國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering
College of Electrical Engineering and Computer Science
National Taiwan University

Master's Thesis

肺癌患者呼出氣體凝結液的呼吸體學分析

Breathomics Analysis of the Exhaled Breath Condensate of Lung Cancer Patients

林首志

Shou-Zhi Lin

指導教授：曾宇鳳 博士

Advisor: Yufeng Jane Tseng, Ph.D.

中華民國 114 年 7 月

July 2025

doi:10.6342/NTU202504043

誌謝

能夠完成這篇論文，首先要感謝我的指導老師曾宇鳳教授還有郭天爵學長的指導，從實驗的設計到論文的撰寫與修改，老師和學長給我的寶貴建議與協助讓我了解到一個完整的研究應具有的架構以及細節，也讓我學習到之前未曾接觸過的生醫資訊領域的許多知識。也感謝我的口試委員何肇基醫師以及王三源學長對我論文提出的建議，讓我的論文可以更加完整。另外也要感謝實驗室的蘇柏翰學長，在我仍專注在先前的研究主題時給了我很多的建議與幫助。感謝實驗室、資訊系、學校認識的同學們，他們給了我不少精神上的支持以及課業之外的啟發。最後要感謝家人們對我生活上的支持，讓我可以專心的完成我的研究。

doi:10.6342/NTU202504043

中文摘要

肺癌的早期偵測可以大幅改善病人的治療結果。當前肺癌篩檢的主要方法是低劑量電腦斷層掃描，但仍有開發更快速且更具成本效益的篩檢方法的需求。本研究探討了呼出氣體中揮發性有機化合物的代謝體學分析作為肺癌早期偵測工具的潛力。

我們使用頂空固相微萃取法和氣相層析質譜法，分析了 72 名肺癌患者和 13 名健康個體呼出氣體凝結液樣本中揮發性有機化合物的組成。透過單變量以及多變量分析，我們發現肺癌患者與健康個體之間揮發性有機化合物的組成有顯著差異，而肺癌不同分期及兩種肺癌亞型（腺癌和鱗狀細胞癌）之間的差異則不顯著，我們提出了 18 種有潛力作為肺癌診斷生物標記的揮發性有機化合物和 5 種有潛力作為肺癌分期生物標記的揮發性有機化合物。然而，仍需進一步的研究來驗證這些潛在生物標記的可靠性。

關鍵詞：肺癌偵測、呼吸代謝體學、呼氣分析、代謝體學、揮發性有機化合物、呼出氣體凝結液、生物標記

doi:10.6342/NTU202504043

Abstract

Early detection of lung cancer can significantly improve patient outcomes. Low-dose computed tomography (LDCT) is the main lung cancer screening method, yet developing other quicker and more cost-effective methods is desirable. This study investigated the potential of breathomic analysis of volatile organic compounds (VOCs) in exhaled breath as a tool for the early detection of lung cancer.

We analyzed the VOC composition in exhaled breath condensate (EBC) samples collected from 72 lung cancer patients and 13 healthy individuals using headspace solid-phase microextraction (HS-SPME) and gas chromatography–mass spectrometry (GC–MS). By univariate and multivariate analysis, we found significant differences in VOC profiles between lung cancer patients and healthy individuals. In contrast, the differences between lung cancer stages and the two lung cancer subtypes, adenocarcinoma (AC) and squamous cell carcinoma (SCC), were nonsignificant. We proposed 18 VOCs as potential biomarkers for lung cancer diagnosis and 5 for cancer staging. However, additional studies are required to validate these potential biomarkers.

Keywords: Lung cancer detection, breathomics, breath analysis, metabolomics, volatile organic compounds, exhaled breath condensate, biomarkers

doi:10.6342/NTU202504043

Contents

doi:10.6342/NTU202504043

doi:10.6342/NTU202504043

List of Figures

doi:10.6342/NTU202504043

List of Tables

doi:10.6342/NTU202504043

Glossary

AC	Adenocarcinoma
ANCOVA	Analysis of Covariance
EBC	Exhaled Breath Condensate
EMM	Estimated Marginal Mean
FC	Fold Change
GC–MS	Gas Chromatography–Mass Spectrometry
HS-SPME	Headspace Solid-phase Microextraction
LDCT	Low-dose Computed Tomography
m/z	Mass-to-charge Ratio
PCA	Principal Component Analysis
PLS-DA	Partial Least Squares Discriminant Analysis
QC	Quality Control
SCC	Squamous Cell Carcinoma
TOF	Time-of-flight
VIP	Variable Importance in Projection
VOC	Volatile Organic Compound

doi:10.6342/NTU202504043

Chapter 1 Introduction

Lung cancer was the second most diagnosed cancer and the leading cause of cancer death worldwide in 2020, with an estimated 2.2 million new cases and 1.8 million deaths. This cancer accounts for 11.4% of cancer incidence and 18% of cancer mortality[1]. Because the majority of lung cancer cases are diagnosed at later stages, the five-year survival rate of lung cancer patients after diagnosis is only 10% to 20% in most countries[2]. However, early detection of lung cancer can substantially improve survival. If lung cancer is diagnosed at stage I or stage II, the five-year survival rate can surpass 50%[3]. Thus, early detection of lung cancer is crucial for reducing lung cancer mortality.

Several lung cancer detection approaches have been developed to accurately detect and characterize early-stage lung cancers[4], [5]. Chest X-ray is a comparatively accessible way to detect lung cancer. Although it can successfully detect some early-stage cases, this tool is not considered effective for lung cancer screening[6]. Low-dose computed tomography (LDCT) has been proven to be effective in detecting early-stage lung cancer and is currently widely used for lung cancer screening[7], [8]. However, the whole LDCT examination process can be time-consuming, from scheduling a scan to receiving a report. Liquid biopsy is an emerging technique for the early detection of lung cancer, but it is still in the research stage, and there is no evidence for its effectiveness as a screening tool[9], [10].

Breathomics is a promising cancer detection method, and it has the advantages of being simple, quick, and noninvasive. By collecting and analyzing exhaled breath or exhaled breath condensate (EBC), volatile organic compound (VOC) metabolites associated with cancer can be identified and quantified. Previous breathomic studies have shown potential in detecting various types of cancer, including breast, head and neck, and gastric cancer[11],

doi:10.6342/NTU202504043

and pulmonary diseases, such as asthma and chronic obstructive pulmonary disease[12]. Many breathomic studies have been devoted to lung cancer as well. Numerous potential lung cancer VOC biomarkers have been proposed in recent decades, although the results were deemed heterogeneous due to different experimental conditions and analytical approaches[13]. The most frequently mentioned potential VOC biomarkers include 2-butanone, 1-propanol, isoprene, ethylbenzene, styrene, and hexanal, according to Saalberg and Wolff[14]. Additional breathomic research efforts are needed to determine which VOCs are more reliable for lung cancer detection. Apart from breathomic studies, canine scent detection[15], [16] and sensor technology[17] have also shown promising results in lung cancer diagnosis, validating the idea of using exhaled breath as a detection tool for lung cancer.

In addition to detecting lung cancer, distinguishing between lung cancer subtypes and stages can be beneficial because therapeutic strategies depend on these subtypes and stages[18]. The most common lung cancer subtypes are adenocarcinoma (AC) and squamous cell carcinoma (SCC)[19]. Some studies have attempted to develop lung cancer subtyping and staging approaches based on exhaled breath. Mazzone *et al.*[20] used a colorimetric sensor array to analyze exhaled breath. Their model for classifying AC and SCC achieved 90% sensitivity and 83% specificity, and their model for classifying early and advanced stages achieved 81% sensitivity and 73% specificity. Peled *et al.*[17] analyzed exhaled breath with a chemical nanoarray. Their model for classifying AC and SCC achieved 92% and 78% sensitivity and specificity, respectively, and their model for classifying early and advanced stages achieved 86% and 88% sensitivity and specificity, respectively. Wang *et al.*[21] studied the possibility of classifying lung cancer subtypes by applying various data

doi:10.6342/NTU202504043

preprocessing techniques and machine learning algorithms to gas chromatography–mass spectrometry (GC–MS) breathomic data. However, the performance of their best classifier was still insufficient for clinical practice.

In this study, we collected EBC samples from lung cancer patients and healthy individuals and analyzed the samples using GC–MS. Using statistical methods, we investigated the differences in the breathome between lung cancer patients and healthy individuals, between AC and SCC, and between lung cancer stages. We aimed to identify potential VOC biomarkers for lung cancer that could help future lung cancer breathomic research.

doi:10.6342/NTU202504043

Chapter 2 Materials and Methods

2.1 Study Subjects

This study involved 85 subjects, including 72 lung cancer patients and 13 healthy control subjects. The lung cancer patients included 50 patients diagnosed with AC and 22 patients diagnosed with SCC. All subjects exhibited normal pulmonary function.

2.2 EBC Sample Collection

We collected EBC samples from the subjects using the RTube™ Breath Condensate Collection Device (Respiratory Research, Charlottesville, VA, USA). The subjects were asked to fast for 8 hours before sample collection. During each sample collection process, the aluminum sleeve of the device was precooled at -80°C for 20 minutes, and each subject was instructed to inhale and exhale using their mouth tidally for 15 minutes through the mouthpiece without wearing a nose clip. The subjects were allowed to temporarily stop sample collection when they needed to swallow saliva or cough. The exhaled breath was condensed and collected at the base of a polypropylene tube. The EBC samples were immediately stored at -80°C until subsequent analysis. To ensure the quality of the whole experiment, we prepared quality control (QC) samples by pooling EBC samples collected from 5-7 individuals and separating the mixture into multiple vials. The QC samples were analyzed across batches to monitor the consistency of the experiment.

