68% vs. 26%) were more frequent among cases. Alcohol use was more frequent among cases (68% vs. 26%; $p < 0.001$), whereas smoking prevalence (54% vs. 44%; $p = 0.42$), sex distribution ($p = 1.00$) and age (median 63.5 vs. 60.5 years; $p = 0.10$) did not differ significantly. Controls nevertheless spanned a wider age range than cases (22–73 against 49–74 years), including individuals below the age at which lung-cancer screening would be considered; as age was a model input, this difference may have contributed to the observed separation independently of biomarker signal. Cases spanned all disease stages (Stage I, n = 7; II, n = 19; III, n = 21; IV, n = 3) and three non-small-cell histological subtypes (adenocarcinoma, n = 25; large-cell carcinoma, n = 14; squamous-cell carcinoma, n = 11).

Table 2
Baseline demographic and clinical characteristics of the pilot cohort, by group

| Characteristic | Lung cancer (n = 50) | Control (n = 50) | p-value |
|---|---|---|---|
| Age, years (median, IQR) | 63.5 (56–70) | 60.5 (54–66) | 0.100 |
| Male / Female | 27 / 23 | 28 / 22 | 1.000 |
| Current smokers, n (%) | 27 (54%) | 22 (44%) | 0.424 |
| Current alcohol use, n (%) | 34 (68%) | 13 (26%) | < 0.001 |
| Stage I | 7 | - | |
| Stage II | 19 | - | |
| Stage III | 21 | - | |
| Stage IV | 3 | - | |
| Adenocarcinoma | 25 | - | |
| Large-cell carcinoma | 14 | - | |
| Squamous-cell carcinoma | 11 | - | |

Table 2. *SD: standard deviation; IQR: interquartile range. Stage and histological subtype were not applicable to controls. p-values are two-sided Mann–Whitney U for age and Fisher's exact test for categorical variables.*

# Between-group differences in serum biomarkers and the NLR

Four of the eight quantitative inputs differed significantly between lung-cancer cases and controls on the two-sided Mann-Whitney U test: the NLR (median 3.00 vs. 2.00; fold-change 1.50; $p < 0.0001$), CA19-9 (median 12.55 vs. 31.85 U/mL; fold-change 0.39; $p = 0.0003$), CYFRA 21 − 1 (median 5.53 vs. 3.35 ng/mL; fold-change 1.65; $p = 0.0069$) and ProGRP (median 50.35 vs. 35.45 pg/mL; fold-change 1.42; $p = 0.0182$) (Table 3). Under a Bonferroni-corrected threshold for eight comparisons ($\alpha = 0.00625$), the NLR and CA19-9 remained significant while CYFRA 21 − 1 and ProGRP did not.

NSE, HE4, CEA, and CA125 did not separate the two groups significantly at the univariate level in this cohort. Several of these analytes were selected during panel discovery for a conditional or subgroup-specific contribution; for example, ProGRP and NSE are the panel's neuroendocrine markers, whereas the pilot cohort's histological mix (adenocarcinoma, large-cell and squamous-cell carcinoma only, with no small-cell disease) under-represents the subpopulation in which these markers are expected to carry most of their signal. Detecting a conditional or interaction effect requires substantially more statistical power than a cohort of 50 cases per arm provides.

Within the lung-cancer cases, the epithelial-lineage markers correlated with one another (Spearman ρ CEA-CYFRA 21 − 1, 0.65; CA125-CEA, 0.63; CA125-CYFRA 21 − 1, 0.56), consistent with a shared biological axis rather than seven independent signals, whereas ProGRP anticorrelated with this cluster ($\rho = -0.44$ with CYFRA 21 − 1; $\rho = -0.43$ with the NLR), the pattern expected of a neuroendocrine analyte in a predominantly non-small-cell cohort.

Table 3
Serum biomarker and NLR concentrations in lung-cancer cases versus non-cancer controls (n = 100)

| Analyte(unit) | Lung cancer, median | Control, median | Fold-change | p-value | Significance |
|---|---|---|---|---|---|
| NLR | 3.00 | 2.00 | 1.50 | < 0.0001 | *** |
| CA19-9(U/mL) | 12.55 | 31.85 | 0.39 | 0.0003 | *** |
| CYFRA 21 − 1(ng/mL) | 5.53 | 3.35 | 1.65 | 0.0069 | ** |
| ProGRP(pg/mL) | 50.35 | 35.45 | 1.42 | 0.0182 | * |
| NSE(ng/mL) | 11.45 | 13.00 | 0.88 | 0.3092 | ns |
| HE4(pmol/L) | 184.00 | 190.00 | 0.97 | 0.5955 | ns |
| CEA(ng/mL) | 6.77 | 5.44 | 1.24 | 0.7638 | ns |
| CA125(U/mL) | 23.40 | 22.85 | 1.02 | 0.8605 | ns |

Table 3 *Concentrations are group medians in assay-specific units; the NLR is dimensionless. p-values are two-sided Mann-Whitney U on log(1 + concentration)-transformed values, uncorrected for multiplicity (Bonferroni-corrected threshold for eight comparisons: α = 0.00625). Significance: ***p < 0.001; **p < 0.01; *p < 0.05; ns not significant.*

# Diagnostic accuracy of the VarSee™ (Lung) score

Under stratified five-fold cross-validation, the VarSee™ (Lung) score correctly classified 98 of 100 participants, detecting all 50 lung-cancer cases (sensitivity 100%, 95% CI 92.9–100) and 48 of 50 controls (specificity 96.0%, 95% CI 86.5–98.9), with a positive predictive value of 96.2% (95% CI 87.0-98.9), a negative predictive value of 100% (95% CI 92.6–100), an accuracy of 98.0% (95% CI 93.0-99.4), and an AUC of 0.997 (95% CI 0.990-1.000) (Table 4, Figs. 4 and 5).

