



Swansea University  
Prifysgol Abertawe



## Cronfa - Swansea University Open Access Repository

---

This is an author produced version of a paper published in:  
*Biochimica et Biophysica Acta (BBA) - General Subjects*

Cronfa URL for this paper:  
<http://cronfa.swan.ac.uk/Record/cronfa29367>

---

### Paper:

O'Shea, K., Cameron, S., Lewis, K., Lu, C. & Mur, L. (2016). Metabolomic-based biomarker discovery for non-invasive lung cancer screening: A case study. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1860(11), 2682-2687.

<http://dx.doi.org/10.1016/j.bbagen.2016.07.007>

---

This item is brought to you by Swansea University. Any person downloading material is agreeing to abide by the terms of the repository licence. Copies of full text items may be used or reproduced in any format or medium, without prior permission for personal research or study, educational or non-commercial purposes only. The copyright for any work remains with the original author unless otherwise specified. The full-text must not be sold in any format or medium without the formal permission of the copyright holder.

Permission for multiple reproductions should be obtained from the original author.

Authors are personally responsible for adhering to copyright and publisher restrictions when uploading content to the repository.

<http://www.swansea.ac.uk/library/researchsupport/ris-support/>

# Metabolomic-based biomarker discovery for non-invasive lung cancer screening: A case study

Keiron O'Shea<sup>a</sup>, Simon J.S. Cameron<sup>a,b</sup>, Keir E Lewis<sup>c</sup>, Chuan Lu<sup>d</sup>, Luis AJ Mur<sup>a,\*</sup>

<sup>a</sup>*Institute of Biological, Environmental and Rural Studies, Aberystwyth University, Aberystwyth, Wales, SY23 3DA, UK*

<sup>b</sup>*Division of Computational and Systems Medicine, Department of Surgery and Cancer, Imperial College London, London, W6 8RP, UK*

<sup>c</sup>*Department of Respiratory Medicine, Prince Philip Hospital, Llanelli, Wales, SA14 8LY, UK*

<sup>d</sup>*Department of Computer Science, Aberystwyth University, Aberystwyth, Wales, SY23 3DA, UK*

---

## Abstract

**Background:** Lung cancer (LC) is one of the leading lethal cancers worldwide, with an estimated 18.4% of all cancer deaths being attributed to the disease. Despite developments in cancer diagnosis and treatment over the previous thirty years, LC has seen little to no improvement in the overall five year survival rate after initial diagnosis.

**Methods:** In this paper, we extended a recent study which profiled the metabolites in sputum from patients with lung cancer and age-matched volunteers smoking controls using flow infusion electrospray ion mass spectrometry. We selected key metabolites for distinguishing between different classes of lung cancer, and employed artificial neural networks and leave-one-out cross-validation to evaluate the predictive power of the identified biomarkers.

**Results:** The neural network model showed excellent performance in classification between lung cancer and control groups with the area under the receiver operating characteristic curve of 0.99. The sensitivity and specificity of for detecting cancer from controls were 96% and 94% respectively. Furthermore, we have identified six putative metabolites that were able to

---

\*Corresponding author

*Email address:* lum@aber.ac.uk (Luis AJ Mur)

discriminate between sputum samples derived from patients suffering small cell lung cancer (SCLC) and non-small cell lung cancer. These metabolites achieved excellent cross validation performance with a sensitivity of 80% and specificity of 100% for predicting SCLC.

**Conclusions:** These results indicate that sputum metabolic profiling may have potential for screening of lung cancer and lung cancer recurrence, and may greatly improve effectiveness of clinical intervention.

*Keywords:*

lung cancer, small vs non-small cell lung cancer, sputum, metabolomics, biomarkers, artificial neural networks

---

## 1. Introduction

The year 2008 saw an estimated 12.7 million new cases of cancer, and 7.6 million cancer-related deaths worldwide [1]. While the incidence and mortality rates of most cancers is decreasing in the developed world, they are rising in emerging economies such as China and India. Migrant studies have found that cancer rates in the descendent generation of migrants tends to shift toward the host country, suggesting that environmental risk factors such as smoking and weight are responsible for the global variance in cancer rates [2].

### 1.1. Lung cancer

Lung cancer is a major cause of death in the developed and developing worlds. It is the leading cause of cancer-related deaths in men, and second only to breast cancer in women. There was an estimated 1.6 million new cases of lung cancer and 1.4 million deaths in 2008. This accounts for 12.6% of all cancer incidence and a staggering 18.4% of all cancer-related deaths [2]. This can be attributed to its poor prognosis, with the five-year survival rate being a mere 15%. Despite recent advances in lung cancer treatment, survival rates are low when compared to other forms of cancer [3]. However, improvements in surgical techniques and chemotherapy over the past twenty years has resulted in one-year lung cancer survival rates drastically improving. Despite this, the overall five-year lung cancer survival rates have remained stagnant at 6% for small cell lung cancer and 18% for non-small cell lung cancer. Unfortunately the vast majority (85%) of lung cancer cases are

24 diagnosed at advanced stages, heavily reducing the effectiveness of treatment  
25 [1].

26 This can be attributed to the difficulty of effectively diagnosing cancer of  
27 the lung at stage early enough to make a real impact. One of the main diffi-  
28 culties is that symptoms of the conditions are often identical to less serious  
29 conditions. This makes the pre-clinical diagnosis of lung cancer particularly  
30 problematic as the observed symptoms are often confused with other respi-  
31 ratory conditions. Prognostic factors may help diagnose patients who show  
32 symptoms of a disease, or have an increased chance of recurrence or progres-  
33 sion to advanced disease which should support clinicians in the creation of  
34 appropriate treatment plans. The World Health Organisation (WHO) have  
35 set out ten key principles to be met by an effective screening procedure in  
36 order for it to be beneficial and cost effective [4]. Currently there are no lung  
37 cancer screening techniques of which meet all of the ten conditions laid out  
38 by the WHO.

