



Development and internal validation of a serum biomarker signature for non-small cell lung cancer: an exploratory case-control study

Angelina V. Zhilenkova^{1,✉}, Polina S. Shilo², Ikenna K. Uchendu^{1,3}, Ekaterina V. Orlova¹, Vladislav S. Skossyrskiy¹, Julia A. Iasnova¹, Nikolay N. Bagmet⁴, Nikolay A. Agafonov¹, Natalya M. Kashakanova¹, Marina I. Sekacheva¹

¹ Sechenov First Moscow State Medical University (Sechenov University)
8/2, Trubetskaya str., Moscow, 119048, Russia

² Lahta Clinic LLC

50, Dibunovskaya str., Saint Petersburg, 197183, Russia

³ University of Nigeria

Enugu Campus, Ademola str., Ogui, Enugu, Enugu State, 401105, Nigeria

⁴ Russian Scientific Center of Surgery named after academician B.V. Petrovsky
2, Abrikosovsky lane, Moscow, 119435, Russia

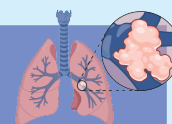
SECHENOV
MEDICAL JOURNAL
GRAPHICAL ABSTRACT



Development and internal validation of a serum biomarker signature for non-small cell lung cancer: an exploratory case-control study

Summary

A diagnostic model incorporating HE4, ApoA4, and age showed high discriminatory ability for non-small cell lung cancer; the associations of HE4 and ApoA4 with NSCLC remained significant after age matching.

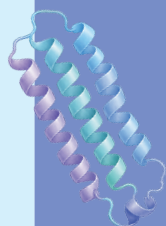


Materials and methods

Retrospective case-control study with prospective sample supplementation



Outcomes



Full cohort

Predictors, association

HE4	positive
ApoA4	inverse
Age	positive

Model performance

Final model:
HE4 + ApoA4 + age
AUC = 0.975
95% CI 0.958–0.992

After age matching

Predictors, association

HE4	retained
ApoA4	retained
Age	n.s.

Model performance

Model 1: HE4 + ApoA4
AUC = 0.961
Model 2: HE4 + ApoA4 + age
AUC = 0.962

Zhilenkova A.V., Shilo P.S., Uchendu I.K., et al. Development and internal validation of a serum biomarker signature for non-small cell lung cancer: an exploratory case-control study. Sechenov Medical Journal. 2026; 17(2): 17–32. <https://doi.org/10.47093/2218-7332.2026.17.2.17-32>

20 minutes
to read



ApoA4 – apolipoprotein A4, HE4 – human epididymis protein 4, n.s. – not significant, NSCLC – non-small cell lung cancer

Abstract

Blood-based biomarkers are increasingly explored for their ability to distinguish non-small cell lung cancer (NSCLC) from non-malignant conditions. Single markers, however, often show limited performance, which has led to growing interest in combining multiple biological features.

Aim. To develop a multivariable diagnostic model based on serum biomarkers and assess its discriminatory performance in an exploratory TRIPOD Type 1a study.

© Zhilenkova A.V., Shilo P.S., Uchendu I.K., Orlova E.V., Skossyrskiy V.S., Iasnova J.A., Bagmet N.N., Agafonov N.A., Kashakanova N.M., Sekacheva M.I., 2026

Materials and methods. This study represents a secondary analysis of a previously published case–control cohort, focusing on multivariable model development rather than single biomarker assessment. A total of 275 participants were enrolled in this retrospective case–control study supplemented by prospectively collected samples, including 85 patients with histologically confirmed NSCLC and 190 controls. Serum concentrations of 20 biomarkers representing tumor-associated, inflammatory, and metabolic pathways were analyzed. Univariable logistic regression was performed, followed by multivariable modeling including variables with $p < 0.10$. Model discrimination was evaluated using receiver operating characteristic analysis and the area under the curve (AUC), and internal validation was performed by bootstrap resampling on complete-case data.

Results. Three independent variables were retained in the final multivariable model: human epididymis protein 4 (HE4), apolipoprotein A4 (ApoA4), and age. The final model showed high apparent discrimination within the study dataset (AUC = 0.975; 95% CI: 0.958–0.992). Bootstrap internal validation demonstrated minimal optimism, with an optimism-corrected AUC of 0.972. HE4 and age were positively associated with NSCLC, whereas ApoA4 showed an inverse association. In the age-matched sensitivity analysis, HE4 and ApoA4 retained essentially unchanged associations while age lost independent significance, supporting a biomarker-driven rather than age-driven contribution to model discrimination. The multivariable model demonstrated improved discrimination compared with individual biomarkers.

Conclusion. This exploratory model derivation study identified a candidate serum biomarker combination associated with NSCLC. The findings should be considered hypothesis-generating. Independent validation in clinically representative populations is required before any clinical interpretation or application.

Keywords: early detection of lung cancer; human epididymis protein 4; apolipoprotein A4; multivariable analysis; logistic regression; ROC curve

MeSH terms:

CARCINOMA, NON-SMALL-CELL LUNG – DIAGNOSIS

CARCINOMA, NON-SMALL-CELL LUNG – PATHOLOGY

BIOMARKERS, TUMOR – ANALYSIS

HISTOLOGICAL TECHNIQUES

For citation: Zhilenkova A.V., Shilo P.S., Uchendu I.K., Orlova E.V., Skossyrskiy V.S., Iasnova J.A., Bagmet N.N., Agafonov N.A., Kashakanova N.M., Sekacheva M.I. Development and internal validation of a serum biomarker signature for non-small cell lung cancer: an exploratory case–control study. *Sechenov Medical Journal*. 2026; 17(2): 17–32. <https://doi.org/10.47093/2218-7332.2026.17.2.17-32>

CONTACT INFORMATION:

Angelina V. Zhilenkova, junior researcher, Institute of Personalized Oncology, Biomedical Science and Technology Park, Sechenov First Moscow State Medical University (Sechenov University).

Address: 8/2, Trubetskaya str., Moscow, 119048, Russia

E-mail: zhilenkova_a_v@staff.sechenov.ru

Ethics statements. This study was performed in line with the principles of the Declaration of Helsinki. The study was approved by the Local Ethics Committee of Sechenov First Moscow State Medical University (Sechenov University), approval No. 02-23, January 26, 2023. Informed consent for the storage of biospecimens and their subsequent use for analysis at the university's discretion was obtained from all participants, including those prospectively enrolled, at the time of sample deposition into the Biobank of Sechenov University; the same standard biobank consent form was used for all participants.

Data availability. The datasets generated or analyzed during the current study are available from the corresponding author on request. This study was not registered in a clinical trial or study registry. No custom code or software was developed for model derivation or implementation; all statistical analyses were performed using standard built-in procedures in SPSS v.23 (IBM Corp.). The final logistic regression model is fully specified in the manuscript, allowing independent implementation and validation without access to proprietary code.

Conflict of interest. The authors declare that there is no conflict of interests.

Financing. This research received no external funding.

Use of artificial intelligence. No artificial intelligence tools were used in the preparation of this manuscript.

Received: 21.05.2026

Accepted: 03.09.2026

Date of publication: 24.09.2026

Разработка и внутренняя валидация комбинации биомаркеров крови, ассоциированной с немелкоклеточным раком легкого: поисковое исследование «случай – контроль»

А.В. Жиленкова^{1,✉}, П.С. Шило², И.К. Ученду^{1,3}, Е.В. Орлова¹, В.С. Скосырский¹, Ю.А. Яснова¹, Н.Н. Багмет⁴, Н.А. Агафонов¹, Н.М. Кашаканова¹, М.И. Секачева¹

¹ ФГАОУ ВО «Первый Московский государственный медицинский университет имени И.М. Сеченова»
Министерства здравоохранения Российской Федерации (Сеченовский Университет)
ул. Трубецкая, д. 8, стр. 2, г. Москва, 119048, Россия

² ООО «Лакта Клиника»
ул. Дибуновская, д. 50, г. Санкт-Петербург, 197183, Россия

³ Университет Нигерии
Кампус Энугу, ул. Адемола, Огуи, г. Энугу, штат Энугу, 401105, Нигерия

⁴ ФГБНУ «Российский научный центр хирургии имени академика Б.В. Петровского»
пер. Абрикосовский, д. 2, г. Москва, 119435, Россия

Аннотация

Биомаркеры крови все чаще изучаются с точки зрения их способности различать немелкоклеточный рак легкого (НМРЛ) и немалигненные состояния. Однако отдельные маркеры нередко демонстрируют ограниченную эффективность, что привело к растущему интересу к комбинированию нескольких биологических признаков.

Цель. Разработать и оценить дискриминационную способность многофакторной диагностической модели на основе сывороточных биомаркеров в поисковом исследовании типа TRIPOD 1a.