Table 4
Diagnostic performance of the VarSee™ (Lung) score at the pre-specified 0.50 probability threshold (n = 100, out-of-fold, five-fold cross-validation)

| Metric | Value | 95% CI | Definition |
|---|---|---|---|
| Sensitivity (recall) | 100.0% | 92.9–100 | TP / (TP + FN) |
| Specificity | 96.0% | 86.5–98.9 | TN / (TN + FP) |
| Positive predictive value | 96.2% | 87.0-98.9 | TP / (TP + FP) |
| Negative predictive value | 100.0% | 92.6–100 | TN / (TN + FN) |
| Accuracy | 98.0% | 93.0-99.4 | (TP + TN) / N |
| ROC-AUC | 0.997 | 0.990-1.000 | Area under the ROC curve |
| Confusion counts (TN, FP, FN, TP) | 48, 2, 0, 50 | - | n = 100, out-of-fold |

Table 4. *95% confidence intervals are Wilson score intervals for binomial proportions. TP: true positive; TN: true negative; FP: false positive; FN: false negative.*

# Performance across tumour stage

Sensitivity was 100% at every tumour stage represented in the cohort: Stage I (7/7), Stage II (19/19), Stage III (21/21) and Stage IV (3/3) (Fig. 6). Because the cohort contained only seven Stage I cases, this internal estimate is dominated by established (Stage II-IV) disease, and the confidence interval around early-stage sensitivity specifically remains wide.

# Feature importance and robustness to classifier choice

Random-forest impurity-based importance across the twelve model inputs was dominated by the NLR (0.501), followed by CA19-9 (0.116), CYFRA 21 − 1 (0.066), ProGRP (0.061), CA125 (0.057), age (0.055), HE4 (0.045), alcohol history (0.039), CEA (0.029), NSE (0.020), smoking history (0.006) and sex (0.004)

(Fig. 7). Because impurity-based importance favours continuous, high-cardinality inputs such as the NLR, this ranking should be read as the order in which the model preferentially splits on each input rather than as an unbiased measure of independent predictive contribution.

Four alternative classifiers were trained and evaluated under the identical five-fold cross-validation protocol (Table 5). Accuracy ranged from 0.950 (logistic regression) to 0.980 (random forest, gradient boosting and the deployed calibrated random forest), and AUC ranged from 0.995 to 0.997, indicating that the discriminating signal resides substantially in the panel and covariate set itself rather than in the choice of classifier.

Table 5
Cross-validated performance of five candidate classification algorithms trained on the identical twelve-input feature set (n = 100, five-fold cross-validation)

| Model | Accuracy | F1 score | ROC-AUC |
|---|---|---|---|
| Logistic regression | 0.950 | 0.950 | 0.995 |
| Random forest (uncalibrated) | 0.980 | 0.980 | 0.997 |
| Gradient boosting | 0.980 | 0.980 | 0.996 |
| Support-vector machine (RBF kernel) | 0.960 | 0.958 | 0.995 |
| Calibrated random forest (deployed VarSee™ (Lung) model) | 0.980 | 0.980 | 0.997 |

Table 5. *All models used the identical twelve-input feature set and stratified five-fold cross-validation scheme*

# Discussion

In this pilot study, a blood-based seven-protein panel integrated with the NLR for VarSee™ (Lung) algorithm discriminated lung cancer from non-cancer controls with high sensitivity and specificity. Because the test requires only a single peripheral blood draw and involves no ionising radiation, it could complement or extend LDCT-based screening, particularly where CT access is limited or repeated imaging is undesirable. Several panel members; CEA, CYFRA 21 – 1, NSE, ProGRP and CA19-9 are already in clinical use as individual markers; the value proposition of VarSee™ (Lung) lies in their algorithmic combination rather than in any single constituent.

The observed sensitivity and specificity numerically overlap with reported LDCT operating characteristics (sensitivity 80–100%, specificity > 75%) [5]; however, this is not a head-to-head comparison, as the LDCT figures derive from a prospective, externally validated screening population with a much lower cancer prevalence, whereas the present estimates are internal, cross-validated results from a balanced case-control cohort. The pilot cohort also provided 50 cancer events against twelve candidate model inputs - approximately 4.2 events per variable, below the conventional ten-events-per-variable heuristic for stable multivariable model fitting, so these internal cross-validated estimates, while genuinely out-of-fold,

should be treated as supportive rather than sufficient evidence of generalisable performance, consistent with TRIPOD guidance on interpreting model performance at low events-per-variable ratios.

Only four of the eight quantitative inputs (NLR, CA19-9, CYFRA 21 – 1 and ProGRP) differed significantly between groups at the univariate level, and only two (NLR and CA19-9) survived correction for multiple comparisons. This is consistent with a panel whose discriminative value is a property of the combined, algorithmically weighted signal rather than of any constituent analyte considered alone, but it also means the univariate evidence for four panel members remains limited in this cohort.

CA19-9 was the second most influential model input, yet its concentration was lower in cases than in controls (fold-change 0.39), which is the reverse of the direction generally reported in NSCLC. Two explanations are plausible and cannot be separated in a cohort of this size. The first is control-group composition: CA19-9 is elevated in benign hepatic, pancreatic and gastrointestinal conditions, so controls drawn from a general population may carry a higher CA19-9 distribution than cases without any tumour biology being involved [19–20]. The second is sampling variation at 50 participants per arm. Because the model draws substantial weight from this input, the direction and effect size of CA19-9 require confirmation in the planned validation cohort, and the marker's contribution should be regarded as provisional until then.

