



Methodological considerations for large-scale breath analysis studies

DOI:

[10.1088/1752-7163/aae557](https://doi.org/10.1088/1752-7163/aae557)

Document Version

Accepted author manuscript

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):

the U-BIOPRED study group (2018). Methodological considerations for large-scale breath analysis studies: lessons from the U-BIOPRED severe asthma project. *Journal of Breath Research*, 13(1).
<https://doi.org/10.1088/1752-7163/aae557>

Published in:

Journal of Breath Research

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.



Methodological considerations for large-scale breath analysis studies: lessons from the U-BIOPRED severe asthma project.

Waqar M. Ahmed¹, Paul Brinkman², Hans Weda³, Hugo H. Knobel⁴, Yun Xu⁵, Tamara M. Nijsen³, Royston Goodacre⁵, Nicholas Rattray⁶, Teunis J. Vink³, Marco Santonico⁷, Giorgio Pennazza⁷, Paolo Montuschi⁸, Peter J. Sterk², Stephen J. Fowler^{1,9*}, on behalf of the U-BIOPRED study group[#]

Author affiliations:

¹. School of Biological Sciences, University of Manchester, United Kingdom.

². Department of Respiratory Medicine, Academic Medical Center (AMC), University of Amsterdam, The Netherlands.

³. Philips Research, Royal Philips B.V., The Netherlands.

⁴. Philips Innovation Laboratories, Philips Lighting B.V., The Netherlands.

⁵. School of Chemistry, Manchester Institute of Biotechnology, University of Manchester, United Kingdom.

⁶. Department of Environmental Health Sciences, Yale School of Public Health, Yale University, New Haven, United States.

⁷. Unit of Electronics for Sensor Systems, Faculty of Engineering, Campus Bio-Medico University of Rome, Rome, Italy.

⁸. Department of Pharmacology, Faculty of Medicine, Catholic University of the Sacred Heart, Rome, Italy.

⁹. Manchester Academic Health Sciences Centre, Manchester University NHS Foundation Trust, United Kingdom.

*Corresponding author

Email: stephen.fowler@manchester.ac.uk

[#]Membership of the U-BIOPRED study group is listed in the supplementary information

Abstract

Methods for breath sampling and analysis require robust quality assessment to minimise the risk of false discoveries. Planning large-scale multi-site breath metabolite profiling studies also requires careful consideration of systematic and random variation as a result of sampling and analysis techniques. In this study we use breath sample data from the recent U-BIOPRED cohort to evaluate and discuss some important methodological considerations such as batch variation and correction, variation between sites, storage and transportation, as well as inter-instrument analytical differences. Based on this we provide a summary of recommended best practices for new large scale multi-site studies.

Keywords

Multi-site studies, breath analysis, metabolomics, quality assurance, quality control, U-BIOPRED

1. Introduction

The deployment of 'omics technologies in clinical research is proving invaluable in unravelling complex biological mechanisms and for the discovery of important markers for inflammatory lung disease[1]. However, for studies to produce clinically meaningful results, a large number of samples are typically required to provide statistical power and hence reduce the risk of false discoveries[2]. Robust sampling and analysis methods are equally important, especially where additional sites and instruments will be prone to increased systematic variation.

In metabolomic studies, a standardised and well-controlled sampling and analysis approach must be adopted throughout the whole experimental pipeline [3–6]. The use of metabolomic workflows in breath research is challenging as breath samples typically contain high intra- and inter-sample variation, a number of sampling variables (such as diet and medical history), and the presence of

1
2
3 exogenous artefact or contaminant compounds. As a result, and because of limited long term
4 storage options, it is usually not possible to form pooled quality control samples of breath to
5 evaluate analytical variation[7]. In addition, breath sample analysis can result in a high number of
6 features, often exceeding the number of samples. This dominant effect, which may be specific to a
7 single site or instrument, can influence downstream data analysis[8]. It is therefore important to
8 consider additional quality assessment procedures in breath research where large samples sizes are
9 used. Studies with multiple sites can introduce new variation from unavoidable inconsistent
10 sampling techniques, sample storage and transportation, and lack of control samples.
11 Methodological considerations have previously been reported for the collection and analysis of
12 breath samples, and international task forces have recommended breath sampling and analysis
13 standardisation[9–13]. However, specific guidance for large scale multi-site studies remains limited.

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
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

The aim of this investigation is to address methodological considerations for large scale multi-site studies that may otherwise be overlooked throughout collection and analysis of breath samples. To show examples of methodological considerations, we will use data from the U-BIOPRED (Unbiased BIOMarkers in PREDiction of respiratory disease outcomes) severe asthma cohort study, and consider the following objectives:

- To identify previously-reported disease-specific and artefact VOCs;
- To use the extracted VOCs to assess variation between sites and patients;
- To assess the effect of variation from storing breath samples in sorbent-packed tubes;
- To evaluate sampling and analytical inter-batch variation within multivariate data and subsequently compare batch correction methods;
- Compare multivariate datasets from different instruments using Procrustes analysis.

2. Materials and Methods

2.1. Study information

The U-BIOPRED study is a large scale European multi-site study with the objective to reveal novel phenotypes and therapeutic targets in severe asthma[14,15]. Asthma subtype was categorised as either mild/moderate or severe, as defined by the Innovative Medicines Initiative (IMI)[16]. Several 'omics technologies were employed in a systems biology approach using samples including blood, urine, sputum supernatants, and breath from adult and paediatric (school-aged and pre-school children) patient cohorts. Extensive clinical data and patient-reported outcomes were also collected. In this work we will focus on data from the breath volatilomics part of the study.

2.2. Breath sampling and analysis

Samples were collected and analysed between 2012 and 2015. After 5 min breathing room air through a VOC filter (A2, North Safety, Middelburg, Netherlands), participants were asked to breathe a single vital capacity into a ten litre Tedlar® bag (SKC Inc, Eighty Four, PA, United States) via a three way valve[17]. Within 10 min, the mixed expiratory breath sample was sequentially purged onto two sorbent tubes containing Tenax GR (stainless steel, 6 mm x 177.8 mm, Gerstel, Mülheim an der Ruhr, Germany) by using a custom-built peristaltic pump (medical grade tubing, bore 4.8 mm, wall thickness 2.4 mm, Williamson Pumps, Brighton, UK) at a flow rate of 250 mL/min. A diagram of this layout is shown in the supplementary information (Figure S1). After local sampling, the tubes were tightened with brass caps, sealed in an air-tight packaging, and transported by air or land to the Academic Medical Centre, Amsterdam, for central analysis and distribution. Sorbent tubes were stored in refrigerators (2 to 8 °C) prior to analysis.

The first sample was used for eNose analysis. VOCs were released from the sorbent tube using a thermal desorption oven (TDS 3, Gerstel, Mülheim an der Ruhr, Germany), after which the sample was transferred into a Tedlar bag with nitrogen as carrier gas. Subsequent analysis was carried out by a composite eNose platform. The sample handing method has been described previously[18]. The

1
2
3 eNose platform consisted of four eNoses from four different brands, using distinct sensor
4 technologies: 1) Cyranose C320 using carbon black-polymer sensors[19] (two instruments were
5 included in this study), 2) Tor Vergata eNose using quartz crystal microbalances (QMB) covered with
6 metalloporphyrins[20], 3) Common Invent eNose using metal oxide semiconductor sensors[21], and
7
8 4) Owlstone Lonestar based on field asymmetric ion mobility spectrometry[22]. A combined total of
9
10 158 sensor response signals were used in the analysis (Comon Invent = 8, Cyranose = 32, Lonestar =
11
12 110, Tor Vegata = 8). The Lonestar was setup to scan between 0 and 100 % dispersion field
13
14 (magnitude of electric field) in 51 steps and a compensation voltage between -6 and +6 Volt in 512
15
16 steps.
17
18
19
20
21
22

23
24 The second sorbent tube was sent to Philips Research in Eindhoven, The Netherlands, in order to be
25 analysed centrally by thermal desorption-gas chromatography-mass spectrometry (TD-GC-MS), using
26 a similar method as previously described[23]. In brief, samples underwent automated TD (Gerstel,
27
28 Mülheim an der Ruhr, Germany) using helium as the carrier gas. The sample was then purged onto a
29
30 packed liner, heated to 300 °C for 3 min, and subsequently transferred to a Tenax TA cold trap (kept
31
32 at -150 °C) for 2 min. The cold trap was then heated to 280°C at 20°C/s, after which the sample was
33
34 injected into the GC column (VF1-ms, 30 m × 0.25 mm, 1 µm, 100% dimethylpolysiloxane, Varian
35
36 Chrompack, Middelburg, The Netherlands) using a splitless injection method with helium as a carrier
37
38 gas at a flow rate of 1.2 mL/min. The GC oven (7890 N GC, Agilent, Santa Clara, CA, USA) was set to a
39
40 ramped temperature programme with the following parameters: 40 °C (5 min hold), ramp to 300 °C
41
42 at 10 °C/min (5 min hold). Compounds were then transferred to a MS with a Time-of-Flight mass
43
44 analyser (LECO Pegasus 4D, LECO, Sint Joseph, MI, USA) in electron ionization positive mode (70eV),
45
46 with a mass spectral acquisition range between 29 to 450 Da.
47
48
49
50
51
52

53 *2.3. Data pre-processing and treatment*

54
55 Raw GC-MS files were converted to netCDF files, and uploaded to an open source knowledge
56
57 management platform (transSMART) for quality checks and to ensure data transparency between
58
59
60

1
2
3 project partners[24]. Both pre-processing and statistical analyses were performed in the R software
4 environment (R Development Core Team 2017, version 3.4.2) using relevant R packages (*XCMS*,
5 *BatchCorrMetabolomics*, and *Vegan*) described in detail previously [25–27]. GC-MS samples were
6 pre-processed to create an ion fragment data matrix for further analysis and extraction of known
7 VOCs.
8
9