### 39 *1.2. Metabolomic insights into lung cancer*

40 An emerging screening methodology to other traditional screening meth-  
41 ods is the utilisation of molecular biomarkers in biofluids. The ease of analy-  
42 sis of biofluids using mass spectrometry (MS) or nuclear magnetic resonance  
43 (NMR) makes metabolomics a well-suited methodology for the non-invasive  
44 detection of biomarkers in lung cancer. Current focus of metabolomics in lung  
45 cancer has been on the exploitation of serum, urine and tumour biopsies. For  
46 example, the analysis of serum using liquid chromatography (LC-MS) and  
47 gas chromatography (GC-MS) approaches have suggested a potential use for  
48 biomarkers of lung cancer. A small-scale pilot study sampling lung cancer pa-  
49 tients before and after surgical intervention, alongside patients without lung  
50 cancer has suggested ten candidate biomarkers for lung cancer, including  
51 sphingosine, oleic acid and serine [5].

52 Sputum has been suggested as a potential biofluid source of biomarkers  
53 in lung cancer [3, 6]. Recent work has used Fourier Transform Infra-Red  
54 (FTIR) spectroscopy as a non-invasive method to detect lung cancer in spu-  
55 tum samples. This work concluded that FTIR was able to sufficiently dis-  
56 tinguish between lung cancer and control samples, and effectively act as a  
57 non-invasive, high-throughput and cost-effective method for screening spu-  
58 tum samples from high-risk patients. Furthermore, it further validated the  
59 use of sputum as an effective biofluid for lung cancer screening [7].

60 *1.3. Artificial Neural Networks*

61 Artificial neural networks (ANNs) are a class of sophisticated computa-  
 62 tional modelling structures that are inspired by biological neurological sys-  
 63 tems, regarding how they are able to learn and process highly non-linear  
 64 information [37]. The past three decades have seen ANNs being widely used  
 65 for biomedical decision support systems [8, 9, 10, 11, 12].

66 In general, an artificial neural network is formed of interconnected pro-  
 67 cessing units, commonly referred to as neurons. Each neuron applies an ac-  
 68 tivation function over the weighted sum of the incoming stimuli (or inputs),  
 69 and generate an output signal, which could be the input signal for other neu-  
 70 rons. Many different neural network architectures exist, in this paper we will  
 71 focus on the popular feed-forward artificial neural network, in particular the  
 72 multi-layer perceptron (MLP) [13], that usually consists of multiple layers  
 73 of neurons - the input layer, one or more hidden layers and the final output  
 74 layer, as illustrated in Figure 1.

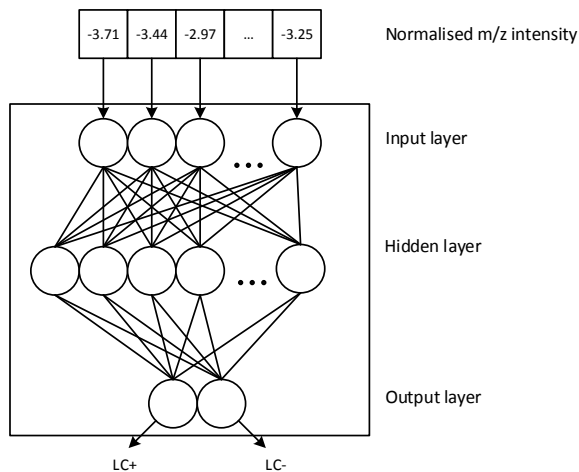


Figure 1: Illustration of a three-layered feed-forward artificial neural network, where each neuron in one layer has connections to the subsequent layer.

75 The design of network architectures involves setting the number of hidden  
 76 layers, the number of neurons within each layer, the connections between  
 77 them, and the type of activation function to use. The connection weights in  
 78 the network could be adjusted through a learning algorithm that minimises  
 79 the amount of error in the outputs compared to the true ones. Generalisation,

80 and to avoid overfitting the training data would be a central issue both in  
 81 network design and training.

## 82 2. Case Study

83 We have recently developed this approach by employing flow-infusion  
 84 electrospray-Mass Spectrometry (FIE-MS) to evaluate the potential of spon-  
 85 taneous sputum as a source of non-invasive metabolomic biomarkers for LC  
 86 status [14]. Spontaneous sputum was collected and processed from 34 pa-  
 87 tients suspected of having LC, and 33 healthy controls. Of the 34 patients,  
 88 23 were subsequently diagnosed with LC (LC+) at various stages of disease  
 89 progression. The clinical characteristics of all samples taken are summarised  
 90 in Table 1.

