

INTRODUCTION

Early screening of LC such as LDCT significantly improves survival rates but faces limitations like high false-positive rates. Since human physiological processes yield a multifaceted array of metabolic byproducts, a novel strategy, breath analysis for **volatile organic compounds (VOCs)** shows up. This study aims to build a VOC panel for LC screening.

Methods

This study was conducted between December 1, 2022 and December 31, 2024 in the First Affiliated Hospital of Guangzhou Medical University. Exhaled breath samples were collected and analysed by high-pressure photon ionisation time-of-flight mass spectrometry. Among 144 detected VOC peak features, based on a meta-analysis already completed, seven of them were selected as potential breath biomarkers. Then developed different classification models using five machine learning algorithms. The model exhibiting superior performance was selected for subsequent subgroup analysis, while factors such as histological type, stage, and maximum tumor diameter were also investigated.

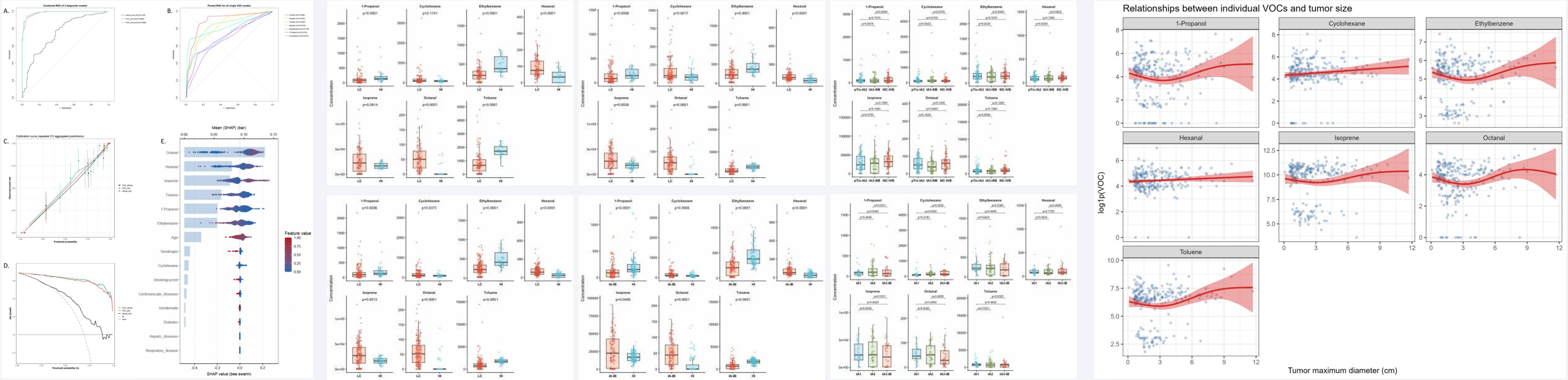
Characteristic	Lung cancer (%)	Healthy individual (%)	P value
Demographic data			
Age, mean(SD) (years)	58.5(11.49)	66.7(11.00)	<0.0001
Gender (female)	124 (50.29)	36 (12.43)	<0.0001
Smoking history	63 (25.45)	60 (24.05)	<0.0001
Current smoking	47 (18.93)	34 (10.63)	0.0146
Comorbidities			
Diabetes	20 (8.06)	21 (13.92)	0.0029
Respiratory disease	29 (10.08)	31 (10.79)	0.9462
Cardiovascular disease	68 (27.42)	49 (16.14)	0.0018
Hepatic disease	26 (10.48)	3 (1.70)	0.0126
Pathological data			
Maximum diameter	26.5(10.93)		
Cancer type	8/5073(42/23/11/710)		
Pathological stage	8/5775(78/281/512/1274/9)		
Tumor size (mm)	3.2(1.4) 9.2(2.3) 39.1(1.2) 49.4(3.6) 47.0(2.7) 8.2(2.8)		
VOC peak intensity of HPPI-TOFMS			
Acetone	7396.61 (1555.69-12539.03)	7686.11 (5502.14-11221.87)	0.1097
1-Propanol	85.4 (24.53-163.39)	150.22 (102.97-247.94)	<0.0001
Dimethyl sulfide	790.63 (306.12-1273.92)	780.16 (560.5-1136.45)	0.4892
Isoprene	2735.49 (1188.61-4022.4)	3483.19 (1307.69-2283.01)	<0.0001
2-Butanone	379.01 (21.76-502.34)	467.76 (404.27-532.32)	<0.0001
1-Butanol	28.3 (10.16-63.14)	34.9 (0-71.56)	0.4488
Benzene	413.23 (240.52-578.88)	448.26 (340.81-645.32)	0.0034
Cyclohexane	92.58 (64.89-120.3)	79.94 (48.91-103.86)	0.0009
Hexanal	80.59 (60.13-146.27)	79.69 (50.56-96.84)	0.1895
Aspirin	487.26 (244.12-947.12)	638.55 (429.83-1061.47)	0.0008
Toluene	752.14 (455.18-1269.12)	1662.65 (1300.2-2053.44)	<0.0001
Cyclohexanone	118.09 (76.34-164.99)	106.99 (86.85-140.32)	0.2676
Hexanol	84.43 (54.04-127.89)	34.68 (24-47.97)	<0.0001
Isoprene	125.17 (86.27-185.63)	91.54 (77.76-122.25)	<0.0001
Ethylbenzene	221.23 (131.51-342.59)	381 (289.73-558.77)	<0.0001
Buty acetate	59.87 (22.52-102.54)	87.48 (36.52-164.5)	0.0131
Propylbenzene	65.24 (40.15-93.66)	81.49 (43.9-123.71)	0.0023
Octanal	50.25 (22.79-77.78)	0 (0-13.43)	<0.0001