10
11
12
13
14
15 Several putative asthma-related VOCs (and other commonly reported VOCs such as C₅ to C₁₅ alkanes,
16 acetone, isoprene, toluene, and internal standards) [28–33] were targeted for further assessment,
17 based on predicted retention indices and mass fragments (extracted from Pubchem and NIST online
18 chemical databases). Retention indices were predicted using known retention times of alkanes[34].
19 These alkanes were assigned a retention index number using their carbon number (e.g. 800 for C₈, or
20 900 for C₉) and subsequently used as anchors for cubic spline interpolation of unknown VOCs, for
21 example, if a VOC had a retention time between octane and nonane, its retention index would fall
22 between 800 and 900. In addition, a retention index threshold of ± 20 was applied.
23
24
25
26
27
28
29
30
31
32

33
34 The identity of compounds that met these search criteria were then confirmed by pre-processing
35 chromatograms using AMDIS (version 2.72) and subsequently NIST14 library search. This meant the
36 identifications conformed to MSI level 2[35]. It is important to note that the Tedlar bag method used
37 in this study may not be adopted in other multi-site studies where other sampling devices may be
38 preferred[36]. Both known and unknown VOCs were therefore evaluated as shown in the fault tree
39 analysis in Figure 1, except for calculating LODs and linear ranges for identified compounds.
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

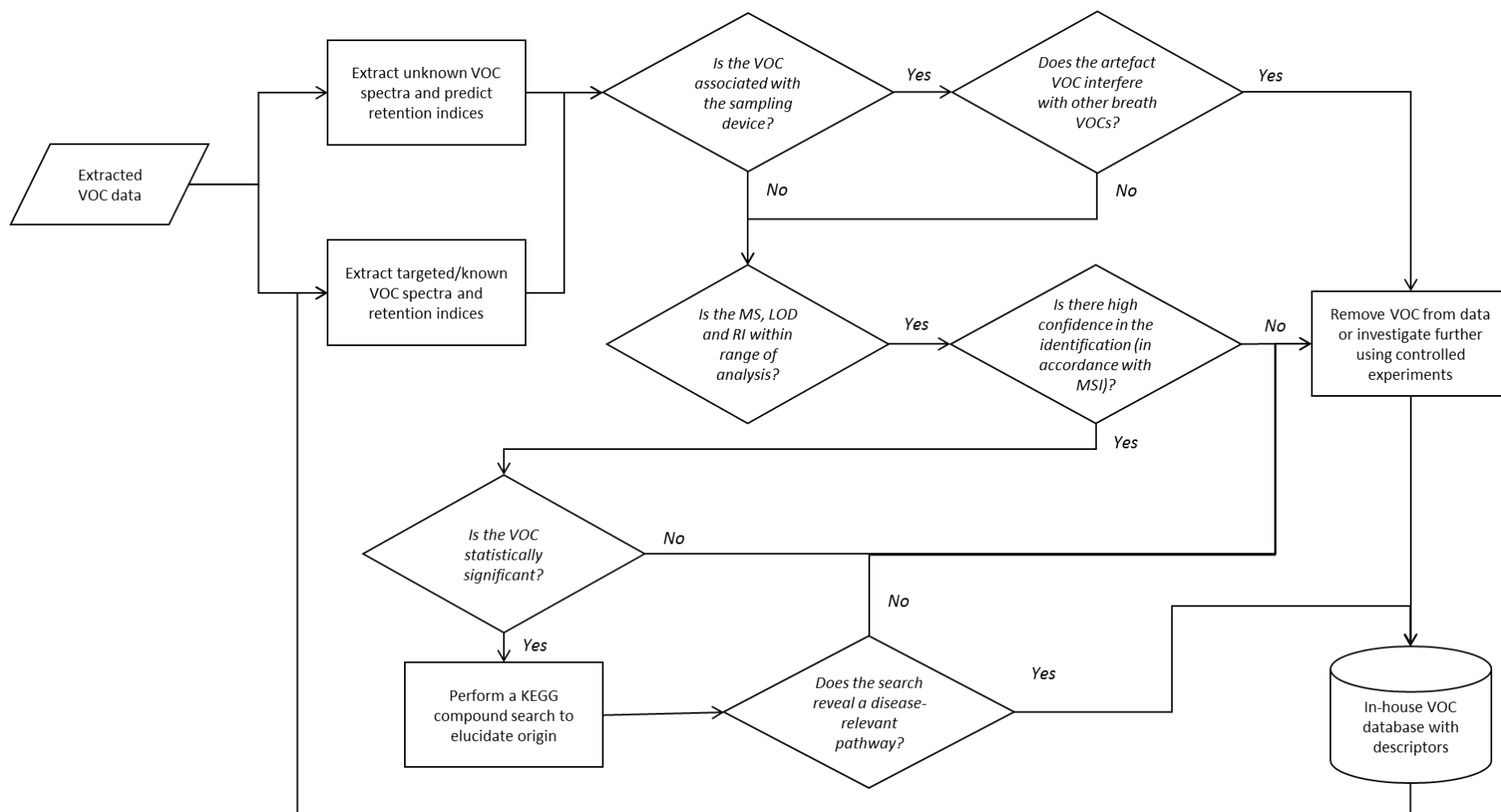


Figure 1. A fault tree analysis of breath sample VOC origin and artefact determination

1
2
3 Data from e-nose instruments were auto scaled ($mean = 0$, $standard\ deviation = 1$) and the ComBat
4 algorithm was used to correct for systematic drift. MS data were normalised by the internal standard
5 toluene-d8, and features were range scaled. Missing values (i.e. zero values) within the GC-MS data
6 matrix were imputed using random forest proximity[37,38] for multivariate analyses and samples
7 with more than 70 % missing values removed.
8
9

15 2.4 Univariate and multivariate analysis

16
17
18 To assess the significance of variation of VOCs between groups, non-parametric Kruskal-Wallis test
19 was used, and if statistically significant, a post-hoc test was performed consisting of a Mann–
20 Whitney U test between two groups with Bonferroni correction. With regard to assessing how
21 breath sample VOCs respond to storage duration, any samples that would be influenced by patient
22 groups or site variation were removed from analysis. Therefore, breath samples from adults with
23 severe asthma from the Amsterdam site were used. These samples had a maximum storage time of
24 39 days and were randomised within each storage batch. Linear regression was used to determine if
25 known exogenous and endogenous VOCs were correlated with duration of sample storage. To
26 evaluate the effectiveness of common batch correction methods, a distance matrix calculated using
27 Bhattacharyya distance measure was used in principal component analysis (PCA), retaining the first
28 three PCs based on scree plot analysis, where the resulting inter-batch metric was the mean
29 Bhattacharyya distance (0 equal to no batch effect). The metric was adopted from and described in
30 more detail by Wehrens *et al.*[27].
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

47 To measure and compare dissimilarity between instruments (i.e. pairwise analysis of GC-MS *versus*
48 e-nose, and e-nose *versus* e-nose), a Procrustes test with principal coordinates analysis (PCoA) was
49 performed. In summary, paired samples were aligned for each feature matrix (i.e. per instrument),
50 and subsequently converted to a distance matrix using the Jaccard measure, as described in previous
51 studies[39,40]. PCoA was then performed on the distance matrices retaining the first 3 PCs, which
52 explained the most variance based on scree plot analysis. As it is important to correct for the
53
54
55
56
57
58
59
60

1
2
3 occurrence of any negative eigenvalues when using the Jaccard measure (non-Euclidean) with PCoA
4 (Euclidean)[39], we used Lingoes' correction to convert negative eigenvalues to non-negative where
5
6 a constant value double the value of negative eigenvalues, and added to all eigenvalues resulting in a
7
8 corrected non-negative eigenvector[41].
9

10
11
12 Pairwise congruence between two instruments for the same patient sample was assessed using
13
14 Procrustes analysis[42,43]. Briefly, an input matrix is geometrically transformed (i.e. rotating and
15
16 stretching/shrinking) to find the optimal superimposition on a target matrix, and such that the sum
17
18 of squared distances (m^2) is minimised and normalised (where 1 = data are completely different, and
19
20 0 = data are exactly the same). Subsequently the correlation r between the two data matrices after
21
22 Procrustes transformation is derived from m^2 ($r = \sqrt{1 - m^2}$). In order to assess the statistical
23
24 significance of the similarity measured by the single Procrustes superimposition (non-symmetric), a
25
26 Procrustean test with 1000 permutations is performed and an empirical p -value was derived by
27
28 counting the number of cases where r of permuted data had been higher than the one calculated
29
30 from the original data and divided by the number of permutations performed (i.e. 1000). Detailed
31
32 explanations and considerations of the Procrustean test applied to MS data have been published
33
34 [42,43].
35
36
37
38

39 40 **3. Results**

41 42 *3.1. Sample and metadata description*

43
44
45 A total of 298 breath samples were collected in the study, of which 164 were first visit (baseline)
46
47 samples and used in this work, each sample comprising breath from a single patient, thereby
48
49 reducing any confounding effect from longitudinal samples. Baseline samples included adults with
50
51 mild/moderate ($n = 11$), and severe ($n = 42$) asthma; school aged children with mild/moderate ($n =$
52
53 15), and severe ($n = 37$) asthma; and pre-school aged children with mild/moderate ($n = 21$), and
54
55 severe ($n = 32$) wheeze. Clinical characteristics and definitions of asthma severity from within the U-
56
57 BIOPRED consortium are described elsewhere[14–16]. The remaining samples were from healthy
58
59
60

1
2
3 subjects ($n = 6$). Breath samples were collected at the following sites: Amsterdam ($n = 78$),
4
5 Copenhagen ($n = 19$), London, ($n = 47$), Manchester ($n = 12$), and Southampton ($n = 8$). With the
6
7 exception of Copenhagen and Southampton sites, which comprised breath samples solely from
8
9 children and adults (respectively), all sites provided a mixture of age and asthma subtype groups.
10
11

12 13 *3.2. Determination of VOCs from breath samples*

14
15 To assess the variation of VOCs using sample groups (i.e. by age, site, asthma severity), we first
16
17 identified known breath VOCs (acetone, isoprene, and toluene), C5-15 alkanes, and putative asthma
18
19 VOCs compiled from selected studies using TD-GC-MS profiling techniques. A final target list of 53
20
21 asthma VOCs was used to search the U-BIOPRED ion fragment matrix. From this search, 42 VOCs
22
23 were extracted, of which 16 had identification confirmed (see Table 1 and Figure S2). Confirmed
24
25 VOCs were classed as aldehydes ($n = 2$), ketone ($n = 1$), sulphide ($n = 1$), furan ($n = 2$), and
26
27 hydrocarbons ($n = 10$), of which seven were branched hydrocarbons.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 1. A list of VOCs putatively identified (MSI level 2) as biomarkers of asthma and extracted from U-BIOPRED breath sample GC-MS data. Also shown are their average molecular weight, base peak mass fragment, the predicted retention index, the KEGG description of that VOC, and the reference from which they were identified.