Characteristics	Lung cancer (LC+)	Symptoms (LC-)	Healthy con- trols
Number	23	11	33
Age (mean $\pm$ SD)	66.6 $\pm$ 8.1	66.5 $\pm$ 14.3	55.3 $\pm$ 14.6
Gender (Male/Female)	11/12	10/1	20/13
Smoking (Current / Ex / Never)	10/10/3	3/0/8	15/18/0
Previous cancer (Yes/No)	3/20	N/A	N/A
Final clinical diagnosis (SCLC/NSCLC/Radiological)	5/17/1	N/A	N/A
CO level (ppm)	4.2 $\pm$ 2.8	3.7 $\pm$ 1.3	N/A

Table 1: Summary of clinical characteristics of patients with Lung Cancer (LC+), Symptoms (LC-) and Healthy Controls

91 In these preliminary analyses, discriminatory metabolites were identified  
 92 using ANOVA and Random Forest and included Ganglioside GM1 which has  
 93 previously been linked to lung cancer [15]. This suggested that the use of  
 94 sputum as a non-invasive source of metabolite biomarkers may aid in the  
 95 development of an at-risk population screening programme for lung cancer  
 96 or enhanced clinical diagnostic pathways. We now demonstrate how further  
 97 data-mining of the FIE-MS data has revealed further metabolite biomark-  
 98 ers, and evaluate further the use of metabolomics to yield biomarkers for  
 99 distinguishing lung cancer type.

### 100 2.1. Ethics statement

101 The MedLung observational study (UKCRN ID 4682) received loco-regional  
 102 ethical approval from the Hywel Dda Health Board (05/WMW01/75). Writ-

103 ten informed consent was obtained from all participants at least 24 hours  
104 before sampling, at a previous clinical appointment, and all data was link  
105 anonymised before analysis.

## 106 *2.2. Mass spectrometry*

107 Frozen sputum samples were thawed before being exposed to 0.5 mL of  
108 dithiothreitol (DTT) to isolate sputum cells. Each sample was mixed using a  
109 vortex mixer for 15 minutes before being centrifuged at 1800g for 10 minutes  
110 before removing the supernatant. Sputum pellets were then analysed using  
111 Flow Infusion Electrospray Ion Mass spectrometry (FIE-MS).

112 Signals identified under 50  $m/z$  were removed, and the resulting FIE-MS  
113 data matrix contains 2,582  $m/z$  values after binning. The data was further  
114 preprocessed by total ion count (TIC) normalisation (to ensure the intensity  
115 values for each spectrum sum up to one), followed by log10 transformation  
116 prior to further data analysis.

## 117 *2.3. Effect of clinical characteristics on the metabolic profiles*

118 To explore the possible effects of the clinical characteristics on the global  
119 metabolic profiles, we conducted the so called 50-50 MANOVA test, which is  
120 essentially a variant of classical MANOVA that can handle multiple highly  
121 correlated responses [16]. We found no significant effect of age, gender or  
122 the CO level on the preprocessed metabolomic data (with  $p$ -values of 0.4,  
123 0.08, and 0.8, respectively), whilst the effect of disease status (LC+/LC-  
124 /Control) is really strong ( $p$ -value = 1e-12). And perhaps not surprisingly,  
125 as tobacco smoking is an important risk factor for lung cancer, there is indeed  
126 a significant effect of the smoking pack numbers per year over the metabolic  
127 profiles for the patients of LC- and LC+ ( $p$ -value = 2e-5).

## 128 *2.4. Diagnostic modelling with artificial neural networks*

129 To find discriminatory  $m/z$  features, Welch's unequal variance  $t$ -test have  
130 been performed using the pre-processed intensity values after log-transformation.  
131 Random forests have also been tried (results not shown), and the top ranked  
132 features identified for both methods are quite similar.

133 Then an ANN was used as a diagnostic model for various binary classifi-  
134 cation problems, taking the selected  $m/z$  signals (the preprocessed intensity  
135 values after log-transform) as the inputs and estimating the probability for  
136 individual classes. The activation function was set to hyperbolic tangent for  
137 both hidden and output layer; and the number of hidden layers was set to

138 two for all problems based on our initial analysis. Regularisation techniques  
 139 such as weight decay [17] have been employed to control the complexity of  
 140 the model parameters in order to avoid overfitting the models to the training  
 141 data.

142 The predictive power of neural network classifiers was evaluated using  
 143 Receiver Operating Characteristic (ROC) analysis and through leave-one out  
 144 (LOO) cross-validation (CV) - the overall ROC curve and the area under the  
 145 curve (AUC) from CV was obtained using the pooled test examples from CV.  
 146 For each binary classification problem and within each round of training, a  
 147 *t*-test would be performed on the training set, only the *m/z* signals with  
 148 resulting *p*-values < 0.05 were selected as input features for ANN modelling.

### 149 3. Results and Discussion

150 Representative spectra of the samples from the LC+, LC- and healthy  
 151 control sample groups are shown in Figure 2. FIE-MS profiles were analysed  
 152 using principal component analysis (PCA) (Figure 4). One can observe no  
 153 clear separation between clinically collected sample groups (LC+ and LC-)  
 154 if all *m/z* signals were used for PCA.

155 Welch *t*-tests provided 445 distinctive *m/z* values for LC+ versus healthy  
 156 controls, and 90 significant *m/z* values for LC+ versus LC- from our pooled  
 157 leave one out cross validation *t*-tests. PCA of both stratifications showed  
 158 good discriminative ability when using the pooled features, as shown in Fig-  
 159 ure 4. The number of neurons in each hidden layer was chosen through grid  
 160 search [18]. The best-performing models and their diagnostic performance  
 161 from LOOCV can be found in Table 2. And Figure 6 shows the resulting  
 162 ROC curves.

Classification	Mean no. of inputs	No. of hidden neurons	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	AUC
LC+ (vs Control)	1730.2	100, 50	96%	94%	92%	97%	0.99
LC+ (vs LC-)	71.9	40, 10	100%	91 %	96%	100%	1.00
SCLC (vs NSCLC)	77.8	50, 20	80%	100%	100%	94%	1.00

Table 2: Results of cross-validation prediction performance of our ANN models. The diagnostic performance (except for AUC) was obtained by using the default probability cutoff value of 0.5 to determine the class labels.