2.3 HS-SPME Sampling

The headspace solid-phase microextraction (HS-SPME) vials were cleaned twice with

doi:10.6342/NTU202504043

deionized water, ethanol, and acetone in a sonicator and then dried under a stream of nitrogen. Afterward, 0.5 mL of EBC sample and 200 mg of NaCl were added to the vials. The headspace of the vials was sampled using a PDMS/DVB fiber and extracted at 45°C for 4 hours. After the extraction, the SPME fiber was immediately transferred to the GC injector port and heated at 250°C for 3 minutes in splitless mode for the thermal desorption of the analytes into the GC column so that the loss of the extracted substances was minimized. To avoid sample carryover, the SPME fiber was cleaned in the GC injection port at 250°C for 30 minutes in split mode before each extraction, and a blank run was conducted to ensure that the fiber had been thoroughly cleaned.

2.4 GC–TOF-MS Analysis

Gas chromatography–time-of-flight mass spectrometry (GC–TOF-MS) analyses were performed on a LECO Pegasus 4D time-of-flight mass spectrometer (Leco Corporation, St. Joseph, MI, USA) equipped with an Agilent 7890a gas chromatograph (Agilent Technologies, Santa Clara, CA, USA). The chromatographic column used was a 30-meter DB-5MS capillary column (5% phenyl/95% dimethylpolysiloxane) with an internal diameter of 250 μm. The oven temperature was maintained at 50°C for 2 minutes, increased at a rate of 10°C/min to 280°C, and held at 280°C for 5 minutes. The helium carrier gas flow rate was set at 1 mL/min. The electron energy was set at 70 eV, and the ion source temperature was at 240°C. The TOF-MS detector was operated at 1500 V in autodetection mode. The data were acquired in full scan mode with a mass range of m/z 40–550.

doi:10.6342/NTU202504043

2.5 Compound Identification

The raw GC–MS data were converted to the netCDF format by LECO ChromaTOF® software (version 4.33). The netCDF files were further processed using MZmine (version 3.3.0)[22]. We used the ADAP algorithms[23] incorporated in MZmine for extracted ion chromatogram (EIC) construction, chromatographic peak detection, spectral deconvolution, and alignment. Afterward, the aligned retention time, peak heights across the samples, and mass spectrum of each aligned compound were exported for later analysis. Table 1 lists the parameters used in the MZmine processing pipeline.

Table 1. Parameters in the MZmine processing pipeline

Step	Parameter	Value
Mass detection	Mass detector	Centroid Noise level = 500 Detect isotope signals below noise level = false
ADAP chromatogram builder	Min group size in # of scans	5
	Group intensity threshold	1000
	Min highest intensity	1000
	Scan-to-scan accuracy	m/z 0.002 or 10 ppm
ADAP feature resolver	Dimension	Retention time
	S/N threshold	4
	S/N estimator	Wavelet Coeff. SN Peak width mult. = 3 abs(wavelet coeffs.) = true
	Min feature height	500
	Coefficient/area threshold	110
	Peak duration range	0–40
	RT wavelet range	0–0.1

doi:10.6342/NTU202504043

Multivariate curve resolution	Deconvolution window width	0.2 min
	Retention time tolerance	0.05 min
	Minimum number of peaks	1
	Adjust apex ret times	False
ADAP aligner (GC)	Min confidence	0.1
	Retention time tolerance	0.3 min
	m/z tolerance	m/z 0.001 or 5 ppm
	Score threshold	0.75
	Score weight	0.1
	Retention time similarity	Retention time difference
Peak finder (multithreaded)	Intensity tolerance	20%
	m/z tolerance	m/z 0.001 or 5 ppm
	Retention time tolerance	0.4 min
	Minimum data points	1

For compound identification, the computed mass spectra were compared with the mass spectra in the NIST 20 mass spectral libraries (*mainlib* and *replib*). To ensure accuracy, we first applied composite weighted cosine similarity[24] to retrieve the most similar library spectra. For each compound, twenty top matches were generated: ten using an intensity weight of 0.5 and m/z weight of 0, and ten using an intensity weight of 0.5 and m/z weight of 2. The compounds were then manually identified. The identification criteria included alignment of peak positions and similarity in intensity patterns between the experimental and library spectra. We also compared the computed retention time to the library retention index data to assess the reasonability of the identification. Compounds without good matches in the library were excluded from future analysis. We excluded all identified compounds that were not organic or likely contaminants. Finally, a total of 43 identified compounds were used for subsequent analysis.

doi:10.6342/NTU202504043

2.6 Statistical Analysis

Univariate and multivariate analyses were conducted to determine differences in the VOC profiles among the healthy subjects, patients with AC, patients with SCC, and patients with different cancer stages.

In the univariate analysis, we used analysis of covariance (ANCOVA) and estimated marginal means (EMMs)[25] to adjust for potential confounders, including age, sex, and smoking history. ANCOVA could reveal whether the VOC levels between the groups were different, and EMMs were used to calculate pairwise comparison *p*-values and fold changes. To mitigate the effect of outliers on model fitting, robust linear models were used as the underlying models of ANCOVA and EMMs. The *p*-values obtained from ANCOVA were adjusted using the Benjamini–Hochberg procedure[26] to control the false discovery rate (FDR). The *p*-values obtained from pairwise comparisons of the EMMs were adjusted using Scheffé's method[27] for each VOC. All univariate analyses were conducted using R (version 4.3.2)[28]. ANCOVA was conducted using the rstatix package (version 0.7.2)[29]. EMMs and pairwise *p*-values were calculated using the emmeans package (version 1.8.9)[30]. Robust linear model fitting was conducted using the rlm function in the MASS package (version 7.3.60)[31].

Multivariate analyses, including principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA), and heatmap analysis, were performed using Metaboanalyst (version 6.0)[32]. The data were normalized by autoscaling before analysis. The hierarchical clustering algorithm for generating the heatmap was Ward’s method with Euclidean distance.

doi:10.6342/NTU202504043

Chapter 3 Results and Discussion

3.1 Overview of the Data

Table 2 and Figure 1 summarize the subject characteristics. There were more male subjects than female subjects in the control and SCC groups, while the sex ratio in the AC group was almost balanced. No control subjects had a smoking history. Approximately one-quarter of the patients with AC and approximately three-quarters of the patients with SCC were current or former smokers. The mean age of subjects in the control group was lower than that of patients in both the AC and the SCC groups. Most of the lung cancer patients were in late cancer stages (stage III and stage IV), which could be due to difficulty in the early diagnosis of lung cancer.

Table 2. Characteristics of the healthy control subjects, patients with adenocarcinoma (AC), and patients with squamous cell carcinoma (SCC)

Characteristics	**Control (*n* = 13)**	**AC (*n* = 50)**	**SCC (*n* = 22)**
Male	10 (76.9%)	24 (48%)	18 (81.8%)
Smoking	0	13 (26%)	16 (72.7%)
Age, mean ± SD	46.00 ± 19.54	62.02 ± 11.12	67.64 ± 9.87
Stage I	*Not applicable*	6 (12%)	1 (4.5%)
Stage II	*Not applicable*	0	2 (9.1%)
Stage III	*Not applicable*	9 (18%)	9 (40.9%)
Stage IV	*Not applicable*	35 (70%)	10 (45.5%)

doi:10.6342/NTU202504043

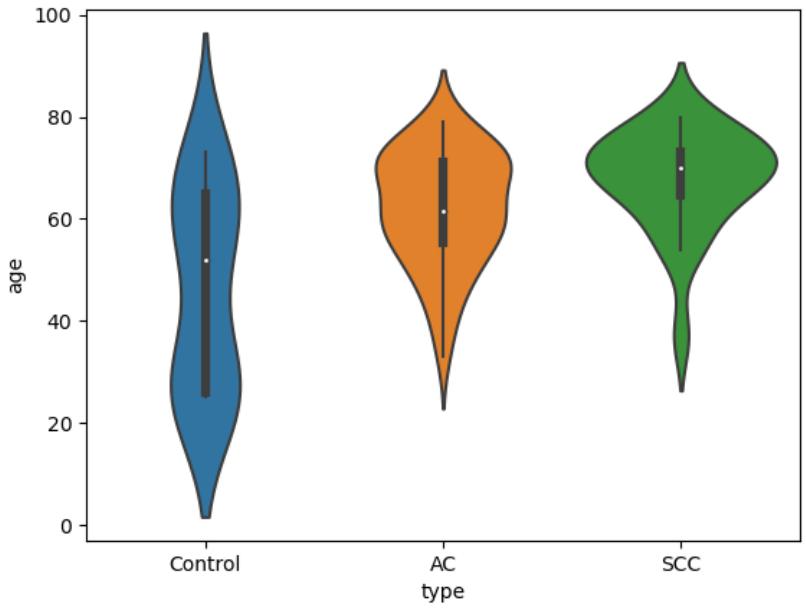

Figure 1. Violin plot of the subjects' ages

A heatmap of the VOCs in the subjects is shown in Figure 2. Each row represents a VOC, and each column represents a VOC profile of a subject. Some names of the VOCs are truncated because of space limitations. The cell color indicates the scaled level (GC–MS peak intensity) of the VOC in the subject sample. Redder cells indicate higher levels, and bluer cells indicate lower levels. The bar on the top of the figure shows each subject's group. Orange represents the control group, green represents the AC group, and blue represents the SCC group. The dendrograms on the top and left of the figure show hierarchical clustering results. The rows and columns are ordered so that similar rows and columns are clustered together.

Figure 2 shows that all the control subjects were clustered together, while the patients with AC and the patients with SCC were not separated by the clustering algorithm. This indicates that the VOC profiles of the control subjects may differ greatly from those of the

doi:10.6342/NTU202504043

lung cancer patients, whereas the distinction between the patients with AC and the patients with SCC may be less apparent.