Материалы и методы. Данное исследование представляет собой вторичный анализ ранее опубликованной когорты «случай – контроль», сфокусированный на разработке многофакторной модели, а не на оценке одиночных биомаркеров. В это ретроспективное исследование с проспективным набором образцов типа «случай – контроль» было включено в общей сложности 275 участников, из которых 85 пациентов имели гистологически подтвержденный НМРЛ, а 190 составили контрольную группу. Были проанализированы сывороточные концентрации 20 биомаркеров, отражающих опухолевые, воспалительные и метаболические пути. Сначала был выполнен однофакторный логистический регрессионный анализ, после чего проведено многофакторное моделирование с включением переменных при $p < 0,10$. Дискриминационная способность модели оценивалась с использованием ROC-анализа и площади под кривой (AUC), а внутренняя валидация выполнялась методом бутстреп-ресемплинга на полных данных.

Результаты. В окончательной многофакторной модели были сохранены три независимые переменные: белок эпидидимиса человека 4 (HE4, human epididymis protein 4), аполипопротеин A4 (ApoA4, apolipoprotein A4) и возраст. Итоговая модель продемонстрировала высокую дискриминационную способность в пределах исследуемого набора данных (AUC = 0,975; 95% ДИ: 0,958–0,992). Внутренняя валидация с использованием бутстреп показала минимальный оптимизм, со скорректированным значением AUC 0,972. HE4 и возраст положительно ассоциировались с НМРЛ, тогда как ApoA4 демонстрировал обратную связь. В рамках апостериорного анализа чувствительности модели на возраст-подобранной подвыборке ассоциации HE4 и ApoA4 с наличием заболевания оставались практически неизменными, тогда как возраст утрачивал независимую ассоциацию с НМРЛ, что свидетельствует о преимущественном вкладе биомаркеров, а не возрастного дисбаланса, в дискриминационную способность модели. Многофакторная модель показала лучшую дискриминацию по сравнению с отдельными биомаркерами.

Закключение. В данном поисковом исследовании по построению модели была выявлена кандидатная комбинация сывороточных биомаркеров, связанная с НМРЛ. Полученные результаты следует рассматривать как гипотезообразующие. Для клинической интерпретации и применения необходима независимая валидация в репрезентативных популяциях.

Ключевые слова: ранняя диагностика рака лёгкого; белок эпидидимиса человека 4; аполипопротеин A4; многофакторный анализ; логистическая регрессия; ROC-кривая

Рубрики MeSH:

КАРЦИНОМА НЕМЕЛКОКЛЕТОЧНАЯ ЛЕГКОГО – ДИАГНОСТИКА

КАРЦИНОМА НЕМЕЛКОКЛЕТОЧНАЯ ЛЕГКОГО – ПАТОЛОГИЯ

БИОМАРКЕРЫ НОВООБРАЗОВАНИЙ – АНАЛИЗ

ГИСТОЛОГИЧЕСКИЕ МЕТОДЫ

Для цитирования: Жиленкова А.В., Шило П.С., Ученду И.К., Орлова Е.В., Скосырский В.С., Яснова Ю.А., Багмет Н.Н., Агафонов Н.А., Кашаканова Н.М., Секачева М.И. Разработка и внутренняя валидация комбинации биомаркеров крови, ассоциированной с немелкоклеточным раком легкого: поисковое исследование «случай – контроль». Сеченовский вестник. 2026; 17(2): 17–32. <https://doi.org/10.47093/2218-7332.2026.17.2.17-32>

КОНТАКТНАЯ ИНФОРМАЦИЯ:

Жиленкова Ангелина Владимировна, младший научный сотрудник Института персонализированной онкологии Научно-технологического парка биомедицины ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).

Адрес: ул. Трубецкая, д. 8, стр. 2, г. Москва, 119048, Россия

E-mail: zhilenkova_a_v@staff.sechenov.ru

Соответствие принципам этики. Исследование было проведено в соответствии с принципами Хельсинкской декларации. Исследование одобрено Локальным этическим комитетом ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет), протокол № 02-23 от 26 января 2023 г. Информированное согласие на хранение биологических образцов и их последующее использование для анализа по усмотрению университета было получено от всех участников, включая проспективно включенных в исследование, в момент передачи образцов в Биобанк Сеченовского Университета; для всех участников использовалась единая стандартная форма информированного согласия Биобанка.

Доступ к данным исследования. Наборы данных, полученные или проанализированные в рамках настоящего исследования, доступны у автора для корреспонденции по запросу. Исследование не было зарегистрировано в реестре клинических исследований или иных реестрах исследований. Специальный программный код или программное обеспечение для разработки или реализации модели не создавались; все статистические анализы выполнялись с использованием стандартных встроенных процедур SPSS версии 23 (IBM Corp.). Окончательная логистическая регрессионная модель полностью представлена в рукописи, что позволяет независимо реализовать и валидировать ее без доступа к проприетарному программному коду.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Финансирование. Исследование не получало внешнего финансирования.

Использование искусственного интеллекта. Инструменты искусственного интеллекта не использовались при подготовке данной рукописи.

Поступила: 21.05.2026

Принята: 03.09.2026

Дата печати: 24.09.2026

Abbreviations:

AFP – alpha-fetoprotein

ApoA1 – apolipoprotein A1

ApoA2 – apolipoprotein A2

ApoA4 – apolipoprotein A4

ApoB – apolipoprotein B

AUC – area under the receiver operating characteristic curve

B2M – β 2-microglobulin

CA 15-3 – carbohydrate antigen 15-3

CA 19-9 – carbohydrate antigen 19-9

CA 125 – carbohydrate antigen 125

CEA – carcinoembryonic antigen

CI – confidence interval

CYFRA 21-1 – cytokeratin 19 fragment

HE4 – human epididymis protein 4

hsCRP – high-sensitivity C-reactive protein

LDCT – low-dose computed tomography

LRG-1 – leucine-rich alpha-2-glycoprotein 1

NSCLC – non-small cell lung cancer

OR – odds ratio

RANTES – regulated upon activation, normal T-cell expressed and secreted

ROC – receiver operating characteristic

sVCAM-1 – soluble vascular cell adhesion molecule 1

tPSA – total prostate-specific antigen

TTR – transthyretin

VEGFR1 – vascular endothelial growth factor receptor 1

HIGHLIGHTS	КЛЮЧЕВЫЕ ПОЛОЖЕНИЯ
An exploratory diagnostic model derivation study identifies a combination of human epididymis protein 4, apolipoprotein A4, and age associated with non-small cell lung cancer.	В поисковом исследовании по разработке диагностической модели выявлена комбинация белка эпидидимиса человека 4, аполипопротеина A4 и возраста, ассоциированная с немелкоклеточным раком легкого.
The developed model demonstrates high apparent discrimination, with an area under the curve of 0.975 and 0.972 after bootstrap internal validation.	Разработанная модель демонстрирует высокую дискриминационную способность в исследуемой выборке: площадь под кривой составляет 0,975, а после внутренней бутстреп-валидации – 0,972.
After age matching, independent associations of human epididymis protein 4 and apolipoprotein A4 with non-small cell lung cancer are preserved.	После сопоставления по возрасту сохраняются независимые ассоциации белка эпидидимиса человека 4 и аполипопротеина A4 с немелкоклеточным раком легкого.
The identified serum biomarker combination provides a basis for subsequent independent validation of the diagnostic model.	Выявленная комбинация сывороточных биомаркеров представляет собой основу для последующей независимой валидации диагностической модели.

Lung cancer is the leading cause of cancer-related mortality worldwide. According to global statistics, approximately 2.4 million new cases of lung cancer and 1.8 million related deaths were registered in 2022 [1]. The most common histological type is non-small cell lung cancer (NSCLC), accounting for ~85% of all cases [2]. Despite advances in oncology, the overall prognosis for patients with advanced lung cancer remains poor. A direct correlation between prognosis and disease stage at diagnosis has been demonstrated: for a localized process (stage I), five-year overall survival can reach 60–70% or higher (exceeding 90% for stage IA), whereas in the presence of distant metastases (stage IV), it decreases to less than 10% [3].

Early detection of lung cancer remains a key challenge, which has driven the development of screening programs. Evidence from large randomized trials, including the National Lung Screening Trial (NLST) in the USA and the Nederlands-Leuvens Longkanker Screenings Onderzoek (NELSON) in Europe, indicates that low-dose computed tomography (LDCT) screening can reduce mortality by increasing the detection of tumors at earlier, potentially curable stages.

National lung cancer screening programs based on LDCT are already operating in a number of countries. In particular, such programs have been implemented in the United States since the mid-2010s, and in recent years, they have been introduced in Europe (e.g., the Czech Republic, Poland, and Croatia) and Asia (China, Taiwan, and South Korea). These programs have substantially increased the proportion of cancers detected at an early stage, reaching up to 70% within organized screening cohorts [4–6].

Despite its success in reducing mortality, LDCT screening has significant limitations. First, according to NLST data, suspicious lung findings were detected by LDCT in 24% of the screened individuals, but malignancy was confirmed in only 1–2%. Second, even with a reduced radiation dose, regular computed tomography scanning is associated with cumulative

radiation exposure. Third, LDCT screening requires substantial resources, such as high-quality tomography equipment and highly qualified personnel, which are not always available in all regions.

Fourth, limited patient awareness also contributes to low uptake of screening programs. For instance, in the USA in 2015, only about 4% of individuals who met high-risk criteria underwent LDCT screening. These observations highlight the need for additional diagnostic approaches that are more accessible and offer improved specificity for early lung cancer detection [7].

