



PEARL

**Visualisation of exhaled breath metabolites reveals distinct diagnostic signatures for acute cardiorespiratory breathlessness**

Wilde, Michael J.; Ibrahim, Wadah; Cordell, Rebecca L.; Richardson, Matthew; Free, Robert C.; Zhao, Bo; Singapuri, Amisha; Hargadon, Beverley; Gaillard, Erol A.; Suzuki, Toru; Ng, Leong L.; Coats, Tim; Thomas, Paul; Monks, Paul S.; Brightling, Christopher E.; Greening, Neil J.; Siddiqui, Salman

**Published in:**

Science Translational Medicine

**DOI:**

[10.1126/scitranslmed.abl5849](https://doi.org/10.1126/scitranslmed.abl5849)

**Publication date:**

2022

**Link:**

[Link to publication in PEARL](#)

**Citation for published version (APA):**

Wilde, M. J., Ibrahim, W., Cordell, R. L., Richardson, M., Free, R. C., Zhao, B., Singapuri, A., Hargadon, B., Gaillard, E. A., Suzuki, T., Ng, L. L., Coats, T., Thomas, P., Monks, P. S., Brightling, C. E., Greening, N. J., & Siddiqui, S. (2022). Visualisation of exhaled breath metabolites reveals distinct diagnostic signatures for acute cardiorespiratory breathlessness. *Science Translational Medicine*, 0(0). <https://doi.org/10.1126/scitranslmed.abl5849>

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40

This is the author’s version of the work. It is posted here by permission of the AAAS for personal use, not for redistribution. The definitive version was published in Science Translational Medicine on [18/10/2022], DOI: 10.1126/scitranslmed.abl5849

## Visualisation of exhaled breath metabolites reveals distinct diagnostic signatures for acute cardiorespiratory breathlessness

Wadah Ibrahim<sup>1\*</sup>, Michael J. Wilde<sup>2,3\*\*</sup>, Rebecca L. Cordell<sup>2\*</sup>, Matthew Richardson<sup>1</sup>, Dahlia Salman<sup>4</sup>, Robert C. Free<sup>1,2</sup>, Bo Zhao<sup>5,6</sup>, Amisha Singapuri<sup>1</sup>, Beverley Hargadon<sup>1</sup>, Erol A. Gaillard<sup>1</sup>, Toru Suzuki<sup>7,8</sup>, Leong L. Ng<sup>7</sup>, Tim Coats<sup>9</sup>, Paul Thomas<sup>4</sup>, Paul S. Monks<sup>2</sup>, Christopher E. Brightling<sup>1</sup>, Neil J. Greening<sup>1</sup>, Salman Siddiqui<sup>10†</sup>, on behalf of the EMBER consortium

*\*equal contribution and first authorship*

*†joint corresponding authorship. Email: [michael.wilde@plymouth.ac.uk](mailto:michael.wilde@plymouth.ac.uk) (M.J.W.); [s.siddiqui@imperial.ac.uk](mailto:s.siddiqui@imperial.ac.uk) (S.S.)*

### Affiliations:

<sup>1</sup>Department of Respiratory Sciences, University of Leicester, Leicester, LE1 7RH UK

<sup>2</sup>School of Chemistry, University of Leicester, Leicester, LE1 7RH UK

<sup>3</sup>School of Geography, Earth and Environmental Sciences, University of Plymouth, Plymouth, PL4 8AA, UK

<sup>4</sup>Department of Chemistry, Loughborough University, Loughborough, LE11 3TT UK

<sup>5</sup>Leverhulme Centre for Demographic Science, University of Oxford, Oxford, OX1 1JD United Kingdom

<sup>6</sup>Nuffield College, University of Oxford, Oxford, OX1 1NF United Kingdom

<sup>7</sup>Department of Cardiovascular Sciences, University of Leicester, Cardiovascular Research Centre, Glenfield General Hospital, Leicester, LE3 9QP UK

<sup>8</sup>The Institute of Medical Science, The University of Tokyo Shirokane-dai, Minato-ku 4-6-1, 108-8639 Tokyo, Japan

<sup>9</sup>Emergency Medicine Academic Group, Department of Cardiovascular Sciences, University of Leicester, University Road, Leicester LE1 7RH, UK.

<sup>10</sup>National Heart and Lung Institute, Imperial College, London, SW3 6LY UK

41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81

**Abstract**

Acute cardiorespiratory breathlessness accounts for 1 in 8 of all emergency hospitalisations. Early, non-invasive diagnostic testing is a clinical priority that allows rapid triage and treatment. Here, we sought to discover and replicate diagnostic breath volatile organic compound (VOC) biomarkers of acute cardiorespiratory disease and understand breath metabolite network enrichment in acute disease, with a view to gaining mechanistic insight of breath biochemical derangements. We collected and analysed exhaled breath samples from 277 participants presenting with acute cardiorespiratory exacerbations and aged matched healthy volunteers. Topological data analysis (TDA) phenotypes differentiated acute disease from health and acute cardiorespiratory exacerbation subtypes [acute heart failure, acute asthma, acute Chronic Obstructive Pulmonary Disease (COPD) and community-acquired pneumonia]. A multi-biomarker score (101 breath biomarkers) demonstrated good diagnostic sensitivity and specificity ( $\geq 80\%$ ) in both discovery and replication sets and was associated with all-cause mortality at 2 years. In addition, VOC biomarker scores differentiated metabolic subgroups of cardiorespiratory exacerbation. Louvain clustering of VOCs coupled with metabolite enrichment and similarity assessment revealed highly specific enrichment patterns in all acute disease subgroups, for example selective enrichment of correlated  $C_{5-7}$  hydrocarbons and  $C_{3-5}$  carbonyls in heart failure and selective depletion of correlated aldehydes in acute asthma. This study identified breath VOCs that differentiate acute cardiorespiratory exacerbations and associated subtypes and metabolic clusters of disease-associated VOCs.

**Word count: 216**

**One sentence summary**

82 Exhaled VOCs can distinguish acute cardiorespiratory exacerbations and associated subtypes and  
83 map underlying metabolic clusters.

## 84 **1. Introduction**

85  
86

87 Breathlessness due to cardiorespiratory diseases accounts for more than 1 in 8 of all emergency admissions  
88 to hospital (1). Despite the same presenting symptom, the aetiology of acute breathlessness is highly  
89 varied, with diverse disease trajectories and therapeutic options. Diagnostic evaluation of acute  
90 breathlessness is heavily reliant on investigations such as blood-based biomarkers [e.g. C-reactive protein  
91 (CRP), B-type natriuretic peptide] and radiological procedures. These biomarkers have clinical utility  
92 primarily in patients with single pathologies, but have poor discriminatory power in patients with  
93 multifactorial presentations of acute breathlessness and are particularly challenging to interpret in the  
94 context of pre-admission treatment exposure (e.g. antibiotics for pneumonia and admission CRP values) (2).

95 Breathomics, the characterisation of volatile organic compounds (VOCs) in exhaled breath, enables the  
96 evaluation of diagnostic and prognostic biomarkers in acute breathlessness, directly from the lung as well as  
97 incorporating metabolites from the systemic circulation (3). The assessment of exhaled, low-molecular  
98 weight biochemicals, chemically classified as VOCs, has been presented as a new paradigm for the  
99 development of rapid, non-invasive diagnostic and prognostic biomarkers. However, the scarcity of robustly  
100 powered clinical studies, combined with a lack of standardisation in sample collection and analysis as well  
101 as data and chemometric processing, have delayed further translation of this technology to clinical settings.

102 Notwithstanding these challenges, the potential of breathomics is becoming increasingly recognised in  
103 research and therapeutic development in respiratory diseases. The emergence of powerful high-resolution  
104 mass spectrometry and multidimensional separation technologies such as comprehensive two-dimensional  
105 gas chromatography coupled with mass spectrometry (GCxGC-MS), which provides visual readouts of  
106 breath-based biomarkers (4, 5), has facilitated research advances. Although chemometric analyses play a  
107 vital role in this field, the enhanced dimensionality of GCxGC-MS data enriches established chemometric  
108 and imaging-based characterisation methods for visualising, extracting and quantifying VOC markers from  
109 complex and previously unresolved matrices.

110 Herein, we present a real-world, prospective study of acutely unwell hospitalised patients presenting with  
111 breathlessness due to severe exacerbations of cardiorespiratory aetiology (asthma, COPD, heart failure or  
112 pneumonia) and healthy controls. By isolating and visualizing exhaled VOCs with GCxGC-MS, coupled with  
113 rigorous clinical phenotyping, exhaled breath metabolites were shown to have high diagnostic accuracy for  
114 severe cardiorespiratory exacerbations (including in the presence of diagnostic uncertainty) and to be  
115 dysregulated across several pertinent volatile classes in different clinical subtypes of cardiorespiratory  
116 exacerbation. This research provides pivotal evidence that shows how breath biomarker platforms may be  
117 used in acute care and demonstrates the potential for translation of this technology into a real-world  
118 clinical setting.

119

120

121

122

123

124

## 125 **2. Results**

126

### 127 **2.1. Participant demographics and clinical characteristics**

128 As part of the East Midlands Breathomics Pathology Node (EMBER), exhaled breath from 277 participants  
129 recruited from acutely breathless hospitalised patients and matched healthy controls was sampled (**Figure**  
130 **1**). Sample size calculations are detailed in (**Methods section ‘sample size estimation’ and Table S1**).

131 Breath samples were analysed to identify dysregulation of metabolic classes in cardiorespiratory disease  
132 and investigate whether exhaled VOC profiles could predict acute cardio respiratory exacerbations despite  
133 diagnostic uncertainty, and thus have a potential role in phenotyping acute cardiorespiratory  
134 breathlessness (**fig. S1**). Participants’ mean (SD) age was  $60.8 \pm (16.8)$  years, 51% were males, 30 patients  
135 required supplemental oxygen on admission and the mean admission modified early warning score (mEWS-  
136 2 score) was 2. The cohort was made up of patients presenting with the following exacerbation subtypes:  
137 acute severe asthma (n= 65), acute severe COPD (n= 58), acute severe heart failure (n=44), community  
138 acquired pneumonia (n=55), and healthy volunteers (n=55), recruited between May 2017 and December  
139 2018. Participants’ demographic and clinical characteristics are summarised in (**Table 1**). Breath samples  
140 were collected using a ReCIVA device, adopting a standardised sampling and gated protocol that enriches  
141 alveolar volatiles (6), and analysed using thermal desorption (TD) coupled to comprehensive two-  
142 dimensional gas chromatography (GCxGC) with dual flame ionisation detection (FID) and mass spectrometry  
143 (MS).