It is worth noting that some compound names in our analysis represent isomers. These isomers were indistinguishable according to our GC–MS analysis. For example, xylene can include o-xylene, m-xylene, or p-xylene. According to the calculated mass spectrum, C_3-Benzenes can include trimethylbenzenes, ethyltoluenes, or cumene, but not other isomers.

doi:10.6342/NTU202504043

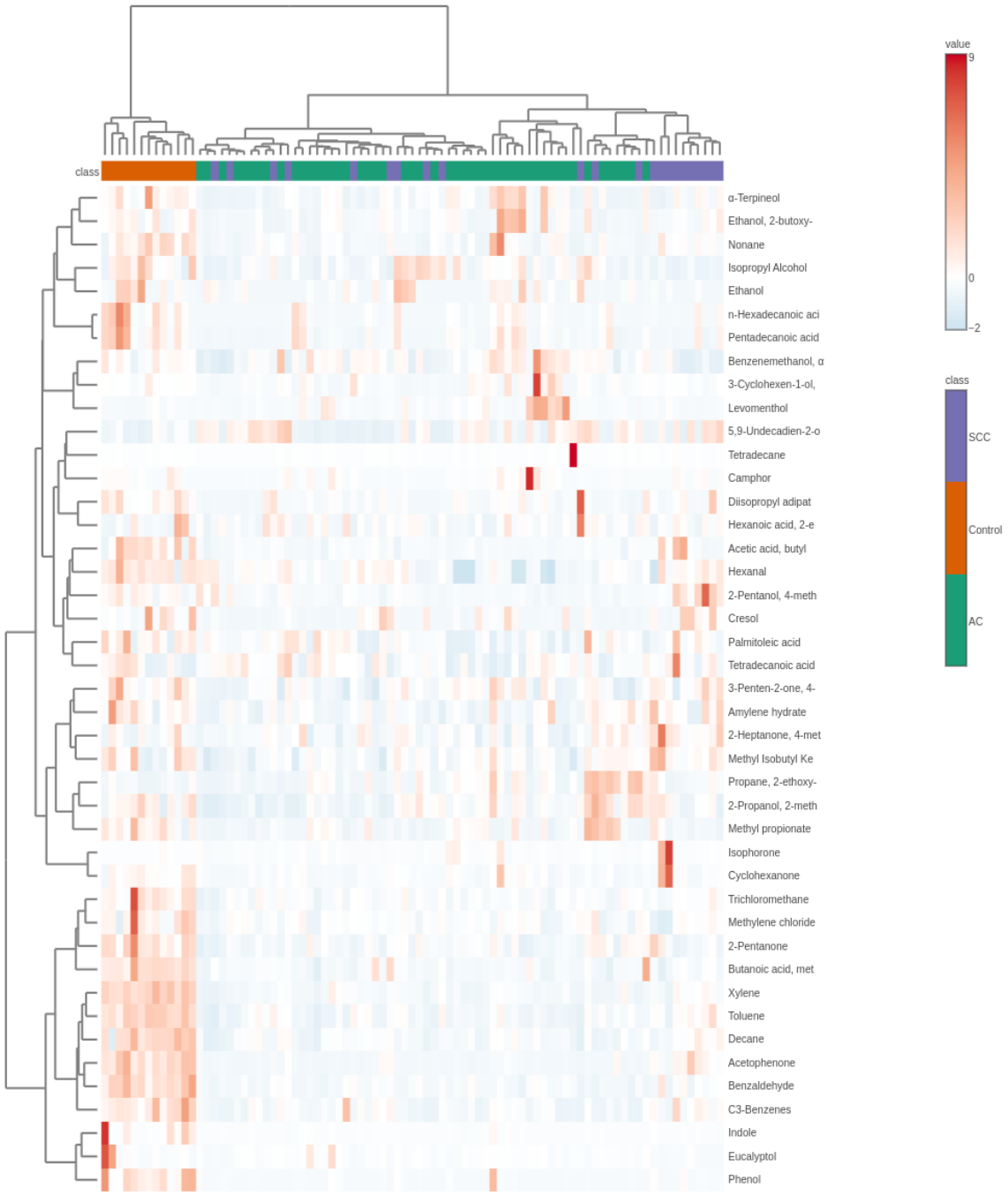

Figure 2. Heatmap of the detected VOCs

doi:10.6342/NTU202504043

3.2 Univariate Analysis of the VOC Profiles

3.2.1 ANCOVA Identified Some VOCs That Have Different Levels among Control Subjects, Patients with Adenocarcinoma, and Patients with Squamous Cell Carcinoma

Table 3 shows the results of ANCOVA and pairwise comparisons among three groups: control subjects (Group C), patients with AC (Group AC), and patients with SCC (Group SCC). ANCOVA revealed that 25 of the 43 VOCs exhibited significant differences between the groups. According to the results of pairwise comparisons, the most significant differences were observed between Group C and either Group AC or Group SCC. When we compared Group AC with Group C, 16 VOCs exhibited significant *p*-values and fold changes. When we compared Group SCC with Group C, 16 VOCs exhibited significant *p*-values and fold changes. When we compared Group AC with Group SCC, only 1 VOC exhibited a significant *p*-value and a significant fold change.

Table 3. Results of ANCOVA and pairwise comparisons among the control subjects, patients with AC, and patients with SCC

VOC	p_{ANCOVA}	$FC_{AC/C}$	$p_{AC–C}$	$FC_{SCC/C}$	$p_{SCC–C}$	$FC_{AC/SCC}$	$p_{AC–SCC}$
Ethanol	0.508	0.57	0.511	0.56	0.615	1.03	0.998
Isopropyl Alcohol	0.08	0.59	0.075	0.53	0.101	1.11	0.913
Methylene chloride	**<0.0001**	0.66	**<0.0001**	0.77	**0.005**	0.85	0.046
2-Propanol, 2-methyl-	0.126	0.66	0.116	0.74	0.433	0.89	0.817
Propane, 2-ethoxy-2-methyl-	0.425	1.39	0.653	1.05	0.995	1.32	0.587
Trichloromethane	**<0.0001**	0.5	**<0.0001**	0.57	**<0.0001**	0.88	0.43

doi:10.6342/NTU202504043

Methyl propionate	**<0.0001**	**0.42**	**<0.0001**	**0.24**	**<0.0001**	1.72	0.098
Amylene hydrate	**0.000467**	0.56	**0.000413**	0.74	0.154	0.76	0.125
2-Pentanone	**<0.0001**	**0.39**	**<0.0001**	**0.38**	**<0.0001**	1.02	0.995
Butanoic acid, methyl ester	**<0.0001**	**0.34**	**<0.0001**	**0.38**	**<0.0001**	0.89	0.31
Methyl Isobutyl Ketone	0.018	0.67	0.039	0.86	0.678	0.78	0.17
2-Pentanol, 4-methyl-	**<0.0001**	**0.38**	**<0.0001**	0.66	**0.004**	0.58	**0.000199**
Toluene	**<0.0001**	**0.36**	**<0.0001**	**0.47**	**<0.0001**	0.76	0.015
3-Penten-2-one, 4-methyl-	**0.003**	0.65	0.014	0.89	0.735	0.73	0.042
Hexanal	**0.000129**	0.65	**<0.0001**	0.78	0.075	0.83	0.123
Acetic acid, butyl ester	**<0.0001**	**0.27**	**<0.0001**	**0.31**	**<0.0001**	0.88	0.341
Xylene	**<0.0001**	**0.25**	**<0.0001**	**0.31**	**<0.0001**	0.79	0.055
Cyclohexanone	**0.000129**	0.53	**<0.0001**	0.54	**0.000861**	0.98	0.988
Nonane	**<0.0001**	**0.28**	**<0.0001**	**0.3**	**<0.0001**	0.93	0.947
Ethanol, 2-butoxy-	0.184	0.61	0.168	0.58	0.256	1.04	0.988
2-Heptanone, 4-methyl-	0.036	0.83	0.661	1.23	0.625	0.67	0.028
Benzaldehyde	**<0.0001**	**0.19**	**<0.0001**	**0.22**	**<0.0001**	0.85	0.664
Phenol	**<0.0001**	**0.05**	**<0.0001**	**0.04**	**<0.0001**	1.14	0.941
Decane	**<0.0001**	**0.43**	**<0.0001**	**0.47**	**<0.0001**	0.93	0.582
C_3-benzenes	**<0.0001**	**0.38**	**<0.0001**	**0.45**	**<0.0001**	0.85	0.457
Eucalyptol	0.058	1.16	0.834	0.63	0.531	1.85	0.041
Acetophenone	**<0.0001**	**0.07**	**<0.0001**	**0.19**	**<0.0001**	**0.37**	**0.002**
Cresol	0.01	**0.46**	0.01	0.67	0.311	0.68	0.32
Hexanoic acid, 2-ethyl-	0.066	1.79	0.335	**2.57**	0.055	0.7	0.177
Isophorone	0.066	0.67	0.048	0.72	0.216	0.94	0.912
3-Cyclohexen-1-ol, 1-methyl-4-(1-methylethyl)-	0.076	0.6	0.057	0.65	0.226	0.92	0.932

doi:10.6342/NTU202504043

Benzenemethanol, α-ethyl-	0.024	0.98	0.99	0.58	0.143	1.69	0.015
Camphor	0.08	**0.44**	0.064	**0.49**	0.226	0.88	0.952
Levomenthol	0.186	1.23	0.723	0.83	0.892	1.48	0.207
α-Terpineol	0.024	0.57	0.016	0.51	0.028	1.11	0.882
Indole	**<0.0001**	**0.16**	**<0.0001**	**0.13**	**<0.0001**	1.26	0.705
Tetradecane	**0.000963**	0.68	**0.002**	0.56	**0.000316**	1.21	0.239
Diisopropyl adipate	**0.000147**	0.69	**<0.0001**	0.71	**0.002**	0.96	0.891
5,9-Undecadien-2-one, 6,10-dimethyl-	**0.003**	**30.88**	0.07	**57.56**	**0.002**	0.54	0.031
Tetradecanoic acid	0.416	1.06	0.982	1.37	0.599	0.77	0.411
Pentadecanoic acid	**<0.0001**	**0.33**	**<0.0001**	**0.22**	**<0.0001**	1.54	0.268
Palmitoleic acid	0.014	0.57	0.012	0.72	0.298	0.78	0.388
n-Hexadecanoic acid	**<0.0001**	**0.16**	**<0.0001**	**0.08**	**<0.0001**	1.94	0.336

p_{ANCOVA}: ANCOVA *p*-values, adjusted using the Benjamini–Hochberg procedure

p_{AC-C}, p_{SCC-C}, p_{AC-SCC}: pairwise *p*-values, adjusted using Scheffé's method for each VOC