VOC	Molecular Weight (g/mol)	Extracted MS fragments	Predicted RI	KEGG description	Reference
carbon disulfide	76.13	44, 76	537	Non-specific	[30]
2-butanone	72.11	43, 57, 72	577	Non-specific	[32]
2-methylpentane	86.18	57, 86	584	Lipid peroxidation	[29]
2-methylfuran	82.10	39, 53	588	Carcinogen	[28]
3-methylfuran	82.10	39, 53, 82	608	Carcinogen	[28]
benzene	78.11	51, 78	660	Xenobiotic	[31,33]
2-methylhexane	100.20	85, 100	677	Lipid peroxidation	[29]
octane	114.23	43, 57, 114	800	Lipid peroxidation	[29]
p-xylene	106.17	91, 106	872	Non-specific	[30,31]
cumene	120.20	105, 120	928	Carcinogen	[30]
octanal	128.22	128	978	Lipid peroxidation	[28]
2-methyldecane	156.31	43, 57, 128, 156	1061	Lipid peroxidation	[32]
nonanal	142.24	57	1084	Lipid peroxidation	[28]
dodecane	170.34	170	1200	Lipid peroxidation	[33]
2,6,11-trimethyldodecane	212.42	57, 71, 212	1275	Lipid peroxidation	[32]
2,6,10-trimethyldodecane	212.42	57, 71	1332	Lipid peroxidation	[29,32]

Known analytical artefacts including polydimethylsiloxanes, phenol, and *N,N*-dimethylacetamide (DMAC) were identified. Figure 2 shows a combined mass spectrum highlighting sampling and instrument artefacts using the mean of sample intensities. Reproducible and stable artefact peaks have been used to calibrate an instrument for qualitative analysis, as previously shown[44].

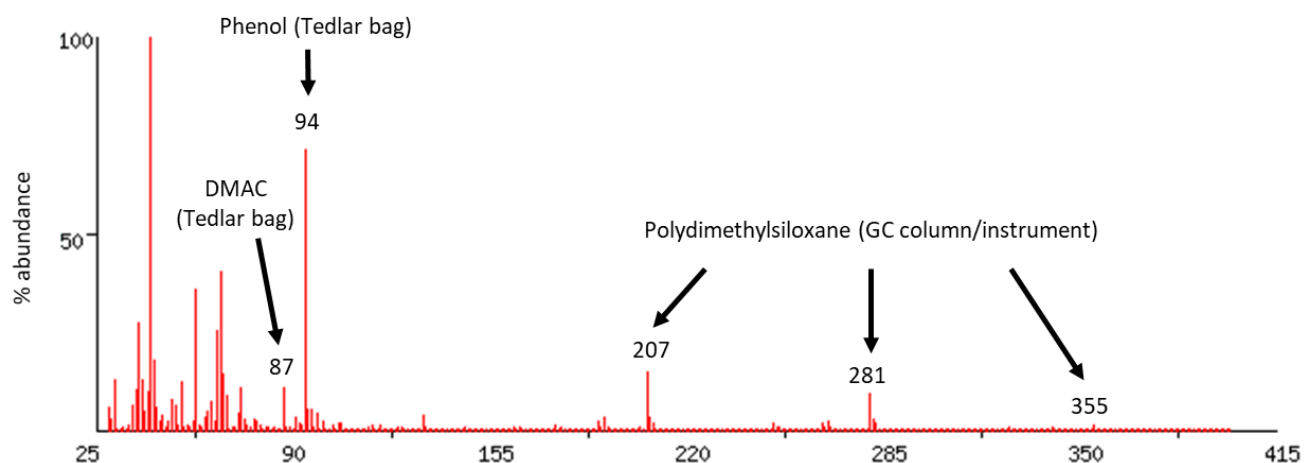


Figure 2. A combined mass spectrum of all fragments within the MS acquisition range (m/z 29-400) and their mean percentage abundance, normalised to the most abundant peak. Highlighted are Tedlar bag sampling artefacts (phenol, and DMAC), and instrument artefacts (polydimethylsiloxane).

3.3. Variation from patient factors and between sites

Out of the previously identified VOCs, we found increased levels for dodecane (Figure 3a) for the London site (adults = 4, school =16, pre-school = 20), in comparison to the Amsterdam (adults = 14, school =11, pre-school = 10) and Manchester (adults = 8, school =3, pre-school = 1) sites (Kruskal-Wallis $p < 0.001$). Similarly, pentane (shown in Figure 3b) has shown difference (Kruskal-Wallis $p = 0.029$) in intensity between London and Copenhagen sites (Mann–Whitney U, Bonferroni corrected $p = 0.027$) within the pre-school cohort only.

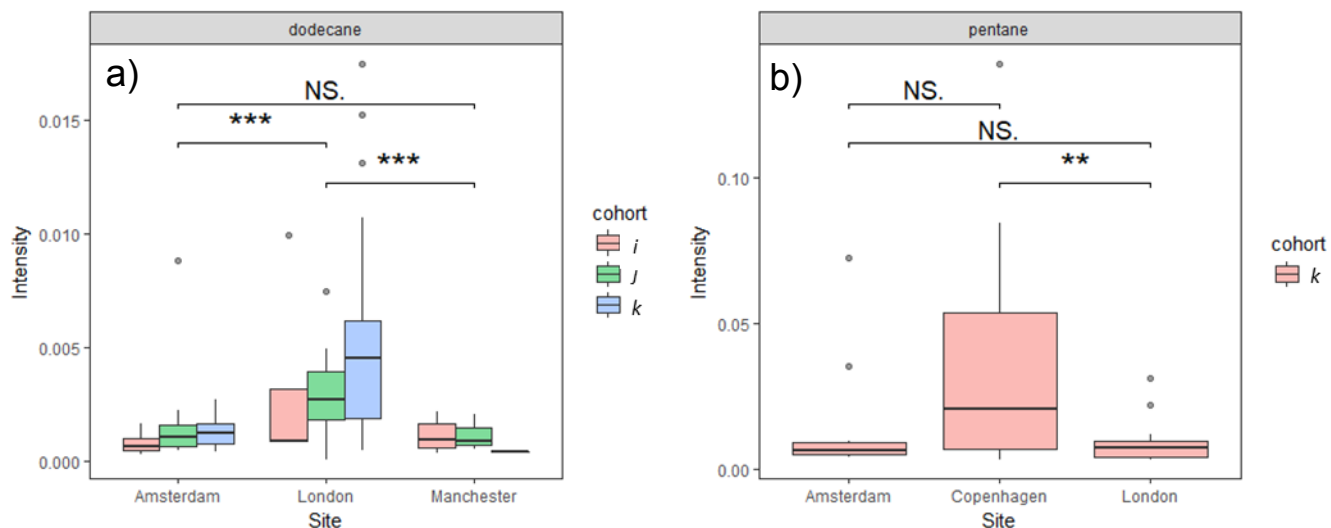


Figure 3. Relative mass fragment intensities in patient breath samples for (a) dodecane across three sites including Amsterdam ($n = 35$), London ($n = 40$), and Manchester ($n = 12$), and (b) pentane across three sites including Amsterdam ($n = 10$), Copenhagen ($n = 16$), and London ($n = 20$). Significance between these groups are highlighted by asterisks or not significant (NS) after a Kruskal-Wallis test followed by a Mann-Whitney-U test with Bonferroni correction. Also shown for dodecane (a) are patient age groups (where i = adults, j = school children, and k = pre-school children) within each site.

Using the adult cohort only (from all sites), we found octanal to be increased (Mann-Whitney U $p = 0.048$) for patients with mild-to-moderate asthma ($n = 11$) when compared to patients with severe asthma ($n = 22$), as shown in Figure 4.

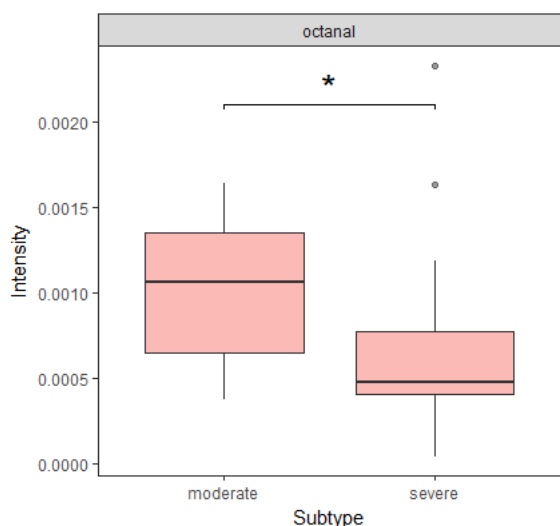


Figure 4. Relative mass fragment intensity showing increased octanal in patient breath samples for moderate ($n = 11$) compared to severe ($n = 22$) asthma subtypes, with a significance of $p = 0.049$ after a Mann-Whitney-U test.

3.4. Variation from sample storage

After analysis of storage variation, we found no significant variation for breath samples stored refrigerated for up to 39 days (Kruskal-Wallis $p = 0.514$). Several known VOCs were selected, and their intensity shift analysed against sample storage duration. These included several alkanes including hexane, octane, decane, dodecane, and tetradecane; common exogenous VOCs benzene and toluene, and common endogenous VOCs acetone and isoprene. Figure 5 illustrates variation of the VOCs described against sample storage duration.