144

### 145 **2.2. Unbiased discovery using TDA identifies breath markers of acute disease**

146 Topological data analysis is an unsupervised machine-learning tool used for the analysis of large-  
147 scale, high-dimensional, complex datasets. It is highly sensitive to patterns that are often overlooked  
148 by other data reduction tools like Principal Component Analysis (PCA) (7).

149 TDA is a well-established data analytic technique for unbiased data driven discovery based  
150 phenotyping (7). TDA has proven to be a powerful tool, yielding critical insights in the prognostic  
151 phenotyping (8), cancer imaging biomarker stratification (9), disease classification using pathology  
152 biomarkers (10), omics based cancer phenotyping (11). Several publications have reported the use of  
153 TDA in the metabolomics field, for example, unbiased lipid phenotyping of lung epithelial lining fluid  
154 (12).

155 To achieve an unbiased discovery of exhaled VOCs predictive of the acute disease groups, patients were  
156 block randomised *post-hoc* into a discovery cohort of 139 participants (acute asthma n= 33, acute COPD n=  
157 29, acute heart failure n=22, community acquired pneumonia n=28, healthy volunteers n=27), and a  
158 replication cohort of 138 participants (acute asthma n= 32, acute COPD n= 29, acute heart failure n=22,  
159 community acquired pneumonia n=27, healthy volunteers n=28). Randomisation allowed internal  
160 replication of diagnostic breath biomarkers, whilst adjusting for relevant confounders. Details of the  
161 randomisation and further clinical characteristics of the cohorts can be found in (**tables S2-S3**).

162 Chemometric analysis and quantification of VOCs was performed blinded to clinical diagnosis by two

163 analytical chemists (MW and RC), with biostatistical analyses linking subject identifier to chemometric  
164 biomarkers performed following data lock by an independent statistician (MR).

165 805 unique chromatographic features (peaks) were detected across the breath sample set using TD-  
166 GCxGC-FID/MS, with 404 features detected on average in each sample. Topological data analysis (TDA)  
167 applied to these 805 chromatographic features yielded topologically distinct networks that distinguished  
168 underlying causes of acute breathlessness whilst anchoring to corresponding blood-based biomarkers in  
169 both the discovery and replication cohorts (**Figure 2**). Specifically, healthy volunteers and patients with  
170 acute heart failure formed distinct topological groupings in both discovery and replication populations.  
171 Respiratory admissions due to acute asthma, acute COPD and pneumonia formed a topological continuum  
172 albeit within distinct regions of a single network in the replication cohort; similar findings were observed in  
173 the discovery cohort, with the exception of acute asthma forming a distinct grouping.

174

175

### 176 **2.3. Breath biomarker clinical prediction scores**

177 To create a concatenated list of exhaled breath biomarkers suitable for diagnostic application, we applied a  
178 threshold of 80% feature-presence per patient group, below which features were removed to effectively  
179 reduce the number of features used in subsequent models with more than 20% of zero values for peak  
180 areas (**fig. S2**). We found that the zero-valued peak areas were randomly distributed across the disease  
181 groups in all but seven features. The exclusion of the seven features where there was some evidence that  
182 zero-valued peak areas were not randomly distributed across the disease groups did not alter the results of  
183 the regression models.

184 Further filtering steps using least absolute shrinkage and selection operator (LASSO) and elastic net  
185 regression methods, followed by removal of 38 peaks that were considered to be chemical and material  
186 artefacts (e.g. siloxanes), generated a final panel of 101 exhaled breath volatiles (**tables S4-S8**). Therefore,  
187 the analysis plan permitted the identification of a rich and chemically diverse response in the VOC profile as  
188 opposed to only a handful of individual VOC markers and afforded the generation of biomarker scores. The  
189 data was examined for batch effects and was adjusted accordingly. Batch effects detected related to major  
190 instrument maintenance events, which occurred twice creating three groups. No contributions were  
191 observed based on the ReCIVA device used, operator, time of day, or volume of breath sample collected,  
192 most likely nullified by the simultaneous and consecutive recruitment across all cohorts throughout the  
193 study to reduce potential biases (**fig. S3-4**).

194 The value of the generated acute disease VOC biomarker score was found to be higher in acute  
195 cardiorespiratory patients compared to healthy volunteers (**Figure 3A**). For the discovery cohort (n=139),  
196 the acute disease VOC biomarker score effectively differentiated participants with acute cardiorespiratory  
197 exacerbations from age-matched healthy controls with an area under the curve (AUC) of 1.00 (1.00-1.00)  $P$   
198  $< 0.0001$ , sensitivity 1.00 (1.00-1.00), specificity (1.00-1.00), positive predictive value (PPV) 1.00 (1.00-1.00),  
199 negative predictive value (NPV) (1.00-1.00). For the replication cohort (n=138), the same VOC biomarker  
200 score differentiated participants with acute disease from healthy controls with AUC 0.90 (0.83-0.96)  $P$   
201  $< 0.0001$ , sensitivity 0.88 (0.82-0.94), specificity 0.79 (0.63-0.94), PPV of 0.95 (0.91-0.99), NPV of 0.51 (0.36-  
202 0.65) (**Figure 3B**).

203 To evaluate the impact of potential confounders on our model classification, we re-ran our statistical  
204 models, adjusting for the following factors: (i) smoking status (current, ex-smoker or never smoker); (ii)  
205 time between hospital admission and the acquisition of the breath samples, as this time period is often the

206 period within which acute treatments are delivered; (iii) the modified early warning score 2 (mEWS-2),  
207 which is a composite acuity score combining respiratory rate, oxygen saturations, systolic blood pressure,  
208 heart rate, degree of consciousness, confusion and body temperature for each patient; and (iv) prior  
209 exposure to either antibiotics or steroids for cardiorespiratory illness in the fortnight prior to the index  
210 admission. We observed improved diagnostic accuracy in the replication cohort [AUC 1.00 (1.00-1.00),  $P$   
211  $<0.0001$ ] when considering these adjustments, which would be expected with the inclusion of acuity  
212 markers for the classification of acute illness.

213 Following a clinical adjudication process (**Methods: section 'clinical adjudication'**), each patient was  
214 assigned a degree of clinical diagnostic uncertainty using a 100-mm visual analogue scale (VAS) at the point  
215 of clinical triage (**Figure 3C**). Diagnostic uncertainty was defined as patients with values higher than or equal  
216 to the upper quartile of 20 mm on the VAS. The acute disease VOC biomarker score was able to identify  
217 acute disease with an AUC 0.96 (0.92-0.99)  $P <0.0001$ , sensitivity 0.90 (0.82-0.97), specificity 0.92 (0.85-  
218 0.99), PPV 0.93 (0.86-0.99), NPV 0.89 (0.81-0.97) (**Figure 3D**).

219

#### 220 **2.4. Exhaled breath biomarker disease-specific scores correlate with blood-based biomarkers and** 221 **admission observations**

222 As previously described, VOC biomarker scores were generated for each of the acute disease  
223 subgroups and healthy subjects without cardiorespiratory breathlessness. There was a weak but  
224 positive correlation in the combined discovery and replication cohorts ( $n=277$ ) between the VOC  
225 subgroup scores for pneumonia and CRP ( $n=277$ ,  $r=0.33$ ,  $P <0.0001$ ) and acute heart failure and Brain  
226 Natriuretic Peptide (BNP) ( $n=277$ ,  $r=0.33$ ,  $P <0.0001$ ), in addition to a negative correlation between  
227 the healthy-state VOC score and CRP and BNP ( $n=277$ ,  $r= -0.15$ ,  $P <0.0001$ , and  $-0.21$ ,  $P <0.0001$   
228 respectively) (**Figure 4A**). Correlations were also identified between the acute disease VOC score and  
229 vital observations carried out during triage (**Figure 4B**).

230

231 The acute disease VOC score was also associated with 2-year all-cause mortality, but not with the risk  
232 of 60-day readmission (**fig. S5**).

233

#### 234 **2.5. Diagnostic accuracy of breath biomarker scores in cardiorespiratory disease subgroups**

235 A multinomial regression model using elastic net regularization was fitted to the matrix of 101 breath  
236 biomarkers with the 10-fold cross validation repeated 1,000 times. Linear combinations of the most  
237 stable features from the multinomial regression model fitted to the 101 biomarkers formed a set of  
238 scores for predicting probability of belonging to the different disease groups (acute asthma, acute  
239 COPD, pneumonia, heart failure or healthy volunteers).

240 The overall classification accuracy for the statistical model generated from 101 breath biomarkers  
241 was assessed by comparing the balanced accuracy of model trained using the true class labels versus  
242 the balanced accuracy of the same model tested using randomly shuffled class labels. This process  
243 was repeated 1,000 times. The balanced accuracy is reported in (**fig. S6A**) the acute disease  
244 biomarker score in the discovery cohort, (**fig. S6B**) the acute disease biomarker score in the  
245 replication cohort and (**fig. S6C**) the multinomial biomarker scores for the five subgroups acute  
246 asthma, acute COPD, heart failure, pneumonia and healthy volunteers. NB: replication was not

247 evaluated in the subgroups as the study was not powered to do this.

248 For the pooled cohort ( $n = 277$ ), the overall classification accuracy using all five biomarker scores was  
249 0.72, 95% CI (0.67 - 0.77). The balanced accuracy for acute asthma was 0.83, for acute COPD 0.78, for  
250 heart failure 0.80, for community acquired pneumonia 0.79, and for healthy controls was 0.93 (**fig**  
251 **S5**).

252 Further comparative ROC analyses were performed based upon the observed separation of asthma  
253 from pneumonia/COPD acute groups, and heart failure from other acute exacerbation groups in the  
254 discovery and replication TDA analyses. The diagnostic AUC accuracy of the asthma biomarker score  
255 against pooled Pneumonia and COPD cohorts was AUC: 0.70 (0.62-0.78)  $P < 0.0001$ , sensitivity 0.72  
256 (0.64-0.83), specificity 0.64 (0.55-0.73), positive predictive value (PPV) 0.54 (0.43-0.64), negative  
257 predictive value (NPV) 0.80 (0.72-0.88). Receiver operating curve (ROC) analysis to assess the  
258 diagnostic value of the heart failure biomarker score against other acute disease groups was AUC:  
259 0.78 (0.70-0.86)  $P < 0.0001$ , sensitivity 0.77 (0.64-0.89), specificity 0.71 (0.64-0.78), PPV 0.40 (0.29-  
260 0.50), NPV 0.92 (0.88-0.97) (**fig. S7**).