$FC_{AC/C}$, $FC_{SCC/C}$, $FC_{AC/SCC}$: pairwise fold changes

p-values below 0.01 and FC below 0.5 or above 2 are considered significant and marked in bold

3.2.2 ANCOVA Identified Some VOCs That Have Different Levels between Lung Cancer Stages

We divided the lung cancer patients into two groups: early-stage patients (Group E), including stage I and stage II patients, and late-stage patients (Group L), including stage III and stage IV patients. There were nine patients in Group E and 63 patients in Group L. We then analyzed the two groups and the control subjects (Group C). Table 4 shows the results of ANCOVA and pairwise comparisons. ANCOVA revealed that 28 of the 43 VOCs

doi:10.6342/NTU202504043

exhibited significant differences across the groups. In terms of the number of VOCs showing significant *p*-values and fold changes in pairwise comparisons, the largest groupwise difference was observed between Group L and Group C (17 VOCs), followed by Group E and Group C (15 VOCs), and then Group L and Group E (5 VOCs).

Table 4. Results of ANCOVA and pairwise comparisons among the control subjects, early-stage patients, and late-stage patients

VOC	p_{ANCOVA}	$FC_{E/C}$	p_{E-C}	$FC_{L/C}$	p_{L-C}	$FC_{L/E}$	p_{L-E}
Ethanol	0.332	0.93	0.988	0.53	0.45	0.57	0.548
Isopropyl Alcohol	0.039	0.8	0.688	0.55	0.048	0.69	0.38
Methylene chloride	**<0.0001**	0.71	**0.003**	0.67	**<0.0001**	0.94	0.812
2-Propanol, 2-methyl-	0.031	0.99	0.999	0.64	0.101	0.65	0.1
Propane, 2-ethoxy-2-methyl-	0.034	**2.42**	0.061	1.25	0.869	0.52	0.036
Trichloromethane	**<0.0001**	0.53	**<0.0001**	0.52	**<0.0001**	0.98	0.99
Methyl propionate	**<0.0001**	**0.47**	**0.002**	**0.36**	**<0.0001**	0.78	0.655
Amylene hydrate	**0.001**	0.7	0.118	0.57	**0.000542**	0.81	0.452
2-Pentanone	**<0.0001**	**0.47**	**<0.0001**	**0.38**	**<0.0001**	0.8	0.51
Butanoic acid, methyl ester	**<0.0001**	**0.34**	**<0.0001**	**0.35**	**<0.0001**	1.03	0.964
Methyl Isobutyl Ketone	0.033	0.86	0.708	0.68	0.042	0.79	0.327
2-Pentanol, 4-methyl-	**<0.0001**	**0.42**	**<0.0001**	**0.41**	**<0.0001**	0.99	0.999
Toluene	**<0.0001**	**0.42**	**<0.0001**	**0.37**	**<0.0001**	0.89	0.664
3-Penten-2-one, 4-methyl-	**<0.0001**	1.08	0.848	0.66	**0.006**	0.61	**0.000285**
Hexanal	**0.000478**	0.64	**0.002**	0.67	**0.00021**	1.06	0.887
Acetic acid, butyl ester	**<0.0001**	**0.29**	**<0.0001**	**0.28**	**<0.0001**	0.97	0.967

doi:10.6342/NTU202504043

Xylene	**<0.0001**	**0.31**	**<0.0001**	**0.25**	**<0.0001**	0.81	0.236
Cyclohexanone	**<0.0001**	0.77	0.173	0.5	**<0.0001**	0.65	0.016
Nonane	**<0.0001**	**0.42**	**<0.0001**	**0.28**	**<0.0001**	0.66	0.214
Ethanol, 2-butoxy-	**<0.0001**	1.38	0.324	0.52	0.05	**0.38**	**<0.0001**
2-Heptanone, 4-methyl-	0.788	0.85	0.787	0.89	0.825	1.06	0.962
Benzaldehyde	**<0.0001**	**0.21**	**<0.0001**	**0.19**	**<0.0001**	0.9	0.895
Phenol	**<0.0001**	**0.06**	**<0.0001**	**0.04**	**<0.0001**	0.74	0.741
Decane	**<0.0001**	**0.48**	**<0.0001**	**0.43**	**<0.0001**	0.91	0.577
C_3-benzenes	**<0.0001**	**0.47**	**<0.0001**	**0.39**	**<0.0001**	0.83	0.51
Eucalyptol	0.955	0.99	1	1.06	0.979	1.07	0.969
Acetophenone	**<0.0001**	**0.12**	**<0.0001**	**0.08**	**<0.0001**	0.65	0.308
Cresol	**0.000158**	1.03	0.992	**0.42**	**0.004**	**0.41**	**0.001**
Hexanoic acid, 2-ethyl-	0.341	1.73	0.412	1.62	0.353	0.93	0.963
Isophorone	**0.006**	0.97	0.981	0.64	0.027	0.67	0.043
3-Cyclohexen-1-ol, 1-methyl-4-(1-methylethyl)-	0.065	0.77	0.608	0.59	0.066	0.76	0.555
Benzenemethanol, α-ethyl-	0.033	1.39	0.302	0.86	0.785	0.62	0.022
Camphor	0.031	0.75	0.688	**0.43**	0.037	0.57	0.323
Levomenthol	0.661	1.38	0.634	1.17	0.866	0.84	0.776
α-Terpineol	**<0.0001**	1.34	0.188	0.51	**0.003**	**0.38**	**<0.0001**
Indole	**<0.0001**	**0.26**	**<0.0001**	**0.15**	**<0.0001**	0.57	0.074
Tetradecane	**0.005**	0.6	**0.007**	0.67	**0.003**	1.11	0.791
Diisopropyl adipate	**<0.0001**	0.76	0.011	0.67	**<0.0001**	0.89	0.344
5,9-Undecadien-2-one, 6,10-dimethyl-	0.025	**4.25**	0.55	**7.37**	0.023	1.74	0.381
Tetradecanoic acid	0.16	0.62	0.613	1.2	0.811	1.94	0.143
Pentadecanoic acid	**<0.0001**	0.58	**0.002**	**0.27**	**<0.0001**	**0.46**	**0.002**
Palmitoleic acid	0.026	0.5	0.028	0.62	0.033	1.24	0.702

doi:10.6342/NTU202504043

n-Hexadecanoic acid	**<0.0001**	**0.38**	**<0.0001**	**0.13**	**<0.0001**	**0.33**	**0.005**

p_{ANCOVA}: ANCOVA *p*-values, adjusted using the Benjamini–Hochberg procedure

$p_{E–C}$, $p_{L–C}$, $p_{L–E}$: pairwise *p*-values, adjusted using Scheffé's method for each VOC

$FC_{E/C}$, $FC_{L/C}$, $FC_{L/E}$: pairwise fold changes

p-values below 0.01 and FC below 0.5 or above 2 are considered significant and marked in bold

3.3 Multivariate Analysis of the VOC Profiles

3.3.1 PCA and PLS-DA Could Discriminate between VOC Profiles in Lung Cancer Patients and Control Subjects

Figure 3 shows all subjects' two-dimensional PCA projection of the VOC profiles. The x-axis and y-axis represent the first and second principal components, respectively. The number in parentheses next to each axis label represents the proportion of data variance explained by the corresponding principal component. The ellipses under the data points represent the 95% confidence regions for the corresponding classes.

This figure shows that the projected data of the control subjects and those of the lung cancer patients each form a cluster, but the projected data of the patients with AC and those of the patients with SCC highly overlap. This demonstrates that PCA could easily distinguish the VOC profiles in control subjects from those in lung cancer patients. Still, PCA might be insufficient for discriminating between patients with AC and SCC.