The development of noninvasive diagnostic tests based on blood biomarkers is increasingly considered as a potential approach for lung cancer screening and early detection [3]. Biomarkers can be understood as measurable biological indicators associated with the presence of malignant processes. Their use in screening is being explored in several contexts, including earlier cancer detection and reduction of missed cases.

In addition, they may help refine diagnostic pathways by lowering the number of false-positive LDCT findings through noninvasive assessment. Another potential application is risk stratification, where biomarkers could assist in identifying smokers with the highest likelihood of malignancy and guide decisions regarding computed tomography referral [7].

For clinical application, a biomarker should combine high sensitivity, allowing detection of as many true cancer cases as possible, with adequate specificity to limit false-positive results. Practical considerations are also important, including measurement reproducibility and feasibility for routine use in terms of cost and accessibility [7].

A wide range of biomarkers has been studied for lung cancer screening, but their routine clinical use remains limited. One reason is the overlap in marker levels between malignant and benign conditions, leading to reduced specificity. In addition, variability between patients and insufficient validation in large independent cohorts continue to restrict their application [5]. Combining

different types of biomarkers has been explored as a way to overcome the limitations of individual tests. Several studies indicate that such approaches may improve diagnostic accuracy and support earlier detection of disease. This, in turn, could facilitate the development of noninvasive screening methods that complement LDCT and enhance early detection of NSCLC [3].

In our previous work using this cohort [8], we reported the univariable diagnostic performance of 20 individual serum biomarkers. However, single biomarkers often lack sufficient sensitivity and specificity for clinical decision-making. The present study is a secondary, in-depth analysis of the same cohort, with a shift in focus from univariate description to multivariable prediction. Here, we aimed to develop and internally validate a parsimonious multimarker signature for NSCLC case-control discrimination. The analysis was conducted in an exploratory, hypothesis-generating framework to identify a combination of biomarkers for further independent prospective validation.

MATERIALS AND METHODS

Patients and biomarker measurements

This case-control study, based primarily on retrospective biobank samples and supplemented by a small number of prospectively enrolled NSCLC patients, was conducted at a single academic center, Sechenov First Moscow State Medical University (Sechenov University).

Biological samples from both NSCLC patients and controls were retrieved retrospectively from the biobank of the Institute for Personalized Oncology, Sechenov University; the NSCLC group was additionally supplemented with prospectively collected samples from patients enrolled at two affiliated university clinical hospitals, University Clinical Hospital No. 1 and University Clinical Hospital No. 4.

A total of 275 participants were included in the analysis, comprising 85 patients with histologically confirmed NSCLC and 190 controls. Serum or plasma levels of 20 biomarkers – alpha-fetoprotein (AFP), carcinoembryonic antigen (CEA), carbohydrate antigens 19-9 (CA 19-9), 125 (CA 125), and 15-3 (CA 15-3), human epididymis protein 4 (HE4), total prostate-specific antigen (tPSA), β 2-microglobulin (B2M), high-sensitivity C-reactive protein (hsCRP), D-dimer, cytokeratin 19 fragment (CYFRA 21-1), apolipoproteins A1 (ApoA1), A2 (ApoA2), B (ApoB), and A4 (ApoA4), transthyretin (TTR), soluble vascular cell adhesion molecule 1 (sVCAM-1), regulated upon activation normal T-cell expressed and secreted (RANTES), vascular endothelial growth factor receptor 1 (VEGFR1), and leucine-rich alpha-2-glycoprotein 1 (LRG-1) – were quantified using standard laboratory methods.

The exclusion criteria included acute infection or inflammatory conditions, recent surgery or trauma,

pregnancy, and any history of malignant disease. Blood sampling was performed prior to any oncologic treatment. Control participants were healthy volunteers recruited during routine annual medical check-ups; the biobank records did not include a smoking-history questionnaire.

Serum levels of AFP, CA 15-3, CA 19-9, CA 125, HE4, CEA, CYFRA 21-1, and tPSA were measured via an electrochemiluminescence immunoassay on a Cobas e411 analyzer (Roche Diagnostics, Germany). The levels of hsCRP, ApoA1, ApoB, and TTR were measured on an Advia 1800 analyzer via an immunoturbidimetric method (Siemens Healthcare, Germany).

Serum concentrations of sVCAM-1, RANTES, VEGFR1, ApoA4, and LRG-1 were determined using enzyme-linked immunosorbent assay with Quantikine® kits (R&D Systems, USA) and read on a Biochrom Anthos 2020 microplate reader (Biochrom, UK).

The levels of B2M and D-dimer were measured via a chemiluminescent immunoassay on an Immulite 2000 analyzer (Siemens Medical Solutions, USA). ApoA2 was measured via an enzymatic colorimetric method (Randox Laboratories, UK).

Sample size considerations

At the planning stage, the sample size was primarily considered in relation to the precision of receiver operating characteristic (ROC) curve estimation for individual biomarkers. Based on published data and preliminary observations, several candidate biomarkers were expected to demonstrate an area under the ROC curve (AUC) of approximately 0.80 or higher. According to commonly used recommendations for diagnostic studies, detecting an AUC ≥ 0.8 versus the null hypothesis of AUC = 0.5 with a two-sided significance level $\alpha = 0.05$ and statistical power of 80%, assuming a case-control ratio close to 1:2, requires approximately 50–70 cases and 100–150 controls.

Statistical analysis

This study followed the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) recommendations and corresponds to a TRIPOD Type 1a model development study (model derivation without external validation).

Univariable and multivariable logistic regression analyses of 20 serum biomarkers were performed to identify the most promising predictors for constructing multimarker models. In the first stage, univariable analysis (binary logistic regression using the Enter method) was conducted in SPSS v.23. The dependent variable was the presence of lung cancer (0 – control, 1 – NSCLC). For each biomarker, the odds ratio (OR), 95% confidence interval (CI), and p value were calculated. Statistical significance was defined as $p < 0.05$.

In the second stage, multivariable analysis (logistic regression) for biomarkers and demographic variables

(age and sex) was performed: variables with $p < 0.10$ in univariable analyses were entered into the multivariable model. Predictors for the final model were selected using forward stepwise selection (Forward: Conditional method, SPSS v.23), in which candidate predictors were sequentially added to the model, with inclusion and retention determined by the statistical significance of their independent contribution (entry threshold $p < 0.05$; removal threshold $p > 0.10$). The predictors retained by the stepwise procedure were subsequently re-estimated using the forced-entry (Enter) method to derive the final regression equation.

The performance of the resulting model was assessed via the Nagelkerke R^2 coefficient and ROC curve analysis.

Missing data were assessed for all variables. ApoA4 concentrations were unavailable for 10 of the 275 enrolled participants (9 of 85 NSCLC patients and 1 of 190 controls) because enzyme-linked immunosorbent assay reagent kits required for this assay were

temporarily unavailable at the time of testing owing to disruptions in the international supply of laboratory reagents. These participants were not excluded from the study and contributed to all analyses not requiring ApoA4 data. Analyses incorporating ApoA4, including the multivariable model, were therefore performed using complete-case analysis ($n = 265$: 76 NSCLC patients and 189 controls) (Fig. 1). The proportion of missing values for key clinical variables related to the TNM (tumor, node, metastasis) classification is reported in the Results section.

Model performance assessment, internal validation, and sensitivity analyses

Model discrimination was quantified using the AUC with 95% CI. Predicted probabilities were calculated from the final model. The threshold was determined within the same dataset and therefore represents an internal descriptive cut-off rather than a clinically validated decision threshold.