Several panel members are not specific to lung cancer and have been reported to be elevated in other malignancies and in benign inflammatory or pulmonary conditions, including chronic obstructive pulmonary disease, interstitial lung disease, pneumonia and tuberculosis [21–23]. The control arm in this pilot comprised cancer-free volunteers rather than patients with benign respiratory disease; consequently, the specificity reported here should be regarded as an upper bound rather than representative of performance in the intended screening population, where benign pulmonary disease is a more clinically relevant comparator. Evaluating VarSee™ (Lung) against benign-disease controls, and within a population enriched for the risk factors that motivate screening, is accordingly an essential next step before any claim of clinical specificity can be made.

This study has further limitations. It was a single-centre pilot with a small sample (50 cases, 50 controls), limiting precision and generalisability; the stage distribution was weighted toward established disease (43 of 50 cases were Stage II-IV), so early-stage performance, while numerically 100% in the seven Stage I cases observed, carries a wide confidence interval; and all reported performance is internal (cross-validated on the fitting cohort) rather than externally validated on an independent sample, which STARD and TRIPOD both identify as necessary before a diagnostic or prognostic claim can be considered established. Of the baseline characteristics, only alcohol use differed significantly between arms (68% against 26%; $p < 0.001$); age, sex and smoking prevalence did not. While these covariates were included as model inputs, residual confounding cannot be excluded at this sample size. The relative importance assigned to the clinical covariates warrants caution. Self-reported alcohol history carried greater importance than smoking history, which is inconsistent with the established aetiology of lung cancer and more plausibly reflects the difference in alcohol-use prevalence between the two arms

(68% of cases against 26% of controls) than a biological relationship. Because impurity-based importance rewards inputs that separate the arms irrespective of mechanism, this ordering should be read as a feature of the present cohort rather than a generalisable finding, and the contribution of the lifestyle covariates requires re-estimation in a cohort where the arms are better matched. The two arms were also verified against different reference standards; histopathology and imaging for cases versus clinical history review alone for controls, a differential-verification design that likely inflates apparent specificity and accuracy relative to a study in which both arms receive the same reference standard.

These findings motivate a larger, multi-centre, prospectively registered validation study, enriched for early-stage disease and incorporating benign pulmonary disease controls, with a pre-specified, locked model and an independent test set, before VarSee™ (Lung) can be considered for any clinical or screening role.

## Conclusions

A blood-based, seven-protein serum panel combined with the neutrophil-to-lymphocyte ratio, integrated by a calibrated machine-learning classifier (VarSee™ (Lung)), discriminated histologically confirmed lung cancer from non-cancer controls with a sensitivity of 100% and a specificity of 96.0% in a 100-participant pilot study, with performance robust across five candidate classification algorithms. These pilot findings support further investigation of VarSee™ (Lung) as a radiation-free, blood-based adjunct for lung-cancer detection, but they derive from a small, single-centre cohort with cancer-free rather than disease-matched controls and no external validation, and must be confirmed in a larger, multi-centre, prospectively registered study before any diagnostic or screening claim can be supported.

## Abbreviations

| Abbreviation | Definition |
|---|---|
| AUC | Area under the (receiver-operating-characteristic) curve |
| CA125 | Cancer antigen 125 (mucin 16, MUC16) |
| CA19-9 | Carbohydrate antigen 19-9 |
| CEA | Carcinoembryonic antigen |
| CI | Confidence interval |
| CLIA | Chemiluminescence immunoassay |
| cfDNA | Cell-free DNA |
| CTC | Circulating tumour cell |
| CTRI | Clinical Trials Registry – India |
| CYFRA 21-1 | Cytokeratin-19 fragment |
| HE4 | Human epididymis protein 4 (WFDC2) |
| LDCT | Low-dose computed tomography |
| NCCN | National Comprehensive Cancer Network |
| NLR | Neutrophil-to-lymphocyte ratio |
| NPV | Negative predictive value |
| NSCLC | Non-small-cell lung cancer |
| NSE | Neuron-specific enolase |
| PPV | Positive predictive value |
| ProGRP | Progastrin-releasing peptide |
| ROC | Receiver operating characteristic |
| SD | Standard deviation |
| STARD | Standards for Reporting of Diagnostic Accuracy Studies |
| TRIPOD | Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis |

# Declarations

### Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Ethics Committee of Shlok Superspeciality Hospital. The study was registered with the

Clinical Trials Registry - India (CTRI/2026/03/105171); registration was retrospective, as enrolment began before the trial was registered. Written informed consent to participate was obtained from all participants at enrolment.

**Consent for publication**

Not applicable. This manuscript does not contain any individual person's identifiable data, images, or other personal or clinical details.

# Competing Interests

J.K., K.B., K.G., A.J., M.K. and S.S. are employed by BioMarkIQ Scientific Technologies Pvt. Ltd., which develops and markets the VarSee™ (Lung) platform evaluated in this article. BioMarkIQ holds a pending Indian patent application (Application No. 202421079476, "Machine Learning Method for Curating Biomarker Panel, for Screening of Cancer(s) and System Therefor") covering the biomarker-panel curation method described in the Methods.

# Additional files

Additional file 1. STARD 2015 checklist (provided as a separate document).

Additional file 2. TRIPOD checklist (provided as a separate document).

# Funding

BioMarkIQ Scientific Technologies is funded under the Startup India Seed Fund Scheme by the Department for Promotion of Industry and Internal Trade and NIDHI-PRAYAS and NIDHI-SSS under the Department of Science & Technology, India. It is registered under DPIIT.