doi:10.6342/NTU202504043

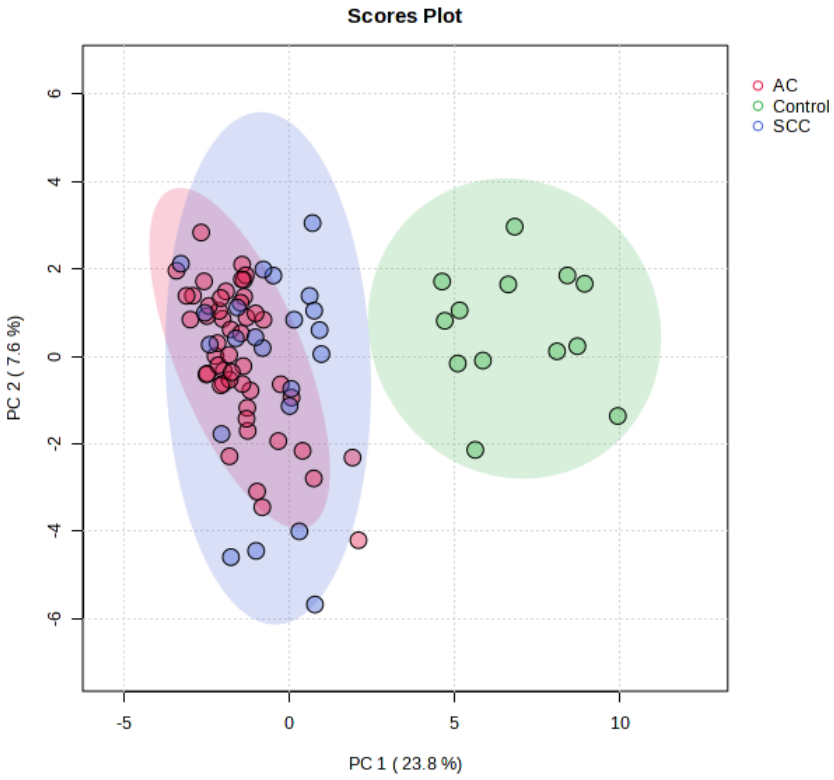

Figure 3. PCA score plot of the VOC profiles

We further conducted PLS-DA to build a classification model for discriminating between the control subjects and the whole group of lung cancer patients. Figure 4 shows the two-dimensional PLS-DA projection of the VOC profiles. The figure format is similar to that of Figure 3. The control subjects and the lung cancer patients are well separated on the plot. Figure 5 shows the 5-fold cross-validation scores of the PLS-DA model using different numbers of the first PLS-DA components. The R^2 score assesses how well the model explains the training data, while the accuracy and Q^2 score assess how well the model performs on the validation data. The asterisk indicates the highest Q^2 score. This figure shows that this model always attains high accuracies and Q^2 scores, indicating that it can sufficiently discriminate between control subjects and lung cancer patients and may have good generalizability.

doi:10.6342/NTU202504043

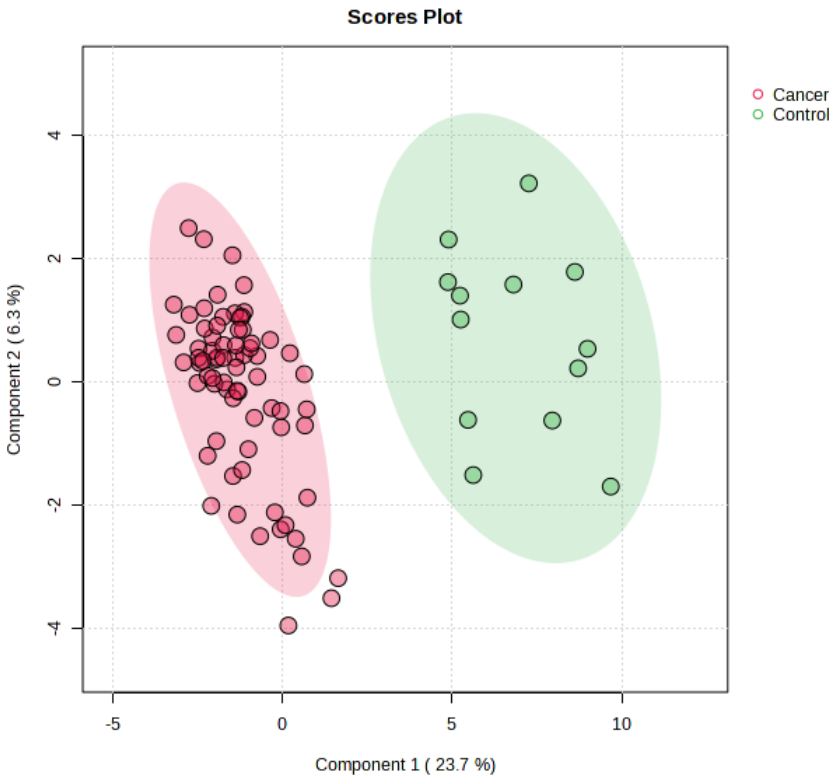

Figure 4. PLS-DA score plot of the VOC profiles in the control subjects and the lung cancer patients

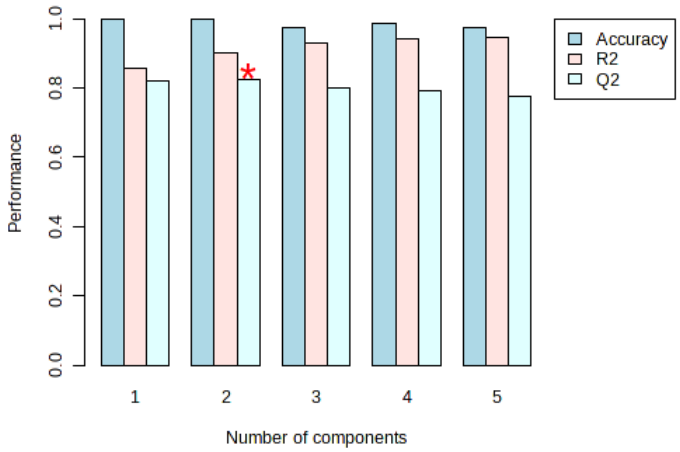

Figure 5. Cross-validation of the PLS-DA model for the control subjects and the lung cancer patients

doi:10.6342/NTU202504043

3.3.2 PLS-DA Failed to Discriminate between VOC Profiles in Patients with AC and Patients with SCC

We subsequently built a PLS-DA model for discriminating between patients with AC and patients with SCC. Figure 6 shows the two-dimensional PLS-DA projection of their VOC profiles. The patients with different cancer subtypes are not well separated on the plot. In addition, the first two PLS-DA components explain only 16.7% of the variance, which may indicate poor performance of the model. Figure 7 shows the 5-fold cross-validation scores of the model using different numbers of the first PLS-DA components. The model achieves the highest accuracy and Q^2 score when only the first PLS-DA component is used. However, considering that the majority class classifier, which always predicts the "AC" class regardless of the input, can achieve an accuracy of 69.4%, the accuracy of our model is not high. Additionally, the highest Q^2 score is low, showing that the model has poor generalizability. Our PLS-DA model cannot distinguish between AC and SCC patients.

doi:10.6342/NTU202504043

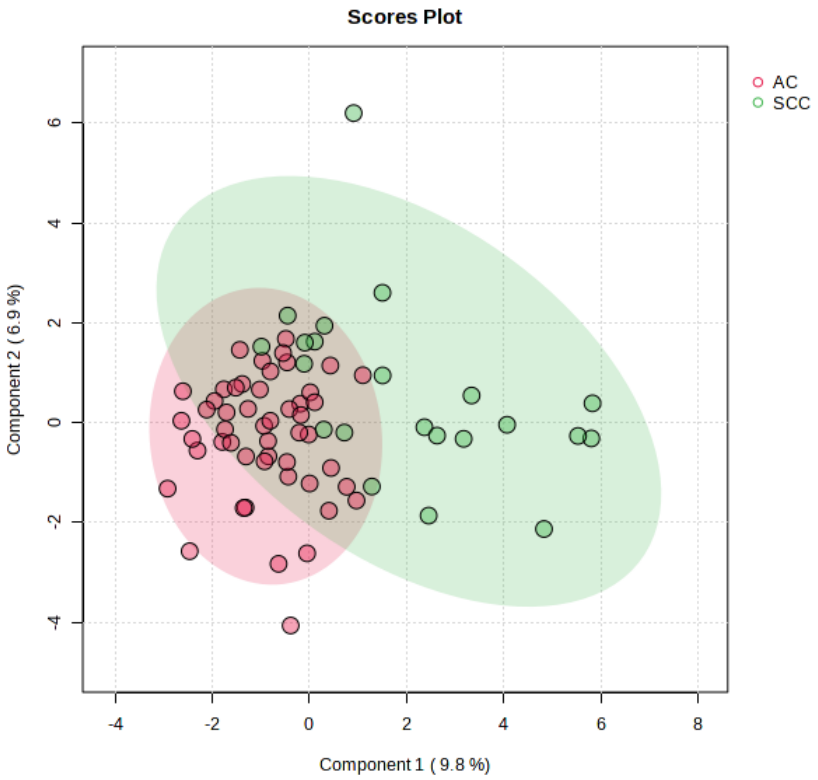

Figure 6. PLS-DA score plot of the VOC profiles in the patients with different cancer subtypes

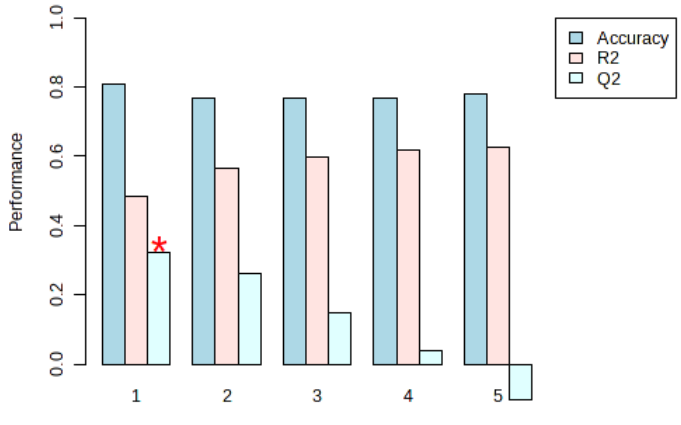

Figure 7. Cross-validation of the PLS-DA model for the patients with different cancer subtypes

doi:10.6342/NTU202504043

3.3.3 PLS-DA Failed to Discriminate between VOC Profiles in Patients with Different Cancer Stages

We also used PLS-DA to construct a model for distinguishing patients with different cancer stages. Figure 8 shows the two-dimensional PLS-DA projection of their VOC profiles. The patients in different stages are not well separated on the plot. Moreover, only 17.7% of the variance is explained by the first two PLS-DA components, indicating that the model may perform poorly. Figure 9 shows the 5-fold cross-validation of the model using different numbers of the first PLS-DA components. Considering that the majority class classifier, which always predicts the "late" class regardless of the input, will have an accuracy of 87.5%, the accuracy of our model is low, regardless of how many PLS-DA components are used.