261 The median values of the exhaled breath VOC scores and their distribution across disease subgroups  
262 are detailed in (**fig. S8**). **Figure S9** is a Venn diagram demonstrating the distribution of the final panel  
263 of 101 exhaled breath biomarkers across the different disease groups.

264  
265 We also ran our models adjusting for the following factors: (i) smoking status (current, ex-smoker or never  
266 smoker; (ii) time between hospital admission and the acquisition of the breath samples, as this time period  
267 is often the period within which acute treatments are delivered; (iii) the modified early warning score 2  
268 (mEWS-2), which is a composite acuity score combining respiratory rate, oxygen saturations, systolic blood  
269 pressure, heart rate, level of consciousness and confusion for each patient; and (iv) prior exposure to either  
270 antibiotics or steroids for cardiorespiratory illness in the fortnight prior to the index admission. We observed  
271 only marginally improved diagnostic accuracy; acute asthma - AUC 0.88 (0.831,0.933),  $P < 0.0001$ , COPD -  
272 AUC 0.86, (0.808,0.918) ,  $P < 0.0001$ , heart failure - AUC 0.91 (0.849,0.969)  $P < 0.0001$ , community acquired  
273 pneumonia – AUC 0.91 (0.863,0.953),  $P < 0.0001$ , and healthy controls AUC 1.0, suggesting limited  
274 confounding influence of disease acuity on our biomarker scores (**Auxiliary supp table 1**). Replication was  
275 not performed in the subgroups, as the EMBER study was not powered for disease subgroup diagnostic  
276 accuracy.

277

## 278 **2.6. Chemical classification of predictive markers in disease groups**

279 Chemical identification of the 101-biomarker panel involved comparison with an authentic reference  
280 compound in accordance with the Metabolomics Standard Initiative (MSI) Level 1 criteria for  
281 metabolite identification. The most common chemical classes associated with acute breathlessness in  
282 this study included straight-chain and methyl-branched hydrocarbons (30%), ketones (10%),  
283 aldehydes (8%) and terpenes (13%), followed by sulphur-containing VOCs (7%), alcohols (6%),  
284 aromatics (5%), esters (3%), nitrogen-containing VOCs (3%), ethers (2%), halogen-compounds(1%),  
285 and an assortment of other less prevalent and less relevant classes such as acrylates (12%) (**Table S9**).

286

## 287 **2.7. Metabolite set enrichment and chemical similarity analysis**



288 Unlike functional indications, which are reliant on mapping metabolites with known, well-annotated  
289 metabolic pathways, metabolic changes indicative of response can be derived independently. To  
290 derive clues of responsive indication, the panel of 101 features was assessed for co-varying clusters  
291 (i.e. metabolite sets).

292 Metabolite sets were derived based on Ward hierarchical cluster analysis using the ChemRICH  
293 method reported previously (13) (**Figure 5A & figure S10**), and broader communities were derived  
294 from Louvain cluster analysis (**Figure 5B and tables S10-S13**) to help interpret the correlation graphs.  
295 Overall, twenty metabolite sets were identified using ChemRICH, eleven of which were enriched  
296 during acute cardiorespiratory exacerbations. The seven metabolite sets that were upregulated  
297 consisted of predominantly acyclic and branched hydrocarbons (**sets 3, 5, 7 and 9 in figure S10**). The  
298 results from the analysis herein demonstrated enriched, co-expression of hydrocarbons with high  
299 chemical similarity providing primary evidence of exhaled VOCs indicative of disease response  
300 measured in vivo. This is clearly seen in Figure 5A, with the metabolite sets (inner tree) labelled by  
301 broader chemical classifications (outer ring); C<sub>5-7</sub>, C<sub>8-10</sub> and C<sub>11-16</sub> form clusters based on carbon  
302 number also exhibiting the highest change during acute exacerbation. Owing to the increased  
303 separation power afforded by GCxGC-MS, it was possible to map the VOC signatures back to the  
304 multidimensional chromatograms for the visualisation of exhaled breath metabolites which revealed  
305 distinct diagnostic signatures for acute cardio-respiratory breathlessness (**Figure 5C**).

306  
307

### 308 **3. Discussion**

309 In this pragmatic, acute-care study, we evaluated the validity of breath biomarker profiling in high-  
310 acuity patients presenting with acute cardiorespiratory breathlessness. Using GCxGC-MS, we  
311 observed that robust and validated sampling of alveolar breath coupled with GCxGC-MS biomarker  
312 characterisation demonstrated high diagnostic accuracy for acute cardiorespiratory exacerbations.  
313 We have also identified putative biomarker scores from subsets of breath VOC biomarkers that  
314 classify cardiorespiratory exacerbation subtypes and warrant validation in appropriately powered  
315 replication studies. Furthermore, we have identified several classes of VOCs that are highly correlated  
316 and selectively enriched or suppressed in acute disease (including subgroups) compared to health,  
317 providing potential insights into broad dysregulation of the metabolome in acute cardiorespiratory  
318 exacerbations.

319 The analytical methods described herein were underpinned by robust biomarker development  
320 protocols using TD-GCxGC-FID/MS, integral to the standardisation and integration of breath analysis  
321 in large translational studies (5, 14). Several potential confounders including batch variation were  
322 addressed in detail. Furthermore, biomarker quantification of the 101 VOCs followed the  
323 recommendations of the MSI, with 58 compounds identified against pure and traceable standards  
324 (level I), 21 putative identities based on mass spectral and retention index library matches (level 2),  
325 and 22 classified on mass spectral data (15). Markers that appeared to localise to individual  
326 cardiorespiratory conditions could be readily visualised using TDA.

327

328 The identification of hydrocarbons and carbonyls as the major chemical classes was consistent with current  
329 mechanistic understanding, postulated as chemical endpoints of lipid peroxidation resulting from oxidative  
330 stress during inflammation. Aldehydes such as nonanal, decanal and hexanal were predictive for asthma;  
331 ketones included 2-pentanone (asthma), cyclohexanone (pneumonia) and 2,3- butanedione (COPD) which  
332 were all previously reported (4, 16-20). Individual hydrocarbons such as 2,4- and 2,2-dimethylpentane, 2-  
333 methylbutane, 4-methyldecane, 5-methylnonane and isoprene have been previously reported as predictive  
334 for pneumonia and heart failure (18, 21). Sulphur-containing VOCs, such as 3-methylthiophene, allyl methyl  
335 sulphide and carbonyl sulphide (found to be predictive of COPD) are associated with bacterial metabolism,  
336 postulated to originate from the gut (22) and on occasions as a result of radiation injury (23); however, 2,3-  
337 butanedione, also predictive of COPD, has been identified as a metabolic product of bacterial isolates from  
338 patients with cystic fibrosis (CF) (22) and postulated to be an important metabolite in monitoring lung  
339 infection in CF, COPD and pneumonia. We acknowledge that the biological origin of most VOCs within our  
340 biomarker signature has yet to be fully elucidated. Future studies combining carbon labelling of glucose  
341 with in vitro headspace analysis of primary cells will be required to more precisely establish the molecular  
342 origins of VOCs identified in this report.

343 Not all compounds were considered to be endogenous VOCs, with 27 possibly attributed to potential  
344 cosmetics. Eleven of the features predictive of the control group were assigned as either possible  
345 fragrances (e.g. alpha isomethyl ionone) or waxy long-chain chemicals used in cosmetics as emollients and  
346 surfactants (e.g. stearyl vinyl ether and isopropyl myristate). These may have been captured in the breath  
347 sample because of the proximity of the sorbent tubes to the patients' faces. It should be noted that a  
348 frequent problem with ascribing the origin of VOCs is that those compounds often identified in cosmetics  
349 are natural products, therefore there is uncertainty about the precise origin of these markers. The  
350 downregulation in acute disease of several of these markers may be indicative of them being biomarkers as  
351 opposed to exogenous confounders from cosmetics.

352  
353 Co-expression and enrichment analysis of the Louvain clusters on the correlation graph revealed a  
354 set of highly correlated metabolites significantly enriched in specific disease groups. Comparison of  
355 the Louvain clusters with the metabolite sets identified using the method previously  
356 described (13) demonstrated strong overlap. The metabolites enriched in heart failure were  
357 a cluster of highly correlated C<sub>5-7</sub> hydrocarbons and C<sub>3-5</sub> carbonyls with high chemical  
358 similarity (based on Tanimoto coefficients as determined in (fig. S10)). The cluster included  
359 2,4- and 2,2-dimethylpentane, 2- methylbutane, 2-methyl-1,3-butadiene (isoprene), 3-  
360 methylpentane, hexane and cyclohexane. These hydrocarbons (2,4- and 2,2-dimethylpentane, 2-  
361 methylbutane, and isoprene) have been individually reported and associated with heart failure and  
362 pneumonia (17, 20). However, the analysis herein captured the collective response and  
363 demonstrated enriched, co-expression of these hydrocarbons.

364 The analysis also revealed a separate set of highly correlated aldehydes (nonanal, decanal,  
365 undecanal, and a methyldecanal isomer), found to be potentially depleted in acute exacerbations of  
366 asthma compared with acute exacerbations of COPD and pneumonia. Depletion of VOCs during in  
367 vitro experiments has been reported as a consequence of metabolic activity by immune cells (24-26),  
368 but the association herein is tentative and should be interpreted with caution due to the correlation  
369 between inhaled air and exhaled air concentrations of these compounds (median Spearman rank =  
370 0.60), also previously observed (27).

371

372 Our study has some limitations. Although internally replicated, the results presented here for acute VOC  
373 biomarker scores and cardiorespiratory exacerbation subtype biomarker scores are limited by the lack of  
374 external replication and internal replication respectively. The single centre design of this study may have  
375 introduced nonpathogenic biases related to diet, environment and lifestyle that might be absent in a multi-  
376 center study. The cardiorespiratory exacerbation disease subgroups pre-selected in this study were chosen  
377 as the commonest reported causes of cardiorespiratory breathlessness (28, 29) and there was a relatively  
378 high degree of clinical certainty in the diagnostic labels. For these findings to be generalisable, the identified  
379 markers will need to be validated in unselected cardiorespiratory populations and patients presenting with  
380 mixed acute pathologies.