# Author Contribution

S.S., M.K. & J.K. conceived and designed the study. J.K., K.B. and K.G. developed the methodology. A.J. developed the analysis software. K.G. and S.S. performed validation. J.K. and A.J. performed the formal analysis. A.D. supervised and ran the CLIA experiments in his laboratory. K.B., K.G., A.J. and S.S. conducted the investigation and data collection. M.K. provided resources and supervision. K.G. and M.K. curated the data. S.S. wrote the original draft. All authors reviewed and edited the manuscript. A.J. prepared the figures. S.S. administered the project. All authors read and approved the final manuscript.

# Acknowledgement

The authors thank the clinical, nursing, and laboratory staff of Shlok Superspeciality Hospital, Gujarat, and of Dhande Pathlab Diagnostics Pvt. Ltd, Pune, for their contribution to participant recruitment, sample collection, and biomarker analysis. The funding support for this study was provided by the Startup India Seed Fund Scheme, Government of India.

# Data Availability

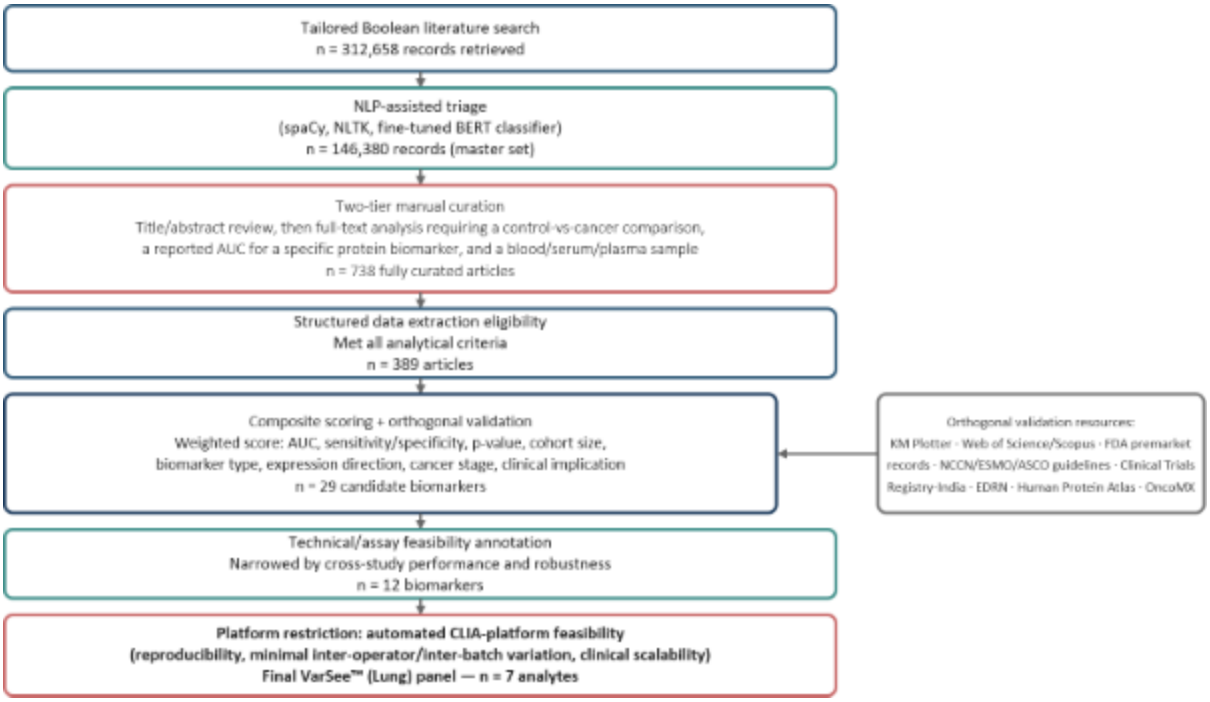
The datasets analysed during the current study are available from the corresponding author on reasonable request, subject to institutional and privacy restrictions governing clinical patient data.

## References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74:229–63.
2. Siegel RL, Kratzer TB, Wagle NS, Sung H, Jemal A. Cancer statistics, 2026. CA Cancer J Clin. 2026. 10.3322/caac.70043.
3. Wood DE, Kazerooni EA, Aberle DR et al. NCCN Guidelines® Insights: Lung Cancer Screening, Version 1.2025. J Natl Compr Canc Netw. 2025;23(1). 10.6004/jnccn.2025.0002. PMID: 39819601.
4. Hill W, Lim EL, Weeden CE, et al. Lung adenocarcinoma promotion by air pollutants. Nature. 2023;616:159–67. 10.1038/s41586-023-05874-3.
5. National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, et al. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. N Engl J Med. 2011;365(5):395–409.
6. Lancaster HL, Heuvelmans MA, Oudkerk M. Low-dose computed tomography lung cancer screening: Clinical evidence and implementation research. J Intern Med. 2022;292(1):68–80.
7. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144(5):646–74.
8. Bettegowda C, Sausen M, Leary RJ, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med. 2014;6(224):224ra24.
9. Cohen JD, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test (CancerSEEK). Science. 2018;359(6378):926–30.
10. Cristofanilli M, Budd GT, Ellis MJ, et al. Circulating tumor cells, disease progression, and survival in metastatic breast cancer. N Engl J Med. 2004;351(8):781–91.
11. Garcia-Murillas I, Schiavon G, Weigelt B, et al. Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. Sci Transl Med. 2015;7(302):302ra133.
12. Abbosh C, Birkbak NJ, Wilson GA, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature. 2017;545(7655):446–51.