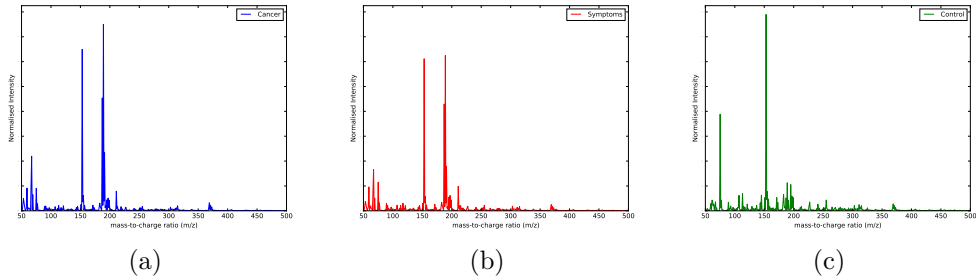


Figure 2: Typical FIE-MS spectra of sputum obtained from sputum obtained from (a) a patient with lung cancer , (b) a patient with symptoms of lung cancer and (c) a healthy control sample.

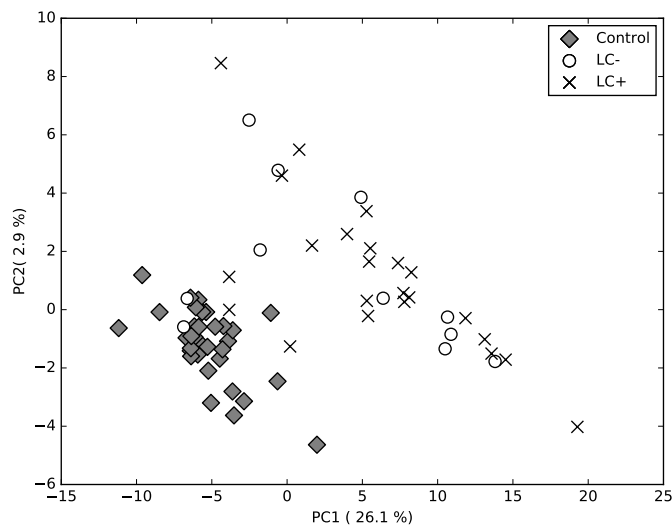


Figure 3: PCA score plots of FIE-MS data using all normalised  $m/z$  intensity values in negative ionisation mode, showing no clear separation between LC+ and LC- samples.

163 *3.1. Analysis of small-cell lung cancer and non-small cell lung cancer*

164 Determining the type of lung cancer that has developed in a patient  
 165 is a key component of determining the correct treatment and management  
 166 pathway. For lung cancer, two broad classes of classification exist: non-

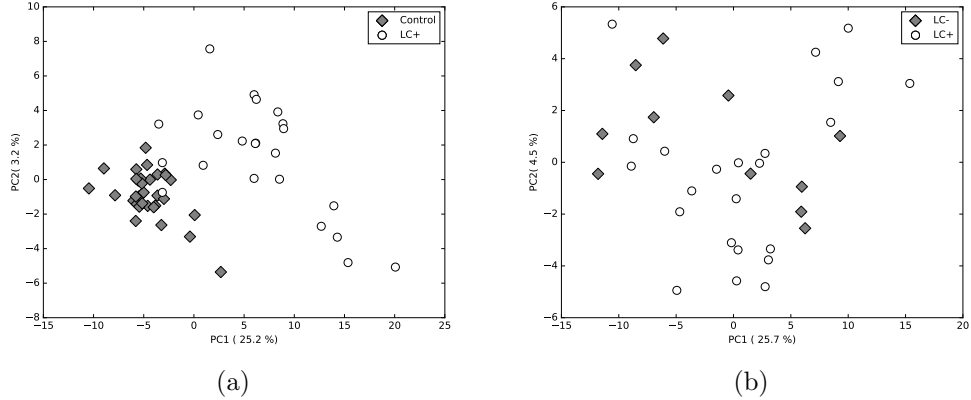


Figure 4: PCA score plots of the FIE-MS data with (a) only 970 selected  $m/z$  signals for LC+ and healthy controls, and (b) 125 selected  $m/z$  signal for LC+ and LC-. Using  $m/z$  features taken from  $t$ -tests clearly differentiates between relevant classes.

167 small-cell lung cancer (NSCLC) and small-cell lung cancer (SCLC). Patients  
 168 with NSCLC are usually classified as one of three main subtypes: adenocar-  
 169 cinoma, squamous-cell carcinoma and large-cell carcinoma. Of these, ade-  
 170 noma is the most common and is characterised by overproduction of  
 171 mucin. Squamous-cell carcinoma is the second most common form of lung  
 172 cancer and typically occurs in the centre of the lungs. Large-cell carcinoma  
 173 is less common and is characterised by cancerous cells that are large, with  
 174 excess cytoplasm and large nuclei. The extent of NSCLCs is reported using  
 175 the TNM format, which is important for prognosis and treatment planning.  
 176 The TNM format ranges from Stage 0 to Stage IV, with the relevant stage  
 177 determined through assessment of the primary tumour, involvement of re-  
 178 gional lymph nodes, and the extent of distant metastasis against set criteria  
 179 [19].