381

382 In conclusion, we have conducted an acute care volatile breath biomarker study using robust clinical  
383 and analytical technology and have identified biomarkers with high combined diagnostic sensitivity  
384 and specificity in acute cardiorespiratory disease. In addition, we have used methods enabling robust  
385 biomarker identification and mechanistic association. Future clinical studies in acute  
386 cardiorespiratory patients at initial presentation and triage using near patient sensor platforms  
387 capable of detecting the volatiles identified in this report are warranted to maximise the clinical  
388 impact of our discovery biomarker approach.

389

390

## 391 **4. Materials and Methods**

### 392 **4.1. Study design**

393 The study design, eligibility criteria and methodology have been described in detail previously (30).  
394 This is a prospective, real-world, observational study (ClinicalTrials.gov Identifier NCT03672994),  
395 carried out in a tertiary cardiorespiratory centre in Leicester, United Kingdom. Participants were  
396 recruited year-round from May 2017 through to December 2018.

397 Patients with self-reported acute breathlessness, requiring admission and/or a change in baseline  
398 treatment, presenting within University Hospitals of Leicester (UHL) were approached for study  
399 participation. Following triage and senior clinical assessment, if a primary clinical diagnosis of (i) acute  
400 decompensation of heart failure, (ii) exacerbation of asthma/COPD, or (iii) adult community acquired  
401 pneumonia was suspected by the triage nurse/attending clinician at triage, members of the research  
402 team would evaluate patients against predefined eligibility criteria for study participation.

403 A total of 277 participants were included in the final analysis. Sample size attrition from the recruited  
404 455 participants is detailed in **(Figure 1)**. This was mainly due to the delayed deployment of GCxGC-  
405 MS and analytical QC/QA. These decisions were made objectively during the discovery phase of the  
406 program, prioritising the optimisation of a robust sampling and analysis pathway. Sample size  
407 calculations were informed based on estimation for adequate sensitivity and or specificity as detailed  
408 in **(table S1)**.

409 The 277 subjects were randomised *post-hoc* to Discovery and Replication cohorts in a 1:1 ratio  
410 through block random assignment. Randomisation was stratified based on (i) adjudicated clinical  
411 diagnosis, (ii) time to breath-testing from the point of hospital admission, and (iii) clinical diagnostic  
412 uncertainty score. The R package randomizr was used to perform block random assignment.

413 After block randomisation there were 139 and 138 subjects in the discovery and replication sets  
414 respectively.

415

416 Inclusion and exclusion criteria and study objectives are outlined in detail in 'study design' and 'study  
417 objectives' sections of the Supplementary material. Informed consent was obtained in all participants  
418 within 24 hours of hospitalisation. Age- and/or home environment-matched healthy volunteers were  
419 recruited. Where environment-matched controls were unsuitable, healthy volunteers were recruited  
420 from local recruitment databases and via advertising. Healthy volunteers were defined as participants  
421 with no prior history of asthma, COPD, heart failure and had not been admitted to hospital with  
422 community acquired pneumonia within 6 weeks of the baseline study visit. The diagnostic accuracy of  
423 the reported exhaled breath VOCs was tested following the Standards for reporting of Diagnostic  
424 Accuracy Studies guidelines (31) (table S14). Statistical procedures presented here were carried out  
425 as complete case analysis with no imputations. Transparent Reporting of multivariate prediction  
426 model for Individual Prognosis or Diagnosis (TRIPOD) was followed for multivariate prediction models  
427 (32, 33) (table S15).

428

429

430 The trial was conducted in accordance with the ethics and principles of the declaration of Helsinki  
431 and Good Clinical Practice Guidelines. All patients provided written consent. The National Research  
432 Ethics Service Committee East Midlands has approved the study protocol (REC number: 16/LO/1747).  
433 Integrated Research Approval System (IRAS) 198921.

434

#### 435 **4.2. Clinical adjudication**

436 A clinical adjudication process was introduced to precisely define and quantify the diagnostic labels in  
437 the study, addressing any potential misclassification. A panel of two senior clinical adjudicators (SS &  
438 NG) reviewed all available case notes and imaging and determined the primary diagnosis for each  
439 case by discussion to reach a concordance. The degree of diagnostic uncertainty was marked on a  
440 100-mm visual analogue scale (VAS scale), blinded to given diagnosis and blood biomarkers.

441 The process was implemented with emphasis on mirroring an acute triage pathway, where all  
442 pathology data required to support the diagnosis e.g. CRP, BNP are not available at the initial clinical  
443 review. The degree of diagnostic uncertainty obtained from the clinical adjudication process was  
444 factored into the block randomisation and subjects with higher diagnostic uncertainty ( $\geq$ upper  
445 quartile = 20mm) were assessed separately as previously described (Figure 3C-D).

446

#### 447 **4.3. Breath collection and analysis**

##### 448 **4.3.1. Collection of breath samples**

449 Exhaled breath collection was attempted in all consented participants using a CE marked breath  
450 sampling device 'Respiration Collector for In Vitro Analysis' RECIVA (Owlstone Nanotech Ltd), in  
451 combination with a dedicated clean air supply unit (34). Breath sampling was well tolerated by all  
452 participants (6).

##### 453 **4.3.2. Sample storage and preparation**

454 Samples were dry purged on arrival for two minutes using nitrogen (chemically pure grade with inline  
455 trap, BOC) at a flow rate of 50 mL min<sup>-1</sup> and then stored in refrigeration at 2 °C until analysis. Before  
456 analysis, samples were left to reach room temperature before being spiked with a 0.6 µL aliquot of 20  
457 µg mL<sup>-1</sup> standard solution containing deuterated toluene and octane, into a flow of nitrogen at a flow  
458 rate of 100 mL min<sup>-1</sup> for 2 min, purging the excess solvent.

#### 459 **4.3.3. Exhaled Breath analysis**

460 Breath samples were analysed by thermal desorption with comprehensive two-dimensional gas  
461 chromatography (GCxGC) using flow modulation and coupled to dual flame ionisation detection and  
462 mass spectrometry (MS). Dual detection, with the use of MS and flame ionisation detection (FID),  
463 utilises the excess flow from the flow-based modulator suited for volatile analyses, providing both  
464 quantitative and qualitative results.

465 Analysis by GCxGC was optimised and conducted as described previously (5), using an Agilent 7890A  
466 gas chromatogram, fitted with a CFT flow modulator and 5799B mass spectrometer with a high  
467 efficiency EI ion source (Agilent Technologies Ltd). The instrument was coupled to a TD-100xr thermal  
468 desorption auto-sampler (Markes International Ltd). Samples were analysed in trays; typically six per  
469 tray along with a reference mixture containing n-alkanes and aromatics run every tray and a  
470 reference indoor air VOC mixture run every four trays. Data was acquired in MassHunter GC-MS  
471 Acquisition B.07.04.2260 (Agilent) and processed (i.e. baseline correction, alignment, feature  
472 extraction) with a workflow previously developed and optimised, using GC Image™ v2.8 suite (GC  
473 Image, LLC.) and Python (14). The sorbent tubes used were Tenax/TA with Carbograph 1TD  
474 (Hydrophobic, Markes International Ltd) with matching cold trap. Chromatographic features arising  
475 from analytical artefacts were removed from the peak table. (e.g. ubiquitous siloxanes). For purposes  
476 of quality control, samples were analysed in accordance with a previously published workflow and a  
477 detailed sample history, metadata and experimental data were recorded at every stage of the  
478 collection and analysis using the open-access LabPipe toolkit (5, 35).

#### 479 **4.3.4. Chemical speciation of identified breath biomarkers**

480 The chemical nature of volatile metabolites exhaled in breath comprises a diverse mixture of non-  
481 novel, low-molecular weight compounds. Thus, for most features, chemical identification involved  
482 comparison with an authentic reference compound in accordance with the Metabolomics Standard  
483 Initiative (MSI) Level 1 criteria for metabolite identification outlined in table S9. Identification was  
484 based on a minimum of two independent and orthogonal identifiers including primary and secondary  
485 retention time, mass spectral similarity match and calculated retention index. When an authentic  
486 reference compound was unavailable, chemical identification was compliant with MSI Level 2 for  
487 putative annotations. The highly structured chromatographic data and group-type separation  
488 afforded by GCxGC, alongside a well-characterised chromatographic space from analysing an  
489 extensive library of authentic compounds, gave increased confidence in the tentative assignments  
490 made. The orthogonal separation of GCxGC also meant chemical identification of unknown  
491 metabolites could be made, at minimum, in compliance with MSI Level 3 for putative chemical  
492 classification.

#### 493 **4.3.5. Sample analysis quality control/quality assurance (QC/QA) procedures**

494 For purposes of quality control, samples were analysed in accordance with a previously published  
495 workflow and a detailed sample history, metadata and experimental data were recorded at every

496 stage of the collection and analysis using the open-access LabPipe toolkit (35). The chromatographic  
497 method was optimised for peak shape, sensitivity and separation; quality control charts of the  
498 internal standards were used to track the stability of the TD-GCxGC-FID/MS analysis, and instrument  
499 performance was evaluated following the assessment of the variation of retention times, peak area  
500 and shapes of VOCs in two standard reference mixtures every six samples (5). Before being  
501 conditioned and sent to clinic, the number of heat cycles and weight for each tube was recorded to  
502 monitor tube age and integrity. For each conditioning cycle, all tubes were given a batch number and  
503 a batch blank was analysed to monitor contamination from the beginning of the sample preparation  
504 process. Furthermore, all batches were given an expiry of two weeks to ensure routine monitoring.

505 To minimise the influence of biological and analytical confounders (e.g. circadian rhythm, sample  
506 stability), potential effects due to the operator, date of analysis, time of day collected, storage time  
507 before dry purging, sample storage time after dry purging and collection volume were assessed and  
508 where necessary accounted for in the batch correction. In addition to the routine analysis of  
509 reference standards, used to monitor retention shift and instrument response, the TD-GCxGC  
510 analytical system underwent a programmed heat cycle between each sample to reduce potential  
511 issues arising from sample carry-over, and a TD-trap blank and empty sorbent tube were analysed  
512 every six samples to monitor the instrument baseline signal.