Number of Participants, lung cancer: 248; Healthy individual: 111.
Data of VOCs are the median (interquartile range). P values are from Wilcoxon matched pairs signed-rank sum tests.
SD: standard deviation; VOC: volatile organic compound; HPPI-TOFMS: high-pressure photon ionisation time-of-flight mass spectrometry; AS: Adenocarcinoma in situ; MIA: Minimally invasive adenocarcinoma; LUAD: Lung adenocarcinoma; LUOC: Lung squamous cell carcinoma; NSCLC: Non-small cell lung cancer; not otherwise specified; NSCLC: Small cell lung cancer.

Results

Key Findings

- Developed a meta-evidence-guided 7-VOC breathomics panel for lung cancer detection using HPPI-TOFMS and machine learning. Unlike purely data-driven screening based solely on the 144 peaks identified in the current cohort, this study first integrated lung cancer-associated VOCs consistently reported in previous meta-analyses. Subsequent differential analysis and LASSO regression were used to refine the selection, ultimately identifying seven VOCs: **1-propanol, isoprene, cyclohexane, toluene, hexanal, ethylbenzene, and octanal**.
- The 7-VOC panel demonstrates strong discriminatory power for lung cancer, with the primary diagnostic information derived from the VOCs themselves. The 7-VOC model maintains stable performance across various clinical subgroups, including early-stage lung cancer. The model achieved an AUC of approximately 0.98 across subgroups including women, non-smokers, patients without comorbidities, those with lung adenocarcinoma, and patients with early-stage disease (stages IA–IB). This indicates that its diagnostic capability is not solely dependent on smoking status, age, or characteristics typical of high-risk populations, suggesting potential utility for early detection.
- The VOC signal appears to indicate the "presence of lung cancer" rather than reflecting tumor burden or disease progression. While the 7-VOC panel effectively distinguishes lung cancer patients from healthy controls, it shows limited ability to differentiate between histological subtypes or TNM stages, and it lacks a consistent linear or monotonic correlation with maximum tumor diameter.



DISCUSSION AND CONCLUSION

Summary

- A 7-VOCs breathomics panel guided by previous meta-analysis measured showed strong internally validated performance for distinguishing lung cancer from healthy controls. It also remained robust across key subgroups, but showed restrictive association with histological subtypes, stage, and maximum tumor diameter.
- Further external multicentre validation in clinically realistic populations is needed before clinical application.