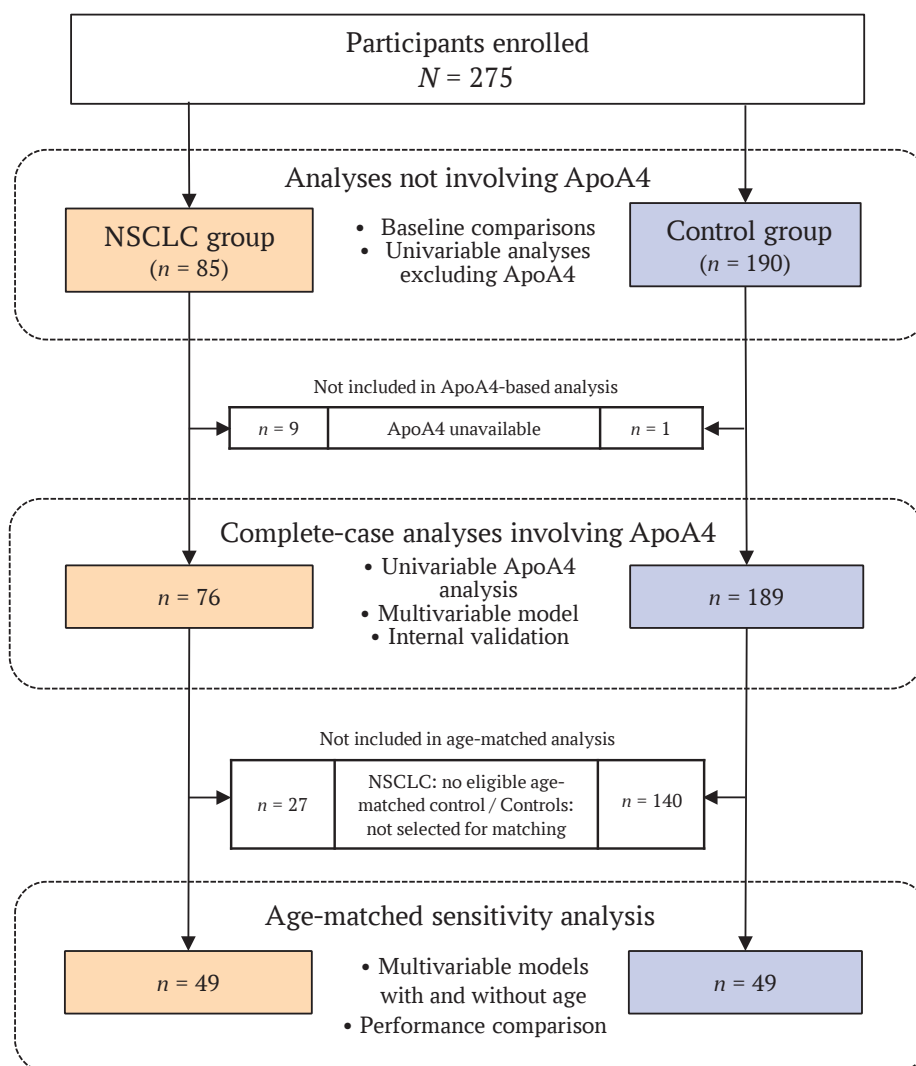


FIG. 1. Flow diagram of patient inclusion in the study.

Note: ApoA4 – apolipoprotein A4; NSCLC – non-small cell lung cancer.

To assess internal validity and quantify optimism due to potential overfitting, bootstrap resampling was performed ($B = 500$). In each bootstrap sample (same size as the original dataset, sampled with replacement), the logistic regression model including HE4, ApoA4, and age was refitted. Discrimination was quantified using the AUC. For each bootstrap replicate, AUC was calculated in the bootstrap sample ("apparent") and in the original dataset ("test"). Optimism was defined as the difference between apparent and test AUC and averaged across 500 replicates. The optimism-corrected AUC was obtained by subtracting the mean optimism from the apparent AUC of the model fitted to the original dataset. Complete-case analysis was used.

Age-matched sensitivity analysis

A post hoc sensitivity analysis was undertaken to determine whether the observed age imbalance between cases and controls, rather than the biomarkers themselves, accounted for the model's discriminatory performance.

Among complete-case subjects (76 NSCLC patients and 189 controls with non-missing HE4, ApoA4, and age values), each patient was matched without replacement to a control of similar age. Optimal matching using the Hungarian algorithm was applied to minimize the summed absolute age difference across all pairs under a caliper of ± 3 years; patients without an eligible control within this window were left unmatched and excluded from the resulting subsample.

HE4- and ApoA4-based logistic regression models were then refitted in the matched subsample, both with and without age as a covariate, and their performance (AUC, Nagelkerke R^2 , sensitivity, and specificity at the Youden-optimal cut-off) was compared with the results from the full dataset. This analysis was intended as a robustness check rather than a replacement for the primary model.

Study design and relationship to previous publication

This study is a secondary analysis of data from a previously published our case-control study [8]. The original study described the univariate diagnostic accuracy of 20 serum biomarkers using ROC analysis. For detailed descriptive statistics of individual biomarkers, readers are referred to the original publication [8]. The current study focuses on multivariable model development and does not duplicate the univariate findings of the original publication.

RESULTS

Cohort characteristics

The clinical and demographic characteristics of the study cohort are summarized in Table 1. Among the lung cancer patients, 63.5% ($n = 54$) were men and 36.5% ($n = 31$) were women. In the control group, men constituted 52.1% ($n = 99$) and women 47.9% ($n = 91$)

of the participants. The mean age was 62.14 years (95% CI: 60.07–64.21) in the NSCLC group and 48.86 years (95% CI: 47.97–49.74) in the control group.

Among the NSCLC patients, T1 status was identified in 28.2% ($n = 24$), T2 in 49.4% ($n = 42$), T3 in 10.6% ($n = 9$), T4 in 5.9% ($n = 5$), and Tx in 5.9% ($n = 5$). N0 status was recorded in 45.9% ($n = 39$), N1 in 5.9% ($n = 5$), N2 in 9.4% ($n = 8$), and Nx in 38.8% ($n = 33$). M0 status was defined in 71.8% ($n = 61$), M1 in 7.1% ($n = 6$), and Mx in 21.2% ($n = 18$).

Owing to the predominantly retrospective, biobank-based nature of the study, the availability of clinical data was limited, which necessitated the use of the "x" designation in the TNM classification to indicate an undetermined value for a specific parameter.

Univariable associations between biomarkers and non-small cell lung cancer

The comparison of median biomarker concentrations between NSCLC patients and controls using the Mann-Whitney U test, along with Bonferroni correction for multiple comparisons, has been reported previously [8]. After adjustment, the following biomarkers showed statistically significant differences: HE4, D-dimer, CYFRA 21-1, CA 125, B2M, hsCRP, ApoA1, ApoA2, ApoA4, TTR, LRG-1, and VEGFR1. Detailed descriptive statistics (medians, interquartile ranges) can be found in the original publication [8].

Extending our previous ROC-based analysis of this cohort [8], we next performed univariable logistic regression to quantify associations as ORs. Statistically significant associations within the univariable analysis were obtained for most of the studied biomarkers (Table 2, Fig. 2). For a number of biomarkers, direct associations ($OR > 1$) were identified, whereas for others, inverse associations ($OR < 1$) were noted with NSCLC status in this dataset.

Among tumor-associated antigens, statistically significant associations were identified for CEA (OR 1.418; 95% CI: 1.208–1.664; $p < 0.001$), CA 19-9 (OR 1.053; 95% CI: 1.020–1.086; $p = 0.001$), CA 125 (OR 1.115; 95% CI: 1.071–1.160; $p < 0.001$), HE4 (OR 1.083; 95% CI: 1.062–1.106; $p < 0.001$), tPSA (OR 1.339; 95% CI: 1.019–1.761; $p = 0.036$), CA 15-3 (OR 1.067; 95% CI: 1.030–1.105; $p < 0.001$), and CYFRA 21-1, which demonstrated the highest OR value (OR 4.931; 95% CI: 3.138–7.748; $p < 0.001$).

Among the markers of inflammation, immune response, and coagulation, statistically significant associations were identified for hsCRP (OR 1.174; 95% CI: 1.099–1.253; $p < 0.001$), D-dimer (OR 1.006; 95% CI: 1.004–1.009; $p < 0.001$), B2M (OR 1.003; 95% CI: 1.002–1.003; $p < 0.001$), sVCAM-1 (OR 1.001; 95% CI: 1.000–1.003; $p = 0.019$), as well as RANTES (OR 1.000; 95% CI: 1.000–1.000; $p = 0.020$) and LRG-1 (OR 1.000; 95% CI: 1.000–1.000; $p < 0.001$).

Among apolipoprotein-related markers, inverse associations with NSCLC status were identified for

Table 1. Clinical and demographic characteristics of patients with non-small cell lung cancer and controls

Parameter	NSCLC group (n = 85)	Control group (n = 190)
Sex		
female	31 (36.5)	91 (47.9)
male	54 (63.5)	99 (52.1)
Age, years	62.14 (60.07; 64.21)	48.86 (47.97; 49.74)
T status		
T1	24 (28.2)	
T2	42 (49.4)	
T3	9 (10.6)	
T4	5 (5.9)	
Tx	5 (5.9)	
N status		
N0	39 (45.9)	
N1	5 (5.9)	
N2	8 (9.4)	
Nx	33 (38.8)	
M status		
M0	61 (71.8)	
M1	6 (7.1)	
Mx	18 (21.2)	

Notes: quantitative variables are presented as means with 95% confidence interval, categorical variables are presented as the absolute number of patients with the characteristic and the proportion within the group, expressed as a percentage (in parentheses).

M – metastasis; N – node; NSCLC – non-small cell lung cancer; T – tumor; x – the corresponding parameter was not determined owing to limited availability of clinical data.