513

#### 514 **4.3.1. Topological data analysis in the discovery and replication sets**

515 In topological data analysis, the x-y coordinate position of a particular patient within a TDA cluster  
516 cannot be directly compared between discovery and replication TDA graphs, as the graphs represent  
517 a simple 2-dimensional projection of a higher dimensional structure. Prior to performing TDA, each  
518 feature was  $\log(x + 1)$  transformed. TDA parameters were set as: number of hypercubes=20, where  
519 the number of hypercubes refers to the number of overlapping intervals of the projection.

520 The distance between data points was measured using the Euclidean distance. The first two linear  
521 discriminant functions (LD1) and (LD2) were used as the projection. Clustering on the overlapping  
522 intervals on the projection was done using agglomerative (bottom up) hierarchical clustering with  
523 complete linkage. TDA was performed using Kepler Mapper 1.4.0 (36) with Python 3.5.

524 Herein, we computed the equivalence between topological data shapes generated using 805 volatile  
525 features extracted from the GCxGC-MS peak data, in both the discovery and replication cohorts.

#### 526 **4.3.2. Breath biomarker score generation**

527 Feature selection was implemented via Lasso and Elastic-Net Regularized Generalized Linear Models  
528 (GLMNET) using the glmnet package in R. After removing features present in <80% of all samples  
529 from the  $(x + 1)$  transformed discovery GCxGC-MS peak data a 735-feature matrix was obtained. A  
530 multinomial regression model using LASSO regularization was fitted to the 735-feature matrix in the  
531 discovery set using 10-fold cross validation, with the dependent variable in the model being clinical  
532 diagnosis (acute asthma, acute COPD, pneumonia, heart failure, or healthy volunteers). The 10-fold  
533 cross validation was repeated 100 times; features that had a non-zero regression coefficient in more  
534 than 80 of the cross validation runs were considered as being stable candidate features predictive of  
535 the outcome (clinical diagnosis), and this resulted in 278 stable candidate features. For validation,  
536 predictors were calculated using the *Predict Function* of (GLMNET).

537 A multinomial regression model using elastic net regularization was fitted to the 278 features with  
538 the dependent variable in the model being clinical diagnosis. Following the chemometric inspection  
539 detailed above and the lasso and elastic regression analysis, a final set of 101 exhaled breath volatile  
540 compounds was generated.

541 A multinomial regression model using elastic net regularization was fitted to the matrix of 101 breath  
542 biomarkers with the 10-fold cross validation repeated 100 times. The R package glmnetUtils was  
543 used to determine the optimal value of  $\alpha$  the elastic net penalty, the best value for  $\alpha$  was 0 (Ridge  
544 regression). Ridge regression with a logit link function (binary logistic regression) was fitted to the  
545 101 breath relevant features; the dependent variable was 'acute disease', as a binary outcome. The  
546 linear predictor from the combination of the most stable features was used to as a score to predict  
547 acute disease. Linear combinations of the most stable features from the multinomial regression  
548 model fitted to the 101 biomarkers formed a set of scores for predicting probability of belonging to  
549 the different disease groups (acute Asthma, acute COPD, pneumonia, heart failure or healthy  
550 volunteers). Sensitivity analysis for the interactive elastic net regression approach and justification of  
551 the optimal  $\alpha$  values are provided in **(figs. S11-S12 and tables S6-S8)**.

552 Figure S13 is a graphical probability distribution of the final 101 exhaled breath features in the  
553 GCxGC-MS peak data. The features largely follow a similar distribution. Some features contained a  
554 mixture of zero and non-zero values, which have arisen owing to the measurement being below the  
555 instrument's lower limit of detection. Constant features (all zero values) were removed prior to  
556 fitting the main model.

557

#### 558 **4.3.3. Breath biomarker co-expression and feature enrichment analysis**

559 It was of interest to investigate if within the final set of 101 features, sets of 'co expressed' features  
560 existed, i.e. sets containing features that are correlated. Considering sets of co-expressed features  
561 has value in terms of reducing the dimensions of a problem and mitigating the multiple testing  
562 problem through the use of enrichment score. Co-expression and feature enrichment analysis are  
563 described in the **(Supplementary material section 'co-expression and feature enrichment analysis')**.  
564 Metabolite sets were derived based on Ward hierarchical cluster analysis using the ChemRICH  
565 method reported by (13), and broader communities were derived from Louvain cluster analysis to  
566 help interpret the correlation graphs **(Supplementary material section 'co-expression and feature  
567 enrichment analysis')**. Covariation among metabolites lacks evidential value on its own, therefore,  
568 set-level significance was established using the Kolmogorov-Smirnov test (K-S test) as described using  
569 the ChemRICH method (13), Tanimoto coefficients were calculated to assess intra-set chemical  
570 similarity using Metabox (37), and the frequency of occurrence in the published literature and  
571 relevant databases considered (KEGG, ChEBI, Human Metabolome Database, Human Breathomics  
572 Database and microbial VOC database). Chemical similarity is of interest because compounds derived  
573 from similar pathways may also share common structural features or chemical groups. This  
574 combined data-driven and chemistry-driven approach has been shown to improve enrichment  
575 analysis (13, 38), and allowed further interpretation of core findings herein **(fig. S10)**.

576

577

578

579 ***Statistical procedures:***

580

581 Statistical analysis was performed using R [3.6.1 and 4.0.0, R Core Team (2019)]. This research used  
582 the SPECTRE High Performance Computing Facility at the University of Leicester. Baseline data and  
583 figures were presented as mean  $\pm$  (SD), and median (IQ range). Data was analysed using (ANOVA) to  
584 assess the differences between groups for normally or approximately normally-distributed variables  
585 and Kruskal-Wallis for non-normally distributed variables. Pearson chi-squared and Fisher's exact  
586 were used to assess the differences in categorical variables. All *P* values are two sided and significant  
587 at the 0.05 level, unless reported otherwise.

588

589

590 **List of supplementary materials:**

591 Materials and methods

592 Figs. S1 to S13

593 Tables S1 To S15

594 MDAR Reproducibility Checklist

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617



619 **References**

- 620 1. A. Hutchinson, A. Pickering, P. Williams, J. M. Bland, M. J. Johnson, Breathlessness and presentation  
621 to the emergency department: a survey and clinical record review. *BMC Pulmonary Medicine* **17**, 53  
622 (2017); published online Epub2017/03/20 (10.1186/s12890-017-0396-4).
- 623 2. M. B. Parshall, R. M. Schwartzstein, L. Adams, R. B. Banzett, H. L. Manning, J. Bourbeau, P. M.  
624 Calverley, A. G. Gift, A. Harver, S. C. Lareau, D. A. Mahler, P. M. Meek, D. E. O'Donnell, D. American  
625 Thoracic Society Committee on, An official American Thoracic Society statement: update on the  
626 mechanisms, assessment, and management of dyspnea. *American journal of respiratory and critical*  
627 *care medicine* **185**, 435-452 (2012)10.1164/rccm.201111-2042ST).
- 628 3. W. Ibrahim, L. Carr, R. Cordell, M. J. Wilde, D. Salman, P. S. Monks, P. Thomas, C. E. Brightling, S.  
629 Siddiqui, N. J. Greening, Breathomics for the clinician: the use of volatile organic compounds in  
630 respiratory diseases. *Thorax*, thoraxjnl-2020-215667 (2021)10.1136/thoraxjnl-2020-215667).
- 631 4. F. N. Schleich, D. Zanella, P. H. Stefanuto, K. Bessonov, A. Smolinska, J. W. Dallinga, M. Henket, V.  
632 Paulus, F. Guissard, S. Graff, C. Moermans, E. F. M. Wouters, K. Van Steen, F. J. van Schooten, J. F.  
633 Focant, R. Louis, Exhaled Volatile Organic Compounds are Able to Discriminate between  
634 Neutrophilic and Eosinophilic Asthma. *Am J Respir Crit Care Med*, (2019); published online EpubApr  
635 11 (10.1164/rccm.201811-2210OC).
- 636 5. M. J. Wilde, R. L. Cordell, D. Salman, B. Zhao, W. Ibrahim, L. Bryant, D. Ruszkiewicz, A. Singapuri, R.  
637 C. Free, E. A. Gaillard, C. Beardsmore, C. L. P. Thomas, C. E. Brightling, S. Siddiqui, P. S. Monks,  
638 Breath analysis by two-dimensional gas chromatography with dual flame ionisation and mass  
639 spectrometric detection - Method optimisation and integration within a large-scale clinical study. *J*  
640 *Chromatogr A* **1594**, 160-172 (2019); published online EpubJun 7 (10.1016/j.chroma.2019.02.001).
- 641 6. K. A. Holden, W. Ibrahim, D. Salman, R. Cordell, T. McNally, B. Patel, R. Phillips, C. Beardsmore, M.  
642 Wilde, L. Bryant, A. Singapuri, P. Monks, C. Brightling, N. Greening, P. Thomas, S. Siddiqui, E. A.  
643 Gaillard, Use of the ReCIVA device in breath sampling of patients with acute breathlessness: a  
644 feasibility study. *ERJ Open Research* **6**, 00119-02020 (2020)10.1183/23120541.00119-2020).
- 645 7. P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, G.  
646 Carlsson, Extracting insights from the shape of complex data using topology. *Scientific Reports* **3**,  
647 1236 (2013); published online Epub2013/02/07 (10.1038/srep01236).
- 648 8. J. L. Nielson, J. Paquette, A. W. Liu, C. F. Guandique, C. A. Tovar, T. Inoue, K.-A. Irvine, J. C. Gensel, J.  
649 Kloke, T. C. Petrossian, P. Y. Lum, G. E. Carlsson, G. T. Manley, W. Young, M. S. Beattie, J. C.  
650 Bresnahan, A. R. Ferguson, Topological data analysis for discovery in preclinical spinal cord injury  
651 and traumatic brain injury. *Nature communications* **6**, 8581-8581 (2015)10.1038/ncomms9581).
- 652 9. E. Somasundaram, A. Litzler, R. Wadhwa, S. Owen, J. Scott, Persistent homology of tumor CT scans  
653 is associated with survival in lung cancer. *Med Phys* **48**, 7043-7051 (2021); published online  
654 EpubNov (10.1002/mp.15255).
- 655 10. S. Siddiqui, A. Shikotra, M. Richardson, E. Doran, D. Choy, A. Bell, C. D. Austin, J. Eastham-Anderson,  
656 B. Hargadon, J. R. Arron, A. Wardlaw, C. E. Brightling, L. G. Heaney, P. Bradding, Airway pathological  
657 heterogeneity in asthma: Visualization of disease microclusters using topological data analysis. *J*  
658 *Allergy Clin Immunol* **142**, 1457-1468 (2018); published online EpubNov  
659 (10.1016/j.jaci.2017.12.982).
- 660 11. M. Nicolau, A. J. Levine, G. Carlsson, Topology based data analysis identifies a subgroup of breast  
661 cancers with a unique mutational profile and excellent survival. *Proc Natl Acad Sci U S A* **108**, 7265-  
662 7270 (2011); published online EpubApr 26 (10.1073/pnas.1102826108).
- 663 12. J. Brandsma, V. M. Goss, X. Yang, P. S. Bakke, M. Caruso, P. Chanez, S. E. Dahlén, S. J. Fowler, I.  
664 Horvath, N. Krug, P. Montuschi, M. Sanak, T. Sandström, D. E. Shaw, K. F. Chung, F. Singer, L. J.  
665 Fleming, A. R. Sousa, I. Pandis, A. T. Bansal, P. J. Sterk, R. Djukanović, A. D. Postle, Lipid phenotyping