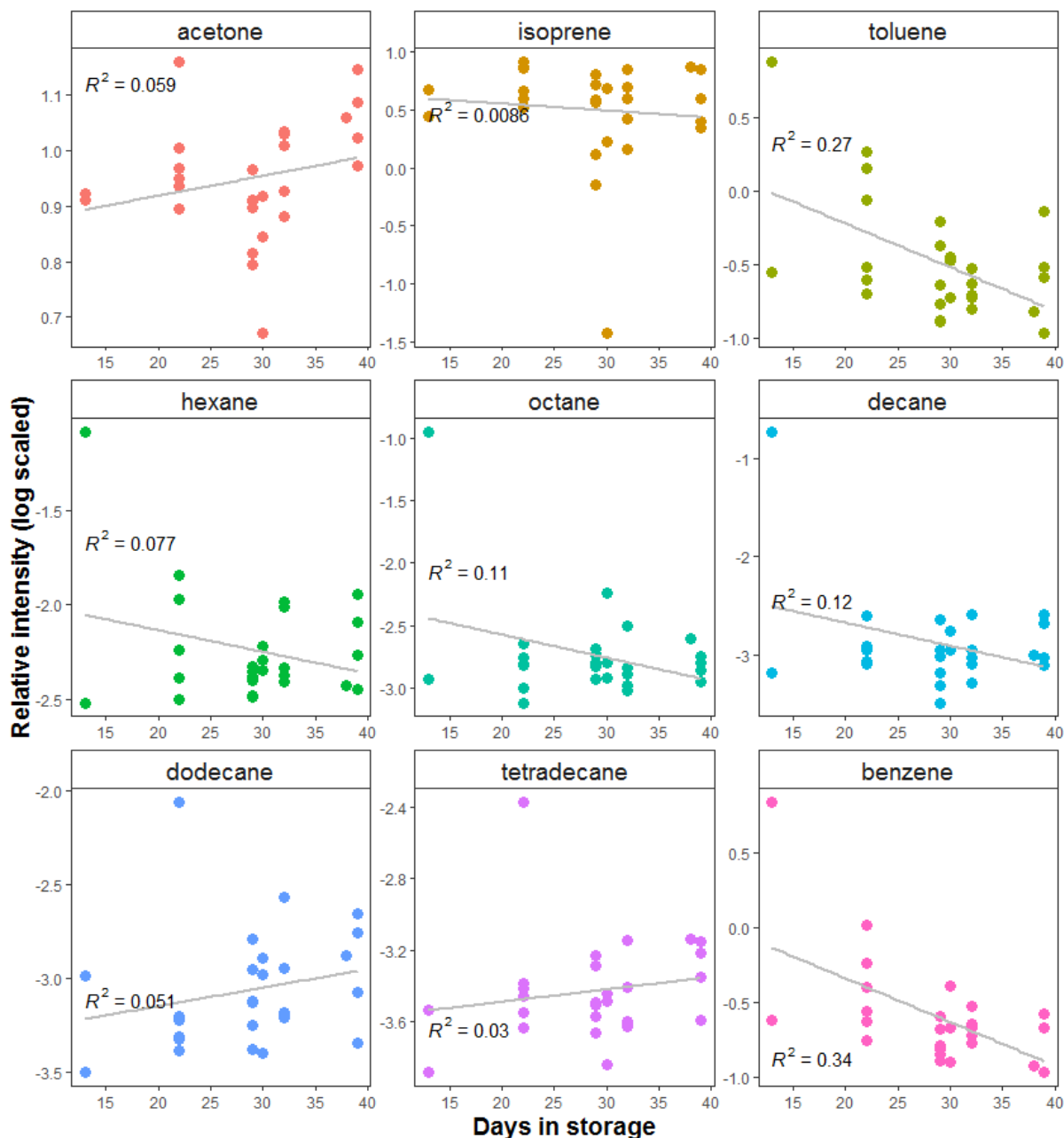


Figure 5. Selected VOCs (normalised and log scaled) for samples stored for up to 39 days (mean of 28 days), annotated with R^2 derived from a linear fit. Samples were limited to severe asthma adult patients from the Amsterdam site.

3.5. Assessment of multivariate datasets: Comparison of batch correction methods

To compare correction methods of these batches, we adopted a method described by Wehrens *et al.* [27]. In the U-BIOPRED study, breath was sampled in uncontrolled batches, where one batch may be associated to one site within a patient visit period. Samples were then analysed in controlled batches, and each sample was spiked with a four-component deuterated internal standard mixture. Mean inter-batch distances (IBD) are shown in Table 2.

Table 2. Comparison of selected normalisation methods using the mean Bhattacharyya distance within sampling and analytical batches, where low inter-batch distance indicates larger point cloud overlap between batches, which in turn means an overall low batch effect.

Normalisation method	Analytical batch distance	Sample batch distance	Combined average
No normalisation	0.75	0.69	0.72
toluene- d_8	0.43	0.54	0.49
acetone- d_6	2.52	2.91	2.72
Sum	0.79	0.63	0.71
Mean	0.79	0.63	0.71
Median	0.78	0.75	0.77
Sum of squares	0.87	0.90	0.89

When compared to untransformed data (combined IBD = 0.72), we show that using an internal standard (toluene- d_8) improved batch correction (0.49) more than using scaling factors (between 0.71 and 0.89). In our case, scaling factors such as normalisation by sample sum or sample mean (0.71) performed better than normalisation by sample median (0.77) or sample variation (0.89). The combined IBD for acetone- d_6 was 2.72. Figure 6 illustrates the sampling batch distance for data

normalised by acetone- d_6 (Figure 6a), and toluene- d_8 (Figure 6b), where less distance between eclipses is visible for the latter meaning less variation between batches.

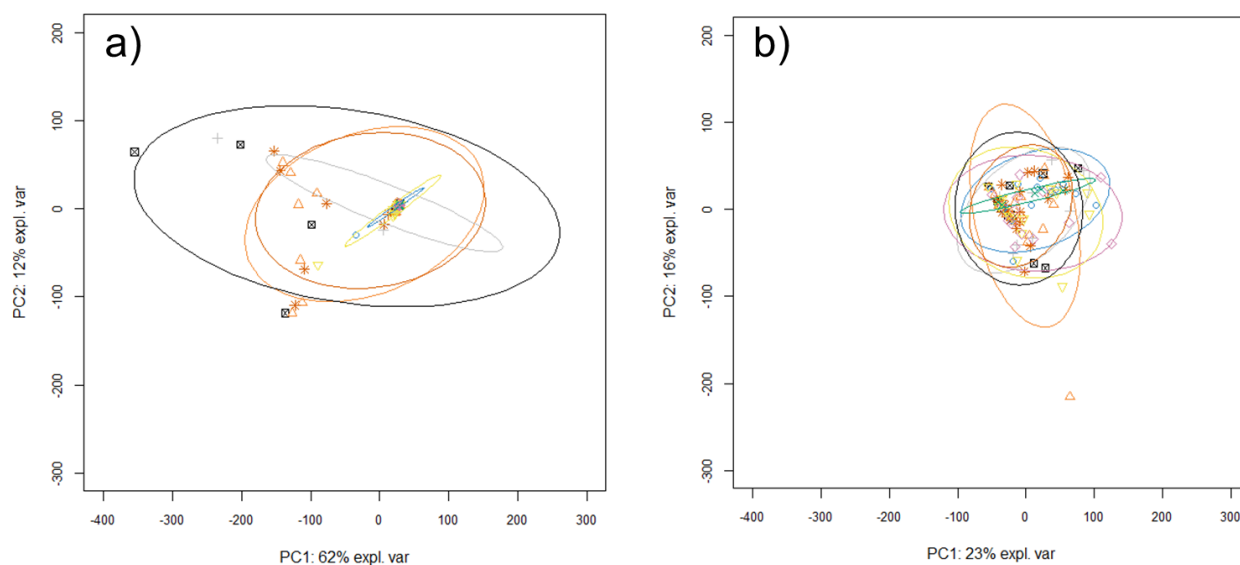


Figure 6. Visual representations of GC-MS data normalised by internal standards a) acetone- d_6 , and b) toluene- d_8 , where batches are represented with the same colour and shape. The mean inter-batch Bhattacharyya distance is shown for each scores plot.

3.6. Assessment of multivariate datasets: Comparison between instruments

After central analysis by TD-GC-MS and an e-nose platform, Procrustes analysis was performed. Table 3 shows a pairwise matrix of correlations between instruments (Procrustes error m^2 and 95% confidence intervals shown in Table S3).

Table 3. Pairwise matrix of similarity correlations derived from the Procrustes test, between GC-MS and e-nose instruments.

	GC-ToF-MS	Lonestar	Cyranose	Tor Vergata	Comon Invent
GC-ToF-MS	–				
Lonestar	R 0.252 (p = 0.132)	–			
Cyranose	R 0.112 (p = 0.878)	R 0.183 (p = 0.204)	–		
Tor Vergata	R 0.211 (p = 0.255)	R 0.189 (p = 0.397)	R 0.745 (p = 0.001)	–	
Comon Invent	R 0.173 (p = 0.525)	R 0.355 (p = 0.003)	R 0.392 (p = 0.003)	R 0.216 (p = 0.184)	–

Significant values ($\alpha = 0.05$) are highlighted in bold

No significant similarities were found between GC-MS and e-nose data. For e-nose comparisons, Cyranose and Tor Vergata instruments showed correlation ($R = 0.745$, $p = 0.001$). Less strong correlations were found between Lonestar and Comon Invent ($R = 0.355$, $p = 0.003$), and between Cyranose and Comon Invent ($R = 0.392$, $p = 0.003$). To illustrate this Procrustes transformation, Figure 7 shows an example of a Procrustes superimposition plot for dissimilarity – between GC-MS and Lonestar instruments, and similarity – between Cyranose and Tor Vergata instruments.