Table 2. Univariable and multivariable logistic regression analyses of biomarkers and demographic factors associated with non-small cell lung cancer

Parameter	Univariable OR (95% CI)	p-value	Multivariable OR (95% CI)	p-value
AFP, IU/mL	1.017 (0.908–1.140)	0.771		
CEA, ng/mL	1.418 (1.208–1.664)	<0.001*		
CA 19-9, U/mL	1.053 (1.020–1.086)	0.001*		
CA 125, U/mL	1.115 (1.071–1.160)	<0.001*		
HE4, pmol/L	1.083 (1.062–1.106)	<0.001*	1.077 (1.043–1.112)	<0.001*
tPSA, ng/mL	1.339 (1.019–1.761)	0.036*		
CA 15-3, U/mL	1.067 (1.030–1.105)	<0.001*		
B2M, ng/mL	1.003 (1.002–1.003)	<0.001*		
hsCRP, mg/L	1.174 (1.099–1.253)	<0.001*		
D-dimer, ng/mL	1.006 (1.004–1.009)	<0.001*		
CYFRA 21-1, ng/mL	4.931 (3.138–7.748)	<0.001*		
ApoA1, g/L	1.019 (0.907–1.145)	0.747		
ApoA2, g/L	1.019 (0.964–1.093)	0.414		
ApoB, g/L	0.264 (0.087–0.800)	0.019*		
TTR, mg/dL	0.796 (0.747–0.849)	<0.001*		
sVCAM-1, ng/mL	1.001 (1.000–1.003)	0.019*		
ApoA4, mcg/mL	0.943 (0.927–0.959)	<0.001*	0.898 (0.864–0.934)	<0.001*
RANTES, pg/mL	1.000 (1.000–1.000)	0.020*		
VEGFR1, pg/mL	1.002 (0.997–1.006)	0.427		
LRG-1, ng/mL	1.000 (1.000–1.000)	<0.001*		
Age, years	1.219 (1.163–1.278)	<0.001*	1.203 (1.114–1.298)	<0.001*
Sex (female vs. male)	0.625 (0.369–1.056)	0.079*		

Notes: * $p < 0.10$ indicating variables considered as candidates for inclusion in the multivariable model.

AFP – alpha-fetoprotein; ApoA1 – apolipoprotein A1; ApoA2 – apolipoprotein A2; ApoA4 – apolipoprotein A4; ApoB – apolipoprotein B; B2M – β 2-microglobulin; CA 15-3 – carbohydrate antigen 15-3; CA 19-9 – carbohydrate antigen 19-9; CA 125 – carbohydrate antigen 125; CEA – carcinoembryonic antigen; CI – confidence interval; CYFRA 21-1 – cytokeratin 19 fragment; HE4 – human epididymis protein 4; hsCRP – high-sensitivity C-reactive protein; LRG-1 – leucine-rich alpha-2-glycoprotein 1; OR – odds ratio; RANTES – regulated upon activation, normal T-cell expressed and secreted; sVCAM-1 – soluble vascular cell adhesion molecule 1; tPSA – total prostate-specific antigen; TTR – transthyretin; VEGFR1 – vascular endothelial growth factor receptor 1.

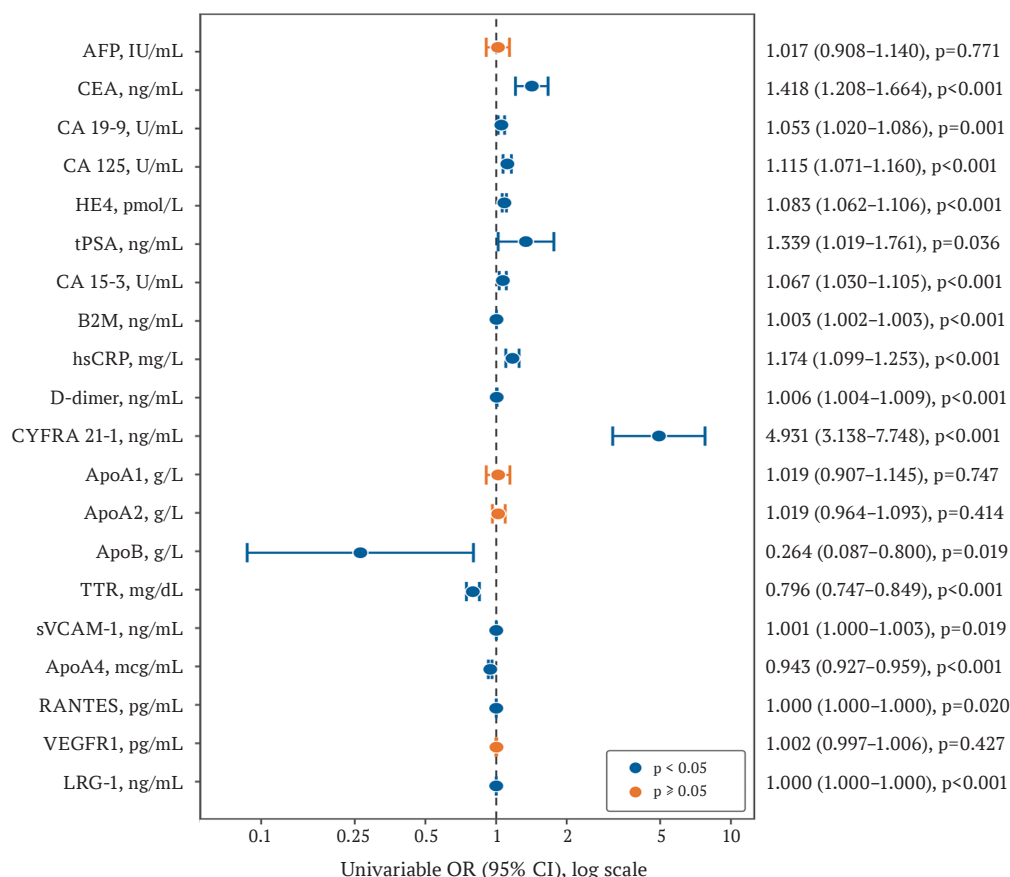


FIG. 2. Forest plot of univariable logistic regression analysis of biomarkers in non-small cell lung cancer.

Notes: the x-axis is on a logarithmic scale. Biomarkers with $p < 0.05$ are shown in red and those with $p \geq 0.05$ in grey; the vertical dashed line corresponds to OR = 1.

AFP – alpha-fetoprotein; ApoA1 – apolipoprotein A1; ApoA2 – apolipoprotein A2; ApoA4 – apolipoprotein A4; ApoB – apolipoprotein B; B2M – β 2-microglobulin; CA 15-3 – carbohydrate antigen 15-3; CA 19-9 – carbohydrate antigen 19-9; CA 125 – carbohydrate antigen 125; CEA – carcinoembryonic antigen; CI – confidence interval; CYFRA 21-1 – cytokeratin 19 fragment; HE4 – human epididymis protein 4; hsCRP – high-sensitivity C-reactive protein; LRG-1 – leucine-rich alpha-2-glycoprotein 1; OR – odds ratio; RANTES – regulated upon activation, normal T-cell expressed and secreted; sVCAM-1 – soluble vascular cell adhesion molecule 1; tPSA – total prostate-specific antigen; TTR – transthyretin; VEGFR1 – vascular endothelial growth factor receptor 1.

ApoB (OR 0.264; 95% CI: 0.087–0.800; $p = 0.019$), ApoA4 (OR 0.943; 95% CI: 0.927–0.959; $p < 0.001$), and TTR (OR 0.796; 95% CI: 0.747–0.849; $p < 0.001$).

Biomarkers such as AFP (OR 1.017; 95% CI: 0.908–1.140; $p = 0.771$), ApoA1 (OR 1.019; 95% CI: 0.907–1.145; $p = 0.747$), ApoA2 (OR 1.019; 95% CI: 0.964–1.093; $p = 0.414$), and VEGFR1 (OR 1.002; 95% CI: 0.997–1.006; $p = 0.427$) did not show a statistically significant association.

Multivariable model development, internal validation, and performance

At the first stage, univariable logistic regression identified a number of biomarkers significantly associated with NSCLC status. These findings provided the basis for the subsequent multivariable models aimed at identifying independent predictors and developing a combined multimarker diagnostic panel. Four biomarkers did not meet the prespecified candidate-screening criterion

of $p < 0.10$ and were therefore excluded from further consideration: AFP ($p = 0.771$), ApoA1 ($p = 0.747$), ApoA2 ($p = 0.414$), and VEGFR1 ($p = 0.427$).

At the second stage, the remaining 16 biomarkers that had met the prespecified candidate-screening criterion of $p < 0.10$ in the univariable analysis were entered as candidates into a forward stepwise (Forward: Conditional) logistic regression, together with age and sex as potential demographic factors that could influence the observed associations, yielding a total of 18 candidate predictors.

The stepwise selection procedure retained three independent predictors in the final model: age, HE4 concentration, and ApoA4 level. Inclusion of the remaining candidate variables did not produce a statistically significant improvement in model fit, and they were therefore excluded from the final regression equation. All three retained predictors remained statistically significant in the multivariable model: HE4

(OR 1.077; 95% CI: 1.043–1.112; $p < 0.001$), ApoA4 (OR 0.898; 95% CI: 0.864–0.934; $p < 0.001$), and age (OR 1.203; 95% CI: 1.114–1.298; $p < 0.001$). Once the composition of the multimarker panel had been determined, the final logistic regression model was constructed using the forced-entry (Enter) method.

The linear predictor (LP) was calculated as:

$$LP = -9.652 + 0.074 \times HE4 - 0.107 \times ApoA4 + 0.185 \times age,$$

where HE4 is expressed in pmol/L, ApoA4 in mcg/mL, and age in years.

The predicted probability of NSCLC was calculated as:

$$p(NSCLC) = 1 / (1 + e^{-LP}),$$

where e is the base of the natural logarithm ($e \approx 2.718281828$).