- 666 of lung epithelial lining fluid in healthy human volunteers. *Metabolomics* **14**, 123 (2018); published  
667 online EpubSep 17 (10.1007/s11306-018-1412-2).
- 668 13. D. K. Barupal, O. Fiehn, Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to  
669 biochemical pathway mapping for metabolomic datasets. *Scientific Reports* **7**, 14567 (2017);  
670 published online Epub2017/11/06 (10.1038/s41598-017-15231-w).
- 671 14. M. J. Wilde, B. Zhao, R. L. Cordell, W. Ibrahim, A. Singapuri, N. J. Greening, C. E. Brightling, S.  
672 Siddiqui, P. S. Monks, R. C. Free, Automating and extending comprehensive two-dimensional gas  
673 chromatography data processing by interfacing open-source and commercial software. *Anal Chem*,  
674 (2020); published online EpubSep 28 (10.1021/acs.analchem.0c02844).
- 675 15. L. W. Sumner, A. Amberg, D. Barrett, M. H. Beale, R. Beger, C. A. Daykin, T. W. M. Fan, O. Fiehn, R.  
676 Goodacre, J. L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. N. Lane, J. C.  
677 Lindon, P. Marriott, A. W. Nicholls, M. D. Reily, J. J. Thaden, M. R. Viant, Proposed minimum  
678 reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics  
679 Standards Initiative (MSI). *Metabolomics : Official journal of the Metabolomic Society* **3**, 211-221  
680 (2007)10.1007/s11306-007-0082-2).
- 681 16. M. Basanta, B. Ibrahim, R. Dockry, D. Douce, M. Morris, D. Singh, A. Woodcock, S. J. Fowler, Exhaled  
682 volatile organic compounds for phenotyping chronic obstructive pulmonary disease: a cross-  
683 sectional study. *Respir Res* **13**, 72 (2012); published online EpubAug 23 (10.1186/1465-9921-13-72).
- 684 17. S. J. Fowler, M. Basanta-Sanchez, Y. Xu, R. Goodacre, P. M. Dark, Surveillance for lower airway  
685 pathogens in mechanically ventilated patients by metabolomic analysis of exhaled breath: a case-  
686 control study. *Thorax* **70**, 320-325 (2015); published online EpubApr (10.1136/thoraxjnl-2014-  
687 206273).
- 688 18. A. Pizzini, W. Filipiak, J. Wille, C. Ager, H. Wiesenhofer, R. Kubinec, J. Blaško, C. Tschurtschenthaler,  
689 C. A. Mayhew, G. Weiss, R. Bellmann-Weiler, Analysis of volatile organic compounds in the breath  
690 of patients with stable or acute exacerbation of chronic obstructive pulmonary disease. *Journal of*  
691 *breath research* **12**, 036002 (2018); published online EpubMar 2018 (
- 692 19. R. Peltrini, R. Cordell, W. Ibrahim, M. Wilde, D. Salman, A. Singapuri, B. Hargadon, C. E. Brightling, C.  
693 L. P. Thomas, P. Monks, S. Siddiqui, Volatile organic compounds in a headspace sampling system  
694 and asthmatics sputum samples. *J Breath Res*, (2020); published online EpubNov 23  
695 (10.1088/1752-7163/abcd2a).
- 696 20. D. Zanella, M. Henket, F. Schleich, T. Dejong, R. Louis, J.-F. Focant, P.-H. Stefanuto, Comparison of  
697 the effect of chemically and biologically induced inflammation on the volatile metabolite  
698 production of lung epithelial cells by GC×GC-TOFMS. *The Analyst* **145**, 5148-5157 (2020); published  
699 online EpubAug 2020 (
- 700 21. R. Schnabel, R. Fijten, A. Smolinska, J. Dallinga, M.-L. Boumans, E. Stobberingh, A. Boots, P.  
701 Roekaerts, D. Bergmans, F. J. van Schooten, Analysis of volatile organic compounds in exhaled  
702 breath to diagnose ventilator-associated pneumonia. *Scientific reports* **5**, 17179 (2015); published  
703 online EpubNov 2015 (
- 704 22. J. Phan, S. Meinardi, B. Barletta, D. R. Blake, K. Whiteson, Stable isotope profiles reveal active  
705 production of VOCs from human-associated microbes. *J Breath Res* **11**, 017101 (2017); published  
706 online EpubFeb 6 (10.1088/1752-7163/aa5833).
- 707 23. D. Salman, M. Eddleston, K. Darnley, W. H. Nailon, D. B. McLaren, A. Hadjithelki, D. Ruskiewicz, J.  
708 Langejuergen, Y. Alkhalifa, I. Phillips, C. L. P. Thomas, Breath markers for therapeutic radiation. *J*  
709 *Breath Res* **15**, 016004 (2020); published online EpubOct 24 (10.1088/1752-7163/aba816).
- 710 24. A. Sponring, W. Filipiak, T. Mikoviny, C. Ager, J. Schubert, W. Miekisch, A. Amann, J. Troppmair,  
711 Release of volatile organic compounds from the lung cancer cell line NCI-H2087 in vitro. *Anticancer*  
712 *Res* **29**, 419-426 (2009); published online EpubJan (
- 713 25. W. Filipiak, A. Sponring, A. Filipiak, C. Ager, J. Schubert, W. Miekisch, A. Amann, J. Troppmair, TD-  
714 GC-MS analysis of volatile metabolites of human lung cancer and normal cells in vitro. *Cancer*

- 715 *Epidemiol Biomarkers Prev* **19**, 182-195 (2010); published online EpubJan (10.1158/1055-9965.Epi-  
716 09-0162).
- 717 26. A. Sponring, W. Filipiak, C. Ager, J. Schubert, W. Miekisch, A. Amann, J. Troppmair, Analysis of  
718 volatile organic compounds (VOCs) in the headspace of NCI-H1666 lung cancer cells. *Cancer*  
719 *Biomark* **7**, 153-161 (2010)10.3233/cbm-2010-0182).
- 720 27. M. Koureas, P. Kirgou, G. Amoutzias, C. Hadjichristodoulou, K. Gourgoulanis, A. Tsakalof, Target  
721 Analysis of Volatile Organic Compounds in Exhaled Breath for Lung Cancer Discrimination from  
722 Other Pulmonary Diseases and Healthy Persons. *Metabolites* **10**,  
723 (2020)10.3390/metabo10080317).
- 724 28. M. K. David Ponka, Top differential diagnoses in family medicine Dyspnoea Can Fam Physician.  
725 *Canadian Family Physician*, (2008).
- 726 29. S. Laribi, G. Keijzers, O. van Meer, S. Klim, J. Motiejunaite, W. S. Kuan, R. Body, P. Jones, M.  
727 Karamercan, S. Craig, V. P. Harjola, A. Holdgate, A. Golea, C. Graham, F. Verschuren, J. Capsec, M.  
728 Christ, L. Grammatico-Guillon, C. Barletta, L. Garcia-Castrillo, A. M. Kelly, Epidemiology of patients  
729 presenting with dyspnea to emergency departments in Europe and the Asia-Pacific region. *Eur J*  
730 *Emerg Med* **26**, 345-349 (2019); published online EpubOct (10.1097/mej.0000000000000571).
- 731 30. W. Ibrahim, M. Wilde, R. Cordell, D. Salman, D. Ruzskiewicz, L. Bryant, M. Richardson, R. C. Free, B.  
732 Zhao, A. Yousuf, C. White, R. Russell, S. Jones, B. Patel, A. Awal, R. Phillips, G. Fowkes, T. McNally, C.  
733 Foxon, H. Bhatt, R. Peltrini, A. Singapuri, B. Hargadon, T. Suzuki, L. L. Ng, E. Gaillard, C. Beardsmore,  
734 K. Ryanna, H. Pandya, T. Coates, P. S. Monks, N. Greening, C. E. Brightling, P. Thomas, S. Siddiqui,  
735 Assessment of breath volatile organic compounds in acute cardiorespiratory breathlessness: a  
736 protocol describing a prospective real-world observational study. *BMJ open* **9**, e025486 (2019);  
737 published online EpubMar 2019 (
- 738 31. P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. Irwig, J. G. Lijmer, D.  
739 Moher, D. Rennie, H. C. de Vet, H. Y. Kressel, N. Rifai, R. M. Golub, D. G. Altman, L. Hooft, D. A.  
740 Korevaar, J. F. Cohen, STARD 2015: an updated list of essential items for reporting diagnostic  
741 accuracy studies. *BMJ* **351**, h5527 (2015); published online EpubOct 28 (10.1136/bmj.h5527).
- 742 32. G. S. Collins, J. B. Reitsma, D. G. Altman, K. G. Moons, Transparent Reporting of a multivariable  
743 prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. *Br J Surg*  
744 **102**, 148-158 (2015); published online EpubFeb (10.1002/bjs.9736).
- 745 33. K. G. M. Moons, D. G. Altman, J. B. Reitsma, J. P. A. Ioannidis, P. Macaskill, E. W. Steyerberg, A. J.  
746 Vickers, D. F. Ransohoff, G. S. Collins, Transparent Reporting of a multivariable prediction model for  
747 Individual Prognosis Or Diagnosis (TRIPOD): Explanation and ElaborationThe TRIPOD Statement:  
748 Explanation and Elaboration. *Annals of Internal Medicine* **162**, W1-W73 (2015)10.7326/m14-0698).
- 749 34. S. Kitchen, A. Edge, R. Smith, P. Thomas, S. Fowler, S. Siddiqui, M. van der Schee, LATE-BREAKING  
750 ABSTRACT: Breathe free: Open source development of a breath sampler by a consortium of breath  
751 researchers. *European Respiratory Journal* **46**, (2015)10.1183/13993003.congress-2015.PA3987).
- 752 35. B. Zhao, L. Bryant, R. Cordell, M. Wilde, D. Salman, D. Ruzskiewicz, W. Ibrahim, A. Singapuri, T.  
753 Coats, E. Gaillard, C. Beardsmore, T. Suzuki, L. Ng, N. Greening, P. Thomas, P. Monks, C. Brightling,  
754 S. Siddiqui, R. C. Free, LabPipe: an extensible bioinformatics toolkit to manage experimental data  
755 and metadata. *BMC Bioinformatics* **21**, 556 (2020); published online Epub2020/12/02  
756 (10.1186/s12859-020-03908-5).
- 757 36. H. J. v. V. a. N. S. a. D. E. a. S. W. Mangham, Kepler Mapper: A flexible Python implementation of  
758 the Mapper algorithm. *Zenodo*, (2021)<https://doi.org/10.5281/zenodo.4754451>  
759 ).
- 760 37. K. Wanichthanarak, S. Fan, D. Grapov, D. K. Barupal, O. Fiehn, Metabox: A Toolbox for Metabolomic  
761 Data Analysis, Interpretation and Integrative Exploration. *PLoS One* **12**, e0171046  
762 (2017)10.1371/journal.pone.0171046).