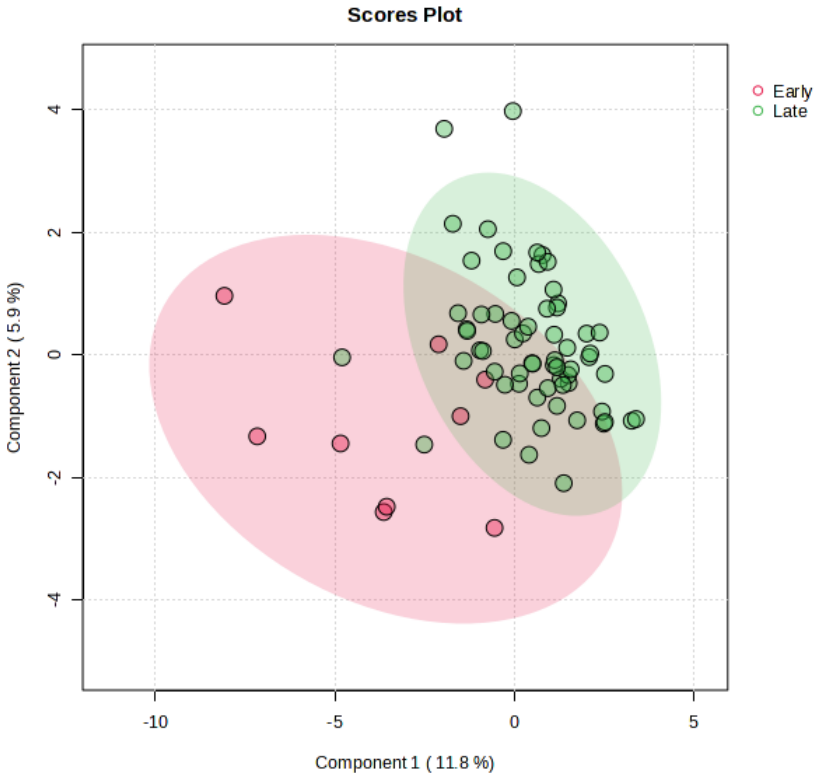

Figure 8. PLS-DA score plot of the VOC profiles in the patients with different cancer stages

doi:10.6342/NTU202504043

Furthermore, the highest Q^2 score is low, indicating that our model does not generalize well. Our PLS-DA model could not distinguish between patients with different cancer stages.

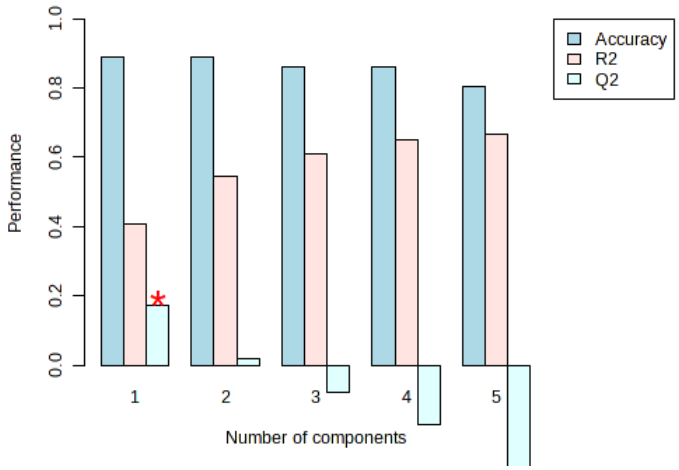

Figure 9. Cross-validation of the PLS-DA model for patients with different cancer stages

3.4 Potential VOC Biomarkers of Lung Cancer

Univariate and multivariate analyses demonstrated that some VOCs are more notable than others when we aim to differentiate between lung cancer patients and healthy individuals. These VOCs could be potential biomarkers for lung cancer diagnosis. We selected some potential biomarkers based on two criteria. First, the fold changes and *p*-values should always be significant (fold changes above 2 or below 0.5; *p*-values below 0.01) when we compare the control subjects with AC and SCC patients. Second, when we compare the control subjects and the lung cancer patients by using PLS-DA (Section 3.3.1), the variable importance in projection (VIP) score of the VOC should be greater than 1. Any VOC that met either criterion was selected and summarized in Table 5. It is worth noting that

doi:10.6342/NTU202504043

acetophenone exhibited a significant difference between AC and SCC patients ($FC_{AC/SCC}$ = 0.37; *p*-value = 0.002), suggesting its potential as a biomarker capable of simultaneously distinguishing among healthy individuals, AC patients, and SCC patients.

Table 5. Potential biomarkers for distinguishing between lung cancer patients and healthy individuals

VOC	$FC_{AC/C}$	$p_{AC–C}$	$FC_{SCC/C}$	$p_{SCC–C}$	VIP
Methylene chloride	0.66	**<0.0001**	0.77	**0.005**	**1.1109**
Trichloromethane	0.5	**<0.0001**	0.57	**<0.0001**	**1.3013**
Methyl propionate	**0.42**	**<0.0001**	**0.24**	**<0.0001**	0.6705
2-Pentanone	**0.39**	**<0.0001**	**0.38**	**<0.0001**	**1.489**
Butanoic acid, methyl ester	**0.34**	**<0.0001**	**0.38**	**<0.0001**	**1.5465**
Toluene	**0.36**	**<0.0001**	**0.47**	**<0.0001**	**1.8674**
Hexanal	0.65	**<0.0001**	0.78	0.075	**1.1148**
Acetic acid, butyl ester	**0.27**	**<0.0001**	**0.31**	**<0.0001**	**1.1436**
Xylene	**0.25**	**<0.0001**	**0.31**	**<0.0001**	**2.0557**
Nonane	**0.28**	**<0.0001**	**0.3**	**<0.0001**	0.96868
Benzaldehyde	**0.19**	**<0.0001**	**0.22**	**<0.0001**	**1.9016**
Phenol	**0.05**	**<0.0001**	**0.04**	**<0.0001**	**1.4723**
Decane	**0.43**	**<0.0001**	**0.47**	**<0.0001**	**1.7927**
C_3-benzenes	**0.38**	**<0.0001**	**0.45**	**<0.0001**	**1.4562**
Acetophenone	**0.07**	**<0.0001**	**0.19**	**<0.0001**	**1.8123**
Indole	**0.16**	**<0.0001**	**0.13**	**<0.0001**	0.98017
Pentadecanoic acid	**0.33**	**<0.0001**	**0.22**	**<0.0001**	**1.1275**
n-Hexadecanoic acid	**0.16**	**<0.0001**	**0.08**	**<0.0001**	**1.2123**

$FC_{AC/C}$, $FC_{SCC/C}$: fold changes between the patients with AC or SCC and the control subjects
$p_{AC–C}$, $p_{SCC–C}$: *p*-values obtained from the comparisons between patients with AC or SCC and the control subjects
VIP: PLS-DA variable importance in projection

doi:10.6342/NTU202504043

Most of the listed potential biomarkers have been reported in previous lung cancer breath analysis studies. Table 6 shows the number of occurrences and reported directions of level changes of these VOCs in the literature according to the scoping review by Schmidt *et al.*[33]. In our study, all proposed potential biomarkers exhibited lower levels in lung cancer patients than in healthy subjects. However, some of the previous studies reported opposite results. The discrepancies may be due to different experimental designs and require further research.

Table 6. The number of occurrences and reported directions of level changes in lung cancer patients of our proposed potential lung cancer biomarkers in the literature

VOC	**Count**	**# Increase**	**# Decrease**	**# Equal**	**# N/A**	**Our study**
Methylene chloride	2	0	0	0	2	Decreased
Trichloromethane	1	1	0	0	0	Decreased
Methyl propionate	0	0	0	0	0	Decreased
2-Pentanone	10	5	1	1	3	Decreased
Butanoic acid, methyl ester	0	0	0	0	0	Decreased
Toluene	18	8	3	1	6	Decreased
Hexanal	17	7	1	4	5	Decreased
Acetic acid, butyl ester	3	1	1	0	1	Decreased
Xylene*	19	7	2	5	5	Decreased
Nonane	5	3	0	1	1	Decreased
Benzaldehyde	9	2	0	0	7	Decreased
Phenol	4	4	0	0	0	Decreased
Decane	7	1	2	2	2	Decreased
C_3-benzenes**	11	3	0	2	6	Decreased
Acetophenone	5	1	0	0	4	Decreased
Indole	2	0	1	0	1	Decreased

doi:10.6342/NTU202504043

Pentadecanoic acid	0	0	0	0	0	Decreased
n-Hexadecanoic acid	0	0	0	0	0	Decreased

*Xylene isomers were indistinguishable in our study. All of the possible isomers were accounted for

**According to the calculated mass spectrum, C_3-benzenes could include trimethylbenzenes, ethyltoluenes, or cumene. They were indistinguishable in our study. All of these compounds were accounted for.

Several factors are needed for a VOC to be a favorable lung cancer biomarker. One factor is whether the VOC is specific to lung cancer. It is known that some of the previously reported potential lung cancer biomarkers are also related to other lung diseases, including asthma and chronic obstructive pulmonary disease[34], and other types of cancer[35]. Another factor is whether the variation in the VOC level can be attributed to nonpathological causes. Some aromatic compounds, such as toluene and xylenes, are believed to be exogenous contaminants[33]. Isoprene, which was frequently reported in previous lung cancer breath analysis studies, has been linked to physical activity[36].

The metabolic pathways of lung cancer-related VOCs are mostly unclear, although some discussions exist. Alkanes, such as nonane and decane, could be linked to lipid peroxidation[36]. Aldehydes such as hexanal could be linked to alcohol metabolism, lipid peroxidation, and tobacco metabolism[36]. Benzaldehyde could be linked to fatty acid metabolism, glycolysis, glyconeogenesis, and tryptophan metabolism[37].