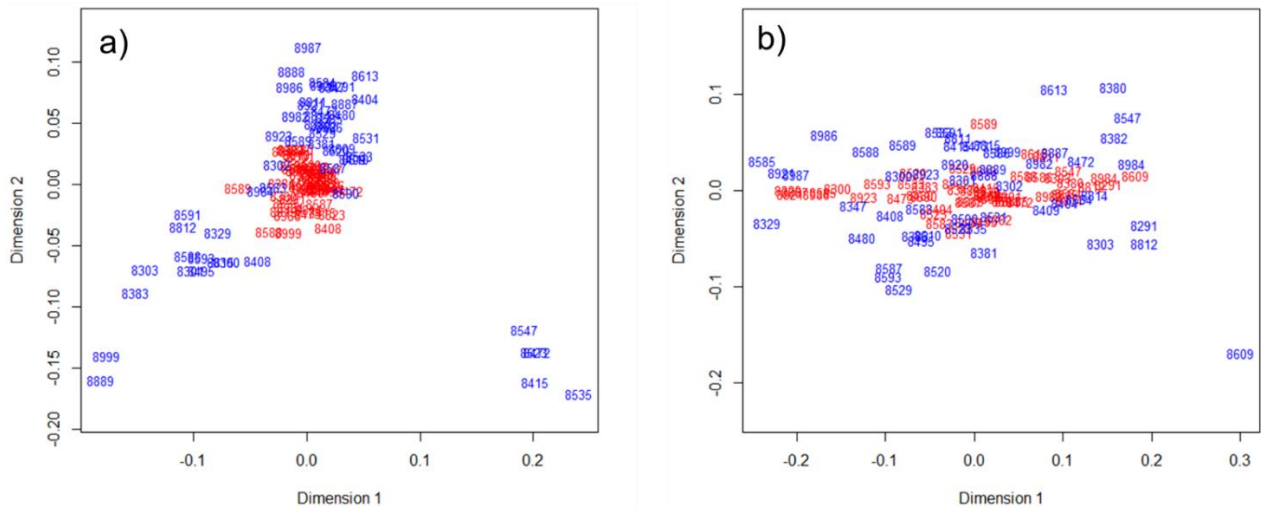


Figure 7. Procrustes superimposition plots using distance matrices of aligned samples (paired numbers represent patient identification) where a) Lonestar data (red) transformed onto GC-MS data (blue), and b) Cyranose e-nose data (red) transformed onto Tor Vergata e-nose data.

4. Discussion

4.1. Summary of findings

In this study, we have assessed breath sample data from the recent U-BIOPRED severe asthma cohort study, and have shown possible sources of variation which must be considered when planning a large scale or multi-site study.

To achieve this, we used a targeted approach, where VOCs have been described in literature and linked to asthma, and used the KEGG database to search for their metabolic origins. It is important to stress that KEGG is curated by experts and links several chemical, biological, and pathway databases however, information of breath metabolites is limited as there are few databases that provide this information. With regard to VOCs found within U-BIOPRED breath data, many hydrocarbons may have an exogenous source and not be linked to internal metabolism, as are many VOCs found in breath samples. A KEGG search for benzene and *p*-xylene revealed these VOCs as

1
2
3 xenobiotic. Furan based cyclic compounds such as 2- and 3-methylfuran and cumene were described
4
5 in KEGG as carcinogenic. These previously identified compounds may not be directly related to
6
7 asthma but may constitute epiphenomena such as differing exposures or dietary influence. It is also
8
9 possible that furans may be derived from aldehydes (in similarity to the Paal-Knorr furan synthesis
10
11 reaction) where multiple carbonyl functional groups are reduced to form an ether functional group.
12
13 Carbon disulfide, *p*-xylene, and 2-butanone are non-specific as they have multiple origins other than
14
15 host inflammation, such as products of microbial metabolism[45], and the human airway contains a
16
17 complex community of bacteria whose metabolic outputs will also be measured in breath sampling.
18
19

20
21 Data from both KEGG and the selected studies suggest that alkanes (both saturated and
22
23 unsaturated) are associated with lipid peroxidation, a hallmark of host inflammation, where volatile
24
25 alkanes would be breakdown products of fatty acids (FAs). In addition, volatile aldehydes or ketones
26
27 are also linked to lipid peroxidation, as carbonyl group breakdown products of FAs. Patients with
28
29 severe asthma had a high daily dose of corticosteroids, and their lower abundance of octanal when
30
31 compared to mild/moderate (Figure 4) may be linked to suppression of lipid peroxidation. Although
32
33 octanal may arise from other latent factors such as diet or age differences, this suppression is also
34
35 evident in a previous study investigating breath VOCs from patients with chronic obstructive
36
37 pulmonary disease (COPD) – which shares some clinical characteristics with severe asthma (such as
38
39 airflow obstruction, frequent exacerbations, and high doses of inhaled steroids) – where a lower
40
41 number of heavier VOCs were identified in comparison to non-COPD controls[46].
42
43
44
45

46
47 Similar volatile metabolic products, such as 2,6,11,15-tetramethyl-hexadecane and nonanal, have
48
49 also been found in breath VOC profiles for patients with lower respiratory tract infections [47]. This
50
51 may indicate the increased production of long chain unsaturated FAs due to a change in membrane
52
53 fluidity, or defective efferocytosis in asthmatic patients [48]. Additionally, unsaturated FAs have
54
55 been found not to induce cytokine release *in vitro* when compared to saturated FAs [49]. Shorter
56
57 methylated alkanes, such as 2-methylhexane, may arise as breakdown products. It was not possible
58
59
60

1
2
3 to NIST-match several methylated alkanes because many share similar mass fragments and retention
4
5 time, and therefore requires targeted analysis for accurate identification.
6
7

8 Sampling devices may introduce artefact VOCs depending on the type of material used, for example
9
10 gas sampling bag contaminants such as phenol and *N,N*-dimethylacetamide [36]. Both the latter
11
12 compounds were found in breath samples in this study, as breath was collected using Tedlar bags,
13
14 and their associated signals data were removed to prevent a confounding effect during statistical
15
16 analyses. Latent variables and confounders must be considered in multi-site studies, and it is
17
18 important to assess any variation due to patient factors, asthma subtype, and sampling sites.
19
20 Previous studies have shown variation in breath profiles due to geographical location[50,51], and
21
22 these findings are supported by our analysis, where we found significant differences in levels of
23
24 pentane and dodecane between Copenhagen and London sites. It is likely that the difference may
25
26 be due to exogenous environmental alkanes before sampling breath, but this potential
27
28 contamination is especially relevant as these VOCs may also be linked to lipid peroxidation. Variation
29
30 between sites can also occur from differences in site-specific sampling differences or patient
31
32 demographics such as age[52], or gender[53], as demonstrated using metabolomics on human
33
34 serum[54]. In the example shown for dodecane, levels for pre-school and school aged children were
35
36 increased compared to adults, and therefore age may be main contributing variable rather than a
37
38 method-related effect, as the majority of school-aged patients used the same method as adult
39
40 patients.
41
42
43
44
45
46

47 We found no VOCs strongly correlate with sample storage duration for up to 39 days, although we
48
49 show that benzene and toluene (the known exogenous compounds used for this test) show the
50
51 strongest correlation from the VOCs selected and decrease over time in breath samples stored less
52
53 than 40 days. However this may be attributed to sampling, storage, and analytical conditions, where
54
55 the presence of intrinsic and unidentifiable artefacts may have influenced other VOCs within
56
57 storage. It is also important to state that the acetone and isoprene are not adequately trapped on
58
59
60

1
2
3 Tenax GR sorbent material. The analysis carried out was a basic measurement, and we suggest using
4 quality assurance samples which include known breath compounds (stored and analysed alongside
5 breath samples) to measure variation by sample storage. Previous studies have investigated the
6 effect of sample storage on VOCs. Using e-nose devices to measure variation, van der Schee *et al.*
7 found no variation where breath samples were stored for up to two weeks[18]. Kang *et al.* further
8 investigated breath sample storage for a longer duration analysed using TD-GC-MS. They
9 recommended storage duration may be extended up to 1.5 months[55], however this was specific to
10 samples frozen at -80°C and therefore not comparable to other studies where samples were
11 refrigerated or stored at room temperature. In metabolomic experiments, -80 °C is recommended as
12 metabolites do not react with each other, however this is dependent on the type of metabolites
13 within a sample[4], and the time taken after freeze thaw and before thermal desorption.

14
15 In metabolomic studies, batch variation is intrinsic to sampling and analysis, especially for
16 longitudinal studies. Data were batch corrected using normalisation to the internal standard. In our
17 test we found the worst performing method was normalisation by an internal standard (acetone-*d*₆,
18 Figure 6a) unsuitable for the sorbent material used in our analysis, as its characteristics mean it is
19 outside the ideal Tenax GR capture range. This indicates significant “over-normalisation” of data,
20 thereby producing a false batch effect. It is important to monitor any batch variation effect to
21 prevent misclassification. A common batch correction method is to normalise samples by an internal
22 standard or scaling factors[27,56]. Other methods such as ComBat or retention time alignment may
23 also help to correct known batch variation[57].

24
25 Breath samples may be analysed centrally or on-site (i.e. near patient analysis), therefore it is useful
26 to assess intra-sample similarity between different instruments[58,59]. The expectation is
27 instruments should show similar results for the same sample, however the instrument mechanism
28 may overshadow any sample comparability. Analysis by GC-MS produces high resolution ion
29 fragment data. In contrast, e-nose data is limited to the response to several sensors. This is shown in
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Figure 7a, where GC-MS form additional clusters of data compared to the Lonestar instrument.
4
5 Regarding e-nose comparisons, although e-nose sensing mechanisms are different, we have shown
6
7 good superimposition (and therefore similarity) between Cyranose and Tor Vergata, where the
8
9 reactive layer of both sensors may have similar sensing properties despite their different technology
10
11 (in asthmatic patients). We have shown that the sensing mechanism may overshadow biological
12
13 differences, however the response signal may also account for environmental influences such as
14
15 humidity and temperatures, especially for those sensors with a polymer-based reactive layer which
16
17 may also show co-linearity in the sensor drift. This is especially important for multi-site studies as
18
19 clinical sites may have different parameters and limits for controlling temperature and humidity
20
21 levels. Furthermore, using e-nose sensors in multi-site studies may negatively affect reproducibility
22
23 for the same sensor or between different sensors. Additional controlled experiments are required to
24
25 elucidate the origin of sensor responses.
26
27
28

30 *4.2. Recommendations for future research*

31
32
33 Based on our data, we provide several recommendations for breath volatile analysis for large-scale
34
35 and multi-site studies from study initiation, sample collection, and sample analysis:
36
37