180 Small-cell lung cancers are less common than NSCLC, with approximately  
 181 10% of all lung cancers classified as SCLC. These lung cancers are charac-  
 182 terised by their small cells, with minimal cytoplasm, and poorly-defined cell  
 183 borders. Cancerous cells are usually rounded, oval and spindle-shaped. Typ-  
 184 ically, patients with SCLC present when the disease has metastasised from  
 185 the lungs and symptoms frequently this, such as issues with bone marrow  
 186 and the liver because of metastasis. Small-cell lung cancers are staged differ-

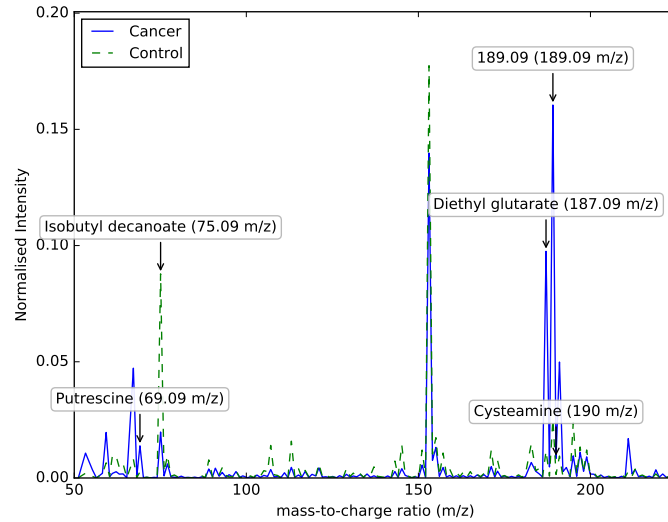


Figure 5: Mean FIE-MS spectra illustrating five key distinguishable metabolites between patients with lung cancer (LC) and healthy controls.

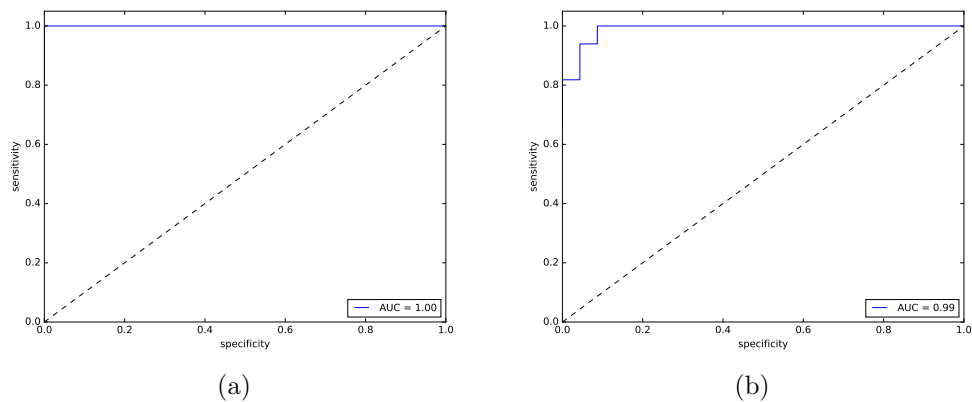


Figure 6: Receiver operating characteristic curves obtained from ANN models using leave one out cross-validation for classifying: (a) LC+ against LC-, and (b) LC against healthy controls.

187 ently to NSCLC. Although the TNM format can be used, it does not predict  
 188 survival and other outcomes well. Typically, SCLC is staged as either lim-

189 ited or extensive disease, with the latter equivalent to Stage IV of the TNM  
 190 staging format for NSCLC [19].

191 A total of nine  $m/z$  values provided strong differentiation between SCLC  
 192 and NSCLC with  $p$ -values less than 0.05 from Welch  $t$ -tests. Out of these  
 193 9  $m/z$  values, 6 were identified and brought forward for further analysis as  
 194 potential biomarkers of NSCLC and SCLC. Their relative values and  $p$ -values  
 195 are shown in Table 3. Furthermore, ranges of each metabolite shown as box-  
 196 and-whisker plots can be found in Figure 7. These both indicate that the 6  
 197 biomarker candidates show that the median levels of the 6 markers are higher  
 198 in NSCLC samples to patients with SCLC.

Metabolite	$m/z$	Normalised Intensity		$p$ -value
		NSCLC	SCLC	
Phenylacetic acid	137.09	$-3.31 \pm 0.21$	$-2.85 \pm 0.42$	0.001
L-Fucose	165.09	$-3.19 \pm 0.12$	$-2.86 \pm 0.30$	0.001
Caprylic acid	145.18	$-2.81 \pm 0.44$	$-2.12 \pm 0.31$	0.001
Acetic acid	61.09	$-2.96 \pm 0.19$	$-2.56 \pm 0.40$	0.002
Propionic acid	75.09	$-1.91 \pm 0.18$	$-1.56 \pm 0.34$	0.003
Glycine	76.09	$-3.17 \pm 0.12$	$-2.91 \pm 0.21$	0.004

Table 3: Identified metabolites that are significantly different between patients with Non-Small Cell Lung Cancer (NSCLC) and Small Cell Lung Cancer (SCLC) using Welch  $t$ -tests.

199 PCA analysis (Figure 8a) showed good separation capabilities between  
 200 NSCLC and SCLC samples, 6 metabolites were selected as input features  
 201 to build a second ANN. Due to the small sample size, leave-one-out cross  
 202 validation was performed to estimate the generalisation performance of the  
 203 model. Our MLP model was able to distinguish between NSCLC and SCLC  
 204 with a sensitivity of 80% and a specificity of 100% for predicting SCLC from  
 205 cross-validation (see Table 2 and Figure 8b).