Among the potential biomarkers we proposed, the International Agency for Research on Cancer has evaluated the carcinogenicity of several compounds. Due to inadequate evidence for carcinogenicity in humans and animals, toluene, xylene, and phenol were classified as Group 3 (not classifiable) agents. However, some studies have demonstrated the

doi:10.6342/NTU202504043

potential genotoxicity of phenol and toluene[38]. More recent studies have suggested that exposure to toluene and xylene may be associated with an increased risk of lung cancer after adjusting for smoking. For example, Warden *et al.*[39] observed a slight increase in lung cancer risk among male subjects with occupational exposure, while Khorrami *et al.*[40] reported a positive association between ambient exposure to these compounds and lung cancer incidence. Trichloromethane and cumene, a possible C_3-benzene isomer, were classified as Group 2B (possibly carcinogenic) agents since there is sufficient evidence for their carcinogenicity in animals but inadequate evidence in humans[41], [42]. Methylene chloride was classified as a Group 2A (probably carcinogenic) agent because there is sufficient evidence for its carcinogenicity in animals, limited evidence in humans, and strong mechanistic evidence[43].

Although we failed to construct a PLS-DA model capable of discriminating between lung cancer stages, the univariate analysis still identified some VOCs that exhibited significant differences between early- and late-stage patients. Table 7 summarizes these VOCs.

Table 7. Potential biomarkers for distinguishing between early-stage patients and late-stage patients

VOC	**$FC_{L/E}$**	**$p_{L–E}$**
Ethanol, 2-butoxy-	**0.38**	**<0.0001**
Cresol	**0.41**	**0.001**
α-Terpineol	**0.38**	**<0.0001**
Pentadecanoic acid	**0.46**	**0.002**

doi:10.6342/NTU202504043

n-Hexadecanoic acid	**0.33**	**0.005**

$FC_{L/E}$: fold changes

p_{L-E}: *p*-values

To our knowledge, few breath analysis studies have reported potential VOC biomarkers for lung cancer staging, and none of our proposed potential biomarkers for staging have been mentioned in those studies. However, some studies have explored their relationship to cancer. Ethanol, 2-butoxy-, commonly used in paints and cleaning products, may be carcinogenic to mice but lacks sufficient evidence for its carcinogenicity to humans[44]. Wang *et al.*[45] found elevated levels of exhaled ethanol, 2-butoxy- in lung cancer patients both before and after lung cancer resection, and suggested that it may be unrelated to the tumor and likely of exogenous origin. Cresol is found in various sources such as food, automobile exhaust, tobacco smoke, preservatives, and disinfectants. There is no clear evidence of its carcinogenicity in mice and rats[46]. Peralbo-Molina *et al.*[47] noted lower p-cresol levels in the EBC of lung cancer patients than in healthy controls and suggested that this could stem from differences in toluene metabolism. α-Terpineol, widely used as a fragrance and a food flavoring, might have anticarcinogenic properties[48]. Pentadecanoic acid, found in dairy, plants, and fish, may inhibit lung cancer cell proliferation[49]. n-Hexadecanoic acid, abundantly present in the human body and originating from dietary intake and endogenous synthesis, has mixed evidence regarding its role in cancer. Some studies suggested that it may promote metastasis[50], [51], while others indicated that it may inhibit tumor cell proliferation and metastasis[52]. Blood metabolomic studies have reported inconsistent results on the n-hexadecanoic acid level differences between lung cancer patients and healthy controls. For instance, Miyamoto *et al.*[53] reported increased levels in lung cancer patients,

doi:10.6342/NTU202504043

while Qi *et al.*[54] reported decreased levels. Although the roles of these compounds in cancer are not definitive, these studies provide initial insights for further research.

3.5 Limitations

First, the sample size in our study was limited, and the distribution of the characteristics of the subjects was not well-balanced. In particular, the control group had a small sample size, no smoking history, and a lower average age than the lung cancer patients. Although the statistical methods we used can mitigate these challenges to some degree, having a larger and more balanced statistical sample could improve the generalizability of our findings. In addition, while we tried to control for several potential confounding factors, there could be other confounding factors that were not accounted for, which could impact the interpretation of our results.

Second, because different EBC analysis methods may yield different detected compounds, comparisons between our findings and those of other studies should be made with caution. In addition, the concentration of collected EBC samples may vary, which could affect the obtained levels of VOCs in subsequent GC–MS analysis. Studies have aimed to develop standardization techniques for non-volatile compounds in EBC, but no gold standard has been established[55].

Finally, the compound identification process from GC–MS data could be improved. The algorithms used to process GC–MS data occasionally generate inaccurate results, thus requiring manual verification. This can make the compound identification process laborious and prone to errors.

doi:10.6342/NTU202504043

Chapter 4 Conclusion

This study collected EBC samples from 72 lung cancer patients and 13 healthy individuals. The samples were then extracted using HS-SPME and analyzed by GC–MS. We found significant differences in the exhaled VOC profiles between lung cancer patients and healthy individuals through univariate and multivariate analyses. However, the differences between different lung cancer subtypes and stages were not pronounced. We identified 18 VOCs that may be potential biomarkers for lung cancer diagnosis. Additionally, 5 VOCs showed potential as biomarkers for lung cancer staging.

doi:10.6342/NTU202504043

References

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA A Cancer J Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] C. Allemani *et al.*, "Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries," *The Lancet*, vol. 391, no. 10125, pp. 1023–1075, Mar. 2018, doi: 10.1016/S0140-6736(17)33326-3.

[3] P. Goldstraw *et al.*, "The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer," *Journal of Thoracic Oncology*, vol. 11, no. 1, pp. 39–51, Jan. 2016, doi: 10.1016/j.jtho.2015.09.009.

[4] S. Blandin Knight, P. A. Crosbie, H. Balata, J. Chudziak, T. Hussell, and C. Dive, "Progress and prospects of early detection in lung cancer," *Open Biol.*, vol. 7, no. 9, p. 170070, Sep. 2017, doi: 10.1098/rsob.170070.

[5] L. M. Seijo *et al.*, "Biomarkers in Lung Cancer Screening: Achievements, Promises, and Challenges," *Journal of Thoracic Oncology*, vol. 14, no. 3, pp. 343–357, Mar. 2019, doi: 10.1016/j.jtho.2018.11.023.

[6] M. M. Oken *et al.*, "Screening by Chest Radiograph and Lung Cancer Mortality: The Prostate, Lung, Colorectal, and Ovarian (PLCO) Randomized Trial," *JAMA*, vol. 306, no. 17, p. 1865, Nov. 2011, doi: 10.1001/jama.2011.1591.

[7] H. J. De Koning *et al.*, "Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial," *N Engl J Med*, vol. 382, no. 6, pp. 503–513, Feb. 2020, doi: 10.1056/NEJMoa1911793.

[8] S. J. Adams, E. Stone, D. R. Baldwin, R. Vliegenthart, P. Lee, and F. J. Fintelmann, "Lung cancer screening," *The Lancet*, vol. 401, no. 10374, pp. 390–408, Feb. 2023, doi: 10.1016/S0140-6736(22)01694-4.

[9] C. Freitas *et al.*, "The Role of Liquid Biopsy in Early Diagnosis of Lung Cancer," *Front. Oncol.*, vol. 11, p. 634316, Apr. 2021, doi: 10.3389/fonc.2021.634316.

[10] W. Li *et al.*, "Liquid biopsy in lung cancer: significance in diagnostics, prediction, and treatment monitoring," *Mol Cancer*, vol. 21, no. 1, p. 25, Dec. 2022, doi: 10.1186/s12943-022-01505-z.

[11] A. Krilaviciute, J. A. Heiss, M. Leja, J. Kupcinskas, H. Haick, and H. Brenner, "Detection of cancer through exhaled breath: a systematic review," *Oncotarget*, vol. 6, no. 36, pp. 38643–38657, Nov. 2015, doi: 10.18632/oncotarget.5938.

[12] W. Ibrahim *et al.*, "Breathomics for the clinician: the use of volatile organic compounds in respiratory diseases," *Thorax*, vol. 76, no. 5, pp. 514–521, May 2021, doi: 10.1136/thoraxjnl-2020-215667.

[13] P. Wang *et al.*, "Identification of lung cancer breath biomarkers based on perioperative breathomics testing: A prospective observational study," *eClinicalMedicine*, vol. 47, p. 101384, May 2022, doi: 10.1016/j.eclinm.2022.101384.

[14] Y. Saalberg and M. Wolff, "VOC breath biomarkers in lung cancer," *Clinica Chimica Acta*, vol. 459, pp. 5–9, Aug. 2016, doi: 10.1016/j.cca.2016.05.013.