- 38 • To minimise variation between all sampling and analysis sites, quality assurance procedures
39
40 should be incorporated into a study design. Such measures may include instrument
41
42 calibration or maintenance schedule, quality monitoring, data auditing at predefined
43
44 intervals with open access to metadata, and regular staff training events.
45
46
- 47 • Sampling devices (e.g. phenol and DMAC from Tedlar bags) or instrument connections (e.g.
48
49 system leaks, loose connections) can be prone to contamination and therefore must be
50
51 regularly serviced and/or cleaned.
52
53
- 54 • Strict storage and transportation methods should be in place where samples are stored
55
56 consistently across sites without long-term storage. In addition, sample storage conditions,
57
58
59
60

1
2
3 dry-purging of breath samples, and transport parameters need to be defined within the
4
5 study design.
6

- 7 • Data from multiple instruments should be compared to investigate similarities and
8
9 differences in the analytical methods used.
10

11
12
13 Additional recommendations influenced from metabolomics-based literature useful for future
14
15 studies are as follows:
16

- 17 • Perform regular instrument proficiency testing and use quality control samples to assess
18
19 variation between instruments, as in previous studies[58,60].
20
- 21 • Consider data pre-treatment methods, for example, log transformation is not suitable for
22
23 high RSD or missing values, and Pareto-scaling can be sensitive to high sample variance[61],
24
25 a common occurrence in breath analysis[8].
26
- 27 • Consider using several models for multivariate analysis, especially for e-nose data analysis,
28
29 as shown by Gromski *et al.* and Leopold *et al.* [62,63].
30
- 31 • Breath samples analysed with GC-MS are known to contain high numbers of metabolites
32
33 that are collinear in nature, and any multivariate models must be optimised and validated
34
35 prior to reporting[2,64,65].
36
- 37 • Consider using multiple databases, as highlighted by Vinaixa *et al.*[66] and the use of
38
39 external standards, as recommended by the Metabolomic Standards Initiative [35,67], to
40
41 verify the identification of a VOC.
42
- 43 • For e-nose analysis consider decentralised usage of exchangeable sensor systems, together
44
45 with a central cloud-based analysis[68].
46
47
48
49
50
51

52 To summarise the main methodological considerations for breath sampling and analysis with focus
53
54 on large-scale and multi-site studies, a cause-and-effect diagram was produced, shown in Figure 8. A
55
56 checklist is also provided in the supplementary information (Table S4) for new multi-site and/or
57
58 longitudinal studies.
59
60

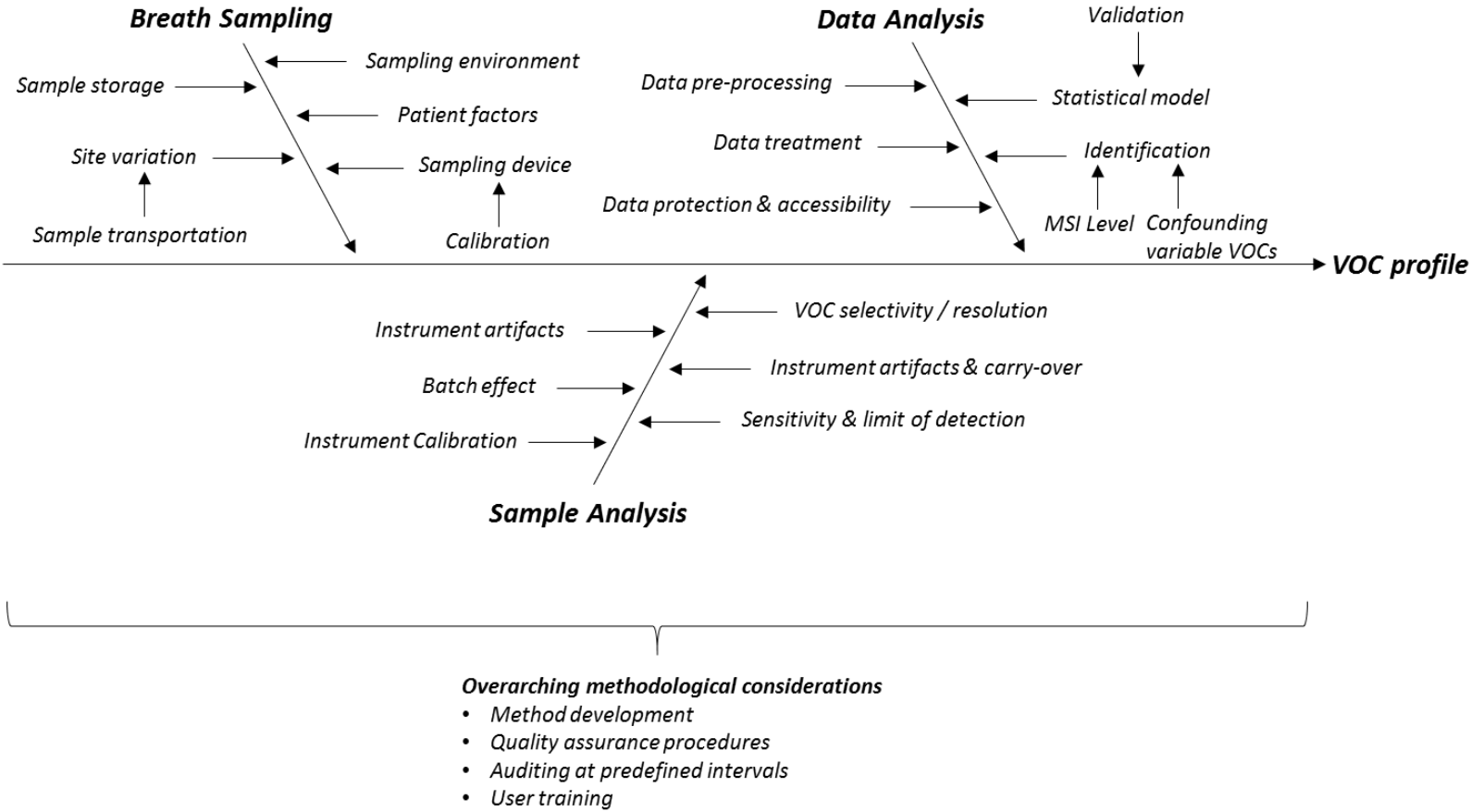


Figure 8. Cause-and-effect diagram highlighting the main methodological considerations for large-scale and multi-site studies, where MSI is Metabolomics Standards Initiative.

5. Conclusions

Exhaled breath analysis is a rapidly developing field aiming for diagnostic use within the clinical environment. In the near future we envisage personalised treatment of people with asthma based on molecular phenotypes and detected by point-of-care sensors. However, the requirement for large scale multi-centre studies will introduce new sources of variation, and in this work, we have explored several issues that may affect the results and interpretation of such as studies. Using examples from the recent U-BIOPRED severe asthma study, we have shown the importance of assessing variation which may arise between sites, patients, and instruments (which could be exploited to obtain complementary information), with the overall aim to reduce the risk of false discoveries when interpreting results. Finally, we also submit recommendations for researchers to consider if they are performing breath analysis using metabolomics approaches.

Acknowledgements

We would like to thank and acknowledge members of the U-BIOPRED consortium members (see Table S5), industrial partners, and contributors. U-BIOPRED has received funding from the Innovative Medicines Initiative (IMI) Joint Undertaking under grant agreement no. 115010, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007–2013) and European Federation of Pharmaceutical Industries and Associations (EFPIA) companies' in-kind contribution (www.imi.europa.eu). Additionally, we would also like to thank Lieuwe Bos (Academic Medical Centre, The Netherlands), Maxim Wilkinson (University of Manchester, United Kingdom), Tom Wenseleers (University of Leuven, Belgium), and Yuanyue Wang (Royal Philips B.V, The Netherlands) for their assistance with data pre-processing.

References

- [1] Wheelock C E, Goss V M, Balgoma D, Nicholas B, Brandsma J, Skipp P J, Snowden S, Burg D, D'Amico A, Horvath I, Chaiboonchoe A, Ahmed H, Ballereau S, Rossios C, Chung K F, Montuschi P, Fowler S J, Adcock I M, Postle A D, Dahlen S-E, Rowe A, Sterk P J, Auffray C and Djukanovic R 2013 Application of 'omics technologies to biomarker discovery in inflammatory lung diseases *Eur. Respir. J.* **42** 802–25
- [2] Broadhurst D I and Kell D B 2007 Statistical strategies for avoiding false discoveries in metabolomics and related experiments *Metabolomics* **2** 171–96
- [3] Dunn W B, Wilson I D, Nicholls A W and Broadhurst D 2012 The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans *Bioanalysis* **4** 2249–64
- [4] Dunn W B, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, Brown M, Knowles J D, Halsall A, Haselden J N, Nicholls A W, Wilson I D, Kell D B and Goodacre R 2011 Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **6** 1060–83
- [5] Fiehn O, Robertson D, Griffin J, van der Werf M, Nikolau B, Morrison N, Sumner L W, Goodacre R, Hardy N W, Taylor C, Fostel J, Kristal B, Kaddurah-Daouk R, Mendes P, van Ommen B, Lindon J C and Sansone S-A 2007 The metabolomics standards initiative (MSI) *Metabolomics* **3** 175–8
- [6] Rocca-Serra P, Salek R M, Arita M, Correa E, Dayalan S, Gonzalez-Beltran A, Ebbels T, Goodacre R, Hastings J, Haug K, Koulman A, Nikolski M, Oresic M, Sansone S-A, Schober D, Smith J, Steinbeck C, Viant M R and Neumann S 2016 Data standards can boost metabolomics research, and if there is a will, there is a way *Metabolomics* **12** 14
- [7] Broadhurst D, Goodacre R, Reinke S N, Kuligowski J, Wilson I D, Lewis M R and Dunn W B 2018 Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies *Metabolomics* **14** 1–17
- [8] Xu Y, Fowler S J, Bayat A and Goodacre R 2013 Chemometrics models for overcoming high between subject variability: applications in clinical metabolic profiling studies *Metabolomics* **10** 375–85
- [9] Horváth I, Barnes P J, Högman M, Olin A, Amann A, Antus B, Baraldi E, Bikov A, Boots A W, Bos L D, Brinkman P, Bucca C, Carpagnano G E, Corradi M and Cristescu S 2017 A European Respiratory Society technical standard : exhaled biomarkers in lung disease *Eur Respir J* **49** 1–26
- [10] Herbig J and Beauchamp J 2014 Towards standardization in the analysis of breath gas volatiles. *J. Breath Res.* **8** 037101
- [11] Lourenço C and Turner C 2014 Breath Analysis in Disease Diagnosis: Methodological Considerations and Applications *Metabolites* **4** 465–98
- [12] Bikov A, Paschalaki K, Logan-Sinclair R, Horváth I, Kharitonov S a, Barnes P J, Usmani O S and Paredi P 2013 Standardised exhaled breath collection for the measurement of exhaled volatile organic compounds by proton transfer reaction mass spectrometry. *BMC Pulm. Med.* **13** 43