#### 206 4. Biomarker analysis

207 Metabolic profiling recognised and provided identifications of 6 candidate  
 208 metabolites that offered superb predictive values. Amongst the targeted  
 209 metabolites are examples which have already been linked to lung cancer.  
 210 The enzyme glycine decarboxylase (GLDC) is involved in the degradation of

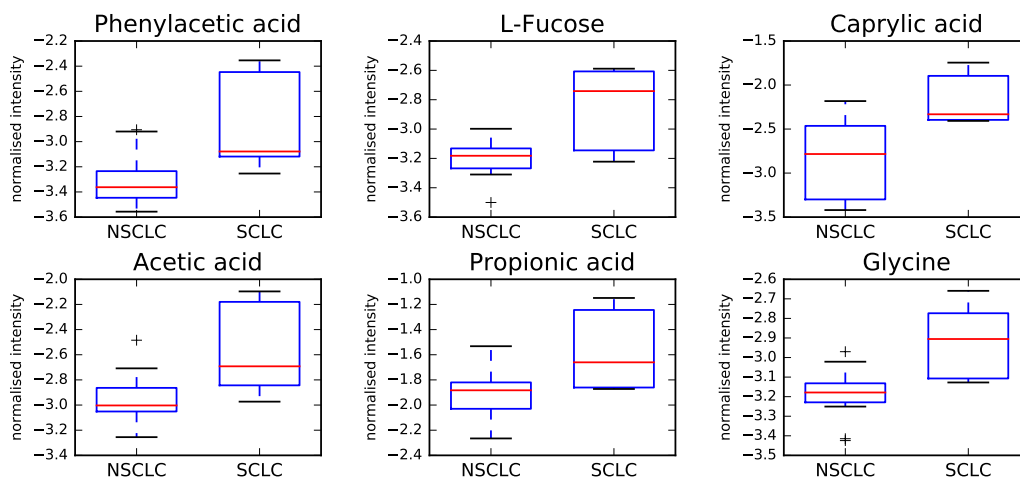


Figure 7: Box-and-whisker plots of the candidate metabolite biomarkers for discrimination between NSCLC and SCLC. The y axis represents the normalised intensity of each metabolite. Horizontal lines in the middle portion of the box illustrates the median, bottom and top boundaries of boxes represent the lower and upper quartile, whiskers depict the 5th and 95th percentiles and plus signs depicts the outliers.

211 glycine which is coupled to the generation of methylgroups which can be used  
 212 in (for example) purine biosynthesis. GLDC expression was increased in cells  
 213 isolated from NSCLC tumours with concomitant decreases in glycine [20].  
 214 These authors showed that GLDC expression could serve as a biomarker,  
 215 we now provide evidence that relative decreases in glycine is a feature of  
 216 NSCLC in sputum. This biofluid represents a less invasive and potentially  
 217 cost-effective means of clinically assessing patient LC status.

218 Fucose (6-deoxy-L-galactose) is N- and O-linked to a range of glycolipids  
 219 and glycopeptides produced by mammalian cells. Increases in fucosylated

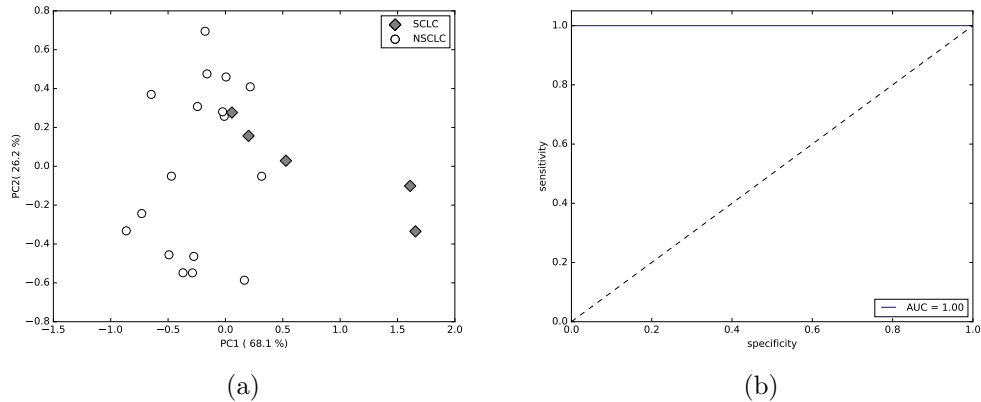


Figure 8: PCA score plot of the 6 identified metabolites for (a) NSCLC and SCLC, and (b) the ROC curve obtained from ANN LOOCV.

220 proteins, for example,  $\alpha$ -fetoprotein are used in the diagnosis of hepatocel-  
 221 lular carcinoma [21]. Fucosylation is dependent on the availability of the  
 222 substrate guanosine 5'-diphospho-fucose (GDP-Fucose) and associated gly-  
 223 cosyltransferases to transfer the fucose motif on to the protein / lipid. Fucose  
 224 is a precursor to GDP-Fucose production [22] so that increases in fucosylation  
 225 could lead to a relative depletion in fucose as noted in our study. Increased  
 226 fucosylation has been previously linked to NSCLC biopies [23] but ours is the  
 227 first suggestion that decreases in fucose pools in sputum could be clinically  
 228 suitable marker.

229 Other key metabolites were volatile short chain fatty s acetic (C2), pro-  
 230 pionic (C3) and caprylic [octanoic ] (C8), These could be derived as a result  
 231 of lipid peroxidation [24] but irrespective of their means of generation would  
 232 provide further support for efforts that are attempting to sample breath as  
 233 an non-invasive method for lung cancer detection [24, 25, 26].