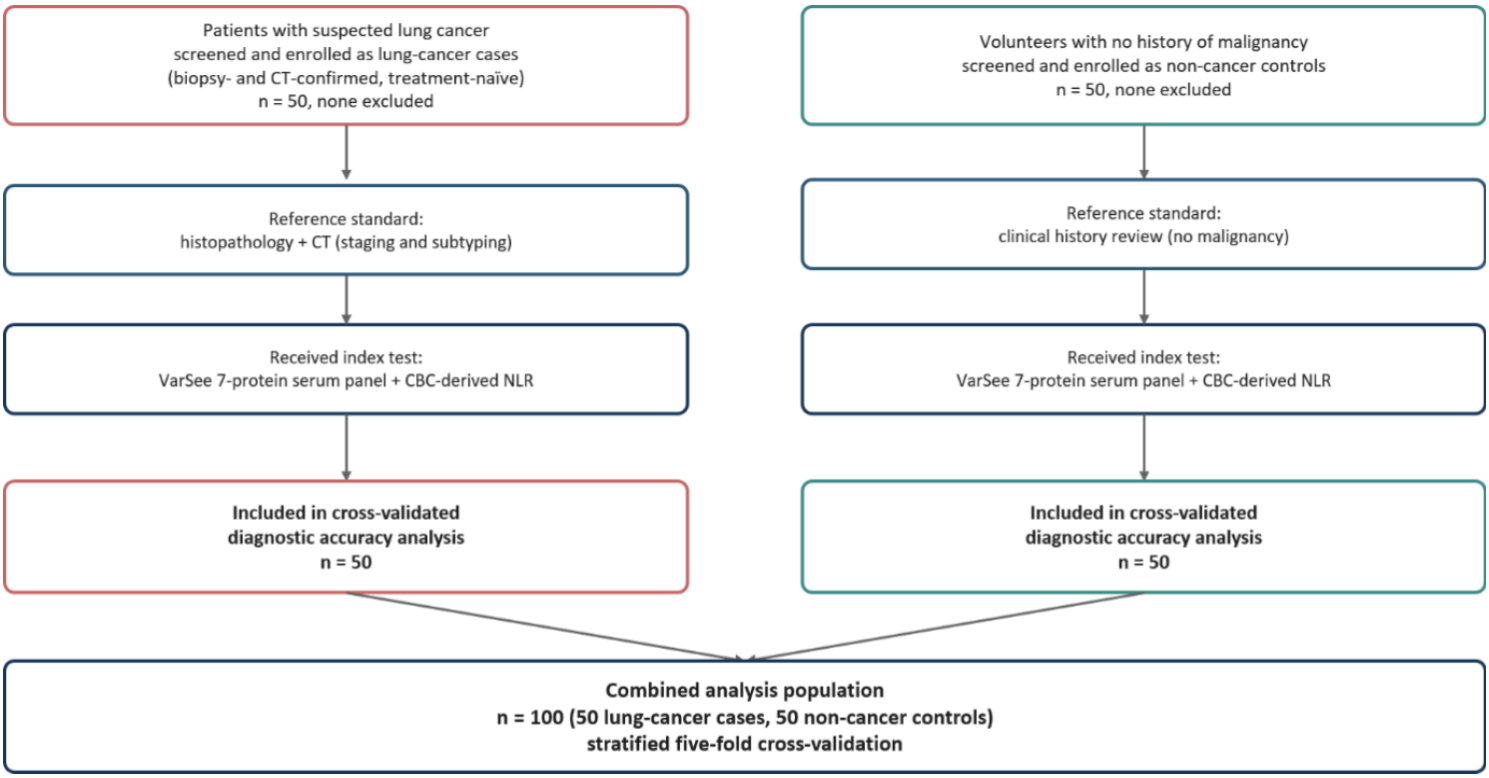
13. Alix-Panabières C, Pantel K. Clinical applications of circulating tumor cells and circulating tumor DNA as liquid biopsy. Cancer Discov. 2016;6(5):479–91.
14. Papier K, Atkins JR, Tong TYN, et al. Identifying proteomic risk factors for cancer using prospective and exome analyses of 1463 circulating proteins and risk of 19 cancers in the UK Biobank. Nat Commun. 2024;15:4010. 10.1038/s41467-024-48017-6.
15. Carrasco-Zanini J, Pietzner M, Davitte J, et al. Proteomic signatures improve risk prediction for common and rare diseases. Nat Med. 2024;30:2489–98.
16. Davies MPA, Sato T, Ashoor H, et al. Plasma protein biomarkers for early prediction of lung cancer. EBioMedicine. 2023;93. 10.1016/j.ebiom.2023.104686.
17. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351. 10.1136/BMJ.H5527.
18. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350. 10.1136/BMJ.G7594.
19. Kim S, Park BK, Seo JH, Choi J, Choi JW, Lee CK, Chung JB, Park Y, Kim DW. Carbohydrate antigen 19 – 9 elevation without evidence of malignant or pancreatobiliary diseases. Sci Rep. 2020;10(1):8820.
20. Kim K, Yong SH, Lee SH, Lee SH, Leem AY, Kim SY, Chung K, Kim EY, Jung JY, Park MS, Kim YS, Lee HJ, Kang YA. Correlation between serum carbohydrate antigen 19 – 9 levels and computed tomography severity score in patients with nontuberculous mycobacterial pulmonary disease. Sci Rep. 2021;11(1):2777.
21. Yang Y, Xu M, Huang H et al. Serum carcinoembryonic antigen elevation in benign lung diseases. Scientific Reports 2021 11:1. 2021;11:19044-. 10.1038/s41598-021-98513-8
22. Kwon BS, Kim ES, Lim SY et al. The significance of elevated tumor markers among patients with interstitial lung diseases. Scientific Reports 2022 12:1. 2022;12:16702.
23. Zhao P, Yu Q, Zhang A, et al. Serum CA-125 for the diagnosis of pulmonary tuberculosis: a systematic review and meta-analysis. BMC Infect Dis. 2021;21:1091.
24. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Cancer Inst. 2008;100(20):1432–8. 10.1093/jnci/djn326.