- 1
2
3 [13] Guallar-hoyas C, Turner M a and Wilson I D 2012 A workflow for the
4 metabolomic/metabonomic investigation of exhaled breath using thermal desorption GC–MS
5 *Bioanalysis* **4** 2227–37
6
- 7 [14] Shaw D E, Sousa A R, Fowler S J, Fleming L J, Roberts G, Corfield J, Pandis I, Bansal A T, Bel E
8 H, Auffray C, Compton C H, Bisgaard H, Bucchioni E, Caruso M, Chanez P, Dahlén B, Dahlen S-
9 E, Dyson K, Frey U, Geiser T, Gerhardsson de Verdier M, Gibeon D, Guo Y, Hashimoto S,
10 Hedlin G, Jeyasingham E, Hekking P-P W, Higenbottam T, Horváth I, Knox A J, Krug N,
11 Erpenbeck V J, Larsson L X, Lazarinis N, Matthews J G, Middelveld R, Montuschi P, Musial J,
12 Myles D, Pahus L, Sandström T, Seibold W, Singer F, Strandberg K, Vestbo J, Vissing N, von
13 Garnier C, Adcock I M, Wagers S, Rowe A, Howarth P, Wagener A H, Djukanovic R, Sterk P J
14 and Chung K F 2015 Clinical and inflammatory characteristics of the European U-BIOPRED
15 adult severe asthma cohort *Eur. Respir. J.* **46** 1308–1321
16
17
- 18 [15] Fleming L, Murray C, Bansal A T, Hashimoto S, Bisgaard H, Bush A, Frey U, Hedlin G, Singer F,
19 Van Aalderen W M, Vissing N H, Zolkipli Z, Selby A, Fowler S, Shaw D, Chung K F, Sousa A R,
20 Wagers S, Corfield J, Pandis I, Rowe A, Formaggio E, Sterk P J and Roberts G 2015 The burden
21 of severe asthma in childhood and adolescence: Results from the paediatric U-BIOPRED
22 cohorts *Eur. Respir. J.* **46** 1322–33
23
- 24 [16] Bel E H, Sousa A, Fleming L, Bush A, Chung K F, Versnel J, Wagener A H, Wagers S S, Sterk P J
25 and Compton C H 2011 Diagnosis and definition of severe refractory asthma: an international
26 consensus statement from the Innovative Medicine Initiative (IMI) *Thorax* **66** 910–7
27
- 28 [17] Dragonieri S, Schot R, Mertens B J A, Le Cessie S, Gauw S A, Spanevello A, Resta O, Willard N
29 P, Vink T J, Rabe K F, Bel E H and Sterk P J 2007 An electronic nose in the discrimination
30 of patients with asthma and controls *J. Allergy Clin. Immunol.* **120** 856–62
31
- 32 [18] van der Schee M P, Fens N, Brinkman P, Bos L D, Angelo M D, Nijssen T M, Raabe R, Knobel H
33 H, Vink T J and Sterk P J 2013 Effect of transportation and storage using sorbent tubes of
34 exhaled breath samples on diagnostic accuracy of electronic nose analysis *J. Breath Res.* **7**
35 16002
36
- 37 [19] Lewis N S 2004 Comparisons between mammalian and artificial olfaction based on arrays of
38 carbon black-polymer composite vapor detectors *Acc. Chem. Res.* **37** 663–72
39
- 40 [20] Di Natale C, Paolesse R and D’Amico A 2007 Metalloporphyrins based artificial olfactory
41 receptors *Sensors Actuators, B Chem.* **121** 238–46
42
- 43 [21] Bos L D J, van Walree I C, Kolk A H J, Janssen H-G, Sterk P J and Schultz M J 2013 Alterations
44 in exhaled breath metabolite-mixtures in two rat models of lipopolysaccharide-induced lung
45 injury *J. Appl. Physiol.* **115** 1487–95
46
- 47 [22] Arasaradnam R P, Mcfarlane M J, Ryan-Fisher C, Westenbrink E, Hodges P, Thomas M G,
48 Chambers S, O’Connell N, Bailey C, Harmston C, Nwokolo C U, Bardhan K D and Covington J A
49 2014 Detection of colorectal cancer (CRC) by urinary volatile organic compound analysis *PLoS*
50 *One* **9**
51
- 52 [23] Fens N, de Nijs S B, Peters S, Dekker T, Knobel H H, Vink T J, Willard N P, Zwinderman A H,
53 Krouwels F H, Janssen H-G, Lutter R and Sterk P J 2011 Exhaled air molecular profiling in
54 relation to inflammatory subtype and activity in COPD *Eur. Respir. J.* **38** 1301–9
55
- 56 [24] Szalma S, Koka V, Khasanova T and Perakslis E D 2010 Effective knowledge management in
57 translational medicine *J. Transl. Med.* **8** 1–9
58
- 59 [25] Smith C A, Want E J, O ’maille G, Abagyan R and Siuzdak G 2006 XCMS: Processing Mass
60

- 1
2
3 Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and
4 Identification *Anal. Chem.* **78** 779–87
5
- 6 [26] Oksanen J 2015 Multivariate analysis of ecological communities in R: vegan tutorial *R Doc.* 43
7
- 8 [27] Wehrens R, Hageman J A, van Eeuwijk F, Kooke R, Flood P J, Wijnker E, Keurentjes J J B,
9 Lommen A, van Eekelen H D L M, Hall R D, Mumm R and de Vos R C H 2016 Improved batch
10 correction in untargeted MS-based metabolomics *Metabolomics* **12**
11
- 12 [28] Van Vliet D, Smolinska A, Van Horck M, Jöbsis Q, Rosias P, Muris J, Dallinga J, Van Schooten F-
13 J and Dompeling E 2017 Can exhaled volatile organic compounds predict asthma
14 exacerbations in children? *J. Breath Res.* **11**
15
- 16 [29] Smolinska A, Klaassen E M M, Dallinga J W, van de Kant K D G, Jobsis Q, Moonen E J C, van
17 Schayck O C P, Dompeling E and van Schooten F J 2014 Profiling of Volatile Organic
18 Compounds in Exhaled Breath As a Strategy to Find Early Predictive Signatures of Asthma in
19 Children *PLoS One* **9** e95668
20
- 21 [30] Dallinga J W, Robroeks C M H H T, Van Berkel J J B N, Moonen E J C, Godschalk R W L, Jöbsis
22 Q, Dompeling E, Wouters E F M and Van Schooten F J 2010 Volatile organic compounds in
23 exhaled breath as a diagnostic tool for asthma in children *Clin. Exp. Allergy* **40** 68–76
24
- 25 [31] Robroeks C M, Van Berkel J J, Jöbsis Q, Van Schooten F J, Dallinga J W, Wouters E F and
26 Dompeling E 2013 Exhaled volatile organic compounds predict exacerbations of childhood
27 asthma in a 1-year prospective study *Eur. Respir. J.* **42** 98–106
28
- 29 [32] Ibrahim B, Basanta M, Cadden P, Singh D, Douce D, Woodcock A and Fowler S J 2011 Non-
30 invasive phenotyping using exhaled volatile organic compounds in asthma *Thorax* **66** 804–9
31
- 32 [33] Meyer N, Dallinga J W, Nuss S J, Moonen E J C, van Berkel J J B N, Akdis C, van Schooten F J
33 and Menz G 2014 Defining adult asthma endotypes by clinical features and patterns of
34 volatile organic compounds in exhaled air. *Respir. Res.* **15** 136
35
- 36 [34] Gonzalez F R and Nardillo A M 1999 Retention index in temperature-programmed gas
37 chromatography *J. Chromatogr. A* **842** 29–49
38
- 39 [35] Sumner L W, Amberg A, Barrett D, Beale M H, Beger R, Daykin C A, Fan T W-M, Fiehn O,
40 Goodacre R, Griffin J L, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane A N, Lindon
41 J C, Marriott P, Nicholls A W, Reily M D, Thaden J J and Viant M R 2007 Proposed minimum
42 reporting standards for chemical analysis *Metabolomics* **3** 211–21
43
- 44 [36] Beauchamp J, Herbig J, Gutmann R and Hansel A 2008 On the use of Tedlar® bags for breath-
45 gas sampling and analysis *J. Breath Res.* **2** 046001
46
- 47 [37] Gromski P S, Xu Y, Kotze H L, Correa E, Ellis D I, Armitage E G, Turner M L and Goodacre R
48 2014 Influence of missing values substitutes on multivariate analysis of metabolomics data.
49 *Metabolites* **4** 433–52
50
- 51 [38] Stekhoven D J and Bühlmann P 2012 Missforest-Non-parametric missing value imputation for
52 mixed-type data *Bioinformatics* **28** 112–8
53
- 54 [39] Xu Y, Dixon S J, Brereton R G, Soini H A, Novotny M V., Trebesius K, Bergmaier I, Oberzaucher
55 E, Grammer K and Penn D J 2007 Comparison of human axillary odour profiles obtained by
56 gas chromatography/mass spectrometry and skin microbial profiles obtained by denaturing
57 gradient gel electrophoresis using multivariate pattern recognition *Metabolomics* **3** 427–37
58
- 59 [40] AlRabiah H, Xu Y, Rattray N J W, Vaughan A a, Gibreel T, Sayqal A, Upton M, Allwood J W and
60 Goodacre R 2014 Multiple metabolomics of uropathogenic *E. coli* reveal different information