763 38. J. R. Ash, M. A. Kuenemann, D. Rotroff, A. Motsinger-Reif, D. Fourches, Cheminformatics approach  
764 to exploring and modeling trait-associated metabolite profiles. *J Cheminform* **11**, 43 (2019);  
765 published online EpubJun 24 (10.1186/s13321-019-0366-3).

#### 766 **Acknowledgments**

767 This research was funded by the Medical Research Council (MRC), Engineering and Physical Sciences  
768 Research Council (EPSRC) Stratified Medicine Grant for Molecular Pathology Nodes (Grant No.  
769 MR/N005880/1), Midlands Asthma and Allergy Research Association (MAARA) and British Lung  
770 Foundation (Grant No. BLFPHD17-1). The work was carried out at the University Hospitals of Leicester  
771 NHS Trust, University of Leicester and Loughborough University, supported by the NIHR Leicester  
772 Biomedical Research Centre and the NIHR Leicester Clinical Research Facility. The views expressed  
773 are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of  
774 Health and Social Care. The authors would also like to thank the wider East Midlands Breathomics  
775 Pathology Node consortium.

776 Access to anonymised core patient level data used to generate results in this manuscript are  
777 available via request from senior author (Prof S. Siddiqui - email: s.siddiqui@imperial.ac.uk) and are  
778 subject to MRC EMBER steering group approval and fully executed material transfer agreement with  
779 the University of Leicester as the study sponsor.

#### 780 **The EMBER Consortium**

781 In addition to EMBER Consortium members who are authors (Wadah Ibrahim, Michael J. Wilde,  
782 Rebecca L. Cordell, Matt Richardson, Dahlia Salman, Bo Zhao, Rob Free, Amisha Singapuri, Beverley  
783 Hargadon, Erol A. Gaillard, Toru Suzuki, Leong L. Ng, Tim Coats, Paul S. Monks, Paul Thomas,  
784 Christopher E. Brightling, Neil Greening, and Salman Siddiqui), the following EMBER Consortium  
785 members are collaborators who have contributed to the study design, data analysis, and  
786 interpretation:

787 Caroline S. Beardsmore<sup>1</sup>, Rachel Munton<sup>11</sup>, John Le Quesne<sup>1</sup>, Alison H. Goodall<sup>1</sup>, Hitesh C. Pandya<sup>1</sup>,  
788 James C. Reynolds<sup>4</sup>, Martha R. J. Clokie<sup>1</sup>, Nilesh J. Samani<sup>1</sup>, Andrew Wardlaw<sup>1</sup>, Michael R. Barer<sup>1</sup>, and  
789 Jacqueline A. Shaw<sup>1</sup>

790 Affiliations 1 to 10 can be found on the first page of the paper.

791 11. East Midlands Academic health Science Network, University of Nottingham Innovation Park,  
792 Nottingham, NG7 2TU UK

793

#### 794 **Funding:**

795 Research funding was provided by the MRC and EPSRC Stratified Medicine Grant for Molecular  
796 Pathology Nodes (Grant No. MR/N005880/1). The research was supported by the National Institute  
797 for Health Research (NIHR) Leicester Biomedical Research Centre and NIHR, Leicester Clinical  
798 Research Facility, the Leicester Wellcome Trust ISSF (Award No. 204801/Z/16/Z), and the Midlands  
799 Asthma and Allergy Research Association (MAARA) to whom we are extremely grateful. N.J.G. is  
800 funded by a NIHR post-doctoral fellowship (PDF-2017-10-052).

801

#### 802 **Author contribution**

803 The first 3 authors, W.I, M.J.W., and R. L.C, contributed equally.  
804 S.S, C.E.B, N.J.G, P.T and P.S.M. conceived the study, obtained funding, wrote the study protocol,  
805 obtained ethical and MHRA approvals for the study and coordinated the deployment of analytical  
806 testing methods for breath analysis. W.I. has led planning and recruitment of study participants as  
807 well as taking the lead in writing the manuscript, with support from S.S, R.L.C., N.J.G, M.J.W, and  
808 M.R. Analytical chemistry team formed of M.J.W and R.L.C led the analytical method development,  
809 the development of all breath sampling and analytical protocols, pre- and post-clinic preparation and  
810 analysis of breath samples, and data processing of chemical and analytical data. M.R, a senior  
811 statistician, constructed a statistics and data analysis plan in conjunction with S.S and M.J.W.  
812 Bioinformatics pipeline and electronic CRFs developed by R.C.F and B.Z. All authors, including D.S.,  
813 A.S., B. H., E.A.G, T.S., L.L.N, and T.C. contributed to the study design. All authors contributed to and  
814 approved the manuscript.

815

### 816 **Competing interests**

817 C.E.B. has received consultancy and or grants paid to his Institution from GlaxoSmithKline (GSK),  
818 Astrazeneca (AZ), Boehringer Ingelheim (BI), Novartis, Chiesi, Genentech, Roche, Sanofi, Regeneron,  
819 TEVA Pharmaceuticals, MSD, Mologic, CSL Behring, Gossamer and 4Dpharma. SS has received  
820 funding from the MRC/EPSRC and University of Leicester for the research program presented in the  
821 manuscript. S.S. has engaged in consultancies/received speaker fees related to asthma, COPD, lung  
822 physiology and eosinophilic airway diseases from the following companies: Boehringer Ingelheim,  
823 Chiesi, Novartis, GSK, AZ, ERT Medical, Owlstone Medical, CSL Behring, Mundipharma, Knopp  
824 biotech. The rest of the co-authors have no competing interests to declare. E.G reports consultancy  
825 work for Boehringer Ingelheim with money paid to the institution (University of Leicester);  
826 Investigator led research grant from Circassia Group, Gilead Sciences, Chiesi Limited and Propeller  
827 Health; Research collaboration with Medimmune and Adherium (NZ) Limited; and speaker fees  
828 Circassia Group. The work presented in this paper has been filed by S.S, R.C, M.W, C.E.B, D.S, and P.T  
829 as part of UK Patent Application No. 2110365.0 and International Patent Application No.  
830 PCT/GB2022/051858.

831

### 832 **Data and materials availability**

833 All data associated with this study are available in the main text or the supplementary materials.

834 R and Python codes have been archived in Zenodo (<https://doi.org/10.5281/zenodo.6956451>).

835

### 836 **Figure captions:**

837 **Fig. 1. Study Consort diagram.** Consort diagram outlining the acute study recruitment and number of  
838 analysable GCxGC-MS breath samples.

839

840

841 **Fig. 2. Topological data analysis (TDA) representing the various acute disease groups annotated by blood**  
842 **biomarkers.** Each circle or 'node' in the TDA graph represents a subject or group of subjects. Similar

843 subjects are grouped together in the same node and the relative similarity of the subjects is represented by  
844 the proximity of the nodes. The size of each node is determined by the number of subjects within it. **A:**  
845 Visual mapping of the acute disease groups in the discovery cohort (n=139), based on the discriminatory  
846 805 features and coloured by proportion of acute COPD exacerbations in each node. **B:** The network is  
847 colour coded by the average values of CRP in each node in the discovery cohort (n=139). Higher CRP values  
848 corresponded topologically with the COPD and pneumonia patients. **C:** The network is colour coded by the  
849 average values of BNP in each node in the discovery cohort (n=139). Higher BNP values corresponded  
850 topologically with the heart failure patients. **D:** The network is coloured by proportion of acute COPD  
851 exacerbations in each node in the replication cohort (n=138). In replication cohort, Pneumonia and COPD  
852 exacerbation subjects occupied polar ends of the same TDA network. **E:** The networks are coloured by the  
853 average values of CRP in each node. High CRP values corresponded topologically with the pneumonia  
854 subjects. **F:** The networks are coloured by the average values of BNP in each node. High BNP values  
855 corresponded topologically with the heart failure subjects.

856  
857 **Fig. 3. Diagnostic accuracy of an acute VOC biomarker score.** **A** Scatter plot demonstrating significant  
858 difference between breath VOC biomarker score values in acute cardiorespiratory patients compared to  
859 healthy volunteers. The black horizontal line within the scatter plot represents the median value of the  
860 biomarker score. Mann Whitney test  $*P < 0.0001$ . **B:** Receiver operating characteristic (ROC) curve of  
861 participants in the discovery [black line - AUC 1.00 (1.00-1.00)] and replication [blue line - AUC 0.89 (0.82-  
862 0.95)] cohorts  $P < 0.0001$ . **C:** Histogram showing the number of patients with higher diagnostic uncertainty  
863 (blue bars with values  $>$  upper quartile value of 20 mm). **D:** ROC curve assessing the discriminatory power of  
864 exhaled breath VOCs in participants with higher diagnostic uncertainty. AUC 0.96 (0.92- .99)  $P < 0.0001$

865

866 **Fig. 4. Correlation of VOC biomarker score with blood biomarkers and disease acuity.** **A:** Pearson's  
867 correlation of disease-specific VOC scores and blood-based biomarkers. Pearson correlation  
868 demonstrating the positive and negative correlations between breath VOC scores and blood-based  
869 biomarkers.  $*P < 0.05$ . **B:** Pearson's correlation of disease-specific VOC scores and admission  
870 observations. Pearson correlation between the VOC biomarker score and admission vital signs. VAS:  
871 Visual Analogue Scale (100 mm), participants were asked to rate their breathlessness on a 100 mm  
872 VAS on admission.