# Figures




## Figure 1

*Biomarker discovery and prioritisation pipeline for the VarSee™ (Lung) panel.*




## Figure 2

*STARD participant flow diagram for the pilot study (n = 100)*

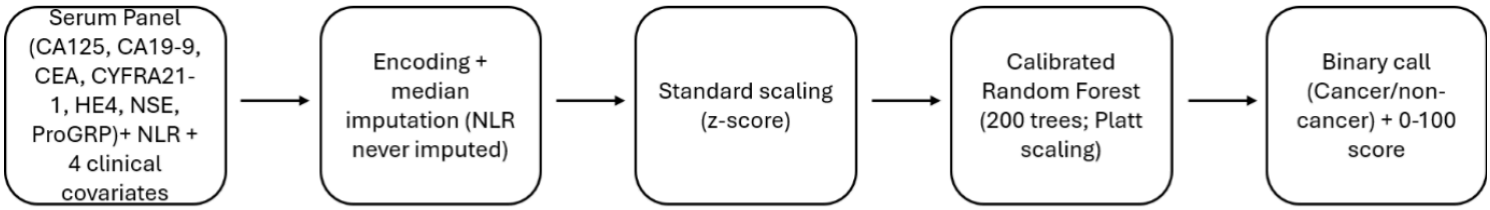



**Figure 3**

*End-to-end pipeline of the locked VarSee™ (Lung) inference model. Serum-marker concentrations, the CBC-derived neutrophil-to-lymphocyte ratio, and four clinical covariates are encoded, median-imputed where applicable (except the NLR), standard-scaled, and passed to a calibrated random-forest classifier that returns a binary diagnosis and a 0-100 VarSee™ (Lung) score.*

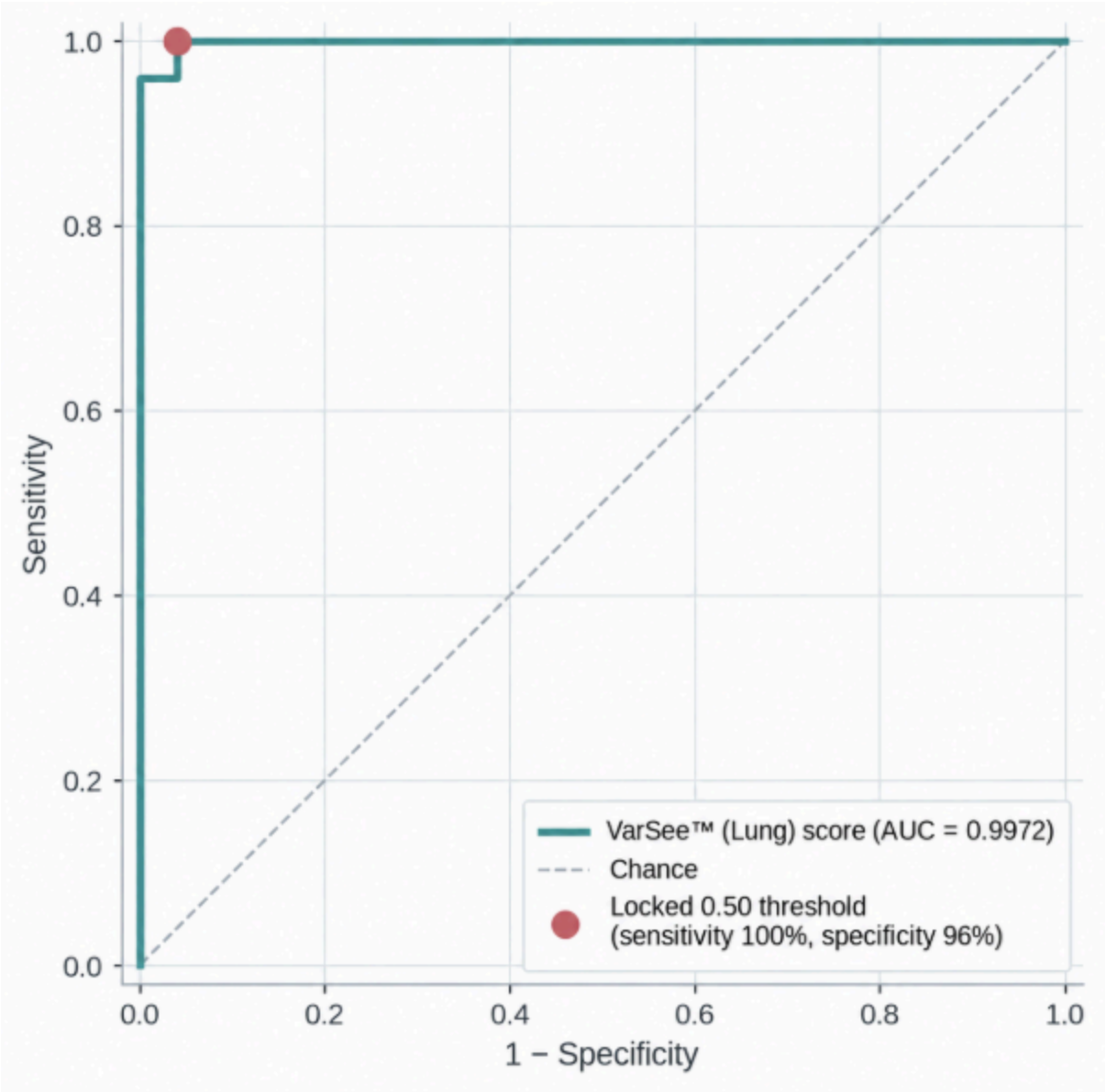



**Figure 4**

*Receiver-operating-characteristic curve for the VarSee™ (Lung) score under stratified five-fold cross-validation (n = 100).*