- 1
2
3 content in terms of metabolic potential compared to virulence factors. *Analyst* **139** 4193–9
4
- 5 [41] Legendre P and Anderson M J 1999 Distance-Based Redundancy Analysis: Testing
6 Multispecies Responses in Multifactorial Ecological Experiments *Ecol. Monogr.* **69** 1–24
7
- 8 [42] Andrade J M, Gómez-Carracedo M P, Krzanowski W and Kubista M 2004 Procrustes rotation
9 in analytical chemistry, a tutorial *Chemom. Intell. Lab. Syst.* **72** 123–32
10
- 11 [43] Peres-Neto P R and Jackson D A 2001 How well do multivariate data sets match? The
12 advantages of a procrustean superimposition approach over the Mantel test *Oecologia* **129**
13 169–78
14
- 15 [44] Reynolds J C, Blackburn G J, Guallar-Hoyas C, Moll V H, Bocos-Bintintan V, Kaur-Atwal G,
16 Howdle M D, Harry E L, Brown L J, Creaser C S and Thomas C L P 2010 Detection of volatile
17 organic compounds in breath using thermal desorption electrospray ionization-ion mobility-
18 mass spectrometry. *Anal. Chem.* **82** 2139–44
19
- 20 [45] Ahmed W M, Lawal O, Nijsen T M, Goodacre R and Fowler S J 2017 Exhaled Volatile Organic
21 Compounds of Infection: A Systematic Review *ACS Infect. Dis.* **3** 695–710
22
- 23 [46] Basanta M, Ibrahim B, Dockry R, Douce D, Morris M, Singh D, Woodcock A and Fowler S J
24 2012 Exhaled volatile organic compounds for phenotyping chronic obstructive pulmonary
25 disease: a cross-sectional study *Respir. Res.* **13** 72
26
- 27 [47] Fowler S J, Basanta-Sanchez M, Xu Y, Goodacre R and Dark P M 2015 Surveillance for lower
28 airway pathogens in mechanically ventilated patients by metabolomic analysis of exhaled
29 breath: a case-control study *Thorax* **70** 320–5
30
- 31 [48] Grabiec A M, Denny N, Doherty J A, Happonen K E, Hankinson J, Connolly E, Fife M E,
32 Fujimori T, Fujino N, Goenka A, Holden S, Tavernier G, Shah R, Cook P C, MacDonald A S,
33 Niven R M, Dahlbäck B, Fowler S J, Simpson A and Hussell T 2017 Diminished airway
34 macrophage expression of the Axl receptor tyrosine kinase is associated with defective
35 efferocytosis in asthma *J. Allergy Clin. Immunol.*
36
- 37 [49] Gupta S, Knight A G, Gupta S, Keller J N and Bruce-Keller A J 2012 Saturated long-chain fatty
38 acids activate inflammatory signaling in astrocytes *J. Neurochem.* **120** 1060–71
39
- 40 [50] Amal H, Leja M, Broza Y Y, Tisch U, Funka K, Liepniece-Karele I, Skapars R, Xu Z-Q, Liu H and
41 Haick H 2013 Geographical variation in the exhaled volatile organic compounds. *J. Breath Res.*
42 **7** 047102
43
- 44 [51] Gaida A, Holz O, Nell C, Schuchardt S, Lavae-Mokhtari B, Kruse L, Boas U, Langejuergen J,
45 Allers M, Zimmermann S, Vogelmeier C, Koczulla A R and Hohlfeld J M 2016 A dual center
46 study to compare breath volatile organic compounds from smokers and non-smokers with
47 and without COPD *J. Breath Res.* **10**
48
- 49 [52] Phillips M, Cataneo R N, Greenberg J, Gunawardena R, Naidu A and Rahbari-Oskoui F 2000
50 Effect of age on the breath methylated alkane contour, a display of apparent new markers of
51 oxidative stress *J. Lab. Clin. Med.* **136** 243–9
52
- 53 [53] Das M K, Bishwal S C, Das A, Dabral D, Varshney A, Badireddy V K and Nanda R 2014
54 Investigation of gender-specific exhaled breath volatome in humans by GCxGC-TOF-MS *Anal.*
55 *Chem.* **86** 1229–37
56
- 57 [54] Dunn W B, Lin W, Broadhurst D, Begley P, Brown M, Zelena E, Vaughan A A, Halsall A,
58 Harding N, Knowles J D, Francis-McIntyre S, Tseng A, Ellis D I, O'Hagan S, Aarons G, Benjamin
59 B, Chew-Graham S, Moseley C, Potter P, Winder C L, Potts C, Thornton P, McWhirter C, Zubair
60

- 1
2
3 M, Pan M, Burns A, Cruickshank J K, Jayson G C, Purandare N, Wu F C W, Finn J D, Haselden J
4 N, Nicholls A W, Wilson I D, Goodacre R and Kell D B 2015 Molecular phenotyping of a UK
5 population: defining the human serum metabolome *Metabolomics* **11** 9–26
6
- 7 [55] Kang S and Paul Thomas C L 2016 How long may a breath sample be stored for at -80°C ? A
8 study of the stability of volatile organic compounds trapped onto a mixed Tenax:Carbograph
9 trap adsorbent bed from exhaled breath *J. Breath Res.* **10** 026011
10
- 11 [56] Katajamaa M and Orešič M 2007 Data processing for mass spectrometry-based
12 metabolomics *J. Chromatogr. A* **1158** 318–28
13
- 14 [57] Leek J T, Johnson W E, Parker H S, Jaffe A E and Storey J D 2012 The SVA package for
15 removing batch effects and other unwanted variation in high-throughput experiments
16 *Bioinformatics* **28** 882–3
17
- 18 [58] Nakhleh M K, Amal H, Jeries R, Broza Y Y, Aboud M, Gharra A, Ivgi H, Khatib S, Badarneh S,
19 Har-Shai L, Glass-Marmor L, Lejbkovicz I, Miller A, Badarny S, Winer R, Finberg J, Cohen-
20 Kaminsky S, Perros F, Montani D, Girerd B, Garcia G, Simonneau G, Nakhoul F, Baram S, Salim
21 R, Hakim M, Gruber M, Ronen O, Marshak T, Doweck I, Nativ O, Bahouth Z, Shi D, Zhang W,
22 Hua Q, Pan Y, Tao L, Liu H, Karban A, Koifman E, Rainis T, Skapars R, Sivins A, Ancans G,
23 Liepniece-Karele I, Kikuste I, Lasina I, Tolmanis I, Johnson D, Millstone S Z, Fulton J, Wells J W,
24 Wilf L H, Humbert M, Leja M, Peled N and Haick H 2017 Diagnosis and Classification of 17
25 Diseases from 1404 Subjects via Pattern Analysis of Exhaled Molecules *ACS Nano* **11** 112–5
26
- 27 [59] Chin S, Romano A, Doran S L F and Hanna G B 2018 Cross-platform mass spectrometry
28 annotation in breathomics of oesophageal-gastric cancer *Sci. Rep.* **8** 1–10
29
- 30 [60] van Oort P M P, Nijssen T, Weda H, Knobel H, Dark P, Felton T, Rattray N J W, Lawal O, Ahmed
31 W, Portsmouth C, Sterk P J, Schultz M J, Zakharkina T, Artigas A, Povoia P, Martin-Loeches I,
32 Fowler S J and Bos L D J 2017 BreathDx – molecular analysis of exhaled breath as a diagnostic
33 test for ventilator-associated pneumonia: protocol for a European multicentre observational
34 study *BMC Pulm. Med.* **17** 1
35
- 36 [61] van den Berg R A, Hoefsloot H C J, Westerhuis J A, Smilde A K and van der Werf M J 2006
37 Centering, scaling, and transformations: improving the biological information content of
38 metabolomics data. *BMC Genomics* **7** 142
39
- 40 [62] Gromski P S, Correa E, Vaughan A A, Wedge D C, Turner M L and Goodacre R 2014 A
41 comparison of different chemometrics approaches for the robust classification of electronic
42 nose data *Anal. Bioanal. Chem.* **406** 7581–90
43
- 44 [63] Leopold J H, Bos L D J, Sterk P J, Schultz M J, Fens N, Horvath I, Bikov A, Montuschi P, Di
45 Natale C, Yates D H and Abu-Hanna A 2015 Comparison of classification methods in breath
46 analysis by electronic nose. *J. Breath Res.* **9** 046002
47
- 48 [64] Smolinska A, Hauschild A-C, Fijten R R R, Dallinga J W, Baumbach J and van Schooten F J 2014
49 Current breathomics—a review on data pre-processing techniques and machine learning in
50 metabolomics breath analysis *J. Breath Res.* **8** 027105
51
- 52 [65] Gromski P S, Muhamadali H, Ellis D I, Xu Y, Correa E, Turner M L and Goodacre R 2015 A
53 tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of
54 convenience or a shotgun wedding *Anal. Chim. Acta* **879** 10–23
55
- 56 [66] Vinaixa M, Schymanski E L, Neumann S, Navarro M, Salek R M and Yanes O 2016 Mass
57 spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future
58 prospects *TrAC - Trends Anal. Chem.* **78** 23–35
59
60

- 1
2
3 [67] Goodacre R, Broadhurst D, Smilde A K, Kristal B S, Baker J D, Beger R, Bessant C, Connor S,
4 Capuani G, Craig A, Ebbels T, Kell D B, Manetti C, Newton J, Paternostro G, Somorjai R,
5 Sjöström M, Trygg J and Wulfert F 2007 Proposed minimum reporting standards for data
6 analysis in metabolomics *Metabolomics* **3** 231–41
7
- 8 [68] de Vries R, Dagelet Y W F, Spoor P, Snoey E, Jak P M C, Brinkman P, Dijkers E, Bootsma S K,
9 Elskamp F, de Jongh F H C, Haarman E G, in 't Veen J C C ., Maitland-van der Zee A-H and Sterk
10 P J 2018 Clinical and inflammatory phenotyping by breathomics in chronic airway diseases
11 irrespective of the diagnostic label *Eur. Respir. J.* **51** 1701817
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
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60