The resulting model showed measurable association with case-control status. The Nagelkerke R^2 coefficient was 0.822. ROC analysis demonstrated the high discriminatory ability of the model (AUC 0.975; 95% CI: 0.958–0.992; $p < 0.001$). The comparative characteristics of the isolated HE4 and the model (HE4 + ApoA4 + age) are presented in Table 3.

The optimal probability threshold determined using the Youden index was 0.32. Patients with predicted probabilities ≥ 0.32 were categorized as test-positive within the study dataset. At this cut-off, the model demonstrated a sensitivity of 88.2% and a specificity of 93.1%.

In complete-case analysis ($n = 265$), the model including HE4, ApoA4, and age showed high apparent discrimination (AUC = 0.975). Bootstrap internal validation ($B = 500$) indicated minimal optimism for AUC (mean optimism = 0.003), yielding an optimism-corrected AUC of 0.972.

A nomogram (Fig. 3) based on HE4, ApoA4, and age was constructed to estimate individual probability of NSCLC. The predicted probability is obtained by summing the points assigned to each variable and projecting the total score onto the probability scale. A threshold probability of 0.32 was used to categorize

observations as test-positive within the study dataset. The nomogram is provided as a visualization of the fitted regression model to facilitate reproducibility and future external validation rather than for immediate clinical use.

Age-matched sensitivity analysis

Given the substantial mean age difference between NSCLC patients and controls (Table 1), an age-matched subsample was constructed to assess whether this imbalance contributed to the apparent discriminatory performance of the model. Using optimal 1:1 matching within a ± 3 -year caliper, 49 of 76 complete-case NSCLC patients (64.5%) were matched to controls of comparable age; the remaining 27 patients, predominantly the oldest in the cohort, could not be matched because control donors older than 68 years were not represented in the available dataset. In the resulting matched subsample ($n = 98$), mean age was similar between groups (57.41 ± 6.40 vs. 56.20 ± 5.73 years; $p = 0.329$), confirming adequate balancing.

In this age-matched subsample, the model based on HE4 and ApoA4 alone showed an AUC of 0.961 (95% CI: 0.925–0.989; Nagelkerke $R^2 = 0.793$). Adding age to the model did not materially improve discrimination (AUC = 0.962; 95% CI: 0.925–0.989; Nagelkerke $R^2 = 0.793$).

At the Youden-optimal threshold, both models achieved identical sensitivity (93.9%) and specificity (91.8%), with an overall accuracy of 92.9%. Notably, once cases and controls were balanced by age, age was no longer independently associated with NSCLC status (OR 0.995; 95% CI: 0.881–1.124; $p = 0.935$), whereas HE4 (OR 1.094; 95% CI: 1.043–1.149; $p < 0.001$) and ApoA4 (OR 0.869; 95% CI: 0.812–0.930; $p < 0.001$) retained ORs and significance levels essentially unchanged from the full unmatched dataset. A comparable pattern was observed in the full complete-case dataset ($n = 265$), where the HE4 + ApoA4 model alone (without age) yielded an AUC of 0.961 (95% CI: 0.939–0.982), close to that of the age-matched subsample and only modestly lower than the full three-variable model (AUC = 0.975).

Taken together, these findings indicate that the discriminatory performance attributable to HE4 and ApoA4 was preserved after removing the age imbalance

Table 3. Discriminatory performance of human epididymis protein 4 alone and the multivariable model within the study dataset

Parameter	HE4	Multivariable model (HE4 + ApoA4 + age)
AUC	0.903	0.975
95% CI for AUC	0.858–0.947	0.958–0.992
Overall accuracy, %	86.5	91.6
Specificity, %	87.9	93.1
Sensitivity, %	83.5	88.2
Nagelkerke R^2	0.602	0.822

Note: ApoA4 – apolipoprotein A4; AUC – area under the receiver operating characteristic curve; CI – confidence interval; HE4 – human epididymis protein 4.

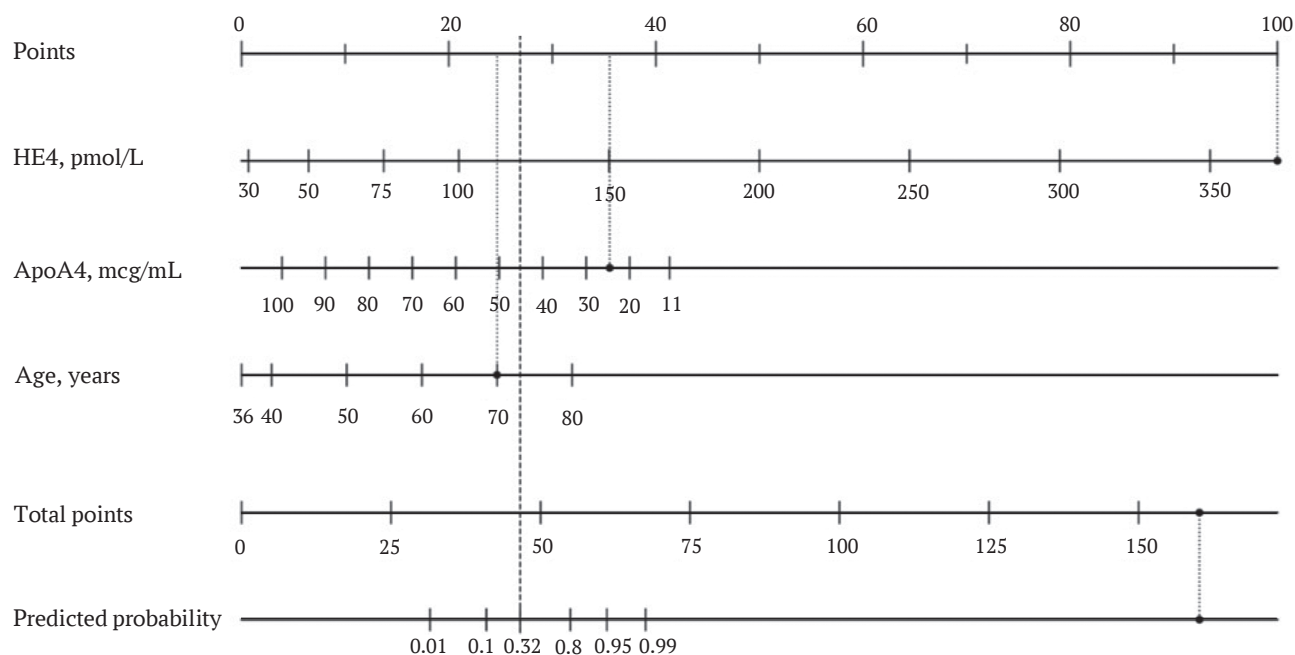


FIG. 3. Harrell-style nomogram for estimation of NSCLC probability based on human epididymis protein 4, apolipoprotein A4, and age.

Notes: To estimate individual NSCLC probability, locate each patient's value on the HE4, ApoA4, and age axes and read the corresponding number of points from the "Points" scale at the top; the resulting point values are then summed to obtain the "Total points", which is projected onto the "Predicted probability" axis to obtain the estimated probability of NSCLC. The dotted lines illustrate the probability calculation for an individual patient based on the values of the predictors included in the model. The vertical dashed line indicates the Youden-optimal probability threshold of 0.32 used to classify observations as test-positive within the study dataset.

ApoA4 – apolipoprotein A4; HE4 – human epididymis protein 4; NSCLC – non-small cell lung cancer.

between groups, supporting a biomarker-driven rather than purely age-driven contribution to model discrimination.

DISCUSSION

Unlike our prior univariate analysis of this cohort [8], which identified individual biomarkers with diagnostic potential, the present secondary analysis demonstrates that combining HE4, ApoA4, and age in a multivariable model substantially improves discrimination. Notably, ApoA1, ApoA2, and VEGFR1, which were significantly different between groups by the Mann-Whitney *U* test with Bonferroni correction in the original analysis [8], did not reach significance in the univariable logistic regression performed here. This discrepancy likely reflects differences in statistical sensitivity between the two methods: the rank-based Mann-Whitney test detects any consistent distributional shift, whereas the univariable OR, estimated per unit of raw biomarker concentration, is more sensitive to distributional shape, extreme values, and measurement scale.

Multimarker analyses are primarily used in exploratory biomarker research to identify candidate variables associated with disease status rather than to establish clinical diagnostic tests. Model development studies represent an early research phase aimed at determining whether measurable biological signals exist that justify further validation [9]. Several multimarker

panels have already been developed and implemented, such as the OVA1® test for ovarian cancer detection, which is based on five plasma biomarkers, and the multitarget DNA test Cologuard® for colorectal cancer screening [10, 11].

The growth of computational capabilities and the spread of machine learning methods allow for more efficient use of combinations of clinical data and biomarkers to improve the detection of neoplasms. For instance, H.I. Yoon et al. [12] demonstrated the possibility of algorithmic risk assessment for lung cancer on the basis of risk factors (age, sex, etc.) and biomarker analysis to optimize disease diagnosis. Thus, the development of noninvasive multimarker tests remains a relevant direction for improving the early diagnosis of lung cancer.