[15] M. McCulloch, T. Jezierski, M. Broffman, A. Hubbard, K. Turner, and T. Janecki,

doi:10.6342/NTU202504043

"Diagnostic Accuracy of Canine Scent Detection in Early- and Late-Stage Lung and Breast Cancers," *Integr Cancer Ther*, vol. 5, no. 1, pp. 30–39, Mar. 2006, doi: 10.1177/1534735405285096.
[16] R. Ehmann *et al.*, "Canine scent detection in the diagnosis of lung cancer: revisiting a puzzling phenomenon," *European Respiratory Journal*, vol. 39, no. 3, pp. 669–676, Mar. 2012, doi: 10.1183/09031936.00051711.
[17] N. Peled *et al.*, "Non-invasive Breath Analysis of Pulmonary Nodules," *Journal of Thoracic Oncology*, vol. 7, no. 10, pp. 1528–1533, Oct. 2012, doi: 10.1097/JTO.0b013e3182637d5f.
[18] H. Lemjabbar-Alaoui, O. U. Hassan, Y.-W. Yang, and P. Buchanan, "Lung cancer: Biology and treatment options," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1856, no. 2, pp. 189–210, Dec. 2015, doi: 10.1016/j.bbcan.2015.08.002.
[19] Y. Zhang *et al.*, "Global variations in lung cancer incidence by histological subtype in 2020: a population-based study," *The Lancet Oncology*, vol. 24, no. 11, pp. 1206–1218, Nov. 2023, doi: 10.1016/S1470-2045(23)00444-8.
[20] P. J. Mazzone *et al.*, "Exhaled Breath Analysis with a Colorimetric Sensor Array for the Identification and Characterization of Lung Cancer," *Journal of Thoracic Oncology*, vol. 7, no. 1, pp. 137–142, Jan. 2012, doi: 10.1097/JTO.0b013e318233d80f.
[21] C. Wang *et al.*, "Exploratory study on classification of lung cancer subtypes through a combined K-nearest neighbor classifier in breathomics," *Sci Rep*, vol. 10, no. 1, p. 5880, Apr. 2020, doi: 10.1038/s41598-020-62803-4.
[22] R. Schmid *et al.*, "Integrative analysis of multimodal mass spectrometry data in MZmine 3," *Nat Biotechnol*, vol. 41, no. 4, pp. 447–449, Apr. 2023, doi: 10.1038/s41587-023-01690-2.
[23] A. Smirnov, Y. Qiu, W. Jia, D. I. Walker, D. P. Jones, and X. Du, "ADAP-GC 4.0: Application of Clustering-Assisted Multivariate Curve Resolution to Spectral Deconvolution of Gas Chromatography–Mass Spectrometry Metabolomics Data," *Anal. Chem.*, vol. 91, no. 14, pp. 9069–9077, Jul. 2019, doi: 10.1021/acs.analchem.9b01424.
[24] S. E. Stein and D. R. Scott, "Optimization and testing of mass spectral library search algorithms for compound identification," *J. Am. Soc. Mass Spectrom.*, vol. 5, no. 9, pp. 859–866, Sep. 1994, doi: 10.1016/1044-0305(94)87009-8.
[25] S. R. Searle, F. M. Speed, and G. A. Milliken, "Population Marginal Means in the Linear Model: An Alternative to Least Squares Means," *The American Statistician*, vol. 34, no. 4, pp. 216–221, Nov. 1980, doi: 10.1080/00031305.1980.10483031.
[26] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, Jan. 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.
[27] H. Scheffe, "A Method for Judging all Contrasts in the Analysis of Variance," *Biometrika*, vol. 40, no. 1/2, p. 87, Jun. 1953, doi: 10.2307/2333100.
[28] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2023. [Online]. Available: https://www.R-project.org/
[29] A. Kassambara, *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. 2023. [Online]. Available: https://CRAN.R-project.org/package=rstatix

doi:10.6342/NTU202504043

[30] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*. 2023. [Online]. Available: https://CRAN.R-project.org/package=emmeans
[31] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, Fourth. New York: Springer, 2002. [Online]. Available: https://www.stats.ox.ac.uk/pub/MASS4/
[32] Z. Pang *et al.*, "MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation," *Nucleic Acids Research*, p. gkae253, Apr. 2024, doi: 10.1093/nar/gkae253.
[33] F. Schmidt *et al.*, "Mapping the landscape of lung cancer breath analysis: A scoping review (ELCABA)," *Lung Cancer*, vol. 175, pp. 131–140, Jan. 2023, doi: 10.1016/j.lungcan.2022.12.003.
[34] I. A. Ratiu, T. Ligor, V. Bocos-Bintintan, C. A. Mayhew, and B. Buszewski, "Volatile Organic Compounds in Exhaled Breath as Fingerprints of Lung Cancer, Asthma and COPD," *JCM*, vol. 10, no. 1, p. 32, Dec. 2020, doi: 10.3390/jcm10010032.
[35] S. Janfaza, B. Khorsand, M. Nikkhah, and J. Zahiri, "Digging deeper into volatile organic compounds associated with cancer," *Biology Methods and Protocols*, vol. 4, no. 1, p. bpz014, Jan. 2019, doi: 10.1093/biomethods/bpz014.
[36] M. Hakim *et al.*, "Volatile Organic Compounds of Lung Cancer and Possible Biochemical Pathways," *Chem. Rev.*, vol. 112, no. 11, pp. 5949–5966, Nov. 2012, doi: 10.1021/cr300174a.
[37] D. Zimmermann, M. Hartmann, M. P. Moyer, J. Nolte, and J. I. Baumbach, "Determination of volatile products of human colon cell line metabolism by GC/MS analysis," *Metabolomics*, vol. 3, no. 1, pp. 13–17, Mar. 2007, doi: 10.1007/s11306-006-0038-y.
[38] International Agency for Research on Cancer, Ed., *Re-evaluation of some organic chemicals, hydrazine and hydrogen peroxide: this publication represents the views and expert opinions of an IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, which met in Lyon, 17 - 24 February 1998*. in IARC monographs on the evaluation of carcinogenic risks to humans, no. 71. Lyon: IARC, 1999.
[39] H. Warden, H. Richardson, L. Richardson, J. Siemiatycki, and V. Ho, "Associations between occupational exposure to benzene, toluene and xylene and risk of lung cancer in Montréal," *Occup Environ Med*, vol. 75, no. 10, pp. 696–702, Oct. 2018, doi: 10.1136/oemed-2017-104987.
[40] Z. Khorrami *et al.*, "Multiple air pollutant exposure and lung cancer in Tehran, Iran," *Sci Rep*, vol. 11, no. 1, p. 9239, Apr. 2021, doi: 10.1038/s41598-021-88643-4.
[41] International Agency for Research on Cancer, Ed., *Some chemicals that cause tumours of the kidney or urinary bladder in rodents and some other substances: this publication represents the views and expert opinions of an IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, which met in Lyon, 13 - 20 October 1998*. in IARC monographs on the evaluation of carcinogenic risks to humans, no. 73. Lyon: IARC, 1999.
[42] Centre international de recherche sur le cancer, Ed., *Some chemicals present in industrial and consumer products, food and drinking-water*. in IARC monographs on the evaluation of carcinogenic risks to humans, no. 101. Lyon: International agency for research on cancer, 2013.
[43] *Some chemicals used as solvents and in polymer manufacture*. Lyon, France: International Agency for Research on Cancer, World Health Organization, 2017.

doi:10.6342/NTU202504043

[44] International Agency for Research on Cancer, Ed., *Formaldehyde, 2-Butoxyethanol and 1-tert-Butoxypropan-2-ol: this publication represents the views and expert opinions of an IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, which met in Lyon, 2 - 9 June 2004*. in IARC monographs on the evaluation of carcinogenic risks to humans, no. 88. Lyon: IARC, 2006.
[45] C. Wang *et al.*, "Exhaled volatile organic compounds as lung cancer biomarkers during one-lung ventilation," *Sci Rep*, vol. 4, no. 1, p. 7312, Dec. 2014, doi: 10.1038/srep07312.
[46] J. M. Sanders, J. R. Bucher, J. C. Peckham, G. E. Kissling, M. R. Hejtmancik, and R. S. Chhabra, "Carcinogenesis studies of cresols in rats and mice," *Toxicology*, vol. 257, no. 1–2, pp. 33–39, Mar. 2009, doi: 10.1016/j.tox.2008.12.005.
[47] A. Peralbo-Molina, M. Calderón-Santiago, F. Priego-Capote, B. Jurado-Gámez, and M. D. Luque De Castro, "Identification of metabolomics panels for potential lung cancer screening by analysis of exhaled breath condensate," *J. Breath Res.*, vol. 10, no. 2, p. 026002, Mar. 2016, doi: 10.1088/1752-7155/10/2/026002.
[48] A. Sales, L. D. O. Felipe, and J. L. Bicas, "Production, Properties, and Applications of α-Terpineol," *Food Bioprocess Technol*, vol. 13, no. 8, pp. 1261–1279, Aug. 2020, doi: 10.1007/s11947-020-02461-6.
[49] M. K. Ediriweera, N. B. To, Y. Lim, and S. K. Cho, "Odd-chain fatty acids as novel histone deacetylase 6 (HDAC6) inhibitors," *Biochimie*, vol. 186, pp. 147–156, Jul. 2021, doi: 10.1016/j.biochi.2021.04.011.
[50] G. Pascual *et al.*, "Dietary palmitic acid promotes a prometastatic memory via Schwann cells," *Nature*, vol. 599, no. 7885, pp. 485–490, Nov. 2021, doi: 10.1038/s41586-021-04075-0.
[51] X. Zhang *et al.*, "Palmitic Acid Promotes Lung Metastasis of Melanomas via the TLR4/TRIF-Peli1-pNF-κB Pathway," *Metabolites*, vol. 12, no. 11, p. 1132, Nov. 2022, doi: 10.3390/metabo12111132.
[52] X. Wang, C. Zhang, and N. Bao, "Molecular mechanism of palmitic acid and its derivatives in tumor progression," *Front. Oncol.*, vol. 13, p. 1224125, Aug. 2023, doi: 10.3389/fonc.2023.1224125.
[53] S. Miyamoto *et al.*, "Systemic Metabolomic Changes in Blood Samples of Lung Cancer Patients Identified by Gas Chromatography Time-of-Flight Mass Spectrometry," *Metabolites*, vol. 5, no. 2, pp. 192–210, Apr. 2015, doi: 10.3390/metabo5020192.
[54] S. Qi *et al.*, "High-resolution metabolomic biomarkers for lung cancer diagnosis and prognosis," *Sci Rep*, vol. 11, no. 1, p. 11805, Jun. 2021, doi: 10.1038/s41598-021-91276-2.
[55] P. Kubáň and F. Foret, "Exhaled breath condensate: Determination of non-volatile compounds and their potential for clinical diagnosis and monitoring. A review," *Analytica Chimica Acta*, vol. 805, pp. 1–18, Dec. 2013, doi: 10.1016/j.aca.2013.07.049.

doi:10.6342/NTU202504043