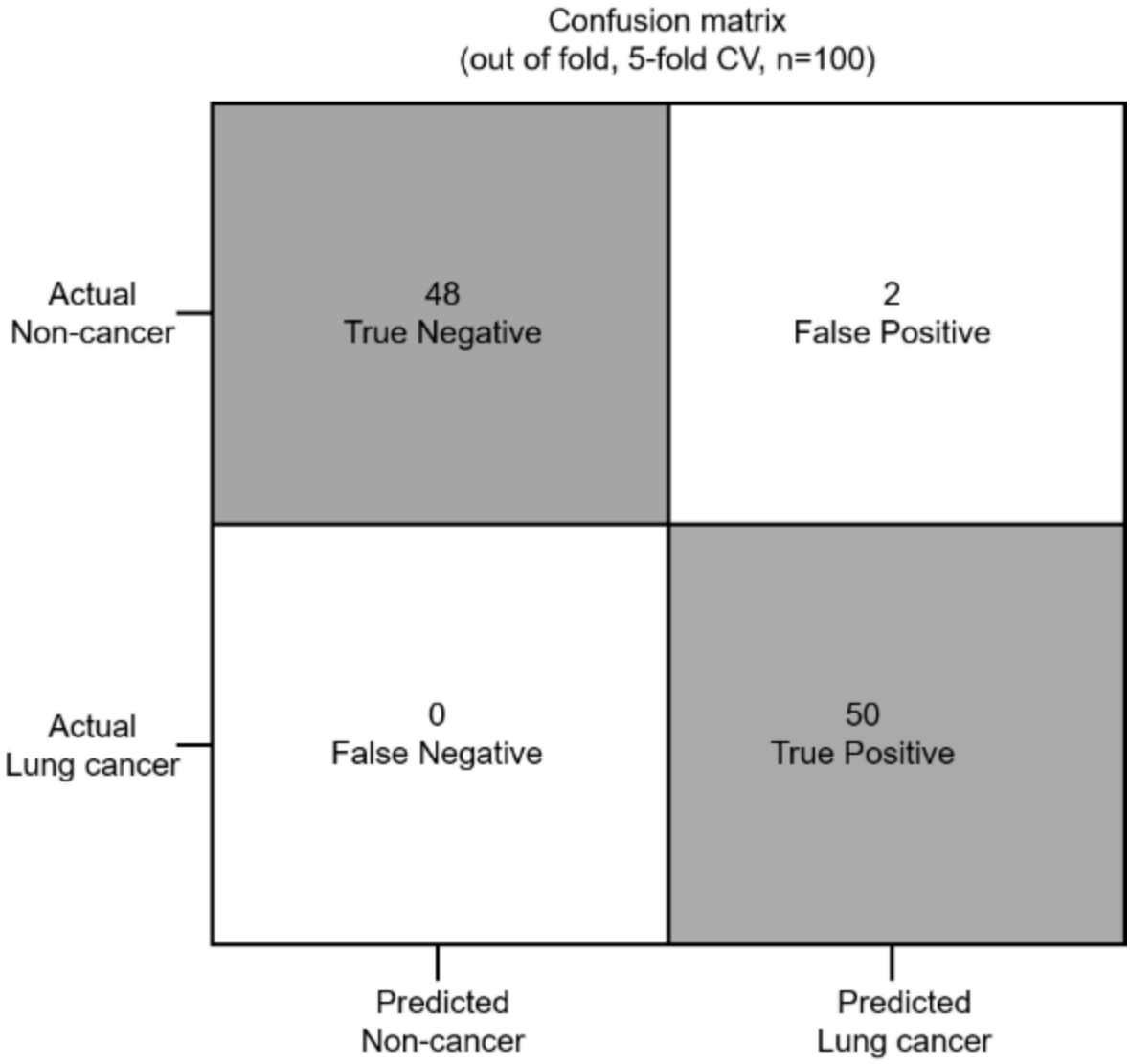



**Figure 5**

*Confusion matrix of out-of-fold VarSee™ (Lung) classifications at the locked 0.50 threshold (n = 100; five-fold cross-validation).*