In the present study, a simple combination model including two biomarkers (HE4 and ApoA4) and patient age was developed for NSCLC detection. Biomarker distributions in clinical datasets are often right-skewed, and logarithmic transformation is commonly applied to stabilize variance and improve probability calibration in regression models. Transformations mainly change the scale of predictors but do not substantially alter the ranking of observations, and their impact on discrimination metrics such as the AUC is usually limited. In this study, the focus was on identifying a discriminative biomarker signature rather than developing a fully calibrated risk

model. For this reason, biomarkers were analyzed using their original measurement scales. Keeping variables in their native laboratory units also improves interpretability and makes the model easier to apply in practice. Clinicians can use routinely reported values without additional transformations when estimating predicted probabilities. A non-transformed model was therefore retained to preserve simplicity and support reproducibility.

The obtained results demonstrated performance metrics for sensitivity (88.2%) and specificity (93.1%) that surpassed those of single markers. The observed AUC reflects discrimination within the case-control dataset and should not be interpreted as real-world diagnostic performance. Retrospective two-gate designs are known to produce optimistic estimates of discrimination; therefore, the present results should be considered exploratory evidence of a detectable biomarker signal rather than validation of a clinical test. For comparison, the single biomarker HE4 in the literature has an average sensitivity of approximately 73% and specificity of approximately 86% [13], meaning that the combined model substantially improves both metrics.

Furthermore, our three-factor model is comparable in diagnostic value to more complex multimarker models. For example, in the study by H.I. Yoon et al. combining six markers (including HE4) with age yielded an AUC of 0.99 and a sensitivity of 93%, with a specificity of 92%. The authors also reported that the multimarker approach significantly increases the sensitivity of lung cancer detection compared to any single indicator [12]. Thus, our data are consistent with the results of other groups and support the concept that combining biologically distinct biomarkers may improve case-control discrimination compared with single-marker approaches in exploratory datasets.

Assessment of individual predictors in the model indicated a particularly strong contribution from HE4 and ApoA4. Among them, HE4 appears to be especially informative. Previous studies have reported that its diagnostic sensitivity in lung cancer is higher than that of many commonly used markers, and its levels are less influenced by tumor stage compared with several other antigens [12, 14]. In this dataset, HE4 was significantly associated with differentiation between NSCLC cases and controls. Similar results have been reported in meta-analyses that highlight its diagnostic relevance [13].

ApoA4 is not a conventional biomarker in lung cancer diagnostics and has been linked to alterations in lipid metabolism. Data on its role in NSCLC remain inconsistent and appear to vary depending on histological subtype, with increased expression reported in squamous cell carcinoma and decreased levels observed in adenocarcinoma [15]. Our results indicate a negative association of ApoA4 with the presence of lung cancer, which is consistent with the literature data. The inclusion of ApoA4 in our model reflects the link

between metabolic disorders and the carcinogenesis process. Earlier epidemiological studies have also demonstrated an association between components of metabolic syndrome (e.g., low high-density lipoprotein cholesterol levels, closely linked to the ApoA profile) and an increased risk of lung cancer [16]. The combined use of the tumor antigen HE4 and the metabolic marker ApoA4 in a panel covers different pathophysiological mechanisms of NSCLC and may reflect complementary biological processes associated with disease presence.

The demographic factor of age made a substantial contribution to the discrimination model alongside the biomarkers. Age is one of the key risk factors for lung cancer development, and its increase is accompanied by an exponential increase in incidence [16]. However, the pronounced age difference between cases and controls raised the possibility that apparent model discrimination was inflated by this imbalance despite statistical adjustment.

To address this concern directly, we performed a post hoc age-matched sensitivity analysis, which showed that HE4 and ApoA4 retained virtually unchanged ORs and statistical significance once cases and controls were balanced by age, while age itself lost independent association with NSCLC status in the matched subsample. This finding suggests that the biomarker signal is not primarily attributable to the underlying age imbalance, although the reduced size of the matched subsample (49 pairs) limits precision, and residual confounding by unmeasured or partially measured factors, including smoking history, cannot be fully excluded.

H.I. Yoon et al. reported that inclusion of age alongside a panel of six biomarkers was associated with improved classification accuracy [12]. In the present dataset, age showed a strong association with NSCLC status in the full, unmatched cohort, but this association was not sustained once the age distributions of the two groups were rendered comparable. Consequently, while demographic variables such as age may still add practical value in a real-world screening population with a broader, more representative age distribution, the present results indicate that the diagnostic signal carried by HE4 and ApoA4 does not depend on the age imbalance specific to this case-control dataset.

Several limitations should be considered, largely related to the study design. The analysis was based on a single-center case-control model derivation approach that combined retrospective biobank samples (all controls and 76 of 85 NSCLC patients) with a small number of prospectively enrolled NSCLC patients ($n = 9$); no prospective recruitment was performed for the control group. While this design is suitable for identifying potential biomarker signals, it does not reflect real-world diagnostic performance. Smoking history was not available for a substantial proportion of controls and therefore was not included in model development. In such sampling schemes, disease prevalence is artificially

determined, and discrimination metrics, including AUC, sensitivity, and specificity, are typically overestimated compared with those observed in prospective diagnostic cohorts. Consequently, the proposed model should not be interpreted as a clinical diagnostic or screening tool.

The age imbalance between cases and controls and the lack of key clinical covariates, particularly smoking history, introduce a risk of residual confounding despite statistical adjustment. An age-matched sensitivity analysis was performed to partially address this concern; however, matching was feasible for only 49 of 76 complete-case NSCLC patients, as the available control donors were predominantly younger than 60 years and could not be matched to the oldest patients in the cohort. This reflects a structural limitation of the control population available for this study rather than a deficiency of the matching procedure itself and underscores the need for a prospectively age-matched or age-representative control population in future validation studies.

In addition, no external validation was performed, and missing biomarker values were handled using complete-

case analysis, which may further affect generalizability. Therefore, the present findings should be considered evidence of a detectable biomarker association rather than a validated clinical prediction model, and prospective validation in representative diagnostic populations is required before clinical application.

CONCLUSION

In this exploratory model derivation study, a multivariable combination of HE4, ApoA4, and age was identified as a candidate serum biomarker profile associated with non-small cell lung cancer. Owing to the two-gate case-control design and the absence of external validation, the discrimination observed within the studied dataset (AUC = 0.975) should be regarded as preliminary and potentially optimistic rather than as evidence of clinical performance. The proposed model should therefore not be interpreted as a clinical diagnostic test but rather as a hypothesis-generating biomarker combination requiring independent prospective validation in clinically relevant populations.

AUTHOR CONTRIBUTIONS

Angelina V. Zhilenkova: study concept and design, biosample collection, statistical analysis, drafting the manuscript. **Polina S. Shilo:** study concept and design, statistical analysis. **Ikenna K. Uchendu, Julia A. Iasnova, Vladislav S. Skossyrskiy:** drafting the manuscript. **Ekaterina V. Orlova, Nikolay A. Agafonov:** study concept and design, critical revision of the manuscript. **Natalya M. Kashakanova, Nikolay N. Bagmet, Marina I. Sekacheva:** study concept and design, laboratory analysis, biosample collection, critical revision of the manuscript. All authors reviewed and approved the final version of the manuscript.

ВКЛАД АВТОРОВ

А.В. Жиленкова: разработка концепции и дизайна исследования, сбор биоматериала, статистический анализ, подготовка текста рукописи. **П.С. Шило:** разработка концепции и дизайна исследования, статистический анализ. **И.К. Ученду, Ю.А. Яснова, В.С. Скосырский:** подготовка текста рукописи. **Е.В. Орлова, Н.А. Агафонов:** разработка концепции и дизайна исследования, критический пересмотр рукописи. **Н.М. Кашаканова, Н.Н. Багмет, М.И. Секачева:** разработка концепции и дизайна исследования, проведение лабораторных исследований, сбор биоматериала, критический пересмотр рукописи. Все авторы утвердили окончательную версию статьи.