873

874 **Fig. 5. VOC biomarker chemical enrichment in acute cardiorespiratory exacerbations.** **A:** Circular  
875 correlation tree generated based on metabolite set enrichment and chemical similarity analysis of 101  
876 breath volatiles associated with acute breathlessness. Branches depict metabolite sets derived using the  
877 ChemRICH; bar graphs portray  $-\log_{10}(p)$  and  $\log_2(\text{fold change})$  values of 101 features extracted using LASSO  
878 regression (table S4) in acute breathlessness compared with control group. The arcs represent the Louvain  
879 clusters, derived from the correlation graph (green for upregulated, red for not significant, blue for  
880 downregulated according to K-S test result). Chemical names are coloured based on their chemical  
881 classification and coloured regions used to summarise broader chemical groups. **B:** Correlation graph  
882 showing metabolite communities identified using Louvain clustering, with the identity and location of the  
883 cluster enriched in heart failure projected onto the circular dendrogram. **C:** i) Example GCxGC  
884 chromatogram showing complex profile of breath metabolites; ii) 3D render of chromatogram showing  
885 visualisation of breath markers; and iii) phenotypic differences based on features included in the breath  
886 biomarker scores (table S9) (yellow, asthma; red, pneumonia; magenta, COPD; cyan, heart failure). Created  
887 in part using the iTOL online <https://itol.embl.de/>.

888

889

890  
891  
892  
893  
894  
895  
896  
897

**Study tables:**

	<b>Total number</b>	<b>Healthy controls</b>	<b>Acute asthma</b>	<b>Acute COPD</b>	<b>Pneumonia</b>	<b>Heart failure</b>	<b>p value</b>
<b>Total number of participants (n=)</b>	<b>277</b>	<b>55</b>	<b>65</b>	<b>58</b>	<b>55</b>	<b>44</b>	
<b>Demographics</b>							
Age *, years	60.8 ± (16.8)	63.05 ± (11.78)	44.3 ± (17.93)	69.82 ± (8.16)	60.67 ± (16.50)	70.72 ± (11.04)	.124
Gender Male (n=) (%)	143 (51%)	26 (47%)	25 (38%)	33 (56%)	27 (49%)	32 (72%)	<b>.008 ¥</b>
Body Mass Index (BMI)* <sup>a</sup>	29.5 ± (7.3)	28.2 ± (4.5)	31.5 ± (9.0)	27.5 ± (7.7)	29.2 ± (6.9)	31.5 ± (6.5)	.767
Smoking Current smoker (n=) (%)	53 (19%)	4 (7%)	13 (20%)	21 (36%)	11 (20%)	4 (9%)	<b>.001 ¥</b>
<b>Vital signs</b>							
Temperature (Celsius)*	36.7 ± (0.6)	36.1 ± (0.4)	36.8 ± (0.5)	36.7 ± (0.5)	37.1 ± (0.7)	36.5 ± (0.3)	<b>.000</b>
Heart rate (beats/min)*	87.2 ± (18.5)	68.1 ± (9.54)	99.6 ± (17.2)	92.9 ± (15.6)	90.3 ± (15.4)	81.3 ± (15.6)	<b>.005</b>
Respiratory rate (breaths/min)*	18.9 ± (4.2)	13.0 ± (1.8)	20.5 ± (3.4)	21 ± (2.5)	20.4 ± (4.6)	19.1 ± (1.8)	<b>.000</b>
Oxygen saturations (%)*	95.8 ± (3.0)	97.7 ± (1.3)	96.1 ± (2.5)	94.0 ± (2.9)	94.5 ± (0.5)	96.5 ± (1.9)	<b>.001</b>
Systolic Blood Pressure (mmHg)*	131.5 ± (19.2)	134 ± (15.7)	133 ± (17.7)	133 ± (20.5)	126 ± (19.4)	128 ± (22.2)	.515
Total mEWS-2 score <sup>^b</sup>	1 (0-3)	0 (0-1)	2 (1-3.5)	3 (1-5)	2 (1-3)	1 (0-2)	<b>.000</b>
<b>Breath sampling</b>							
Time from admission to breath sampling (hours) <sup>^</sup>	16 (3.0–23.0)	1 (1-1)	16 (9.2–22.7)	18 (12.5-23.0)	18 (11.0-23.0)	23 (19.0-26.0)	.000
<b>Symptoms assessment</b>							
Breathlessness VAS score (mm)* <sup>c</sup>	58.1 ± (31.6)	6.2 ± (9.3)	76.6 ± (14.2)	71.6 ± (19.2)	67.8 ± (22.1)	67.9 ± (20.0)	.000**
Cough VAS score (mm)* <sup>c</sup>	43.3 ± (33.2)	8.7 ± (14.3)	64.5 ± (26.7)	57.8 ± (27.0)	53.6 ± (30.6)	24.3 ± (25.2)	.000**
Wheeze VAS score (mm)	41.8 ±	3.4 ±	66.2 ±	60.3 ±	45.1 ±	28.1 ±	.000**

*c	(34.9)	(6.4)	(24.5)	(29.0)	(34.8)	(28.6)	
<b>eMRCd score (n=) (%)</b>							
<b>1</b>	17 (6%)		1 (1.5%)	8 (13%)	7 (12%)	1 (2%)	.000¥
<b>2</b>	6 (2%)		0 (0%)	0 (0%)	5 (9%)	1 (2%)	.000¥
<b>3</b>	15 (5%)		6 (10%)	0 (0%)	7 (12%)	2 (4.5%)	.000¥
<b>4</b>	50 (18%)		16 (25%)	11 (19%)	6 (11%)	17 (38.5%)	.000¥
<b>5a</b>	112 (40%)		38 (51%)	32 (55%)	22 (41%)	20 (46%)	.000¥
<b>5b</b>	21 (7%)		3 (4.5%)	7 (13%)	8 (15%)	3 (7%)	.000¥
<b>Exposure to antibiotics and steroids within 2 weeks of hospital admission</b>							
<b>Antibiotics (n=) (%)</b>	61	n=0 (0%)	n=24 (36.9%)	n=23 (39.6%)	n=10 (18.2%)	n=4 (9.0%)	.002¥
<b>Steroids (n=) (%)</b>	57	n=0 (0%)	n=28 (43.0%)	n=24 (41.3%)	n=3 (5.4%)	n=2 (4.5%)	.000¥
<b>Morbidity and mortality measures</b>							
<b>Length of hospital stay (days) ^</b>	3 (2-6)		2.0 (1.0-3.0)	4.0 (2.0-6.0)	4.0 (2.0-5.0)	7.0 (4.0-11)	.000**
<b>30-60 days hospital readmission (n=)</b>	29		7	9	6	7	.461¥
<b>1-year all-cause mortality</b>	12	0	1	5	1	5	.078¥
<b>Laboratory parameters</b>							
<b>C-reactive protein (CRP) (mg/L)^</b>	11 (5.0-34.2)	5 (5-5)	10.0 (5.0-23.0)	12.0 (5.0-20.7)	108.0 (53.5-245.3)	11.0 (5.0-22.0)	.000**
<b>Blood Eosinophil count 109/L^</b>	0.13 (0.06-0.24)	0.17 (0.09-0.24)	0.18 (0.06-0.42)	0.13 (0.06-0.24)	0.08 (0.04-0.14)	0.13 (0.08-0.23)	.000**
<b>Troponin T (ng/l)^</b>	3.3 (1.0-11.4)	2.05 (1.0-2.7)	1.55 (1.0-3.4)	3.75 (2.6-10.9)	4.3 (2.18-11.3)	20.2 (13.4-59.6)	.000**
<b>Brain natriuretic peptide (BNP) (ng/l)^</b>	40.5 (20.6-98.9)	28.40 (17.60-39.88)	20.4 (12.1-40.0)	56.3 (24.3-95.0)	56.3 (27.4-132.1)	611.8 (172.1-1259.1)	.000**
<b>Questionnaires</b>							
<b>Asthma Quality of Life Questionnaire (AQLQ) total*</b>	65		117.3 ± (37.3)				
<b>COPD Assessment test (CAT) *</b>	58			26.7 ± (7.3)			
<b>COPD Decaf score *</b>	58			1.7 ± (0.8)			
<b>CURB65 score^</b>	55				2 (1-3)		
<b>NYHA score^</b>	44					2 (1-3)	

898

899 **Table 1. Demographics and clinical characteristics of study participants.** Continuous variables are  
900 presented as mean ± standard deviation. Categorical variables are presented as numbers (%).

901 <sup>a</sup> The body mass index (BMI) is the weight in kilograms divided by the square of the height in meters.



902 <sup>b</sup> Modified Early warning score - 2 (MEWS-2) is a guide widely used by medical services to determine the  
903 degree of illness of a patient based on their vital signs including respiratory rate, oxygen saturations,  
904 temperature, blood pressure, and heart rate. Vital signs collected at the point of admission for acute  
905 disease groups.

906 <sup>c</sup> Participants were asked to determine their degree of breathlessness, cough and wheeze on a 100mm  
907 visual analogue scale (VAS) on admission. Higher scores indicate worse symptoms.

908 <sup>d</sup> Extended Medical research Council (eMRC) scale is a validated measure of perceived respiratory disability,  
909 scored from 1 to 5b. Higher scores indicate worse disability.

910 \* Data is expressed as mean (SD) or n (%) ± (SD), ^ Data expressed as median (IQ range), \*\* Kruskal-Wallis  
911 test comparing non-parametric data, † Pearson Chi Squared and Fisher's Exact test.

912 ANOVA was used to assess the differences between groups for normally distributed continuous  
913 variables and Kruskal-Wallis for non-parametric continuous variables. Pearson chi-squared and  
914 Fisher's exact were used to assess the differences in categorical variables. The results were  
915 considered statistically significant at *p*-values <0.05.