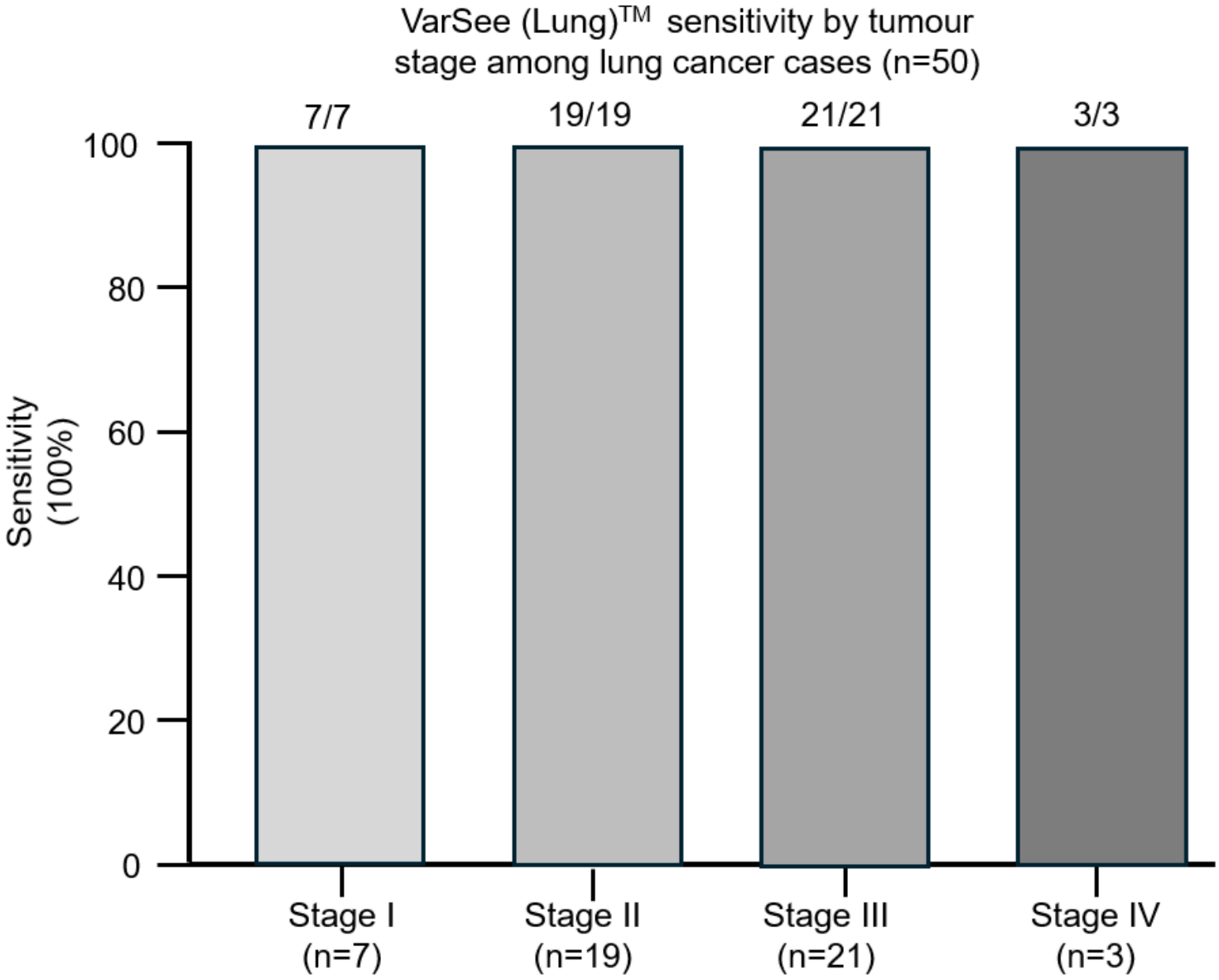



**Figure 6**

*VarSee™ (Lung) sensitivity stratified by tumour stage among the 50 lung-cancer cases.*

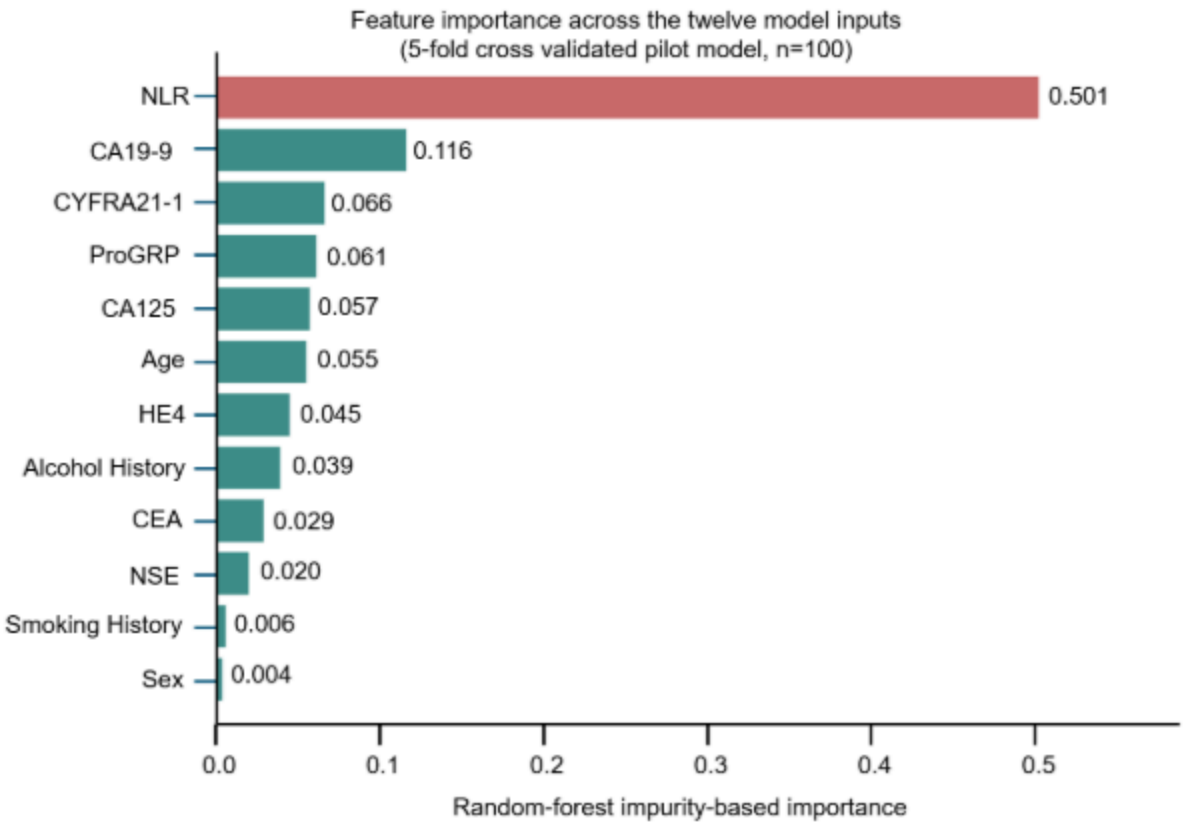



**Figure 7**

*VarSee™ (Lung) sensitivity stratified by tumour stage among the 50 lung-cancer cases.*

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.docx