234 A somewhat surprising observation was the detection of phenylacetic .  
 235 This is classically associated with phenylketonuria (PKU); an inherited dis-  
 236 order of amino metabolism. PKU arises from a deficiency of the liver enzyme  
 237 phenylalanine-4-hydroxylase which production of tyrosine from phenylala-  
 238 nine. If this enzyme is non-functional a range of alternative metabolites are  
 239 produced, including phenylacetic [27]. However, beyond PKU, phenylacetate  
 240 accumulates in patients with chronic kidney disease and during renal failure  
 241 where it can inhibit nitric oxide generation [28] and macrophage intracellular

242 killing of bacteria [29]. Further, disease-associated phenylacetate accumula-  
243 tion can contribute to an inflammatory responses [30]. To our knowledge,  
244 phenylacetate has not been associated with cancer; indeed quite the opposite  
245 it has a history of being tested for its anti-tumour properties [31]. However,  
246 it may be that NSCLC has particularly phenotypic / biochemical features  
247 which lead to altered phenylalanine metabolism leading to the accumulation  
248 of phenylacetate to contribute to inflammatory events and a reduced ability  
249 to deal with the lung microbiome. This is currently under investigation in  
250 our group.

## 251 5. Conclusions

252 A metabolomics approach based on FIE-MS coupled with univariate  
253 Welch *t*-test based feature selection and artificial neural networks provides  
254 an efficient methodology for metabolomic-based profiling of sputum to dif-  
255 ferentiate between non-small cell and small-cell lung cancer. This paper has  
256 identified 6 candidate metabolites markers, including L-fucose, phenylacetic ,  
257 caprylic , acetic, propionic acid, and glycine, which were found to have good  
258 discriminatory abilities and low *p*-values. Excellent sensitivity and specificity  
259 was also shown using these markers through leave-one-out cross validation,  
260 which further indicates the promise of metabolomic analysis of sputum for  
261 non-invasive screening for LC. Further analysis involving a larger number of  
262 samples is required to determine both the precision and applicability of this  
263 approach in guiding the diagnosis and treatment of LC and respective forms.

## 264 Acknowledgements

265 The work described in this paper was funded by two Aberystwyth Uni-  
266 versity PhD scholarships to Keiron O'Shea and Simon Cameron. Keiron  
267 O'Shea would like to thank Nicholas Dimonaco and Manfred Beckmann of  
268 Aberystwyth University for their helpful support offered during this project.

## 269 References

- 270 [1] J. Ferlay, H.-R. Shin, F. Bray, D. Forman, C. Mathers, D. M. Parkin,  
271 Estimates of worldwide burden of cancer in 2008: Globocan 2008, In-  
272 ternational journal of cancer 127 (2010) 2893–2917.

- 273 [2] A. Jemal, R. Siegel, J. Xu, E. Ward, *Cancer statistics, 2010*, CA: a  
274 cancer journal for clinicians 60 (2010) 277–300.
- 275 [3] L. Guzmán, M. S. Depix, A. M. Salinas, R. Roldán, F. Aguayo, A. Silva,  
276 R. Vinet, Analysis of aberrant methylation on promoter sequences of  
277 tumor suppressor genes and total dna in sputum samples: a promising  
278 tool for early detection of copd and lung cancer in smokers, *Diagn Pathol*  
279 7 (2012) 1596–7.
- 280 [4] A. Andermann, I. Blancquaert, S. Beauchamp, V. Déry, Revisiting  
281 wilson and jungner in the genomic age: a review of screening criteria  
282 over the past 40 years, *Bulletin of the World Health Organization* 86  
283 (2008) 317–319.
- 284 [5] Y. Chen, Z. Ma, A. Li, H. Li, B. Wang, J. Zhong, L. Min, L. Dai,  
285 Metabolomic profiling of human serum in lung cancer patients using liq-  
286 uid chromatography/hybrid quadrupole time-of-flight mass spectrom-  
287 etry and gas chromatography/mass spectrometry, *Journal of cancer*  
288 research and clinical oncology 141 (2015) 705–718.
- 289 [6] A. Hubers, C. Prinsen, G. Sozzi, B. Witte, E. Thunnissen, Molecular  
290 sputum analysis for the diagnosis of lung cancer, *British journal of*  
291 cancer 109 (2013) 530–537.
- 292 [7] P. D. Lewis, K. E. Lewis, R. Ghosal, S. Bayliss, A. J. Lloyd, J. Wills,  
293 R. Godfrey, P. Kloer, L. A. Mur, Evaluation of ftir spectroscopy as a  
294 diagnostic tool for lung cancer using sputum, *BMC cancer* 10 (2010) 1.
- 295 [8] P. J. Lisboa, A. F. Taktak, The use of artificial neural networks in  
296 decision support in cancer: a systematic review, *Neural networks* 19  
297 (2006) 408–415.
- 298 [9] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann,  
299 F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, et al., Classifica-  
300 tion and diagnostic prediction of cancers using gene expression profiling  
301 and artificial neural networks, *Nature medicine* 7 (2001) 673–679.
- 302 [10] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E.  
303 Metz, Artificial neural networks in mammography: application to de-  
304 cision making in the diagnosis of breast cancer., *Radiology* 187 (1993)  
305 81–87.