REFERENCES / ЛИТЕРАТУРА

1. Zhou J., Xu Y., Liu J., et al. Global burden of lung cancer in 2022 and projections to 2050: incidence and mortality estimates from GLOBOCAN. *Cancer Epidemiol.* 2024 Dec; 93: 102693. <https://doi.org/10.1016/j.canep.2024.102693>. Epub 2024 Nov 13. PMID: 39536404. EDN: DDEFLU
2. Molina J.R., Yang P., Cassivi S.D., et al. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin Proc.* 2008 May; 83(5): 584–594. <https://doi.org/10.4065/83.5.584>. PMID: 18452692. EDN: MNEYDX
3. Ning J., Ge T., Jiang M., et al. Early diagnosis of lung cancer: which is the optimal choice? *Aging (Albany NY)*. 2021 Feb; 13(4): 6214–6227. <https://doi.org/10.18632/aging.202504>. Epub 2021 Feb 11. PMID: 33591942. EDN: RDRVXA
4. Hardavella G., Frille A., Sreter K.B., et al. Lung cancer screening: where do we stand? *Breathe (Sheff)*. 2024 Aug; 20(2): 230190. <https://doi.org/10.1183/20734735.0190-2023>. PMID: 39193459. EDN: VYHFFK
5. Ru Zhao Y., Xie X., de Koning H.J., et al. NELSON lung cancer screening study. *Cancer Imaging*. 2011 Oct; 11 Spec No A(1A): S79–S84. <https://doi.org/10.1102/1470-7330.2011.9020>. PMID: 22185865
6. National Lung Screening Trial Research Team, Aberle D.R., Adams A.M., Berg C.D., et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med*. 2011 Aug; 365(5): 395–409. <https://doi.org/10.1056/NEJMoa1102873>. Epub 2011 Jun 29. PMID: 21714641
7. Marmor H.N., Zorn J.T., Deppen S.A., et al. Biomarkers in lung cancer screening: a narrative review. *Curr Chall Thorac Surg*. 2023 Feb; 5: 5. <https://doi.org/10.21037/ccts-20-171>. Epub 2021 Mar 1. PMID: 37016707. EDN: VRZPVH
8. Жиленкова А.В., Воронова В.М., Орлова Е.В. и др. Диагностическая ценность биомаркеров крови для диагностики рака легкого. *Наука и инновации в медицине*. 2026; 11(1): 31–37. <https://doi.org/10.35693/SIM698688>. EDN: EVHITS / Zhilenkova A.V., Voronova V.M., Orlova E.V., et al. Diagnostic value of blood biomarkers for the diagnosis of lung cancer. *Science and Innovations in Medicine*. 2026; 11(1): 31–37 (In Russian). <https://doi.org/10.35693/SIM698688>. EDN: EVHITS
9. Voronova V., Glybochko P., Svistunov A., et al. Diagnostic value of combinatorial markers in colorectal carcinoma. *Front Oncol*. 2020 May; 10: 832. <https://doi.org/10.3389/fonc.2020.00832>. PMID: 32528895. EDN: HBISGG
10. Zhang Z. An in vitro diagnostic multivariate index assay (IVDMIA) for ovarian cancer: harvesting the power of multiple biomarkers. *Rev Obstet Gynecol*. 2012; 5(1): 35–41. <https://doi.org/10.3909/riog0182>. PMID: 22582125
11. Imperiale T.F., Ransohoff D.F., Itzkowitz S.H., et al. Multitarget stool DNA testing for colorectal-cancer screening. *N Engl J*

- Med. 2014 Apr; 370(14): 1287–1297. <https://doi.org/10.1056/NEJMoa1311194>. Epub 2014 Mar 19. PMID: 24645800. EDN: XIMGGZ
12. Yoon H.I., Kwon O.R., Kang K.N., et al. Diagnostic value of combining tumor and inflammatory markers in lung cancer. *J Cancer Prev.* 2016 Sep; 21(3): 187–193. <https://doi.org/10.15430/JCP.2016.21.3.187>. Epub 2016 Sep 30. Erratum in: *J Cancer Prev.* 2016; 21(4): 302. <https://doi.org/10.15430/JCP.2016.21.4.302>. PMID: 27722145
 13. He Y.P., Li L.X., Tang J.X., et al. HE4 as a biomarker for diagnosis of lung cancer: a meta-analysis. *Medicine (Baltimore).* 2019 Sep; 98(39): e17198. <https://doi.org/10.1097/MD.00000000000017198>. PMID: 31574828. EDN: SDNOVH
 14. Zare M.E., Nasir Kansestani A., Wu X., et al. Serum human epididymis protein-4 outperforms conventional biomarkers in the early detection of non-small cell lung cancer. *iScience.* 2024 Oct; 27(11): 111211. <https://doi.org/10.1016/j.isci.2024.111211>. PMID: 39524348. EDN: JHOTXQ
 15. Darwish N.M., Al-Hail M.K., Mohamed Y., et al. The role of apolipoproteins in the commonest cancers: a review. *Cancers (Basel).* 2023 Nov; 15(23): 5565. <https://doi.org/10.3390/cancers15235565>. PMID: 38067270. EDN: MYWQXO
 16. Li M., Cao S.M., Dimou N., et al. Association of metabolic syndrome with risk of lung cancer: a population-based prospective cohort study. *Chest.* 2024 Jan; 165(1): 213–223. <https://doi.org/10.1016/j.chest.2023.08.003>. Epub 2023 Aug 10. PMID: 37572975. EDN: BHSKKP

INFORMATION ABOUT THE AUTHORS / ИНФОРМАЦИЯ ОБ АВТОРАХ

Angelina V. Zhilenkova[✉], junior researcher, Institute of Personalized Oncology, Biomedical Science and Technology Park, Sechenov First Moscow State Medical University (Sechenov University).
ORCID: <https://orcid.org/0000-0002-0060-2197>

Polina S. Shilo, oncologist, chemotherapist, Lahta Clinic LLC.
ORCID: <https://orcid.org/0009-0001-1482-4604>

Ikenna K. Uchendu, student, Institute of Personalized Oncology, Biomedical Science and Technology Park, Sechenov First Moscow State Medical University (Sechenov University); lecturer, Department of Medical Laboratory Science, Faculty of Health Science and Technology, College of Medicine University of Nigeria.
ORCID: <https://orcid.org/0000-0002-1503-3759>

Ekaterina V. Orlova, Cand. of Sci. (Medicine), senior researcher, Institute of Personalized Oncology, Biomedical Science and Technology Park, Sechenov First Moscow State Medical University (Sechenov University).
ORCID: <https://orcid.org/0000-0002-1684-8781>

Vladislav S. Skossyrskiy, training specialist, Accreditation and Simulation Center, student at the Institute of Clinical Medicine, Sechenov First Moscow State Medical University (Sechenov University).
ORCID: <https://orcid.org/0000-0002-3417-2359>

Julia A. Iasnova, student, Sechenov First Moscow State Medical University (Sechenov University).
ORCID: <https://orcid.org/0009-0005-5970-5296>

Nikolay N. Bagmet, Dr. of Sci. (Medicine), associate professor, chief researcher, oncologist, Department of Coloproctology and Urogynecology, surgeon, Diagnostic and Treatment Department, Research and Clinical Center No. 2, Petrovsky National Research Center of Surgery.
ORCID: <https://orcid.org/0000-0001-8325-4409>

Жиленкова Ангелина Владимировна[✉], младший научный сотрудник Института персонализированной онкологии Научно-технологического парка биомедицины ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).
ORCID: <https://orcid.org/0000-0002-0060-2197>

Шило Полина Сергеевна, врач-онколог, химиотерапевт ООО «Лахта Клиника».
ORCID: <https://orcid.org/0009-0001-1482-4604>

Ученду Икенна Кингсли, студент Института персонализированной онкологии Научно-технологического парка биомедицины ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет); преподаватель кафедры медицинских лабораторных наук факультета наук о здоровье и технологий Колледжа медицины Университета Нигерии.
ORCID: <https://orcid.org/0000-0002-1503-3759>

Орлова Екатерина Вадимовна, канд. мед. наук, старший научный сотрудник Института персонализированной онкологии Научно-технологического парка биомедицины ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).
ORCID: <https://orcid.org/0000-0002-1684-8781>

Скосырский Владислав Сергеевич, учебный мастер аккредитационного центра, студент Института клинической медицины ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).
ORCID: <https://orcid.org/0000-0002-3417-2359>

Яснова Юлия Александровна, студентка ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).
ORCID: <https://orcid.org/0009-0005-5970-5296>

Багмет Николай Николаевич, д-р мед. наук, доцент, главный научный сотрудник, врач-онколог отделения колопроктологии и урогинекологии, врач-хирург лечебно-диагностического отделения Научно-клинического центра № 2 ФГБНУ «РНЦХ им. акад. Б.В. Петровского».
ORCID: <https://orcid.org/0000-0001-8325-4409>

Nikolay A. Agafonov, research associate, Institute of Personalized Oncology, Biomedical Science and Technology Park, Sechenov First Moscow State Medical University (Sechenov University).

ORCID: <https://orcid.org/0009-0007-0982-5032>

Natalya M. Kashakanova, Head of the Centralized Laboratory and Diagnostic Service, Clinical Center, Sechenov First Moscow State Medical University (Sechenov University).

ORCID: <https://orcid.org/0009-0009-9879-2117>

Marina I. Sekacheva, Dr. of Sci. (Medicine), Professor, Department of Oncology, Radiotherapy and Reconstructive Surgery, director, Institute of Personalized Oncology, Biomedical Science and Technology Park, director, Department of Innovative Healthcare Development, Sechenov First Moscow State Medical University (Sechenov University).

ORCID: <https://orcid.org/0000-0003-0015-7094>

Агафонов Николай Александрович, научный сотрудник Института персонализированной онкологии Научно-технического парка биомедицины ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).

ORCID: <https://orcid.org/0009-0007-0982-5032>

Кашаканова Наталья Михайловна, руководитель Централизованной лабораторно-диагностической службы Клинического центра ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).

ORCID: <https://orcid.org/0009-0009-9879-2117>

Секачева Марина Игоревна, д-р мед. наук, профессор кафедры онкологии, радиотерапии и реконструктивной хирургии; директор Института персонализированной онкологии Научно-технического парка биомедицины; директор департамента инновационного развития здравоохранения ФГАОУ ВО «Первый МГМУ им. И.М. Сеченова» Минздрава России (Сеченовский Университет).

ORCID: <https://orcid.org/0000-0003-0015-7094>