- 306 [11] C. Lu, J. De Brabanter, S. Van Huffel, I. Vergote, D. Timmerman, Using  
307 artificial neural networks to predict malignancy of ovarian tumors, in:  
308 Engineering in Medicine and Biology Society, 2001. Proceedings of the  
309 23rd Annual International Conference of the IEEE, volume 2, IEEE, pp.  
310 1637–1640.
- 311 [12] R. Dybowski, V. Gant, Artificial neural networks in pathology and  
312 medical laboratories, *The Lancet* 346 (1995) 1203–1207.
- 313 [13] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning internal rep-  
314 resentations by error propagation, Technical Report, DTIC Document,  
315 1985.
- 316 [14] S. J. Cameron, K. E. Lewis, M. Beckmann, G. G. Allison, R. Ghosal,  
317 P. D. Lewis, L. A. Mur, The metabolomic detection of lung cancer  
318 biomarkers in sputum, *Lung Cancer* 94 (2016) 88–95.
- 319 [15] T. Brezicka, B. Bergman, S. Olling, P. Fredman, Reactivity of mon-  
320 oclonal antibodies with ganglioside antigens in human small cell lung  
321 cancer tissues, *Lung Cancer* 28 (2000) 29–36.
- 322 [16] Ø. Langsrud, 5050 multivariate analysis of variance for collinear re-  
323 sponses, *Journal of the Royal Statistical Society: Series D (The Statis-  
324 tician)* 51 (2002) 305–317.
- 325 [17] R. C. S. L. L. Giles, Overfitting in neural nets: Backpropagation, conju-  
326 gate gradient, and early stopping, in: *Advances in Neural Information  
327 Processing Systems 13: Proceedings of the 2000 Conference*, volume 13,  
328 MIT Press, p. 402.
- 329 [18] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimiza-  
330 tion, *The Journal of Machine Learning Research* 13 (2012) 281–305.
- 331 [19] W. D. Travis, C. Harris, Pathology and genetics of tumours of the lung,  
332 pleura, thymus and heart, Feance: IARC Press, 2004, 2004.
- 333 [20] W. C. Zhang, N. Shyh-Chang, H. Yang, A. Rai, S. Umashankar, S. Ma,  
334 B. S. Soh, L. L. Sun, B. C. Tai, M. E. Nga, et al., Glycine decarboxy-  
335 lase activity drives non-small cell lung cancer tumor-initiating cells and  
336 tumorigenesis, *Cell* 148 (2012) 259–272.

- 337 [21] E. Miyoshi, K. Moriwaki, N. Terao, C.-C. Tan, M. Terao, T. Nakagawa,  
338 H. Matsumoto, S. Shinzaki, Y. Kamada, Fucosylation is a promising  
339 target for cancer diagnosis and therapy, *Biomolecules* 2 (2012) 34–45.
- 340 [22] D. J. Becker, J. B. Lowe, Fucose: biosynthesis and biological function  
341 in mammals, *Glycobiology* 13 (2003) 41R–53R.
- 342 [23] X. Zeng, B. L. Hood, M. Sun, T. P. Conrads, R. S. Day, J. L. Weissfeld,  
343 J. M. Siegfried, W. L. Bigbee, Lung cancer serum biomarker discovery  
344 using glycoprotein capture and liquid chromatography mass spectrometry,  
345 *Journal of proteome research* 9 (2010) 6440–6449.
- 346 [24] C. Wang, R. Dong, X. Wang, A. Lian, C. Chi, C. Ke, L. Guo, S. Liu,  
347 W. Zhao, G. Xu, et al., Exhaled volatile organic compounds as lung cancer  
348 biomarkers during one-lung ventilation, *Scientific reports* 4 (2014).
- 349 [25] M. Phillips, K. Gleeson, J. M. B. Hughes, J. Greenberg, R. N. Cataneo,  
350 L. Baker, W. P. McVay, Volatile organic compounds in breath as markers  
351 of lung cancer: a cross-sectional study, *The Lancet* 353 (1999) 1930–  
352 1933.
- 353 [26] M. Phillips, R. N. Cataneo, A. R. Cummin, A. J. Gagliardi, K. Gleeson,  
354 J. Greenberg, R. A. Maxfield, W. N. Rom, Detection of lung cancer with  
355 volatile markers in the breath, *Chest Journal* 123 (2003) 2115–2123.
- 356 [27] A. Bajtarevic, C. Ager, M. Pienz, M. Klieber, K. Schwarz, M. Ligor,  
357 T. Ligor, W. Filipiak, H. Denz, M. Fiegl, et al., Noninvasive detection  
358 of lung cancer by analysis of exhaled breath, *BMC cancer* 9 (2009) 1.
- 359 [28] J. Jankowski, M. Van Der Giet, V. Jankowski, S. Schmidt, M. Hemeier,  
360 B. Mahn, G. Giebing, M. Tölle, H. Luftmann, H. Schlüter, et al., In-  
361 creased plasma phenylacetic acid in patients with end-stage renal failure  
362 inhibits inos expression, *The Journal of clinical investigation* 112 (2003)  
363 256–264.
- 364 [29] S. Schmidt, T. H. Westhoff, P. Krauser, R. Ignatius, J. Jankowski,  
365 V. Jankowski, W. Zidek, M. Van der Giet, The uraemic toxin pheny-  
366 lacetic acid impairs macrophage function, *Nephrology Dialysis Trans-  
367 plantation* 23 (2008) 3485–3493.

- 368 [30] G. Cohen, J. Raupachova, W. H. Hörl, The uraemic toxin phenylacetic  
369 acid contributes to inflammation by priming polymorphonuclear leuco-  
370 cytes, *Nephrology Dialysis Transplantation* 28 (2013) 421–429.
- 371 [31] A. Thibault, D. Samid, M. R. Cooper, W. D. Figg, A. C. Tompkins,  
372 N. Patronas, D. J. Headlee, D. R. Kohler, D. J. Venzon, C. E. Myers,  
373 Phase i study of phenylacetate administered twice daily to patients with  
374 cancer, *Cancer* 75 (1995) 2932–2938.